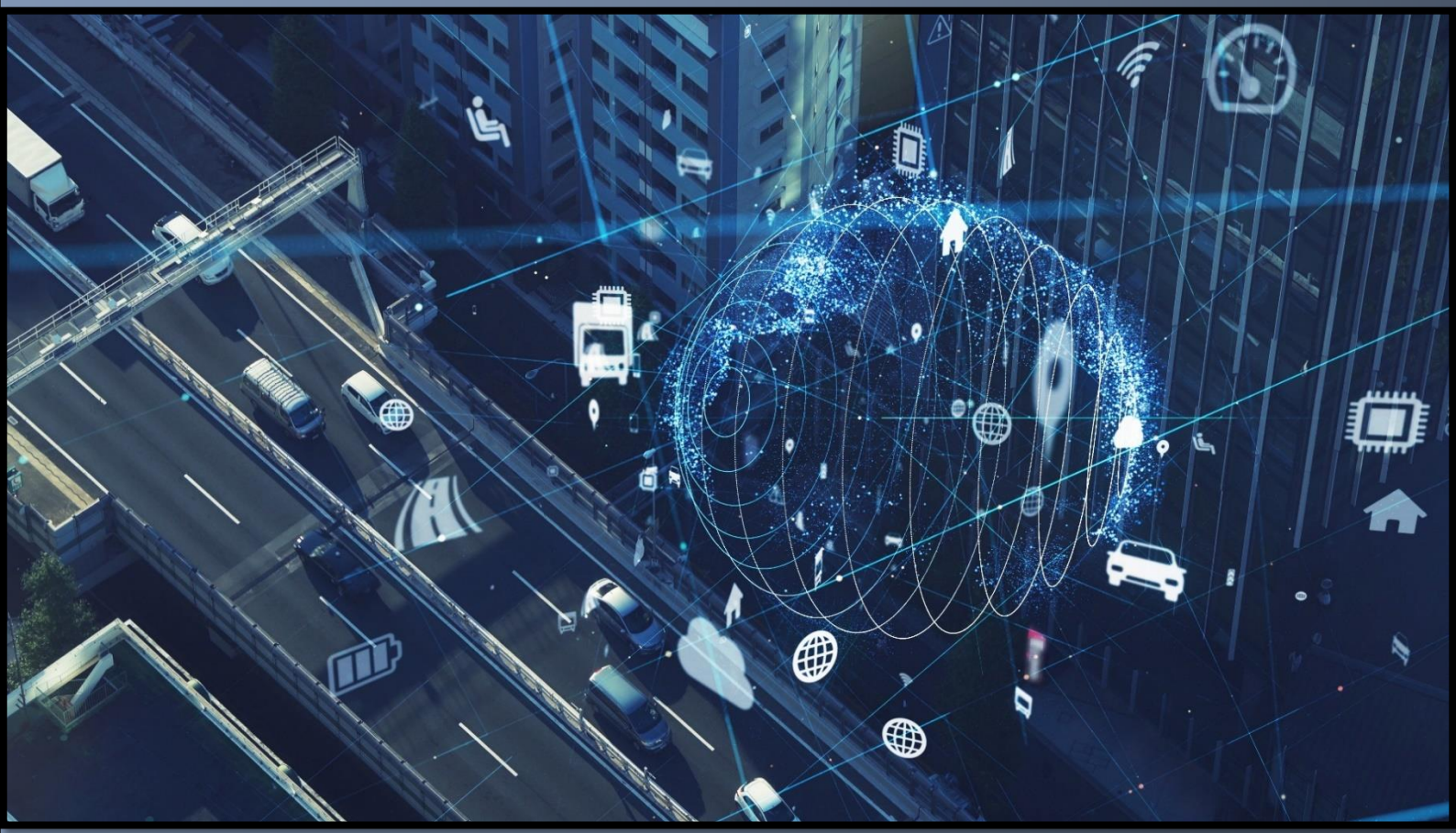
ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
Σχολή Πολιτικών Μηχανικών
Τομέας Μεταφορών και Συγκοινωνιακής Υποδομής

Ανάλυση Μηχανικής Μάθησης ανισόρροπων δεδομένων τηλεματικής για την πρόβλεψη της συμπεριφοράς του οδηγού



Κωστόπουλος Αντώνης

Διπλωματική Εργασία

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

Αθήνα, Νοέμβριος 2022

Ευχαριστίες

Σε αυτό το σημείο, θα ήθελα να αποτυπώσω τις σκέψεις και τα συναισθήματά μου, με την συγγραφή της Διπλωματικής μου εργασίας και την ολοκλήρωση των σπουδών μου. Μια διαδρομή γεμάτη με ποικίλα και ανάμεικτα συναισθήματα, με ατέρμονες δυσκολίες και εμπόδια και όμως μεγάλη ικανοποίηση και αγαλλίαση, εν τέλει.

Αρχικά, θα ήθελα να ευχαριστήσω θερμά τον κ. Γιώργο Γιαννή, Καθηγητή της Σχολής Πολιτικών Μηχανικών ΕΜΠ, τόσο για την εμπιστοσύνη που μου έδειξε με την ανάθεση της παρούσας εργασίας και την καθοδήγησή του σε αυτήν, όσο και για τα γνωστικά εφόδια που μου παρείχε καθ' όλη την διάρκεια των σπουδών μου και για τα ωραία του λόγια που θα θυμάμαι για πάντα.

Ένα μεγάλο ευχαριστώ θα ήθελα να απευθύνω και στον Δρ. Χρήστο Κατρακάζα για την υποστήριξη και καθοδήγησή του, για τις συμβουλές, τις γνώσεις και τον επαγγελματισμό του και για την εν γένει συνεργασία, σε όλη τη διάρκεια της εκπόνησης της εργασίας.

Ένα τεράστιο ευχαριστώ στους γονείς μου και την αδερφή μου Ήρα, για την υποστήριξη, τα εφόδια, τις συμβουλές τους και την παρότρυνσή τους και ιδιαίτερα στην μητέρα μου, Μαρία, που με υπέμεινε και που χωρίς αυτήν δεν θα ήταν τίποτα εφικτό.

Αμέριστο ευχαριστώ στους συνοδοιπόρους μου σε αυτή τη διαδρομή Χρήστο, Λάμπρο, Άννα, Θοδωρή και Στέλιο, που με βοήθησαν ειλικρινά στην προσωπική και ακαδημαϊκή μου εξέλιξη όσο κανείς. Τεράστιο ευχαριστώ οφείλω και στους παιδικούς μου φίλους για την ακρόαση, την κατανόηση και τις συμβουλές τους.

Κλείνοντας, θα ήθελα να αναφερθώ σε όλο το ερευνητικό προσωπικό του Τομέα Μεταφορών και Συγκοινωνιακής Υποδομής, εκφράζοντας τον θαυμασμό και τις ειλικρινείς μου ευχαριστίες για το αξιέπαινο ακαδημαϊκό τους έργο, μέσα από πάρα πολλές ώρες αφανούς εργασίας και για την ανυπολόγιστη συνεισφορά τους στην εξέλιξη της επιστήμης. Χωρίς το δικό τους έργο, η ολοκλήρωση της παρούσας εργασίας θα ήταν δυσεπίτευκτη.

Αθήνα, Νοέμβριος 2022

Αντώνης Κωστόπουλος

Ανάλυση Μηχανικής Μάθησης ανισόρροπων δεδομένων τηλεματικής για την πρόβλεψη της συμπεριφοράς του οδηγού

Κωστόπουλος Αντώνης

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

Σύνοψη

Αντικείμενο της παρούσας Διπλωματικής Εργασίας αποτελεί η ανάλυση δεδομένων φυσικής οδήγησης (Naturalistic Driving Data – NDD) για την ταξινόμηση και πρόβλεψη της οδικής συμπεριφοράς και των σοβαρών περιστατικών (harsh events), με τεχνικές Μηχανικής Μάθησης (Machine Learning). Αξιοποιήθηκαν δεδομένα τηλεματικής της εταιρίας OSeven, για την ταξινόμηση και πρόβλεψη της Οδικής συμπεριφοράς, με τη χρήση των δεικτών Επικίνδυνης Οδικής συμπεριφοράς απότομων επιταχύνσεων και επιβραδύνσεων σε αναγωγή 100 χιλιομέτρων διαδρομής. Πιο συγκεκριμένα, επιδιώκεται ο προσδιορισμός του βαθμού επιρροής των οδικών δεδομένων στις καταστάσεις εμφάνισης απότομων περιστατικών, μέσω της διαδικασίας Επιλογής Χαρακτηριστικών (Feature Selection) και η ταξινόμηση των απότομων επιταχύνσεων (harsh accelerations) και απότομων επιβραδύνσεων (harsh brakings) σε δύο επίπεδα ασφαλείας μέσα από τεχνικές Μηχανικής Μάθησης. Η ομαδοποίηση με K-means κατέδειξε ότι οι οδηγοί με περισσότερες από 48 απότομες επιταχύνσεις και 45 απότομες επιβραδύνσεις ανά 100χλμ. οδήγησης εμφάνισαν την πιο επικίνδυνη συμπεριφορά. Τα αποτελέσματα της εργασίας ανέδειξαν την συνολική διανυθείσα απόσταση διαδρομής ως την μεταβλητή με την μεγαλύτερη επιρροή σε απότομα περιστατικά, ενώ τις καλύτερες μετρικές αξιολογήσεις ταξινόμησης σε κλάσεις για το συγκεκριμένο πρόβλημα Μη Ισορροπημένης Μάθησης έδωσαν οι αλγόριθμοι Gradient Boosting και Multilayered Perceptrons, με αξιόλογες επιδόσεις για την εξαγωγή χρήσιμων συμπερασμάτων σχετικά με τα επίπεδα ασφάλειας της Οδικής συμπεριφοράς.

Λέξεις κλειδιά: ανάλυση οδικής συμπεριφοράς, ταξινόμηση οδικής συμπεριφοράς, δεδομένα φυσικής οδήγησης, πρόβλεψη ατυχημάτων, μοντέλο παλινδρόμησης, μοντέλο ταξινόμησης, μηχανική μάθηση, μη ισορροπημένη μάθηση, πολυεπίπεδα δεδομένα, σημαντικότητα χαρακτηριστικών, απότομα περιστατικά, απότομες επιταχύνσεις, απότομες επιβραδύνσεις, Τεχνητά Νευρωνικά Δίκτυα, Βαθιά Μάθηση

Imbalanced learning analysis for driving behaviour prediction using naturalistic driving data

Kostopoulos Antonis

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

Abstract

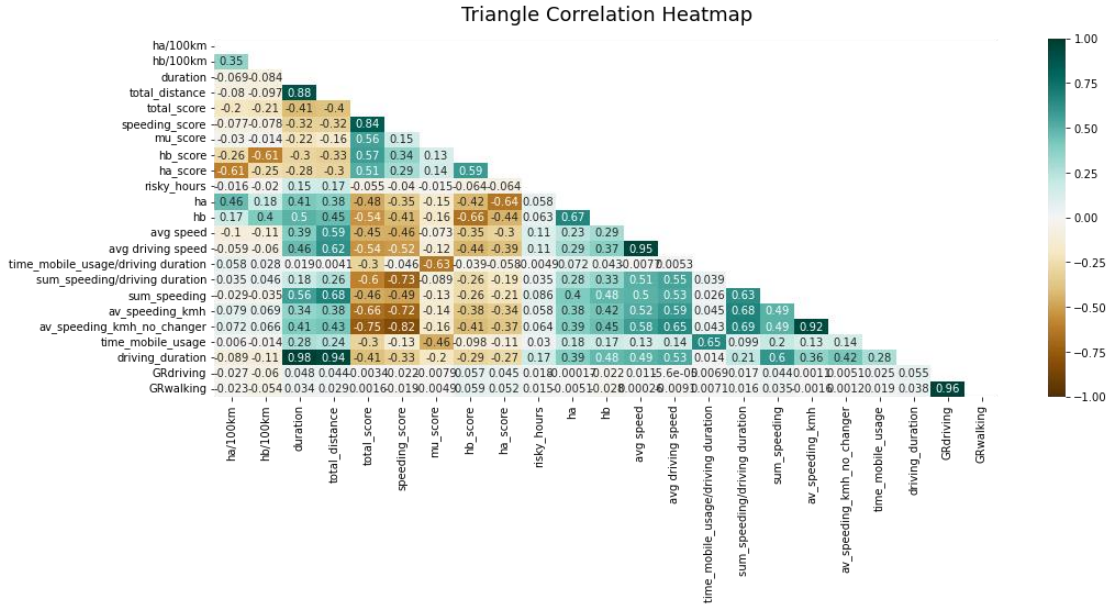
The objective of this Diploma Thesis is the exploitation of imbalanced learning for the task of classifying and predicting driving behaviour and harsh events, using naturalistic driving data. Data was collected through the telematics company OSeven, in order to classify and predict driving behaviour in terms of harsh accelerations and brakings occurences. More precisely, this thesis intends to determine the most crucial predictors for the occurrence of harsh events, through a feature selection process and to identify two safety levels for harsh accelerations and brakings using Machine Learning techniques. K-means clustering revealed that users with more than 48 harsh accelerations and more than 45 harsh brakings per 100 km of driving were deemed the most dangerous. The imbalanced classification results showcased that the total driving distance was the more impactful variable to harsh events occurence, whilst the best techniques for this particular imbalanced learning process, were achieved by Gradient Boosting and Multilayered Perceptrons algorithms.

Key words: driver behaviour analysis, driver behaviour classification, naturalistic driving data, dangerous driving prediction, regression model, classification model, machine learning, imbalanced learning, contextual data, feature importance, harsh accelerations, harsh brakings, harsh events, Artificial Neural Networks, Deep Learning

Περίληψη

Στόχος της παρούσας Διπλωματικής Εργασίας είναι ο εντοπισμός των σημαντικότερων παραγόντων επιρροής στην εμφάνιση Επικίνδυνης Οδικής συμπεριφοράς και η ανάπτυξη αποτελεσματικών μοντέλων ταξινόμησης και πρόβλεψης της Οδικής συμπεριφοράς, αξιοποιώντας τεχνικές Μηχανικής Μάθησης και Νευρωνικών Δικτύων. Τα οδικά δεδομένα που υπέστησαν επεξεργασία παραχωρήθηκαν από την εταιρία OSeven Telematics© και συλλέχθηκαν σε πραγματικές οδικές συνθήκες μέσω των κινητών τηλεφώνων των οδηγών. Στο σύνολό τους τα δεδομένα που αναλύθηκαν ήταν κυκλοφοριακής φύσης και δείκτες κίνησης και οδικής συμπεριφοράς. Προτού ξεκινήσει η ανάλυση των δεδομένων, καθορίστηκαν, ως αντιπροσωπευτικοί δείκτες Επικίνδυνης Οδικής συμπεριφοράς, οι απότομες επιταχύνσεις και επιβραδύνσεις που πραγματοποιεί ο οδηγός σε αναγωγή 100 χιλιομέτρων, οι οποίες και αποτέλεσαν τις εξαρτημένες μεταβλητές της έρευνας, με την ανάπτυξη των μοντέλων να πραγματοποιείται και για τις δύο μεταβλητές ξεχωριστά. Η ανάλυση εφαρμόστηκε με την βοήθεια της γλώσσας προγραμματισμού Python, σε προγραμματιστικό περιβάλλον Jupyter Notebook και Google Colab.

Στο πρώτο μέρος της ανάλυσης, επιχειρήθηκε ο εντοπισμός των σημαντικότερων παραγόντων στην εμφάνιση απότομων περιστατικών, μέσω της διαδικασίας Επιλογής Χαρακτηριστικών. Η διαδικασία αυτή περιλάμβανε τον υπολογισμό του συντελεστή συσχέτισης Pearson των εξαρτημένων μεταβλητών με τα υπόλοιπα οδικά δεδομένα και την διεργασία Σημαντικότητας Χαρακτηριστικών (Feature Importance), μέσω ανάπτυξης μοντέλων Παλινδρομήσεων για να ποσοτικοποιηθεί ο βαθμός επιρροής τους, με τεχνικές Μηχανικής Μάθησης. Τα μοντέλα Παλινδρομήσεων που αναπτύχθηκαν ήταν τέσσερα. Συγκεκριμένα, αναπτύχθηκαν Γραμμικές Παλινδρομήσεις, Παλινδρομήσεις Δένδρων Απόφασης, Τυχαίων Δασών, Ακραίας Ενίσχυσης Κλίσης και Γραμμικές Παλινδρομήσεις Μηχανών Διανυσμάτων Υποστήριξης και αξιολογήθηκαν, βάσει του συντελεστή προσδιορισμού τους R^2. Τα συνδυαστικά αποτελέσματα των Παλινδρομήσεων και του συντελεστή συσχέτισης ανέδειξαν ως σημαντικότερες μεταβλητές τις εξής: η διανυθείσα απόσταση, η συνολική διάρκεια της οδήγησης εν κινήσει, η μέση ταχύτητα οδήγησης, το σκορ χρήσης κινητού τηλεφώνου και το σκορ υπέρβασης ορίου ταχύτητας.



Γράφημα Περίληψης 1: *Συντελεστής συσχέτισης Pearson των εξετασθέντων δεδομένων*

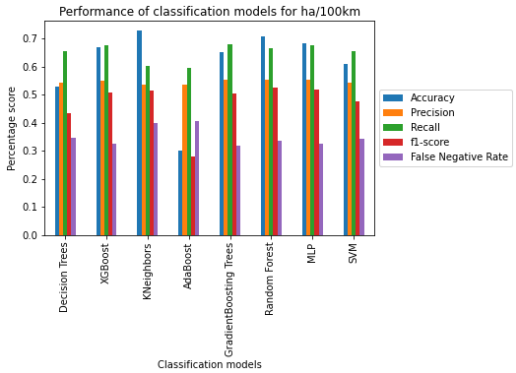
Ως ευρύτερο μοντέλο ταξινόμησης και πρόβλεψης της επικίνδυνης Οδικής συμπεριφοράς, επιλέχθηκαν ως τάξεις κατανομής οι εξής δύο: Επικίνδυνη Οδική συμπεριφορά και Μη Επικίνδυνη Οδική συμπεριφορά.

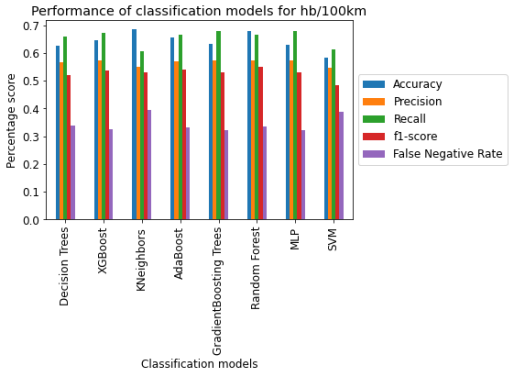
Προτού αναπτυχθούν οι αλγόριθμοι ταξινόμησης, τα δεδομένα απαιτούταν να προ-επεξεργαστούν, προκειμένου να αναβαθμιστούν τα προβλεπτικά μοντέλα. Η διαδικασία της προ-επεξεργασίας διακρίθηκε στις τεχνικές ομαδοποίησης των εξαρτημένων μεταβλητών και στην τεχνική Υπερδειγματοληψίας της μειονοτικής τάξης. Με την χρήση του αλγορίθμου Κ-μέσου, εντοπίστηκαν όρια τιμών (thresholds). Άνω και κάτω αυτών των ορίων, οι εξαρτημένες μεταβλητές μετατράπηκαν σε δυαδική μορφή, προσβλέποντας στην κατανομή τους στις επιλεγείσες κλάσεις ταξινόμησης Οδικής συμπεριφοράς. Στη συνέχεια, εφαρμόστηκε η τεχνική Υπερδειγματοληψίας Συνθετικής Μειονοτικής (SMOTE), προκειμένου να επιλυθεί το πρόβλημα της άνισης κατανομής των δεδομένων εκπαίδευσης στις δύο κλάσεις.

Για την διαδικασία της ταξινόμησης αναπτύχθηκαν οκτώ αλγόριθμοι για τις δύο εξαρτημένες μεταβλητές. Οι ταξινομήσεις περιλάμβαναν τους αλγορίθμους ταξινόμησης Δένδρων Απόφασης, Ενίσχυσης Κλίσης, Ακραίας Ενίσχυσης Κλίσης, Τυχαίων Δασών, Προσαρμοστική Ενδυνάμωση, Κ-πλησιέστερων γειτόνων, Μηχανών Διανυσμάτων Υποστήριξης και Πολυεπίπεδων Αισθητήρων. Για την αξιολόγηση των προβλεπτικών μοντέλων που αναπτύχθηκαν, αξιοποιήθηκαν στατιστικές μετρικές αξιολογήσεις, με κριτήριο την Οδική Ασφάλεια. Η σύγκριση των αλγορίθμων παρατίθεται στα ακόλουθους Πίνακες και Γραφήματα.

Πίνακας Περίληψης 1: *Σύγκριση μετρικών αξιολόγησης ταξινόμησης για τις απότομες επιταχύνσεις ανά 100χλμ.*

Αλγόριθμος ταξινόμησης	**Ορθότητα**	**Ακρίβεια**	**Ανάκληση**	**FNR**	**AUC score**
Decision Trees	53.03%	54.12%	65.35%	34.65%	70.48%
GradientBoosting	65.28%	55.15%	68.05%	31.95%	75.10%
XGBoost	66.76%	55.09%	67.46%	32.54%	74.26%
Random Forests	70.83%	55.16%	66.39%	33.61%	73.98%
AdaBoost	29.97%	53.51%	59.44%	40.56%	59.44%
KNeighbors	72.70%	53.46%	60.08%	39.92%	64.55%
SVM	61.07%	54.30%	65.60%	34.40%	65.60%
MLP	68.16%	55.26%	67.65%	32.35%	74.67%





Πίνακας Περίληψης 2 : *Σύγκριση μετρικών αξιολόγησης ταξινόμησης για τις απότομες επιβραδύνσεις ανά 100χλμ.*

Αλγόριθμος ταξινόμησης	**Ορθότητα**	**Ακρίβεια**	**Ανάκληση**	**FNR**	**AUC score**
Decision Trees	62.51%	56.58%	66.03%	33.97%	72.35%
GradientBoosting	63.36%	57.36%	67.91%	32.09%	74.88%
XGBoost	64.53%	57.20%	67.30%	32.70%	74.28%
Random Forests	67.78%	57.20%	66.48%	33.52%	73.62%
AdaBoost	65.51%	57.03%	66.66%	33.34%	73.04%
KNeighbors	68.45%	54.88%	60.55%	39.45%	65.00%
SVM	58.35%	54.58%	61.33%	38.67%	61.33%
MLP	62.96%	57.29%	67.80%	32.20%	74.69%

Οι επιδόσεις των αλγόριθμων πρόγνωσης και ταξινόμησης κυμάνθηκαν σε παρεμφερή επίπεδα, καθιστώντας τα μοντέλα που αναπτύχθηκαν χρήσιμα για την ταξινόμηση της Οδικής συμπεριφοράς. Ο αλγόριθμος Ενίσχυσης Κλίσης έδωσε τα καλύτερα αποτελέσματα και εκ του σύνεγγυς ακολούθησαν οι Πολυεπίπεδοι Αισθητήρες (MLP).

Βάσει των αποτελεσμάτων που προέκυψαν σε όλο το εύρος της παρούσας Εργασίας, ανέκυψαν ορισμένα χρήσιμα συμπεράσματα, άμεσα σχετιζόμενα με τον στόχο της. Πιο συγκεκριμένα:

1. Η συνολική διανυθείσα απόσταση αποτελεί την σημαντικότερη μεταβλητή για την αναγνώριση της Οδικής συμπεριφοράς.

2. Κατά βάση οι οδικές συμπεριφορές δεν ήταν υπερβολικά επιθετικές, καθώς οι σχετικοί διάμεσοι των απρόοπτων περιστατικών ήταν μηδενικοί και οι οδηγοί εναρμονίζονται με τους κανόνες κυκλοφορίας. Απρόοπτα περιστατικά εμφανίζονται σε διαδρομές επιβαρυμένες σε χρόνο ή και απόσταση.

3. Οι μέση ταχύτητα κίνησης, η χρήση κινητού τηλεφώνου και η υπέρβαση ορίου ταχύτητας είναι άρρηκτα συνδεδεμένες με την Επικίνδυνη Οδική συμπεριφορά και την εμφάνιση ατυχημάτων και μπορούν να αξιοποιηθούν στο μέλλον ως μεταβλητές αντίστοιχης ταξινόμησης.

4. Οι αλγόριθμοι Ενίσχυσης Κλίσης (Gradient Boosting) και Πολυεπίπεδων Αισθητήρων (Multilayered Perceptrons) που αναπτύχθηκαν ξεπέρασαν σε επιδόσεις τους υπόλοιπους αλγόριθμους, όμως οι διαφορές δεν ήταν μεγάλες. Συμπέρασμα αποτελεί ότι οι εξαρτημένες μεταβλητές των απρόοπτων περιστατικών ήταν καλά ομαδοποιημένες, με διακριτές κλάσεις, με τη μέθοδο του K-μέσου σε δυαδική κατανομή και επίπεδα ασφαλείας και ότι οι συγκεκριμένοι αλγόριθμοι μπορούν να αναπτυχθούν για αποδοτικές ταξινομήσεις σε επίπεδα ασφαλείας δύο κλάσεων.

5. Η Συνθετική Μειονοτική (SMOTE) αποδείχτηκε πιο αποτελεσματική μέθοδος από την Προσαρμοστική Συνθετική (ADASYN) σε καταστάσεις μεγάλων και πολυεπίπεδων δεδομένων και διακριτών κλάσεων, επιβεβαιώνοντας την εγχώρια και διεθνή βιβλιογραφία.

6. Επιβεβαιώνεται η επίδοση της μεθόδου πυρήνα Radial Basis Function στις Μηχανές Διανυσμάτων Υποστήριξης, συγκριτικά με τις εναλλακτικές μεθόδους πυρήνα, επαληθεύοντας την Γκαουσιανή κατανομή των εξετασθέντων στοιχείων.

7. Η μέθοδος ομαδοποίησης K-μέσου εντόπισε ως βέλτιστο threshold για την κατάταξη των απότομων περιστατικών στο Επικίνδυνο επίπεδο ασφαλείας τις 48.82 απότομες επιταχύνσεις ανά 100χλμ. και τις 45.40 απότομες επιβραδύνσεις ανά 100χλμ., παράγοντας πρωτότυπα αποτελέσματα για τα όρια δύο κλάσεων ταξινόμησης της οδικής συμπεριφοράς.

Πίνακας Περιεχομένων

1. Εισαγωγή

1.1 Γενική Ανασκόπηση

Με την άνοδο στις οδικές μεταφορές να είναι διαρκής και με τον ρυθμό μεταβολής τους ολοένα και αυξανόμενο, οι χερσαίες μεταφορές αποτελούν την κύρια μέθοδο μετακίνησης, ανά την ευρωπαϊκή επικράτεια και παγκοσμίως. Απόρροια των μετακινήσεων συνιστά η ύπαρξη οδικών ατυχημάτων, πολλών εκ των οποίων αποβαίνουν θανατηφόρα. Σύμφωνα με την Ευρωπαϊκή Επιτροπή και την Γενική Διεύθυνση Κινητικότητας και Μεταφορών (Directorate-General for Mobility and Transport), σημειώθηκαν, προσεγγιστικά, 19.800 απώλειες στους ευρωπαϊκούς δρόμους για το έτος 2021, με χαρακτηριστική μείωση από το έτος 2019 σε συνθήκες κινητικότητας προ πανδημίας, της τάξεως του -13%. Οι τραυματισμοί από οδικά ατυχήματα αποτελούν την κύρια αιτία για θανάτους σε παιδιά και νέους ενήλικες ηλικίας 5-29 ετών, με το σύνολο των θανατηφόρων περιστατικών, που προκαλούνται άμεσα ή έμμεσα από οδικά ατυχήματα, να φτάνει στα 1.3 εκατομμύρια ανθρώπους ανά έτος (WHO, 2022).

Πίνακας *1.1: Θανατηφόρα ατυχήματα στην Ευρώπη (2010-2020)*

Πηγή*: NTUA – Road Safety Observatory, [Available: https://www.nrso.ntua.gr/]*

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

Πίνακας 1.2: *Σύνοψη βασικών στατιστικών στοιχείων Οδικής Ασφάλειας στην Ελλάδα (2010-2019)*

Πηγή: *Hellenic Statistical Authority (ELSTAT), [Available: https://www.statistics.gr/]*
Traffic Police,

Επεξεργασία: *NTUA – Road Safety Observatory, [Available: https://www.nrso.ntua.gr/]*

Greece 2010 - 2020

Basic Road Safety Figures

www.nrso.ntua.gr

	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	2020/2019	2020/2010	2016/2020
Injury Road crashes	15,032	13,849	12,398	12,109	11,690	11,440	11,318	10,848	10,737	10,712	9,105	-15.0%	-39.4%	-19.6%
Fatalities	1,258	1,141	988	879	795	793	824	731	700	688	579	-15.8%	**-54.0%**	-29.7%
Serious Injuries	1,709	1,626	1,399	1,212	1,016	999	879	706	727	652	487	-25.3%	-71.5%	-44.6%
Slight Injuries	17,399	15,633	14,241	13,963	13,548	13,097	12,946	12,565	12,422	12,350	10,130	-18.0%	-41.8%	-21.8%
Vehicle Fleet (x1000)	8,062	8,087	8,070	8,035	8,048	8,076	8,173	8,263	8,237	8,402	8,519	1.4%	5.7%	4.2%
Fatalities per million vehicles	156	141	122	109	99	98	101	88	85	82	68	-17.0%	-56.4%	-32.6%
Speed infringements	263,382	238,033	186,675	178,816	156,892	173,476	176,592	208,190	213,333	234,169	206,554	-11.8%	-21.6%	17.0%
Drink & drive infringements	38,033	34,992	30,707	30,853	29,597	29,191	33,192	32,964	33,394	31,557	19,096	-39.5%	-49.8%	-42.5%
Seat belt infringements	49,703	37,120	33,722	35,478	34,526	29,611	34,831	31,510	33,380	34,594	30,174	-12.8%	-39.3%	-13.4%
Helmet infringements	51,526	47,250	47,736	58,122	54,354	52,783	63,971	59,405	52,706	52,089	46,394	-10.9%	-10.0%	-27.5%

Γράφημα 1.1: *Θανατηφόρα ατυχήματα και στόλος οχημάτων στην Ελλάδα (2010-2020)*

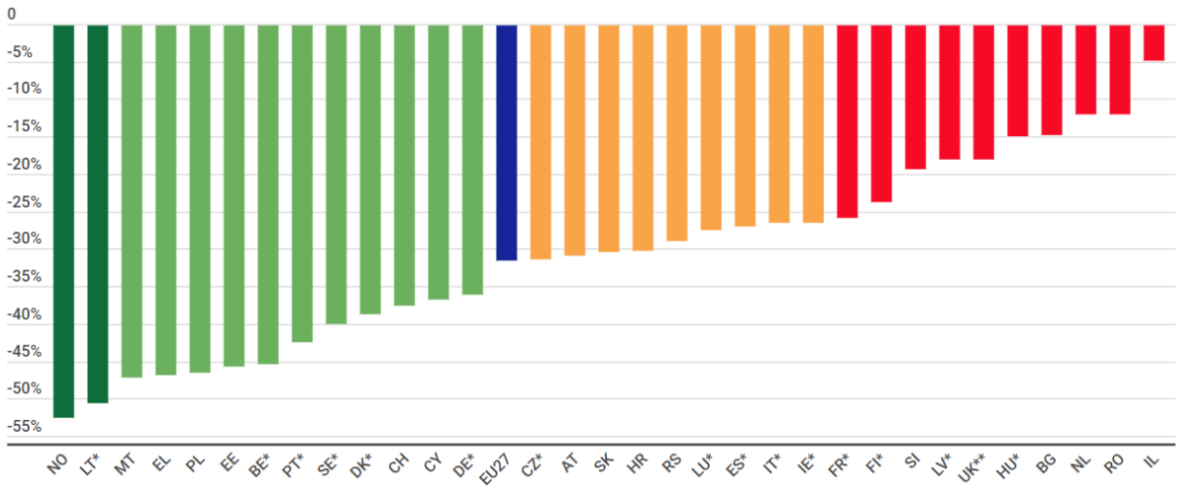
Πηγή: *Hellenic Statistical Authority (ELSTAT), [Available: https://www.statistics.gr/]*

Επεξεργασία: *NTUA – Road Safety Observatory, [Available: https://www.nrso.ntua.gr/]*

Επομένως, η Οδική Ασφάλεια αποτελεί ζήτημα μείζονος σημασίας για τους θεσμούς της Ευρωπαϊκής Ένωσης και τις επιμέρους εθνικές της συνιστώσες.

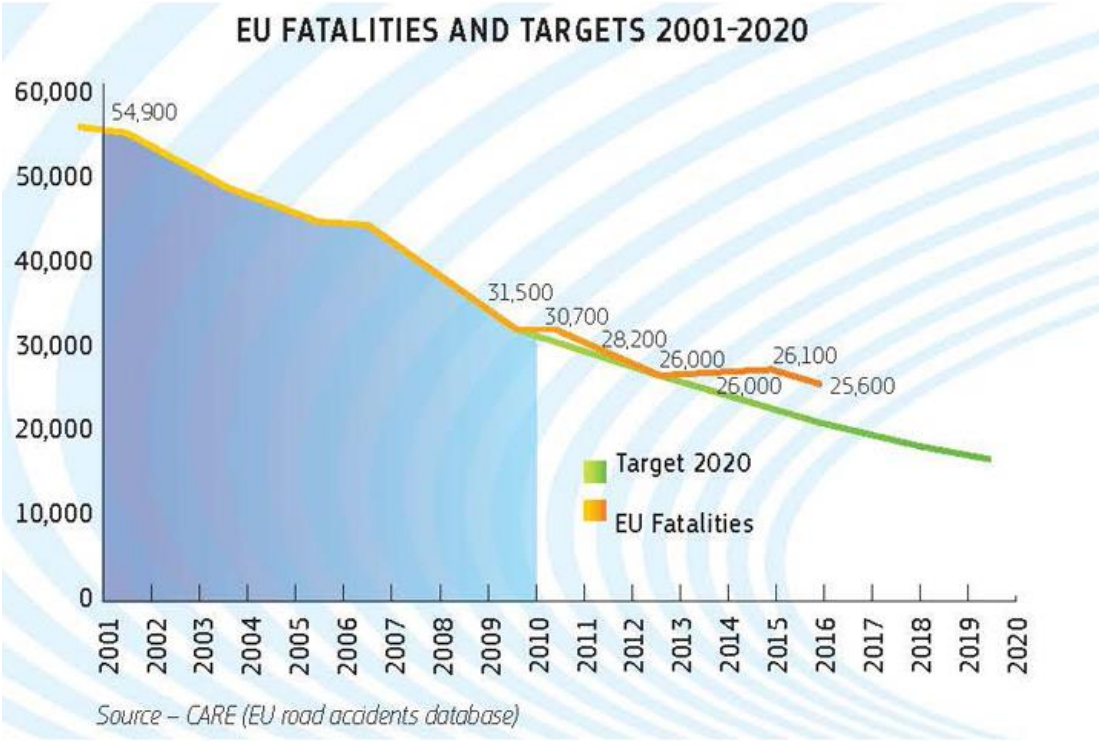
Τα τελευταία έτη και με γνώμονα την βελτίωση της Οδικής Ασφάλειας, η Ευρωπαϊκή Ένωση έχει θεσπίσει αξιοπρόσεκτα μέτρα, χάρη στα οποία, εκτιμάται ότι σημειώθηκαν 50.000 λιγότερα θανατηφόρα τροχαία ατυχήματα. Χαρακτηριστικά αναφέρεται ότι, η Ελλάδα ήταν η μοναδική ευρωπαϊκή χώρα, μεταξύ των κρατών μελών, η οποία ανταπεξήλθε στον τεθειμένο στόχο για μείωση 50% των θανατηφόρων τροχαίων ατυχημάτων την δεκαετία 2010-2020, με συνολική μείωση περίπου 52%.

Η κυριότερη πρωτοβουλία που έχει ήδη ληφθεί είναι το «Όραμα για Μηδενικές Απώλειες» (Vision Zero), σύμφωνα με το οποίο τίθεται στόχος να προσεγγιστούν οι μηδενικοί θάνατοι και σοβαροί τραυματισμοί, μέχρι το έτος στόχο 2050. Μεσοπρόθεσμη επιδίωξη αποτελεί η μείωση κατά 50% των σοβαρών περιστατικών έως το έτος 2030, σύμφωνα με την Διακήρυξη της Βαλέτα και το καταστατικό ‘EU Road Safety Policy Framework 2021-2030 – Next steps towards “Vision Zero” (SWD (2019) 0283), το οποίο βασίζεται στην προσέγγιση «Ασφαλούς Συστήματος». Κύριος άξονας της πρωτοβουλίας Vision Zero είναι η δημιουργία ενός διασυνδεδεμένου συστήματος προληπτικής Οδικής Ασφάλειας, με το εγχείρημα να ξεκινά από την Σουηδία το 1994 και να εγκρίνεται από το εθνικό Κοινοβούλιο το 1997, πριν υιοθετηθεί εξ ολοκλήρου από την Ευρωπαϊκή Επιτροπή. Τα ευφυή δεδομένα συγκροτούν ένα σημαντικό πλεονέκτημα, όσον αφορά την επίτευξη των στόχων του Vision Zero και οι πληροφορίες που βασίζονται σε δεδομένα αποτελούν σημαίνον πλεονέκτημα για κάθε στρατηγικό σχέδιο οδικής ασφάλειας. Εν τέλει, ο κατάλληλος χειρισμός των ευφυών και πολυεπίπεδων δεδομένων μπορεί να συμβάλλει σημαντικά στο να γίνουν τα κρατικά οδικά συστήματα ασφαλέστερα για ταξίδια.



Γράφημα 1.2: *Σχετικές μεταβολές (%) στα θανατηφόρα ατυχήματα στην Ευρώπη (2011-2021)*

Πηγή: *ETSC PIN report, 2021, [Available: https://etsc.eu/projects/pin/]*



Γράφημα 1.3: *Θανατηφόρα ατυχήματα (2001-2016) και στόχος για το έτος 2020 στην Ευρώπη*

Πηγή: *CARE (EU road accidents database), [Available: https://road-safety.transport.ec.europa.eu/statistics-and-analysis/methodology-and-research/care-database_en]*

Πέραν των γεωμετρικών χαρακτηριστικών της οδού (χάραξη, μηκοτομή, διατομή, οδόστρωμα, κλπ.) και της κυκλοφορίας (ταχύτητες, φόρτοι, σήμανση, σηματοδότηση, κλπ.), η ανθρώπινη συμπεριφορά αποτελεί την κύρια αιτία πρόκλησης σοβαρών τροχαίων περιστατικών. Ανά τα χρόνια, έχει πραγματοποιηθεί πληθώρα μελετών για να εντοπιστεί πόσο επιδρά κάθε στοιχείο του οδικού δικτύου και της ανθρώπινης Οδικής συμπεριφοράς στον αριθμό και την σοβαρότητα των οδικών ατυχημάτων. Παρόλο που οι μελέτες έχουν δείξει ότι υπάρχουν στατιστικά σημαντικές σχέσεις μεταξύ των εξεταζόμενων παραγόντων, τα συμπεράσματα τις προηγούμενες δεκαετίες, χαρακτηρίζονταν επισφαλή, κυρίως λόγω της πολυπλοκότητας της ανάλυσης δεδομένων (Bärgman, 2015). Με την διαρκή ανάπτυξη των τεχνολογιών, η Ευρωπαϊκή Ένωση θεσπίζει διαρκώς νέες καινοτομίες και χρηματοδοτεί έρευνες στον τομέα της Οδικής Ασφάλειας και της επεξεργασίας στατιστικών δεδομένων συναφών με αυτή, προκειμένου να εξαχθούν χρήσιμα συμπεράσματα αναφορικά με την επικίνδυνη οδική συμπεριφορά, που είναι η κύρια αιτία για την πρόκληση τροχαίων ατυχημάτων (Bieńkowska, 2018). Κύριος άξονας για την βελτίωση της οδικής ασφάλειας, αποτελεί η ανάλυση της Οδικής συμπεριφοράς με την ανάπτυξη κατάλληλων αλγορίθμων μηχανικής μάθησης.

Σημειώνεται ότι, η υπερβολική ταχύτητα και οι σχέσεις επιτάχυνσης-επιβράδυνσης αποτελούν βασικό παράγοντα στο 30% των θανατηφόρων ατυχημάτων, προσεγγιστικά, με την Ευρωπαϊκή Επιτροπή να καλεί τα κράτη μέλη να δώσουν προτεραιότητα σε έναν κεντρικό άξονα διαχείρισης της ταχύτητας, προκειμένου να προλαμβάνονται περιστατικά υπερβολικών επιταχύνσεων και επιβραδύνσεων. Η ταχύτητα έχει άμεση επιρροή στην συχνότητα ατυχημάτων και στην σοβαρότητα αυτών. Μία αύξηση της τάξεως του 1% στην μέση ταχύτητα οδηγεί στην αύξηση κατά 2% περίπου της συχνότητας εμφάνισης ήπιων τραυματισμών από οδικά ατυχήματα, κατά 3% της συχνότητας εμφάνισης σοβαρών τραυματισμών και 4% των θανατηφόρων περιστατικών (Nilsson, 1981 & 2004). Επομένως, για λόγους πρόληψης και ταξινόμησης, η χωροχρονική

διάσταση με την οποία πραγματοποιούνται οι επιταχύνσεις και οι επιβραδύνσεις, χρίζεται ιδιαιτέρως σημαντική και αποδίδεται αξία στην κατηγοριοποίησή τους σε κλάσεις «Επικίνδυνες» και «Μη επικίνδυνες».

1.2 Στόχος

Η παρούσα Διπλωματική Εργασία αποσκοπεί στον εντοπισμό, ταξινόμηση και πρόβλεψη της επικίνδυνης οδικής συμπεριφοράς και των σοβαρών περιστατικών μέσω μάθησης ανισόρροπων δεδομένων και αλγορίθμων ταξινόμησης.

Πιο συγκεκριμένα, έγινε καθορισμός του βαθμού επιρροής των ανεξάρτητων μεταβλητών στις εξαρτημένες, όπως αυτές είχαν προκαθοριστεί, με την διαδικασία της Επιλογής Χαρακτηριστικών (Feature Selection) και της σημαντικότητας (μετάθεσης των) χαρακτηριστικών [Feature (Permutation) Importance], με την συμβολή μοντέλων Παλινδρομήσεων. Μετέπειτα, αναπτύχθηκαν μοντέλα ταξινόμησης, προκειμένου οι εξαρτημένες μεταβλητές να ταξινομηθούν στα καθορισμένα επίπεδα ασφαλείας και να κατασκευαστεί ένα μοντέλο πρόβλεψης σοβαρών περιστατικών.

Τα συμπεράσματα που θα προκύψουν από την ανάλυση συμβάλλουν στην διερεύνηση εναλλακτικών μορφών καθορισμού και πρόβλεψης των επικίνδυνων οδικών περιστατικών, συγκριτικά με τις υπάρχουσες. Εν κατακλείδι, η εκπόνηση της παρούσας Διπλωματικής Εργασίας προσβλέπει στην συμβολή σε τρέχουσες και μελλοντικές έρευνες και στην εξέλιξη του τομέα της Οδικής Ασφάλειας.

1.3 Μεθοδολογία

Στην συγκεκριμένη ενότητα, περιγράφεται συνοπτικά το μεθοδολογικό πλαίσιο που ακολουθήθηκε, για την εκπόνηση της παρούσας Διπλωματικής Εργασίας.

Αρχικά, οριστικοποιήθηκε το θέμα της εργασίας και καθορίστηκε ο στόχος που δυνητικά πρόκειται να εκπληρώσει, μέσω καθορισμού των κύριων αξόνων και ερευνητικών ερωτημάτων. Ακολούθησε η διαδικασία της βιβλιογραφικής ανασκόπησης συναφών ερευνών και επιστημονικών εργασιών, για την άντληση πληροφοριών σχετικές με το θέμα της Διπλωματικής Εργασίας, όπως μέθοδοι χειρισμού των δεδομένων που χρησιμοποιήθηκαν, κατάλληλα μοντέλα για την επεξεργασία και ανάλυση των δεδομένων και τυχόν ελλείψεις και μειονεκτήματα που αυτές οι εργασίες/ διατριβές παρουσιάζουν, έτσι ώστε να αποφευχθούν.

Στη συνέχεια, πραγματοποιήθηκε η συλλογή και επεξεργασία των στοιχείων. Παρουσιάζονται οι πηγές και οι μέθοδοι συλλογής τους και πραγματοποιείται μια προκαταρκτική περιγραφική και στατιστική ανάλυση των δεδομένων, για την κατάλληλη προεργασία τους, προκειμένου να υποβληθούν μετέπειτα προς επεξεργασία. Μετά την συλλογή και προκαταρκτική επεξεργασία των στοιχείων, ακολούθησε η ανάπτυξη των μοντέλων Μηχανικής Μάθησης για Παλινδρόμηση και ταξινόμηση τους, καθώς και η συνοπτική παρουσίαση των αποτελεσμάτων. Η ανάπτυξη των μοντέλων πραγματοποιήθηκε με την γλώσσα προγραμματισμού Python σε προγραμματιστικό περιβάλλον Jupyter Notebook και Google Colab, με την αρωγή κατάλληλων βιβλιοθηκών ανάλυσης δεδομένων και Μηχανικής Μάθησης.

Τέλος, έγινε αναλυτική παρουσίαση, σύγκριση και αξιολόγηση των αποτελεσμάτων και εξήχθησαν ορισμένα χρήσιμα ερευνητικά συμπεράσματα από τις αναλύσεις. Σε αυτό το στάδιο, ο

συγγραφέας υποβάλλει συγκεκριμένες προτάσεις, που θα μπορούσαν να λειτουργήσουν ως εφαλτήριο για περαιτέρω ερευνητικό έργο και την εν γένει εξέλιξη του γνωστικού αντικειμένου.

Στο παρακάτω Γράφημα (Γράφημα 1.4), παρουσιάζεται το μεθοδολογικό πλαίσιο, με τα επιμέρους διαδοχικά στάδιά του, που ακολουθήθηκε για την εκπόνηση της Διπλωματικής Εργασίας.

Γράφημα 1.4*: Διάγραμμα ροής – Σύνοψη μεθοδολογικού πλαισίου της Διπλωματικής Εργασίας*

1.4 Δομή της Διπλωματικής Εργασίας

Στην παρούσα ενότητα, παρουσιάζεται η δομή της Διπλωματικής Εργασίας και η αδρομερής περιγραφή των διακριτών της ενοτήτων.

Στο πρώτο Κεφάλαιο, επιδιώκεται μια πρώτη επαφή του αναγνώστη με το γνωστικό αντικείμενο που πραγματεύεται η Διπλωματική Εργασία, διαμέσως του υποκεφαλαίου της Γενικής Ανασκόπησης και τεκμαίρεται η συνεισφορά της παρούσας ερευνητικής αποστολής. Παρατίθενται συγκεντρωτικά στατιστικά στοιχεία για την Οδική Ασφάλεια και την επιρροή των απότομων περιστατικών σε αυτή, καθώς και πρακτικές που έχουν υιοθετηθεί την τελευταία δεκαετία στην Ευρώπη, προς την βελτίωση των υφισταμένων συνθηκών. Συγκεντρωτικά, τεκμαίρεται η αξία της πρόληψης και της πρόβλεψης των επικίνδυνων περιστατικών, συναρτήσει των δεδομένων φυσικής οδήγησης και του ανθρωπίνου παράγοντα.

Στο δεύτερο Κεφάλαιο, συντάσσεται η βιβλιογραφική ανασκόπηση που λειτούργησε ως πρόδρομος για την εκπόνηση της Εργασίας. Παρουσιάζονται συναφείς έρευνες και αποτελέσματα, καθώς και συμπεράσματα, τα οποία διαφέρουν μεταξύ τους ως προς τα ποσοτικά και ποιοτικά τους χαρακτηριστικά , με τις αντίστοιχες μεθοδολογίες που ακολουθήθηκαν. Αξιοποιείται το

ερευνητικό έργο από ένα σημαντικό υποσύνολο της Διεθνούς και της Ελληνικής Επιστημονικής Κοινότητας, με τις επιμέρους παραθέσεις των ερευνητών και του διακεκριμένου έργου τους.

Στο τρίτο Κεφάλαιο, παρουσιάζεται το θεωρητικό υπόβαθρο, που αξιοποιήθηκε για το ερευνητικό αντικείμενο της παρούσας Διπλωματικής Εργασίας. Πραγματοποιείται αναλυτική περιγραφή της διαδικασίας που ακολουθήθηκε, με τα επιμέρους θεωρητικά στοιχεία που την συνοδεύουν, οι μέθοδοι ανάλυσης και επεξεργασίας των δεδομένων με ταυτόχρονη επισήμανση της σημαντικότητας κάθε διακριτής τεχνικής. Αναλύονται οι μαθηματικές και στατιστικές θεωρίες και οι θεωρίες ανάπτυξης αλγορίθμων Μηχανικής Μάθησης (Machine Learning), Τεχνητών Νευρωνικών Δικτύων (Artificial Neural Networks) και Βαθιάς Μάθησης (Deep Learning), στις οποίες βασίζονται οι μέθοδοι που εξετάστηκαν και γίνεται αναφορά, εν συντομία, σε εναλλακτικές μεθόδους που θα μπορούσαν να εκτελεσθούν. Εν τέλει, εισάγονται οι έννοιες των μετρικών αξιολογήσεων (metrics) και των κριτηρίων αποδοχής των μοντέλων, προκειμένου να επιλεγεί η κατάλληλη μέθοδος ανάλυσης.

Στο τέταρτο Κεφάλαιο, πραγματοποιείται η περιγραφή των δεδομένων προς επεξεργασία, καθώς και η διαδικασία συλλογής και επιλογής τους στην τελική βάση δεδομένων της OSeven Telematics©. Στη συνέχεια, τα δεδομένα υποβάλλονται προς μιας πρώτης τάξεως προκαταρκτική επεξεργασία, έτσι ώστε να γίνει η κατάλληλη προεργασία, πριν το στάδιο της τελικής τους ανάλυσης.

Στο πέμπτο Κεφάλαιο, πραγματοποιείται η κύρια επεξεργασία και ανάλυση των δεδομένων, κατά τα επιμέρους μοντέλα που αξιοποιήθηκαν. Το συγκεκριμένο κεφάλαιο διαρθρώνεται σε δύο κυρίως τμήματα, αυτό της Σημαντικότητας Χαρακτηριστικών μέσω του υπολογισμού της επιρροής των εξαρτημένων μεταβλητών στις ανεξάρτητες (Feature Importance) και αυτό της διαδικασίας ταξινόμησης (Classification). Αναλύονται λεπτομερώς τα βήματα και η εν γένει διαδικασία που ακολουθήθηκε για τον καταρτισμό των μοντέλων και πραγματοποιείται μια συνοπτική περιγραφική και συγκριτική ανάλυση των αποτελεσμάτων με κριτήριο τις μετρικές τους αξιολογήσεις (metrics).

Στο έκτο Κεφάλαιο, γίνεται λεπτομερής παρουσίαση των τελικών αποτελεσμάτων, μέσω χαρακτηριστικών πινάκων και γραφημάτων, που προέκυψαν από την ανάλυση που προηγήθηκε στο Κεφαλαίο 5. Επιδιώκεται η κατά μέρη κατάλληλη σύνθεση των αποτελεσμάτων, προκειμένου να εξαχθούν χρήσιμα συμπεράσματα για το ερευνητικό έργο και τα δεδομένα. Στο τέλος, παρατίθενται τα συμπεράσματα που ανέκυψαν από το παρόν ερευνητικό έργο, καθώς και ορισμένες προτάσεις που χρήζουν περαιτέρω διερεύνησης κατά τον συγγραφέα, βάσει της εμπειρίας που αποκόμισε από την εκπόνηση της παρούσας Διπλωματικής Εργασίας.

2. Βιβλιογραφική Ανασκόπηση

2.1 Εισαγωγή

Σκοπός αυτής της ενότητας, είναι η ανασκόπηση και αξιολόγηση ερευνών, τεχνικών και μεθοδολογιών συναφών με το αντικείμενο της παρούσας Διπλωματικής Εργασίας. Πιο συγκεκριμένα, πραγματοποιήθηκε διερεύνηση της διεθνούς και εγχώριας βιβλιογραφίας σε δημοσιευμένες έρευνες, που πραγματεύονται το αντικείμενο της Οδικής Ασφάλειας, μέσω ανάλυσης της Οδικής συμπεριφοράς, επεξεργασίας δεδομένων φυσικής οδήγησης και ικανότητας πρόβλεψης συμπεριφοράς των μοντέλων Μηχανικής Μάθησης. Το ερευνητικό έργο που πρόκειται να παρουσιαστεί λειτούργησε ως προπομπός για την εκπόνηση της παρούσας Διπλωματικής Εργασίας, αφού σύμφωνα με αυτό καθορίστηκε ο στόχος και καταρτίστηκε η μεθοδολογία της. Οι επιλεχθείσες έρευνες εξετάζουν προβλήματα ταξινόμησης της Οδικής συμπεριφοράς με τεχνικές Μηχανικής Μάθησης και μεθόδους διαχείρισης και ανάλυσης μη ισορροπημένων πολυεπίπεδων δεδομένων. Στο τέλος, επιχειρείται η σύγκριση των εναλλακτικών μεθοδολογιών και η καταγραφή τυχόν ελλείψεων ή μειονεκτημάτων που αυτές παρουσιάζουν, καθώς και συνοπτικές προτάσεις για περαιτέρω έρευνα σε κάθε μια από τις ακόλουθες δημοσιεύσεις.

2.2 Ταξινόμηση της Οδικής συμπεριφοράς

Στην παρούσα υποενότητα, εξετάζεται το ερευνητικό έργο σε προβλήματα ταξινόμησης Οδικής συμπεριφοράς. Η ταξινόμηση της Οδικής συμπεριφοράς αποτελεί ιδιάζουσα πρόκληση για την βελτίωση της Οδικής Ασφάλειας, καθώς επιτρέπει την ανάπτυξη και την εξέλιξη τεχνολογικών συστημάτων ασφαλείας, που έχουν προβλεπτικό, διορθωτικό και αποτρεπτικό χαρακτήρα, όσων αφορά την επιθετική και την εν γένει Επικίνδυνη Οδική συμπεριφορά.

Η δυνητική επικινδυνότητα της οδικής συμπεριφοράς είναι άρρηκτα συνδεδεμένη με τις συνήθειες του οδηγού, κατά την διάρκεια του ταξιδιού του. Το παραπάνω αποδεικνύει η έρευνα των Papadimitriou et al., 2019 η οποία επιδιώκει την ποσοτικοποίηση της συσχέτισης της χρήσης κινητού τηλεφώνου κατά τη διάρκεια μιας διαδρομής και της επικίνδυνης οδικής συμπεριφοράς. Τα δεδομένα που χρησιμοποιήθηκαν προήλθαν από πραγματικά οδικά πειράματα δεδομένων φυσικής οδήγησης και για την επεξεργασία τους αξιοποιήθηκαν τεχνικές Μηχανικής Μάθησης και υπεισήλθαν δεδομένα όπως η ταχύτητα, η επιτάχυνση, οι απότομες επιβραδύνσεις και η χρήση κινητού τηλεφώνου στην διάρκεια της διαδρομής. Τα αποτελέσματα καταδεικνύουν την επικινδυνότητα με την οποία σχετίζεται η χρήση του κινητού τηλεφώνου με φαινόμενα υπέρβασης του ορίου ταχύτητας και απότομα περιστατικά, κυρίως απότομες στροφές. Η προβλεπτική ικανότητα του μοντέλου δυαδικής λογιστικής παλινδρόμησης που αναπτύχθηκε, φτάνει το 70% και κρίνεται ιδιαίτερα ικανοποιητική.
Η έρευνα των K. Yang et al., 2021 στοχεύει στην ταξινόμηση της Οδικής συμπεριφοράς σε καθορισμένα επίπεδα ασφαλείας και στην αξιολόγησή της σε πραγματικό χρόνο. Η προτεινόμενη μεθοδολογία επικεντρώνεται στην εύρεση του βέλτιστου αριθμού και των βέλτιστων ορίων των επιπέδων ασφαλείας για την αξιολόγηση της Οδικής συμπεριφοράς και αποσκοπεί στην εξέλιξη των Προχωρημένων Συστημάτων Οδικής Βοήθειας για την εξασφάλιση καλύτερου επιπέδου Οδικής Ασφάλειας. Εφαρμόστηκαν τρεις τεχνικές ομαδοποίησης των δεδομένων: K-μέσου, ιεραρχική ομαδοποίηση και μείγμα Γκαουσιανών μοντέλων (GMM) με αλγόριθμο

Μεγιστοποίησης Προσδοκίας (EM algorithm). Τα αποτελέσματα κατέδειξαν ότι ο βέλτιστος αριθμός επιπέδων ασφαλείας είναι 4: ομαλή οδήγηση, οδήγηση χαμηλής επικινδυνότητας, μεσαίας επικινδυνότητας και υψηλής επικινδυνότητας, με πιο αποτελεσματικό μοντέλο να αποδεικνύεται αυτό του K-μέσου. Για την ταξινόμηση των οδικών στοιχείων αξιοποιήθηκε ο αλγόριθμος SVM μαζί με την χρήση των Γκαουσιανών μοντέλων, με την ακρίβεια του μοντέλου να φτάνει το 97.9%.

Στην δημοσίευσή τους, οι Zhang et al., 2016 επιχειρούν την ανάλυση Οδικής συμπεριφοράς από δεδομένα που προέρχονται μόνο από αισθητήρες χαμηλού επιπέδου, όπως αυτούς του αυτοδιαγνωστικού συστήματος του οχήματος (OBD) και του κινητού τηλεφώνου. Τα δεδομένα που αναλύθηκαν στην συγκεκριμένη έρευνα προήλθαν από ελεγχόμενες δοκιμές δεδομένων φυσικής προέλευσης και εξετάστηκαν ανεξάρτητα και συνδυαστικά για κάθε αισθητήρα. Ο συνδυασμός αισθητήρων κινητού τηλεφώνου και ενσωματωμένων αισθητήρων του οχήματος αποδείχτηκε ο πιο αποτελεσματικός, με ποσοστό ακρίβειας ταξινόμησης έως και 97.5% για τα δεδομένα ενός οχήματος και η ταξινόμηση με προσέγγιση πραγματικών συνθηκών οδήγησης για όλα τα εξεταζόμενα οχήματα έφτασε το 86.67%, με την βοήθεια Μηχανών Διανυσμάτων Υποστήριξης.

Η ταξινόμηση Οδικής συμπεριφοράς στην έρευνα των Ghandour et al., 2021 βασίζεται στην ανάλυση εναλλακτικών συναισθηματικών και ψυχολογικών συνθηκών του οδηγού. Πληθώρα ερευνών έχει ήδη εκπονηθεί βάσει αυτής της ταξινόμησης, όμως με χαρακτηριστική έλλειψη ακρίβειας σε έλεγχο πραγματικών οδικών συνθηκών. Η μεθοδολογία των Ghandour et al., περιλαμβάνει την ανάπτυξη μοντέλων ταξινόμησης μηχανικής μάθησης όπως ταξινόμηση Λογιστικής Παλινδρόμησης, Random Forests, Τεχνητών Νευρωνικών Δικτύων και Gradient Boosting, με τις επιμέρους κλάσεις ταξινόμησης να χωρίζονται σε ομαλή συμπεριφορά, επιθετική και νυσταγμένη. Η έρευνα πραγματοποιήθηκε ξεχωριστά για δύο βάσεις δεδομένων διαφορετικής προέλευσης. Η μία βάση περιλαμβάνει δεδομένα ανίχνευσης λωρίδας με τα δεδομένα της να περιέχουν στοιχεία θέσης και κατεύθυνσης του οχήματος, ενώ η δεύτερη περιλαμβάνει δεδομένα συνθηκών φόρτου, με στοιχεία σχετικά με το περιβάλλον του οχήματος. Η βάση δεδομένων συνθηκών φόρτου αποδεικνύεται καταλυτικότερη και ακριβέστερη στην ταξινόμηση στις 3 ψυχολογικές κλάσεις, με την ταξινόμηση Gradient Boosting να δίνει χρήσιμα συμπεράσματα για τον προσδιορισμό της επικινδυνότητας.

Το ερευνητικό έργο των Mumcuoglu et al., 2019 ασχολείται με την αναγνώριση μοτίβων Οδικής συμπεριφοράς, με γνώμονα την εξέλιξη των Συστημάτων Οδικής Ασφαλείας. Η κατασκευή του μοντέλου ταξινόμησης συμπεριφοράς πραγματοποιήθηκε με δεδομένα που συλλέχθηκαν από προσομοιωτές οδήγησης φορτηγού οχήματος. Οι Mumcuoglu et al., ανέπτυξαν ένα μοντέλο μακροχρόνιας βραχυπρόθεσμης μνήμης (LSTM) Νευρωνικών Δικτύων, που αξιοποιεί τα οδικά σήματα που παρέχονται από συστήματα Αδρανειακής Μονάδας οχήματος (IMU), GPS και Radar/LiDAR και αποστέλλει ο προσομοιωτής. Τα επιμέρους στοιχεία που αξιοποιήθηκαν είναι στοιχεία, όπως η διαμήκης και πλευρική επιτάχυνση, η ταχύτητα, η κατανάλωση καυσίμου, ο αισθητήρας πεντάλ επιτάχυνσης και της πίεσης πεντάλ επιβράδυνσης. Η έρευνα επικεντρώθηκε στην άμεση ανάλυση και ταξινόμηση της συμπεριφοράς, με τα αποτελέσματα να καταδεικνύουν ότι το χρονικό παράθυρο ανάλυσης των δεδομένων επηρεάζει σημαντικά τα αποτελέσματα ταξινόμησης, με το βέλτιστο παράθυρο να υπολογίστηκε στα 15 δευτερόλεπτα. Σε δεύτερο χρόνο, ο ίδιος αλγόριθμος LSTM που κατασκευάστηκε από δεδομένα προσομοιωτή, ελέγχθηκε με οδικά δεδομένα σε ρεαλιστικά μοντελοποιημένο δρόμο. Συμπερασματικά, η συγκεκριμένη έρευνα κατέδειξε ότι ο προτεινόμενος αλγόριθμος LSTM απεδείχθη ιδιαίτερα αποδοτικός στην

ταξινόμηση και αναγνώριση σχέσεων δυναμικής από οδικά σήματα και πρόκειται να χρησιμοποιηθεί ευρέως σε μελλοντικές αναλύσεις.

Στον Πίνακα 2.1 παρουσιάζονται αδρομερώς οι ελλείψεις που παρατηρήθηκαν στην ανασκόπηση των ερευνών Ταξινόμησης της Οδικής συμπεριφοράς, καθώς και προτάσεις για περαιτέρω διερεύνηση, κατά τον ερευνητή.

Πίνακας 2.1: *Ελλείψεις και Προτάσεις από την ανασκόπηση ερευνών Ταξινόμησης Οδικής συμπεριφοράς*

Έρευνα	**Ελλείψεις έρευνας**	**Προτάσεις για μελλοντική έρευνα**
Papadimitriou et al., 2021	Περιορισμένη ανάλυση σε όγκο μεταβλητών και τεχνικών ταξινόμησης.	Διερεύνηση με δεδομένα προέλευσης από μεγαλύτερη ποικιλία αισθητήρων.
K. Yang et al., 2021	Τα δεδομένα δεν συμπεριέλαβαν μεταβλητές δημογραφικών και ψυχολογικών στοιχείων και αντιληπτικής ικανότητας των οδηγών.	Περαιτέρω διερεύνηση στον εντοπισμό των σημαντικών μεταβλητών για κάθε επίπεδο ασφαλείας και την σχέση μεταξύ τους.
Zhang et al., 2016	Περιορισμός πληροφορίας από τους αισθητήρες οχήματος από το πρωτόκολλο OBD-II. Το δείγμα οδηγών και οχημάτων ήταν ιδιαίτερα περιορισμένο.	Έλεγχος με προσέγγιση δικτύων Μπεϋζιανής βελτιστοποίησης, για την βελτίωση της ακρίβειας ταξινόμησης. Περαιτέρω ανάλυση με δεδομένα από περισσότερους οδηγούς και οχήματα.
Ghandour et al., 2021	Τα δεδομένα προήλθαν μόνο από οδήγηση σε αυτοκινητόδρομο. Περιορισμένος αριθμός κλάσεων ταξινόμησης για την ψυχολογική κατάσταση του οδηγού. Μη συμπερίληψη μεταβλητών όπως ο πνευματικός φόρτος και τα όρια ταχύτητας.	Ανάλυση με υβριδικό σύστημα ταξινόμησης με συνδυασμό τεχνικών. Η ανάλυση έγινε με τυχαία δείγματα από το σύνολο των δεδομένων, συνεπώς απαιτείται ολοκληρωμένη ανάλυση με όλο το πλήθος τους. Βελτίωση του συνόλου της βάσης δεδομένων, με συνδυασμό των δύο που εξετάστηκαν και επιπρόσθετων.
Mumcuoglu et al., 2019	Περιορισμένη ανάλυση σε κλάσεις ταξινόμησης. Υψηλή πιθανότητα Παρατήρηση για κοινή συμπεριφορά των διαφορετικών κλάσεων στην πλειονότητα των λάθος ταξινομημένου δείγματος.	Εφαρμογή του προτεινόμενου αλγορίθμου LSTM σε οχήματα με αισθητήρες υψηλής ευαισθησίας. Εφαρμογή του μοντέλου σε πραγματικές οδικές συνθήκες.

2.3 Μη Ισορροπημένη Μάθηση στην ταξινόμηση της Οδικής συμπεριφοράς

Λόγω της μη ισορροπημένης φύσης των δεδομένων που ταξινομούν και προβλέπουν την επικίνδυνη οδική συμπεριφορά, η μειονοτική τάξη, που αφορά τα δείγματα υψηλής επικινδυνότητας, συνήθως δεν επεξεργάζεται και δεν αναλύεται με βέλτιστο τρόπο, από τους συνήθεις αλγορίθμους ταξινόμησης. Σε αυτήν την ενότητα εξετάζονται διαφορετικές προσεγγίσεις στην μεταχείριση της μειονοτικής τάξης με παρουσίαση εναλλακτικών τεχνικών και μεθοδολογιών, από παραπλήσιο ερευνητικό έργο ταξινόμησης Οδικής συμπεριφοράς.

Η αναγνώριση και η πρόβλεψη των επικίνδυνων περιστατικών ήταν ο κυρίαρχος στόχος της μελέτης Wang et al., 2021. Το σημαντικότερο πρόβλημα που αντιμετωπίζει η επιστημονική κοινότητα ως προς την ταξινόμηση των δυνητικά επικίνδυνων οδικών φαινομένων είναι η έλλειψη κατάλληλου χειρισμού των δεδομένων που ανήκουν στην μειονοτική τάξη, λόγω της ποσοτικής ανισορροπίας στις διαφορετικές τάξεις που παράγουν τα οδικά δεδομένα. Προκειμένου να αντιμετωπιστεί το παραπάνω φαινόμενο, οι ερευνητές προτείνουν ένα καινοτόμο πλαίσιο αυτοματοποιημένης βελτιστοποίησης των υπερπαραμέτρων που υπεισέρχονται στους αλγορίθμους μηχανικής μάθησης, που εν τέλει επιδιώκουν την αναγνώριση της επικίνδυνης οδήγησης. Τα δεδομένα που χρησιμοποιήθηκαν για την παραγωγή του αυτοματοποιημένου μοντέλου συλλέχθηκαν από πραγματικά δεδομένα τροχιάς αυτοκινούμενων και σε αυτό υπεισήλθαν οι διακεκριμένοι συντελεστές μετασχηματισμού Fourier των διαμηκών και πλευρικών ταχυτήτων και της απόστασης δύο διαδοχικών αυτοκινούμενων. Οι υπερπαράμετροι ελέγχονται από αλγορίθμους Μπεϋζιανής βελτιστοποίησης, ενώ η Υπερδειγματοληψία (oversampling) χειρίστηκε με την Τεχνική Συνθετικής Μειονοτικής Υπερδειγματοληψίας βασισμένη σε SVM (Supported Vector Machine based Synthetic Minority Oversampling Technique – SVMSMOTE). Εξετάστηκαν 5 τεχνικές Υποδειγματοληψίας και 5 τεχνικές Υπερδειγματοληψίας, με την καλύτερη μέθοδο δειγματοληψίας να αποδεικνύεται η SVMSMOTE, με σταθμισμένη διασταυρούμενη εντροπία (weighted cross-entropy) ως συνάρτηση απώλειας.

Η έρευνα των L.Yang et al., 2017 στοχεύει στην ταξινόμηση της Οδικής συμπεριφοράς μέσω συσχέτισής της με δεδομένα ηλεκτροεγκεφαλογραφήματος (EEG) του οδηγού. Παρεμφερείς έρευνες είχαν ήδη εκπονηθεί (Zeng et al., 2018; Atilla et al., 2021), όμως η συγκεκριμένη χρησιμοποίησε μεθόδους μάθησης δύο επιπέδων με την συμβολή δεδομένων ηλεκτροεγκεφαλογραφήματος. Στο πρώτο επίπεδο, η Οδική συμπεριφορά ταξινομήθηκε βάσει δισδιάστατων οδικών χαρακτηριστικών αξιοποιώντας ομαδοποίηση K-μέσων (K-means) και αναδρομικό αποκλεισμό χαρακτηριστικών με Supported Vector Machines (SVM). Στο δεύτερο επίπεδο ταξινόμησης με Μηχανική Μάθηση, τα αποτελέσματα της ταξινόμησης του πρώτου επιπέδου εφαρμόστηκαν σε ταξινομητή K-nearest Neighbors (KNN), χρησιμοποιώντας παράλληλα ταχύ μετασχηματισμό Fourier και ανεξάρτητη ανάλυση στοιχείων. Η συλλογή των δεδομένων πραγματοποιήθηκε από πείραμα με προσομοιωτή οδήγησης, ενώ η μειονοτική τάξη υπέστη βελτιστοποίηση με την τεχνική Υπερδειγματοληψίας Προσαρμοστική Συνθετική (ADASYN). Τα αποτελέσματα της έρευνας κατέδειξαν εμφανή συσχέτιση μεταξύ των διαφορετικών μοτίβων δεδομένων ηλεκτροεγκεφαλογραφήματος του οδηγού και της οδικής του συμπεριφοράς, σε ποσοστό που έφτασε και το 83.5%.

Η διαχείριση της ανισορροπίας δεδομένων στην ταξινόμηση Οδικής συμπεριφοράς αποτελεί σημαντική πρόκληση στην ταξινόμηση, με την έρευνα των Zhu et al., 2022 να επιδίδεται σε αυτή. Οι περισσότερες έρευνες χρησιμοποιούν τεχνικές Μηχανικής Μάθησης και στατιστική ανάλυση

προκειμένου να αποκαταστήσουν το συγκεκριμένο πρόβλημα, εν αντιθέσει με την έρευνα των Zhu et al., 2022, η οποία επιχειρεί μια εναλλακτική προσέγγιση διαχείρισης της ανισορροπίας. Στην συγκεκριμένη έρευνα, προτείνεται η ταξινόμηση της επικίνδυνης Οδικής συμπεριφοράς, μέσω μη ισορροπημένων δειγμάτων χρονοσειρών. Αρχικά, εφαρμόστηκε η τεχνική MeanShift (Schnell, 1964; Fukunaga, Hostetler, 1975), ένας μη παραμετρικός αλγόριθμος που αξιοποιεί το μέγιστο της συνάρτησης πυκνότητας των δεδομένων, συνδυαζόμενη με αυτόματο εύρος ζώνης (bandwidth) για την συλλογή δειγμάτων και την αύξηση του μεγέθους τους με τεκμήριο την ομοιότητά τους. Σε δεύτερη φάση, χρησιμοποιήθηκε μοντέλο τοπικής διαίρεσης για την προσομοίωση του εξωτερικού περιβάλλοντος του οχήματος και έγινε εξαγωγή χαρακτηριστικών για τις διαφορετικές μεταβατικές καταστάσεις του οχήματος, σύμφωνα με το μοντέλο Markov (MFE). Τα παραπάνω δεδομένα επεξεργάστηκαν με συνδυασμό τριών Συνελικτικών Νευρωνικών Δικτύων (CNN), με τα πειραματικά αποτελέσματα να επιβεβαιώνουν την απόδοση των χρονοσειρών για την παραγωγή δεδομένων, σε προβλήματα μη ισορροπημένης μάθησης.

Η πρόβλεψη οδικών ατυχημάτων μπορεί να πραγματοποιηθεί όταν τα δεδομένα συγκρίνονται με καταστάσεις ομαλής Οδικής συμπεριφοράς. Όμως, στις περισσότερες έρευνες δεν εξετάζεται η δυναμική κατάσταση των ατυχημάτων, ούτε αξιολογείται η ανισορροπία των οδικών δεδομένων σε κάθε τάξη, κάτι που μπορεί να οδηγήσει σε εσφαλμένες προβλέψεις σε πραγματικό χρόνο. Η έρευνα των Katrakazas et al., 2017 επιτυγχάνει να λύσει τα προβλήματα ανισορροπίας δεδομένων, χρησιμοποιώντας συνδυαστικά δεδομένα χρονοσειράς ανεπεξέργαστης ταχύτητας και τεχνικές Μη Ισορροπημένης Μάθησης. Η ταξινόμηση της Οδικής συμπεριφοράς προέκυψε από την σύγκριση των αποτελεσμάτων ανάλυσης της πρωτότυπης χρονοσειράς, με την ανάλυση δεδομένων χρονοσειράς, στην οποία εφαρμόστηκαν τεχνικές Υποδειγματοληψίας (Undersampling) και Υπερδειγματοληψίας (Oversampling). Τα αποτελέσματα της έρευνας καταδεικνύουν ότι σε προβλήματα ταξινόμησης στα οποία οι επιμέρους κλάσεις δεν είναι εύκολα διακριτές, η πιο αξιόπιστη τεχνική είναι ο συνδυασμός των αλγορίθμων Υπερδειγματοληψίας Repeated Edited Nearest Neighbors (RENN) και Synthetic Minority Oversampling Technique (SMOTE), η οποίες επιτυγχάνουν καλή διάκριση των τάξεων και βελτιώνουν τα αποτελέσματα ταξινόμησης. Αφού εφαρμοστεί η διαδικασία Υπερδειγματοληψίας SMOTE, ο αλγόριθμος ENN, σε μια επαναληπτική διαδικασία (Repeated), καθαρίζει τα δεδομένα, αφαιρώντας κάθε στοιχείο, του οποίου οι 3 κοντινότεροι του γείτονες έχουν λανθασμένα ταξινομηθεί. Η συγκεκριμένη μέθοδος διαχείρισης της ανισορροπίας δεδομένων αποδεικνύεται ιδιαίτερα χρήσιμη σε βάσεις δεδομένων με σχετικώς μικρή αναλογία στοιχείων στην θετική τάξη της ταξινόμησης, δηλαδή σε σπάνια γεγονότα, πχ. σε τάσεις σύγκρουσης οχημάτων.

Στον Πίνακα 2.2 που ακολουθεί, παρουσιάζονται τυχόν ελλείψεις που εντοπίστηκαν στην ανασκόπηση των ερευνών για Μη Ισορροπημένη Μάθηση στην ταξινόμηση της Οδικής συμπεριφοράς, καθώς και προτάσεις για περαιτέρω διερεύνηση, κατά τον ερευνητή.

Πίνακας 2.2: *Ελλείψεις και Προτάσεις από την ανασκόπηση ερευνών Μη ισορροπημένης Μάθησης στην ταξινόμηση Οδικής συμπεριφοράς*

Έρευνα	**Ελλείψεις έρευνας**	**Προτάσεις για μελλοντική έρευνα**
Wang et al., 2021	Περιορισμένος αριθμός μοντέλων ταξινόμησης.	Εφαρμογή σε αλγορίθμους με στοιχεία από πραγματικές συνθήκες και περαιτέρω ανάλυσή τους, για την εξακρίβωση της αξιοπιστίας του αυτοματοποιημένου μοντέλου.
L.Yang et al., 2017	Η ακρίβεια των μοντέλων ταξινόμησης είναι περιορισμένη. Εξετάστηκαν δεδομένα σχετικά μόνο με την διαδοχικότητα των οχημάτων, χωρίς άλλες παραμέτρους οδικών συνθήκων.	Εξέταση των αλγορίθμων με βάση δεδομένων παραμέτρων από αλλαγή λωρίδας ή και στρεφουσών κινήσεων. Εφαρμογή σε πραγματικά οδικά πειράματα, για την εξάλειψη μεροληψίας της αντιληπτικότητας των οδηγών.
Zhu et al., 2022	Οι παράμετροι με τους οποίους πραγματοποιήθηκε η τοπική διαίρεση (RD) και το μοντέλο μετάβασης Markov (MFE) δεν ήταν αποσαφηνισμένοι.	Ανάλυση με περισσότερα πραγματικά δεδομένα για συνθήκες οδήγησης χαμηλής επικινδυνότητας.
Katrakazas et al., 2017	Εφαρμογή περιορισμένων μοντέλων ταξινόμησης.	Ανάλυση με δεδομένα που συλλέχθηκαν από πραγματικές οδικές συνθήκες. Εφαρμογή τεχνικών Μπεϋζιανών Δικτύων και Βαθιάς Μάθησης για την αντιμετώπιση του θορύβου των χρονοσειρών μικρών χρονικών περιόδων. Περαιτέρω ανάλυση με μοντέλα μορφής δένδρου απόφασης, καθώς τα Τυχαία Δάση ήταν πολύ αποδοτικά.

2.4 Σύνοψη ανασκόπησης

Η συγκεκριμένη Διπλωματική Εργασία αξιοποιεί παρεμφερείς τεχνικές και μεθοδολογίες, που αναφέρθηκαν στο παρόν Κεφάλαιο και εξετάζει πληθώρα μοντέλων ταξινόμησης, για τον εντοπισμό του βέλτιστου αλγορίθμου. Ο καθορισμός του στόχου της εργασίας προέκυψε από την διαρκή ανάγκη αναλύσεων ταξινόμησης Οδικής συμπεριφοράς, καθώς διαλευκαίνονται προβλήματα αντιμετώπισης, μεταχείρισης δεδομένων και αξιολόγησης της ταξινόμησης, για όσο το δυνατόν πληρέστερη και ακριβέστερη συνεισφορά στην βελτίωση και εξέλιξη της Οδικής Ασφάλειας. Τα κυριότερα ζητήματα και ελλείψεις που συναντήθηκαν κατά τη διάρκεια της Βιβλιογραφικής Ανασκόπησης σχετίζονται με την φύση των δεδομένων που αξιοποιήθηκαν, αφού στην πλειονότητά τους επρόκειτο για πειράματα οδήγησης και όχι φυσικά στοιχεία, καθώς και με την έλλειψη εφαρμογής ποικίλλων μοντέλων ταξινόμησης. Επιπλέον, οι περισσότερες έρευνες προσανατολίστηκαν στον εντοπισμό και ταξινόμηση σε περισσότερα από δύο επίπεδα ασφαλείας.

Η παρούσα Διπλωματική Εργασία επικεντρώνεται στην ανάλυση πληθώρας δεδομένων φυσικής οδήγησης με πολλούς δείκτες οδικής συμπεριφοράς και φόρτου κυκλοφορίας, ενώ οι αναλύσεις ταξινόμησης δύο επιπέδων ασφαλείας θα πραγματοποιηθούν με μεγάλη ποικιλία, προκειμένου να προκύψει μία ενδεδειγμένη μέθοδος ανάλυσης της Οδικής συμπεριφοράς.

3. Θεωρητικό Υπόβαθρο

3.1 Εισαγωγή

Στο παρόν κεφάλαιο περιγράφεται αναλυτικά το θεωρητικό πλαίσιο που αξιοποιήθηκε για να εκπονηθεί το συγκεκριμένο ερευνητικό έργο. Η συγκεκριμένη ενότητα αποδίδει το μαθηματικό και στατιστικό πλαίσιο που εφαρμόστηκε για την ανάλυση των δεδομένων και των διακριτών τεχνικών της Μηχανικής Μάθησης που αναπτύχθηκαν, καθώς και τα κριτήρια αποδοχής μοντέλου, βάσει των μετρικών τους αξιολογήσεων. Ήτοι, η παρούσα διπλωματική εργασία διαρθρώνεται σε 3 κύρια επιμέρους στάδια. Σε πρώτο στάδιο περιγράφεται το υπόβαθρο για την διαδικασία της επιλογής χαρακτηριστικών “Feature Selection” (Liu et al., 2011), με την συμβολή μοντέλων Παλινδρόμησης. Μετέπειτα, αναλύεται η διαδικασία της προ-επεξεργασίας των δεδομένων, που προέκυψαν από την σημαντικότητα χαρακτηριστικών και οι διαφορετικές τεχνικές που εφαρμόστηκαν. Ακολουθεί η παραγωγή των αποτελεσμάτων με την βοήθεια μοντέλων ταξινόμησης και τα κριτήρια αποδοχής των μοντέλων που ελέγχθηκαν.

3.2 Επιλογή Χαρακτηριστικών

Η διαδικασία της Επιλογής Χαρακτηριστικών βασίζεται στην ιδέα της επιλογής των κατάλληλων εξαρτημένων μεταβλητών, που έχουν την σημαντικότερη επιρροή στην ταξινόμηση των αποτελεσμάτων. Πρόκειται για διαδικασία βελτιστοποίησης της διαδικασίας ταξινόμησης, προκειμένου να μην παρατηρηθούν φαινόμενα εκτροπής των τελικών αποτελεσμάτων λόγω περίσσειας όγκου μεταβλητών εισόδου. Είναι μια διαδικασία απλοποίησης των μοντέλων για βελτίωση της ερμηνείας τους από τους ερευνητές, αποσκοπώντας στον καλύτερο έλεγχο της ανάλυσης, καθώς όταν υπεισέρχονται παραπάνω μεταβλητές, οι μαθηματικοί χώροι της ανάλυσης αυξάνονται με εκθετικό ρυθμό στις Ευκλείδειες διαστάσεις με αποτέλεσμα να παρατηρείται το φαινόμενο της κατάρας διαστάσεων (Bellman, 1957). Ο βασικός άξονας είναι ότι οι μεταβλητές περιέχουν ορισμένα χαρακτηριστικά που δυνητικά αποδεικνύονται περιττά και μπορούν να αφαιρεθούν χωρίς σημαντική απώλεια πληροφορίας. Για την επιλογή των κατάλληλων μεταβλητών πραγματοποιήθηκε συνδυασμός της συσχέτισης τους και της διαδικασίας Σημαντικότητας Χαρακτηριστικών (Feature Importance), με την χρήση μοντέλων Παλινδρόμησης.

3.2.1 Συσχέτιση

Η συσχέτιση (correlation) δύο μεταβλητών αντανακλά τον τρόπο με τον οποίο αυτές συμμεταβάλλονται, με την παραδοχή ότι αυτό συμβαίνει με γραμμικό τρόπο, έτσι ώστε να προσδιοριστεί αν και πόσο υπάρχει αιτιώδης σχέση η μη, μεταξύ τους. Η συσχέτιση περιγράφεται με τον δειγματικό συντελεστή γραμμικής συσχέτισης του **Pearson** (Pearson moment), ο οποία προϋποθέτει κανονική κατανομή των μεταβλητών, συμβολίζεται με r και ορίζεται από την μαθηματική εξίσωση 3.1:

(3.1)

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

Όπου

x_i, y_i : οι τιμές των δύο μεταβλητών

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

Οι αριθμητικές τιμές του συντελεστή κυμαίνονται από -1.00 έως και 1.00, με τις μη μηδενικές τιμές να δηλώνουν κάποια συσχέτιση των μεταβλητών είτε αρνητική, δηλαδή όταν μειώνεται η μία μεταβλητή παρατηρείται αύξηση της άλλης, είτε θετική, δηλαδή παρατηρείται θετική συμμεταβολή των μεταβλητών. Όταν ο συντελεστής r έχει μηδενική τιμή, τότε οι επιλεγείσες μεταβλητές καλούνται ασυσχέτιστες. Πιο συγκεκριμένα, για την διαδικασία επιλογής χαρακτηριστικών, σημαίνοντα ρόλο παίζουν οι μεταβλητές που έχουν απόλυτη τιμή συντελεστή συσχέτισης Pearson ο οποίος προσεγγίζει την μονάδα. Ένα σύνηθες πρόβλημα που αντιμετωπίζεται στην έρευνα είναι η πολυσυγγραμμικότητα, δηλαδή όταν δύο ή περισσότερες μεταβλητές συνδέονται με υψηλή συσχέτιση, και πρόκειται για φαινόμενο που οδηγεί σε επιδείνωση της απόδοσης ορισμένων αλγορίθμων. Το παραπάνω φαινόμενο συναντάται συχνά κατά την Γραμμική Παλινδρόμηση, όπου μία από τις πολυσυγγραμμικές μεταβλητές πρέπει να αφαιρεθεί, προκειμένου να βελτιωθεί η ικανότητα του μοντέλου.

3.2.2 Σημαντικότητα Χαρακτηριστικών

Το δεύτερο κριτήριο με το οποίο θα γίνει η επιλογή χαρακτηριστικών είναι η διαδικασία εντοπισμού της σημαντικότητάς των δεδομένων, η οποία καλείται Σημαντικότητα Χαρακτηριστικών (Feature Importance). Εν προκειμένω, η Σημαντικότητα Χαρακτηριστικών επιτυγχάνεται με την εκπαίδευση αλγορίθμων Μηχανικής μάθησης, βασιζόμενους σε μοντέλα Παλινδρόμησης, αποσκοπώντας στον καθορισμό και ποσοτικοποίηση του βαθμού επιρροής των δεδομένων των ανεξάρτητων μεταβλητών, στις επιλεγείσες εξαρτημένες. Ωστόσο, διαφορετικές μέθοδοι Σημαντικότητας Χαρακτηριστικών είναι πιθανό να υπολογίσουν διαφορετικές βαθμολογίες, ακόμα και για το ίδιο σύνολο δεδομένων και Παλινδρόμησης, με αποτέλεσμα το ενδεχόμενο αστάθειας στα συμπεράσματα (Rajbahadur et al., 2021). Προκειμένου να πραγματοποιηθεί σωστή απόδοση στα συμπεράσματα, η παρούσα Διπλωματική Εργασία εξετάζει πληθώρα αλγορίθμων παλινδρόμησης και τα αποτελέσματα προκύπτουν εν τη συνδυάσει με την συσχέτιση Pearson.

3.3 Αλγόριθμοι Παλινδρόμησης και Ταξινόμησης

Στην επιστήμη της ανάλυσης δεδομένων και της μηχανικής μάθησης, η ανάλυση Παλινδρόμησης (Regression analysis) περιγράφει ένα σύνολο στατιστικών διαδικασιών για την εκτίμηση των σχέσεων μεταξύ εξαρτημένων και ανεξάρτητων μεταβλητών. Συχνά στην μηχανική μάθηση, τα μοντέλα παλινδρόμησης αξιοποιούνται ως μια μέθοδος μοντελοποίησης μιας τιμής στόχου, που βασίζεται σε ανεξάρτητους προγνωστικούς παράγοντες. Αυτή η μέθοδος χρησιμοποιείται κυρίως για την προβλεπτικής της ικανότητα και την εύρεση της σχέσης αιτίου και αποτελέσματος μεταξύ

των μεταβλητών. Οι τεχνικές παλινδρόμησης διαφέρουν ως επί το πλείστον με βάση τον αριθμό των ανεξάρτητων μεταβλητών και τον τύπο σχέσης μεταξύ των ανεξάρτητων και των εξαρτημένων μεταβλητών. Βάσει αυτής της διαφοροποίησής τους, οι αλγόριθμοι Παλινδρόμησης χωρίζονται σε γραμμικοί και μη-γραμμικοί, με τους δεύτερους να έχουν την δυνατότητα να λειτουργήσουν αποτελεσματικότερα σε ετεροσκεδαστικά δεδομένα. Ένα μη γραμμικό μοντέλο έχει γενική μορφή Yi=f(Xi,β)+εi, όπου το Xi είναι το διάνυσμα των προβλεπουσών μεταβλητών και β το διάνυσμα των παραμέτρων.

Επιπλέον, στην Μηχανική Μάθηση, η ταξινόμηση αποτελεί μια διαδικασία αναγνώρισης και ομαδοποίησης στοιχείων εκπαίδευσης σε προκαθορισμένες κλάσεις, αξιοποιώντας ένα ευρύ φάσμα αλγορίθμων για την δημιουργία μοντέλων κατηγοριοποίησης και πρόβλεψης. Η ταξινόμηση εφαρμόζει μια διαδικασία αναγνώρισης μοτίβου με τεχνικές Επιβλεπόμενης Μάθησης (Supervised Learning), κατά την οποία ο αλγόριθμος εκπαιδεύεται από τα παρεχόμενα δεδομένα και βρίσκεται σε θέση να κατηγοριοποιήσει νέα παραγόμενα στοιχεία, με στοιχεία εξόδου μια κλάση, εν αντιθέσει με τις Παλινδρομήσεις που έχουν ως στοιχείο εξόδου μία αριθμητική τιμή.

Στην παρούσα Διπλωματική Εργασία θα αξιοποιηθούν τα μοντέλα Γραμμικής Παλινδρόμησης (Linear Regression), Δένδρων Απόφασης (Decision Trees), Τυχαίων Δασών (Random Forests), Προσαρμοστικής Ενδυνάμωσης (AdaBoost), Ενίσχυσης Κλίσης (Gradient Boosting), Ακραίας Ενίσχυσης Κλίσης (XGBoost), Κ-πλησιέστερων γειτόνων (K-nearest Neighbors), Μηχανών Διανυσμάτων Υποστήριξης (SVM) και Πολυεπίπεδων Αισθητήρων (Multilayered Perceptron).

Στον Πίνακα 3.1, παρουσιάζονται τα μοντέλα Παλινδρόμησης και Ταξινόμησης που αξιοποιήθηκαν με την αντίστοιχη ελληνική και αγγλική ορολογία, καθώς και ο αντίστοιχος συμβολισμός τους που συναντάται σε μεταγενέστερους Πίνακες και Γραφήματα.

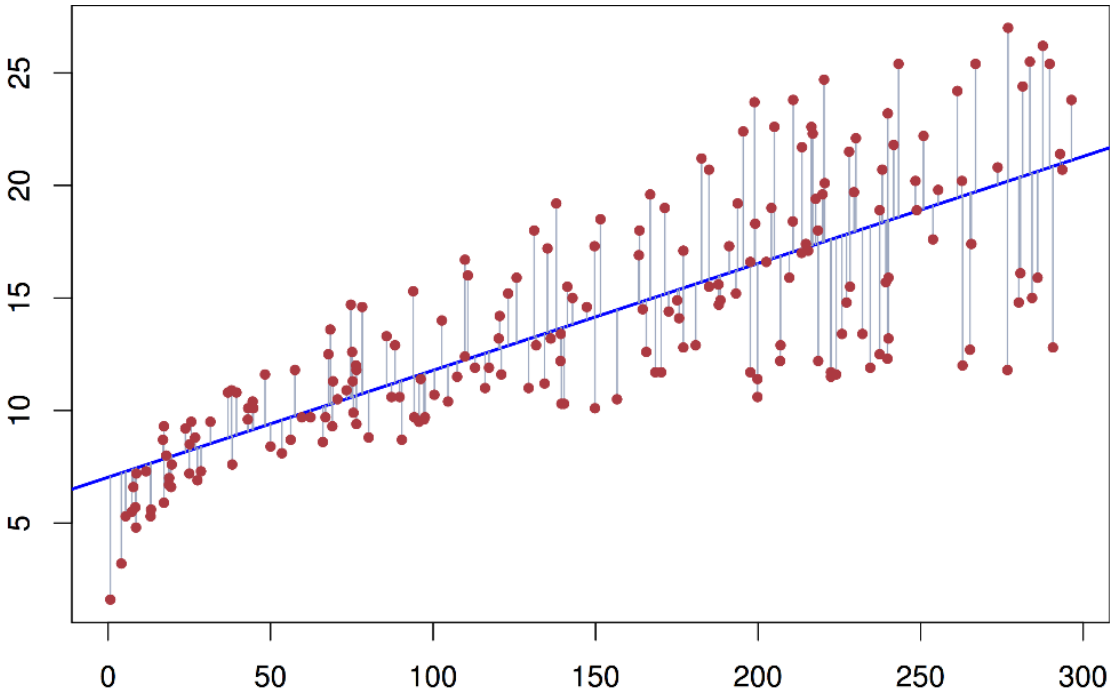
Πίνακας 3.1 : *Ορολογία και συμβολισμός αλγορίθμων*

Ελληνικό όνομα αλγορίθμου	**Αγγλικό όνομα αλγορίθμου**	**Συμβολισμός**
Δένδρα Απόφασης	Decision Trees	DT
Ενίσχυση Κλίσης	Gradient Boosting	GB
Ακραία Ενίσχυση Κλίσης	XGBoost	XGB
Προσαρμοστική Ενδυνάμωση	AdaBoost	AB
Τυχαία Δάση	Random Forests	RF
Κ-πλησιέστεροι γείτονες	K-nearest Neighbors	KNN
Μηχανές Διανυσμάτων Υποστήριξης	Supported Vector Machines	SVM
Πολυεπίπεδοι Αισθητήρες	Multilayered Perceptron	MLP

3.3.1 Γραμμική Παλινδρόμηση

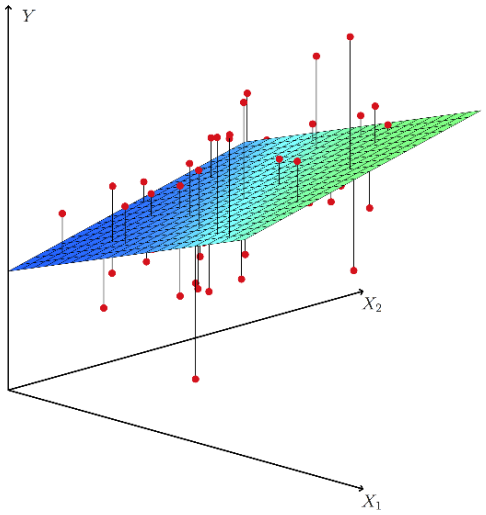
Στη στατιστική, η Γραμμική Παλινδρόμηση είναι μια προσέγγιση για τη μοντελοποίηση της σχέσης μεταξύ μιας βαθμωτής εξαρτημένης μεταβλητής και μίας ή περισσοτέρων ανεξαρτήτων μεταβλητών. Στην Γραμμική Παλινδρόμηση, τα δεδομένα μοντελοποιούνται χρησιμοποιώντας γραμμικές λειτουργίες προγνωστικά, και οι άγνωστες παράμετροι μοντέλου υπολογίζονται από τα δεδομένα. Πιο συγκεκριμένα , η Γραμμική Παλινδρόμηση αναφέρεται σε ένα μοντέλο στο οποίο ο υποθετικός μέσος όρος του Y-εξαρτημένης μεταβλητής δεδομένης της αξίας του X – ανεξάρτητής μεταβλητής, είναι μια συνάρτηση αφινικών Χ, δηλαδή μεταβλητών που διατηρούν

την συγγραμμικότητά και τον λόγο των αποστάσεών τους (Weisstein, 2004). Η Γραμμική Παλινδρόμηση αντιστοιχεί σε μια ευθεία γραμμή ή επιφάνεια, που ελαχιστοποιεί τις αποκλίσεις μεταξύ των προβλεπόμενων και των πραγματικών τιμών εξόδου. Η απλή Γραμμική Παλινδρόμηση είναι ένας τύπος ανάλυσης Παλινδρόμησης, όπου ο αριθμός των ανεξάρτητων μεταβλητών είναι μία και υπάρχει μια γραμμική σχέση μεταξύ της ανεξάρτητης (x) και της εξαρτημένης (y) μεταβλητής.



Εικόνα 3.1: *Μοντελοποίηση Γραμμικής Παλινδρόμησης σε δισδιάστατο χώρο*

Πηγή: *Corrada Bravo, H. (2020) Linear Regression, [Available:https://www.hcbravo.org/IntroDataSci/bookdown-notes/linear-regression.html] (Accessed 17-10-2022)*



Εικόνα 3.2: *Μοντελοποίηση πολυμεταβλητής Γραμμικής Παλινδρόμησης σε τρισδιάστατο χώρο*

Πηγή: *Corrada Bravo, H. (2020) Linear Regression, [Available:https://www.hcbravo.org/IntroDataSci/bookdown-notes/linear-regression.html] (Accessed 17-10-2022)*

Η μπλε γραμμή της Εικόνας 3.1 αναφέρεται ως η καλύτερη ευθεία γραμμή και έχει ως αποτέλεσμα μια δισδιάστατη γραμμική γραφική παράσταση. Με βάση τα παρεχόμενα σημεία δεδομένων, επιχειρείται ο σχεδιασμός μιας γραμμής που μοντελοποιεί τα δεδομένα που αναπαρίστανται από τα σημεία, καλύτερα. Η γραμμή καλύτερης προσαρμογής είναι αυτή για την οποία το συνολικό σφάλμα πρόβλεψης, δηλαδή η απόσταση μεταξύ σημείου και γραμμής Παλινδρόμησης, είναι όσο το δυνατόν μικρότερο. Η γραμμή αυτή περιγράφεται από την σχέση 3.2:

(3.2) $$y_i = \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i = \mathbf{x}_i^{\mathrm{T}} \boldsymbol{\beta} + \varepsilon_i, \qquad i = 1, \ldots, n,$$

Όπου

x_i, y_i : οι τιμές των δύο μεταβλητών

$x_i^{\mathrm{T}}\beta$: το εσωτερικό γινόμενο των διανυσμάτων x_i και β, με $^{\mathrm{T}}$ τον ανάστροφο πίνακα

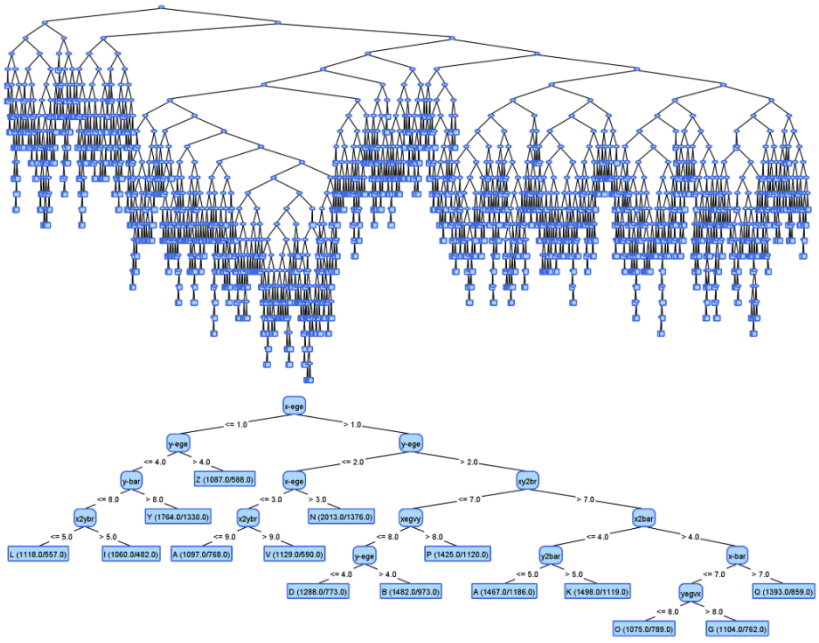
3.3.2 Decision Trees

Ο αλγόριθμος Decision Trees είναι ένα εποπτευόμενο μοντέλο Μηχανικής Μάθησης που αξιοποιείται σε προβλήματα ταξινόμησης και Παλινδρόμησης και υιοθετεί μια προσέγγιση «εκ των άνω προς τα κάτω» (top-down approach) (Apté & Weiss, 1997) . Η λειτουργία του μοντέλου βασίζεται στον αναδρομικό διαμελισμό-recursive partitioning (Lainiotis, 1976), δηλαδή στην δημιουργία δομής με την μορφή ενός δέντρου που αποτελείται από τον πρωταρχικό ριζικό κόμβο, γνωστός και ως πρώτος γονέας, και κατατέμνεται σε επιμέρους θυγατρικούς κόμβους. Οι παράγωγοι κόμβοι με την σειρά τους αποτελούν τους γονικούς κόμβους των επομένων, με διαρκή ακολουθία. Ο βασικός αλγόριθμος για την δημιουργία των Decision Trees ονομάζεται ID3-Iterative Dichotomiser 3 (Quinlan, 1986) και περιγράφεται αναλυτικά στην Εικόνα 3.3. Σε κάθε επαναληπτική διαδικασία, το μοντέλο Decision Trees πραγματοποιεί την βέλτιστη κατάτμηση, βάσει των εισαχθέντων δεδομένων αξιοποιώντας τον δείκτη Gini, ένα μέτρο στατιστικής διασποράς (Deaton, 1982).

Inputs: R: a set of non- target attributes, C: the target **attribute**, S: training data.
Output: returns a decision tree
Start
Initialize to empty tree;
If S is empty **then**
Return a single node failure value
End If
If S is made only for the values of the same target **then**
Return a single node of this value
End if
If R is empty **then**
Return a single node with value as the most common value of the target attribute values found in S
End if
D ← the attribute that has the largest Gain (D, S) among all the attributes of R
{d_j *j = 1, 2, ..., m} ← Attribute values* of D
{S_j *with j = 1, 2, ..., m} ←The subsets of S respectively* constituted of d_j records attribute value D
Return a tree whose root is D and the arcs are labeled by d1, d2, ..., d_m and going to sub-trees ID3 (R-{D}, C, S1), ID3 (R-{D} C, S2), .., ID3 (R-{D}, C, Sm)
End

Εικόνα 3.3: *Ψευδοκώδικας του αλγορίθμου ID3*

Πηγή: *Hssina, Badr & MERBOUHA, Abdelkarim & Ezzikouri, Hanane & Erritali, Mohammed. (2014). A comparative study of decision tree ID3 and C4.5. (IJACSA) International Journal of Advanced Computer Science and Applications. Special Issue on Advances in Vehicular Ad Hoc Networking and Applications, [Available:https://www.researchgate.net/publication/265162251_A_comparative_study_of_decision_tree_ID3_and_C45] (Accessed 17-10-2022)*

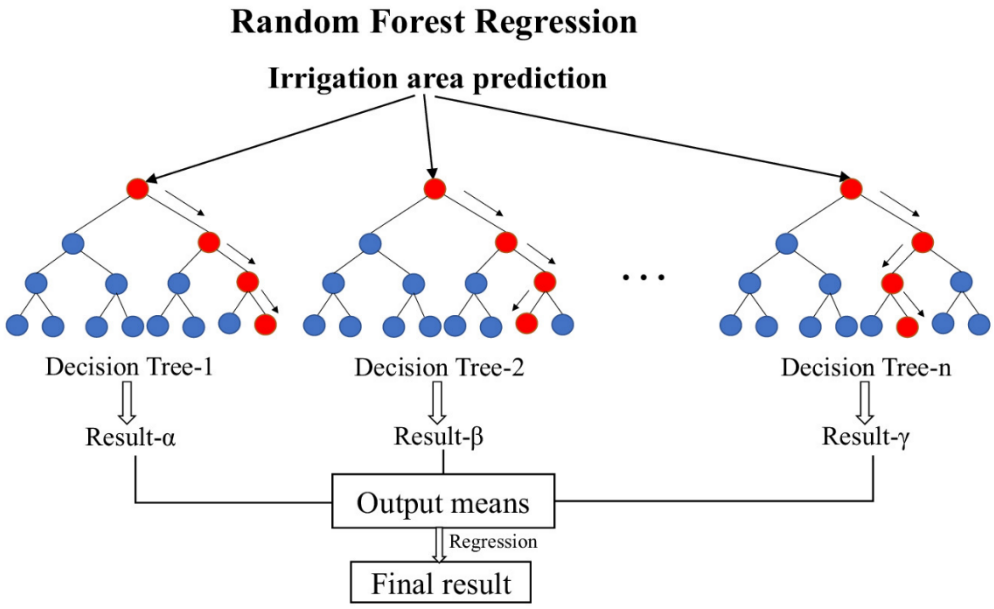


Εικόνα 3.4: *Τυπικό διάγραμμα Decision Trees*

Πηγή: *Stiglic G, Kocbek S, Pernek I, Kokol P. Comprehensive decision tree models in bioinformatics. PLoS One. 2012;7(3):e33812, [Available:https://doi.org/10.1371/journal.pone.0033812] (Accessed 17-10-2022)*

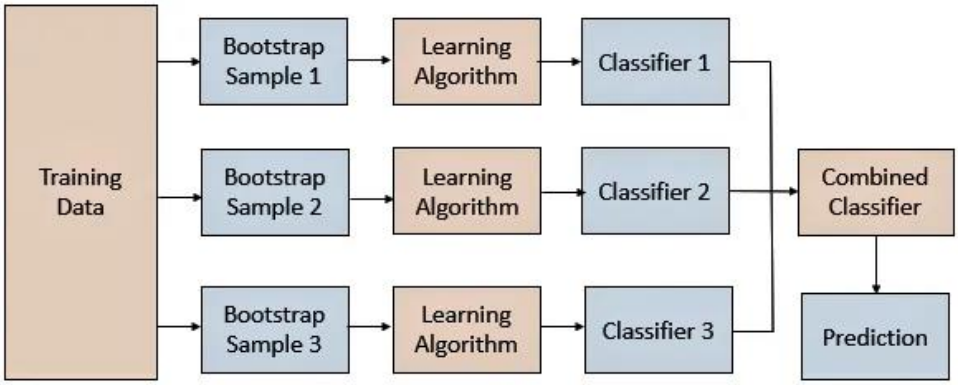
3.3.3 Random Forests

Ο αλγόριθμος Random Forests είναι ένας εποπτευόμενος αλγόριθμος Μηχανικής Μάθησης, που χρησιμοποιεί τη μέθοδο εκμάθησης Συνόλου για Παλινδρόμηση και ταξινόμηση. Η μέθοδος εκμάθησης Συνόλου είναι μια τεχνική που συνδυάζει προβλέψεις από πολλαπλούς αλγόριθμους μηχανικής μάθησης για να κάνει μια πιο ακριβή πρόβλεψη από ένα μεμονωμένο μοντέλο. Πιο εξειδικευμένα, ο συγκεκριμένος αλγόριθμος συνδυάζει μεθόδους εκμάθησης Συνόλου με το πλαίσιο δέντρου αποφάσεων, για να δημιουργήσει πολλαπλά τυχαία σχεδιασμένα δέντρα αποφάσεων από τα δεδομένα, λαμβάνοντας τον μέσο όρο των αποτελεσμάτων για να εξάγει ένα νέο αποτέλεσμα που συχνά οδηγεί σε ισχυρές προβλέψεις και ταξινομήσεις. Η εκπαίδευση των πολλαπλών δένδρων αποφάσεων που κατασκευάζονται, βασίζεται στην τεχνική που ονομάζεται Bootstrap Aggregation και πρόκειται για την διαδικασία τυχαίας δειγματοληψίας υποσυνόλων από ένα ευρύτερο υπερσύνολο, σε έναν δεδομένο αριθμό επαναλήψεων και δεδομένο αριθμό μεταβλητών. Η διαδικασία του Bootstrap Aggregation είναι ένας ειδικά καταρτισμένος μετα-ευρετικός αλγόριθμος προκειμένου να βελτιστοποιεί την Ευστάθεια και την Ακρίβεια του μοντέλου (Brodeur et al., 2020).



Εικόνα 3.5: *Τυπικό διάγραμμα αλγορίθμου Random Forests*

Πηγή: *Wang, M., Huang, H., Gao, G., & Tang, W. Trend prediction of irrigation area using improved random forest regression. Irrigation and Drainage, [Available:https://doi.org/10.1002/ird.2695] (Accessed 17-10-2022)*

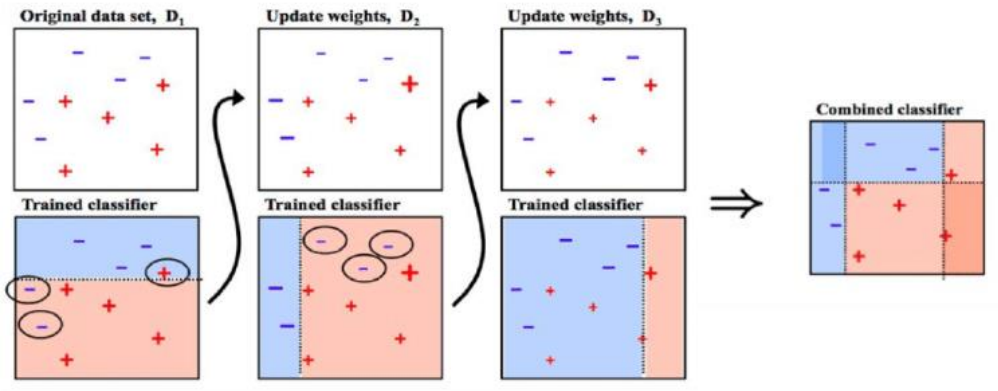


Εικόνα 3.6: *Διάγραμμα Ροής αλγορίθμου Bootstrap Aggregation*

Πηγή: *Packt, Big Data and Business Intelligence, Bootstrap Aggregation. 2022. [Available:https://subscription.packtpub.com/book/big-data-and-business-intelligence/9781789136609/5/ch05lvl1sec29/bootstrap-aggregation] (Accessed 17-10-2022)*

3.3.4 AdaBoost

Το AdaBoost αποτελεί μία τεχνική που χρησιμοποιείται ως Μέθοδος Συνόλου σε διαδικασίες Μηχανικής Μάθησης. Το συγκεκριμένο μετα-ευρετικό μοντέλο αξιοποιεί τον αλγόριθμο Decision Trees ενός επιπέδου, δηλαδή δένδρων με κολομβώματα απόφασης (decision stumps), υπεισέρχοντας αρχικά τα δεδομένα ως ισοβαρή-ισοσταθμισμένα. Στη συνέχεια, ακολουθείται επαναληπτική διαδοχική διαδικασία, κατά την οποία τα δεδομένα που ταξινομήθηκαν με λάθος τρόπο ή και τα σφάλματα πρόβλεψης, προσαρμόζονται, αποκτώντας υψηλότερα βάρη, προκειμένου να διορθωθούν τα λάθη ταξινόμησης. Για αυτόν τον λόγο καλείται και Προσαρμοστικός αλγόριθμος. Σε κάθε επανάληψη της διαδικασίας εκπαίδευσης, βάρος ίσο με το σφάλμα εκχωρείται σε κάθε λανθασμένα ταξινομημένο στοιχείο. Ο αλγόριθμος καταλήγει σε πληθώρα ενισχυμένων δένδρων απόφασης (Boosted Decision Trees -BDT) (Marsh, 2016).

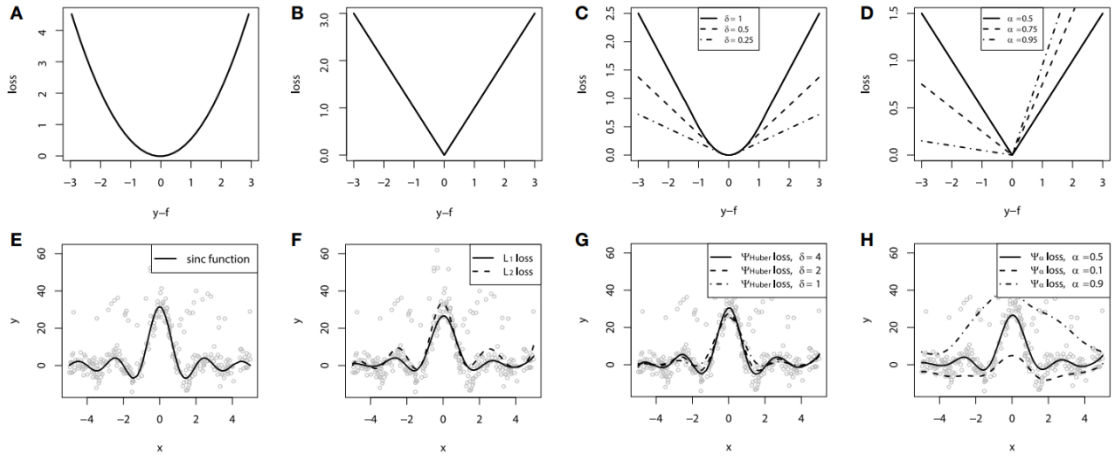


Εικόνα 3.7: *Διάγραμμα Ροής εκπαίδευσης αλγορίθμου AdaBoost*

Πηγή: *Marsh, Brendan. (2016). Multivariate Analysis of the Vector Boson Fusion Higgs Boson. [Available:https://www.researchgate.net/publication/306054843_Multivariate_Analysis_of_the_Vector_Boson_Fusion_Higgs_Boson]*

3.3.5 Gradient Boosting

Ο αλγόριθμος Gradient Boosting αποτελεί μία μέθοδο Συνόλου, με συχνή εφαρμογή σε προβλήματα παλινδρόμησης και ταξινόμησης, βασιζόμενος στην λειτουργία δένδρων αποφάσεων. Ως μέθοδος Συνόλου, αξιοποιεί πολλαπλούς αλγορίθμους εκμάθησης προκειμένου να αποδώσει την καλύτερη προγνωστική δυνατότητα, με την μορφή ενισχυμένου δένδρου απόφασης (BDT). Η κυρίαρχη ιδέα πίσω από αυτόν τον αλγόριθμο είναι η κατάλληλη κατασκευή των πρωταρχικών μοντέλων που χρησιμοποιεί, έτσι ώστε να είναι μέγιστα συσχετισμένα με την αρνητική κλίση (gradient) μιας αυθαίρετης και διαφορίσιμης συνάρτησης απωλειών (Natekin & Knoll, 2013). Η λειτουργία του αλγορίθμου είναι παρεμφερής με αυτή του AdaBoost, με την ειδοποιό διαφορά ότι τα παραγόμενα σφάλματα από ένα μοντέλο πρόβλεψης, προσαρμόζονται στο επόμενο μοντέλο που ακολουθεί, χωρίς να παρατηρείται μετασχηματισμός του βάρους/στάθμισης κάθε δεδομένου.



Εικόνα 3.8: *Συνήθεις Συνεχείς Συναρτήσεις Απωλειών στην Ενίσχυση Κλίσης: (A) L_2 squared; (B) L_1 absolute; (C) Huber loss function; (D) Quantile loss function,*

Προσαρμογή GBM σε συνάρτηση καρδινάλιου ημιτόνου sinc(x) δεδομένων με θόρυβο: (E) original sinc(x) function; (F) ήπιο GBM με απώλειες από L_1, L_2; (G) ήπιο GBM με απώλειες Huber; (H) ήπιο GBM με Quantile απώλειες

Πηγή: *Natekin Alexey, Knoll Alois, Gradient Boosting Machines, A Tutorial, Frontiers in Neurorobotics, vol.7, 2013 [Available: https://doi.org/10.3389/fnbot.2013.00021]*

3.3.6 XGBoost

Ο αλγόριθμος XGBoost αναπτύχθηκε από τους Tianqi Chen και Carlos Guestrin και αποτελεί μια βελτιστοποιημένη μορφή του μοντέλου Gradient Boosting. Πρόκειται για ένα μοντέλο που λειτουργεί ως αλγόριθμος Newton-Raphson, με την βοήθεια δευτέρας τάξεως προσέγγισης Taylor στον συναρτησιακό χώρο, εν αντιθέσει με το Gradient Boosting, που στηρίζεται στην κάθοδο βασισμένη στην κλίση (gradient descent). Πιο συγκεκριμένα, το XGBoost είναι μια υλοποίηση δέντρων αποφάσεων με ενισχυμένη κλίση, στον οποίο τα δέντρα δημιουργούνται με διαδοχικό τρόπο και κατά βάση αποδίδει αρκετά μεγαλύτερη ακρίβεια μοντέλου, σε μικρότερο υπολογιστικό χρόνο εκπαίδευσης, από τα συνήθη μοντέλα μηχανικής μάθησης. Η εφαρμογή του προϋποθέτει τεχνικές συμπίεσης των εκατοντάδων χιλιάδων παραγόμενων δένδρων αποφάσεων σε ένα τελικό. Βασικός παράγοντας που επηρεάζει το συγκεκριμένο μοντέλο είναι η στάθμιση. Η στάθμιση των μεταβλητών που προβλέπονται λανθασμένα αυξάνεται, με αποτέλεσμα την εκ νέου

δημιουργία δένδρου, καταλήγοντας συνδυαστικά στο ενιαίο τελικό. (Chen & Guestrin, 2016). Ο αλγόριθμος περιγράφεται περιληπτικά από την κάτωθι εξίσωση 3.3:

(3.3)

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i), f_k \epsilon \mathcal{F}$$

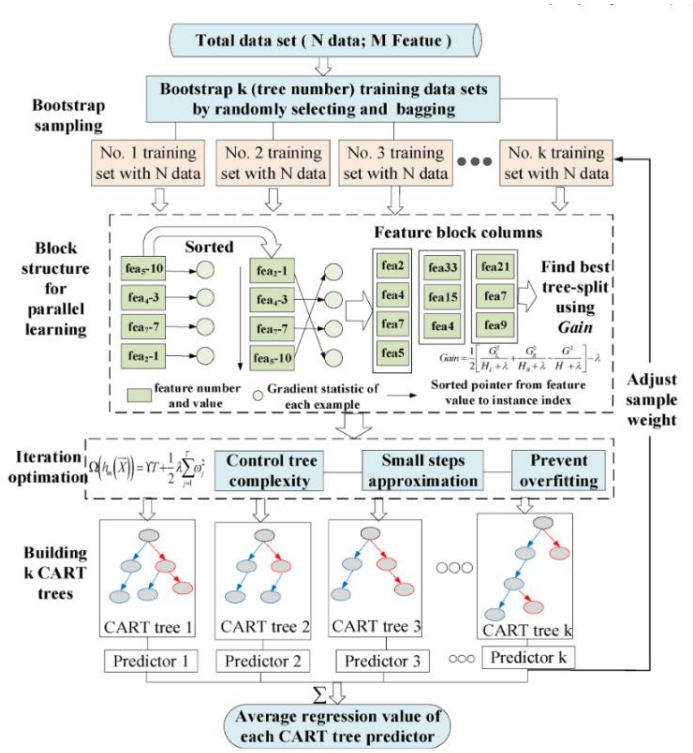
Όπου

x_i, y_i : οι τιμές των δύο μεταβλητών,

K: ο αριθμός των δέντρων,

F: το σύνολο των πιθανών δέντρων

f: ο συναρτησιακός χώρος του F

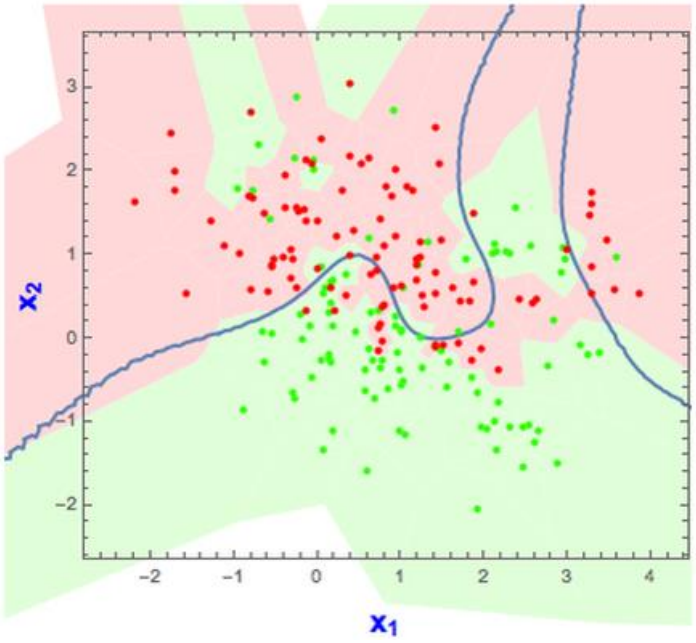


Εικόνα 3.9: *Διάγραμμα ροής μοντέλου XGBoost*

Πηγή: *Zhifen Zhang, Yiming Huang, Rui Qin, Wenjing Ren, Guangrui Wen,XGBoost-based on-line prediction of seam tensile strength for Al-Li alloy in laser welding: Experiment study and modelling, Journal of Manufacturing Processes. 2021, [Available: https://doi.org/10.1016/j.jmapro.2020.12.004] (Accessed 19-10-2022)*

3.3.7 K-nearest Neighbors

Ο αλγόριθμος K-nearest Neighbors (KNN) αποτελεί έναν από τους πιο σημαντικούς και απλούς σε εφαρμογή σε προβλήματα ταξινόμησης, καθώς δεν απαιτεί πρωταρχική γνώση για την διανομή (distribution) των δεδομένων, υπερκαλύπτοντας και το εμπόδιο των παραμετρικών εκτιμήσεων των πυκνοτήτων πιθανότητας που είναι δύσκολο να αποσαφηνιστούν (Peterson, 2009). Για αυτόν τον λόγο, το συγκεκριμένο μοντέλο λειτουργεί αποτελεσματικότερα σε σχετικώς μικρά σύνολα δεδομένων. Ο ταξινομητής KNN βασίζεται στην θεώρηση της Ευκλείδειας απόστασης μεταξύ ενός δείγματος δεδομένων δοκιμής (test data sample) και των δειγμάτων δεδομένων εκπαίδευσης (training data samples), εντοπίζοντας το διάνυσμά τους. Πιο απλουστευμένα, η λειτουργία του αλγορίθμου ταξινομεί τα δεδομένα, βάσει των κοινών τους χαρακτηριστικών, καθώς υποθέτει ότι αυτά βρίσκονται σε κοντινούς χώρους. Το αποτέλεσμα του αλγορίθμου είναι η κατάτμηση του χώρου των δεδομένων στις ζητούμενες κλάσεις και η απεικόνιση του κατατμημένου χώρου περιγράφεται από την ψηφίδωση Voronoi, το 1907.



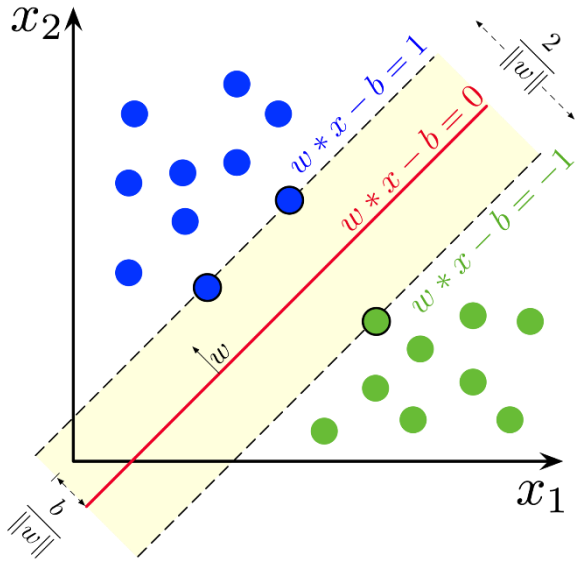
Εικόνα 3.10: *Οπτικοποίηση αλγορίθμου Κ-πλησιέστερων γειτόνων με διάγραμμα Voronoi*

Πηγή: *Ian McLeod, "k-Nearest Neighbor (kNN) Classifier".(2011) [Available:https://demonstrations.wolfram.com/KNearestNeighborKNNClassifier/] (Accessed 18-10-2022)*

3.3.8 Supported Vector Machines

Τα SVM αποτελούν έναν αλγόριθμο Εποπτευόμενης Μηχανικής Μάθησης, εκτελώντας βελτιστοποίηση τετραγωνικού προγραμματισμού (Zhongwen &.Huanghuang, 2017). Απώτερος στόχος του αλγορίθμου, είναι ο εντοπισμός της εξίσωσης ενός βέλτιστου υπερπλάνου (hyperplane), το οποίο διαχωρίζει τα πολυεπίπεδα δεδομένα διαφορετικών κλάσεων. Γενικότερα, η λειτουργία των SVM βασίζεται σε ένα μοντέλο εύρεσης των κοινών χαρακτηριστικών ανάμεσα

στις κλάσεις, εν αντιθέσει με τους συνήθεις αλγόριθμους, οι οποίοι εντοπίζουν τα στοιχεία που τις διαφοροποιούν. Τα δεδομένα με κοινά χαρακτηριστικά μεταξύ δύο κλάσεων καλούνται Διανύσματα Υποστήριξης. Πιο συγκεκριμένα, σε δισδιάστατο χώρο, η εξίσωση του βέλτιστου υπερεπιπέδου περιγράφεται με γραμμική μορφή και ορίζεται ως η ευθεία με την μέγιστη και ισαπέχουσα απόσταση μεταξύ των Διανυσμάτων Υποστήριξης των κλάσεων (Εικόνα 3.11).

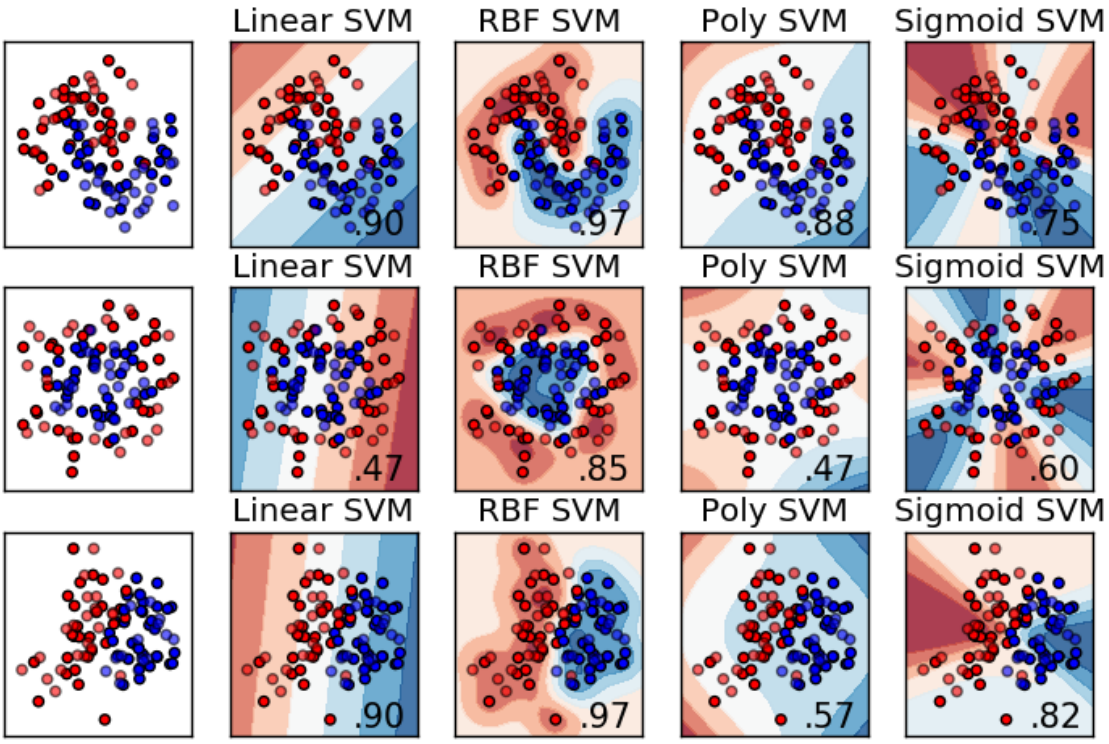


Εικόνα 3.11: *Απεικόνιση βέλτιστου υπερπλάνου δεδομένων δύο κλάσεων σε δισδιάστατο χώρο*

Πηγή: *Wikiwand, Support Vector Machine (2018), [Available: https://www.wikiwand.com/en/Support_vector_machine] (Accessed 18-10-2022)*

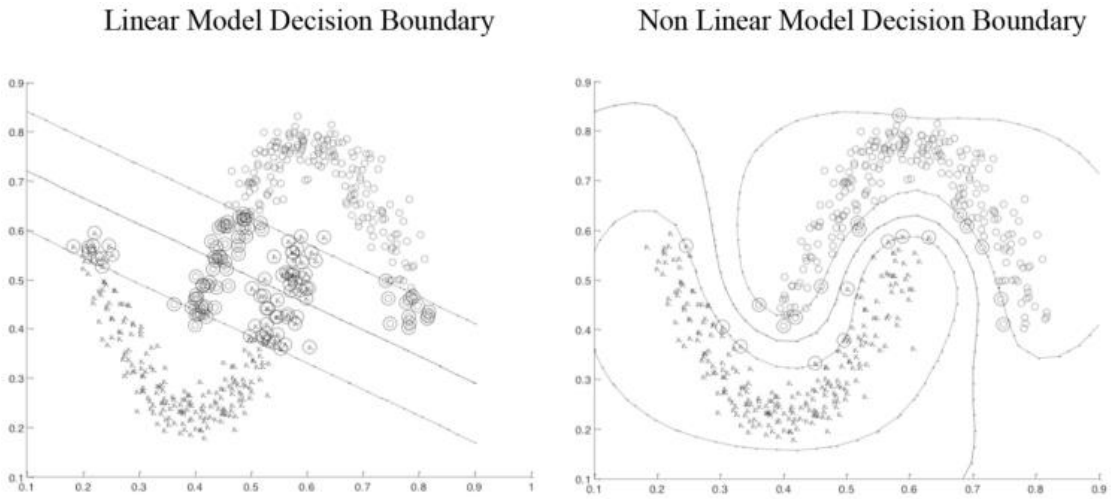
Σε τρισδιάστατο χώρο, η τεχνική την οποία αξιοποιεί ο αλγόριθμος, ονομάζεται μέθοδος των Πυρήνων (Kernels) και συμβάλλει στον κατάλληλο μετασχηματισμό των δεδομένων εισόδου περιορισμένων διαστάσεων, για να αντιστοιχιστούν σε χώρους χαρακτηριστικών πολλαπλών διαστάσεων (Εικόνα 3.12). Στην τελεστική θεωρία, οι συναρτήσεις πυρήνων εισήχθησαν από τον James Mercer και χρησιμοποιούνται εκτεταμένα σε αναλύσεις Fourier, Θεωρία Πιθανοτήτων, Θεωρία Πεπλεγμένων Συναρτήσεων και στην Μηχανική Μάθηση. Οι βασικότερες εκ των συναρτήσεων πυρήνων που αξιοποιούν τα Supported Vector Machines και ορίζονται στον Ευκλείδειο χώρο, περιλαμβάνουν:

- Γραμμικός Πυρήνας: $K(\mathbf{x}, \mathbf{y}) = \mathbf{x}^T \mathbf{y}, \quad \mathbf{x}, \mathbf{y} \in \mathbb{R}^d$
- Πολυωνυμικός Πυρήνας: $K(\mathbf{x}, \mathbf{y}) = (\mathbf{x}^T \mathbf{y} + r)^n, \quad \mathbf{x}, \mathbf{y} \in \mathbb{R}^d, r \geq 0, n \geq 1$
- Συνάρτηση ακτινικής βάσης (RBF): $K(\mathbf{x}, \mathbf{y}) = e^{-\frac{\|\mathbf{x}-\mathbf{y}\|^2}{2\sigma^2}}, \quad \mathbf{x}, \mathbf{y} \in \mathbb{R}^d, \sigma > 0$
- Λαπλασιανός Πυρήνας: $K(\mathbf{x}, \mathbf{y}) = e^{-\alpha\|\mathbf{x}-\mathbf{y}\|}, \quad \mathbf{x}, \mathbf{y} \in \mathbb{R}^d, \alpha > 0$



Εικόνα 3.12: *Ταξινόμηση με SVM εφαρμόζοντας εναλλακτικές Μεθόδους Πυρήνα*

Πηγή: *Kaggle, Kernels and support vector machine regularization. (2017) [Available: https://www.kaggle.com/residentmario/kernels-and-support-vector-machine-regularization] (Accessed 18-10-2022)*



Εικόνα 3.13: *Διαχωρισμός πυρήνα μεταξύ Γραμμικού και Μη γραμμικού πολυδιάστατου μοντέλου SVM*

Πηγή: *Cherkassky, Vladimir & Dhar, Sauptik. (2010). Simple Method for Interpretation of High Dimensional Nonlinear SVM Classification Models.. 267-272, [Available: https://www.researchgate.net/publication/220705021_Simple_Method_for_Interpretation_of_High-Dimensional_Nonlinear_SVM_Classification_Models] (Accessed 18-10-2022)*

Η Γραμμική Παλινδρόμηση SVM (Linear SVR), εν αντιθέσει με τις Μηχανές Διανυσμάτων Υποστήριξης, δεν αξιοποιεί την μέθοδο Πυρήνα. Πληθώρα ερευνών έχουν καταστήσει το συγκεκριμένο μοντέλο αξιόπιστο για δεδομένα μεγάλης κλίμακας με ταχύτερες διαδικασίες εκπαίδευσης (e.g., Keerthi & DeCoste, 2005; Joachims, 2006; Shalev-Shwartz et al., 2007; Hsieh et al., 2008). Με την διαδικασία εύρεσης βέλτιστου υπερπλάνου, ο συγκεκριμένος αλγόριθμος καθίσταται ως το μοναδικό γραμμικό μοντέλο που δύναται να πραγματοποιήσει διαδικασίες Παλινδρόμησης, σε δεδομένα που δεν είναι γραμμικώς διαχωρίσιμα.

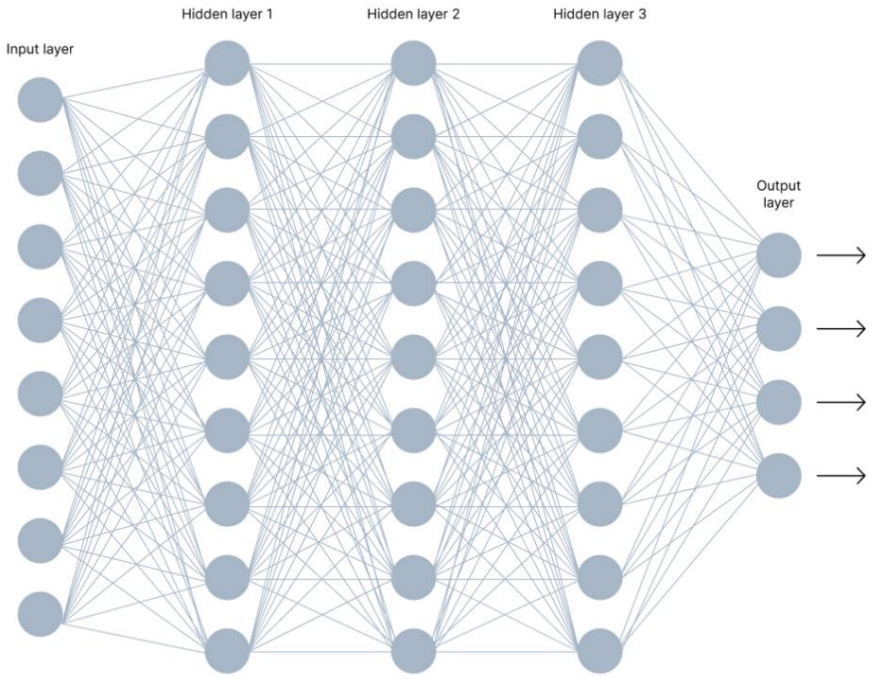
3.3.9 Multilayered Perceptron

Ο αισθητήρας perceptron αποτελεί Νευρωνικό Δίκτυο μονής στρώσης, βασισμένος στους βιολογικούς νευρώνες. Για την κατασκευή τεχνητών νευρώνων, αξιοποιείται η λειτουργία των δενδριτών, υπεισέρχοντας δεδομένα εισόδου και εκχωρώντας τους αντίστοιχα βάρη/στάθμιση, αποστέλλοντάς τα σε επόμενο χρόνο σε συναρτήσεις ενεργοποίησης, για την παραγωγή αποτελεσμάτων. Οι συναρτήσεις ενεργοποίησης (activation functions) είναι ειδικές συναρτήσεις, οι οποίες καθορίζουν το εάν ένας νευρώνας πρέπει να ενεργοποιηθεί ή όχι, βάσει των δεδομένων εισόδου. Ο αισθητήρας perceptron παρέχει μόνο γραμμικές σχέσεις μεταξύ των δεδομένων εισόδου-εξόδου, με αποτέλεσμα να μην επαρκεί για πεπλεγμένες σχέσεις. Αυτό το πρόβλημα έρχονται να λύσουν τα Τεχνητά Νευρωνικά Δίκτυα (Artificial Neural Networks - ANN). Τα Multilayered Perceptrons αποτελούν Τεχνητό Νευρωνικό Δίκτυο (ANN) πρόσω τροφοδότησης, δηλαδή οι επιμέρους συνδέσεις μεταξύ των κόμβων δεν μπορούν να έχουν κυκλική μορφή. Κάθε MLP αποτελείται κατ' ελάχιστον από 3 στρώσεις κόμβων: την στρώση εισόδου, την κρυφή στρώση (hidden layer) και την στρώση εξόδου, όπου κάθε κόμβος, εκτός αυτού της εισόδου, αποτελείται από τεχνητό νευρώνα που χρησιμοποιεί μη γραμμική συνάρτηση ενεργοποίησης. Κάθε νευρώνας κάθε στρώσης συνδέεται με την προηγούμενη και την επόμενη στρώση, με την διαδικασία αυτή να ονομάζεται νευρική σύναψη (synapses). Βασικό στοιχείο απόδοσης στην ταξινόμηση στα MLPs αποτελεί ο αριθμός των κρυφών στρώσεων, μεταξύ αυτών της εισόδου και της εξόδου. Στα MLPs, ο βασικός αλγόριθμος εκμάθησης που χρησιμοποιείται, ονομάζεται Οπισθοδιάδοση (Backpropagation), ένα μοντέλο καθοδικής κλίσης (Gradient descent), που βασίζεται στις θεωρίες του κανόνα αλυσίδας (chain rule) και στην απομνημόνευση (memoization) (Pinkus, 1999). Πιο αναλυτικά, η διαδικασία ενός Multilayered Perceptron περιγράφεται μαθηματικά, ως εξής:

- Η στρώση εισόδου έχει ως αποτέλεσμα εξόδου του στοιχείου j, την τιμή εισόδου x_{0j}.
- Το στοιχείο k της στρώσης i δέχεται αποτέλεσμα εξόδου x_{ij} από κάθε στοιχείο j της (i - 1) στρώσης. Οι τιμές x_{ij} πολλαπλασιάζονται από σταθερές (βάρη/στάθμιση) w_{ijk} και τα αποτελέσματά τους αθροίζονται.
- Μία μετατόπιση θ_{ik}, που ονομάζεται όριο ή μεροληψία, και μετέπειτα μία συνάρτηση ενεργοποίησης σ εφαρμόζονται στο άνωθεν άθροισμα και η τελική τιμή αναπαριστά το αποτέλεσμα εξόδου $x_{i+1,k}$ του στοιχείου k της στρώσης i, το οποίο θα είναι ίσο με:

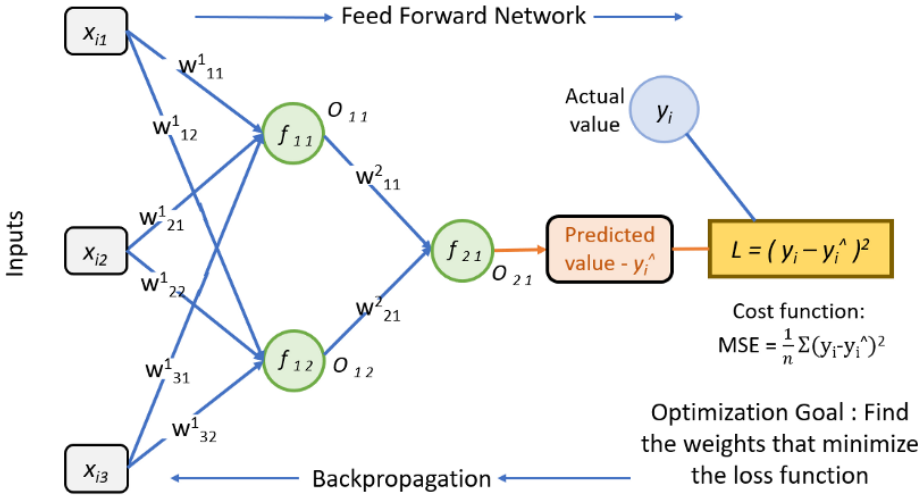
(3.4)

$$x_{i+1,k} = \sigma\left(\sum_j w_{ikj} x_{ij} - \theta_{ik}\right)$$



Εικόνα 3.14: *Αρχιτεκτονική Τεχνητού Νευρωνικού Δικτύου με διασυνδεδεμένους νευρώνες*

Πηγή: *Pragati Baheti, Activation Functions in Neural Networks [12 Types & Use Cases] (2022), [Available: https://www.v7labs.com/blog/neural-networks-activation-functions] (Accessed 19-10-2022)*



Εικόνα 3.15: *Διάγραμμα ροής Πολυεπίπεδου Αισθητήρα NN (MLP NN)*

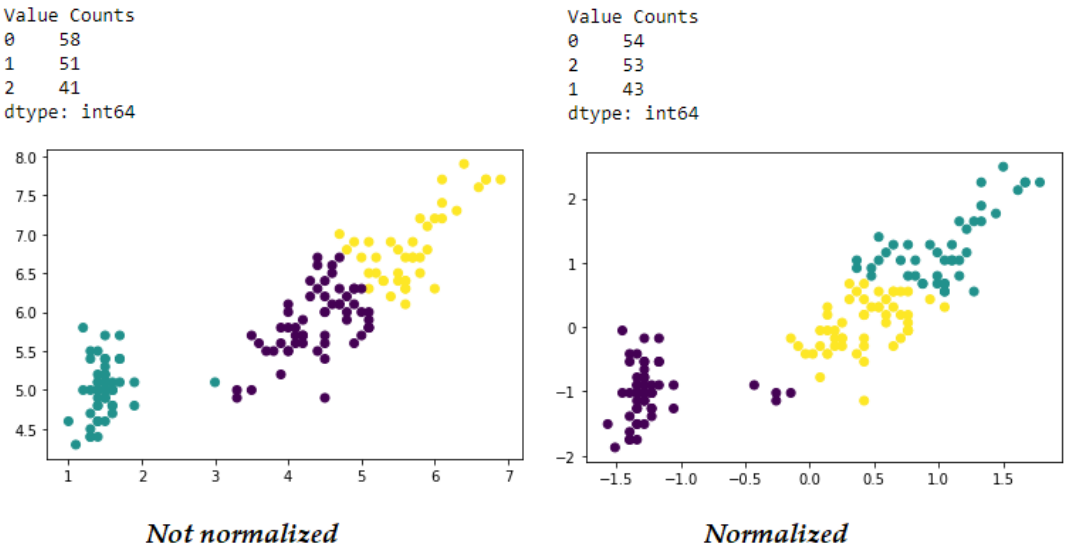
Πηγή: *Saarathi Anbuazhagan, A Complete Guide to train Multi-Layered Perceptron Neural Networks, (2021), [Available: https://paarthasaarathi.medium.com/a-complete-guide-to-train-multi-layered-perceptron-neural-networks-3fd8145f9498] (Accessed 19-10-2022)*

3.4 Τεχνικές Μεταχείρισης Δεδομένων

Για την εκπόνηση της παρούσας Διπλωματικές Εργασίας αξιοποιήθηκαν ορισμένες τεχνικές μεταχείρισης δεδομένων, ως βελτιστοποίηση των διαδικασιών Παλινδρόμησης και ταξινόμησης, επιδιώκοντας στην αξιοπιστία του παρόντος ερευνητικού έργου.

3.4.1 Κανονικοποίηση (Normalization)

Η κανονικοποίηση δεδομένων χρησιμοποιείται στη Μηχανική Μάθηση για να κάνει την εκπαίδευση μοντέλων λιγότερο ευαίσθητη στην κλίμακα εύρους των μεταβλητών. Αυτό, επιτρέπει στο μοντέλο να συγκλίνει σε καλύτερα βάρη και, με τη σειρά του, να οδηγεί σε ένα πιο ακριβές μοντέλο. Η κανονικοποίηση καθιστά τα χαρακτηριστικά πιο συνεπή μεταξύ τους, γεγονός που επιτρέπει στο μοντέλο να προβλέψει τα αποτελέσματα με μεγαλύτερη ακρίβεια. Ως διαδικασία προ-επεξεργασίας, η κανονικοποίηση μετασχηματίζει τα δεδομένα με τρόπο που τα καθιστά αδιάστατα ή και με παρόμοιες κατανομές, δηλαδή προσδίδοντας το ίδιο βάρος σε κάθε επιμέρους μεταβλητή. Εν προκειμένω, κανονικοποίηση εφαρμόστηκε στις ανεξάρτητες μεταβλητές



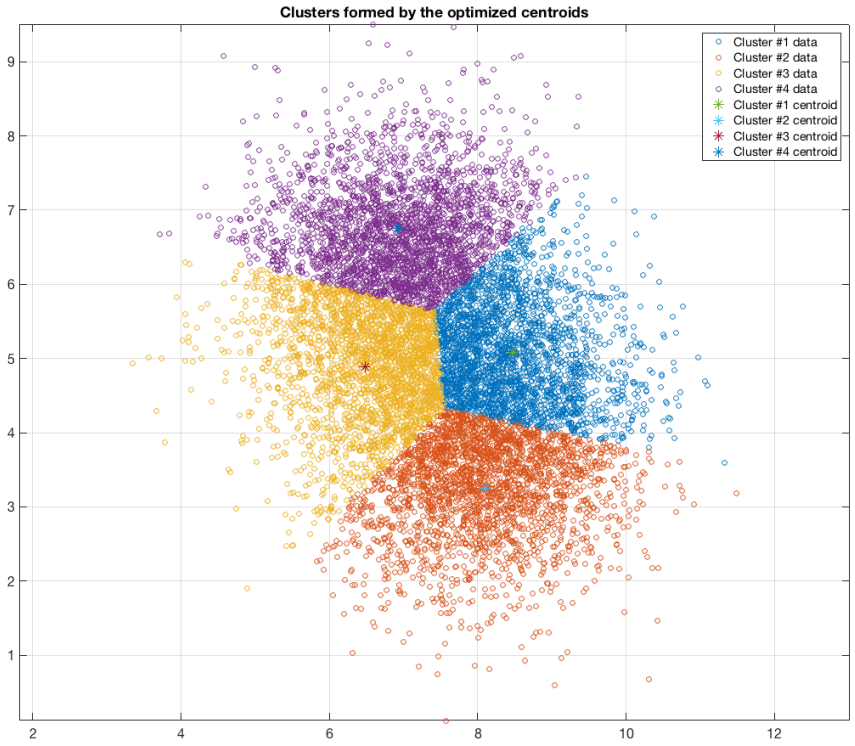
Εικόνα 3.16: *Διαφορές ομαδοποίησης δεδομένων πριν και μετά την κανονικοποίηση*

Πηγή: *Mahbubul Alam, Data normalization in machine learning, TowardsDataScience. (2020), [Available: https://towardsdatascience.com/data-normalization-in-machine-learning-395fdec69d02] (Accessed 19-10-2022)*

3.4.2 Ομαδοποίηση αλγορίθμου K-μέσου

Η διαδικασία της ταξινόμησης των εξαρτημένων μεταβλητών σε τάξεις προϋποθέτει την προ-επεξεργασία και ομαδοποίηση των δεδομένων τους σε δυαδική μορφή. Για να επιτευχθεί αυτή η διαδικασία, πρέπει να οριστούν κάποια όρια τιμών (thresholds) για τον δυαδικό καταμερισμό των

δεδομένων. Ένας τέτοιος αλγόριθμος ομαδοποίησης (clustering) σε πολυεπίπεδα δεδομένα είναι η μέθοδος του K-μέσου, μια ταχεία και απλά εφαρμόσιμη τεχνική (Kodinariya, & Dr. Makwana, 2013). Η ομαδοποίηση K-μέσου πρόκειται για μια διανυσματική κβαντοποίηση των δεδομένων σε δεδομένες συστάδες (clusters), οι οποίες αφορούν τον αριθμό τον κεντροειδών που δημιουργούνται. που κατά την οποία κάθε τιμή μεταβλητής μετασχηματίζεται ώστε να κατανεμηθεί στην συστάδα με τον πλησιέστερο μέσο όρο, δηλαδή στο πλησιέστερο κεντροειδές. Η κατανομή σε κάθε ένα εκ των κεντροειδών πραγματοποιείται με την διαδικασία της μείωσης του αθροίσματος τετραγώνων εντός των συστάδων. Ο αλγόριθμος K-μέσου αποτελεί μια επαναληπτική διαδικασία, κατά την οποία βελτιστοποιείται η θέση των κεντροειδών έως σημείου σταθεροποίησής τους.



Εικόνα 3.17: *Τυπική οπτικοποίηση ομαδοποίησης με αλγόριθμο K-μέσου 4 κεντροειδών*

Πηγή: *sepdek, K-means as a cost function minimization. (2018), [Available: https://georgepavlides.info/k-means-cost-function-minimisation-matlab-octave-approach/] (Accessed 20-10-2022)*

3.4.3 Τεχνικές Υπερδειγματοληψίας δεδομένων σε προβλήματα Μη Ισορροπημένης Μάθησης

Η αξιοπιστία στην διαδικασία της ταξινόμησης οφείλεται και στο γεγονός ότι οι αλγόριθμοι ταξινόμησης προϋποθέτουν κατά βάση την ισορροπία δεδομένων εκπαίδευσης μεταξύ των διαφορετικών τάξεων μεταβλητών. Όμως, στην πλειονότητα των προβλημάτων δεδομένων φυσικής οδήγησης, παρατηρείται ανισορροπία κατανομής στις κλάσεις ενός συνόλου δεδομένων πολλών διαστάσεων, με αποτέλεσμα να αυξάνεται η μεροληψία (bias) των αλγορίθμων προς την κυρίαρχη κλάση (majority class) και την αμφιλεγόμενη αξιοπιστία της ταξινόμησης. Η εφαρμογή τεχνικών Υπερδειγματοληψίας (Oversampling) αντισταθμίζει την άνιση κατανομή του δείγματος

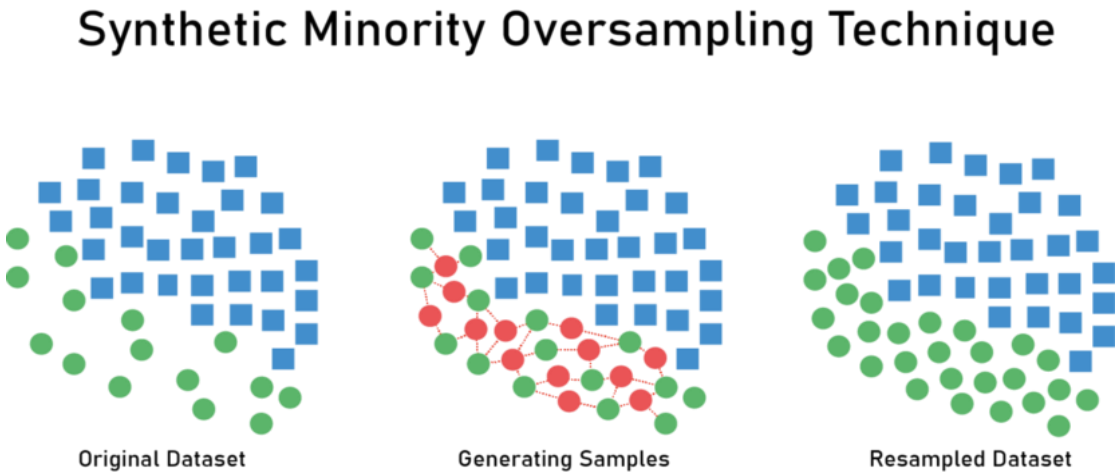
τάξεως μειοψηφίας (minority class) με επανάληψη ή παραγωγή δεδομένων παρεμφερών χαρακτηριστικών (Mohammed et al., 2020). Αντίθετη και ισοδύναμη τεχνική αποτελεί η Υποδειγματοληψία (Undersampling).

Εικόνα 3.18: *Δομικές διαφορές μεταξύ υποδειγματοληψίας και υπερδειγματοληψίας*

Πηγή: *R. Mohammed, J. Rawashdeh and M. Abdullah, "Machine Learning with Oversampling and Undersampling Techniques: Overview Study and Experimental Results," 2020 11th International Conference on Information and Communication Systems (ICICS), 2020, pp. 243-248 [Available at doi: 10.1109/ICICS49469.2020.239556]*

Η συνηθέστερη τεχνική Υπερδειγματοληψίας, η οποία και εφαρμόστηκε στην εκπόνηση της Εργασίας, είναι η μέθοδος Τεχνικής Συνθετικής Μειονοτικής Υπερδειγματοληψίας (Synthetic Minority Oversampling Technique – SMOTE). Η μέθοδος SMOTE κατασκευάστηκε από τους Chawla N. V., Bowyer K. W., Hall, L. O., Kegelmeyer, W. P. το 2002, προσβλέποντας στην επίλυση της ανισορροπίας δεδομένων μειοψηφικής τάξης, παράγοντας τεχνητά σημεία δεδομένων βασιζόμενα στα αρχικά σημεία με ανεπαίσθητες διαφοροποιήσεις. Ο αλγόριθμος SMOTE βασίζεται στις επερχόμενες διαδικασίες:

- Τυχαία επιλογή σημείου από την μειοψηφική τάξη (minority class).
- Προσδιορισμός των K-πλησιέστερων γειτόνων (K-nearest Neighbors) του σημείου.
- Επιλογή ενός από τους K γείτονες και καθορισμός του διανύσματος μεταξύ του επιλεγμένου γείτονα και του τυχαίου σημείου.
- Πολλαπλασιασμός του διανύσματος με τυχαίο αριθμό κυμαινόμενο από 0 έως 1.
- Πρόσθεση του παραγόμενου διανύσματος στο τυχαίο σημείο.
- Παραγωγή του συνθετικού σημείου δεδομένων.

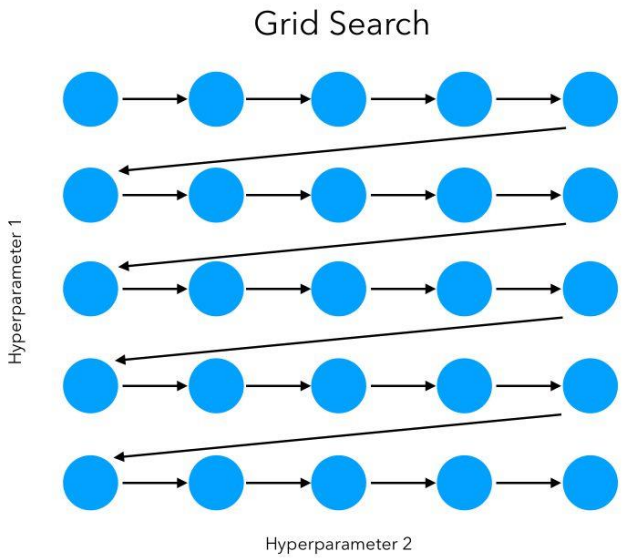


Εικόνα 3.19: *Τεχνική Συνθετικής Μειονοτικής Υπερδειγματοληψίας (SMOTE)*

Πηγή: *Emilia Orellana, SMOTE (2020), [Available: https://emilia-orellana44.medium.com/smote-2acd5dd09948] (Accessed 20-10-2022)*

3.4.4 Αναζήτηση Πλέγματος (GridSearch)

Η Αναζήτηση Πλέγματος πρόκειται για μια διαδικασία βελτιστοποίησης υπερπαραμέτρων (hyperparameters), τεχνική ευρέως διαδομένη στην Μηχανική Μάθηση, με καλύτερα αποτελέσματα από την τυχαία αναζήτηση (Random Search). Οι υπερπαράμετροι μοντέλων αφορούν εξωτερικά χαρακτηριστικά που υπεισέρχονται στους αλγορίθμους παλινδρόμησης και ταξινόμησης κατά την κρίση του ερευνητή, με τις τιμές τους να δίνονται πριν την διαδικασία εκπαίδευσης και είναι ανεξάρτητες των δεδομένων (Ndiaye et al., 2019). Τέτοιες υπερπαράμετροι είναι ο αριθμός συστάδων k για αλγορίθμους K-πλησιέστερων γειτόνων ή K-μέσου, η ρυθμιστική υπερπαράμετρος c και το μοντέλο Πυρήνα (kernel) στις Μηχανές Διανυσμάτων υποστήριξης και ο αριθμός των κρυφών στρώσεων (hidden layers) στα Νευρωνικά Δίκτυα. Η διαδικασία της αναζήτησης πλέγματος (Grid Search) εντοπίζει πλειάδες (tuples) των βέλτιστων υπερπαραμέτρων για καθένα διαφορετικό αλγόριθμο προς ελαχιστοποίηση της προκαθορισμένης συνάρτησης απώλειας (loss function) των δεδομένων δοκιμής.



Εικόνα 3.20: *Αλληλουχία διαδικασίας συντονισμού δύο υπερπαραμέτρων με την αναζήτηση πλέγματος*

Πηγή: *Maël Fabien, A Guide to Hyperparameter Optimization (HPO). (2019), [Available: https://maelfabien.github.io/machinelearning/Explorium4/#] (Accessed 20-10-2022)*

3.5 Μετρικές Αξιολογήσεις

Η αξιολόγηση της απόδοσης των μοντέλων ταξινόμησης και παλινδρόμησης αποτελεί μία από τις κομβικότερες διαδικασίες στην εκπαίδευση των μοντέλων στην Μηχανική Μάθηση. Για αυτό τον λόγο, απαιτείται η ποσοτικοποίηση της ποιότητας απόδοσης κάθε μοντέλου με την είσοδο μετρικών αξιολογήσεων στατιστικής φύσεως (Zhou et al., 2021). Για την εκπόνηση της παρούσας Διπλωματικής Εργασίας εφαρμόστηκαν οι κάτωθι μετρικές.

3.5.1 Συντελεστής προσδιορισμού R^2

Στην στατιστική ο συντελεστής προσδιορισμού R^2 είναι η αναλογία της διακύμανσης των τιμών της εξαρτημένης μεταβλητής που προβλέπονται από τις δεδομένες ανεξάρτητες και υπολογίζεται για να αξιολογηθούν μοντέλα Παλινδρομήσεων και ανάλυσης τάσεων, ως ένα μέτρο καλής προσαρμογής του μοντέλου. Ορίζεται ως ο λόγος του αθροίσματος των τετραγώνων εξαιτίας της παλινδρόμησης προς το συνολικό άθροισμα τετραγώνων. Οι τιμές του συντελεστή προσδιορισμού ανήκουν στο διάστημα (-∞, 1] (Kvålseth, 1985). Συντελεστής προσδιορισμού ίσος με την μονάδα υποδεικνύει ότι οι προβλέψεις ταυτίζονται απόλυτα με τα δεδομένα δοκιμής, ενώ ίσος με το μηδέν υποδεικνύει ότι οι ανεξάρτητες μεταβλητές δεν συμβάλλουν καθόλου στην μεταβλητότητα των εξαρτημένων μεταβλητών, δηλαδή οι προβλέψεις ταυτίζονται με τυχαίους υπολογισμούς, κοντά στον μέσο όρο των δεδομένων δοκιμής. Αρνητικός συντελεστής προσδιορισμού R^2 υποδηλώνει ότι οι προβλέψεις έχουν χειρότερη ερμηνεία και από τυχηματικά γεγονότα.

(3.5)

$$R^2 = 1 - \frac{SS_{\text{res}}}{SS_{\text{tot}}}$$

Όπου

R^2 : ο συντελεστής προσδιορισμού

SS_{res} : το άθροισμα τετραγώνων παλινδρόμησης

SS_{tot} : το όλικό άθροισμα τετραγώνων

3.5.2 Μήτρες Σύγχυσης (Confusion Matrixes)

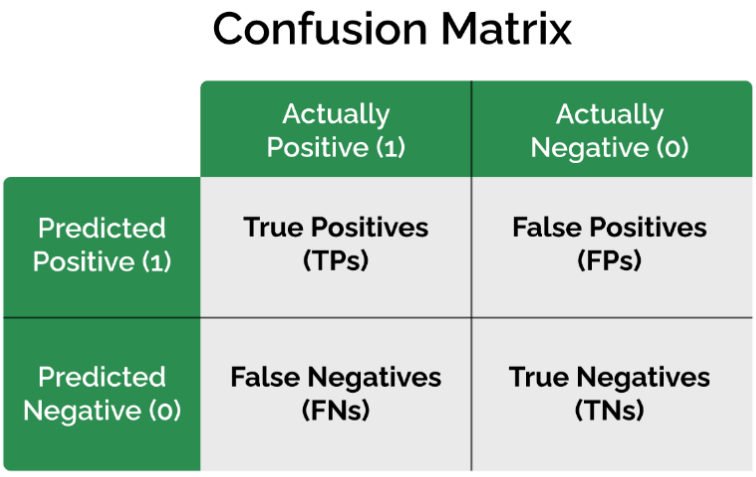
Οι Μήτρες Σύγχυσης αποτελούν μετρικές αξιολογήσεις μέσω της οπτικοποίησης της απόδοσης αλγορίθμων ταξινόμησης σε μορφή πίνακα. Κάθε σειρά του πίνακα αναπαριστά τα διαφορετικά διακριτά ενδεχόμενα των πραγματικών κλάσεων, ενώ κάθε στήλη των προβλεπόμενων κλάσεων, ή το αντίθετο. Η συγκεκριμένη αναπαράσταση συμβάλλει στην αμεσότητα ανάγνωσης της απόδοσης του μοντέλου και εξέταση της πιθανότητας συγχώνευσης των δύο κλάσεων. Τα πιθανά ενδεχόμενα μετά την εκπαίδευση του αλγορίθμου συνοψίζονται ως εξής:

- Ορθώς Θετικό (True Positive – TP): ο αριθμός των προβλέψεων που ο ταξινομητής προέβλεψε την κλάση που ανήκουν σωστά.

- Ορθώς Αρνητικό (True Negative – TN): ο αριθμός των προβλέψεων που δεν ανήκουν στην κλάση και ο ταξινομητής προέβλεψε σωστά.

- Ψευδώς Θετικό (False Positive – FP): ο αριθμός των προβλέψεων που δεν ανήκουν πραγματικά στην κλάση, όμως ταξινομήθηκαν σε αυτή.

- Ψευδώς Αρνητικό (False Negative -FN): ο αριθμός των προβλέψεων που ο ταξινομητής λανθασμένα προέβλεψε ότι ανήκουν σε άλλη κλάση.



Εικόνα 3.21: *Δομή Μήτρας Σύγχυσης*

Πηγή: *Wei Bin Loo, Dominik Czernia, PhD, Jack Bowater, Confusion Matrix Calculator. (2022), [Available: https://www.omnicalculator.com/statistics/confusion-matrix] (Accessed 20-10-2022)*

3.5.3 Ορθότητα

Η Ορθότητα ενός μοντέλου εκπαίδευσης ορίζεται ως η αναλογία των ορθώς ταξινομημένων τιμών στην κλάση που ανήκουν, δηλαδή ποσοτικοποιεί την απόδοση του μοντέλου ταξινόμησης, μέσω του ποσοστού των σωστών του προβλέψεων. Σε δυαδική ταξινόμηση, η Ορθότητα (Accuracy) περιγράφεται από την εξίσωση 3.7:

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

Όπου

TP: Ορθώς Θετικά στοιχεία

TN: Ορθώς Αρνητικά στοιχεία

FP : Ψευδώς Θετικά στοιχεία

FN : Ψευδώς Αρνητικά στοιχεία

Παρόλ'αυτά, διαρκείς έρευνες αποδεικνύουν ότι η ορθότητα ενός μοντέλου δεν επαρκεί για μοντέλα πρόβλεψης σε προβλήματα ταξινόμησης, λόγω του φαινομένου του Παράδοξου της Ορθότητας, σύμφωνα με το οποίο εμφανίζονται περιπτώσεις μεροληψίας λόγω της κυρίαρχης τάξης δεδομένων, επηρεάζοντας την τελική απόδοση του μοντέλου (Uddin, 2019). Για την αντιμετώπιση του Παραδόξου, εισαγάγονται εναλλακτικές αξιολογήσεις.

3.5.4 Ακρίβεια

Η Ακρίβεια περιγράφεται ως την αναλογία της σωστής ταξινόμησης των στοιχείων από το σύνολο των στοιχείων της κλάσης που ταξινομήθηκαν, δηλαδή αποτελεί ένα μέτρο στατιστικής μεταβλητότητας. Ορίζεται σύμφωνα με την εξίσωση 3.8, που ακολουθεί.

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

3.5.5 Ανάκληση

Η Ανάκληση που ονομάζεται και ως ευαισθησία δείγματος, ορίζεται ως το ποσοστό του αριθμού των ορθώς κατανεμημένων στοιχείων σε συγκεκριμένη κλάση προς τα συνολικά στοιχεία που ανήκουν και θα έπρεπε να ανήκουν στην ίδια κλάση, δηλαδή ακόμα και των στοιχείων που δεν ταξινομήθηκαν ορθώς για την συγκεκριμένη κλάση. Σε προβλήματα πρόβλεψης επικινδυνότητας, όπως αυτά που εξετάζει η παρούσα Διπλωματική Εργασία, η ανάκληση υπερισχύει ως αξιολόγηση της ακρίβειας, καθώς περιλαμβάνει τα λανθασμένως κατηγοριοποιημένα στοιχεία (False Positive or False Negative elements), που λόγω του λάθους ταξινόμησης δύναται να αποβούν ιδιαιτέρως κρίσιμα. Συμπερασματικά, σε ανάλυση ταξινόμησης επικινδυνότητας, μοντέλα με χαμηλά μετρικά

ανάκλησης χαρακτηρίζονται ως επισφαλή (Gupta et al., 2021). Η ανάκληση ορίζεται σύμφωνα με την εξίσωση 3.9.

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

3.5.6 F_1-score

Το μετρικό F_1, η αλλιώς συντελεστής Sørensen–Dice, αποτελεί μία συνιστώσα της γενικευμένης του μορφής F_β και ορίζεται ως ο αρμονικός μέσος όρος της ορθότητας και της ακρίβειας. Το συγκεκριμένο μετρικό συνιστά μια σημαντική αξιολόγηση σε περιπτώσεις υπολογισμού ισορροπίας μεταξύ ορθότητας και ακρίβειας σε προβλήματα μη ισορροπημένης μάθησης. Ωστόσο, και εν αντιθέσει με την γενικευμένη του μορφή, σύμφωνα με ερευνητές στερείται διαισθητικής ερμηνείας σαν αξιολόγηση, καθώς δεν λογίζει στάθμιση (weighting) μεταξύ δύο διακριτών εννοιών, όπως η ορθότητα και η ακρίβεια (Hand & Peter, 2018). Οι τιμές του F_1 κυμαίνονται από 0.0 έως 1.0, με την τιμή 1.0 να υποδηλώνει τέλεια ορθότητα και ακρίβεια, ενώ η τιμή 0.0 μηδενικές αντίστοιχα τιμές. Η εξίσωση που περιγράφει το F_1 είναι η ακόλουθη 3.10.

(3.10) $$F_1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall}$$

3.5.7 Εσφαλμένο θετικό ποσοστό (False Positive Ratio -FPR)

Το εσφαλμένο θετικό ποσοστό, η ποσοστό λανθασμένου συναγερμού (False Alarm Ratio – FAR) ορίζεται ως η πιθανότητα λανθασμένης ταξινόμησης σε κλάση μιας αρχικής υπόθεσης και συνιστά την αναλογία μεταξύ των αρνητικών στοιχείων που λανθασμένα ταξινομήθηκαν ως θετικά προς το σύνολο των πραγματικά αρνητικών στοιχείων.

(3.11) $$FPR = \frac{FP}{FP + TN}$$

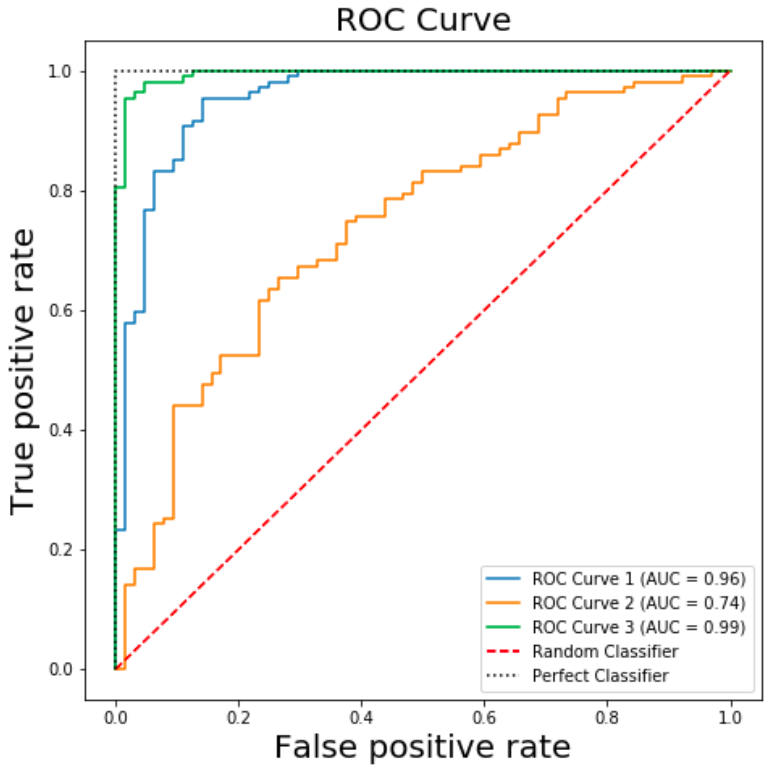
3.5.8 Εσφαλμένο αρνητικό ποσοστό (False Negative Ratio -FNR)

Το εσφαλμένο αρνητικό ποσοστό ορίζεται ως η πιθανότητα λανθασμένης ταξινόμησης στην διαφορετική κλάση από αυτήν που υπολογίζει το εσφαλμένο θετικό ποσοστό. Εν προκειμένω, σε αναλύσεις πρόβλεψης επικινδυνότητας, υψηλό ποσοστό FNR υποδηλώνει πληθώρα λανθασμένα ταξινομημένων πραγματικά επικίνδυνων προβλέψεων.

(3.12) $$FNR = \frac{FN}{FN + TP}$$

3.5.9 Περιοχή κάτω από την Καμπύλη (AUC score)

Πρόκειται για την προτιμότερη μετρική αξιολόγηση για την προβλεπτική ικανότητα ενός μοντέλου ταξινόμησης (Huang & Ling, 2005) και ορίζεται ως το ποσοστό από το συνολικό γράφημα, της περιοχής κάτω από την καμπύλη ROC, δηλαδή της γραφικής δισδιάστατης απεικόνισης του ορθώς θετικού ποσοστού (True Positive Rate) και του εσφαλμένου θετικού ποσοστού (False Positive Rate). Η καμπύλη ROC μπορεί να συμβάλλει στον καθορισμό ενός ορίου απόφασης (optimal threshold) που ελαχιστοποιεί το ποσοστό σφάλματος ή το κόστος εσφαλμένης ταξινόμησης σε δεδομένες κλάσεις, στον προσδιορισμό του καλύτερου ταξινομητή σε δεδομένο εύρος τιμών και περιοχής στην καμπύλη, καθώς και στην απόκτηση βαθμονομημένων εκτιμήσεων, κατά την Μπεϋζιανή στατιστική (Flach, 2016).



Εικόνα 3.22: *Καμπύλες ROC και τιμές AUC σε ταξινόμηση πολλαπλών κλάσεων*

Πηγή: *Ajitesh Kumar, Data Analytic., (2020), [Available: https://vitalflux.com/roc-curve-auc-python-false-positive-true-positive-rate/] (Accessed 21-10-2022)*

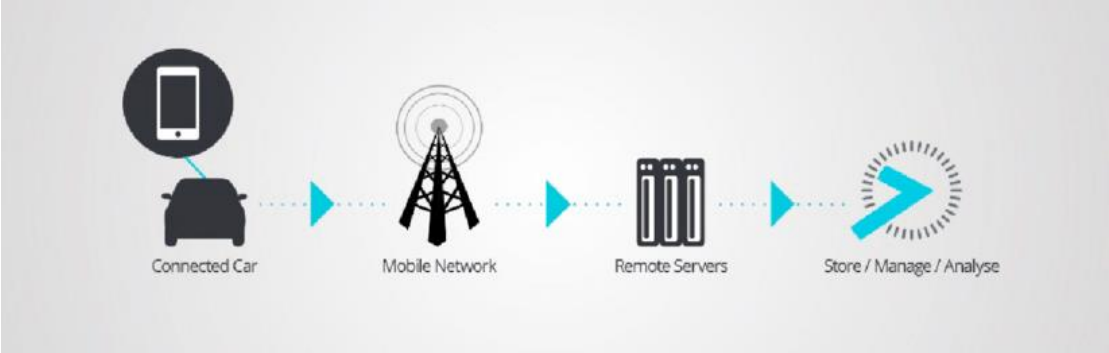
4. Συλλογή και Επεξεργασία των στοιχείων

Αυτή η ενότητα παρέχει τεχνικές πληροφορίες για την διαδικασία συλλογής των δεδομένων, περιγραφικά στατιστικά στοιχεία, διερευνητικές παραμέτρους και διάφορες επιπρόσθετες πληροφορίες, συναφείς με τα δεδομένα που αξιοποιήθηκαν στην παρούσα Διπλωματική Εργασία. Στόχος είναι τα παρεχόμενα δεδομένα να αποσαφηνιστούν και να αποτυπωθούν όσο το δυνατόν πληρέστερα, προκειμένου να προκύψουν χρήσιμα συμπεράσματα και να αναβαθμιστεί η διαδικασία τις κύριας μετέπειτα επεξεργασίας τους.

4.1 Συλλογή των στοιχείων

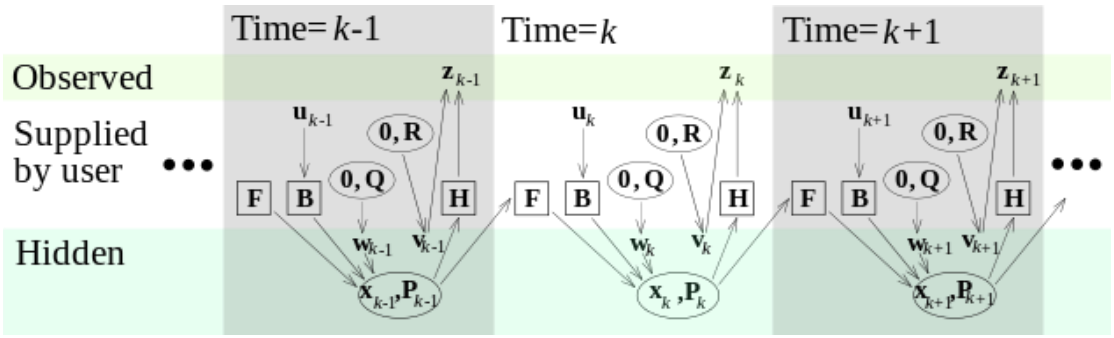
Τα δεδομένα φυσικής οδήγησης υπό πραγματικές οδικές συνθήκες που αξιοποιήθηκαν για την εκπόνηση της εργασίας, συλλέχθηκαν και παρασχέθηκαν από την εταιρία τηλεματικής OSeven Telematics© (https://oseven.io/), μέσω ειδικά κατασκευασμένης και απόλυτα ενσωματωμένης στο ταξίδι, εφαρμογής κινητού τηλεφώνου, η οποία καταγράφει και συλλέγει δεδομένα συνεχόμενα, δίχως παρεμβολές στην οδήγηση. Η ανάπτυξη της εφαρμογής, η οποία παρουσιάστηκε το 2014, ξεκίνησε από την ανάγκη για την συλλογή και ανάλυση δεδομένων οδικής συμπεριφοράς στοιχείων αληθινών οδικών συνθηκών, μεγάλης κλίμακας υπό πραγματικό χρόνο και με άμεση καταγραφή και αποθήκευσή τους. Απώτερος στόχος της ανάπτυξης αυτής της καινοτόμου εφαρμογής είναι η αξιολόγηση και βελτίωση της οδικής συμπεριφοράς και οδικής ασφάλειας.

Η λειτουργία της εφαρμογής βασίζεται στους αισθητήρες hardware της συσκευής smartphone, δίχως την χρήση άλλου εξοπλισμού. Επιπρόσθετα, για την ανάγνωση των δεδομένων των αισθητήρων και την προσωρινή τους αποθήκευση στην βάση δεδομένων του κινητού, προτού μεταδοθούν στην κεντρική βάση δεδομένων (back-end database) της εταιρίας, αξιοποιείται πληθώρα APIs (Application Programming Interface). Η τυπική διαδικασία που ακολουθείται σε κάθε καταγραφή νέου ταξιδιού παρουσιάζεται χαρακτηριστικά στην Εικόνα 4.. Τα συλλεχθέντα δεδομένα είναι ιδιαιτέρως διακεκριμένα χωροχρονικά και μόλις αποθηκευτούν στην τελική βάση δεδομένων μετατρέπονται σε δείκτες οδικής συμπεριφοράς και ασφάλειας δια μέσω επεξεργασίας σημάτων, αλγορίθμους Μηχανικής Μάθησης, συγχώνευση δεδομένων (data fusion) και αλγορίθμους Μαζικών Δεδομένων (Big Data algorithms) (Kontaxi et al., 2022). Πιο συγκεκριμένα για την καταγραφή και παραγωγή των δεικτών, οι αισθητήρες του smartphone που αξιοποιούνται περιλαμβάνουν επιταχυνσιόμετρο, γυροσκόπιο, μαγνητόμετρο και GPS, ενώ οι τεχνικές συγχώνευσης δεδομένων παρέχονται από την iOS© και την Android© με μοντέλα 9 βαθμών ελευθερίας (Yaw, Pitch, Roll) (Tran, 2017), Γραμμικής επιτάχυνσης και βαρύτητας, με τις καταγραφές των δεδομένων να πραγματοποιούνται στην μέγιστη ισχύ του 1Hz.



Εικόνα 4.1: *Ροή δεδομένων συστήματος OSeven*

Πηγή: *Tselentis, Dimitrios & Vlahogianni, Eleni & Yannis, George & Kavouras, Loukas. (2020). Hybrid Data Envelopment Analysis for Large-Scale Smartphone Data Modeling. Transportation Research Procedia. 48. 975-986. [Available at doi: 10.1016/j.trpro.2020.08.126] (Accessed 21-10-2022)*



Εικόνα 4.2: *Συγχώνευση δεδομένων με φίλτρο Kalman πολλαπλών διαστάσεων*

Πηγή: *Sharath Srinivasan, The Kalman Filter: An algorithm for making sense of fused sensor insight. (2018), [Available: https://towardsdatascience.com/kalman-filter-an-algorithm-for-making-sense-from-the-insights-of-various-sensors-fused-together-ddf67597f35e] (Accessed 21-10-2022)*

Στο παρόν ερευνητικό έργο, τα δεδομένα που συλλέχθηκαν από την OSeven© και επεξεργάστηκαν αποτελούνται από καταγραφές 356.162 διαφορετικών οδικών διαδρομών, με τους δείκτες που παράχθηκαν για την κάθε μία διαδρομή να ανέρχονται στους 75. Οι δείκτες και τα δεδομένα αποτυπώθηκαν σε αρχείο τιμών διαχωρισμένων με κόμμα του Microsoft Excel (.csv). Οι καταγραφές των διαδρομών πραγματοποιήθηκαν, ως επί το πλείστον, κατά την διάρκεια της έξαρσης της πανδημίας SARS-CoV-2, με τους δείκτες που αναφέρονται σε αυτήν, όπως ο δείκτης Αυστηρότητας (Stringency index) ή ο δείκτης Περιορισμών (Restrictions index) να μην συμπεριλαμβάνονται στην ανάλυση των δεδομένων, λόγω ανεπίκαιρής τους ιδιότητας. Ως ακολούθως, οι δείκτες με τους οποίους πραγματοποιήθηκε εν τέλει η ανάλυση, όντας οι ανεξάρτητες και εξαρτημένες μεταβλητές, ανέρχονται σε 23 και περιγράφονται στον Πίνακα 4.1 . Τα συλλεχθέντα οδικά δεδομένα είναι απολύτως ανώνυμα, συνεπώς άλλα οδηγοκεντρικά στοιχεία, όπως η ηλικία, το φύλο και το ιστορικό ατυχημάτων είναι παντελώς άγνωστα.

4.2 Περιγραφή των δεδομένων

Σε αυτή την υποενότητα παρατίθεται η περιγραφή των δεδομένων που συλλέχθηκαν σε μορφή Πίνακα (Πίνακας 4.1).

Πίνακας 4.1: *Περιγραφή των δεδομένων που συλλέχθηκαν από την βάση δεδομένων της OSeven*

Μεταβλητή	**Μονάδα μέτρησης**	**Περιγραφή**
duration	sec	Συνολική διάρκεια της διαδρομής
total_distance	km	Συνολική διανυθείσα απόσταση
risky_hours	km	Διανυθείσα απόσταση κατά την διάρκεια επικίνδυνης ζώνης ώρας (00:00 - 05:00)
driving_duration	sec	Συνολική διάρκεια οδήγησης (δεν περιλαμβάνονται καταστάσεις ακινησίας του οχήματος/ στάση, στάθμευση)
ha	-	Απότομων επιταχύνσεις σε μια διαδρομή
hb	-	Απότομες επιβραδύνσεις σε μια διαδρομή
ha/100 km	-	Απότομες επιταχύνσεις στα 100 χιλιόμετρα
hb/100km	-	Απότομες επιβραδύνσεις στα 100 χιλιόμετρα
avg speed	km/h	Μέση ταχύτητα διαδρομής
av_speeding_kmh_no_changer	km/h	Μέση διαφορά μεταξύ της ταχύτητας οδήγησης και του ορίου ταχύτητας
avg driving speed	km/h	Μέση ταχύτητα οδήγησης
av_speeding_kmh	km/h	Μέση ταχύτητα οδήγησης κατά την διάρκεια υπέρβασης ορίου ταχύτητας
sum_speeding	sec	Συνολική διάρκεια οδήγησης με υπερβολική ταχύτητα σε μια διαδρομή (Όριο ταχύτητας + Όρια ανέχειας)
time_mobile_usage	sec	Συνολική διάρκεια χρήσης κινητού σε μια διαδρομή
time_mobile_usage/driving duration	sec/sec	Διάρκεια χρήσης κινητού ανά μονάδα συνολικής διάρκειας οδήγησης

sum_speeding/driving duration	sec/sec	Διάρκεια οδήγησης με υπερβολική ταχύτητα ανά μονάδα συνολικής διάρκειας οδήγησης
speeding_score	%	Σκορ υπερβολικής ταχύτητας
mu_score	%	Σκορ χρήσης κινητού κατά την διάρκεια της διαδρομής
hb_score	%	Σκορ απότομων επιβραδύνσεων
ha_score	%	Σκορ απότομων επιταχύνσεων
total_score	%	Συνολικό σκορ
GRdriving	%	Ποσοστιαία ημερήσια διαφορά φόρτου οδηγών Ι.Χ. συγκριτικά με συνθήκες προ πανδημίας (BL:13 Ιαν 2020)
GRwalking	%	Ποσοστιαία ημερήσια διαφορά φόρτου οδηγών Ι.Χ. συγκριτικά με συνθήκες προ πανδημίας (BL.: 13 Ιαν 2020)

Προκειμένου να αξιολογηθεί η ποιότητα κάθε διαδρομής, το μοντέλο οδικής συμπεριφοράς της OSeven© κατασκευάζει έναν δείκτη σκορ συγκεκριμένων παραμέτρων, για κάθε διαδρομή. Τα επιμέρους σκορ διακρίνονται σε σκορ υπερβολικής ταχύτητας, σκορ απότομων επιταχύνσεων και επιβραδύνσεων (αριθμός και ένταση), σκορ χρήσης κινητού τηλεφώνου και συνολικό σκορ, με τις τιμές τους να κυμαίνονται από 0 (χειρότερο) έως 100 (καλύτερο). Το συνολικό σκορ κάθε χρήστη υπολογίζεται ως ο σταθμισμένος μέσος των καταγραμμένων διαδρομών κάθε χρήστη σε διάστημα 12 μηνών, με την απόσταση να αποτελεί τον συντελεστή στάθμισης.

4.3 Επεξεργασία των δεδομένων

Η επεξεργασία των δεδομένων πραγματοποιήθηκε με την γλώσσα προγραμματισμού Python και την βοήθεια των βιβλιοθηκών ανάλυσης δεδομένων pandas, numpy και seaborn για την απεικόνισή τους.

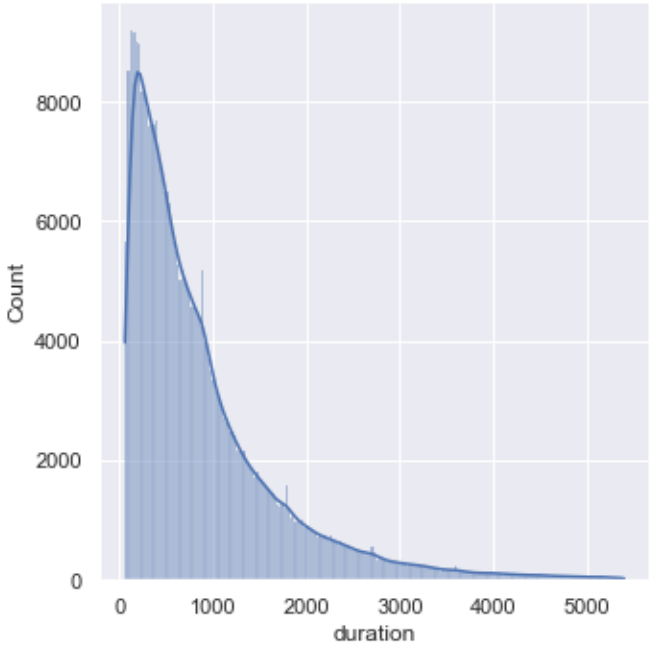
4.3.1 Περιγραφική Στατιστική των δεδομένων

Στην ενότητα που ακολουθεί, παρουσιάζονται τα στοιχεία περιγραφικής στατιστικής που συνοδεύουν τα δεδομένα, όπως η μέση τιμή, η τυπική απόκλιση, η διακύμανση, οι συντελεστές λοξότητας και κύρτωσης και οι ελάχιστες και μέγιστές τους τιμές, καθώς και η απεικόνιση ορισμένων βασικών τους χαρακτηριστικών σε μορφή γραφημάτων.

Πίνακας 4.2: *Περιγραφική στατιστική των μεταβλητών*

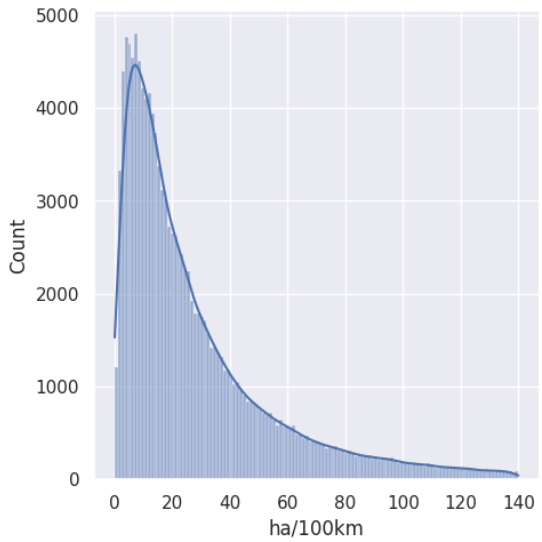
Μεταβλητή	**Μέση τιμή**	**Τυπική απόκλιση**	**Διακύμανση**	**Μέτρο Λοξότητας**	**Μέτρο Κύρτωσης**	**Ελάχιστη τιμή**	**Μέγιστη τιμή**
duration (sec)	962.04	1093.98	1.197E+06	4.044	29.179	61.00	25549.00
total_distance (km)	11.60	22.31	4.979E+02	6.751	70.106	0.50	648.68
total_score (%)	75.66%	23.47	5.510E+02	-0.795	-0.491	10.00%	100.00%
speeding_score (%)	76.52%	32.92	1.084E+03	-1.026	-0.581	10.00%	100.00%
mu_score (%)	80.53%	34.62	1.198E+03	-1.333	-0.073	10.00%	100.00%
hb_score (%)	79.20%	21.35	4.558E+02	-1.005	0.083	10.00%	100.00%
ha_score (%)	81.59%	19.74	3.897E+02	-1.347	1.045	10.00%	100.00%
risky_hours (km)	0.37	4.01	1.606E+01	30.709	1759.217	0.00	427.7
ha (-)	0.89	1.97	3.882E+00	6.888	137.036	0.00	121.00
hb (-)	1.26	2.20	4.856E+00	4.355	45.307	0.00	87.00
sum_speeding (sec)	65.63	194.31	3.776E+04	8.744	139.920	0.00	7697.00
av_speeding_kmh (km/h)	4.00	6.03	3.633E+01	2.746	32.347	0.00	314.16
time_mobile_usage (sec)	39.11	159.27	2.537E+04	11.273	240.711	0.00	9901.00
driving_duration (sec)	769.97	967.15	9.354E+05	4.576	35.074	61.00	23900.00
ha/100 km (-)	11.95	27.86	7.763E+02	4.500	31.378	0.00	597.01
hb/100km (-)	16.39	29.76	8.858E+02	3.560	21.975	0.00	819.67
avg speed (km/h)	35.13	18.89	3.567E+02	1.309	2.037	1.96	262.52
time_mobile_usage/driving duration (sec/sec)	0.05	0.14	2.079E-02	4.090	18.010	0.00	0.99
sum_speeding/driving duration (sec/sec)	0.06	0.11	1.133E-02	2.753	9.078	0.00	1.00
av_speeding_kmh_no_changer (km/h)	9.51	11.11	1.234E+02	0.864	2.258	0.00	329.16
avg driving speed (km/h)	42.57	17.58	3.091E+02	1.436	2.715	5.57	323.91
GRdriving (%)	109.64	55.88	3.123E+03	0.250	-0.431	0.00	241.14
GRwalking (%)	114.35	61.94	3.837E+03	0.297	-0.809	0.00	254.21

Τα περιγραφικά στατιστικά στοιχεία που προέκυψαν από την αρχική ανάλυση των μεταβλητών, προσφέρουν επιπρόσθετη πληροφορία, σημαντική για την κατανόηση των δεδομένων. Μετά την αρχική επεξεργασία των μεταβλητών, οι διατιθέμενες διαδρομές υπολογίστηκαν να έχουν μέση διάρκεια 962.04 δευτερολέπτων, ήτοι 16.03 λεπτά. Ενώ παρατηρήθηκαν διαδρομές με αυξημένη διάρκεια, αυτές ήταν περιορισμένες και δυνητικά οφείλονται στον κυκλοφοριακό φόρτο της ώρας κίνησης, συνδυαζόμενες με μια κυκλική διαδρομή αποφυγής του. Αυτή η διακύμανση παρατηρείται στο Γράφημα 4.1.

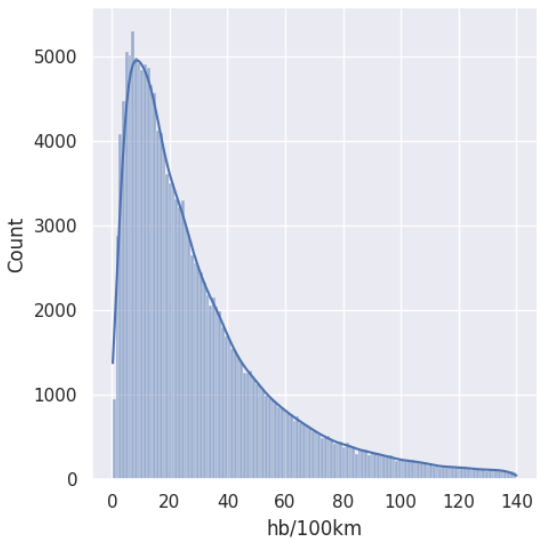


Γράφημα 4.1: *Διακύμανση της κατανομής της διάρκειας διαδρομής*

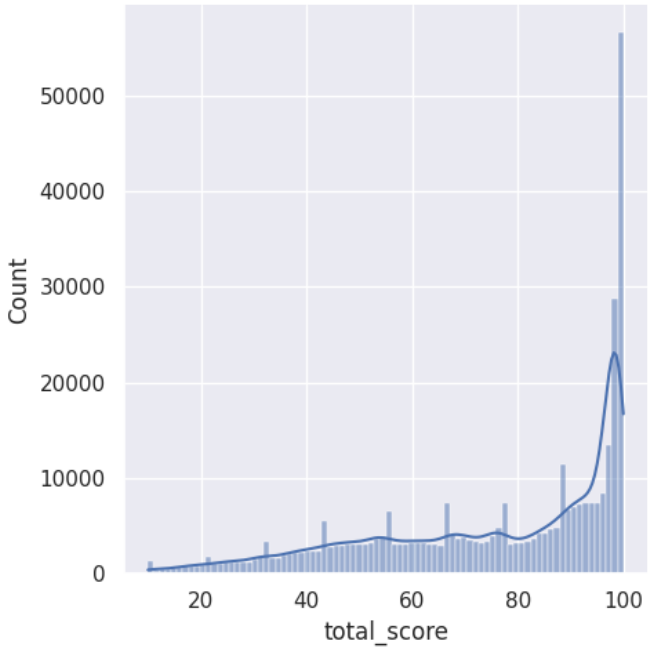
Σε αυτές τις διαδρομές, σημειώθηκαν ορισμένα απότομα συμβάντα (harsh events), με τα περιγραφικά στοιχεία να αποδεικνύουν ότι οι οδικές συμπεριφορές δεν πρόκειται για υπερβολικά επιθετικές, καθώς η συχνότητα εμφάνισής τους είναι σχετικά χαμηλή και οι σχετικοί διάμεσοί τους είναι μηδενικοί (Γράφημα 4.2, 4.3). Το μέτρο κύρτωσης των απότομων συμβάντων είναι αξιόλογα υψηλός, γεγονός που καταδεικνύει ότι η συχνότητα των απότομων συμβάντων φαίνεται να ακολουθεί αρκετά ασύμμετρη κατανομή. Το δείγμα των οδηγών από τους οποίους συλλέχθηκαν τα δεδομένα, φαίνεται να ακολουθούν κατά βάση τους οδικούς κώδικες και κανόνες, καθώς η μεταβλητή total_score έχει διάμεσο 84.00/100 και η διακύμανσή της περιγράφεται στο Γράφημα 4.4. Οι τιμές του μέτρου κύρτωσης παραμένουν μεγάλες συνολικά, πέραν των μεταβλητών score και των φόρτων, ενώ ακόμα χαμηλή είναι και για την μέση ταχύτητα οδήγησης και διαδρομής, κάτι που τις χαρακτηρίζει ως αποκλίνουσες μεταβλητές με αρκετές χαρακτηριστικά έκτοπες τιμές, όπως οι μέγιστές τους.



Γράφημα 4.2: *Διακύμανση απότομων επιταχύνσεων ανά 100χλμ.*



Γράφημα 4.3: *Διακύμανση απότομων επιβραδύνσεων ανά 100χλμ.*

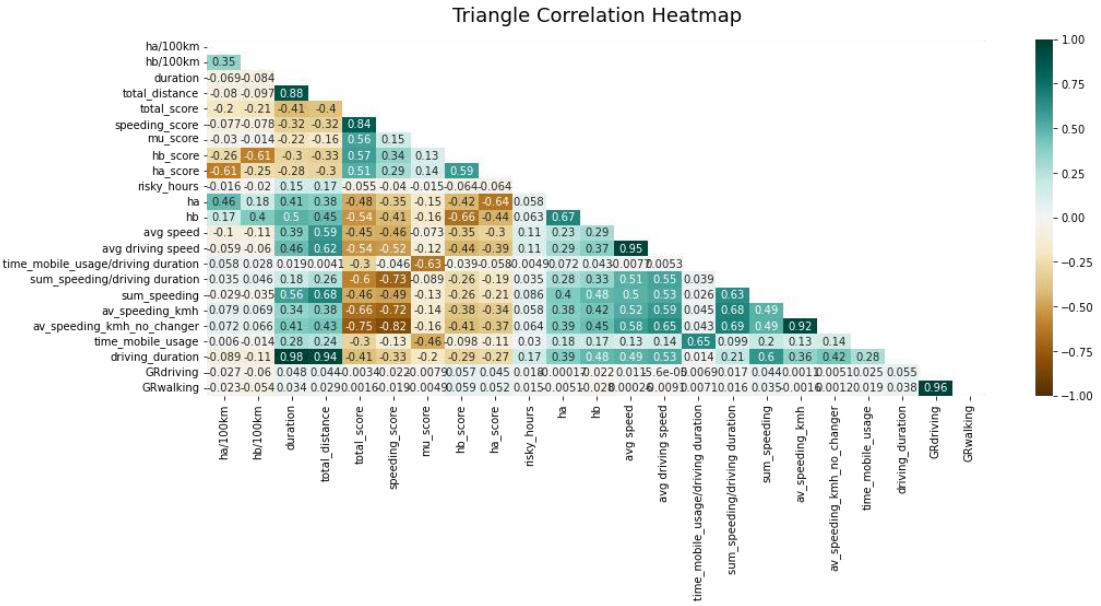


Γράφημα 4.4: *Διακύμανση του συνολικού σκορ διαδρομής*

4.3.2 Συσχέτιση Pearson

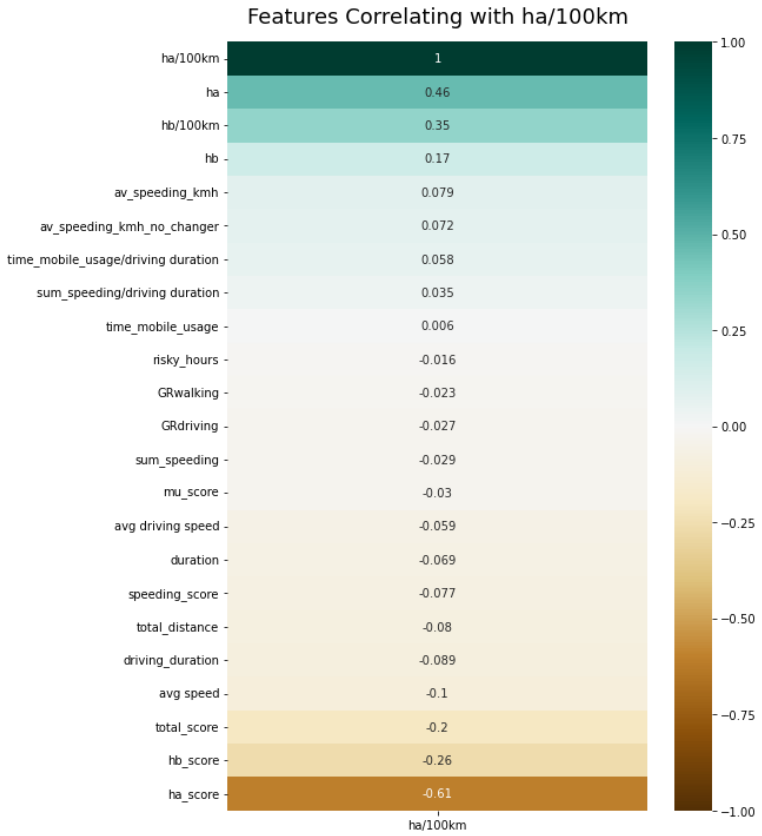
Προκειμένου να διερευνηθούν οι σχέσεις μεταξύ των μεταβλητών, υπολογίστηκε ο συντελεστής συσχέτισης Pearson, ο οποίος απεικονίζεται στο Γράφημα 4.5. Για την καλύτερη κατανόηση του γραφήματος, γίνεται ποιοτικός χαρακτηρισμός του εύρους τιμών του συντελεστή Pearson r:

- | r | = 0, καμία συσχέτιση μεταξύ των μεταβλητών
- 0 < | r | < 0.25, κακή συσχέτιση μεταξύ των μεταβλητών
- 0.26 < | r | < 0.50, ανίσχυρη συσχέτιση μεταξύ των μεταβλητών
- 0.51 < | r | < 0.75, μέτρια συσχέτιση μεταξύ των μεταβλητών
- 0.76 < | r | < 0.99, ισχυρή συσχέτιση μεταξύ των μεταβλητών
- | r | = 1.00, τέλεια συσχέτιση μεταξύ των μεταβλητών

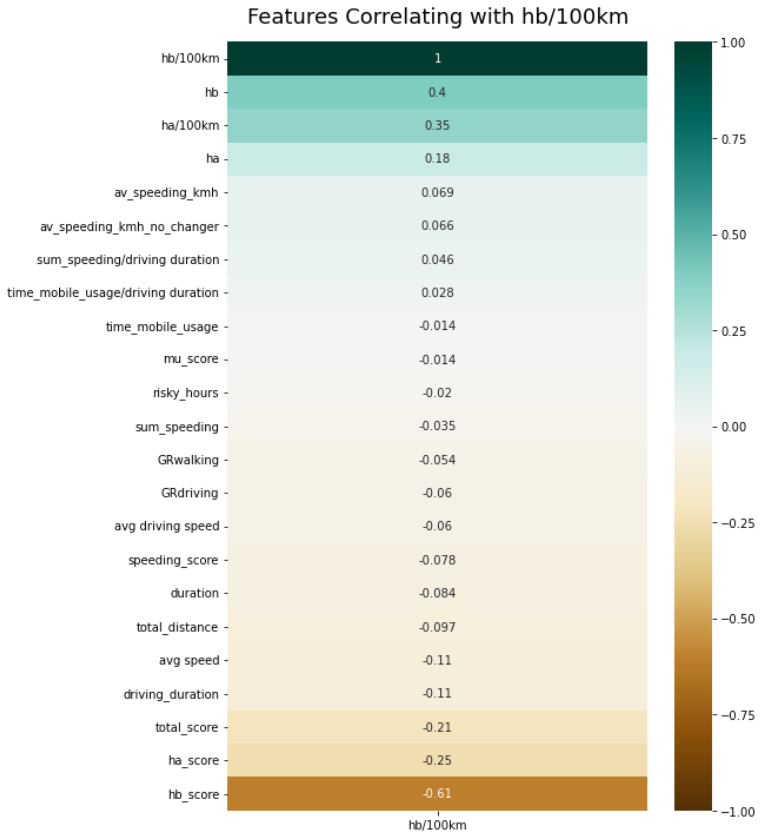


Γράφημα 4.5: *Τριγωνικός πίνακας θερμότητας συσχέτισης μεταβλητών κατά Pearson*

Στα γραφήματα που ακολουθούν παρουσιάζονται απομονωμένοι οι δείκτες συσχέτισης Pearson για τις επιλεγμένες εξαρτημένες μεταβλητές της έρευνας, με τις αντίστοιχες ανεξάρτητες. Υπενθυμίζεται ότι οι εξαρτημένες μεταβλητές είναι οι απότομες επιταχύνσεις και απότομες επιταχύνσεις ανά 100 χιλιόμετρα (ha/100km, hb/100km). Οι συγκεκριμένες μεταβλητές επιλέχθηκαν για την έρευνα, καθώς συνιστούν δείκτες οδικής επικινδυνότητας, αφού είναι αναγωγική τιμή ως προς την διανυθείσα απόσταση.



Γράφημα 4.6: *Συσχέτιση Pearson ανεξάρτητων μεταβλητών με απότομες επιταχύνσεις ανά 100 χλμ.*



Γράφημα 4.7: *Συσχέτιση Pearson ανεξάρτητων μεταβλητών με απότομες επιβραδύνσεις ανά 100 χλμ.*

Όπως παρατηρείται από τα Γραφήματα 4.5, 4.6, 4.7, τα απότομα γεγονότα συμμεταβάλλονται σχετικά ασήμαντα και αρνητικά με τους φόρτους κυκλοφορίας και πεζών, ενώ φαίνεται η σημαντική επιρροή της χρήσης του κινητού τηλεφώνου κατά την διαδρομή. Η οδήγηση κατά την διάρκεια της επικίνδυνης ζώνης δεν φαίνεται να παρουσιάζει υψηλή συσχέτιση με τα απότομα γεγονότα, κάτι που την καθιστά οριακά ασυσχέτιστη με αυτά. Επιπλέον, τα απότομα περιστατικά ανά 100χλμ. είναι λογικά θετικά συσχετισμένα με τις ονομαστικές τιμές των απότομων περιστατικών και αρνητικά με τα επιμέρους σκορ τους και τον δείκτη συνολικού σκορ διαδρομής.

5. Επεξεργασία - Αναλύσεις

Το συγκεκριμένο Κεφάλαιο περιλαμβάνει την εφαρμογή της μεθοδολογίας που περιεγράφηκε αναλυτικά στο 3ο Κεφάλαιο, με την ανάπτυξη των μοντέλων και τεχνικών Ισορροπημένης και Μη Μάθησης. Το συγκεκριμένο κεφάλαιο διαρθρώνεται σε δύο κυρίως τμήματα, αυτό του ελέγχου Σημαντικότητας της επιρροής των εξαρτημένων μεταβλητών στις ανεξάρτητες (Feature Importance) και αυτό της διαδικασίας Ταξινόμησης (Classification). Πιο συγκεκριμένα, θα αξιολογηθεί η σημαντικότητα καθεμιάς από την ανεξάρτητες μεταβλητές και θα επιλεγούν οι σημαντικότερες εξ αυτών οι οποίες στη συνέχεια θα αξιοποιηθούν ως δεδομένα για την ταξινόμηση των απότομων περιστατικών. Προκειμένου να αξιολογηθούν τα εξεταζόμενα μοντέλα παλινδρόμησης και ταξινόμησης θα πραγματοποιηθεί συγκριτική ανάλυση με κριτήριο της μετρικές τους αξιολογήσεις.

Για την ολοκλήρωση της εν λόγω ενότητας, αξιοποιήθηκε η γλώσσα προγραμματισμού Python, με την βοήθεια των βιβλιοθηκών ανάλυσης και επεξεργασίας pandas και numpy, την βιβλιοθήκη μηχανικής εκμάθησης λογισμικού scikit-learn και την βιβλιοθήκη γραφικής απεικόνισης matplotlib.

5.1 Σημαντικότητα Χαρακτηριστικών

Η συγκεκριμένη διαδικασία αποτελεί κομβική παράμετρο στο σύνολο της συγκεκριμένης Διπλωματικής Εργασίας. Η σημαντικότητα χαρακτηριστικών αποσκοπεί στην ποσοτικοποίηση της επιρροής των οδικών δεδομένων που έχουν επιλεγεί ως ανεξάρτητες μεταβλητές, προκειμένου να θεσπιστούν οι βάσεις για βέλτιστη διαδικασία ταξινόμησης. Η διαδικασία περιλαμβάνει την ανάπτυξη μοντέλων Παλινδρόμησης, την αξιολόγηση των Παλινδρομήσεων και, εν τέλει, τον καθορισμό των ανεξάρτητων μεταβλητών με τις οποίες θα πραγματοποιηθεί η ταξινόμηση. Από το σύνολο των ανεξάρτητων μεταβλητών, θα προκριθούν οι 5 σημαντικότερες εξ αυτών, γεγονός που αποτελεί διεργασία βελτιστοποίησης της απόδοσης των προγνωστικών μοντέλων. Η τελική Επιλογή Χαρακτηριστικών θα πραγματοποιηθεί βάσει της διαδικασίας Σημαντικότητας Χαρακτηριστικών και του συντελεστή συσχέτισης Pearson.

Για τον εντοπισμό της σημαντικότητας χαρακτηριστικών, αξιοποιήθηκε η τεχνική Σημαντικότητας (Μετάθεσης) Χαρακτηριστικών [Feature (Permutation) Importance], με την οποία ποσοτικοποιείται ο βαθμός επιρροής των ανεξάρτητων μεταβλητών στις εξαρτημένες.

Η παρούσα Διπλωματική Εργασία περιλαμβάνει την ανάλυση και ταξινόμηση δύο διακριτών εξαρτημένων μεταβλητών, των απότομων επιταχύνσεων ανά 100 χλμ. και των απότομων επιβραδύνσεων ανά 100 χλμ. Συνεπώς, η διαδικασία σημαντικότητας χαρακτηριστικών θα εφαρμοστεί για κάθε μία μεταβλητή ξεχωριστά, με τα αποτελέσματα να ομαδοποιούνται και να προκύπτουν οι 5 σημαντικότερες μεταβλητές και για τις δύο εξαρτημένες.

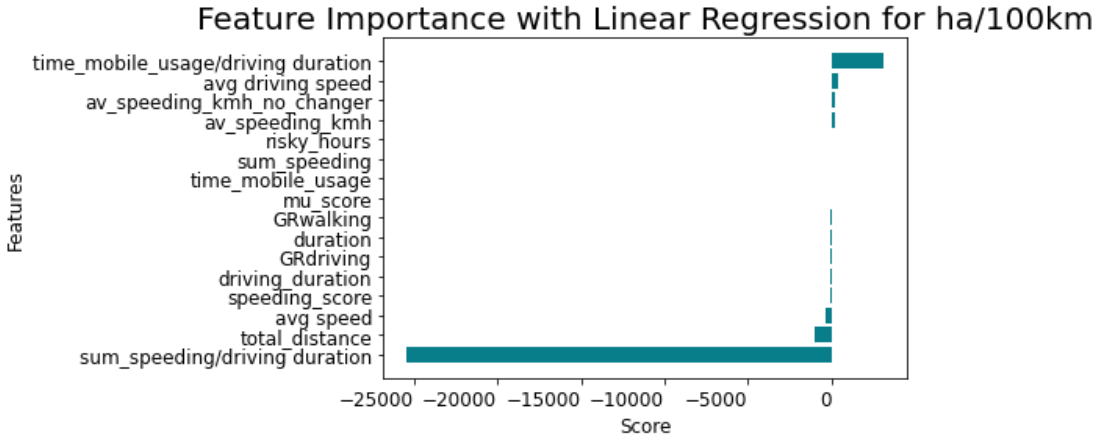
Σε πρώτη φάση, οι μεταβλητές διακρίθηκαν σε 3 υποσύνολα. Το πρώτο περιλαμβάνει τις ανεξάρτητες μεταβλητές, το δεύτερο τις απότομες επιταχύνσεις ανά 100 χλμ. (ha/100km) και το τρίτο τις απότομες επιβραδύνσεις ανά 100 χλμ. (hb/100km). Το σύνολο των ανεξάρτητων μεταβλητών επεξεργάστηκε με την διαδικασία της κανονικοποίησης. Στην συνέχεια,

αναπτύχθηκαν οι αλγόριθμοι που αποτελούνται από τις ακόλουθες Παλινδρομήσεις: Γραμμική Παλινδρόμηση, Decision Trees, XGBoost, Random Forests, Linear SVR.

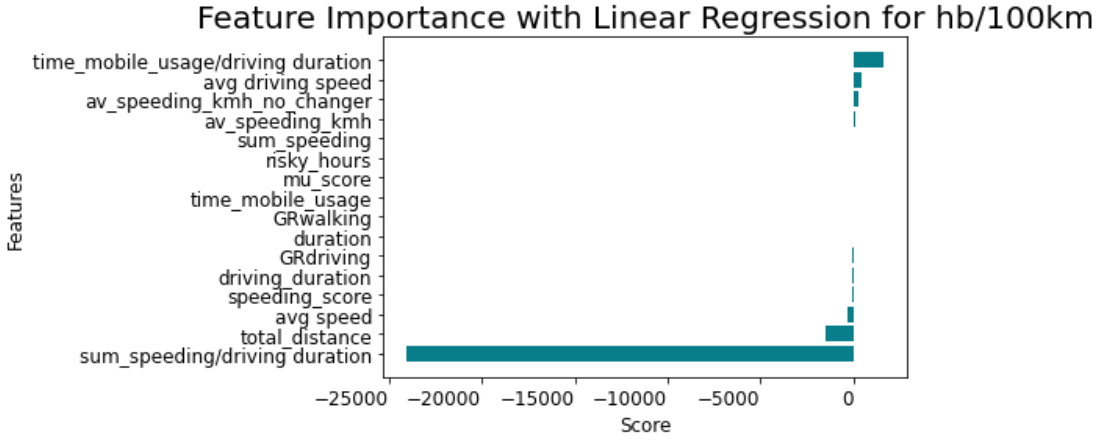
5.1.1 Γραμμική Παλινδρόμηση

Για την διαδικασία της Γραμμικής Παλινδρόμησης, η σημαντικότητα υπολογίστηκε βάσει του συντελεστή σημαντικότητας (.coef_), με τα αποτελέσματα να δύνανται να παίρνουν και αρνητικές τιμές. Οι αρνητικές τιμές σημαντικότητας καταδεικνύουν την ανισορροπία των δεδομένων μειονοτικής τάξης ή και την εμφάνιση φαινομένων θορύβου.

Στα Γραφήματα 5.1 και 5.2, απεικονίζεται η επιρροή των ανεξάρτητων μεταβλητών στις εξαρτημένη ha/100km, hb/100km, σε φθίνουσα κατάταξη. Επιπλέον, σημειώνονται οι υπολογισμένες αυτές σημαντικότητες σε μορφή Πίνακα.



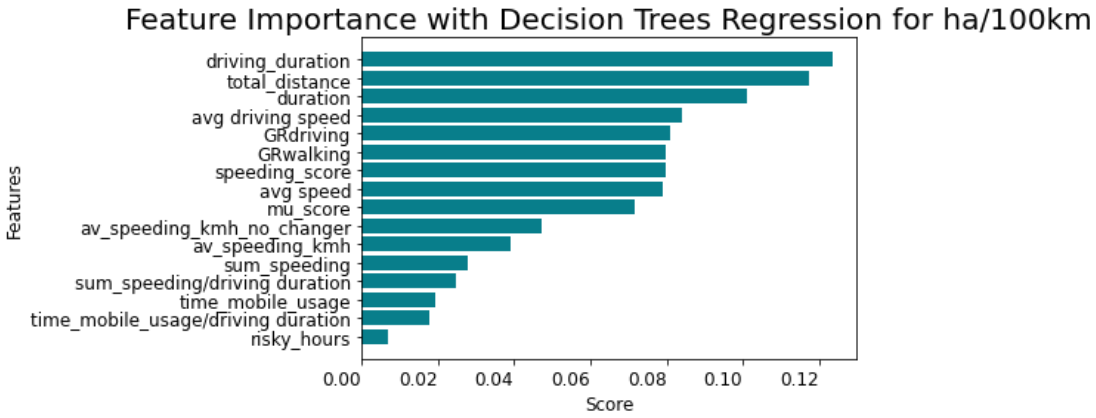
Γράφημα 5.1: *Σημαντικότητα Χαρακτηριστικών για τις απότομές επιταχύνσεις ανά 100χλμ. με Γραμμική Παλινδρόμηση*



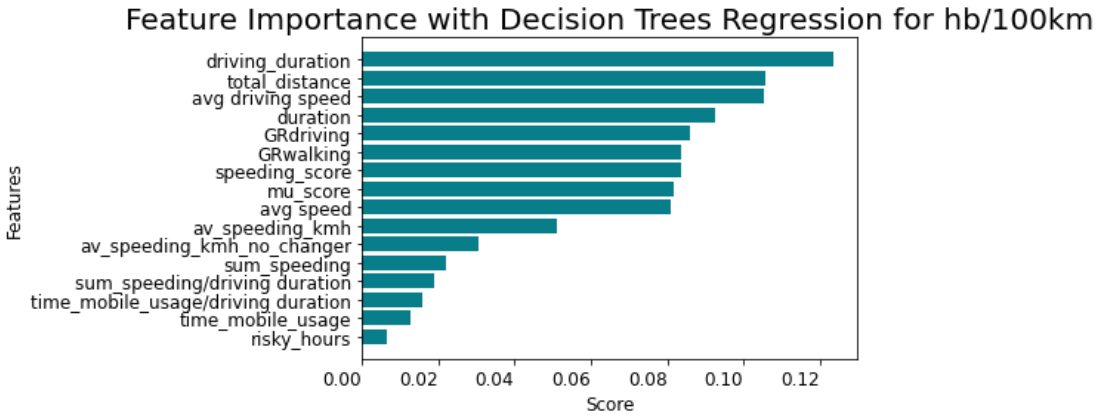
Γράφημα 5.2: *Σημαντικότητα Χαρακτηριστικών για τις απότομές επιβραδύνσεις ανά 100χλμ. με Γραμμική Παλινδρόμηση*

Ο συντελεστής προσδιορισμού R^2 για την πρώτη Γραμμική Παλινδρόμηση προέκυψε ίσος με 0.073, ενώ για την δεύτερη Παλινδρόμηση ίσος με 0.077. Οι συγκεκριμένες τιμές συντελεστή προσδιορισμού αποδεικνύουν την απουσία γραμμικής συμπεριφοράς των μεταβλητών.

5.1.2 Decision Trees Regression



Γράφημα 5.3: *Σημαντικότητα Χαρακτηριστικών για τις απότομές επιταχύνσεις ανά 100χλμ. με Decision Trees*

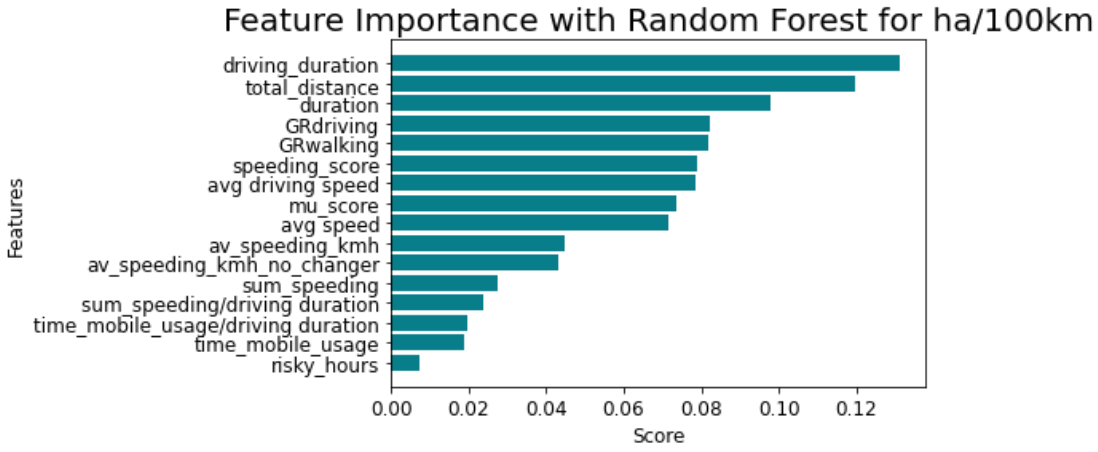


Γράφημα 5.4: *Σημαντικότητα Χαρακτηριστικών για τις απότομες επιβραδύνσεις ανά 100χλμ. με Decision Trees*

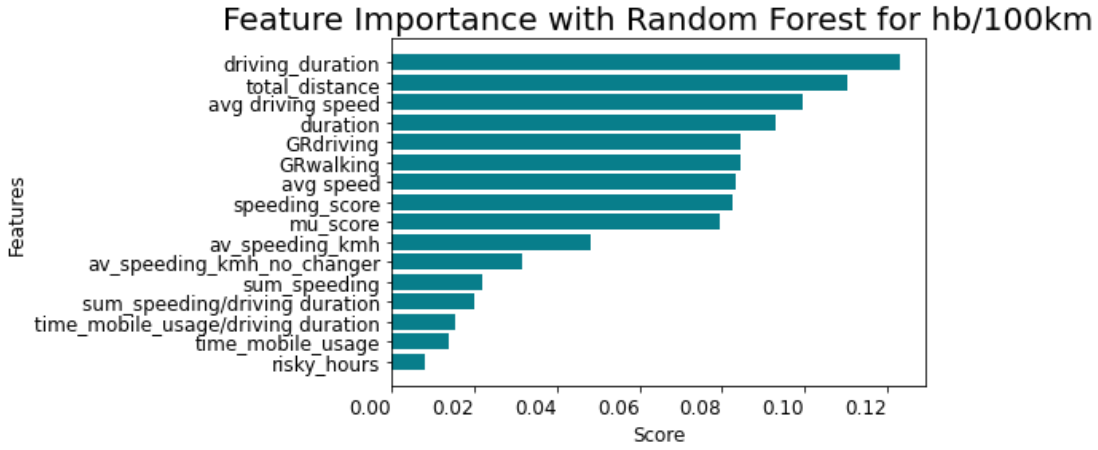
Ο συντελεστής προσδιορισμού R^2 προέκυψε ίσος με 0.999 και για τις δύο Παλινδρομήσεις Decision Trees, καταδεικνύοντας την σχεδόν απόλυτη προσαρμογή μεταβλητών και την προβλεπτική ικανότητα μοντέλου.

5.1.3 Random Forests Regression

Αντίστοιχη διαδικασία ακολουθήθηκε για την Παλινδρόμηση Random Forests που πρόκειται για Παλινδρόμηση Συνόλου.



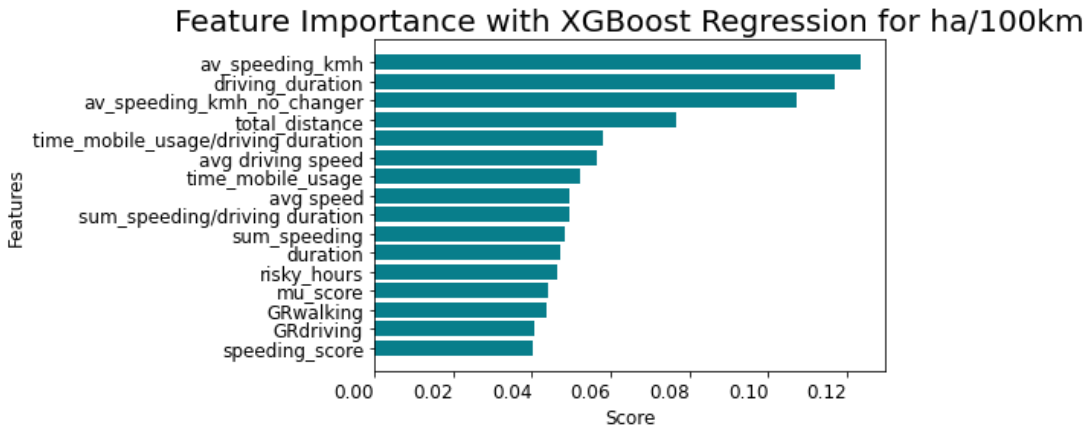
Γράφημα 5.5: *Σημαντικότητα Χαρακτηριστικών για τις απότομες επιταχύνσεις ανά 100χλμ. με Random Forests*



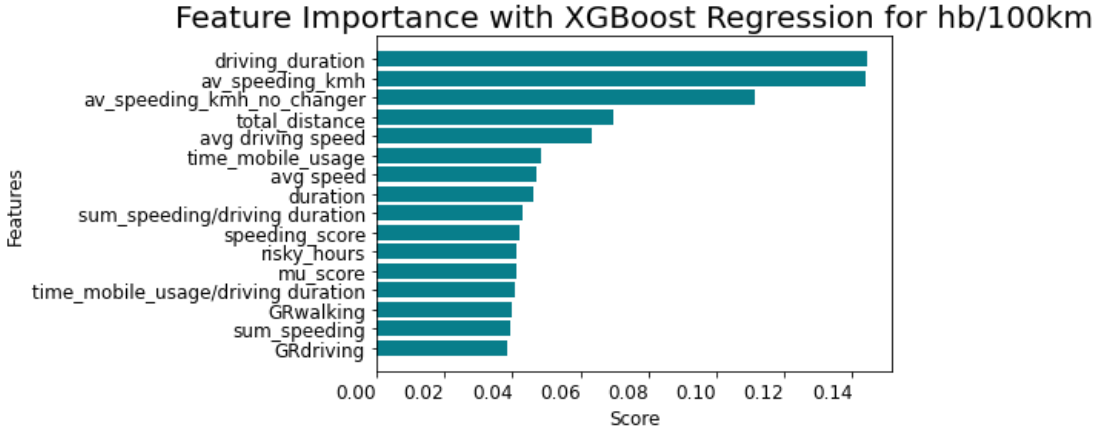
Γράφημα 5.6: *Σημαντικότητα Χαρακτηριστικών για τις απότομες επιβραδύνσεις ανά 100χλμ. με Random Forests*

Ο συντελεστής προσδιορισμού R^2 προέκυψε ίσος με 0.869 για την πρώτη Παλινδρόμηση και 0.871 για την δεύτερη, με τις τιμές να είναι απόλυτα ικανοποιητικές.

5.1.4 XGBoost Regression



Γράφημα 5.7: *Σημαντικότητα Χαρακτηριστικών για τις απότομές επιταχύνσεις ανά 100χλμ. με XGBoost*

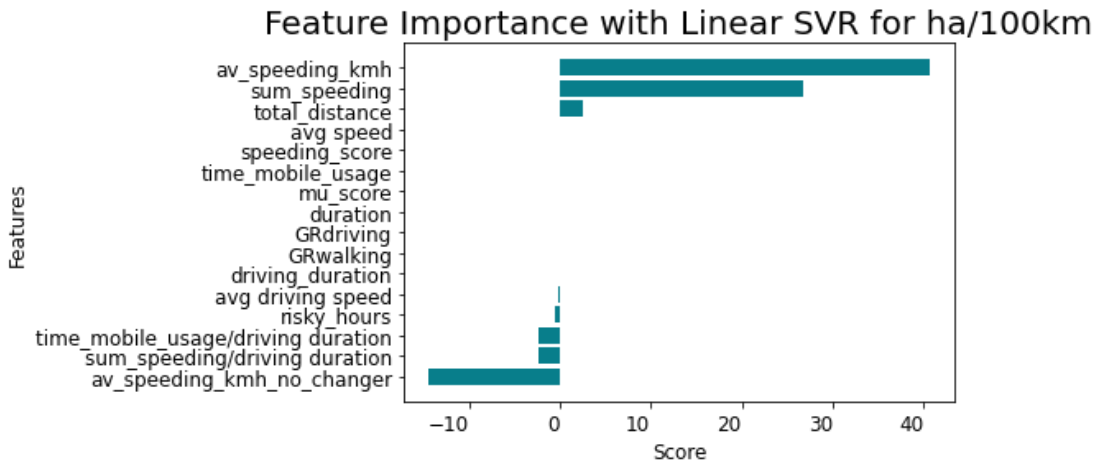


Γράφημα 5.8: *Σημαντικότητα Χαρακτηριστικών για τις απότομές επιβραδύνσεις ανά 100χλμ. με XGBoost*

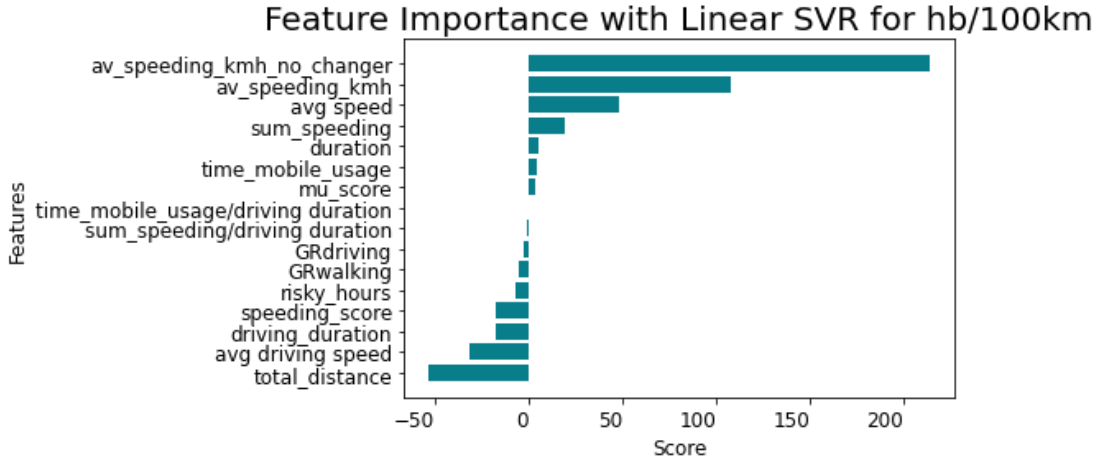
Ο συντελεστής προσδιορισμού R^2 προέκυψε ίσος με 0.165 για τις απότομες επιταχύνσεις ανά 100χλμ και 0.171 για τις απότομες επιβραδύνσεις ανά 100χλμ. Ενώ παρατηρείται προσαρμογή του μοντέλου στις μεταβλητές, τα δεδομένα μπορούν να περιγραφούν καλύτερα από άλλους αλγορίθμους.

5.1.5 Linear SVR

Η σημαντικότητα χαρακτηριστικών υπολογίστηκε με συντελεστή σημαντικότητας (.coef_) για με τον αλγόριθμο Linear SVR, με αποτέλεσμα να προκύπτουν εκ νέου αρνητικές τιμές για τις ανεξάρτητες μεταβλητές.



Γράφημα 5.9: *Σημαντικότητα Χαρακτηριστικών για τις απότομές επιταχύνσεις ανά 100χλμ. με Linear SVR*



Γράφημα 5.10: *Σημαντικότητα Χαρακτηριστικών για τις απότομές επιβραδύνσεις ανά 100χλμ. με Linear SVR*

Ο συντελεστής προσδιορισμού R^2 προέκυψε ίσος με -0.157 για την πρώτη Παλινδρόμηση και ίσος με -0.107 για την δεύτερη, επιβεβαιώνοντας την ύπαρξη θορύβου σε αυτά, αλλά και ακόμα το χαρακτηριστικό πρόβλημα της υπερπροσαρμογής (overfitting), που αποτελεί συνήθης πρόκληση σε προβλήματα Μη Ισορροπημένης Μάθησης.

Συνοψίζοντας τα παραπάνω αποτελέσματα που προέκυψαν από την εξέταση των μοντέλων Παλινδρόμησης, του συντελεστή προσδιορισμού R^2 και του συντελεστή συσχέτισης Pearson των μεταβλητών, προέκυψαν οι 5 σημαντικότερες μεταβλητές, οι οποίες θα χρησιμοποιηθούν εν συνεχεία, ως τα χαρακτηριστικά εισόδου για την διαδικασία της ταξινόμησης. Επισημαίνεται ότι ο συντελεστής προσδιορισμού R^2 είναι η πιο κατατοπιστική απλή μετρική στην αξιολόγηση αναλύσεων παλινδρόμησης (Chicco et al., 2021). Οι 5 προκριθείσες μεταβλητές είναι:

- total_distance: Συνολική διανυθείσα απόσταση
- speeding_score: Σκορ υπερβολικής ταχύτητας
- driving_duration: Συνολική διάρκεια οδήγησης εν κινήσει
- mu_score: Σκορ χρήσης κινητού τηλεφώνου κατά την οδήγηση
- avg driving speed: Μέση ταχύτητα οδήγησης

Εν προκειμένω, οι ανεξάρτητες μεταβλητές μειώθηκαν από 21 σε 5, αποσκοπώντας στην ανάπτυξη αλγορίθμων ταξινόμησης με μεγαλύτερη ακρίβεια πρόβλεψης και ταξινόμησης.

Ειδική μνεία αξίζει στις μεταβλητές που συγκροτούνται από τους φόρτους πεζών και οχημάτων, αφού σύμφωνα με τον συντελεστή συσχέτισης Pearson δεν συμμεταβάλλονταν ισχυρά με τις εξαρτημένες μεταβλητές, όμως σύμφωνα με την Σημαντικότητα Χαρακτηριστικών, παίζουν σημαίνοντα ρόλο στην εμφάνιση απότομων περιστατικών. Η συγκεκριμένη αντίθεση επιβεβαιώνει την πολυεπίπεδη φύση των δεδομένων.

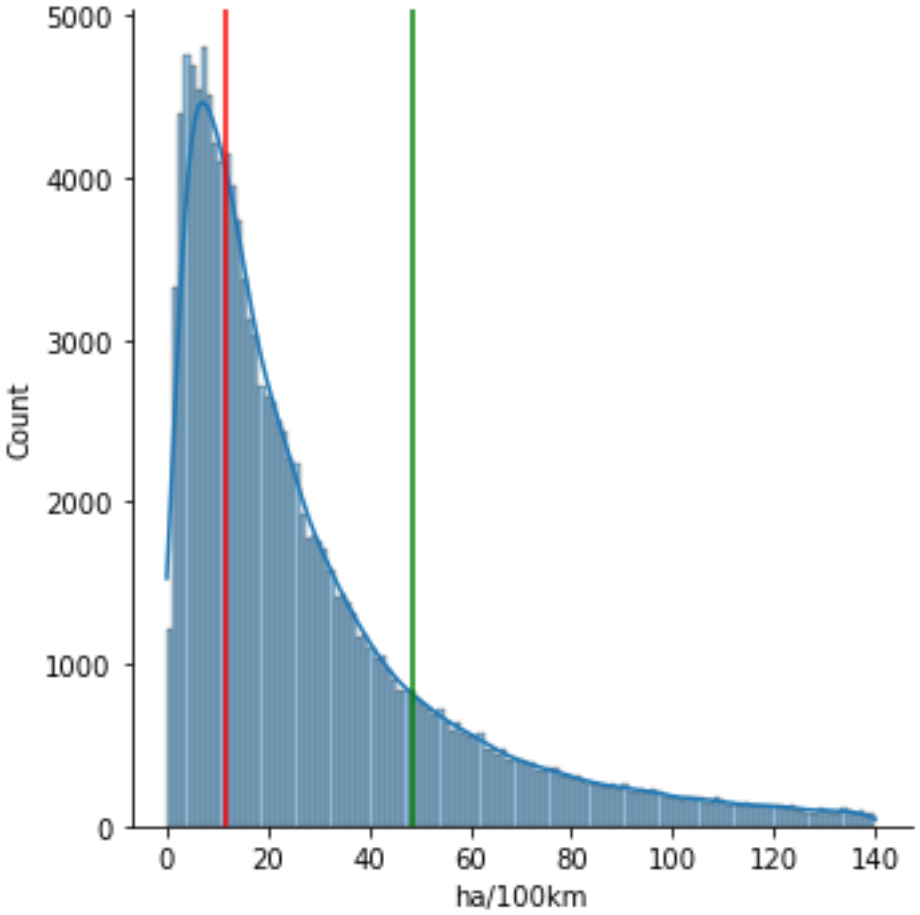
5.2 Προεπεξεργασία δεδομένων

5.2.1 Ομαδοποίηση με αλγόριθμο K-μέσου

Η διαδικασία της ταξινόμησης των επικίνδυνων οδικών συμπεριφορών προϋποθέτει τον διαχωρισμό των δεδομένων εισόδου σε προκαθορισμένες κλάσεις για κατάταξη και πρόγνωση. Στην παρούσα Διπλωματική Εργασία οι εξετασθείσες κλάσεις είναι δύο, οι Μη Επικίνδυνη Οδική συμπεριφορά (Non Dangerous Behaviour) και οι Επικίνδυνη Οδική συμπεριφορά (Dangerous Behaviour). Για να καταστεί εφικτός ο διαχωρισμός των δεδομένων σε δυαδική μορφή και ακολούθως η ομαδοποίησή τους (clustering), προϋποτίθεται η εύρεση ορισμένων ορίων (thresholds). Με την χρήση της βιβλιοθήκης sklearn.cluster, επιλέχθηκε ο αλγόριθμος K-μέσου για την εύρεση των συγκεκριμένων thresholds για κάθε μία εξαρτημένη μεταβλητή, με τις συστάδες να παίζουν τον ρόλο των προκαθορισμένων κλάσεων. Συνεπώς, οι συστάδες επιλέχθηκαν ως δύο και αντίστοιχα δύο είναι και τα κεντροειδή τους.

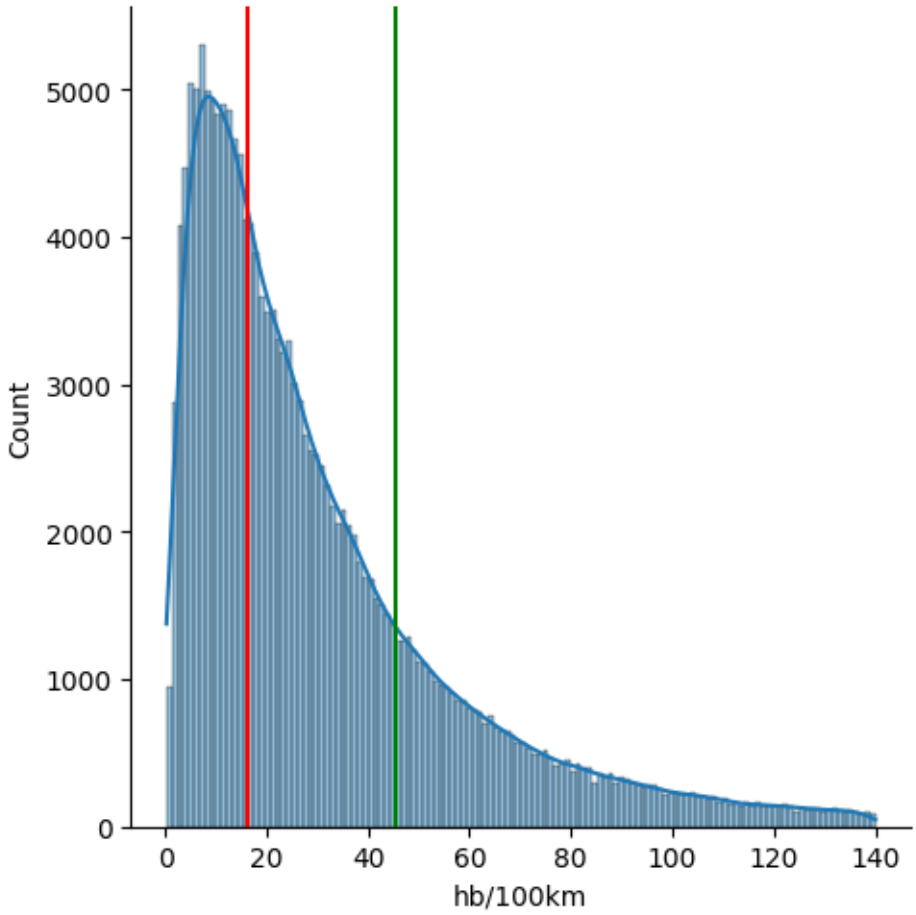
Για την διαδικασία της προεπεξεργασίας των δεδομένων που προηγείται της ταξινόμησης, αναλύθηκαν εκ νέου 3 υποσύνολα. Το ένα υποσύνολο περιλαμβάνει τις επιλεγείσες ανεξάρτητες μεταβλητές, ενώ τα άλλα δύο τις εξαρτημένες. Τα δεδομένα υποβλήθηκαν εκ νέου σε Κανονικοποίηση.

Για το σύνολο τιμών των απότομων επιταχύνσεων ανά 100χλμ., ο αλγόριθμος K-μέσου πραγματοποίησε διανυσματική κβαντοποίηση των δεδομένων σε δύο συστάδες με δεδομένα εξόδου τους Πίνακες 1x1 [5.69291253] και [91.94184525], με το δοθέν όριο τιμών δυαδικού καταμερισμού να δίνεται από τον μέσο όρο των αριθμητικών τιμών των Πινάκων. Το όριο αυτό προέκυψε ίσο με 48.817378890489884. Συνεπώς, τα δεδομένα της μεταβλητής ορίζονται ως 0 για τιμές κάτω του ορίου και ως 1 για τιμές άνω αυτού.



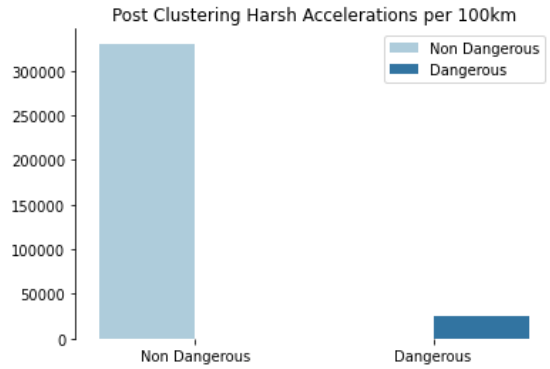
Γράφημα 5.11: *Κεντροειδή και εύρεση threshold για τις απότομες επιταχύνσεις ανά 100χλμ. με αλγόριθμο Κ-μέσου*

Αντίστοιχη διαδικασία πραγματοποιήθηκε και για την δεύτερη μεταβλητή των απότομων επιβραδύνσεων ανά 100χλμ. Ο αλγόριθμος ομαδοποίησης Κ-μέσου για ομαδοποίηση σε 2 συστάδες προτείνει τους μοναδιαίους πίνακες [7.97464008] και [82.83498114], με αποτέλεσμα το όριο κβαντοποίησης να καθοριστεί στην τιμή 45.404810610177094. Συνεπώς, τα δεδομένα της μεταβλητής hb/100km ορίζονται ως 0 για τιμές κάτω του συγκεκριμένου ορίου και ως 1 για τιμές άνω αυτού.

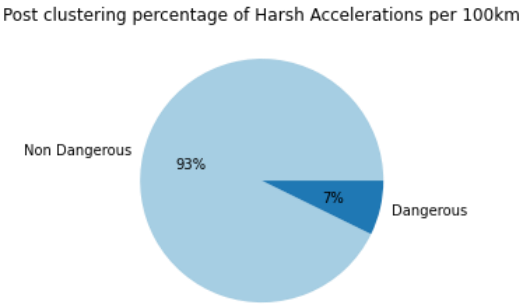


Γράφημα 5.12: *Κεντροειδή και εύρεση threshold για τις απότομες επιβραδύνσεις ανά 100χλμ. με αλγόριθμο Κ-μέσου*

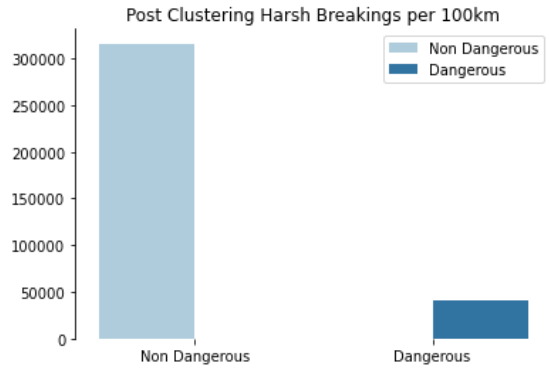
Έπειτα από την διαδικασία ομαδοποίησης με αλγόριθμο Κ-μέσου προέκυψαν οι συστάδες που αποτελούνται από 330.395 στοιχεία Μη Επικίνδυνης Οδικής συμπεριφοράς και 25.767 Επικίνδυνης, για τις απότομες επιταχύνσεις ανά 100χλμ. Αντίστοιχα, προκύπτουν 315.986 Μη Επικίνδυνα στοιχεία και 40.176 Επικίνδυνα, βάσει των απότομων επιβραδύνσεων ανά 100χλμ.



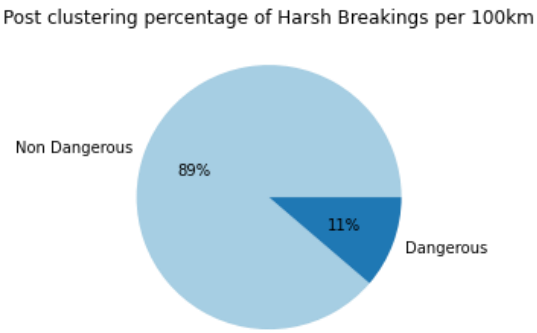
Γράφημα 5.13: *Clustering απότομων επιταχύνσεων ανά 100χλμ. μετά την ομαδοποίηση*



Γράφημα 5.14: *Ποσοστό κατανομής σε τάξεις των απότομων επιταχύνσεων ανά 100χλμ. μετά την ομαδοποίηση*



Γράφημα 5.15: *Clustering απότομων επιβραδύνσεων ανά 100χλμ. μετά την ομαδοποίηση*



Γράφημα 5.16: *Ποσοστό κατανομής σε τάξεις των απότομων επιβραδύνσεων ανά 100χλμ. μετά την ομαδοποίηση*

5.2.2 Μη Ισορροπημένη Μάθηση

5.2.2.1 Διαχωρισμός σε δεδομένα εκπαίδευσης και εξέτασης

Τα δεδομένα που συνθέτουν την παρούσα Διπλωματική Εργασία συνιστούν πρόβλημα Μη Ισορροπημένης Μάθησης, καθώς τα δείγματα που προβλέπουν και ταξινομούν την επικίνδυνη Οδική συμπεριφορά ανήκουν σε μειονοτική τάξη, λόγω της μη ισορροπημένης φύσης τους. Η διαδικασία της ταξινόμησης προϋποθέτει τον διαχωρισμό των δεδομένων σε σύνολα δεδομένων εκπαίδευσης (train data) και εξέτασης (test data). Η συγκεκριμένη διεργασία πραγματοποιήθηκε με την τεχνική train test split της βιβλιοθήκης επεξεργασίας sklearn.model_selection, η οποία διαχωρίζει ένα υπερσύνολο δεδομένων σε πίνακες από τυχαία κατανεμημένα στοιχεία εκπαίδευσης και εξέτασης. Εκ της σύστασής τους, τα δεδομένα εκπαίδευσης αξιοποιούνται για την εκμάθηση των αλγορίθμων Μη Επιβλεπόμενης Μάθησης (Unsupervised Learning), εν αντιθέσει με τα δεδομένα εξέτασης, τα οποία αξιολογούν την απόδοση των μοντέλων ταξινόμησης και διασφαλίζουν ότι το μοντέλο δύναται να προβλέψει και να ταξινομήσει αποτελεσματικά. Η αναλογία που επιλέχθηκε για την εκπαίδευση και την εξέταση των μοντέλων ήταν 75% - 25%, υπέρ του συνόλου εκπαίδευσης (train data). Για τις απότομες επιταχύνσεις ανά 100χλμ., μετά την διαδικασία διαχωρισμού σε σύνολα εκπαίδευσης και ελέγχου, προέκυψαν τα εξής σύνολα εκπαίδευσης: Μη Επικίνδυνα στοιχεία: 247.796, Επικίνδυνα στοιχεία: 19.325, για τις απότομες επιταχύνσεις ανά 100χλμ. και αντίστοιχα Μη Επικίνδυνα: 236.989 Επικίνδυνα: 30.132, για τις απότομες επιβραδύνσεις ανά 100χλμ.

5.2.2.2 Υπερδειγματοληψία

Η μεταχείριση της μειονοτικής τάξης αποτελεί κομβικό ζήτημα στην ανάπτυξη των αλγορίθμων ταξινόμησης και στην βελτίωση της προβλεπτικής ικανότητας των μοντέλων, όντας παράμετρος που επηρεάζει αισθητά την απόδοσή τους. Γενικά, τα μοντέλα ταξινόμησης βασίζουν την λειτουργία τους σε προβλήματα Ισορροπημένης Μάθησης, καθιστώντας τα ασυνεπή σε

προβλήματα άνισων κατανομών. Πιο συγκεκριμένα, τα δεδομένα που ανήκουν στην Επικίνδυνη κλάση είναι σαφώς μικρότερα σε όγκο από αυτά που ανήκουν στην Μη Επικίνδυνη. Εν προκειμένω, για την αποδοτικότερη αξιοποίηση των αλγορίθμων ταξινόμησης απαιτείται Επαναδειγματοληψία των δεδομένων εκπαίδευσης. Η μέθοδος της Υποδειγματοληψίας (Undersampling) δεν προκρίθηκε στην παρούσα έρευνα, υπό τον φόβο της απώλειας σημαντικής πληροφορίας, καθώς η διακύμανση των τιμών μεταβλητών είναι ισχυρή. Στην παρούσα Διπλωματική Εργασία οι επιλεγείσες κλάσεις ταξινόμησης είναι διακριτές, με δυαδικό διαχωρισμό των εξαρτημένων μεταβλητών, όπερ σημαίνει ότι το ζητούμενο πρόβλημα ταξινόμησης δεν εμπίπτει στο πεδίο των δυσδιάκριτων κλάσεων. Επομένως, προκρίνεται η μέθοδος της Υπερδειγματοληψίας (Oversampling), με επιλεγείσα τεχνική αυτή της Συνθετικής Μειονοτικής Υπερδειγματοληψίας (Synthetic Minority Oversampling – SMOTE). Εναλλακτικές τεχνικές όπως η μέθοδος της Προσαρμοστικής Συνθετικής (ADASYN), δεν προκρίθηκε, λόγω ζητημάτων στάθμισης που απαιτούταν στα δείγματα, καθώς οι διακυμάνσεις ήταν σχετικώς ισχυρές. Ένα εξίσου κρίσιμο ζήτημα της Προσαρμοστικής Συνθετικής, είναι η παραγωγή τεχνητών δεδομένων εκπαίδευσης, με χαρακτηριστικά απολύτως πανομοιότυπα με αυτά των γονικών τους σε καταστάσεις μεγάλης αναλογίας δεδομένων πλειονότητας όπως αυτά που αναλύονται στην συγκεκριμένη έρευνα, με αποτέλεσμα τον κίνδυνο παραγωγής υψηλών ποσοστών Ψευδώς Θετικών.

Αφού εφαρμόστηκε η τεχνική SMOTE στα δύο σύνολα δεδομένων εκπαίδευσης για τις δύο ξεχωριστές εξαρτημένες μεταβλητές, η αναλογία της Επικίνδυνης και Μη Επικίνδυνης τάξης κατέστη 1:1. Για τις απότομες επιταχύνσεις ανά 100χλμ., τα στοιχεία κάθε τάξης προέκυψαν ίσα με 247.796, ενώ για τις απότομες επιβραδύνσεις ανά 100χλμ., ίσα με 236.989.

Γράφημα 5.17: *Ποσοστό κατανομών σε τάξεις απότομων περιστατικών ανά 100χλμ μετά το Oversampling*

5.3 Ταξινόμηση απότομων περιστατικών

Έπειτα από την διαδικασία της Υπερδειγματοληψίας, η διαδικασία της αναγνώρισης, ταξινόμησης και πρόβλεψης της Επικίνδυνης Οδικής συμπεριφοράς ανήχθη σε πρόβλημα Ισορροπημένης Μάθησης, γεγονός που καθιστά τους αλγορίθμους ταξινόμησης που πρόκειται να αξιοποιηθούν, πιο αποτελεσματικούς. Όπως έχει ήδη αναφερθεί, η ταξινόμηση της Οδικής συμπεριφοράς θα έχει δυαδική μορφή σε επιμέρους διακριτές κλάσεις: Επικίνδυνη συμπεριφορά και Μη Επικίνδυνη συμπεριφορά. Στην παρούσα υποενότητα θα γίνει η εκπαίδευση των επιλεγμένων αλγορίθμων

ταξινόμησης με τεχνικές Ισορροπημένης Μηχανικής Μάθησης και θα παρουσιαστούν συνοπτικά τα αποτελέσματα καθενός αλγορίθμου.

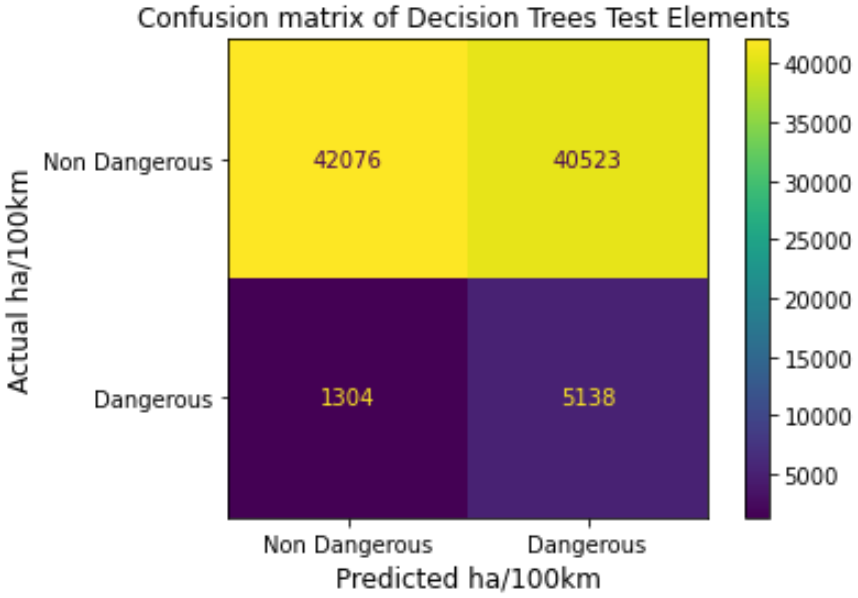
Προκειμένου οι αλγόριθμοι να αποδώσουν το καλύτερο δυνατό αποτέλεσμα για την συγκεκριμένη βάση δεδομένων, πριν την εκπαίδευση των μοντέλων, πραγματοποιήθηκε η βελτιστοποίηση υπερπαραμέτρων, μέσω της Αναζήτησης Πλέγματος (GridSearch). Για κάθε αλγόριθμο ξεχωριστά, η τεχνική Αναζήτησης Πλέγματος από την βιβλιοθήκη sklearn.model_selection εντοπίζει τις καλύτερες υπερπαραμέτρους για το δεδομένο σύνολο τιμών, μεγιστοποιώντας τις επιδόσεις των αλγορίθμων.

Οι αλγόριθμοι ταξινόμησης που εκπαιδεύτηκαν είναι τα Decision Trees, Gradient Boosting, XGBoost, Random Forests, AdaBoost, SVM και Multilayered Perceptrons Η διαδικασία εκπαίδευσης των αλγορίθμων πραγματοποιήθηκε δύο φορές για κάθε εξαρτημένη μεταβλητή και τα αποτελέσματά τους παρουσιάζονται παρακάτω. Επιπλέον, παρατίθενται ως γραφική αναπαράσταση επίδοσης του εκάστοτε μοντέλου, η μήτρα σύγχυσής του (confusion matrix), καθώς και οι μετρικές αξιολογήσεις για τον έλεγχο της ικανότητας πρόβλεψης και ταξινόμησής του. Η ανάλυση των μοντέλων πραγματοποιήθηκε με την βοήθεια της γλώσσας προγραμματισμού Python, σε προγραμματιστικό περιβάλλον Jupyter Notebook και Google Colab.

5.3.1 Decision Trees Classification

5.3.1.1 Απότομες επιταχύνσεις ανά 100χλμ.

Για την ταξινόμηση των απότομων επιταχύνσεων ανά 100χλμ. η τεχνική GridSearch συνέστησε τις υπερπαραμέτρους: criterion=entropy, max_depth=4. Η Ορθότητα εκπαίδευσης του μοντέλου ανήλθε στο 66%. Ο γεωμετρικός μέσος G-mean που καταδεικνύει την ισορροπία μεταξύ των επιδόσεων ταξινόμησης στην μειονοτική τάξη και στην τάξη πλειονότητας είναι ίσος με 0.65.



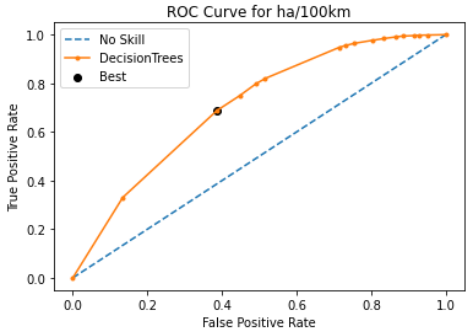
Γράφημα 5.18: *Μήτρα σύγχυσης δεδομένων ελέγχου αλγορίθμου Decision Trees για τις απότομες επιταχύνσεις ανά 100χλμ.*

Η μήτρα σύγχυσης του μοντέλου αποτελείται από μεγάλη αναλογία Ψευδώς Θετικών, καθιστώντας το μοντέλο υπερβολικά συντηρητικό.

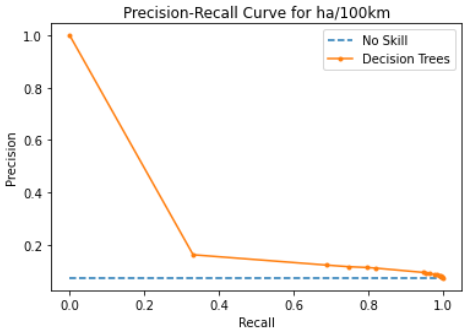
Πίνακας 5.1: *Επίδοση αλγορίθμου Decision Trees για τις απότομες επιταχύνσεις ανά 100χλμ.*

Οδική Συμπεριφορά	**Ακρίβεια**	**Ανάκληση**	**f1-score**	**Σύνολο δεδομένων εξέτασης**
Μη Επικίνδυνη	0.97	0.51	0.67	82599
Επικίνδυνη	0.11	0.80	0.20	6442
Μέσος όρος	0.54	0.63	0.43	89041
Σταθμισμένος μέσος	0.91	0.53	0.63	89041

Σύμφωνα με τον Πίνακα 5.1, το επίπεδο της τάξης Επικίνδυνης Οδικής συμπεριφοράς παρουσιάζει αξιόλογη προβλεπτική ικανότητα της τάξης του 80%, όμως επιβεβαιώνεται η συντηρητική φύση του μοντέλου με την Ανάκληση της Μη Επικίνδυνης τάξης στο 51%. Επιπλέον, η ακρίβεια στην ταξινόμηση της Μη Επικίνδυνης τάξης είναι 97%, ένα ποσοστό που κρίνεται επίσης αξιόλογο. Στα γραφήματα που ακολουθούν γίνεται αναπαράσταση των Καμπύλων ROC και Ακρίβειας-Ανάκλησης του μοντέλου. Το σκορ Περιοχής κάτω από την Καμπύλη (AUC score) υπολογίστηκε στο 70.5%, που κρίνεται ικανοποιητικό.



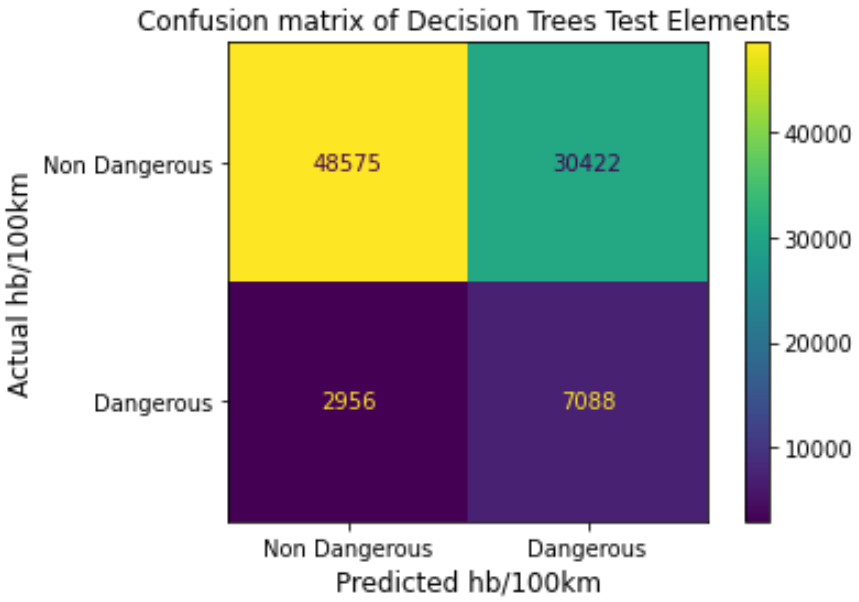
Γράφημα 5.19: *Καμπύλη ROC αλγορίθμου Decision Trees για τις απότομες επιταχύνσεις ανά 100χλμ.*



Γράφημα 5.20: *Καμπύλη Ακρίβειας-Ανάκλησης αλγορίθμου Decision Trees για τις απότομες επιταχύνσεις ανά 100χλμ.*

5.3.1.2 Απότομες επιβραδύνσεις ανά 100χλμ.

Η ρύθμιση υπερπαραμέτρων κατέδειξε ως βέλτιστες τις υπερπαραμέτρους: criterion=gini, max_depth=12. Η Ορθότητα εκπαίδευσης ήταν 71%, ενώ το G-mean ήταν 0.659.



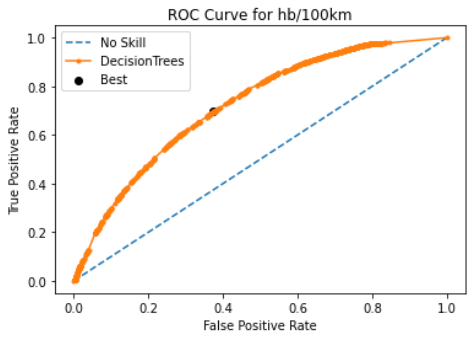
Γράφημα 5.21: *Μήτρα σύγχυσης δεδομένων ελέγχου αλγορίθμου Decision Trees για τις απότομες επιβραδύνσεις ανά 100χλμ.*

Η μήτρα σύγχυσης του μοντέλου δείχνει καλύτερη συμπεριφορά των Ψευδώς Θετικών, όμως υπερδιπλάσια απόδοση των Ψευδώς Αρνητικών.

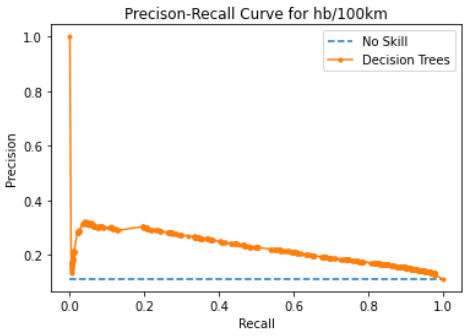
Πίνακας 5.2: *Επίδοση αλγορίθμου Decision Trees για τις απότομες επιβραδύνσεις ανά 100χλμ.*

Οδική Συμπεριφορά	**Ακρίβεια**	**Ανάκληση**	**f1-score**	**Σύνολο δεδομένων εξέτασης**
Μη Επικίνδυνη	0.94	0.61	0.74	78997
Επικίνδυνη	0.19	0.71	0.30	10044
Μέσος όρος	0.57	0.66	0.52	89041
Σταθμισμένος μέσος	0.86	0.63	0.69	89041

Οι μετρικές αξιολογήσεις του Πίνακα 5.2, καταδεικνύουν ότι το μοντέλο δεν είναι εξαιρετικά αξιόπιστο. Το ποσοστό Ψευδώς Αρνητικών φτάνει το 33,97%, ένα μετρικό ιδιαίτερα σημαντικό για το πρόβλημα ταξινόμησης, γεγονός που καθιστά την συνολική ταξινόμηση μη ικανοποιητική. Η συνολική Ανάκληση του μοντέλου, όμως, είναι σχετικώς αποδεκτή.



Γράφημα 5.22: *Καμπύλη ROC αλγορίθμου Decision Trees για τις απότομες επιβραδύνσεις ανά 100χλμ.*



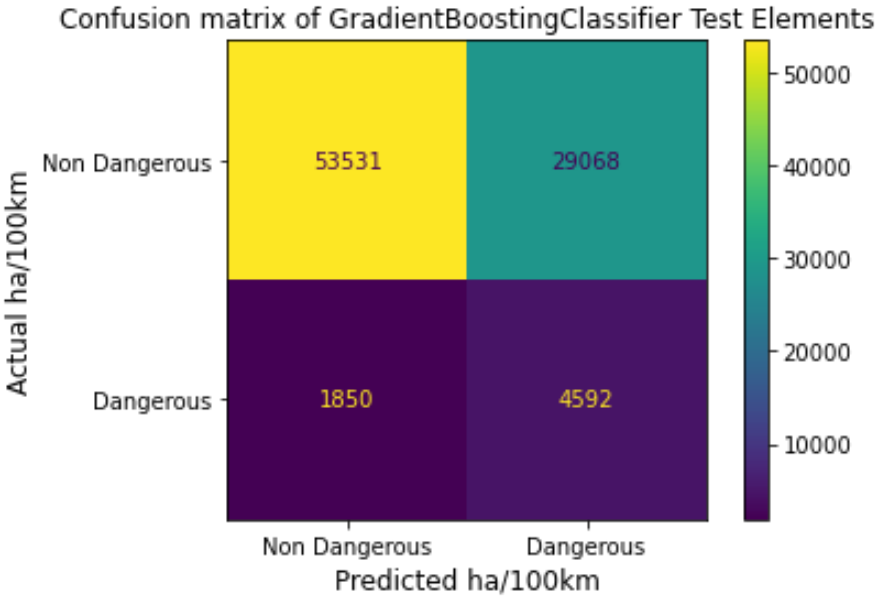
Γράφημα 5.23: *Καμπύλη Ακρίβειας-Ανάκλησης αλγορίθμου Decision Trees για τις απότομες επιβραδύνσεις ανά 100χλμ.*

Το AUC score της καμπύλης ROC είναι ίσο με 72,34%, ένα σχετικά καλό ποσοστό πρόβλεψης, όμως η Καμπύλη Ακρίβειας-Ανάκλησης δεν είναι ικανοποιητική.

5.3.2 Gradient Boosting Classification

5.3.2.1 Απότομες επιταχύνσεις ανά 100χλμ.

Η ρύθμιση υπερπαραμέτρων με το GridSearch συνιστά την χρήση: max_depth=6, n_estimators=200. Η Ορθότητα εκπαίδευσης ήταν 73%, τιμή που κρίνεται σχετικώς ικανοποιητική, ενώ το G-mean ήταν 0.681.



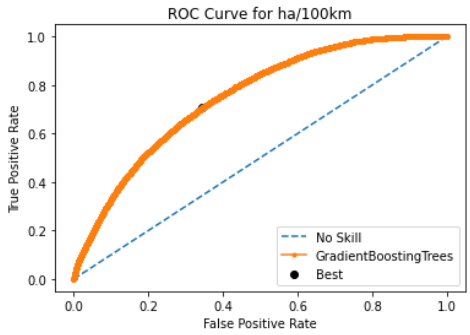
Γράφημα 5.24: *Μήτρα σύγχυσης δεδομένων ελέγχου αλγορίθμου Gradient Boosting για τις απότομες επιταχύνσεις ανά 100χλμ.*

Η μήτρα σύγχυσης δίνει πολύ καλύτερα αποτελέσματα από την αντίστοιχη των Decision Trees, με αρκετά μειωμένο ποσοστό FNR και FPR.

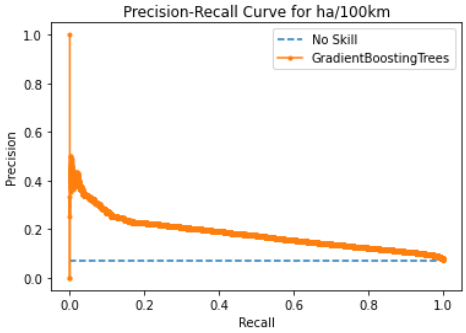
Πίνακας 5.3: *Επίδοση αλγορίθμου Gradient Boosting για τις απότομες επιταχύνσεις ανά 100χλμ.*

Οδική Συμπεριφορά	**Ακρίβεια**	**Ανάκληση**	**f1-score**	**Σύνολο δεδομένων εξέτασης**
Μη Επικίνδυνη	0.97	0.65	0.78	82599
Επικίνδυνη	0.14	0.71	0.23	6442
Μέσος όρος	0.55	0.68	0.50	89041
Σταθμισμένος μέσος	0.91	0.65	0.74	89041

Σύμφωνα με τις μετρικές αξιολογήσεις του Πίνακα 5.3, η σταθμισμένη Ακρίβεια του μοντέλου είναι 91%, ενώ η Ανάκληση στην Επικίνδυνη τάξη είναι 71%. Και οι δύο μετρικές κρίνονται αξιόλογες. Το κρίσιμο για την έρευνα ποσοστό Ψευδώς Αρνητικών του μοντέλου είναι 31.95%.



Γράφημα 5.25: *Καμπύλη ROC αλγορίθμου Gradient Boosting για τις απότομες επιταχύνσεις ανά 100χλμ.*

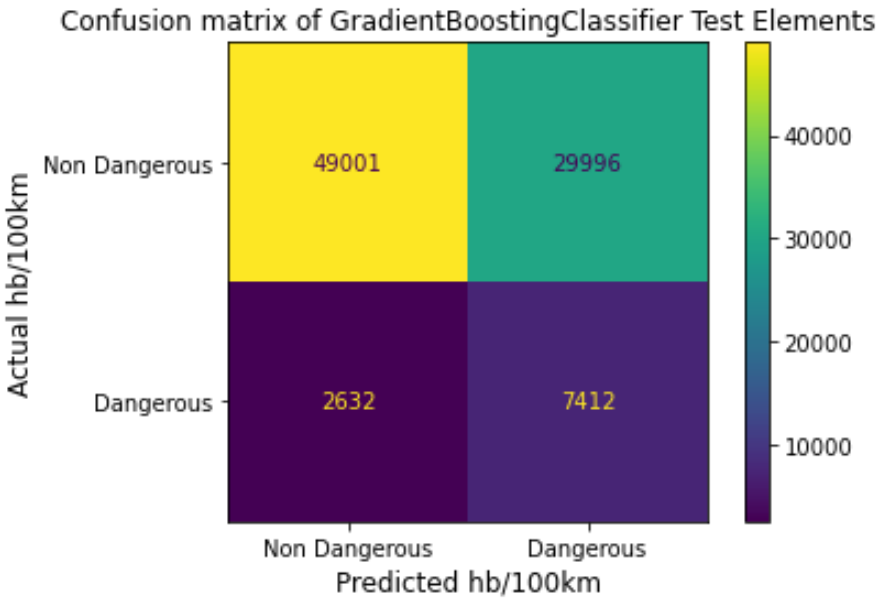


Γράφημα 5.26: *Καμπύλη Ακρίβειας-Ανάκλησης αλγορίθμου Gradient Boosting για τις απότομες επιταχύνσεις ανά 100χλμ.*

Το AUC score της καμπύλης ROC είναι ίσο με 75,1%, καταδεικνύοντας την ικανοποιητική προβλεπτική ικανότητα του αλγορίθμου GradientBoosting, ενώ και η καμπύλη Ακρίβειας-Ανάκλησης προτείνει ένα καλό προβλεπτικά μοντέλο.

5.3.2.2 Απότομες επιβραδύνσεις ανά 100χλμ.

Οι βέλτιστες υπερπαράμετροι ελέγχθηκαν με το GridSearch και επιλέχθηκαν max_depth=6, n_estimators=200. Η Ορθότητα εκπαίδευσης ήταν 71% και το G-mean ίσο με 0.679.



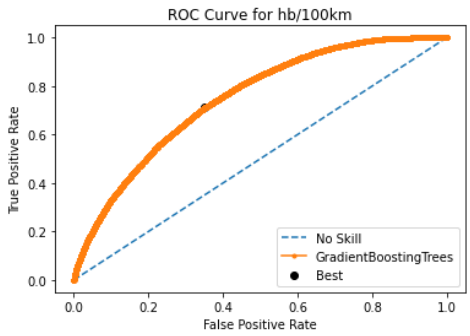
Γράφημα 5.27: *Μήτρα σύγχυσης δεδομένων ελέγχου αλγορίθμου Gradient Boosting για τις απότομες επιβραδύνσεις ανά 100χλμ.*

Η μήτρα σύγχυσης με Gradient Boosting για τις απότομες επιβραδύνσεις ανά 100χλμ δίνει και αυτή καλύτερα αποτελέσματα από την αντίστοιχη των Decision Trees, με καλύτερα ποσοστά FNR και FPR επίσης.

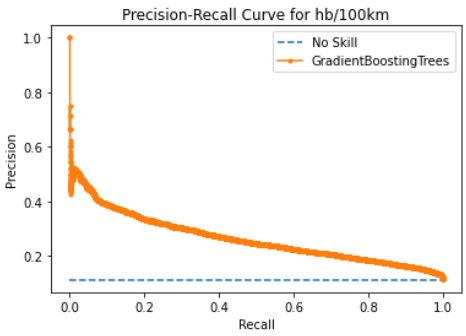
Πίνακας 5.4: *Επίδοση αλγορίθμου Gradient Boosting για τις απότομες επιβραδύνσεις ανά 100χλμ.*

Οδική Συμπεριφορά	**Ακρίβεια**	**Ανάκληση**	**f1-score**	**Σύνολο δεδομένων εξέτασης**
Μη Επικίνδυνη	0.95	0.62	0.75	78997
Επικίνδυνη	0.20	0.74	0.31	10044
Μέσος όρος	0.57	0.68	0.53	89041
Σταθμισμένος μέσος	0.86	0.63	0.70	89041

Η Ανάκληση για την Επικίνδυνη τάξη είναι ικανοποιητική, με την σταθμισμένη Ακρίβεια να διατηρείται σε καλά επίπεδα. Το ποσοστό Ψευδώς Αρνητικών (FNR) είναι 32.1%.



Γράφημα 5.28: *Καμπύλη ROC αλγορίθμου Gradient Boosting για τις απότομες επιβραδύνσεις ανά 100χλμ.*



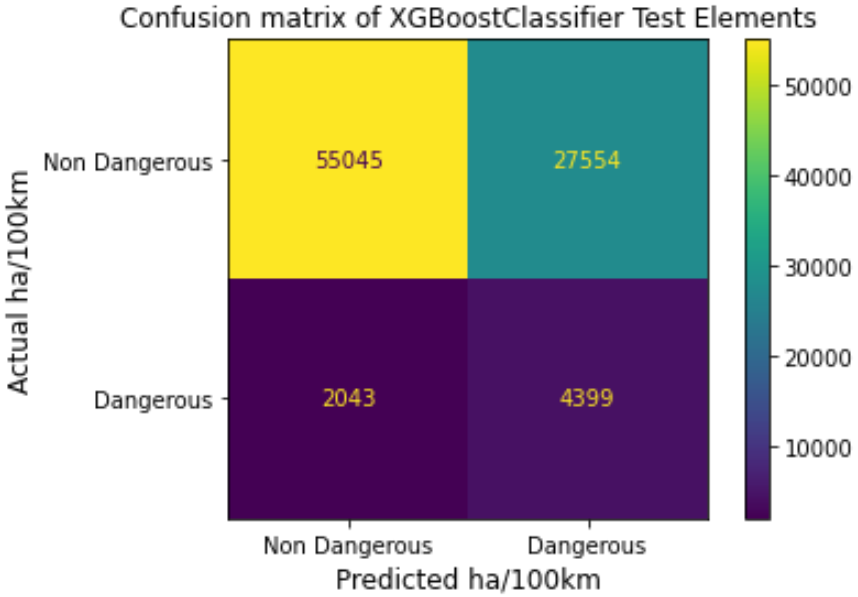
Γράφημα 5.29: *Καμπύλη Ακρίβειας-Ανάκλησης αλγορίθμου Gradient Boosting για τις απότομες επιβραδύνσεις ανά 100χλμ.*

Το AUC score της καμπύλης ROC είναι ίσο με 74,87%. Η συγκεκριμένη τιμή δίνει ένα αξιόπιστο μοντέλο ταξινόμησης και πρόβλεψης. Επιπλέον, η καμπύλη Ακρίβειας-Ανάκλησης που αναπαρίσταται στο Γράφημα 5.29 επιβεβαιώνει τον παραπάνω ισχυρισμό.

5.3.3 XGBoost Classification

5.3.3.1 Απότομες επιταχύνσεις ανά 100χλμ.

Η ρύθμιση υπερπαραμέτρων με GridSearch έδωσε ως βέλτιστες τις υπερπαραμέτρους max_depth=6 και n_estimators=200. Η Ορθότητα ήταν 76% και το G-mean=0.675.



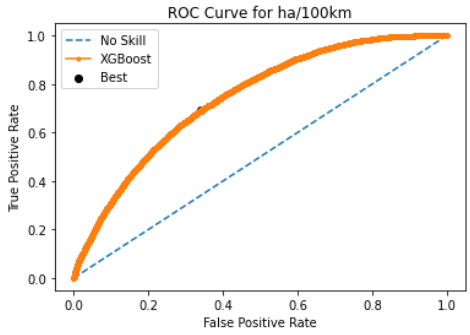
Γράφημα 5.30: *Μήτρα σύγχυσης δεδομένων ελέγχου αλγορίθμου XGBoost για τις απότομες επιταχύνσεις ανά 100χλμ.*

Η μήτρα σύγχυσης του μοντέλου αποτελείται από χαμηλότερο ποσοστό Ψευδώς Θετικών συγκριτικά με το αντίστοιχο μοντέλο Gradient Boosting, όμως το FNR είναι αισθητά μεγαλύτερο.

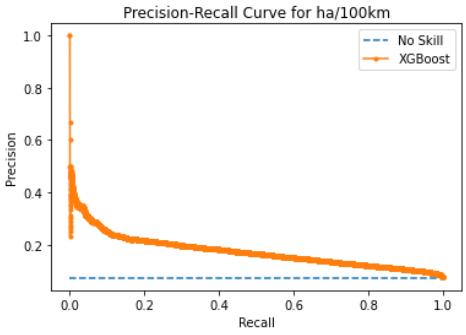
Πίνακας 5.5: *Επίδοση αλγορίθμου XGBoost για τις απότομες επιταχύνσεις ανά 100χλμ.*

Οδική Συμπεριφορά	**Ακρίβεια**	**Ανάκληση**	**f1-score**	**Σύνολο δεδομένων εξέτασης**
Μη Επικίνδυνη	0.96	0.67	0.79	82599
Επικίνδυνη	0.14	0.68	0.23	6442
Μέσος όρος	0.55	0.67	0.51	89041
Σταθμισμένος μέσος	0.90	0.67	0.75	89041

Για την τάξη Μη Επικίνδυνης Οδικής συμπεριφοράς, το ποσοστό λανθασμένων προβλέψεων φτάνει το 33%, ενώ για την Επικίνδυνη τάξη το 32%. Ο σταθμισμένος μέσος της Ακρίβειας είναι επαρκής, ενώ η Ανάκληση στο σύνολο του μοντέλου καθιστά το μοντέλο αποδεκτό. Γενικότερα, φαίνεται να είναι ένα καλό μοντέλο πρόβλεψης.



Γράφημα 5.31: *Καμπύλη ROC αλγορίθμου XGBoost για τις απότομες επιταχύνσεις ανά 100χλμ.*

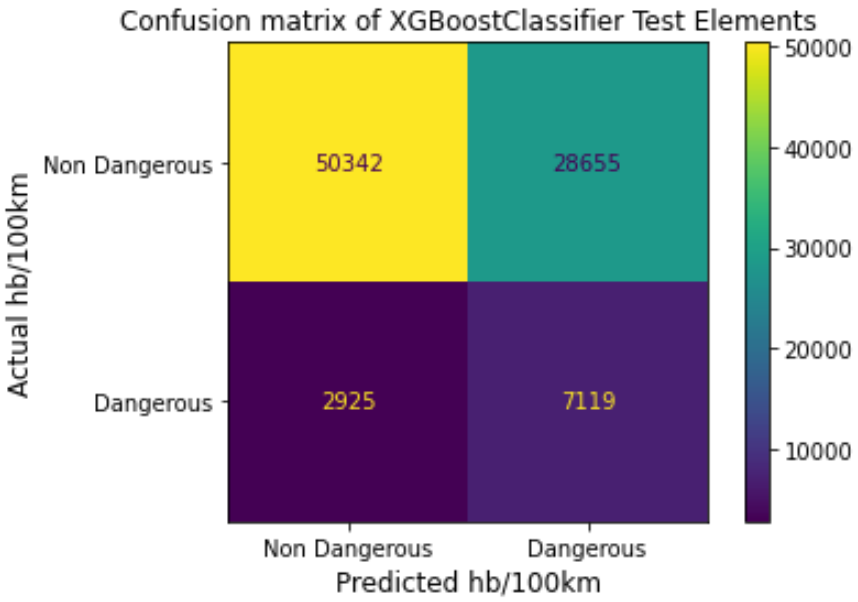


Γράφημα 5.32: *Καμπύλη Ακρίβειας-Ανάκλησης αλγορίθμου XGBoost για τις απότομες επιταχύνσεις ανά 100χλμ.*

Το AUC score της καμπύλης ROC είναι ίσο με 74,25%, τιμή που δείχνει καλή απόδοση, ενώ και η Καμπύλη Ακρίβειας-Ανάκλησης δίνει καλό γραφικό αποτέλεσμα, όμως αισθητά χειρότερο του αντίστοιχου Gradient Boosting.

5.3.3.2 Απότομες επιβραδύνσεις ανά 100χλμ.

Η τεχνική GridSearch κατέδειξε την χρήση των ακόλουθων υπερπαραμέτρων: max_depth=6 και n_estimators=200. Η Ορθότητα εκπαίδευσης του αλγορίθμου ήταν 73% και το G-mean=0.672.



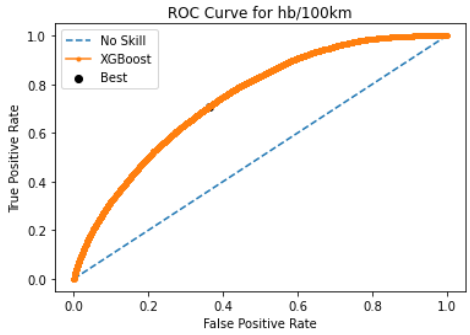
Γράφημα 5.33: *Μήτρα σύγχυσης δεδομένων ελέγχου αλγορίθμου XGBoost για τις απότομες επιβραδύνσεις ανά 100χλμ.*

Η μήτρα σύγχυσης έχει μειωμένα τα επίπεδα FPR, δίνοντας λιγότερα συντηρητικά αποτελέσματα, όμως αισθητά ανεβασμένα αποτελέσματα FNR καταδεικνύοντας κρίσιμα λάθη ταξινόμησης.

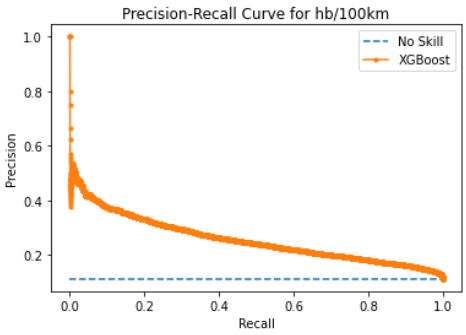
Πίνακας 5.6: *Επίδοση αλγορίθμου XGBoost για τις απότομες επιβραδύνσεις ανά 100χλμ.*

Οδική Συμπεριφορά	**Ακρίβεια**	**Ανάκληση**	**f1-score**	**Σύνολο δεδομένων εξέτασης**
Μη Επικίνδυνη	0.95	0.64	0.76	78997
Επικίνδυνη	0.20	0.71	0.31	10044
Μέσος όρος	0.57	0.67	0.54	89041
Σταθμισμένος μέσος	0.86	0.65	0.71	89041

Η Ακρίβεια της πρώτης τάξης κρίνεται αρκετά ικανοποιητική, ενώ και της δεύτερης τάξης σημειώνει αξιοπρόσεκτες επιδόσεις. Το ποσοστό λανθασμένων προβλέψεων για την Επικίνδυνη τάξη είναι 29%, δηλαδή αποδεικνύεται ένα μοντέλο που έχει καλή προβλεπτική ικανότητα.



Γράφημα 5.34: *Καμπύλη ROC αλγορίθμου XGBoost για τις απότομες επιβραδύνσεις ανά 100χλμ.*



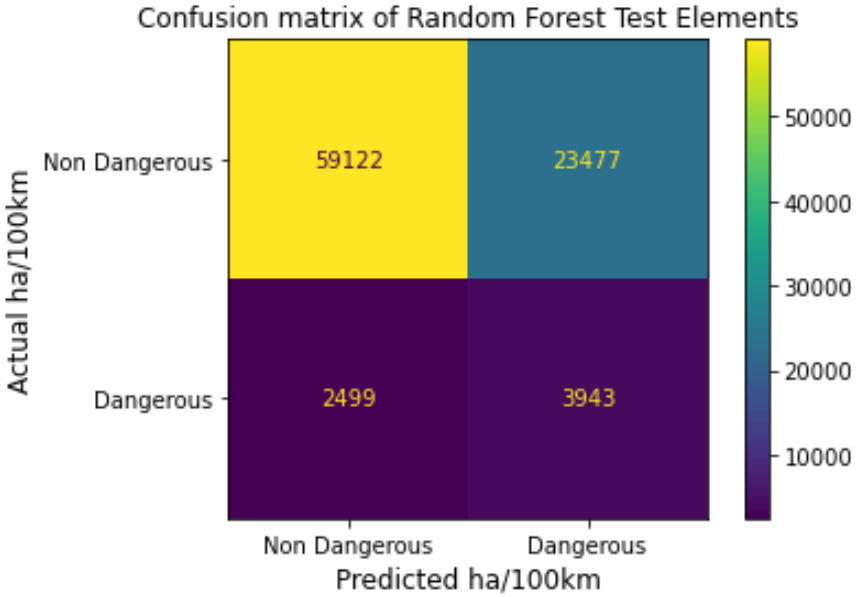
Γράφημα 5.35: *Καμπύλη Ακρίβειας-Ανάκλησης αλγορίθμου XGBoost για τις απότομες επιβραδύνσεις ανά 100χλμ.*

Το AUC score της καμπύλης ROC είναι ίσο με 74,27% επιβεβαιώνοντας την καλή προβλεπτική ικανότητα του αλγορίθμου για τις απότομες επιβραδύνσεις ανά 100χλμ, ενώ και η Καμπύλη Ακρίβειας-Ανάκλησης κυμαίνεται σε αρκετά καλά επίπεδα.

5.3.4 Random Forests Classification

5.3.4.1 Απότομες επιταχύνσεις ανά 100χλμ.

Η μέθοδος ρύθμισης και εύρεσης των βέλτιστων υπερπαραμέτρων έδωσε: max_depth=16 και n_estimators=256. Η Ορθότητα ήταν υψηλή της τάξης του 84% και ο γεωμετρικός μέσος των τάξεων G-mean ήταν 0.672.



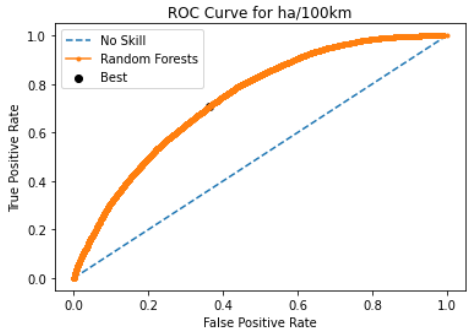
Γράφημα 5.36: *Μήτρα σύγχυσης δεδομένων ελέγχου αλγορίθμου Random Forests για τις απότομες επιταχύνσεις ανά 100χλμ.*

Η μήτρα σύγχυσης δείχνει να ταξινομεί πολύ καλύτερα τα στοιχεία που ανήκαν σε Ψευδώς Θετικά στοιχεία στα προηγούμενα μοντέλα, όμως ο δείκτης FNR είναι ιδιαίτερα υψηλός.

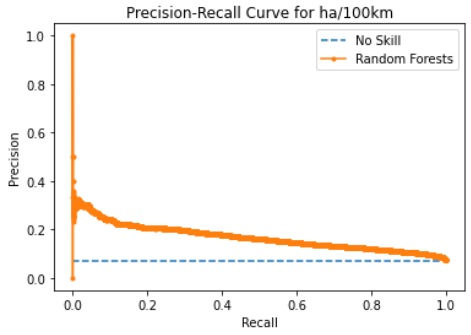
Πίνακας 5.7: *Επίδοση αλγορίθμου Random Forests για τις απότομες επιταχύνσεις ανά 100χλμ.*

Οδική Συμπεριφορά	**Ακρίβεια**	**Ανάκληση**	**f1-score**	**Σύνολο δεδομένων εξέτασης**
Μη Επικίνδυνη	0.96	0.72	0.82	82599
Επικίνδυνη	0.14	0.61	0.23	6442
Μέσος όρος	0.55	0.66	0.53	89041
Σταθμισμένος μέσος	0.90	0.71	0.78	89041

Το ποσοστό λάθος ταξινομημένων στοιχείων της πρώτης τάξης φτάνει το 28%, επιβεβαιώνοντας την καλύτερη απόδοση του μοντέλου σε αυτήν την τάξη. Ο σταθμισμένος μέσος της Ακρίβειας και της Ανάκλησης είναι επίσης αξιόλογος στο 0.90% και στο 71%, αντίστοιχα.



Γράφημα 5.37: *Καμπύλη ROC αλγορίθμου Random Forests για τις απότομες επιταχύνσεις ανά 100χλμ.*

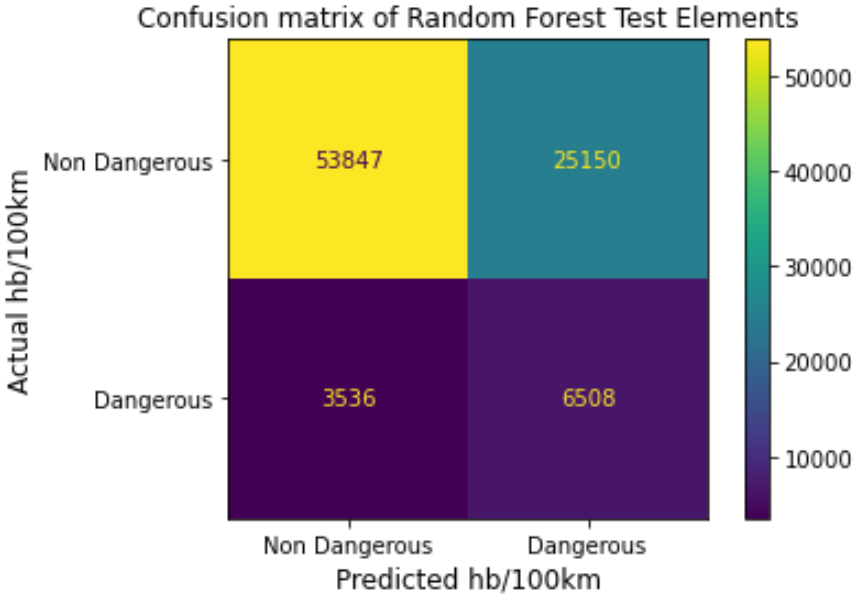


Γράφημα 5.38: *Καμπύλη Ακρίβειας-Ανάκλησης αλγορίθμου Random Forests για τις απότομες επιταχύνσεις ανά 100χλμ.*

Το AUC score της καμπύλης ROC είναι ίσο με 73,98% που δίνει επίσης καλή προβλεπτική εντύπωση. Η Καμπύλη Ακρίβειας-Ανάκλησης, όπως αναπαρίσταται στο Γράφημα 5.38, κυμαίνεται σε χαμηλότερα επίπεδα απόδοσης συγκριτικά με άλλα μοντέλα.

5.3.4.2 Απότομες επιβραδύνσεις ανά 100χλμ.

Η μέθοδος ρύθμισης και εύρεσης των βέλτιστων υπερπαραμέτρων έδωσε: max_depth=16 και n_estimators=128. Η Ορθότητα ήταν σχετικώς υψηλή, της τάξης του 82% και ο γεωμετρικός μέσος των τάξεων G-mean ήταν 0.666.



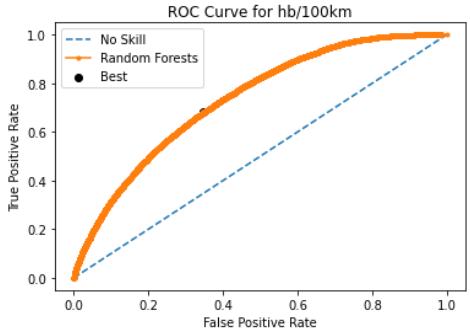
Γράφημα 5.39: *Μήτρα σύγχυσης δεδομένων ελέγχου αλγορίθμου Random Forests για τις απότομες επιβραδύνσεις ανά 100χλμ.*

Η μήτρα σύγχυσης του μοντέλου αποτελείται από αρκετά μειωμένα λάθη ταξινόμησης στην Θετική κλάση, όμως το ποσοστό FNR είναι ιδιαίτερα υψηλό, παρουσιάζοντας ανορθόδοξη προσέγγιση.

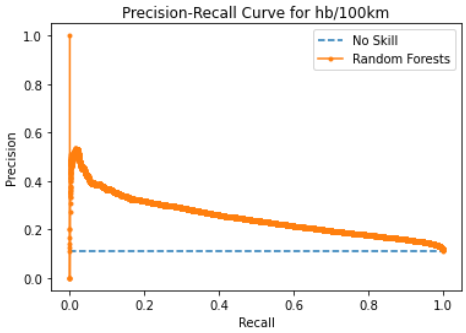
Πίνακας 5.8: *Επίδοση αλγορίθμου Random Forests για τις απότομες επιβραδύνσεις ανά 100χλμ.*

Οδική Συμπεριφορά	**Ακρίβεια**	**Ανάκληση**	**f1-score**	**Σύνολο δεδομένων εξέτασης**
Μη Επικίνδυνη	0.94	0.68	0.79	78997
Επικίνδυνη	0.21	0.65	0.31	10044
Μέσος όρος	0.57	0.66	0.55	89041
Σταθμισμένος μέσος	0.86	0.68	0.74	89041

Η Ακρίβεια της Επικίνδυνης τάξης ξεπερνά τις συγκρίσεις με τα άλλα μοντέλα και ο σταθμισμένος μέσος όρος είναι ικανοποιητικός. Η Ανάκληση παρότι ως μέσος όρος κυμαίνεται σε καλό επίπεδο, καταδεικνύει την μη αξιοπιστία του μοντέλου με μεγάλα ποσοστά λανθασμένα ταξινομημένα στοιχείων.



Γράφημα 5.40: *Καμπύλη ROC αλγορίθμου Random Forests για τις απότομες επιβραδύνσεις ανά 100χλμ.*



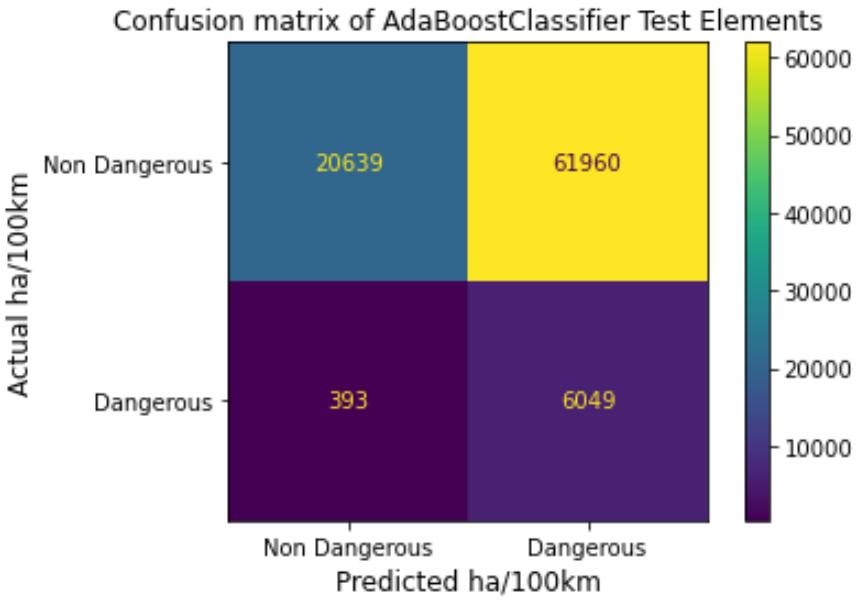
Γράφημα 5.41: *Καμπύλη Ακρίβειας-Ανάκλησης αλγορίθμου Random Forests για τις απότομες επιβραδύνσεις ανά 100χλμ.*

Το AUC score της καμπύλης ROC είναι ίσο με 73,62%, χαμηλότερο δηλαδή σε συγκρίσεις με άλλους αλγορίθμους, ενώ η καμπύλη Ακρίβειας-Ανάκλησης επιβεβαιώνει ότι κατά βάση το μοντέλο προβλέπει καλά, όμως με αρκετά κρίσιμα λάθη ταξινόμησης.

5.3.5 AdaBoost Classification

5.3.5.1 Απότομες επιταχύνσεις ανά 100χλμ.

Η μέθοδος GridSearch έδωσε ως βέλτιστη υπερπαράμετρο: n_estimators=1. Η Ορθότητα του αλγορίθμου ήταν 60% και το G-mean ίσο με 0.484, αποτελέσματα που κρίνονται ανεπαρκή.



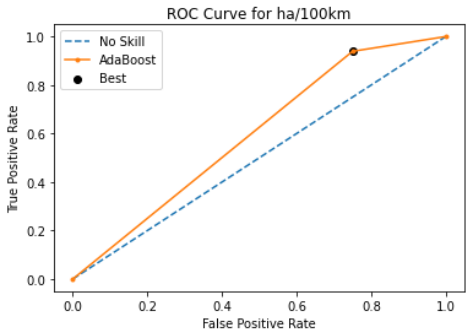
Γράφημα 5.42: *Μήτρα σύγχυσης δεδομένων ελέγχου αλγορίθμου AdaBoost για τις απότομες επιταχύνσεις ανά 100χλμ.*

Η μήτρα σύγχυσης του μοντέλου αποτελείται από μεγάλα παράδοξα. Ενώ το μοντέλο λειτουργεί στην βάση δένδρου απόφασης, τα αποτελέσματά του καταδεικνύουν μεγάλες διαφορές. Το FNR είναι σε σχεδόν ιδανικά επίπεδα, όμως το ποσοστό FPR είναι απογοητευτικό, δείχνοντας ότι το μοντέλο δεν είναι καθόλου αξιόπιστο.

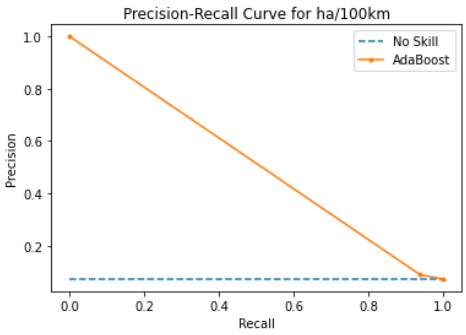
Πίνακας 5.9: *Επίδοση αλγορίθμου AdaBoost για τις απότομες επιταχύνσεις ανά 100χλμ.*

Οδική Συμπεριφορά	**Ακρίβεια**	**Ανάκληση**	**f1-score**	**Σύνολο δεδομένων εξέτασης**
Μη Επικίνδυνη	0.98	0.25	0.40	82599
Επικίνδυνη	0.09	0.94	0.16	6442
Μέσος όρος	0.54	0.59	0.28	89041
Σταθμισμένος μέσος	0.92	0.30	0.38	89041

Ο Πίνακας 5.9 επιβεβαιώνει την έλλειψη αξιοπιστίας του μοντέλου. Η Ακρίβεια της ταξινόμησης Μη Επικίνδυνης συμπεριφοράς είναι πολύ υψηλή, όμως της δεύτερης τάξης είναι μη αποδεκτή και όπως αποδείχτηκε, αντίστροφα κυμαίνεται η Ανάκληση.



Γράφημα 5.43: *Καμπύλη ROC αλγορίθμου AdaBoost για τις απότομες επιταχύνσεις ανά 100χλμ.*

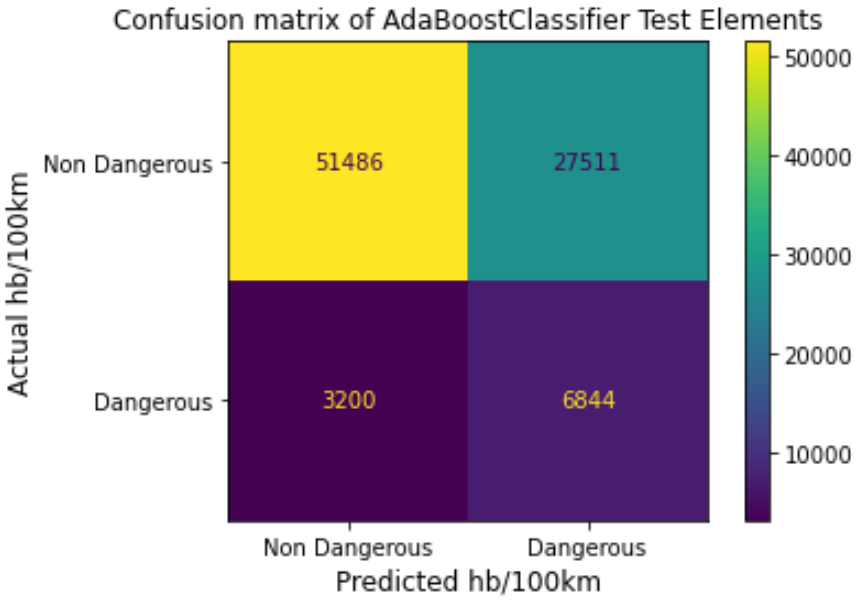


Γράφημα 5.44: *Καμπύλη Ακρίβειας-Ανάκλησης αλγορίθμου AdaBoost για τις απότομες επιταχύνσεις ανά 100χλμ.*

Το AUC score της καμπύλης ROC είναι ίσο με 59,44% , ενώ οι καμπύλες των Γραφημάτων 5.43 και 5.44, επαληθεύουν την κακή ικανότητα ταξινόμησης του μοντέλου.

5.3.5.2 Απότομες επιβραδύνσεις ανά 100χλμ.

Η μέθοδος GridSearch έδωσε ως βέλτιστη υπερπαράμετρο: n_estimators=30. Η Ορθότητα του αλγορίθμου ήταν 67% και το G-mean ίσο με 0.667, συγκριτικά αρκετά καλύτερα από την ταξινόμηση της μεταβλητής απότομων επιταχύνσεων ανά 100χλμ. με το ίδιο μοντέλο.



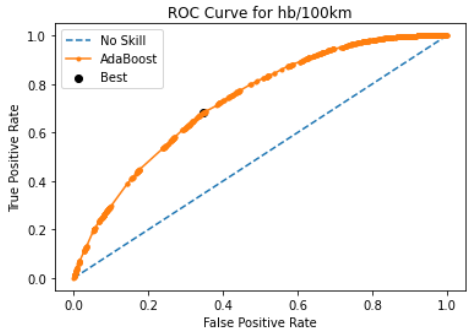
Γράφημα 5.45: *Μήτρα σύγχυσης δεδομένων ελέγχου αλγορίθμου AdaBoost για τις απότομες επιβραδύνσεις ανά 100χλμ.*

Η μήτρα σύγχυσης του μοντέλου για τις απότομες επιβραδύνσεις ανά 100χλμ κυμαίνεται σε αρκετά αξιόλογα επίπεδα συγκριτικά με το αντίστοιχο για τις επιταχύνσεις, όμως είναι αισθητά χαμηλότερης απόδοσης συγκριτικά με άλλα μοντέλα.

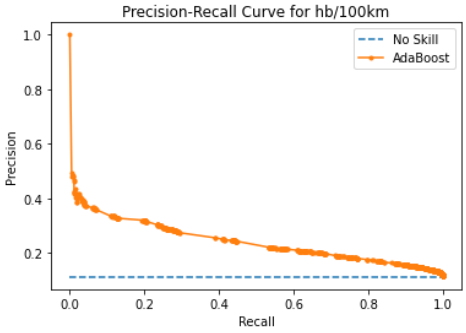
Πίνακας 5.10: *Επίδοση αλγορίθμου AdaBoost για τις απότομες επιβραδύνσεις ανά 100χλμ.*

Οδική Συμπεριφορά	**Ακρίβεια**	**Ανάκληση**	**f1-score**	**Σύνολο δεδομένων εξέτασης**
Μη Επικίνδυνη	0.94	0.65	0.77	78997
Επικίνδυνη	0.20	0.68	0.31	10044
Μέσος όρος	0.57	0.67	0.54	89041
Σταθμισμένος μέσος	0.86	0.66	0.72	89041

Το ποσοστό λάθος ταξινομημένων στοιχείων στην Επικίνδυνη τάξη είναι 32%, ενώ στην Μη Επικίνδυνη 35%. Η Ακρίβεια των δύο τάξεων κυμαίνεται σε ικανοποιητικά επίπεδα, όμως ενώ επιβεβαιώνεται η γενικότερη αξιόλογη απόδοση του AdaBoost για τα συγκεκριμένα δεδομένα εκπαίδευσης, είναι χειρότερο μοντέλο από τα προγενέστερα.



Γράφημα 5.46: *Καμπύλη ROC αλγορίθμου AdaBoost για τις απότομες επιβραδύνσεις ανά 100χλμ.*



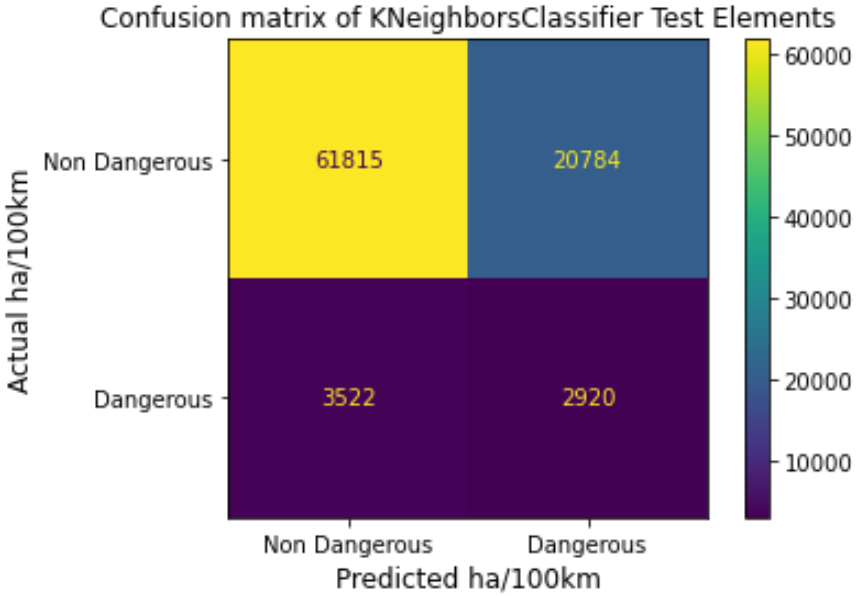
Γράφημα 5.47: *Καμπύλη Ακρίβειας-Ανάκλησης αλγορίθμου AdaBoost για τις απότομες επιβραδύνσεις ανά 100χλμ.*

Το AUC score της καμπύλης ROC είναι ίσο με 73,04% , δηλαδή δίνει καλή σχετικά εντύπωση ταξινόμησης των στοιχείων και η Καμπύλη του Γραφήματος 5.47 καθιστά το μοντέλο καλό για πρόβλεψη, χειρότερο όμως από υπόλοιπα.

5.3.6 K-nearest Neighbors Classification

5.3.6.1 Απότομες επιταχύνσεις ανά 100χλμ.

Η μέθοδος GridSearch έδωσε ως βέλτιστη υπερπαράμετρο: n_neighbors=7. Η Ορθότητα του αλγορίθμου ήταν 87% και το G-mean ίσο με 0.620.



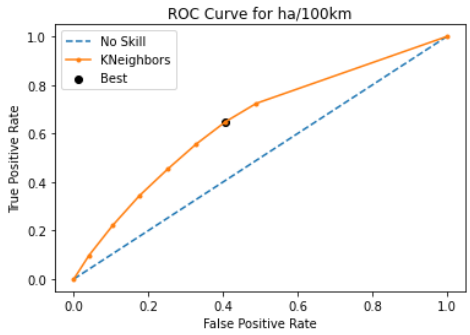
Γράφημα 5.48: *Μήτρα σύγχυσης δεδομένων ελέγχου αλγορίθμου KNN για τις απότομες επιταχύνσεις ανά 100χλμ.*

Η μήτρα σύγχυσης του μοντέλου αποτελείται από μικρό μέγεθος FPR, όμως αρκετά υψηλό FPR, καθιστώντας το μοντέλο αναξιόπιστο.

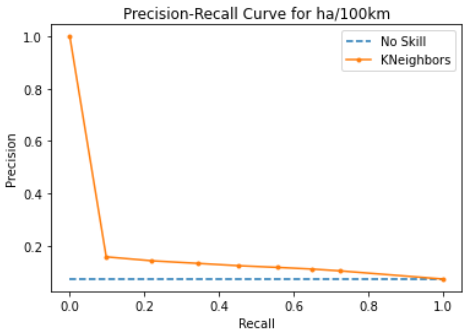
Πίνακας 5.11: *Επίδοση αλγορίθμου KNN για τις απότομες επιταχύνσεις ανά 100χλμ.*

Οδική Συμπεριφορά	**Ακρίβεια**	**Ανάκληση**	**f1-score**	**Σύνολο δεδομένων εξέτασης**
Μη Επικίνδυνη	0.95	0.75	0.84	82599
Επικίνδυνη	0.12	0.45	0.19	6442
Μέσος όρος	0.53	0.60	0.51	89041
Σταθμισμένος μέσος	0.89	0.73	0.79	89041

Το ποσοστό λανθασμένα ταξινομημένων στοιχείων της Μη Επικίνδυνης τάξης είναι 25%, γεγονός που το καθιστά αρκετά ικανοποιητικό. Εκ διαμέτρου αντίθετα είναι όμως τα αποτελέσματα για την Ανάκληση της Επικίνδυνης τάξης με 55% να είναι το ποσοστό των λανθασμένα ταξινομημένων στοιχείων. Ο σταθμισμένος μέσος όρος Ακρίβειας κινείται σε καλά επίπεδα.



Γράφημα 5.49: *Καμπύλη ROC αλγορίθμου KNN για τις απότομες επιταχύνσεις ανά 100χλμ.*

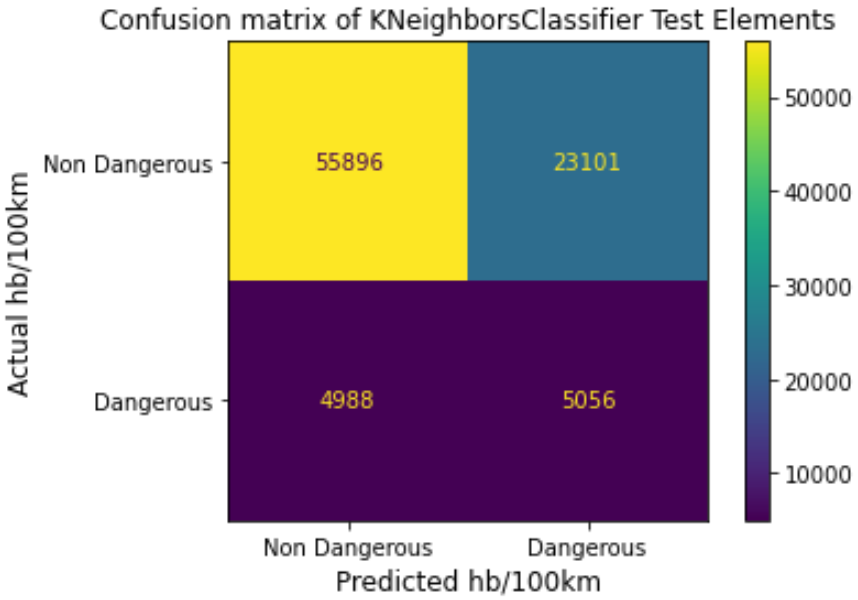


Γράφημα 5.50: *Καμπύλη Ακρίβειας-Ανάκλησης αλγορίθμου KNN για τις απότομες επιταχύνσεις ανά 100χλμ.*

Το AUC score της καμπύλης ROC είναι ίσο με 64,55%, επαληθεύοντας την μέτρια απόδοση του μοντέλου. Στην γραφική αναπαράσταση της Καμπύλης Ακρίβειας-Ανάκλησης επαληθεύεται εκ νέου αυτός ο ισχυρισμός. Γενικά, το μοντέλο δεν προτείνεται.

5.3.6.2 Απότομες επιβραδύνσεις ανά 100χλμ.

Η μέθοδος GridSearch έδωσε ως βέλτιστη υπερπαράμετρο: n_neighbors=7. Η Ορθότητα του αλγορίθμου ήταν 85% και το G-mean ίσο με 0.617.



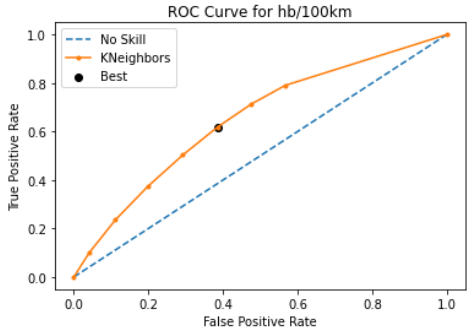
Γράφημα 5.51: *Μήτρα σύγχυσης δεδομένων ελέγχου αλγορίθμου KNN για τις απότομες επιβραδύνσεις ανά 100χλμ.*

Η μήτρα σύγχυσης του αλγορίθμου KNN δίνει εκ νέου καλά αποτελέσματα FPR, αρκετά όμως στοιχεία ταξινομημένα στα Ψευδώς Αρνητικά, δίνοντας την εντύπωση ότι ως μοντέλο είναι ανεπαρκές συγκριτικά με άλλα μοντέλα.

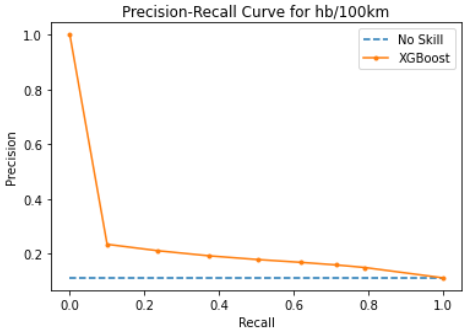
Πίνακας 5.12: *Επίδοση αλγορίθμου KNN για τις απότομες επιβραδύνσεις ανά 100χλμ.*

Οδική Συμπεριφορά	**Ακρίβεια**	**Ανάκληση**	**f1-score**	**Σύνολο δεδομένων εξέτασης**
Μη Επικίνδυνη	0.92	0.71	0.80	78997
Επικίνδυνη	0.18	0.50	0.26	10044
Μέσος όρος	0.55	0.61	0.53	89041
Σταθμισμένος μέσος	0.83	0.68	0.74	89041

Σύμφωνα με τον Πίνακα 5.11, η Ανάκληση της πρώτης τάξης κινείται σε ικανοποιητικά επίπεδα, σε αντίθεση με της τάξης Επικίνδυνης συμπεριφοράς και φυσικά επιβεβαιώνει την μήτρα σύγχυσης του μοντέλου. Οι μετρικές του μοντέλου καθιστούν το μοντέλο αναξιόπιστο.



Γράφημα 5.52: *Καμπύλη ROC αλγορίθμου KNN για τις απότομες επιβραδύνσεις ανά 100χλμ.*



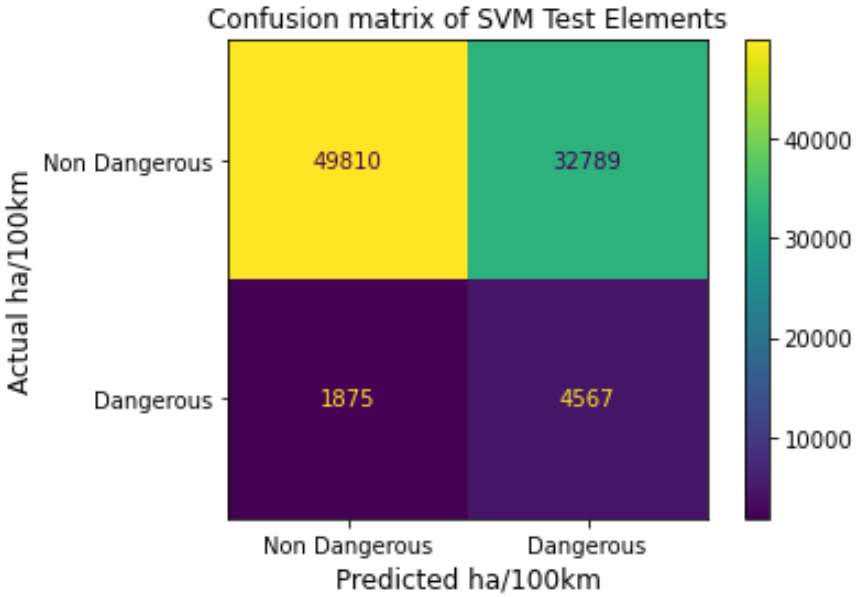
Γράφημα 5.53: *Καμπύλη Ακρίβειας-Ανάκλησης αλγορίθμου KNN για τις απότομες επιβραδύνσεις ανά 100χλμ.*

Το AUC score της καμπύλης ROC είναι ίσο με 65% , ενώ οι γραφικές απεικονίσεις των Καμπυλών δεν είναι ικανοποιητικές.

5.3.7 Supported Vector Machines

5.3.7.1 Απότομες επιταχύνσεις ανά 100χλμ.

Ο εντοπισμός των βέλτιστων υπερπαραμέτρων ανέδειξε τις εξής: kernel=rbf, gamma=0.1, C=100. Η Ορθότητα της εκπαίδευσης έφτασε το 66%, ενώ το G-mean ήταν 0.654.



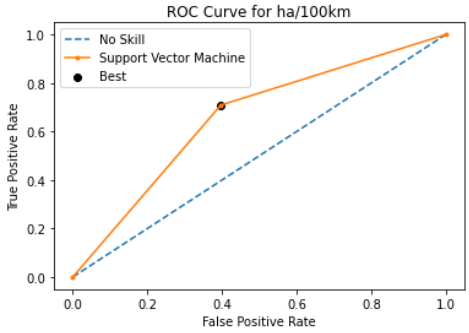
Γράφημα 5.54: *Μήτρα σύγχυσης δεδομένων ελέγχου αλγορίθμου SVM για τις απότομες επιταχύνσεις ανά 100χλμ.*

Η μήτρα σύγχυσης των SVM είναι πολύ καλή στο επίπεδο FNR, με χαμηλές όμως επιδόσεις στο FPR. Φαίνεται να προβλέπει καλά την Επικίνδυνη τάξη.

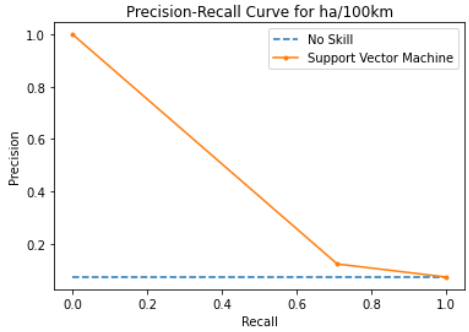
Πίνακας 5.13: *Επίδοση αλγορίθμου SVM για τις απότομες επιταχύνσεις ανά 100χλμ.*

Οδική Συμπεριφορά	**Ακρίβεια**	**Ανάκληση**	**f1-score**	**Σύνολο δεδομένων εξέτασης**
Μη Επικίνδυνη	0.96	0.60	0.74	82599
Επικίνδυνη	0.12	0.71	0.21	6442
Μέσος όρος	0.54	0.66	0.48	89041
Σταθμισμένος μέσος	0.90	0.61	0.70	89041

Η Ακρίβεια της πρώτης τάξης είναι καλή, όμως το μοντέλο υστερεί σε ακρίβεια της δεύτερης Επικίνδυνης τάξης. Το ποσοστό των λανθασμένα ταξινομημένων δεδομένων της Επικίνδυνης τάξης είναι 29%, που φαίνεται να προβλέπει σε καλό βαθμό την Επικίνδυνη συμπεριφορά. Αντιθέτως, η Ανάκληση της πρώτης τάξης δεν είναι καλή, δηλαδή το μοντέλο αποδίδει μέτρια.



Γράφημα 5.55: *Καμπύλη ROC αλγορίθμου SVM για τις απότομες επιταχύνσεις ανά 100χλμ.*

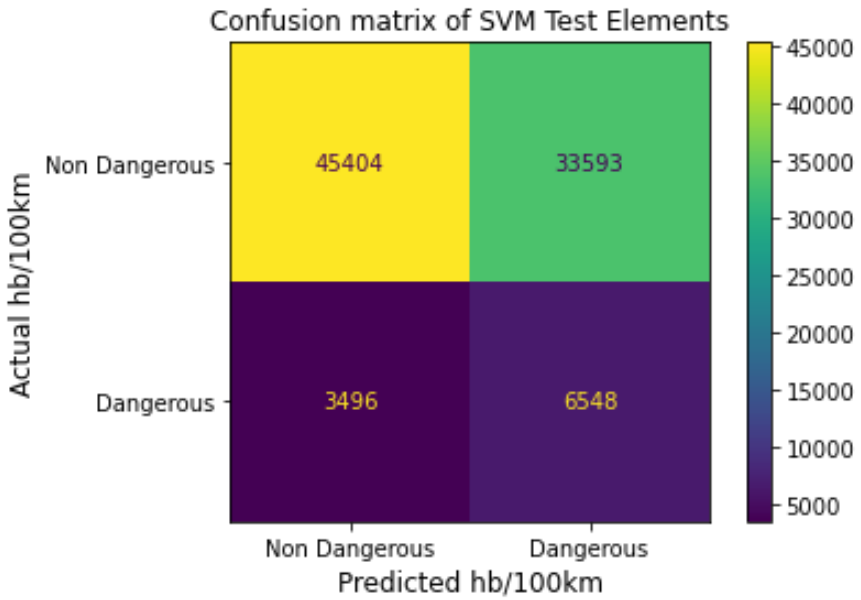


Γράφημα 5.56: *Καμπύλη Ακρίβειας-Ανάκλησης αλγορίθμου SVM για τις απότομες επιταχύνσεις ανά 100χλμ.*

Το AUC score της καμπύλης ROC είναι ίσο με 65,6%, επιβεβαιώνοντας την μέτρια απόδοση των SVM. Εντούτοις, το μοντέλο μπορεί να χρησιμοποιηθεί, όμως υπάρχουν σαφώς καλύτερα εκπαιδευμένοι αλγόριθμοι.

5.3.7.2 Απότομες επιταχύνσεις ανά 100χλμ.

Η τεχνική GridSearch έδωσε ως βέλτιστες υπερπαραμέτρους τις εξής: kernel=rbf, gamma=0.01, C=10. Η Ορθότητα ήταν 61% και το G-mean ίσο με 0.612. Η Ορθότητα εκπαίδευσης του αλγορίθμου δεν είναι καλή.



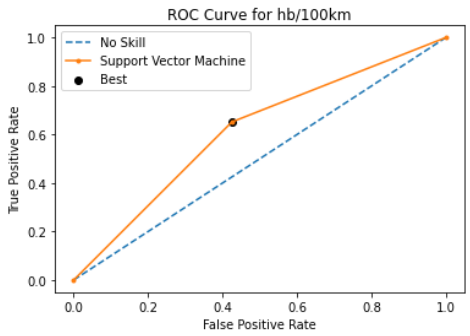
Γράφημα 5.57: *Μήτρα σύγχυσης δεδομένων ελέγχου αλγορίθμου SVM για τις απότομες επιβραδύνσεις ανά 100χλμ.*

Η μήτρα σύγχυσης δεν παρουσιάζει καλά αποτελέσματα, συγκριτικά με τα αντίστοιχα των απότομων επιταχύνσεων με SVM. Παρατηρούνται πολλά λάθη ταξινόμησης με πολύ ανεβασμένα επίπεδα FPR και FNR.

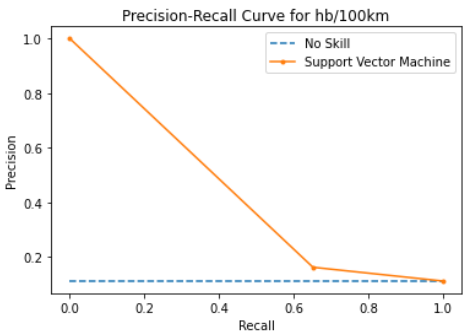
Πίνακας 5.14: *Επίδοση αλγορίθμου SVM για τις απότομες επιβραδύνσεις ανά 100χλμ.*

Οδική Συμπεριφορά	**Ακρίβεια**	**Ανάκληση**	**f1-score**	**Σύνολο δεδομένων εξέτασης**
Μη Επικίνδυνη	0.93	0.57	0.71	78997
Επικίνδυνη	0.16	0.65	0.26	10044
Μέσος όρος	0.55	0.61	0.49	89041
Σταθμισμένος μέσος	0.84	0.58	0.66	89041

Ενώ η Ακρίβεια στην ταξινόμηση της Μη Επικίνδυνης τάξης είναι ικανοποιητική, το ποσοστό Ανάκλησης για κάθε τάξη και στο σύνολο δεν είναι επαρκές για να θεωρηθεί η ταξινόμηση αξιόλογη. Γενικά, το μοντέλο δεν παρουσιάζει καλή εκπαίδευση.



Γράφημα 5.58: *Καμπύλη ROC αλγορίθμου SVM για τις απότομες επιβραδύνσεις ανά 100χλμ.*



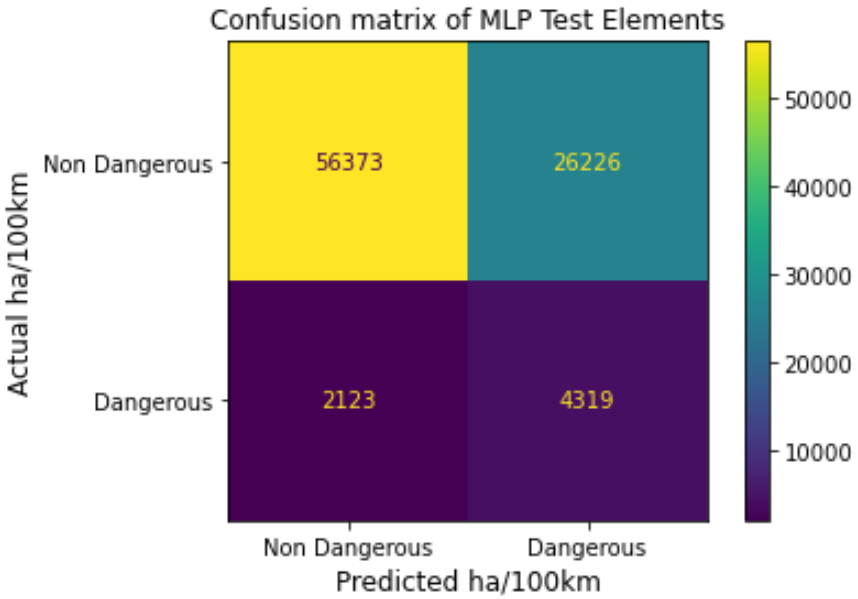
Γράφημα 5.59: *Καμπύλη Ακρίβειας-Ανάκλησης αλγορίθμου SVM για τις απότομες επιβραδύνσεις ανά 100χλμ.*

Το AUC score της καμπύλης ROC είναι ίσο με 61,33%, καθιστώντας αδύναμη την συγκεκριμένη μετρική αξιολόγηση. Η Καμπύλη στο Γράφημα 5.59 αναπαριστά και τις χαμηλές επιδόσεις εκπαίδευσης των SVM.

5.3.8 Multilayered Perceptrons

5.3.8.1 Απότομες επιταχύνσεις ανά 100χλμ.

Οι καλύτεροι υπερπαράμετροι για την ταξινόμηση της συγκεκριμένης μεταβλητής κρίθηκαν από την τεχνική GridSearch και είναι οι εξής: activation=tanh, hidden_layer_sizes=(10,30,10), learning_rate=adaptive, solver=sgd, alpha=0.0001. Η Ορθότητα εκπαίδευσης ήταν 68% και το G-mean ίσο με 0.678.



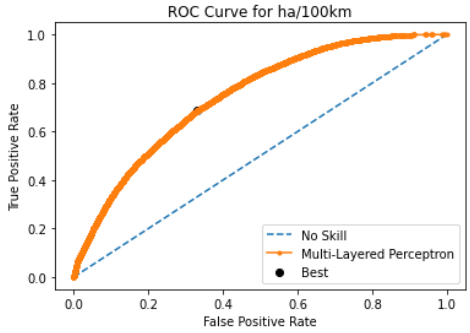
Γράφημα 5.60: *Μήτρα σύγχυσης δεδομένων ελέγχου αλγορίθμου MLP για τις απότομες επιταχύνσεις ανά 100χλμ.*

Η μήτρα σύγχυσης των MLP δίνει αρκετά αξιόλογα αποτελέσματα με αισθητά μειωμένα επίπεδα FPR και FNR, προτείνοντας την χρήση του μοντέλου.

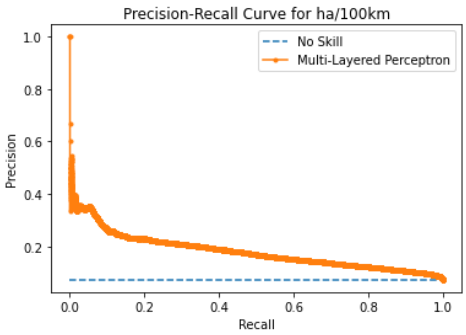
Πίνακας 5.15: *Επίδοση αλγορίθμου MLP για τις απότομες επιταχύνσεις ανά 100χλμ.*

Οδική Συμπεριφορά	**Ακρίβεια**	**Ανάκληση**	**f1-score**	**Σύνολο δεδομένων εξέτασης**
Μη Επικίνδυνη	0.96	0.68	0.80	82599
Επικίνδυνη	0.14	0.67	0.23	6442
Μέσος όρος	0.55	0.68	0.52	89041
Σταθμισμένος μέσος	0.90	0.68	0.76	89041

Ο Πίνακας 5.13 παρουσιάζει τις μετρικές αξιολογήσεις των MLP που στο σύνολό τους δίνουν αρκετά καλή εκπαίδευση του μοντέλου. Τα επίπεδα Ακρίβειας είναι ισχυρά, ενώ και η Ανάκληση παρότι δεν κυμαίνεται σε καταπληκτικά επίπεδα, είναι επαρκής. Το μοντέλο φαίνεται αξιόπιστο.



Γράφημα 5.61: *Καμπύλη ROC αλγορίθμου MLP για τις απότομες επιταχύνσεις ανά 100χλμ.*

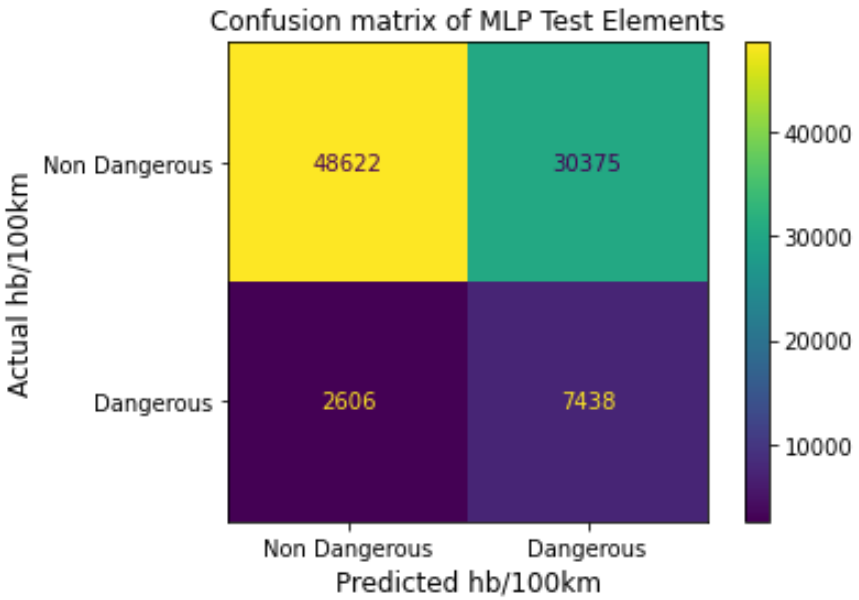


Γράφημα 5.62: *Καμπύλη Ακρίβειας-Ανάκλησης αλγορίθμου MLP για τις απότομες επιταχύνσεις ανά 100χλμ.*

Το AUC score της καμπύλης ROC είναι ίσο με 74,67%, επαληθεύοντας την καλή ικανότητα ταξινόμησης και πρόβλεψης των MLPs, ενώ και η Καμπύλη του Γραφήματος 5.62 έχει καλή συμπεριφορά. Το μοντέλο ικανοποιεί.

5.3.8.2 Απότομες επιβραδύνσεις ανά 100χλμ.

Η μέθοδος ρύθμισης υπερπαραμέτρων ανέδειξε τα εξής: activation=relu, hidden_layer_sizes=(10,30,10), learning_rate=constant, solver=sgd και alpha=0.0001. Η Ορθότητα του μοντέλου ήταν 68% και το G-mean ήταν 0.677.



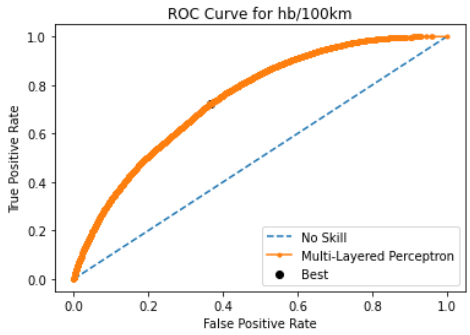
Γράφημα 5.63: *Μήτρα σύγχυσης δεδομένων ελέγχου αλγορίθμου MLP για τις απότομες επιβραδύνσεις ανά 100χλμ.*

Η μήτρα σύγχυσης δίνει χειρότερα αποτελέσματα από τα αντίστοιχα των επιταχύνσεων με MLPs, όμως σε γενικές γραμμές το μοντέλο έχει καλή συμπεριφορά.

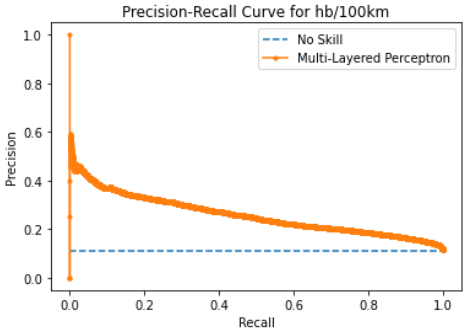
Πίνακας 5.16: *Επίδοση αλγορίθμου MLP για τις απότομες επιβραδύνσεις ανά 100χλμ.*

Οδική Συμπεριφορά	**Ακρίβεια**	**Ανάκληση**	**f1-score**	**Σύνολο δεδομένων εξέτασης**
Μη Επικίνδυνη	0.95	0.62	0.75	78997
Επικίνδυνη	0.20	0.74	0.31	10044
Μέσος όρος	0.57	0.68	0.53	89041
Σταθμισμένος μέσος	0.86	0.63	0.70	89041

Η Ακρίβεια ταξινόμησης στις δύο τάξεις κινείται σε πολύ καλά επίπεδα, ενώ και τα ποσοστά Ανάκλησης επιβεβαιώνουν την καλή εκπαίδευση του μοντέλου, καθιστώντας το χρήσιμο.



Γράφημα 5.64: *Καμπύλη ROC αλγορίθμου MLP για τις απότομες επιβραδύνσεις ανά 100χλμ.*



Γράφημα 5.65: *Καμπύλη Ακρίβειας-Ανάκλησης αλγορίθμου MLP για τις απότομες επιβραδύνσεις ανά 100χλμ.*

Το AUC score της καμπύλης ROC των MLP είναι ίσο με 74,69%, και η Καμπύλη Ανάκλησης-Ακρίβειας κυμαίνεται σε αρκετά ικανοποιητικά επίπεδα. Γενικότερα, το μοντέλο μπορεί να αξιοποιηθεί.

5.4 Σύγκριση μοντέλων ταξινόμησης

Αφού πραγματοποιήθηκε η διαδικασία της ταξινόμησης και προέκυψαν οι μετρικές αξιολογήσεις κάθε αλγορίθμου, προσδιορίζεται η βέλτιστη προβλεπτική ικανότητα μοντέλου, σύμφωνα με την εκπαίδευσή του στα υπάρχοντα δεδομένα. Οι τεχνικές επεξεργασίας και μεταχείρισης των δεδομένων, καθώς και οι ρυθμίσεις βελτιστοποίησης υπερπαραμέτρων που υπέστη κάθε διαφορετικός αλγόριθμος στόχευσαν στην βελτίωση των προγνωστικών μοντέλων και της απόδοση της ταξινόμησης. Προκειμένου να αξιολογηθούν σωστά τα μοντέλα, επισημαίνεται ότι η λανθασμένη ταξινόμηση στοιχείων που στην πραγματικότητα ανήκουν στην τάξη Επικίνδυνης Οδικής συμπεριφοράς αποτελεί την κρισιμότερη αξιολόγηση, λόγω της δυνητικής επικινδυνότητας που παρουσιάζει το συγκεκριμένο λάθος. Συνεπώς, η Ανάκληση (Recall), η Περιοχή κάτω από την Καμπύλη (AUC score), καθώς και η αναλογία Ψευδώς Αρνητικών αποτελούν της σημαντικότερες παραμέτρους αξιολόγησης του κάθε μοντέλου. Στον Πίνακα 5.17 και στον Πίνακα 5.18 που ακολουθούν παρουσιάζονται συγκεντρωτικά τα αποτελέσματα της ταξινόμησης των απότομων περιστατικών με τις κρισιμότερες για την σύγκρισή τους μετρικές αξιολογήσεις τους.

Πίνακας 5.17: *Συγκεντρωτικός Πίνακας αποτελεσμάτων μοντέλων ταξινόμησης για τις απότομες επιταχύνσεις ανά 100χλμ..*

Αλγόριθμος ταξινόμησης	**Ορθότητα**	**Ακρίβεια**	**Ανάκληση**	**FNR**	**f-1 score**	**AUC score**
Decision Trees	53.03%	54.12%	65.35%	34.65%	43.26%	70.48%
GradientBoosting	65.28%	55.15%	68.05%	31.95%	50.25%	75.10%
XGBoost	66.76%	55.09%	67.46%	32.54%	50.86%	74.26%
Random Forests	70.83%	55.16%	66.39%	33.61%	52.64%	73.98%
AdaBoost	29.97%	53.51%	59.44%	40.56%	28.04%	59.44%
KNeighbors	72.70%	53.46%	60.08%	39.92%	51.47%	64.55%
SVM	61.07%	54.30%	65.60%	34.40%	47.52%	65.60%
MLP	68.16%	55.26%	67.65%	32.35%	51.63%	74.67%

Πίνακας 5.18: *Συγκεντρωτικός Πίνακας αποτελεσμάτων μοντέλων ταξινόμησης για τις απότομες επιβραδύνσεις ανά 100χλμ..*

Αλγόριθμος ταξινόμησης	**Ορθότητα**	**Ακρίβεια**	**Ανάκληση**	**FNR**	**f-1 score**	**AUC score**
Decision Trees	62.51%	56.58%	66.03%	33.97%	52.12%	72.35%
GradientBoosting	63.36%	57.36%	67.91%	32.09%	53.13%	74.88%
XGBoost	64.53%	57.20%	67.30%	32.70%	53.60%	74.28%
Random Forests	67.78%	57.20%	66.48%	33.52%	55.09%	73.62%
AdaBoost	65.51%	57.03%	66.66%	33.34%	53.93%	73.04%
KNeighbors	68.45%	54.88%	60.55%	39.45%	53.19%	65.00%
SVM	58.35%	54.58%	61.33%	38.67%	48.55%	61.33%
MLP	62.96%	57.29%	67.80%	32.20%	52.88%	74.69%

Σύμφωνα με τον Πίνακα 5.17, το σημαντικότερο ποσοστό Ορθότητας αναδεικνύεται από την ταξινόμηση με τον αλγόριθμο K-nearest Neighbors και η χαμηλότερη με το AdaBoost. Τα ποσοστά Ορθότητας των μοντέλων είναι στο σύνολό τους ικανοποιητικά. Η υψηλότερη Ανάκληση παρατηρείται στην ανάπτυξη του αλγορίθμου Gradient Boosting σε συνολικό ποσοστό 68.05% και σε συνδυασμό με την Ακρίβεια του μοντέλου συνεπάγονται υψηλή ικανότητα αναγνώρισης και πρόβλεψης της πραγματικά Επικίνδυνης κλάσης. Επισημαίνεται, ότι ο αλγόριθμος Gradient Boosting παρουσιάζει και τα χαμηλότερα ποσοστά Ψευδώς Αρνητικών , καθώς και το υψηλότερο σκορ AUC. Σε ικανότητα ταξινόμησης και πρόβλεψης, με παρεμφερείς, όμως ελάχιστα χαμηλότερες μετρικές αξιολογήσεις ακολουθεί ο αλγόριθμος Multilayered Perceptrons. Η Ανάκληση του μοντέλου είναι αρκετά ικανοποιητική, όπως επίσης και το ποσοστό Ψευδώς Αρνητικών και το σκορ AUC, μετρικές που καθιστούν το μοντέλο αποδοτικό. Αντίθετα, η ταξινόμηση με το AdaBoost αποδεικνύεται η πιο επισφαλής, με μετρικές που καταδεικνύουν την ακαταλληλότητα του συγκεκριμένου αλγορίθμου.

Σύμφωνα με τον Πίνακα 5.18, παρατηρείται σχετικώς κοινή απόδοση συγκριτικά με τους αλγορίθμους που ταξινόμησαν τις απότομες επιταχύνσεις ανά 100 χλμ. Η Ορθότητα των μοντέλων, καθώς και η Ανάκληση στην συνολική εικόνα των μοντέλων παρουσιάζουν αποδοτική συμπεριφορά. Επιπλέον, η αξιολόγηση του σκορ AUC χαρακτηρίζει εν γένει τα μοντέλα σχετικώς ασφαλή. Ισχυρότερο μοντέλο φαίνεται το Gradient Boosting και ακολουθούν τα Multilayered Perceptrons, με τα υψηλότερα ποσοστά Ανάκλησης και τα χαμηλότερα Ψευδώς Αρνητικών. Αντίθετα, μη ικανοποιητική απόδοση φαίνεται να έχουν οι ταξινομήσεις με τον αλγόριθμο K-nearest Neighbors και Supported Vector Machines, με μετρικές αξιολογήσεις που τους καθιστούν ανεπαρκείς για ασφαλή συμπεράσματα.

6. Συμπεράσματα

Στο συγκεκριμένο Κεφάλαιο θα πραγματοποιηθεί η ανασκόπηση της παρούσας Διπλωματικής Εργασίας, με σύνοψη των τελικών αποτελεσμάτων, καθώς και συγκρότηση των συμπερασμάτων που προκύπτουν από την ανάλυση με τεχνικές Μηχανικής Μάθησης. Επιπλέον, θα συνταχθούν συγκεκριμένες προτάσεις προς ερευνητική διερεύνηση, που ανέκυψαν από την ενασχόληση με την Εργασία, καθώς και ορισμένες προεκτάσεις της για αξιοποίηση.

6.1 Σύνοψη Αποτελεσμάτων

Στόχος της παρούσας Διπλωματικής Εργασίας είναι η ανίχνευση, ταξινόμηση και πρόβλεψη των απρόοπτων περιστατικών με εφαρμογή Μη Ισορροπημένης Μάθησης. Η ανασκόπηση της Βιβλιογραφίας κατέδειξε την αναγκαιότητα για περαιτέρω αναλύσεις οδικών δεδομένων και ταξινόμηση της Οδικής συμπεριφοράς, μέσα από καινοτόμες προσεγγίσεις. Δείκτες της Επικίνδυνης Οδικής συμπεριφοράς αποτελούν τα απότομα περιστατικά όπως οι απότομες επιταχύνσεις και απότομες επιβραδύνσεις κατά την διάρκεια μιας διαδρομής. Προκειμένου να καταστεί πιο αξιόπιστη η ανάλυση, οι επιλεγείσες εξαρτημένες μεταβλητές ήταν τα παραμετροποιημένα ως προς συγκεκριμένη μονάδα απόστασης απότομα περιστατικά, ήτοι τα απότομα περιστατικά ανά 100χλμ, διαχωρισμένα σε δύο εξαρτημένες μεταβλητές. Τα δεδομένα συλλέχθηκαν από την βάση δεδομένων της εταιρίας OSeven Telematics©.

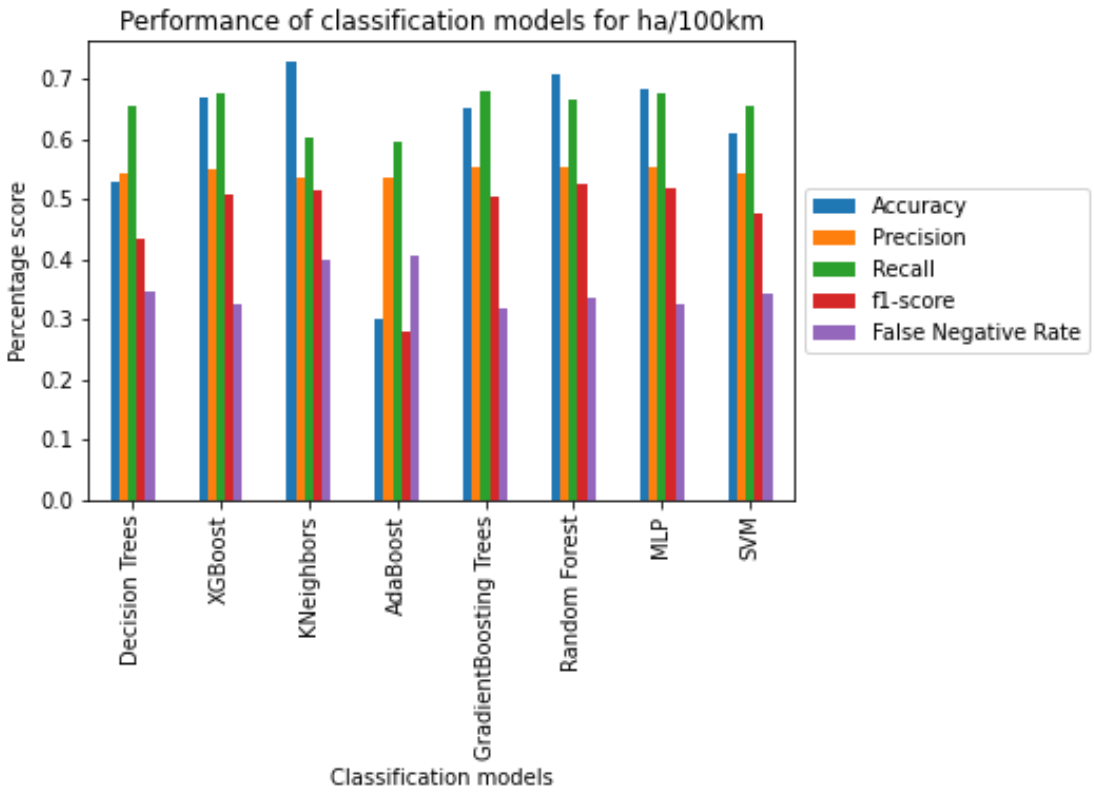
Στο πρώτο μέρος της ανάλυσης, καταρτίστηκαν οι σημαντικότεροι παράγοντες επιρροής των απότομων περιστατικών, μέσω της διαδικασίας Επιλογής Χαρακτηριστικών. Η Επιλογή Χαρακτηριστικών πραγματοποιήθηκε με γνώμονα τον συντελεστή συσχέτισης Pearson των εξαρτημένων μεταβλητών με τις ανεξάρτητες και την διαδικασία της Σημαντικότητας Χαρακτηριστικών, μέσα από επαναληπτικές Παλινδρομήσεις για τον εντοπισμό του βαθμού επιρροής καθεμιάς. Οι παραπάνω διαδικασίες ανέδειξαν ως σημαντικότερες μεταβλητές που επηρεάζουν τα απότομα περιστατικά την συνολική διανυθείσα απόσταση, την συνολική διάρκεια οδήγησης εν κινήσει, την μέση ταχύτητα οδήγησης και τα σκορ υπερβολικής ταχύτητας και χρήσης κινητού τηλεφώνου. Από την τελική επιλογή χαρακτηριστικών αποκλείστηκαν οι μεταβλητές στις οποίες υπεισερχόταν οι παράμετροι των απότομων περιστατικών. Αυτές ήταν οι ονομαστικές απότομες επιταχύνσεις και επιβραδύνσεις, τα αντίστοιχα σκορ τους και το συνολικό σκορ οδήγησης. Οι Παλινδρομήσεις πραγματοποιήθηκαν και για τις δύο εξαρτημένες μεταβλητές με χρήση των αλγορίθμων Γραμμικής Παλινδρόμησης, Decision Trees, Random Forests, XGBoost και Linear SVR. Οι Παλινδρομήσεις με Decision Trees είχαν συντελεστή προσδιορισμού R^2 ίσο με 0.999 μονάδα και οι Παλινδρομήσεις με Random Forests ίσο με 0.871 και 0.869, δηλαδή το προβλεπτικό τους μοντέλο εφάρμοζε πολύ καλά με το δείγμα των παρατηρήσεων. Οι χαμηλές τιμές συντελεστή R^2 των Γραμμικών Παλινδρομήσεων επιβεβαιώνουν την απουσία γραμμικότητας μεταξύ των μεταβλητών.

Στη συνέχεια, αξιοποιώντας τις σημαντικότερες μεταβλητές αναπτύχθηκαν 8 αλγόριθμοι Μηχανικής εκμάθησης, αποσκοπώντας στην ταξινόμηση της Οδικής συμπεριφοράς σε δύο επίπεδα ασφαλείας, αυτό της Επικίνδυνης και αυτό της Μη Επικίνδυνης Οδικής συμπεριφοράς. Προκειμένου να καταστεί εφικτή η διαδικασία της ταξινόμησης, οι εξαρτημένες μεταβλητές υπέστησαν δυαδική ομαδοποίηση, βάσει συγκεκριμένων ορίων τιμών που τέθηκαν με την βοήθεια του αλγορίθμου Κ-μέσου. Εφαρμόζοντας την τεχνική Υπερδειγματοληψίας Συνθετική Μειονοτική (SMOTE), αντιμετωπίστηκε το πρόβλημα άνισης κατανομής των δεδομένων της μειονοτικής

τάξης των δεδομένων εκπαίδευσης. Έγινε εκπαίδευση 16 μοντέλων ταξινόμησης συνολικά με συγκεντρωτικά τους αποτελέσματα να παρουσιάζονται στους Πίνακες 6.1 και 6.2, παράλληλα με την συγκριτική γραφική απεικόνιση των μετρικών αξιολογήσεών τους.

Πίνακας 6.1: *Συγκεντρωτικός Πίνακας αποτελεσμάτων μοντέλων ταξινόμησης για τις απότομες επιταχύνσεις ανά 100χλμ..*

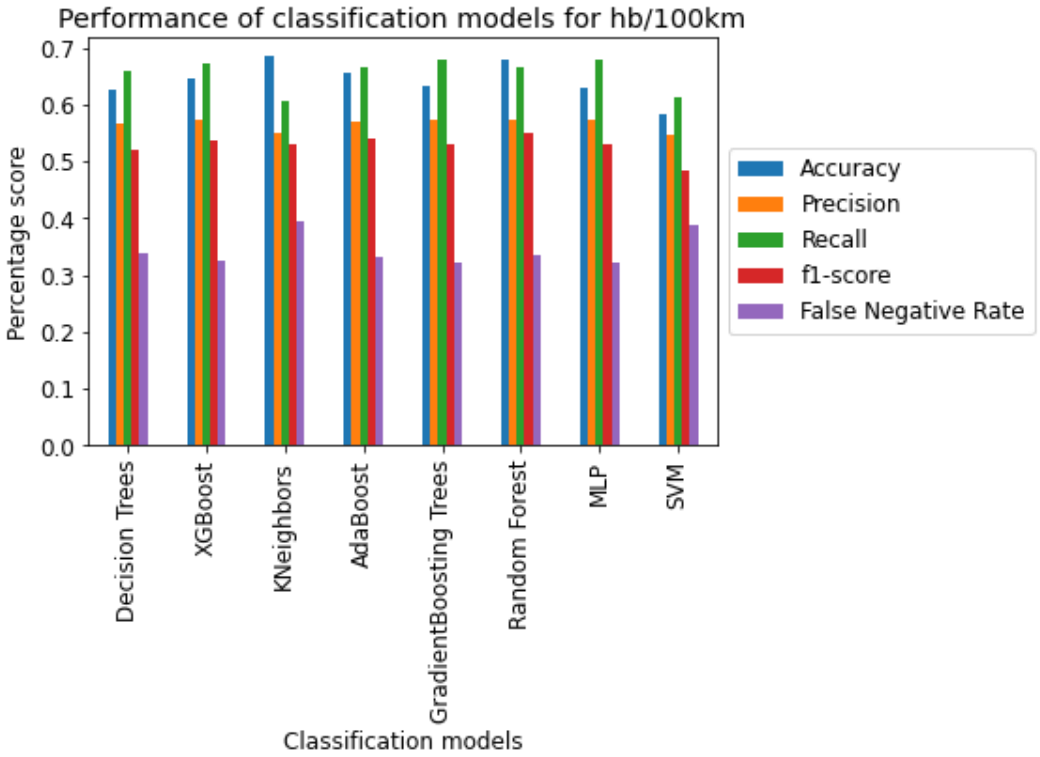
Αλγόριθμος ταξινόμησης	**Ορθότητα**	**Ακρίβεια**	**Ανάκληση**	**FNR**	**f-1 score**	**AUC score**
Decision Trees	53.03%	54.12%	65.35%	34.65%	43.26%	70.48%
GradientBoosting	65.28%	55.15%	68.05%	31.95%	50.25%	75.10%
XGBoost	66.76%	55.09%	67.46%	32.54%	50.86%	74.26%
Random Forests	70.83%	55.16%	66.39%	33.61%	52.64%	73.98%
AdaBoost	29.97%	53.51%	59.44%	40.56%	28.04%	59.44%
KNeighbors	72.70%	53.46%	60.08%	39.92%	51.47%	64.55%
SVM	61.07%	54.30%	65.60%	34.40%	47.52%	65.60%
MLP	68.16%	55.26%	67.65%	32.35%	51.63%	74.67%



Γράφημα 6.1: *Σύγκριση μετρικών αποτελεσμάτων των μοντέλων ταξινόμησης για τις απότομες επιταχύνσεις ανά 100χλμ.*

Πίνακας 6.2: *Συγκεντρωτικός Πίνακας αποτελεσμάτων μοντέλων ταξινόμησης για τις απότομες επιβραδύνσεις ανά 100χλμ..*

Αλγόριθμος ταξινόμησης	**Ορθότητα**	**Ακρίβεια**	**Ανάκληση**	**FNR**	**f-1 score**	**AUC score**
Decision Trees	62.51%	56.58%	66.03%	33.97%	52.12%	72.35%
GradientBoosting	63.36%	57.36%	67.91%	32.09%	53.13%	74.88%
XGBoost	64.53%	57.20%	67.30%	32.70%	53.60%	74.28%
Random Forests	67.78%	57.20%	66.48%	33.52%	55.09%	73.62%
AdaBoost	65.51%	57.03%	66.66%	33.34%	53.93%	73.04%
KNeighbors	68.45%	54.88%	60.55%	39.45%	53.19%	65.00%
SVM	58.35%	54.58%	61.33%	38.67%	48.55%	61.33%
MLP	62.96%	57.29%	67.80%	32.20%	52.88%	74.69%



Γράφημα 6.2: *Σύγκριση μετρικών αποτελεσμάτων των μοντέλων ταξινόμησης για τις απότομες επιβραδύνσεις ανά 100χλμ.*

Το βέλτιστο μοντέλο πρόβλεψης και ταξινόμησης των απότομων περιστατικών αναδείχθηκε ο αλγόριθμος Gradient Boosting, με τον αλγόριθμο Multilayered Perceptrons να αποτελεί και αυτός μια αρκετά αξιόπιστη προσέγγιση.

6.2 Σύνοψη Συμπερασμάτων

Κατά την διάρκεια εκπόνησης της Διπλωματικής Εργασίας και της παρατήρησης των αποτελεσμάτων ανέκυψαν ορισμένα σημαντικά συμπεράσματα για τον τομέα της Οδικής Ασφάλειας και της ανάλυσης της Οδικής Συμπεριφοράς.

1. Σύμφωνα με τα αποτελέσματα των Παλινδρομήσεων, η συνολική διανυθείσα απόσταση αποτελεί την σημαντικότερη μεταβλητή για την αναγνώριση της Οδικής συμπεριφοράς. Η αύξηση της διανυθείσας απόστασης επιδεινώνει χαρακτηριστικά την συμπεριφορά του οδικού στην διαδρομή του, καθώς εμφανίζονται σημάδια κόπωσης, υπνηλίας, αλλοίωσης της αντιληπτικής ικανότητας των εξωτερικών ερεθεισμάτων, επηρεάζει τον χρόνο αντίδρασης και επιβαρύνει τον ψυχοκινητικό συντονισμό.

2. Κατά βάση οι οδικές συμπεριφορές δεν ήταν υπερβολικά επιθετικές, καθώς οι σχετικοί διάμεσοι των απότομων περιστατικών ήταν μηδενικοί και οι οδηγοί εναρμονίζονται με τους κανόνες κυκλοφορίας. Το συγκεκριμένο συμπέρασμα προστίθεται στο παραπάνω, καθώς οι διανυθείσες αποστάσεις και η διάρκεια της

οδήγησης στο σύνολο των παρατηρήσεων θα τις χαρακτήριζαν διαδρομές μικρού έως μεσαίου μεγέθους, δηλαδή με σπάνια εμφάνιση απότομων περιστατικών σε αυτές. Συνεπώς, χαρακτηριστικά απότομα περιστατικά εμφανίζονται σε διαδρομές επιβαρυμένες σε χρόνο ή και απόσταση.

3. Η ταχύτητα του οχήματος αποτελεί σημαντική παράμετρο εμφάνισης απότομων περιστατικών και συντελεί σημαντικά στην εμφάνιση ατυχημάτων και εν γένει Επικίνδυνης Οδικής συμπεριφοράς, επιβεβαιώνοντας την διεθνή βιβλιογραφία. Αυξημένη ταχύτητα ισοδυναμεί με ραγδαία μείωση της αντιληπτικής ικανότητας του οδηγού και την ανάλογη αντίδρασή του σε εξωγενείς παράγοντες.

4. Η οδήγηση κατά την διάρκεια της επικίνδυνης ζώνης ώρας (00:00 – 05:00) δεν συμμεταβάλλεται με ισχυρή σχέση με τα απρόοπτα περιστατικά, συγκριτικά με άλλους παράγοντες. Παρά την γενικότερη εμφάνιση συμπτωμάτων επικίνδυνης οδικής συμπεριφοράς στο συγκεκριμένο χρονικό παράθυρο, οι οδηγοί φαίνεται να μην προβαίνουν σε συχνές απότομες επιταχύνσεις και επιβραδύνσεις, γεγονός που εξηγείται από τους χαμηλούς κυκλοφοριακούς φόρτους και φόρτους πεζών, κατά την νυχτερινή ζώνη ώρας.

5. Για την διαχείριση του προβλήματος της άνισης κατανομής μειονοτικής τάξης των δεδομένων εκπαίδευσης, η τεχνική Υπερδειγματοληψίας SMOTE προκρίθηκε της ADASYN, καθώς οι διακυμάνσεις των δεδομένων ήταν ιδιαίτερα ισχυρές και η αναλογία της τάξης πλειονότητας πολύ μεγάλη. Η SMOTE αποδείχτηκε πιο αποτελεσματική μέθοδος σε καταστάσεις μεγάλων και πολυεπίπεδων δεδομένων, επιβεβαιώνοντας την εγχώρια και διεθνή επιστημονική κοινότητα. Το συγκεκριμένο συμπέρασμα επιβεβαίωσε και η ανάπτυξη των αλγορίθμων ταξινόμησης χωρίς την χρήση τεχνικών Επαναδειγματοληψίας, κατά την οποία παρατηρήθηκε έντονα το φαινόμενο της υπερπροσαρμογής (overfitting).

6. Ο αλγόριθμος Gradient Boosting κατέστη η αποδοτικότερη μέθοδος ταξινόμησης με μεγάλη προβλεπτική ικανότητα για τα εξετασθέντα δεδομένα, ενώ σχεδόν ανάλογες επιδόσεις εμφάνισε και το μοντέλο Multilayered Perceptrons.

7. Οι αλγόριθμοι Gradient Boosting και Multilayered Perceptrons που αναπτύχθηκαν ξεπέρασαν σε επιδόσεις τους υπόλοιπους αλγόριθμους, όμως οι διαφορές δεν ήταν εξέχουσες. Αυτό το φαινόμενο υποστηρίζει την υπόθεση ότι οι εξαρτημένες μεταβλητές των απρόοπτων περιστατικών ήταν καλά ομαδοποιημένες με τη μέθοδο του K-μέσου σε δυαδική κατανομή και επίπεδα ασφαλείας και ότι οι συγκεκριμένοι αλγόριθμοι μπορούν να αναπτυχθούν για αποδοτικές ταξινομήσεις σε επίπεδα ασφαλείας δύο κλάσεων. Επιπλέον, η ποιότητα της ομαδοποίησης με την τεχνική K-μέσου επιβεβαιώνουν την διακριτότητα των κλάσεων και επιβεβαιώνουν την μη αξιοποίηση τεχνικών επαναδειγματοληψίας, όπως η SMOTE-RENN.

8. Οι επιδόσεις των μοντέλων παλινδρόμησης και ταξινόμησης μορφής Δέντρου (CART) ήταν σαφώς ανώτερες από τις αντίστοιχες με αλγορίθμους SVMs συμπεραίνοντας την απουσία γραμμικότητας στα δεδομένα και την αξιοπιστία των συγκεκριμένων μοντέλων.

9. Η υπερπαράμετρος gamma του μοντέλου SVM, η οποία ελέγχει την ελάχιστη μείωση απώλειας που μπορεί να δικαιολογήσει τη δημιουργία διαχωρισμού σε ένα δέντρο, αποδείχτηκε ότι επηρεάζει αισθητά την απόδοση του αλγορίθμου. Επίσης, η παρούσα έρευνα επιβεβαιώνει την επίδοση της μεθόδου πυρήνα Radial Basis Function, συγκριτικά με τις εναλλακτικές μεθόδους πυρήνα, επαληθεύοντας εξίσου την Γκαουσιανή κατανομή των εξετασθέντων στοιχείων.

10. Η μέθοδος ομαδοποίησης K-μέσου εντόπισε ως βέλτιστο threshold για την κατάταξη των απότομων περιστατικών στο Επικίνδυνο επίπεδο ασφαλείας τις 48.82 απότομες επιταχύνσεις ανά 100χλμ. και τις 45.40 απότομες επιβραδύνσεις ανά 100χλμ., παράγοντας πρωτότυπα αποτελέσματα για τα όρια δύο κλάσεων ταξινόμησης της οδικής συμπεριφοράς.

6.3 Προτάσεις για αξιοποίηση των αποτελεσμάτων

Βάσει των αποτελεσμάτων και των συμπερασμάτων που εξήχθησαν κατά την εκπόνηση της παρούσας εργασίας, επιχειρείται η παράθεση συγκεκριμένων προτάσεων προκειμένου να μετουσιωθεί σε όφελος στην εξέλιξη των Συστημάτων Ασφαλούς Προσέγγισης και στο ερευνητικό έργο του τομέα της Οδικής Ασφάλειας γενικότερα. Πιο συγκεκριμένα, προτείνεται:

1. Η αξιοποίηση των μοντέλων ταξινόμησης με τις καλύτερες επιδόσεις πρόβλεψης και ταξινόμησης για την αναγνώριση του επιπέδου ασφαλείας από δεδομένα εξέτασης φυσικής οδήγησης. Τα συγκεκριμένα μοντέλα κατέδειξαν καλή προβλεπτική ικανότητα και δυνητικά μπορούν να αποβούν ιδιαιτέρως χρήσιμα στην εξέλιξη προηγμένων Συστημάτων Πρόβλεψης και Ασφάλειας. Τα Συστήματα θα είναι σε θέση επί τόπου αναγνώρισης της επικίνδυνης οδικής συμπεριφοράς μέσα από αναλύσεις αισθητήρων των πεντάλ του οχήματος, διάρκειας και διανυθείσας απόστασης διαδρομής και άλλων αντίστοιχων παραμέτρων.

2. Η δημιουργία/βελτίωση εφαρμογής κινητών τηλεφώνων για τον εντοπισμό απότομων περιστατικών σε πραγματικό χρόνο με βάση τα οδηγικά χαρακτηριστικά του οδηγού.

3. Η ενημέρωση των χρηστών της οδού σε πραγματικό χρόνο, εάν εντοπιστεί υπέρβαση του ορίου απότομων περιστατικών, μέσω των κέντρων διαχείρισης κυκλοφορίας.

4. Η ανάπτυξη συστημάτων δημιουργίας προφίλ οδηγών ανάλογα με τον αριθμό απότομων περιστατικών στην διαδρομή τους. Εάν ταυτόχρονα ληφθεί υπόψιν και το επιπέδο συμμόρφωσης των οδηγών σε κανόνες και κώδικες κυκλοφορίας και των σημαντικών μεταβλητών για την οδική ασφάλεια, η αξιοποίηση των ευρημάτων της Διπλωματικής θα είναι μεγαλύτερη. Η αξιολόγηση μπορεί να παρουσιάζεται στους ίδιους τους οδηγούς, προκειμένου να αναγνωρίζουν τα λάθη τους και να τα διορθώνουν.

5. Η συνεργασία με την Τροχαία και η αποδοχή των οδηγών για κοινοποίηση των δεδομένων του οχήματος ή του κινητού τους τηλεφώνου, σύμφωνα με τους Γενικούς Κανονισμούς Προστασίας Προσωπικών Δεδομένων (G.D.P.R.) μπορεί να αποβεί καταλυτική στο μέλλον της Οδικής Ασφάλειας. Οι οδηγοί με καλές αξιολογήσεις οδικής συμπεριφοράς μπορεί να

επιβραβεύονται, λειτουργώντας σαν ένα επιπλέον κίνητρο για την βελτίωση της Οδικής Ασφάλειας.

6.4 Προτάσεις για περαιτέρω έρευνα

Η εξέλιξη της Οδικής Ασφάλειας αποτελεί έναν δυναμικό τομέα με διαρκώς αναπτυσσόμενες τάσεις και καινοτομίες, οι οποίες επιχειρούν να αναβαθμίσουν τα επίπεδα ασφαλείας των χρηστών της οδού. Οι τεχνικές μηχανικής μάθησης αποτελούν την σύγχρονη τάση του κλάδου για αυτό το εγχείρημα, με τις αναλύσεις των οδικών στοιχείων να είναι το κυρίαρχο αντικείμενο έρευνας. Η παρούσα εργασία καταπιάστηκε με διάφορα σύνθετα ζητήματα που σχετίζονται με την συλλογή, επεξεργασία, την στατιστική ανάλυση, την πρόβλεψη και την μοντελοποίηση κατάλληλων αλγορίθμων ανά επίπεδο ασφαλείας. Εν προκειμένω, επιχείρησε να εξελίξει το γνωσιακό υπόβαθρο του τομέα μέσα από τις αναλύσεις πολυεπίπεδων δεδομένων και της αξιοποίησης πληθώρας αλγορίθμων Μηχανικής εκμάθησης. Ως εκ τούτου, είναι σαφές ότι ανέκυψαν ορισμένες δυσκολίες και ελλείψεις κατά την διαδικασία εκπόνησής της, οι οποίες παραμένουν ερευνητική πρόκληση και χρήζουν αναφοράς.

1. Εξέταση επιπλέον δεδομένων στην υπάρχουσα βάση δεδομένων. Η συλλογή και επεξεργασία δεδομένων σχετιζόμενα με δημογραφικά στοιχεία του οδηγού, όπως η ηλικία, το φύλο και η οδηγική εμπειρία, με γεωμετρικά και χωρικά στοιχεία, όπως το πλήθος των λωρίδων οδήγησης και το πλάτος της οδού, με στοιχεία διαχείρισης κυκλοφορίας, όπως η ύπαρξη οριζόντιας και κατακόρυφης σήμανσης και η σηματοδότηση, ή ακόμα και με ψυχοσωματικά στοιχεία μπορεί να προσφέρουν χρήσιμες πληροφορίες στην ταξινόμηση της Οδικής συμπεριφοράς. Με την αύξηση της παρεχόμενης πληροφορίας, η μείωση του λάθους ταξινόμησης είναι αναπόφευκτη.

2. Ομαδοποίηση για κατασκευή ολοκληρωμένων προφίλ Οδικής συμπεριφοράς. Τα απρόοπτα περιστατικά, όπως οι επιταχύνσεις και οι επιβραδύνσεις, αποτελούν συνήθεις δείκτες επικίνδυνης συμπεριφοράς και το δείγμα των οδηγών που συλλέχθηκε ήταν επαρκές για την ανάλυση που προηγήθηκε. Ωστόσο, υπάρχει η πιθανότητα το δείγμα των οδηγών που εξετάστηκε να μην είναι αντιπροσωπευτικό στο σύνολό του για την αναγνώριση επιπέδων ασφαλείας. Τα συλλεχθέντα δεδομένα προτείνεται να ομαδοποιηθούν ανά οδηγό προκειμένου να καθοριστούν οι απαιτούμενες διαδρομές, τα απότομα περιστατικά και οι άλλες σημαντικές μεταβλητές, για την σκιαγράφηση συγκεκριμένων προφίλ Οδικής συμπεριφοράς.

3. Εξέταση εναλλακτικών μεθόδων αξιολόγησης της σημαντικότητας χαρακτηριστικών. Η διαδικασία της επιλογής χαρακτηριστικών περιλάμβανε τον καθορισμό του συντελεστή συσχέτισης Pearson και την ανάπτυξη μοντέλων παλινδρόμησης για την ποσοτικοποίηση της σημαντικότητας, με παράλληλη αξιολόγησή του καθενός με τον συντελεστή προσδιορισμού R^2. Καινοτόμες πρακτικές για την σύγκριση των Παλινδρομήσεων έχουν εισαχθεί στην στατιστική επιστήμη, οι οποίες ελαχιστοποιούν το περιθώριο λάθους στην σύγκριση λαμβάνοντας υπόψιν πολυπαραγοντικά κριτήρια. Τέτοιες μέθοδοι οι οποίες και προτείνονται για την αξιολόγηση είναι το Μπεϋζιανό Κριτήριο Πληροφοριών (BIC) και το Κριτήριο Πληροφοριών Akaike (AIC).

4. Διαχωρισμός σε παραπάνω κλάσεις ταξινόμησης. Παρά την ένδειξη για καλή διάκριση των τάξεων συμπεριφοράς σύμφωνα με τα αποτελέσματα, η διεθνής βιβλιογραφία προτείνει ως βέλτιστο τον διαχωρισμό σε 4 επιμέρους κλάσεις. Συνεπώς, προτείνεται η ανάλυση των παραπάνω οδικών δεδομένων να επαναληφθεί για διάκριση 4 τάξεων, αντί για δυαδική και σύγκριση των αποτελεσμάτων τους.

5. Εναλλακτικός τρόπος clustering των εξαρτημένων μεταβλητών. Η ομαδοποίηση των απότομων περιστατικών ανά 100χλμ. πραγματοποιήθηκε με την μέθοδο του K-μέσου με τα αποτελέσματα να καταδεικνύουν την σημαντική συνεισφορά της στην έρευνα. Η διερεύνηση της ομαδοποίησης με την χρήση Γκαουσιανού Μοντέλου Ανάμιξης (GMM) μπορεί να αποδειχθεί ιδιαίτερα χρήσιμη, λόγω και την κατανομής Γκαουσιανής φύσης των μεταβλητών.

6. Εξέταση περισσότερων μοντέλων βαθιάς εκμάθησης. Οι αλγόριθμοι Βαθιάς Μάθησης (Deep Learning) βασίζονται στην δομή του ανθρώπινου εγκεφάλου με την αξιοποίησή τους σε πληθώρα ερευνών να είναι ολοένα και συχνότερη. Συνήθως, αποδίδουν ιδιαίτερα ισχυρά σε πολυεπίπεδα δεδομένα, όπως αυτά που εξετάστηκαν στην παρούσα εργασία. Για την εξέταση των οδικών δεδομένων προτείνεται ο έλεγχος με ένα ειδικό μοντέλο Επαναλαμβανόμενων Νευρωνικών Δικτύων (RNN), τον αλγόριθμο μακροχρόνιας βραχυπρόθεσμης μνήμης (LSTM), λόγω τις ικανότητάς του στην εκμάθηση και αυτοδιόρθωσή του.

7. Βιβλιογραφία

1. Antoniou, C., Koutsopoulos, H. N., & Yannis, G. (2013). Dynamic data-driven local traffic state estimation and prediction. *Transportation Research Part C: Emerging Technologies*, *34*, 89-107. https://doi.org/10.1016/j.trc.2013.05.012
2. Arumugam, S., & Bhargavi, R. (2019). A survey on driving behavior analysis in usage based insurance using big data. *Journal of Big Data*, *6*(1), 1-21. https://doi.org/10.1186/s40537-019-0249-5
3. Bärgman, J., Lisovskaja, V., Victor, T., Flannagan, C., & Dozza, M. (2015). How does glance behavior influence crash and injury risk? A 'what-if'counterfactual simulation using crashes and near-crashes from SHRP2. Transportation Research Part F: Traffic Psychology and Behaviour, 35, 152-169. https://doi.org/10.1016/j.trf.2015.10.011
4. Bellman, R. (1966). Dynamic programming. Science, 153(3731), 34-37. https://doi.org/10.1126/science.153.3731.34
5. Bellman, R. (1960). Directions of mathematical research in nonlinear circuit theory. IRE Transactions on Circuit Theory, 7(4), 542-553. https://doi.org/10.1109/TCT.1960.1086701
6. Bifulco, G. N., Galante, F., Pariota, L., Spena, M. R., & Del Gais, P. (2014). Data collection for traffic and drivers' behaviour studies: a large-scale survey. *Procedia-social and behavioral sciences*, *111*, 721-730. https://doi.org/10.1016/j.sbspro.2014.01.106
7. Bolin, J. H., Edwards, J. M., Finch, W. H., & Cassady, J. C. (2014). Applications of cluster analysis to the creation of perfectionism profiles: a comparison of two clustering approaches. *Frontiers in psychology*, *5*, 343. https://doi.org/10.3389/fpsyg.2014.00343
8. Brodeur, Z. P., Herman, J. D., & Steinschneider, S. (2020). Bootstrap aggregation and cross-validation methods to reduce overfitting in reservoir control policy search. Water Resources Research, 56(8), e2020WR027184. https://doi.org/10.1029/2020WR027184
9. Chen, T., & Guestrin, C. (2016, August). Xgboost: A scalable tree boosting system. In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining (pp. 785-794). https://doi.org/10.1145/2939672.2939785
10. Chu, D., Deng, Z., He, Y., Wu, C., Sun, C., & Lu, Z. (2017). Curve speed model for driver assistance based on driving style classification. *IET Intelligent Transport Systems*, *11*(8), 501-510. https://doi.org/10.1049/iet-its.2016.0294
11. Deaton, A. (1982). Inequality and needs: Some experimental results for Sri Lanka. Population and Development Review, 35-49. https://doi.org/10.2307/2808105
12. Deng, C., Wu, C., Lyu, N., & Huang, Z. (2017). Driving style recognition method using braking characteristics based on hidden Markov model. *PloS one*, *12*(8), e0182419. https://doi.org/10.1371/journal.pone.0182419
13. Dozza, M., & Gonzalez, N. P. E. (2012). Recognizing Safetycritical Events from Naturalistic Driving Data. *Procedia-social and behavioral sciences*, *48*, 505-515. https://doi.org/10.1016/j.sbspro.2012.06.1029
14. European Commission. (2022). Road safety in the EU: fatalities in 2021 remain well below pre-pandemic level. Press release. Available: https://transport.ec.europa.eu/news/preliminary-2021-eu-road-safety-statistics-2022-03-28_en [Accessed 26-09-2022]
15. Flach, P. A. (2016). ROC analysis. In Encyclopedia of machine learning and data mining (pp. 1-8). Springer. https://doi.org/10.1007/978-1-4899-7502-7_739-1

16. Ghandour, R., Potams, A. J., Boulkaibet, I., Neji, B., & Al Barakeh, Z. (2021). Driver Behavior Classification System Analysis Using Machine Learning Methods. Applied Sciences, 11(22), 10562. https://doi.org/10.3390/app112210562
17. Gupta, A., Anand, A., & Hasija, Y. (2021, April). Recall-based Machine Learning approach for early detection of Cervical Cancer. In 2021 6th International Conference for Convergence in Technology (I2CT) (pp. 1-5). IEEE. https://doi.org/10.1109/I2CT51068.2021.9418099
18. Hand, David & Christen, Peter. (2018). A note on using the F-measure for evaluating record linkage algorithms. Statistics and Computing. 28. https://doi.org/10.1007/s11222-017-9746-6.
19. Hellenic Statistical Authority (ELSTAT). (2022). ROAD ACCIDENTS: Year 2020. Press release. Available https://www.statistics.gr/en/statistics/-/publication/SDT04/- [Accessed 26-09-2022] [In Greek].
20. Hsieh, C. J., Chang, K. W., Lin, C. J., Keerthi, S. S., & Sundararajan, S. (2008, July). A dual coordinate descent method for large-scale linear SVM. In Proceedings of the 25th international conference on Machine learning (pp. 408-415). https://doi.org/10.1145/1390156.1390208
21. Hu, J., Zhang, X., & Maybank, S. (2020). Abnormal driving detection with normalized driving behavior data: a deep learning approach. *IEEE transactions on vehicular technology*, *69*(7), 6943-6951. https://doi.org/10.1109/TVT.2020.2993247
22. Huang, J., & Ling, C. X. (2005). Using AUC and accuracy in evaluating learning algorithms. IEEE Transactions on knowledge and Data Engineering, 17(3), 299-310. https://doi.org/10.1109/TKDE.2005.50
23. Joachims, T. (2006, August). Training linear SVMs in linear time. In Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 217-226). https://doi.org/10.1145/1150402.1150429
24. Karlaftis, M. G., & Vlahogianni, E. I. (2011). Statistical methods versus neural networks in transportation research: Differences, similarities and some insights. *Transportation Research Part C: Emerging Technologies*, *19*(3), 387-399. https://doi.org/10.1016/j.trc.2010.10.004
25. Katrakazas, C., Antoniou, C., & Yannis, G. (2019). Time series classification using imbalanced learning for real-time safety assessment. In *Proceedings of the Transportation Research Board (TRB) 98th Annual Meeting, Washington, DC, January* (pp. 13-17). https://www.nrso.ntua.gr/geyannis/wp-content/uploads/geyannis-pc327.pdf
26. Katrakazas, C., Antoniou, C., & Yannis, G. (2020). Identification of driving simulator sessions of depressed drivers: A comparison between aggregated and time-series classification. *Transportation research part F: traffic psychology and behaviour*, *75*, 16-25. https://doi.org/10.1016/j.trf.2020.09.015
27. Katrakazas, C., Michelaraki, E., Sekadakis, M., & Yannis, G. (2020). A descriptive analysis of the effect of the COVID-19 pandemic on driving behavior and road safety. *Transportation research interdisciplinary perspectives*, *7*, 100186. https://doi.org/10.1016/j.trip.2020.100186
28. Katrakazas, C., Quddus, M., & Chen, W. H. (2017). A simulation study of predicting real-time conflict-prone traffic conditions. *IEEE Transactions on Intelligent Transportation Systems*, *19*(10), 3196-3207. https://doi.org/10.1109/TITS.2017.2769158
29. Keerthi, S. S., DeCoste, D., & Joachims, T. (2005). A modified finite Newton method for fast solution of large scale linear SVMs. Journal of Machine Learning Research, 6(3).

30. Khodairy, M. A., & Abosamra, G. (2021). Driving behavior classification based on oversampled signals of smartphone embedded sensors using an optimized stacked-LSTM neural networks. *IEEE Access*, *9*, 4957-4972. https://doi.org/10.1109/ACCESS.2020.3048915
31. Knyazev, G. G., Bocharov, A. V., Levin, E. A., Savostyanov, A. N., & Slobodskoj-Plusnin, J. Y. (2008). Anxiety and oscillatory responses to emotional facial expressions. *Brain Research*, *1227*, 174-188. https://doi.org/10.1016/j.brainres.2008.06.108
32. Kodinariya, T. M., & Makwana, P. R. (2013). Review on determining number of Cluster in K-Means Clustering. International Journal, 1(6), 90-95.
33. Kontaxi, A., Ziakopoulos, A., & Yannis, G. (2021). Trip characteristics impact on the frequency of harsh events recorded via smartphone sensors. *IATSS research*, *45*(4), 574-583. https://doi.org/10.1016/j.iatssr.2021.07.004
34. Kvålseth, T. O. (1985). Cautionary note about R 2. The American Statistician, 39(4), 279-285. https://doi.org/10.1080/00031305.1985.10479448
35. Lainiotis, D. G. (1976). Partitioning: A unifying framework for adaptive systems, I: Estimation. Proceedings of the IEEE, 64(8), 1126-1143. https://doi.org/10.1109/PROC.1976.10284
36. Li, G., Li, S. E., Cheng, B., & Green, P. (2017). Estimation of driving style in naturalistic highway traffic using maneuver transition probabilities. *Transportation Research Part C: Emerging Technologies*, *74*, 113-125. https://doi.org/10.1016/j.trc.2016.11.011
37. Lin, N., Zong, C., Tomizuka, M., Song, P., Zhang, Z., & Li, G. (2014). An overview on study of identification of driver behavior characteristics for automotive control. *Mathematical Problems in Engineering*, *2014*. https://doi.org/10.1155/2014/569109
38. Liu, Y., Wang, G., Chen, H., Dong, H., Zhu, X., & Wang, S. (2011). An improved particle swarm optimization for feature selection. Journal of Bionic Engineering, 8(2), 191-200. https://doi.org/10.1016/S1672-6529(11)60020-6
39. Mantouka, E. G., Barmpounakis, E. N., & Vlahogianni, E. I. (2019). Identifying driving safety profiles from smartphone data using unsupervised learning. *Safety Science*, *119*, 84-90. https://doi.org/10.1016/j.ssci.2019.01.025
40. Marsh, Brendan. (2016). Multivariate Analysis of the Vector Boson Fusion Higgs Boson. https://www.researchgate.net/publication/306054843_Multivariate_Analysis_of_the_Vector_Boson_Fusion_Higgs_Boson
41. Miyajima, C., Nishiwaki, Y., Ozawa, K., Wakita, T., Itou, K., Takeda, K., & Itakura, F. (2007). Driver modeling based on driving behavior and its evaluation in driver identification. *Proceedings of the IEEE*, *95*(2), 427-437. https://doi.org/10.1109/JPROC.2006.888405
42. Mohammed, R., Rawashdeh, J., & Abdullah, M. (2020, April). Machine learning with oversampling and undersampling techniques: overview study and experimental results. In 2020 11th international conference on information and communication systems (ICICS) (pp. 243-248). IEEE. https://doi.org/10.1109/ICICS49469.2020.239556
43. Mumcuoglu, M. E., Alcan, G., Unel, M., Cicek, O., Mutluergil, M., Yilmaz, M., & Koprubasi, K. (2019, July). Driving behavior classification using long short term memory networks. In *2019 AEIT International Conference of Electrical and Electronic Technologies for Automotive (AEIT AUTOMOTIVE)* (pp. 1-6). IEEE. https://doi.org/10.23919/EETA.2019.8804534

44. Natekin, A., & Knoll, A. (2013). Gradient boosting machines, a tutorial. Frontiers in neurorobotics, 7, 21. https://doi.org/10.3389/fnbot.2013.00021
45. Ndiaye, E., Le, T., Fercoq, O., Salmon, J., & Takeuchi, I. (2019, May). Safe grid search with optimal complexity. In International Conference on Machine Learning (pp. 4771-4780). PMLR.
46. Nilsson, G. (2004). Traffic safety dimensions and the power model to describe the effect of speed on safety (Vol. 221). Univ.https://lucris.lub.lu.se/ws/files/4394446/1693353.pdf
47. Oh, J. S., Oh, C., Ritchie, S. G., & Chang, M. (2005). Real-time estimation of accident likelihood for safety enhancement. *Journal of transportation engineering*, *131*(5), 358-363.
48. OSeven Telematics (2022). Official Website. Available: https://oseven.io/ [Accessed 20-10-2022]
49. Osman, O. A., Hajij, M., Karbalaieali, S., & Ishak, S. (2019). A hierarchical machine learning classification approach for secondary task identification from observed driving behavior data. *Accident Analysis & Prevention*, *123*, 274-281. https://doi.org/10.1016/j.aap.2018.12.005
50. Papadimitriou, E., Argyropoulou, A., Tselentis, D. I., & Yannis, G. (2019). Analysis of driver behaviour through smartphone data: The case of mobile phone use while driving. *Safety Science*, *119*, 91-97. https://doi.org/10.1016/j.ssci.2019.05.059
51. Peterson, L. E. (2009). K-nearest neighbor. Scholarpedia, 4(2), 1883. http://dx.doi.org/10.4249/scholarpedia.1883
52. Pinkus, A. (1999). Approximation theory of the MLP model in neural networks. Acta numerica, 8, 143-195. https://doi.org/10.1017/S0962492900002919
53. Quinlan, J. R. (1986). Induction of decision trees. Machine learning, 1(1), 81-106. https://doi.org/10.1007/BF00116251
54. Rajbahadur, G. K., Wang, S., Oliva, G. A., Kamei, Y., & Hassan, A. E. (2021). The impact of feature importance methods on the interpretation of defect classifiers. IEEE Transactions on Software Engineering, 48(7), 2245-2261. https://doi.org/10.1109/TSE.2021.3056941
55. Sahraoui, A., Makhlouf, D., & Roose, P. (2018). Smart Traffic Management System for Anticipating Unexpected Road Incidents in Intelligent Transportation Systems. *International Journal of Grid and High Performance Computing (IJGHPC)*, *10*(4), 67-82. https://doi.org/10.4018/IJGHPC.2018100104
56. Saleh, K., Hossny, M., & Nahavandi, S. (2017, October). Driving behavior classification based on sensor data fusion using LSTM recurrent neural networks. In *2017 IEEE 20th International Conference on Intelligent Transportation Systems (ITSC)* (pp. 1-6). IEEE. https://doi.org/10.1109/ITSC.2017.8317835
57. Scikit-Learn (2022). Official Website. Available: https://scikit-learn.org/stable/ [Accessed 30-10-2022]
58. Shalev-Shwartz, S., Singer, Y., & Srebro, N. (2007, June). Pegasos: Primal estimated sub-gradient solver for svm. In Proceedings of the 24th international conference on Machine learning (pp. 807-814). https://doi.org/10.1145/1273496.1273598
59. Tran, L. D. (2017). Data Fusion with 9 degrees of freedom Inertial Measurement Unit to determine object's orientation. https://digitalcommons.calpoly.edu/eesp/400
60. Uddin, M. F. (2019, November). Addressing accuracy paradox using enhanched weighted performance metric in machine learning. In 2019 Sixth HCT Information Technology Trends (ITT) (pp. 319-324). IEEE. https://doi.org/10.1109/ITT48889.2019.9075071

61. Van Huysduynen, H. H., Terken, J., & Eggen, B. (2018). The relation between self-reported driving style and driving behaviour. A simulator study. *Transportation research part F: traffic psychology and behaviour*, *56*, 245-255. https://doi.org/10.1016/j.trf.2018.04.017
62. Van Ly, M., Martin, S., & Trivedi, M. M. (2013, June). Driver classification and driving style recognition using inertial sensors. In 2013 IEEE Intelligent Vehicles Symposium (IV) (pp. 1040-1045). IEEE. https://doi.org/10.1109/IVS.2013.6629603
63. Vlahogianni, E. I., Karlaftis, M. G., & Golias, J. C. (2006). Statistical methods for detecting nonlinearity and non-stationarity in univariate short-term time-series of traffic volume. *Transportation Research Part C: Emerging Technologies*, *14*(5), 351-367. https://doi.org/10.1016/j.trc.2006.09.002
64. Wang, K., Xue, Q., & Lu, J. J. (2021). Risky driver recognition with class imbalance data and automated machine learning framework. *International journal of environmental research and public health*, *18*(14), 7534. https://doi.org/10.3390/ijerph18147534
65. World Health Organization – WHO. (2022). Road traffic injuries. Available: https://www.who.int/news-room/fact-sheets/detail/road-traffic-injuries
66. XGBoost developer team (2021). XGBoost Documentation. Official website. Available: https://xgboost.readthedocs.io/en/latest/index.html [Accessed 28/09/2022]
67. Xie, J., & Zhu, M. (2019). Maneuver-based driving behavior classification based on random forest. *IEEE Sensors Letters*, *3*(11), 1-4. https://doi.org/10.1109/LSENS.2019.2945117
68. Yan, F., Liu, M., Ding, C., Wang, Y., & Yan, L. (2019). Driving style recognition based on electroencephalography data from a simulated driving experiment. *Frontiers in psychology*, *10*, 1254. https://doi.org/10.3389/fpsyg.2019.01254
69. Yang, K., Al Haddad, C., Yannis, G., & Antoniou, C. (2021, June). Driving behavior safety levels: Classification and evaluation. In *2021 7th International Conference on Models and Technologies for Intelligent Transportation Systems (MT-ITS)* (pp. 1-6). IEEE. https://doi.org/10.1109/MT-ITS49943.2021.9529309
70. Yang, L., Ma, R., Zhang, H. M., Guan, W., & Jiang, S. (2018). Driving behavior recognition using EEG data from a simulated car-following experiment. *Accident Analysis & Prevention*, *116*, 30-40. https://doi.org/10.1016/j.aap.2017.11.010
71. Yi, D., Su, J., Liu, C., Quddus, M., Chen, W.-H., 2019. A machine learning based personalized system for driving state recognition. Transportation Research Part C: Emerging Technologies 105, 241–261. https://doi.org/10.1016/j.trc.2019.05.042
72. Yu, R., Abdel-Aty, M., 2013. Utilizing support vector machine in real-time crash risk evaluation. Accident Analysis & Prevention 51, 252–259. https://doi.org/10.1016/j.aap.2012.11.027
73. Zhang, C., Patel, M., Buthpitiya, S., Lyons, K., Harrison, B., & Abowd, G. D. (2016, March). Driver classification based on driving behaviors. In Proceedings of the 21st International Conference on Intelligent User Interfaces (pp. 80-84). https://doi.org/10.1145/2856767.2856806
74. Zhongwen, Z., & Huanghuang, G. (2017, June). Visualization study of high-dimensional data classification based on PCA-SVM. In 2017 IEEE Second International Conference on Data Science in Cyberspace (DSC) (pp. 346-349). IEEE. https://doi.org/10.1109/DSC.2017.57
75. Zhou, J., Li, C., Arslan, C. A., Hasanipanah, M., & Bakhshandeh Amnieh, H. (2021). Performance evaluation of hybrid FFA-ANFIS and GA-ANFIS models to predict particle

size distribution of a muck-pile after blasting. Engineering with computers, 37(1), 265-274. https://doi.org/10.1007/s00366-019-00822-0

76. Zhu, H., Xiao, R., Zhang, J., Liu, J., Li, C., & Yang, L. (2022). A Driving Behavior Risk Classification Framework via the Unbalanced Time Series Samples. *IEEE Transactions on Instrumentation and Measurement, 71*, 1-12. https://doi.org/10.1109/TIM.2022.3145359
77. Ziakopoulos, A., & Yannis, G. (2020). A review of spatial approaches in road safety. *Accident Analysis & Prevention, 135*, 105323. https://doi.org/10.1016/j.aap.2019.105323
78. Ziakopoulos, A. (2021). Spatial analysis of harsh driving behavior events in urban networks using high-resolution smartphone and geometric data. *Accident Analysis & Prevention, 157,* 106189. https://doi.org/10.1016/j.aap.2021.106189

Παράρτημα

Παράρτημα 1 – Αριθμητικές τιμές Σημαντικότητας Χαρακτηριστικών

Πίνακας Παραρτήματος 1: Αριθμητικές τιμές σημαντικότητας με Γραμμική Παλινδρόμηση για απότομες επιταχύνσεις ανά 100χλμ.

Μεταβλητή	**Μέτρο σημαντικότητας**
duration	-5.37419
total_distance	-982.27635
risky_hours	77.85939
driving_duration	-16.82829
avg speed	-355.30741
av_speeding_kmh_no_changer	250.9501
avg driving speed	417.98748
av_speeding_kmh	223.73726
sum_speeding	26.97048
time_mobile_usage	10.15492
time_mobile_usage/driving duration	3107.93863
sum_speeding/driving duration	-25428.48537
speeding_score	-39.43546
mu_score	9.50493
GRdriving	-9.47116
GRwalking	-2.65892

Πίνακας Παραρτήματος 2: Αριθμητικές τιμές σημαντικότητας με Γραμμική Παλινδρόμηση για απότομες επιβραδύνσεις ανά 100χλμ.

Μεταβλητή	**Μέτρο σημαντικότητας**
duration	-7.90073
total_distance	-1456.55919
risky_hours	12.7044
driving_duration	-18.91987
avg speed	-321.69586
av_speeding_kmh_no_changer	272.64803
avg driving speed	464.33228
av_speeding_kmh	110.43149
sum_speeding	33.02136
time_mobile_usage	1.39288
time_mobile_usage/driving duration	1603.3492

sum_speeding/driving duration	-24035.13322
speeding_score	-49.17894
mu_score	3.37065
GRdriving	-18.39765
GRwalking	-6.48535

Πίνακας Παραρτήματος 3: Αριθμητικές τιμές σημαντικότητας με Παλινδρόμηση Random Forests για απότομες επιταχύνσεις ανά 100χλμ.

Μεταβλητή	**Μέτρο σημαντικότητας**
duration	0.09767
total_distance	0.11955
risky_hours	0.00752
driving_duration	0.13107
avg speed	0.07142
av_speeding_kmh_no_changer	0.04315
avg driving speed	0.07848
av_speeding_kmh	0.04468
sum_speeding	0.02753
time_mobile_usage	0.01893
time_mobile_usage/driving duration	0.0195
sum_speeding/driving duration	0.02394
speeding_score	0.07887
mu_score	0.07361
GRdriving	0.08228
GRwalking	0.08181

Πίνακας Παραρτήματος 4: Αριθμητικές τιμές σημαντικότητας με Παλινδρόμηση Random Forests για απότομες επιβραδύνσεις ανά 100χλμ.

Μεταβλητή	**Μέτρο σημαντικότητας**
duration	0.09305
total_distance	0.11028
risky_hours	0.00798
driving_duration	0.12305
avg speed	0.08343
av_speeding_kmh_no_changer	0.03172
avg driving speed	0.09955
av_speeding_kmh	0.04826
sum_speeding	0.02207
time_mobile_usage	0.01386
time_mobile_usage/driving duration	0.01551
sum_speeding/driving duration	0.02
speeding_score	0.08273
mu_score	0.07957

GRdriving	0.08459
GRwalking	0.08435

Πίνακας Παραρτήματος 5: Αριθμητικές τιμές σημαντικότητας απότομων επιταχύνσεων ανά 100χλμ. με Linear SVR

Μεταβλητή	**Μέτρο σημαντικότητας**
duration	0.00289
total_distance	2.60096
risky_hours	-0.53425
driving_duration	-0.03186
avg speed	0.05886
av_speeding_kmh_no_changer	-14.43063
avg driving speed	-0.18284
av_speeding_kmh	40.6254
sum_speeding	26.70417
time_mobile_usage	0.02154
time_mobile_usage/driving duration	-2.39697
sum_speeding/driving duration	-2.42094
speeding_score	0.03619
mu_score	0.00452
GRdriving	0.00074
GRwalking	0.00024

Πίνακας Παραρτήματος 6: Αριθμητικές τιμές σημαντικότητας απότομων επιβραδύνσεων ανά 100χλμ. με Linear SVR

Μεταβλητή	**Μέτρο σημαντικότητας**
duration	4.94976
total_distance	-53.06561
risky_hours	-7.2846
driving_duration	-17.23078
avg speed	48.11745
av_speeding_kmh_no_changer	213.38284
avg driving speed	-31.70463
av_speeding_kmh	107.51007
sum_speeding	19.73039
time_mobile_usage	4.59024
time_mobile_usage/driving duration	0.30568
sum_speeding/driving duration	-0.39585
speeding_score	-17.02676
mu_score	3.6954
GRdriving	-2.85468

GRwalking	-5.3131

Πίνακας Παραρτήματος 7: Αριθμητικές τιμές σημαντικότητας με Παλινδρόμηση XGBoost για απότομες επιταχύνσεις ανά 100χλμ.

Μεταβλητή	**Μέτρο σημαντικότητας**
duration	0.04703
total_distance	0.07655
risky_hours	0.04656
driving_duration	0.11669
avg speed	0.04941
av_speeding_kmh_no_changer	0.10721
avg driving speed	0.05634
av_speeding_kmh	0.1234
sum_speeding	0.04841
time_mobile_usage	0.05222
time_mobile_usage/driving duration	0.05816
sum_speeding/driving duration	0.04936
speeding_score	0.0401
mu_score	0.04427
GRdriving	0.04057
GRwalking	0.04371

Πίνακας Παραρτήματος 8: Αριθμητικές τιμές σημαντικότητας με Παλινδρόμηση XGBoost για απότομες επιβραδύνσεις ανά 100χλμ.

Μεταβλητή	**Μέτρο σημαντικότητας**
duration	0.04628
total_distance	0.06983
risky_hours	0.04135
driving_duration	0.14432
avg speed	0.04696
av_speeding_kmh_no_changer	0.11113
avg driving speed	0.06328
av_speeding_kmh	0.14371
sum_speeding	0.0394
time_mobile_usage	0.04853
time_mobile_usage/driving duration	0.04077
sum_speeding/driving duration	0.04298
speeding_score	0.04227
mu_score	0.04116
GRdriving	0.0384
GRwalking	0.03965

Πίνακας Παραρτήματος 9: Αριθμητικές τιμές σημαντικότητας με Παλινδρόμηση Decision Trees για απότομες επιταχύνσεις ανά 100χλμ.

Μεταβλητή	**Μέτρο σημαντικότητας**
duration	0.10118
total_distance	0.11747
risky_hours	0.00683
driving_duration	0.12351
avg speed	0.0789
av_speeding_kmh_no_changer	0.0471
avg driving speed	0.0839
av_speeding_kmh	0.03905
sum_speeding	0.0279
time_mobile_usage	0.0195
time_mobile_usage/driving duration	0.01761
sum_speeding/driving duration	0.02483
speeding_score	0.07973
mu_score	0.07172
GRdriving	0.08092
GRwalking	0.07984

Πίνακας Παραρτήματος 10: Αριθμητικές τιμές σημαντικότητας με Παλινδρόμηση Decision Trees για απότομες επιβραδύνσεις ανά 100χλμ.

Μεταβλητή	**Μέτρο σημαντικότητας**
duration	0.09251
total_distance	0.10551
risky_hours	0.0066
driving_duration	0.12348
avg speed	0.08076
av_speeding_kmh_no_changer	0.03061
avg driving speed	0.10543
av_speeding_kmh	0.05099
sum_speeding	0.02214
time_mobile_usage	0.01273
time_mobile_usage/driving duration	0.01571
sum_speeding/driving duration	0.01879
speeding_score	0.0835
mu_score	0.08185
GRdriving	0.08581
GRwalking	0.08358

Παράρτημα 2 – Κώδικας επεξεργασίας δεδομένων

#Κώδικας προεπεξεργασίας δεδομένων και ομαδοποίησης με αλγόριθμο Κ-μέσου

```
Ya=df['ha/100km'].to_numpy()
Yb=df['hb/100km'].to_numpy()
X=df[['total_distance','speeding_score','driving_duration','mu_score','avg driving speed']].to_numpy()
from sklearn.preprocessing import normalize
Xnew=normalize(X)
for cluster_number in range(1,3):
    print(f"\nCluster number: {cluster_number}")
    K_Means=KMeans(n_clusters=cluster_number,random_state=0)
    K_Means.fit(np.expand_dims(Ya, -1))
    K_Means.cluster_centers_

    print(f"Cluster centers: {[x for x in np.sort(K_Means.cluster_centers_)]}")

    ya_threshold_value = np.mean(K_Means.cluster_centers_)
    print(f"Threshold for dangerous driving with respect to ha/100km is: {ya_threshold_value}")
Yaquant = np.where(Ya>ya_threshold_value,1,0)
for cluster_number in range(1,3):
    print(f"\nCluster number: {cluster_number}")
    K_Means=KMeans(n_clusters=cluster_number,random_state=0)
    K_Means.fit(np.expand_dims(Yb, -1))
    K_Means.cluster_centers_

    print(f"Cluster centers: {[x for x in np.sort(K_Means.cluster_centers_)]}")

    yb_threshold_value = np.mean(K_Means.cluster_centers_)
    print(f"Threshold for dangerous driving with respect to hb/100km is: {yb_threshold_value}")
Ybquant = np.where(Yb>yb_threshold_value,1,0)
```

#Κώδικας train-test split

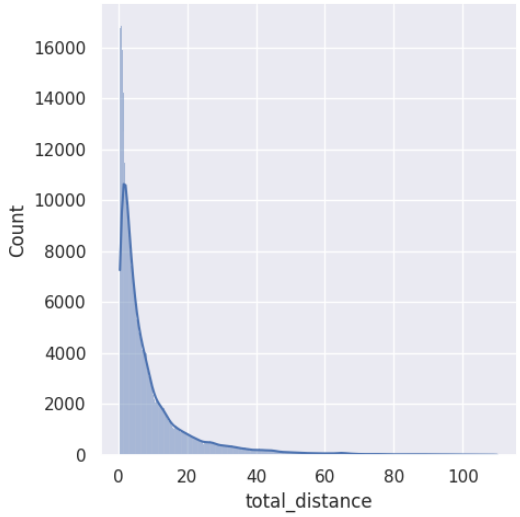
```
from sklearn.model_selection import train_test_split
```

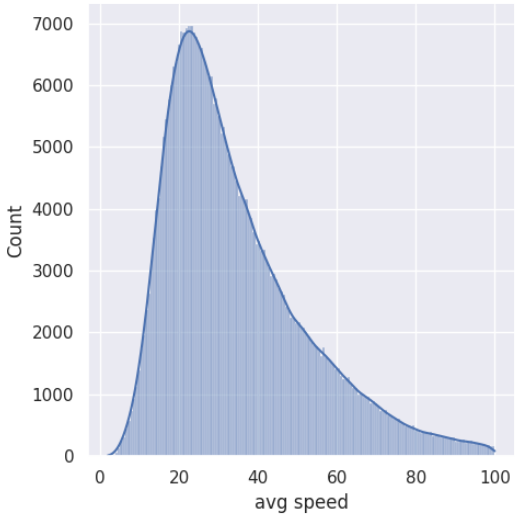
```
X_train, X_test, Yaquant_train, Yaquant_test = train_test_split(Xnew, Yaquant, random_state=42, stratify=Yaquant)

X_train_2, X_test_2, Ybquant_train, Ybquant_test = train_test_split(Xnew, Ybquant, random_state=42, stratify = Ybquant)
```

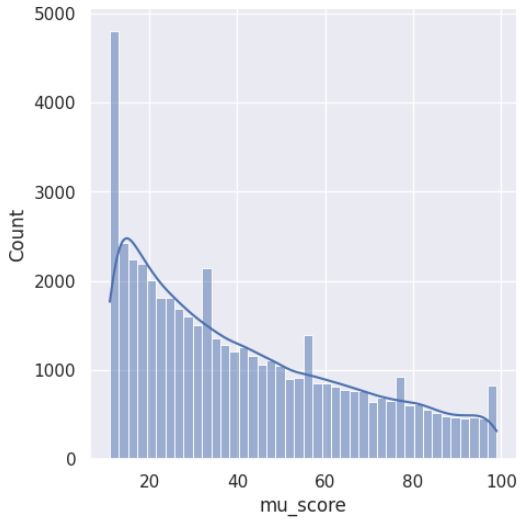
#Κώδικας Υπερδειγματοληψίας

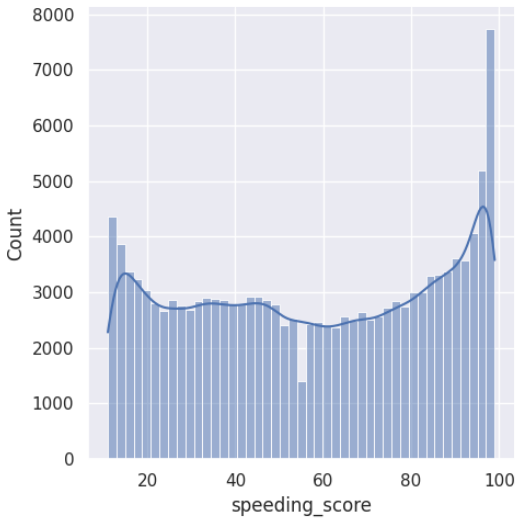
```
pip install imbalanced-learn

conda install -c conda-forge imbalanced-learn

from imblearn.over_sampling import SMOTE

sm=SMOTE(random_state=42)

X_train_sm, Yaquant_train_sm = sm.fit_resample(X_train, Yaquant_train)

X_train_2_sm, Ybquant_train_sm = sm.fit_resample(X_train_2, Ybquant_train)
```

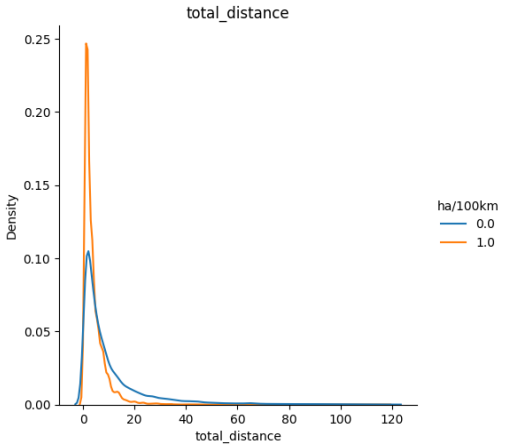
Παράρτημα 3 – Περιγραφικά Γραφήματα μεταβλητών

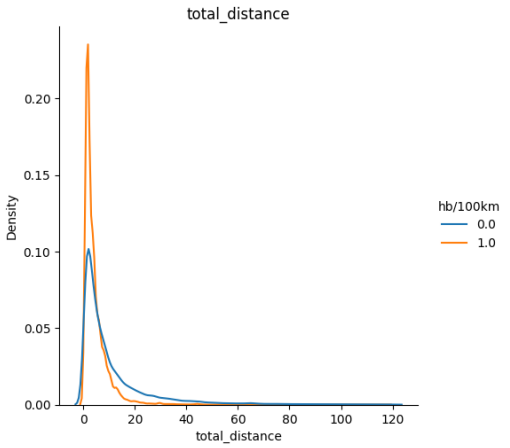


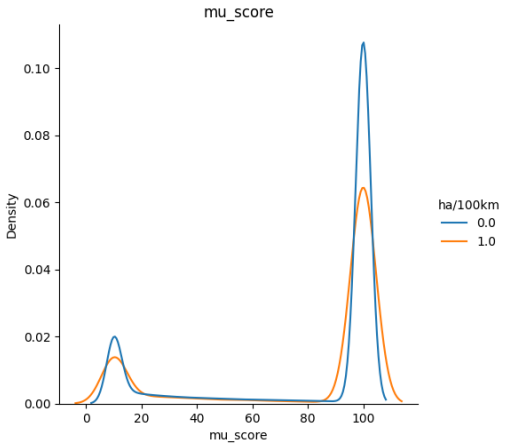


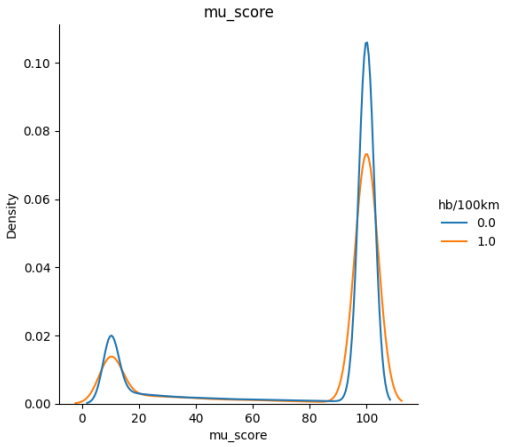


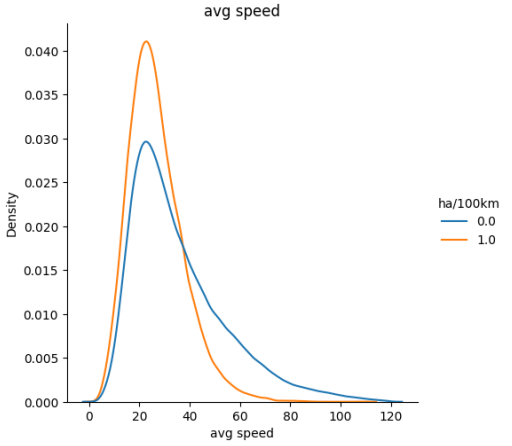


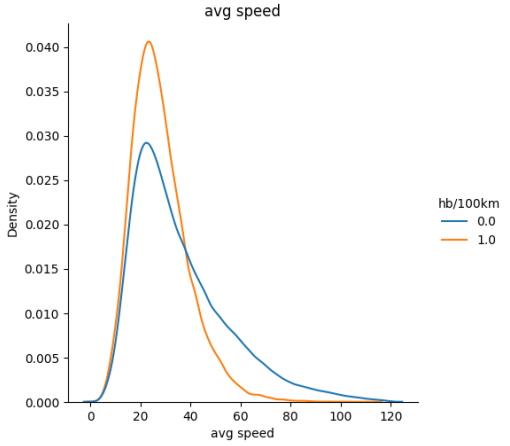

total_distance
Density
total_distance
ha/100km
0.0
1.0


total_distance
Density
total_distance
hb/100km
0.0
1.0


mu_score
Density
mu_score
ha/100km
0.0
1.0


mu_score
Density
mu_score
hb/100km
0.0
1.0


avg speed
Density
avg speed
ha/100km
0.0
1.0


avg speed
Density
avg speed
hb/100km
0.0
1.0