ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
Σχολή Πολιτικών Μηχανικών
Τομέας Μεταφορών και Συγκοινωνιακής Υποδομής

ΔΙΕΡΕΥΝΗΣΗ ΤΗΣ ΕΠΙΡΡΟΗΣ ΤΗΣ ΧΡΗΣΗΣ ΚΙΝΗΤΟΥ ΤΗΛΕΦΩΝΟΥ ΣΤΗ ΣΥΜΠΕΡΙΦΟΡΑ ΤΟΥ ΟΔΗΓΟΥ ΜΕ ΤΕΧΝΙΚΕΣ ΜΗΧΑΝΙΚΗΣ ΜΑΘΗΣΗΣ ΑΝΙΣΟΡΡΟΠΩΝ ΔΕΔΟΜΕΝΩΝ

Κωνσταντίνος-Ειρηναίος Κασελούρης

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

Αθήνα, Μάρτιος 2023

ΕΥΧΑΡΙΣΤΙΕΣ

Η ολοκλήρωση της συγκεκριμένης Διπλωματικής Εργασίας σηματοδοτεί το τέλος την προπτυχιακής μου πορείας στην σχολή Πολιτικών Μηχανικών του ΕΜΠ. Ένα όνειρο που είχα από μικρός, να γίνω πολιτικός μηχανικός και δη συγκοινωνιολόγος, ύστερα από πολλή προσπάθεια και ατέρμονες δυσκολίες στην πορεία, γίνεται πραγματικότητα.

Κατ' αρχάς, θα ήθελα να ευχαριστήσω τον κ. Γιώργο Γιαννή, Καθηγητή της Σχολής Πολιτικών Μηχανικών ΕΜΠ και Διευθυντή του Τομέα Μεταφορών και Συγκοινωνιακής Υποδομής, για την εμπιστοσύνη που μου έδειξε με την ανάθεση της παρούσας διπλωματικής εργασίας, καθώς και για την οργάνωση και καθοδήγησή του ειδικά σε θέματα συγγραφής της εργασίας. Επίσης, τον ευχαριστώ πολύ για το γνωστικό υπόβαθρο που μου μετέδωσε καθ' όλη την διάρκεια φοίτησής μου στην κατεύθυνση του συγκοινωνιολόγου μηχανικού.

Επίσης, θα ήθελα να ευχαριστήσω πολύ τον Δρ. Χρήστο Κατρακάζα για την υποστήριξη, την καθοδήγηση που μου παρείχε όσον αφορά στο κομμάτι του προγραμματισμού, με το οποίο δεν ήμουν εξοικειωμένος έως τώρα, όπως επίσης και για την υπομονή του και την γενικότερη καλή συνεργασία.

Ένα μεγάλο ευχαριστώ χρωστάω και στην Δρ. Εύα Μιχελαράκη για την προθυμία και την υποστήριξή της όσον αφορά σε θέματα παρουσίασης της εργασίας, όπως επίσης και για τα γνωστικά εφόδια που μου παρείχε τόσο εκείνη όσο και το υπόλοιπο διδακτορικό προσωπικό από τα μαθήματα κατεύθυνσης.

Θα ήθελα να ευχαριστήσω και την εταιρεία OSeven Telematics για την παροχή των οδικών δεδομένων που αποτέλεσαν αντικείμενο ανάλυσης και επεξεργασίας στην παρούσα εργασία.

Στην συνέχεια, οφείλω ένα ευχαριστώ στους φίλους μου που ήταν δίπλα μου και με ενθάρρυναν τόσο για την εκπόνηση της συγκεκριμένης εργασίας όσο και καθ' όλη την διάρκεια των σπουδών μου.

Τέλος, σε αυτό το σημείο θα ήθελα να εκφράσω την αμέριστη ευγνωμοσύνη μου στην οικογένειά μου, στον αδερφό μου Νικόλα και τους γονείς μου, Χρήστο και Αθηνά, για την αδιάκοπη ενθάρρυνσή τους τόσο στα σχολικά όσο και στα φοιτητικά έτη, την συμπαράστασή τους στις δυσκολίες, τις προκλήσεις και τα εμπόδια που αντιμετώπιζα τόσο κατά την εκπόνηση της διπλωματικής όσο και καθ' όλη την διάρκεια των σπουδών μου, καθώς και την εμπιστοσύνη που μου έδειχναν και μου δείχνουν πάντα, που χωρίς αυτή τίποτα δεν θα ήταν εφικτό. Κλείνοντας, ένα μεγάλο ευχαριστώ και στην γιαγιά μου, Φωτεινή, για την αγάπη και το ειλικρινές ενδιαφέρον που μου έδειχνε όλα τα χρόνια της ζωής μου, στην οποία και αφιερώνω την παρούσα εργασία.

Αθήνα, Μάρτιος 2023

Κασελούρης Κωνσταντίνος-Ειρηναίος

Διερεύνηση της επιρροής χρήσης κινητού τηλεφώνου στη συμπεριφορά του οδηγού με ανάλυση Μηχανικής Μάθησης ανισόρροπων δεδομένων

Κωνσταντίνος-Ειρηναίος Κασελούρης

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

Σύνοψη

Η συγκεκριμένη Διπλωματική Εργασία αποσκοπεί στη διερεύνηση της επιρροής της χρήσης κινητού τηλεφώνου στη συμπεριφορά του οδηγού μέσω στατιστικής ανάλυσης ανισόρροπων δεδομένων με τεχνικές Μηχανικής Μάθησης (Machine Learning). Για την ταξινόμηση, την παλινδρόμηση και την πρόβλεψη της χρήσης κινητού τηλεφώνου αξιοποιήθηκαν δεδομένα τηλεματικής της εταιρείας OSeven, τα οποία συλλέχθηκαν από μετρήσεις σε πραγματικές οδικές συνθήκες. Ως δείκτης επικίνδυνης συμπεριφοράς ορίστηκε η χρήση κινητού τηλεφώνου και η ταξινόμηση πραγματοποιήθηκε σε δύο επίπεδα συμπεριφοράς του οδηγού (επικίνδυνη και μη επικίνδυνη οδήγηση). Στο πρώτο μέρος των αναλύσεων, αναπτύχθηκαν συνολικά τέσσερις αλγόριθμοι ταξινόμησης, δύο συμπεριλαμβανομένων όλων των υπό εξέταση ανεξάρτητων μεταβλητών και οι δύο ίδιοι αλγόριθμοι με τις πέντε σημαντικότερες εξ' αυτών, όπως εκείνες προέκυψαν από τη μέθοδο Σημαντικότητας Χαρακτηριστικών (Feature Importance). Σημαντικότερες ανεξάρτητες μεταβλητές αναδείχθηκαν μεταβλητές σχετικές με την ταχύτητα διαδρομής, ενώ σύμφωνα με τις μετρικές αξιολόγησης ταξινόμησης καταλληλότερο μοντέλο κρίθηκε εκείνο της 'Γραμμικής Διαχωριστικής Ανάλυσης'. Στο δεύτερο μέρος των αναλύσεων, ακολουθήθηκε πανομοιότυπη διαδικασία για την παλινδρόμηση με εξαρτημένη μεταβλητή τη διάρκεια χρήσης κινητού τηλεφώνου με τον αλγόριθμο της 'Προσαρμοστικής Ενδυνάμωσης' με όλες τις ανεξάρτητες μεταβλητές να παρουσιάζει καλύτερη προβλεπτική ικανότητα και τη μεταβλητή της διάρκειας οδικής διαδρομής να θεωρείται σημαντικότερη.

Λέξεις-Κλειδιά: Μηχανική Μάθηση, ανάλυση συμπεριφοράς του οδηγού, οδική ασφάλεια, πραγματικές οδικές συνθήκες, ταξινόμηση χρήσης κινητού τηλεφώνου, μοντέλα ταξινόμησης, μοντέλο Γραμμικής Διαχωριστικής Ανάλυσης,μοντέλο Λογιστικής Παλινδρόμησης, ανισόρροπο σύνολο δεδομένων, μέθοδοι επαναδειγματοληψίας, Σημαντικότητα Χαρακτηριστικών, μοντέλα παλινδρόμησης, μοντέλο Προσαρμοστικής Ενδυνάμωσης, μοντέλο Γραμμικής Παλινδρόμησης, Βαθιά Μάθηση, Νευρωνικά Δίκτυα

Investigating the influence of mobile phone use on driving behavior with Machine Learning analysis of imbalanced data

Konstantinos-Eirinaios Kaselouris

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

Abstract

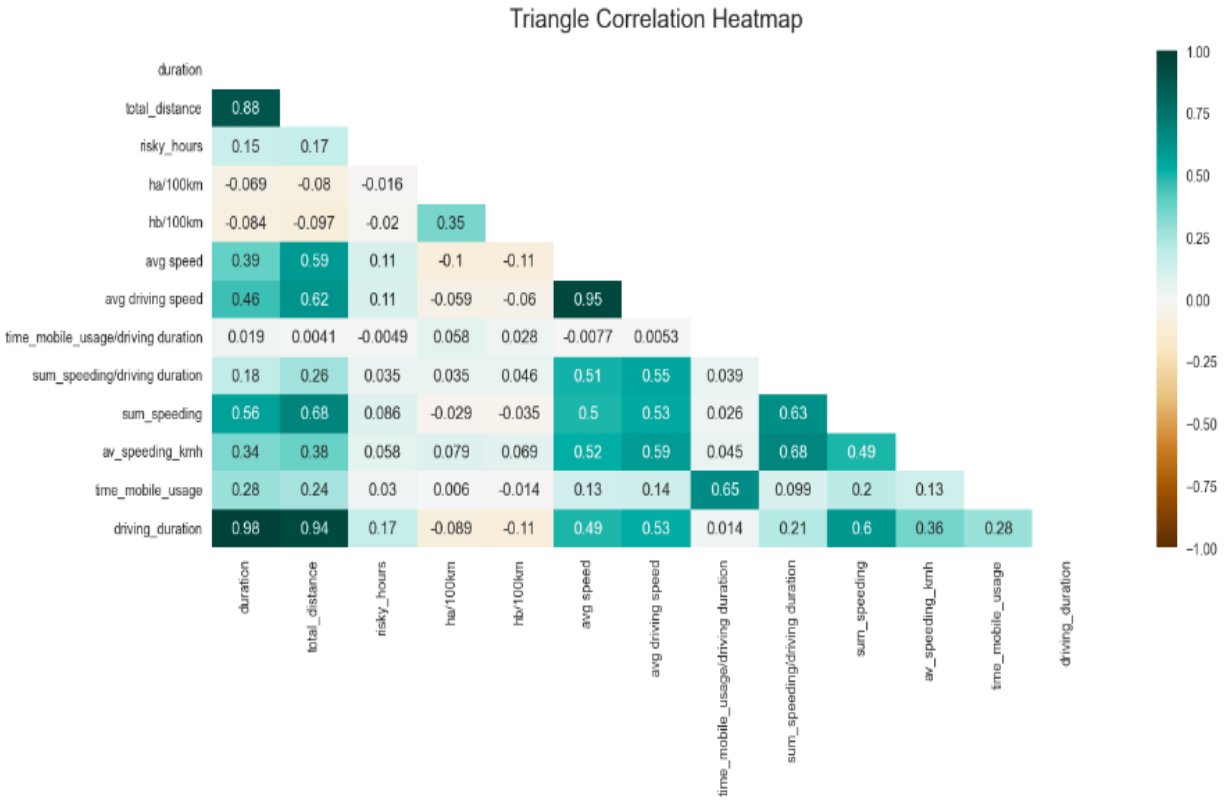
This Diploma Thesis aims to investigate the impact of mobile phone use on driving behaviour through statistical analysis of imbalanced data using Machine Learning techniques. For classification and regression of mobile phone usage, telematics data from the OSeven telematics company, collected from naturalistic measurements, were used. Mobile phone use was defined as an indicator of risky behaviour and classification was performed on two levels of driving behaviour (risky and not risky). In the first part of the analyses, a total of four classification algorithms were developed, two including all the independent variables under consideration and the same two algorithms with the five most significant of them, as derived from the Feature Importance method. Variables related to travel speed were found to be the most significant independent variables, while according to the classification evaluation metrics, the most appropriate model was considered to be that of 'Linear Discriminant Analysis'. In the second part of the analyses, an identical procedure was followed for the regression with the dependent variable of mobile phone usage duration with the 'Adaptive Strengthening' algorithm with all independent variables showing better predictive ability and the variable of road trip duration being considered the most significant.

Keywords: Machine Learning, road behaviour analysis, road safety, naturalistic driving, mobile phone usage classification, classification models, Linear Discriminant Analysis model,Logistic Regression model, imbalanced data set, resampling methods, Feature Importance, regression models, AdaBoost model, Linear Regression model, Deep Learning, Neural Networks

ΠΕΡΙΛΗΨΗ

Στόχος της παρούσας διπλωματικής εργασίας είναι **η διερεύνηση της επιρροής χρήσης του κινητού τηλεφώνου στη συμπεριφορά του οδηγού** και πιο συγκεκριμένα στα μεγέθη της ταχύτητας, της επιτάχυνσης και του χρόνου οδήγησης μέσω της μηχανικής μάθησης ανισόρροπων δεδομένων και αλγορίθμων ταξινόμησης και παλινδρόμησης. Τα δεδομένα που επιλέχθηκαν αντλήθηκαν από τη βάση δεδομένων της εταιρείας OSeven Telematics και καταγράφηκαν σε πραγματικές οδικές συνθήκες μέσω σχετικής εφαρμογής στα κινητά τηλέφωνα των οδηγών. Τα οδικά δεδομένα που αναλύθηκαν, αφορούσαν κυκλοφοριακά δεδομένα με επιμέρους δείκτες κίνησης και συμπεριφοράς του οδηγού. Πριν την έναρξη της ανάλυσης των δεδομένων καθορίστηκε η διάρκεια χρήσης κινητού τηλεφώνου ως δείκτης επικίνδυνης συμπεριφοράς του οδηγού, η οποία αποτέλεσε την εξαρτημένη μεταβλητή. Η ανάλυση πραγματοποιήθηκε στη γλώσσα προγραμματισμού **Python** μέσω του πακέτου PyCaret και σε προγραμματιστικό περιβάλλον JupyterNotebook.

Στο στάδιο της προεπεξεργασίας των δεδομένων συντάχθηκε ο τριγωνικός χάρτης θερμότητας Pearson, ο οποίος αποτέλεσε κριτήριο για την καταρχήν Επιλογή Χαρακτηριστικών.



Γράφημα Π.1:Τριγωνικός χάρτης συσχέτισης μεταβλητών

Σε δεύτερο στάδιο πραγματοποιήθηκε ανάλυση μοντέλων **ταξινόμησης** και από τη σύγκριση των μετρικών αξιολόγησης των μοντέλων αυτών προκρίθηκαν τα μοντέλα **της Γραμμικής Διαχωριστικής Ανάλυσης** και της **Λογιστικής Παλινδρόμησης**. Σημαντικότερη μεταβλητή από την συγκεκριμένη διαδικασία αναδείχθηκε η μέση ταχύτητα οδήγησης είτε με είτε χωρίς υπέρβαση του ορίου ταχύτητας. Η ταξινόμηση πραγματοποιήθηκε τόσο για όλες τις υπό εξέταση μεταβλητές όσο και για τις σημαντικότερες από αυτές. Ως εξαρτημένη μεταβλητή ορίστηκε η χρήση κινητού τηλεφώνου και προέκυψαν δύο επίπεδα οδικής συμπεριφοράς:

- Τιμή 0: Για μη χρήση κινητού τηλεφώνου κατά την διάρκεια της οδήγησης(Μη Επικίνδυνη).

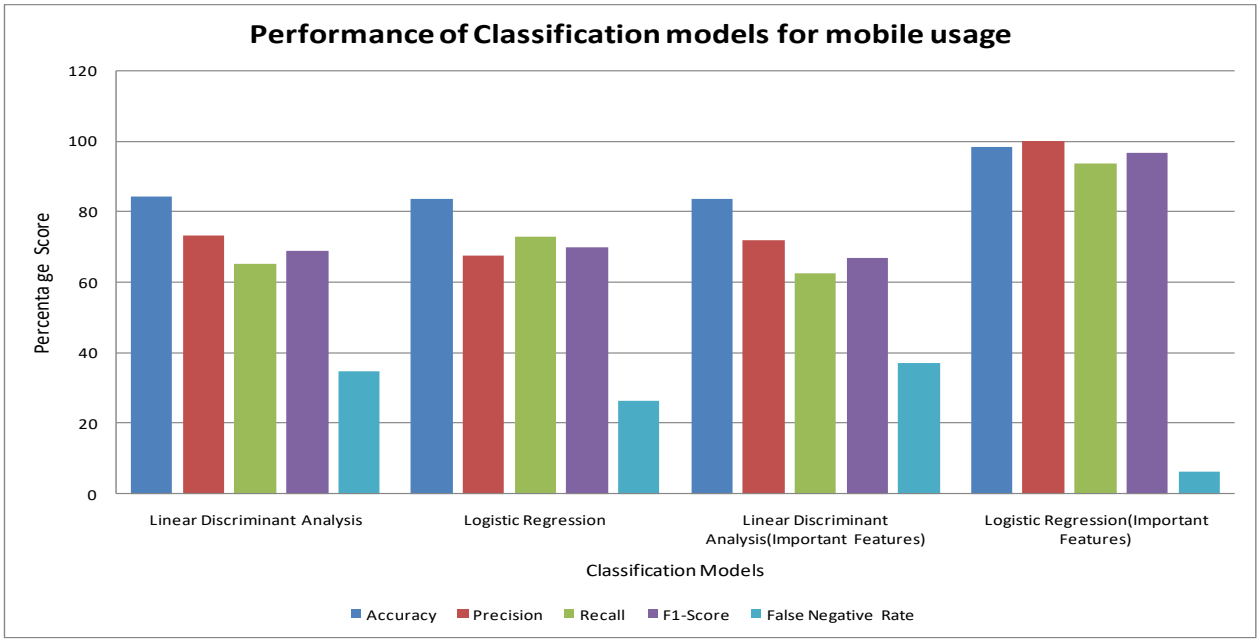
- Τιμή 1: Για χρήση κινητού τηλεφώνου κατά την διάρκεια της οδήγησης(Επικίνδυνη)

Αξίζει να σημειωθεί ότι σε αυτή την φάση λόγω του ανισόρροπου χαρακτήρα του δείγματος με μειονοτική τάξη την χρήση κινητού τηλεφώνου(Τιμή 1) προέκυψε η ανάγκη για την εφαρμογή της τεχνικής της Συνθετικής Μειονοτικής Υπερδειγματοληψίας (SMOTE), προκειμένου να εξισορροπηθεί το δείγμα και να μπορέσει να επιτευχθεί ορθή πρόβλεψη της χρήσης κινητού μέσω των μοντέλων ταξινόμησης.

Παρακάτω παρουσιάζονται συνοπτικά οι μετρικές αξιολόγησης για τα δύο καταλληλλότερα μοντέλα ταξινόμησης τόσο με όλες όσο και με τις σημαντικότερες ανεξάρτητες μεταβλητές.

Πίνακας Π.1: Συγκριτικός πίνακας μετρικών αξιολόγησης μοντέλων ταξινόμησης

ΤΑΞΙΝΟΜΗΣΗ	Αλγόριθμοι Ταξινόμησης											
	Linear Discriminant Analysis						Logistic Regression					
	Ορθότητα	Ακρίβεια	Ανάκληση	FNR	F1-Score	AUC Score	Ορθότητα	Ακρίβεια	Ανάκληση	FNR	F1-Score	AUC Score
Με όλες τις μεταβλητές	84,4%	73,3%	65,1%	34,6%	68,8%	89,5%	83,4%	67,4%	72,4%	26,4%	70,0%	89,1%
Με τις σημαντικότερες μεταβλητές	83,5%	72,0%	62,4%	37,2%	66,9%	87,7%	98,2%	99,8%	93,5%	6,4%	96,6%	99,9%



Γράφημα Π.2: Μετρικές αξιολόγησης μοντέλων ταξινόμησης για την χρήση κινητού τηλεφώνου

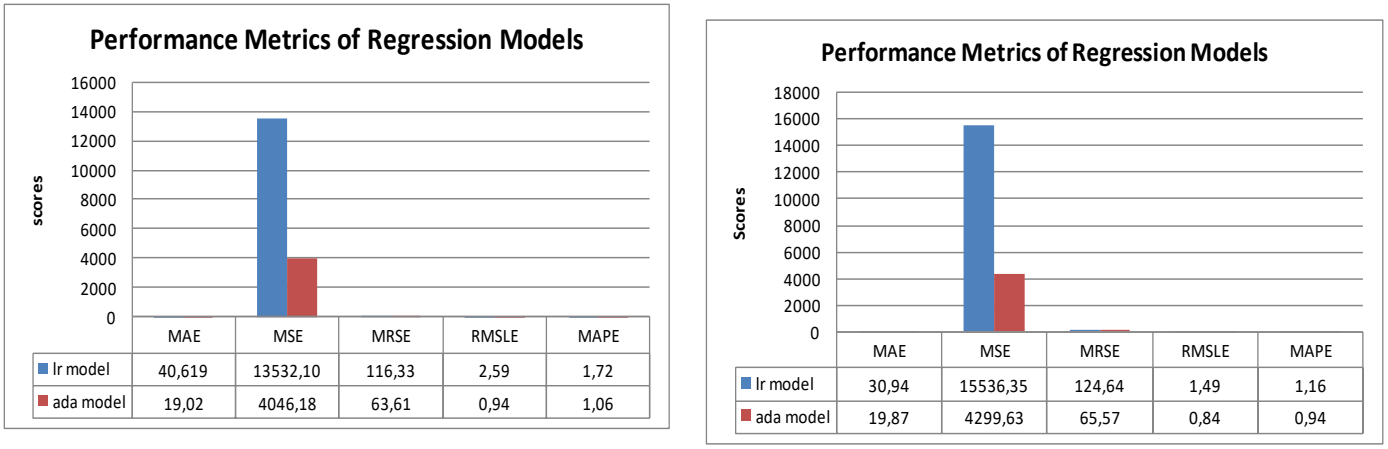
Όλοι οι αλγόριθμοι παρουσίασαν πολύ ικανοποιητική προβλεπτική ικανότητα και οι τρεις εκ των τεσσάρων κυμάνθηκαν σε παρόμοια επίπεδα. Το μοντέλο ταξινόμησης που επιλέχθηκε ως το βέλτιστο εκ των τεσσάρων ήταν εκείνο της Γραμμικής Διαχωριστικής Ανάλυσης λαμβάνοντας υπόψη όλες τις υπό εξέταση ανεξάρτητες μεταβλητές. Το μοντέλο της Λογιστικής Παλινδρόμησης με τις σπουδαιότερες μεταβλητές δεν επιλέχθηκε, καθώς παρουσίασε ορθότητα πολύ κοντά στην μονάδα, γεγονός που δείχνει υπερπροσαρμογή(overfitting) του μοντέλου στα δεδομένα.

Εν συνεχεία, πραγματοποιήθηκε η διαδικασία της παλινδρόμησης, με επιλεχθέντα μοντέλα για τη συγκεκριμένη διαδικασία εκείνα της Προσαρμοστικής Ενδυνάμωσης και η Γραμμικής Παλινδρόμησης. Ως εξαρτημένη μεταβλητή επιλέχθηκε η διάρκεια χρήσης κινητού τηλεφώνου. Όπως και στην ταξινόμηση, μέσω της διαδικασίας της Σημαντικότητας Χαρακτηριστικών προσδιορίστηκαν οι καλύτερες ανεξάρτητες μεταβλητές για τα δύο μοντέλα. Ως σπουδαιότερες μεταβλητές αναδείχθηκαν η διάρκεια οδικής διαδρομής με είτε χωρίς υπέρβαση ορίου ταχύτητας. Συνολικά πραγματοποιήθηκαν τέσσερις αναλύσεις παλινδρόμησης για τα μοντέλα,

τόσο με όλες τις υπό εξέταση ανεξάρτητες μεταβλητές όσο και με τις σπουδαιότερες από αυτές και τα αποτελέσματα της μεθόδου παρουσιάζονται συνοπτικά στον Πίνακα Π.2 και στο Γράφημα Π.3.

Πίνακας Π.2: Συγκριτικός πίνακας μετρικών αξιολόγησης μοντέλων παλινδρόμησης

ΠΑΛΙΝΔΡΟΜΗΣΗ	Αλγόριθμοι Παλινδρόμησης	
	AdaBoost Regressor	Linear Regression
	R2	R2
Με όλες τις μεταβλητές	0,842	0,497
Με τις σημαντικότερες μεταβλητές	0,840	0,422



Γράφημα Π.3: Μετρικές αξιολόγησης μοντέλων παλινδρόμησης με όλες (αριστερά) και με τις σημαντικότερες μεταβλητές(δεξιά)

Από τα παραπάνω μοντέλα επιλέχθηκε εκείνο της Προσαρμοστικής Ενδυνάμωσης με όλες τις υπό εξέταση ανεξάρτητες μεταβλητές, καθώς παρουσίαζε ικανή προβλεπτική ικανότητα με πολύ ικανοποιητική τιμή του συντελεστή προσδιορισμού R^2 και μικρότερα σφάλματα σε σχέση με το μοντέλο της Γραμμικής Παλινδρόμησης.

Κατά την διάρκεια εκπόνησης της Διπλωματικής Εργασίας και της παρατήρησης των αποτελεσμάτων εξήχθησαν αξιοσημείωτα συμπεράσματα τόσο για την ανάλυση της οδικής συμπεριφοράς όσο και για το γενικότερο ερευνητικό πεδίο της Οδικής Ασφάλειας.

1. **Βασικότερη παράμετρος επιρροής** της χρήσης κινητού τηλεφώνου σύμφωνα με τα μοντέλα **ταξινόμησης** αποδείχθηκε **η ταχύτητα** (km/h). Το γεγονός αυτό φαίνεται λογικό, καθώς η διάρκεια χρήσης κινητού τηλεφώνου που μετατράπηκε στη δυαδική μεταβλητή της χρήσης κινητού τηλεφώνου οδηγεί στην απόσπαση της προσοχής του οδηγού και έμμεσα επηρεάζει και την ταχύτητά του είτε αυξάνοντάς την με υπέρβαση του ορίου της είτε μειώνοντας την, επιβεβαιώνοντας τη διεθνή βιβλιογραφία.

2. **Βασικότερες παράμετροι επιρροής** της διάρκειας χρήσης κινητού σύμφωνα με τα μοντέλα **παλινδρόμησης** αναδείχθηκαν τόσο **η συνολική διάρκεια οδήγησης με υπέρβαση του ορίου ταχύτητας και ανοχής ανά μονάδα διάρκειας οδικής διαδρομής χωρίς στάσεις** (sec/sec) όσο και η **διάρκεια οδικής διαδρομής χωρίς στάσεις** (sec). Η διάρκεια χρήσης κινητού τηλεφώνου αυξάνει όσο αυξάνει και μια οδική διαδρομή. Επιπλέον, υπάρχει εξάρτηση μεταξύ της διάρκειας χρήσης κινητού και της διάρκειας οδήγησης με υπέρβαση του ορίου ταχύτητας, καθώς έχει παρατηρηθεί ότι σε οδηγούς, των οποίων η προσοχή αποσπάται με το κινητό τηλέφωνο, αυξάνεται ο χρόνος αντίδρασής τους και οδηγούνται σε πιο απότομη οδηγική συμπεριφορά, άρα και σε υπερβάσεις του ορίου ταχύτητας.

3. Από το παραπάνω συμπέρασμα σε συνδυασμό με τον σχετικό τριγωνικό πίνακα Pearson μπορεί να εξαχθεί έμμεσα το γεγονός ότι όπως και στην ταξινόμηση η **διάρκεια χρήσης κινητού τηλεφώνου** σχετίζεται και με την υπέρβαση των ορίων ταχύτητας, συνεπώς και με την **ταχύτητα**, η οποία αποτελεί τον κρισιμότερο παράγοντα πρόκλησης ατυχημάτων σύμφωνα με την εγχώρια και διεθνή βιβλιογραφία.

4. Στην παρούσα Διπλωματική Εργασία εκπαιδεύτηκαν **δύο αλγόριθμοι ταξινόμησης** και **δύο αλγόριθμοι παλινδρόμησης** τόσο με **όλες** τις υπό εξέταση μεταβλητές όσο και με τις **σπουδαιότερες** εξ' αυτών. Καλύτερος αλγόριθμος για τα μοντέλα ταξινόμησης αναδείχθηκε το μοντέλο Linear Discriminant Analysis, ενώ για την παλινδρόμηση το μοντέλο AdaBoost Regressor έδινε πιο αξιόπιστα αποτελέσματα.

5. Η **συνολική οδηγηθείσα απόσταση** επηρεάζει την διάρκεια χρήσης κινητού τηλεφώνου, όπως αποδεικνύεται από τον τριγωνικό πίνακα Pearson σε συνδυασμό με την σημαντικότητα των μεταβλητών και αυτό εξηγείται από το γεγονός ότι όσο μεγαλύτερη είναι η οδική απόσταση που διανύει ο οδηγός τόσο περισσότερη ώρα διαθέτει για να χρησιμοποιήσει το κινητό του τηλέφωνο. Επιπροσθέτως, με την απόσπαση της προσοχής του οδηγού μέσω του κινητού, μπορεί να ακολουθήσει μια μακρύτερη διαδρομή για να φτάσει στον προορισμό του.

6. Παραδόξως, παρατηρήθηκε ότι η χρήση κινητού τηλεφώνου δεν επηρεάζει σχεδόν καθόλου τα **απότομα περιστατικά** και πιο συγκεκριμένα απότομες επιταχύνσεις και επιβραδύνσεισστους οδηγούς που χρησιμοποιούσαν κινητό. Το γεγονός αυτό ενδεχομένως να οφείλεται στο ότι η ενασχόληση με το κινητό προκαλεί συνήθως μείωση στην ταχύτητα, όπως επιβεβαιώνεται και από την διεθνή βιβλιογραφία(Gazder&Assi, 2021).

7. Σημαντικό ποσοστό των οδηγών, περίπου το **25% εξ' αυτών** που καταγράφηκαν στην βαση δεδομένων της OSeven Telematics, **έκανε χρήση κινητού τηλεφώνου** εν ώρα οδήγηση, μη γνωρίζοντας όμως την δραστηριότητα για την οποία χρησιμοποιούσαν το κινητό τους τηλέφωνο. Αξιοσημείωτο, όμως, παραμένει το γεγονός ότι το ποσοστό είναι αρκετά μικρότερο σε σχέση με τον Ευρωπαϊκό μέσο όρο που κυμαίνεται περίπου στο 33% σύμφωνα με το κεφάλαιο 1(Γράφημα 1.2).

8. Οι **επικίνδυνες ώρες οδήγησης**, δηλαδή το χρονικό διάστημα από 00:00 μέχρι 05:00, φαίνεται να **μην επηρεάζουν** ιδιαίτερα τη **χρήση κινητού** τηλεφώνου. Παρά την γενικότερη εμφάνιση συμπτωμάτων επικίνδυνης οδικής συμπεριφοράς στο συγκεκριμένο χρονικό διάστημα, οι οδηγοί δεν φαίνεται να ασχολούνται ιδιαίτερα με το κινητό τους τηλέφωνο, επομένως πρακτικά δεν οφείλεται εκείνο για την πρόκληση σοβαρών ατυχημάτων εκείνο το χρονικό διάστημα.

9. Παρατηρήθηκε ότι οι αλγόριθμοι **παλινδρόμησης** με τις **καταλληλότερες μεταβλητές** παρουσίασαν **χαμηλότερο** συντελεστή προσδιορισμού $\mathbf{R^2}$ σε σχέση με τους ίδιους αλγόριθμους λαμβάνοντας υπόψη όλες τις υπό εξέταση ανεξάρτητες μεταβλητές. . Το συμπέρασμα αυτό έγκειται στο γεγονός ότι ο συντελεστής R^2 δηλώνει την προβλεπτική ικανότητα του μοντέλου με όσο περισσότερες ανεξάρτητες μεταβλητές.

10. 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.

ΠΙΝΑΚΑΣ ΠΕΡΙΕΧΟΜΕΝΩΝ

ΕΥΡΕΤΗΡΙΟ ΠΙΝΑΚΩΝ

ΕΥΡΕΤΗΡΙΟ ΓΡΑΦΗΜΑΤΩΝ

ΕΥΡΕΤΗΡΙΟ ΕΙΚΟΝΩΝ

1.ΕΙΣΑΓΩΓΗ

1.1 Γενική Ανασκόπηση

Η **οδική ασφάλεια** αποτελεί έναν κρίσιμο τομέα ενδιαφέροντος που διέπει την σημερινή εποχή, η οποία χαρακτηρίζεται από πλήθος συμβάντων τόσο στο αστικό όσο και στο υπεραστικό οδικό δίκτυο. Πλήθος μελετών και σχεδίων ανάπλασης τόσο στην Ελλάδα όσο και στο εξωτερικό αποσκοπούν στην βελτίωση της οδικής ασφάλειας και στην ενίσχυση του αισθήματος της ασφάλειας στους χρήστες της οδού. Αξιοσημείωτο είναι το γεγονός ότι τα τελευταία χρόνια με την ταυτόχρονη αύξηση της κυκλοφορίας στο υφιστάμενο οδικό δίκτυο έχουν προκληθεί αρκετά ατυχήματα που σχετίζονται με παράγοντες όπως η απόσπαση της προσοχής του οδηγού. Σύμφωνα με τον Παγκόσμιο Οργανισμό Υγείας κάθε χρόνο περίπου 1,3 εκατομμύρια άνθρωποι χάνουν την ζωή τους από τροχαία ατυχήματα και 20 έως 50 εκατομμύρια άνθρωποι μένουν σοβαρά τραυματισμένοι ως αποτέλεσμα του ατυχήματος (WHO, 2022). Ειδικότερα, ένας από τους κυριότερους τρόπους απόσπασης της προσοχής του οδηγού αποτελεί η **χρήση του κινητού τηλεφώνου** κατά την διάρκεια της οδήγησης. Όπως απεικονίζεται και στον παρακάτω πίνακα, η ανάγνωση και πληκτρολόγηση μηνυμάτων αυξάνει κατά περίπου 18 φορές τον κίνδυνο ατυχήματος, παρά το γεγονός ότι αυτές οι δραστηριότητες λαμβάνουν χώρα μόλις στο 2% του χρόνου οδήγησης.

Πίνακας 1.1: Επιρροή χρήσης κινητού τηλεφώνου εν ώρα οδήγησης
Πηγή: European Road Safety Observatory (2022)[Available at: https://road-safety.transport.ec.europa.eu/]

Δραστηριότητα	Αύξηση πιθανότητας ατυχήματος	Ποσοστό επί του χρόνου οδήγησης
Συνεχές κράτημα του κινητού εν ώρα οδήγησης	2.05	1.10%
Εύρεση του κινητού τηλεφώνου στο όχημα	4.8	0.58%
Πληκτρολόγηση	12.2	0.14%
Ανάγνωση/Πληκτρολόγηση μηνυμάτων	6.1	1.91%
Πλοήγηση στο διαδίκτυο	2.7	0.73%

Συνεπώς, σύμφωνα με τον παραπάνω πίνακα αποδεικνύεται ότι η χρήση **κινητού τηλεφώνου** δύναται να έχει αρνητικό αντίκτυπο στην συμπεριφορά του οδηγού, καθώς προκαλεί γνωστική,ακουστική,σωματική και οπτική διάσπαση της προσοχής (Ziakopoulos et al.,2016a). Ωστόσο, και τα κινητά τηλέφωνα **με λειτουργία hands-free** επηρεάζουν την οδηγική συμπεριφορά, καθώς παρατηρείται ότι συντελούν με την σειρά τους στην μείωση της συγκέντρωσης του οδηγού, ο οποίος μπορεί να αγνοήσει την σήμανση και την σηματοδότηση της οδού και να οδηγηθεί σε ατύχημα (Caird et al., 2014).Βέβαια, αξίζει να σημειωθεί ότι αρκετές φορές η χρήση κινητού τηλεφώνου με ακουστικά μπορεί να έχει αμφιλεγόμενες ή

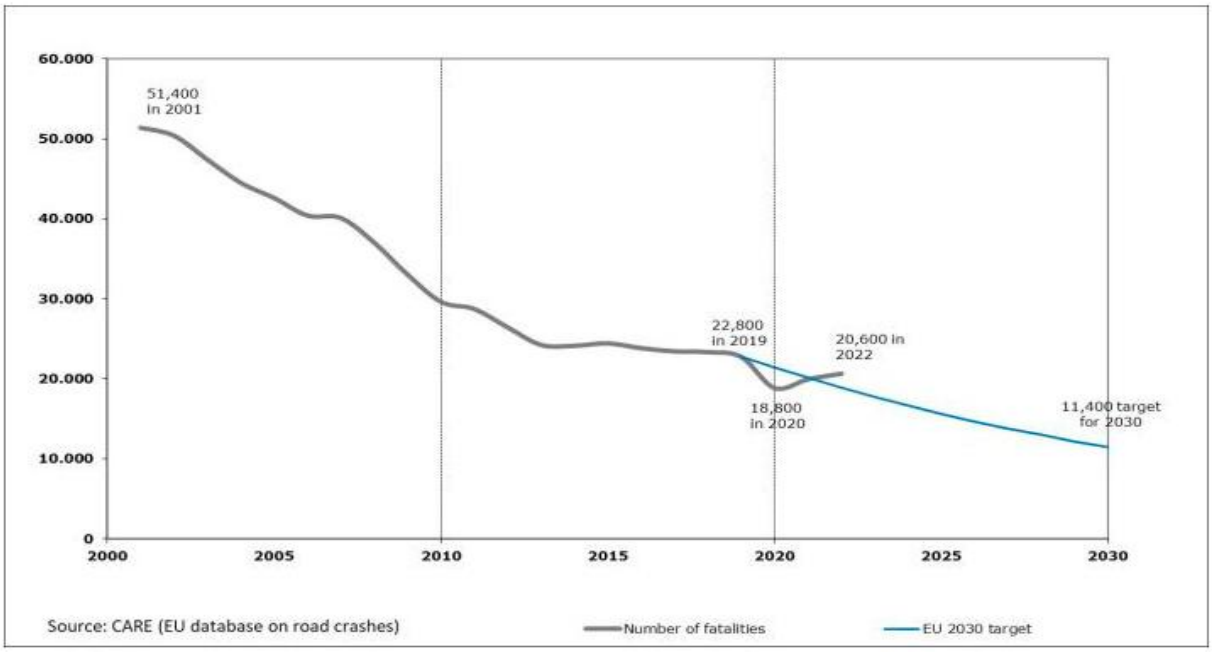
ακόμα και θετικές συνέπειες όσον αφορά στην συμπεριφορά του οδηγού, όπως η μείωση των κρίσιμων οδηγικών καταστάσεων (Ziakopoulos et al., 2018).

Η απόσπαση της προσοχής του οδηγού με την χρήση κινητού τηλεφώνου αποτελεί ένα από τα **πέντε κυριότερα αίτια** για την **πρόκληση ατυχημάτων** μαζί με την ταχύτητα, την οδήγηση υπό την επίρρεια αλκοόλ, τη μη χρήση ζώνης και τη μη χρήση κράνους (DKOA Road-Safety Indroduction,2022). Άλλωστε, από τους τρεις παράγοντες πρόκλησης ατυχημάτων μόνο ο άνθρωπος ευθύνεται σε ποσοστό 65% για την πρόκληση ατυχήματος, ξεπερενώντας την οδό και το όχημα που συνολικά μεταξύ τους ευθύνονται σε ποσοστό μόλις 5,25%, όπως απεικονίζεται στον παρακάτω πίνακα (Γιαννής, 2018).

Πίνακας 1.2 Ποσοστά υπαιτιότητας παραγόντων ατυχημάτων
Πηγή: DKOA Road-Safety Indroduction(2022)

Άνθρωπος Μόνο	65%
Άνθρωπος και Οδός	24%
Άνθρωπος και Όχημα	4.50%
Άνθρωπος, Οδός και Όχημα	1.25%
Οδός Μόνο	2.50%
Οδός και Όχημα	0.25%
Όχημα Μόνο	2.50%

Στο παρακάτω διάγραμμα παρουσιάζεται ο αριθμός των θανάτων από τροχαία ατυχήματα στην ΕΕ από την αρχή του 21^{ου} αιώνα έως το έτος 2022 με πρόβλεψη για το έτος στόχο 2030.

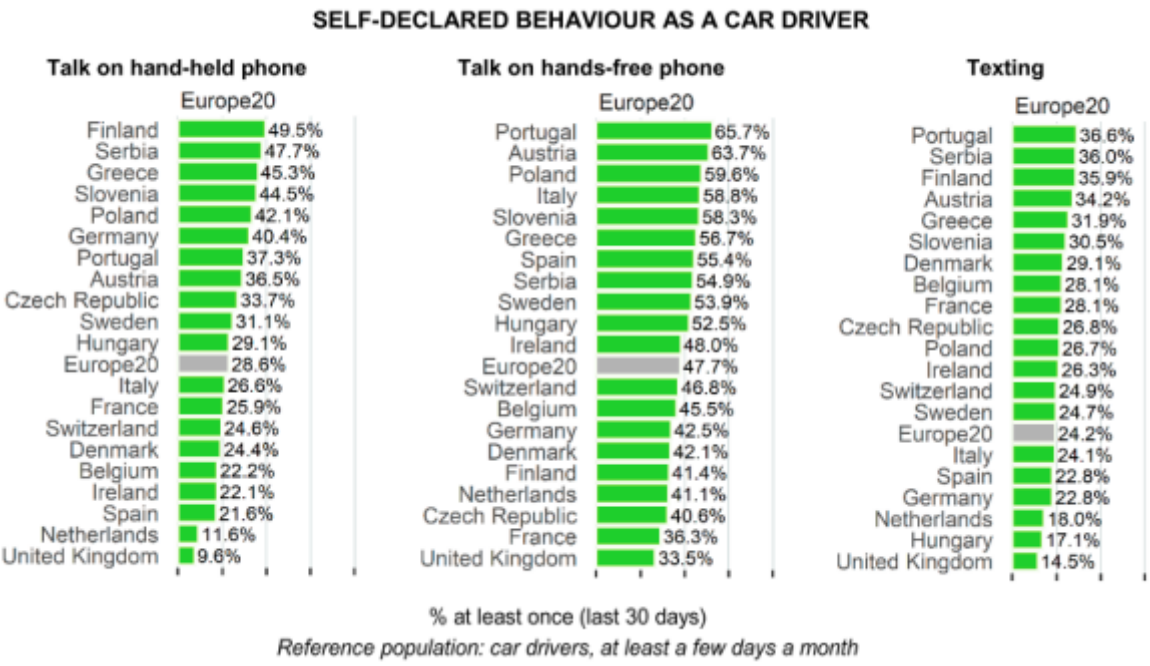


Γράφημα 1.1 Εξέλιξη του αριθμού των θανάτων από τροχαία ατυχήματα στην ΕΕ

Πηγή: CARE (EU database on road crashes) (2022)[Available at: https://transport.ec.europa.eu/]

Ενδιαφέρον αποτελεί το γεγονός ότι το 2020 τα θανατηφόρα ατυχήματα ήταν αρκετά λιγότερα σε σχέση με το προηγούμενο έτος και οφείλεται στην μείωση της κυκλοφορίας εξαιτίας της πανδημίας COVID-19. Επομένως, η αύξηση των ατυχημάτων τα έτη 2021 και 2022 μπορεί να θεωρηθεί πλασματική, καθώς οφείλεται στην απότομη αύξηση της κυκλοφορίας στο οδικό δίκτυο κατά την διακοπή της καραντίνας και την επαναπροσαρμογή στην κανονικότητα.

Σύμφωνα με την έκθεση οδικής ασφάλειας της Ευρωπαϊκής Επιτροπής για το έτος 2022 ανάμεσα σε 20 Ευρωπαικές χώρες προέκυψαν τα ποσοστά των οδηγών που δήλωσαν ότι χρησιμοποίησαν το κινητό τους τηλέφωνο τουλάχιστον μια φορά τις τελευταίες 30 ημέρες είτε για τηλεφωνική επικοινωνία είτε για πληκτρολόγηση μηνυμάτων.



Γράφημα 1.2: Ποσοστά ενασχόλησης με το κινητό τηλέφωνο στην Ευρώπη
Πηγή: Road Safety Observatory (2022)[Available at: https://road-safety.transport.ec.europa.eu/]

Από τα παραπάνω αποτελέσματα προκύπτει η ευρεία χρήση κινητού τηλεφώνου στην Ελλάδα εν ώρα οδήγησης σε σχέση με τον μέσο όρο των ευρωπαικών χωρών που διεξήχθη η έρευνα. Επομένως, το γεγονός αυτό είναι εξαιρετικά κρίσιμο για την οδηγική συμπεριφορά και αξίζει να μελετηθεί εκτενέστερα.

1.2 Στόχος της Διπλωματικής Εργασίας

Στόχο της παρούσας διπλωματικής εργασίας αποτελεί **η διερεύνηση της επιρροής χρήσης του κινητού τηλεφώνου στη συμπεριφορά του οδηγού** και πιο συγκεκριμένα στα μεγέθη της ταχύτητας, της επιτάχυνσης και του χρόνου οδήγησης μέσω της μηχανικής μάθησης ανισόρροπων δεδομένων και αλγορίθμων ταξινόμησης και παλινδρόμησης.

Εικόνα 1.1 Στόχος Διπλωματικής Εργασίας

Ειδικότερα, έγινε καθορισμός του βαθμού που επηρεάζουν οι ανεξάρτητες μεταβλητές τις εξαρτημένες με την διαδικασία της **Επιλογής Χαρακτηριστικών** (Feature Selection) και επιλογή των σημαντικότερων ανεξάρτητων μεταβλητών (Feature Importance).Στην συνέχεια, αναπτύχθηκαν μοντέλα ταξινόμησης και παλινδρόμησης, ώστε να προβλεφθεί ο βαθμός που επηρεάζει η χρήση κινητού τηλεφώνου τις ανεξάρτητες μεταβλητές, όπως θα αναλυθεί και στο κεφάλαιο 4. Η ταξινόμηση μέσω της μηχανικής μάθησης αποτελεί ένα ουσιαστικό εργαλείο για την αναγνώριση της οδηγικής συμπεριφοράς και έμμεσα για την βελτίωση της οδικής ασφάλειας (Wu et al.,2016).

Η **πρόβλεψη** του πλήθους των οδηγών που χρησιμοποιούν κινητό τηλέφωνο με τα μοντέλα ταξινόμησης θα δώσει άμεσα μια εικόνα του προβλήματος της χρήσης κινητού τηλεφώνου κατά την διάρκεια της οδήγησης. Επιπροσθέτως, η ανάλυση της διάρκειας χρήσης κινητού τηλεφώνου με μοντέλα παλινδρόμησης θα επιβεβαιώσει και θα ενισχύσει τα συμπεράσματα σχετικά με την επιρροή των παραγόντων οδήγησης.

Συνεπώς, η συνεισφορά της παρούσας εργασίας έγκειται στο γεγονός πως θα επιχειρήσει να **διευρύνει** την υφιστάμενη γνώση στον τομέα της συμπεριφοράς του οδηγού και το πώς η τελευταία επηρεάζεται από την χρήση κινητής συσκευής.

1.3 Μεθοδολογία της Διπλωματικής Εργασίας

Στο συγκεκριμένο υποκεφάλαιο περιγράφεται συνοπτικά η **μεθοδολογία** που ακολουθήθηκε προκειμένου να επιτευχθεί ο στόχος της παρούσας διπλωματικής εργασίας.

Σε πρώτο στάδιο αφότου προσδιορίστηκε και **οριστικοποιήθηκε η θεματική ενότητα**, στην οποία στηρίζεται η διπλωματική εργασία, καθορίστηκε ο **στόχος** της μελέτης.

Στην συνέχεια, πραγματοποιήθηκε **εκτενής έρευνα** στην ήδη **υπάρχουσα βιβλιογραφία** σχετική με το εξεταζόμενο αντικείμενο με οριζόντια ανασκόπηση των ερευνών, σύγκρισή τους και παράθεση των αποτελεσμάτων τους. Στόχος της μελέτης αυτής αποτέλεσε η κατά το δυνατόν μεγαλύτερη συλλογή στοιχείων, ώστε να υπάρξει βαθύτερη εξοικείωση με το συγκεκριμένο ερευνητικό πεδίο.

Σε επόμενο στάδιο πραγματοποιήθηκε η **συλλογή** και η **επεξεργασία** των στοιχείων.Τα στοιχεία που συλλέχθηκαν προήλθαν από βάση δεδομένων της εταιρίας OSeven Telematics και συλλέχθηκαν σε πραγματικές οδικές συνθήκες μέσω των κινητών τηλεφώνων των οδηγών. Η επεξεργασία των μοντέλων ταξινόμησης και παλινδρόμησης που προέκυψαν από την στατιστική ανάλυση πραγματοποιήθηκε με την βοήθεια της γλώσσας προγραμματισμού Python (συγκεκριμένα μέσω του προγραμματιστικού πακέτου Pycaret) με την αξιοποίηση των κατάλληλων βιβλιοθηκών προσαρμοσμένων στην συγκεκριμένη γλώσσα προγραμματισμού, δηλαδή της βιβλιοθήκης μηχανικής μάθησης scikit-learn και των βιβλιοθηκών ανάλυσης δεδομένων pandas και numpy.

Στο τελευταίο στάδιο, αξιολογήθηκαν τα αποτελέσματα και εξήχθησαν χρήσιμα **συμπεράσματα και προτάσεις για περαιτέρω έρευνα** στο συγκεκριμένο ερευνητικό πεδίο.

Παρακάτω παρουσιάζεται το διάγραμμα ροής της συγκεκριμένης διπλωματικής εργασίας.

Γράφημα 1.3 :Διάγραμμα Ροής Διπλωματικής Εργασίας

1.4 Δομή της Διπλωματικής Εργασίας

Η παρούσα διπλωματική εργασία απαρτίζεται από 7 κεφάλαια, τα οποία περιγράφονται συνοπτικά παρακάτω.

Το **πρώτο κεφάλαιο** αποτελεί την **εισαγωγή** στην γενικότερη θεματική ενότητα που πραγματεύεται η συγκεκριμένη διπλωματική εργασία, ώστε να καταστεί αντιληπτό και εύληπτο στον αναγνώστη το ερευνητικό πεδίο που εξετάζεται. Ειδικότερα, παρατίθενται στατιστικά στοιχεία από έρευνες σχετικές με την οδική ασφάλεια και τους παράγοντες πρόκλησης ατυχημάτων και πραγματοποιείται ομαλή μετάβαση στο κυρίως θέμα, την επιρροή της χρήσης του κινητού τηλεφώνου στην οδηγική συμπεριφορά.Τέλος, περιγράφεται ο στόχος, η μεθοδολογία που ακολουθήθηκε για την επίτευξή του, καθώς και η δομή της διπλωματικής εργασίας.

Στο **δεύτερο κεφάλαιο** περιλαμβάνεται η **βιβλιογραφική ανασκόπηση**, στην οποία παρουσιάζονται περιληπτικά σχετικές έρευνες με το αντικείμενο της διπλωματικής εργασίας προερχόμενες τόσο από την ελληνική όσο και από την διεθνή επιστημονική κοινότητα.

Στο **τρίτο κεφάλαιο** αναφέρεται το **θεωρητικό υπόβαθρο** της παρούσας έρευνας. Πιο συγκεκριμένα, πραγματοποιείται ανάλυση των βασικών εννοιών της στατιστικής, αναφέρονται τα μαθηματικά πρότυπα που αξιοποιήθηκαν για την εκπόνηση της παρούσας εργασίας, καθώς και τα κριτήρια αποδοχής μοντέλων ταξινόμησης και παλινδρόμησης. Στο παρόν κεφάλαιο περιγράφονται και οι μέθοδοι με τις οποίες εξισορροπούνται τα ανομοιογενή δεδομένα.

Στο **τέταρτο κεφάλαιο** περιγράφονται τα **δεδομένα** που χρησιμοποιούνται και η διαδικασία **συλλογής** τους από την βάση δεδομένων της εταιρίας OSeven Telematics. Ακολούθως γίνεται επεξεργασία των δεδομένων αυτών και κρίνεται ποια εξ' αυτών καθίστανται χρήσιμα για περαιτέρω επεξεργασία και ανάλυση.

Το **πέμπτο κεφάλαιο** αποτελεί τον βασικό πυλώνα της διπλωματικής εργασίας, καθώς σε αυτό περιλαμβάνεται η αναλυτική παρουσίαση της **μεθοδολογίας ανάπτυξης των μοντέλων**.Το συγκεκριμένο κεφάλαιο χωρίζεται σε 2 ενότητες, της ταξινόμησης και της παλινδρόμησης.

Στο **έκτο κεφάλαιο** αναφέρονται τα **συμπεράσματα** που απορρέουν από τα αποτελέσματα της ανάλυσης των μοντέλων. Στην συνέχεια, παρουσιάζονται μερικές προτάσεις για περαιτέρω έρευνα.

Το **έβδομο κεφάλαιο** αποτελεί το τελευταίο κεφάλαιο της παρούσας διπλωματικής εργασίας και σε αυτό παρατίθενται με αλφαβητική σειρά οι **βιβλιογραφικές αναφορές** που αξιοποιήθηκαν.

2. ΒΙΒΛΙΟΓΡΑΦΙΚΗ ΑΝΑΣΚΟΠΗΣΗ

2.1 Εισαγωγή

Στο συγκεκριμένο κεφάλαιο παρουσιάζονται συναφείς έρευνες και μεθοδολογίες σχετικές με το αντικείμενο της παρούσας διπλωματικής εργασίας. Ειδικότερα, αναζητήθηκαν στην εγχώρια και διεθνή βιβλιογραφία δημοσιευμένες έρευνες, οι οποίες αναλύουν την οδηγική συμπεριφορά, βασικό πυλώνα του ερευνητικού πεδίου της Οδικής Ασφάλειας. Μελετήθηκαν έρευνες με δεδομένα τόσο σε πραγματικές συνθήκες οδήγησης όσο και σε σνθήκες προσομοίωσης, όπως επίσης και η ικανότητα πρόβλεψης συμπεριφοράς των μοντέλων Μηχανικής Μάθησης.

Το παρόν κεφάλαιο αποτελεί την αφορμή για την εκπόνηση του συγκεκριμένου ερευνητικού έργου, καθώς με βάση τις έρευνες και τις ελλείψεις που εκείνες παρουσιάζουν, διαμορφώνεται ο **στόχος** αυτής της μελέτης και προσδιορίζεται η κατάλληλη **μεθοδολογία** για την επίτευξή του. Κατά την ανάγνωση και σύγκριση των μελετών έγινε αντιληπτό ότι στην πλειοψηφία τους παρατηρείται το πρόβλημα της άνισης κατανομής των δεδομένων στις διαφορετικές τάξεις, που αποτελεί μεγάλη πρόκληση σε προβλήματα Μηχανικής Μάθησης. Συνεπώς, θα πρέπει να να παρουσιαστούν αναλυτικά οι τρόποι αντιμετώπισης του προβλήματος των ανισόρροπων δεδομένων, δηλαδή τεχνικές επαναδειγματοληψίας που έχουν εφαρμοστεί σε παλαιότερες έρευνες.

2.2 Συναφείς έρευνες και μεθοδολογίες

Στο ερευνητικό έργο των **Ziakopoulos et al. (2023)** διερευνώνται οι παράγοντες επιρροής της απόσπασης της προσοχής του οδηγού μέσω της χρήσης κινητού τηλεφώνου σε πραγματικές συνθήκες οδήγησης. Η συλλογή των στοιχείων έγινε τόσο με την χρήση ερωτηματολογίου όσο και μέσω εφαρμογής στο κινητό που κατέγραφε τα οδηγικά χαρακτηριστικά σε συνθήκες φυσικής οδήγησης. Τα δεδομένα της εφαρμογής προέκυψαν από δείγμα 230 οδηγών διαφόρων τύπων οχημάτων και το πείραμα διήρκησε 21 μήνες. Αξίζει να σημειωθεί ότι τέθηκε ένα ελάχιστο όριο 40 διαδρομών στο προαναφερθέν χρονικό διάστημα ως απαραίτητη προϋπόθεση για την προσμέτρηση των οδηγών στα αποτελέσματα του πειράματος. Η διεξαγωγή του πειράματος εκτελέστηκε σε 6 φάσεις και χρησιμοποιήθηκε ο αλγόριθμος ταξινόμησης EXtreme Gradient Boosting (XGBoost). Αφού εφαρμόστηκε η μέθοδος της υπερδειγματοληψίας λόγω του ότι η χρήση κινητού τηλεφώνου αποτελούσε την μειονοτική τάξη εκ των δύο, το μοντέλο της Ενίσχυσης Κλίσης παρουσίασε ικανοποιητική προβλεπτική ικανότητα και με την βοήθεια της ανάλυσης ευαισθησίας (SHAP) για τις έξι σημαντικότερες μεταβλητές αποδείχθηκε ότι η συνολική διανυθείσα απόσταση αποτελεί την σημαντικότερη μεταβλητή. Μέσω του ερωτηματολογίου προέκυψε ότι η χρήση κινητού τηλεφώνου οδηγεί κατά κόρον στην απόσπαση της προσοχής του οδηγού και εξήχθη το συμπέρασμα ότι οι οδηγοί υπερεκτιμούν τις οδηγικές ικανότητές τους, αγνοώντας τις συνέπειες που μπορεί να έχει έστω και μια ελάχιστη εστίαση στο κινητό αντί για την οδό.

Η έρευνα των **Ghandour et al. (2021)** στηρίζεται στην μελέτη της οδηγικής συμπεριφοράς και των διαφορετικών ψυχολογικών συνθηκών του οδηγού. Η μεθοδολογία των Ghandour et al.,

περιλαμβάνει την ανάπτυξη μοντέλων ταξινόμησης μηχανικής μάθησης όπως ταξινόμηση Λογιστικής Παλινδρόμησης (Logistic Regression), Τυχαία Δάση (Random Forests), Τεχνητών Νευρωνικών Δικτύων και Ενίσχυσης Κλίσης (Gradient Boosting), με τις επιμέρους κλάσεις ταξινόμησης να διαχωρίζονται σε τρεις τάξεις, στην ομαλή, επιθετική και νυσταγμένη συμπεριφορά, εν αντιθέσει με την έρευνα των Ziakopoulos et al. (2023), όπου έγινε ταξινόμηση σε δύο τάξεις. Η έρευνα Ghandour et al. πραγματοποιήθηκε ξεχωριστά για δύο βάσεις δεδομένων διαφορετικής προέλευσης. Στην μία βάση εμπεριέχονται δεδομένα που αφορούν στην ανίχνευση λωρίδας με στοιχεία σχετικά με την θέση και κατεύθυνση του οχήματος, ενώ η δεύτερη περιλαμβάνει δεδομένα συνθηκών φόρτου, με στοιχεία που αφορούν στο περιβάλλον του οχήματος. Από τις δύο βάσεις δεδομένων ακριβέστερη στις τρεις ψυχολογικές τάξεις αποδεικνύεται αυτή των συνθηκών φόρτου, με την ταξινόμηση Ενίσχυσης Κλίσης να δίνει ουσιαστικά συμπεράσματα για τον προσδιορισμό της επικινδυνότητας.

Η ερευνητική εργασία του **Jaswanth Reddy Rekkala (2021)** που εκπονήθηκε για λογαριασμό του California State University έχει σκοπό να διερευνήσει την απόδοση του νευρωνικού δικτύου CNN (Convolutional Neural Network) στον εντοπισμό της χρήσης κινητών τηλεφώνων από τους οδηγούς. Σε αντίθεση με τις προηγούμενες μεθόδους, η συλλογή δεδομένων διεξάγεται σε εργαστηριακά εξαρτώμενα οδηγικά δοκιμαστικά περιβάλλοντα, ώστε προσφερθούν όσο το δυνατόν πιο ρεαλιστικές εμπειρίες οδήγησης. Εφαρμόζεται ταξινόμηση σε δέκα επίπεδα απόσπασης της προσοχής του οδηγού και αποδεικνύεται ότι η βαθιά εκμάθηση αποτελεί μία από τις καταλληλότερες μεθόδους για τον προσδιορισμό του επιπέδου απόσπασης προσοχής των οδηγών (ακρίβεια μοντέλου CNN 96,7%). Αξίζει να σημειωθεί ότι το παραπάνω νευρωνικό δίκτυο προσφέρεται κυρίως για δεδομένα που μπορεί να περιλαμβάνουν εικόνες ή βίντεο και συντελεί στην αναγνώρισή τους.

Η μελέτη των **Gazder & Assi (2021)** έχει στόχο να **προσδιορίσει τους παράγοντες απόσπασης προσοχής**, οι οποίοι θεωρούνται πιο επικίνδυνοι για την οδηγική συμπεριφορά, καθώς και τις επιδράσεις εκείνων και της ηλικίας του οδηγού στην ταχύτητα. Πραγματοποιήθηκε έρευνα ερωτηματολογίου, ώστε να διαπιστωθεί η γνώμη των οδηγών για τον πιο επικίνδυνο παράγοντα απόσπασης προσοχής. Στην έρευνα συμμετείχαν 639 συμμετέχοντες και παράλληλα έλαβαν χώρα επιτόπιες μετρήσεις στην οδό της ταχύτητας για 48 οδηγούς, της απόστασης για 36 οδηγούς, της ηλικίας και του τρόπου απόσπασης της προσοχής του οδηγού. Σύμφωνα με τα αποτελέσματα του ερωτηματολογίου πιο επικίνδυνοι παράγοντες χαρακτηρίστηκαν και σε αυτή την έρευνα η χρήση κινητού τηλεφώνου, ο χειρισμός των παιδιών στο όχημα, καθώς και ενδεχόμενα συμβάντα ή τροχαία ατυχήματα που έχουν πραγματοποιηθεί στο οδικό δίκτυο. Για την στατιστική ανάλυση των δεδομένων χρησιμοποιήθηκε το τεστ ανάλυσης διακύμανσης διπλής κατεύθυνσης (ANOVA) σε συνδυασμό με την ανάλυση παλινδρόμησης. Τα αποτελέσματα της μεθόδου έδειξαν ότι τόσο η χρήση κινητού τηλεφώνου όσο και η ηλικία επιδρούν στην απόσπαση της προσοχής του οδηγού. Πιο συγκεκριμένα, η επίδραση της χρήσης κινητού στην ταχύτητα και την απόσταση ήταν μεγαλύτερη σε σχέση με την επίδραση της ηλικίας στις παραπάνω μεταβλητές. Κρισιμότερος τρόπος απόσπασης προσοχής αναδείχθηκε η πληκτρολόγηση μηνυμάτων. Επιπροσθέτως, αξιοσημείωτο αποτελεί το γεγονός ότι με την απόσπαση της προσοχής μειώνεται η ταχύτητα για όλες τις ηλικιακές ομάδες του δείγματος.

Προκειμένου να ανιχνευθεί η έλλειψη προσοχής του οδηγού οι **Zhang et al. (2020)** εισήγαγαν ένα δίκτυο βαθιάς μη επιβλεπόμενης πολυτροπικής συγχώνευσης, το οποίο ονομάζεται UMMFN. Πρόκειται για ένα μοντέλο που απαρτίζεται από τρεις ενότητες, οι οποίες είναι η εκμάθηση πολυτροπικής αναπαράστασης, η σύντηξη χαρακτηριστικών πολλαπλών κλιμάκων

και ανίχνευση της προσοχής του οδηγού χωρίς επίβλεψη. Η πρώτη ενότητα είναι η εκμάθηση χαμηλής διάστασης αναπαράστασης πολλαπλών ετερογενών αισθητήρων με τη χρήση ενσωμάτωσης υποδικτύων. Στόχο της σύντηξης χαρακτηριστικών πολλαπλών κλιμάκων αποτελεί η εκμάθηση τόσο της χρονικής όσο και της χωρικής εξάρτησης από διαφορετικές λειτουργίες. Η τελευταία ενότητα χρησιμοποιεί ένα ConvLSTM κωδικοποιητή-αποκωδικοποιητή για να εκτελέσει ένα έργο ταξινόμησης χωρίς επίβλεψη που δεν επηρεάζεται από νέους τύπους συμπεριφορών του οδηγού. Κατά τη φάση της ανίχνευσης, μπορεί να ληφθεί μια λεπτομερής απόφαση ανίχνευσης μέσω του υπολογισμού του σφάλματος ανακατασκευής του UMMFN ως μετρική αξιολόγησης για κάθε καταγεγραμμένο δεδομένο δοκιμής. Εμπειρικά συγκρίνεται η προσέγγιση αυτή με διάφορες σύγχρονες μεθόδους στο συγκεκριμένο πολυτροπικό σύνολο δεδομένων για απόσπαση προσοχής κατά την οδήγηση, με τα πειραματικά αποτελέσματα να αποδεικνύουν ότι η UMMFN έχει ανώτερη απόδοση σε σχέση με τις υπάρχουσες προσεγγίσεις.

Οι **Phuksuksakul et al. (2021)** ανέλυσαν τους παράγοντες, οι οποίοι επηρεάζουν την συμπεριφορά της χρήσης κινητού τηλεφώνου κατά την οδήγηση, καθώς και τις επιπτώσεις της χρήσης κινητού στην οδηγική απόδοση, δηλαδή στην ταχύτητα, την πλευρική θέση, την απόσταση ακολούθησης, τον χρόνο αντίληψης-αντίδρασης και την κατάσταση εμφάνισης παρ' ολίγον ατυχήματος. Η συλλογή των στοιχείων προέκυψε από έρευνα ερωτηματολογίου σε 1106 ερωτηθέντες από τέσσερις διαφορετικές περιοχές της Ταϊλάνδης και μέσω της Θεωρίας της Σχεδιασμένης Συμπεριφοράς (Planned Behaviour Theory-TPB) έγινε η ερμηνεία αυτών των παραγόντων, με την προσθήκη παραγόντων που αφορούν στην αντίληψη του κινδύνου και την γνώση της νομοθεσίας. Παρόλο που περίπου το 90 % των οδηγών συνειδητοποίησαν ότι η χρήση κινητού τηλεφώνου κατά την οδήγηση ήταν επικίνδυνη και παράνομη, ανέφεραν ότι εξακολουθούν να χρησιμοποιούν το κινητό τηλέφωνο κατά την οδήγηση. Για να προσδιοριστεί η επίδραση της χρήσης κινητού τηλεφώνου στις οδηγικές επιδόσεις, έγινε προσομοίωση σε μια ευθύγραμμη επαρχιακή οδό δύο λωρίδων συνολικά, με ένα προπορευόμενο όχημα και μια απροσδόκητη πινακίδα "STOP", προκειμένου να εξεταστούν οι επιδόσεις οδήγησης των οδηγών χωρίς τηλεφώνο, μιλώντας σε τηλεφωνική κλήση και στέλνοντας μήνυμα. Τα αποτελέσματα απέδειξαν ότι η χρήση κινητού τηλεφώνου κατά την οδήγηση μπορεί να μειώσει την ταχύτητα και την απόσταση ακολούθησης, αλλά να αυξήσει την πλευρική απόκλιση, την απόκλιση διεύθυνσης, την ταχύτητα διεύθυνσης, το χρόνο αντίληψης-αντίδρασης και τον αριθμό των παρ' ολίγον ατυχημάτων που οδηγούν σε υψηλότερους κινδύνους πρόκλησης τροχαίων ατυχημάτων.

2.3 Κριτική Ανάλυση-Οριζόντια Ανασκόπηση

Στον Πίνακα 2.1 επιχειρείται να γίνει μια οριζόντια ανασκόπηση προς ευκολότερη ανάγνωση και σύγκρισαη των ερευνών και μεθοδολογιών που αναλύθηκαν στο προηγούμενο υποκεφάλαιο. Ειδικότερα, παρατίθενται συνοπτικά το όνομα των ερευνητών, ο στόχος της έρευνας, ο τρόπος συλλογής των στοιχείων, η μέθοδος επεξεργασίας τους, τα αποτελέσματα, οι ελλείψεις και τυχόν προτάσεις για περαιτέρω έρευνα.

Πίνακας 2.1:Ανασκόπηση και Σύγκριση ερευνών και μεθοδολογιών

Έρευνα	Στόχος έρευνας	Τρόπος συλλογής στοιχείων	Μέθοδος επεξεργασίας δεδομένων	Συμπεράσματα	Ελλείψεις	Προτάσεις για περαιτέρω έρευνα
Ziakopoulos et al., 2023	Διερεύνηση παραγόντων επιρροής απόσπασης προσοχής οδηγού με χρήση κινητού.	Ερωτηματολόγιο σε συνδυασμό με δεδομένα από πραγματικές οδικές συνθήκες.	Υπερδειγμα-τοληψία, ταξινόμηση σε δύο τάξεις μέσω μηχανικής μάθησης και ανάλυση ευαισθησίας.	Η χρήση κινητού τηλεφώνου οδηγεί κατά κόρον στην απόσπαση προσοχής του οδηγού.	Ερωτηματολόγια: Μειωμένη αντικειμενικότητα Εφαρμογή στο κινητό: Μη δυνατή εξέταση της στατιστικής σχέσης χρήσης κινητού με χαμηλότερες ταχύτητες.	Αξιοποίηση ευφυών συστημάτων μεταφορών ITS και καμερών, που θα μπορούσαν να αναλύουν ανώνυμα και να ανταποκρίνονται στην απόσπαση της προσοχής από την οδήγηση σε σε πραγματικό χρόνο βασιζόμενες περισσότερο σε μετρήσεις πεδίου παρά σε απαντήσεις σε ερωτηματολόγια.
Ghandour et al., 2021	Μελέτη της οδηγικής συμπεριφοράς και των διαφορετικών ψυχολογικών συνθηκών του οδηγού.	Δεδομένα από πραγματικές οδικές συνθήκες.	Εφαρμογή ταξινόμησης σε τρεις ψυχολογικές τάξεις μέσω αλγορίθμων μηχανικής μάθησης και νευρωνικών δικτύων.	Από τις δύο βάσεις δεδομένων ακριβέστερη στις τρεις ψυχολογικές τάξεις αποδεικνύεται αυτή των συνθηκών φόρτου, με την ταξινόμηση Ενίσχυσης Κλίσης να δίνει ουσιαστικά	Οι μελετητές δεν λαμβάνουν υπόψη ορισμένα πρόσθετα χαρακτηριστικά της οδού και της ψυχικής κατάστασης του οδηγού. Επιπροσθέτως, η έρευνα περιορίζεται στην ανάλυση ενός	Με στόχο την βελτίωση των αποτελεσμάτων της ταξινόμησης προτείνεται σε μελλοντική έρευνα να ληφθούν υπόψη πρόσθετοι παράγοντες, όπως το όριο ταχύτητας της οδού και ο ψυχικός φόρτος εργασίας του οδηγού. Επίσης προτείνεται η ανάπτυξη ενός

				συμπεράσματα για τον προσδιορισμό της επικινδυνότητας.	τύπου οδού.	συστήματος ταξινόμησης βασισμένου στον συνδυασμό πολλαπλών μεθόδων.
Rekkala, 2021	Διερεύνηση απόδοσης νευρωνικού δικτύου CNN στον εντοπισμό της χρήσης κινητών τηλεφώνων από τους οδηγούς.	Πείραμα στο εργαστήριο-προσομοίωση.	Εφαρμόζεται ταξινόμηση σε δέκα επίπεδα απόσπασης της προσοχής του οδηγού.	Η βαθιά εκμάθηση αποτελεί μία από τις καταλληλότερες μεθόδους για τον προσδιορισμό του επιπέδου απόσπασης προσοχής των οδηγών.	Αδυναμία ενσωμάτωσης πρόσθετων ρυθμίσεων κάμερας για την ανίχνευση της χρήσης κινητού τηλεφώνου.	Σχέδιο δημιουργίας ενός αποδοτικού και αποτελεσματικού πλαισίου CNN που ενσωματώνει τις πρόσθετες ρυθμίσεις κάμερας και με κατάλληλη προεπεξεργασία να ανιχνεύει την χρήση κινητού τηλεφώνου.
Gazder & Assi, 2021	Προσδιορισμός κρισιμότερων για την οδηγική συμπεριφορά παραγόντων απόσπασης προσοχής, καθώς και επιδράσεις εκείνων και της ηλικίας του οδηγού στην ταχύτητα.	Ερωτηματολόγιο.	Εφαρμόζεται ανάλυση παλινδρόμησης σε συνδυασμό με ANOVA test.	1.Μείωση της ταχύτητας με την χρήση κινητού. 2.Κρισιμότερος τρόπος απόσπασης προσοχής αναδείχθηκε η πληκτρολόγηση μηνυμάτων.	1.Περιορισμένες πληροφορίες από τη Γενική Διεύθυνση Τροχαίας σχετικά με τις αιτίες των παραβάσεων και των ατυχημάτων. 2.Μη αποτελεσματική καταγραφή οδηγικών χαρακτηριστικών για οδηγούς κινούμενους με μεγάλη ταχύτητα.	Δεδομένου ότι οι αφηρημένοι οδηγοί τείνουν να γίνονται προσεκτικοί μειώνοντας την ταχύτητά τους, ως εκ τούτου, η μεγαλύτερη εμπλοκή τους σε ατυχήματα πρέπει να μελετηθεί και από άλλες οπτικές γωνίες.

Zhang et al., 2020	Ανίχνευση της προσοχής του οδηγού χωρίς επίβλεψη.	Πείραμα μέσω αισθητήρων σε πραγματικές συνθήκες οδήγησης.	Χρήση νευρωνικών δικτύων και ταξινόμηση.	Η UMMFN έχει ανώτερη απόδοση σε σχέση με την τις υπάρχουσες προσεγγίσεις ανίχνευσης συμπεριφοράς οδηγού.	Τα αποτελέσματα ταξινόμησης της τάξης 'κανονικής οδήγησης' περιορίζονται από την ποικιλομορφία των δειγμάτων.	Βελτίωση των αποτελεσμάτων με αύξηση του πλήθους και της ποικιλομορφίας των δειγμάτων.
Phuksuksakul et al., 2021	Διερεύνηση παραγόντων επιρροής της χρήσης κινητού κατά την οδήγηση και επιπτώσεις στην απόδοση του οδηγού.	Έρευνα ερωτηματολογίου και προσομοίωση σε επαρχιακή οδό.	Εφαρμογή της Θεωρίας Σχεδιασμένης Συμπεριφοράς και διαχωρισμός των οδηγών σε τρεις κατηγορίες για την εύρεση των επιδόσεών τους.	Η χρήση κινητού τηλεφώνου κατά την οδήγηση μπορεί να μειώσει την ταχύτητα και την απόσταση ακολούθησης, αλλά να αυξήσει την πλευρική απόκλιση, την απόκλιση διεύθυνσης, την ταχύτητα διεύθυνσης, το χρόνο αντίληψης-αντίδρασης και τον αριθμό των παρ' ολίγον ατυχημάτων που οδηγούν σε υψηλότερους κινδύνους πρόκλησης ατυχημάτων.	1.Αποκλίσεις προσομοίωσης από τις πραγματικές συνθήκες οδήγησης. 2.Απουσία κυκλοφορίας στο αντίθετο ρεύμα της οδού δύο λωρίδων στην προσομοίωση. 3.Γενικό ερώτημα στο ερωτηματολόγιο για την χρήση κινητού, χωρίς να γίνει γνωστή η δραστηριότητα για την οποία χρησιμοποιήθηκε το κινητό εν ώρα οδήγησης.	1.Ρεαλιστικότερη προσομοίωση με εισαγωγή κυκλοφορίας και στο αντίθετο ρεύμα. 2.Διευκρύνηση των ερωτημάτων σχετικών με την δραστηριότητα για την οποία ο οδηγός χρησιμοποιεί το κινητό του τηλέφωνο.

2.4 Σύνοψη

Σύμφωνα με τις βιβλιογραφικές έρευνες που αναλύθηκαν διεξοδικά παραπάνω και αφορούσαν στην οδηγική συμπεριφορά και κυρίως στο κρίσιμο ζήτημα της απόσπασης της προσοχής του οδηγού, γίνεται αντιληπτό ότι η **χρήση κινητού τηλεφώνου** αποτελεί **πρωταρχικό παράγοντα απόσπασης της προσοχής**. Για αυτό τον λόγο κρίνεται αναγκαίο να εξεταστεί περαιτέρω στην παρούσα διπλωματική εργασία. Επιπλέον, διαπιστώθηκε ότι σε αρκετές έρευνες χρησιμοποιήθηκε η μεθοδολογία της **υπερδειγματοληψίας** (Ziakopoulos et al., 2023), η οποία ενδείκνυται όταν τα δεδομένα ενός προβλήματος καθίστανται ανισόρροπα και συναντάται συχνά σε προβλήματα Μηχανικής Μάθησης. Τα δεδομένα προήλθαν είτε από πραγματικές συνθήκες κυκλοφορίας είτε εργαστηριακά μέσω προσομοίωσης, αλλά παράλληλα προέκυψαν και δεδομένα μέσω ερωτηματολογίων, τα οποία κρίθηκαν σε πολλές περιπτώσεις ελλιπή και μη αντικειμενικά (Phuksuksakul et al., 2021; Ziakopoulos et al., 2023). Αξίζει να σημειωθεί ότι στις παρούσες έρευνες χρησιμοποιήθηκε πλήθος μεθοδολογιών στατιστικής ανάλυσης με κυριότερες την **ταξινόμηση** σε κλάσεις (Ziakopoulos et al., 2023; Ghandour et al., 2021;Zhang et al,2021), την **παλινδρόμηση** (Gazder & Assi, 2021) και την **βαθιά εκμάθηση** με νευρωνικά δίκτυα (Rekkala, 2021; Zhang et al., 2020).

Από τις προαναφερθείσες μεθόδους επιλέχθηκε η **ταξινόμηση** και η **παλινδρόμηση** για την εκπόνηση της διπλωματικής εργασίας, καθώς στόχος ήταν η ανάλυση των δεδομένων σε δύο τάξεις, στο πλήθος των οδηγών που χρησιμοποιούν το κινητό τηλέφωνο εν ώρα οδήγησης και στο πλήθος των οδηγών που δεν κάνουν χρήση κινητού.

3.ΘΕΩΡΗΤΙΚΟ ΥΠΟΒΑΘΡΟ

3.1 Εισαγωγή

Στο συγκεκριμένο κεφάλαιο αναλύονται διεξοδικά τα βασικά **θεωρητικά ζητήματα** που πραγματεύεται η συγκεκριμένη διπλωματική εργασία.Ειδικότερα,περιγράφεται η στατιστική και μαθηματική ανάλυση η οποία έγινε ώστε να επιτευχθεί το παρόν ερευνητικό έργο. Η στατιστική έρευνα αποτελεί ένα άκρως χρήσιμο εργαλείο αποτύπωσης, απόδειξης, κατανόησης και ερμηνείας διαφόρων φαινομένων. Οι μέθοδοι που επιλέχθηκαν για την στατιστική ανάλυση είναι αυτές της **ταξινόμισης** (classification) και **παλινδρόμισης** (regression). Παρακάτω παρουσιάζονται οι βασικές έννοιες που διέπουν την στατιστική ανάλυση.

3.2 Βασικές έννοιες στατιστικής

Με τον όρο **στατιστικός πληθυσμός (population)** νοείται κάθε σύνολο παρατηρήσεων του χαρακτηριστικού που ενδιαφέρει τη στατιστική έρευνα. Συνήθως το σύνολο των παρατηρήσεων είναι πεπερασμένο, αλλά υπάρχουν περιπτώσεις που το μέγεθος των πληθυσμών είναι τόσο ευρύ ώστε να μην μπορούν να απογραφούν. Ένας πληθυσμός μπορεί να είναι είτε πραγματικός είτε θεωρητικός.

Ο όρος **δείγμα (sample)** αναφέρεται σε ένα αντιπροσωπευτικό υποσύνολο του πληθυσμού τον οποίο η έρευνα μελετά. Συνεπώς, όλα τα στοιχεία που ανήκουν στο δείγμα ανήκουν και στον πληθυσμό, ενώ το αντίστροφο δεν ικανοποείται απαραίτητα. Όταν το δείγμα το οποίο αναλύεται είναι αντιπροσωπευτικό του πληθυσμού τότε τα συμπεράσματα που προκύπτουν από τη στατιστική μελέτη για το δείγμα ισχύουν με ικανοποιητική ακρίβεια και για τον πληθυσμό.

Κατά την συλλογή στοιχείων, όπως θα αναφερθεί και παρακάτω, εισάγουμε χαρακτηριστικά του δείγματος τα οποία αξίζει να καταγραφούν και καλούνται **μεταβλητές (variables).**Οι μεταβλητές διακρίνονται σε:

- **Ποιοτικές ή κατηγορικές μεταβλητές (qualitative variables):** Είναι οι μεταβλητές οι οποίες δεν αντιστοιχούν σε μετρήσιμα μεγέθη , αλλά κατηγοριοποιούν τα στοιχεία ενός πληθυσμού σε ομάδες σαφώς διαφοροποιημένες μεταξύ τους. Η χρήση αριθμητικής κωδικοποίησης στις ποιοτικές μεταβλητές είναι καθαρά συμβολική και αποσκοπεί στην ευκολότερη ταυτοποίησή τους.
- **Ποσοτικές μεταβλητές (quantitave variables)**: Είναι οι άμεσα μετρήσιμες μεταβλητές και διακρίνονται σε:
  - **Διακριτές (discrete variables) ή ασυνεχείς μεταβλητές (discontinuous variables):** Οι μεταβλητές αυτές παίρνουν τιμές πεπερασμένου πλήθους, συνήθως ακέραιες, χωρίς να έχουν την δυνατότητα να πάρουν ενδιάμεσες τιμές μεταξύ των ακεραίων. Διακριτές είναι οι μεταβλητές οι οποίες παίρνουν τιμές μόνο φυσικούς αριθμούς όπως για παράδειγμα ο αριθμός των οδηγών που χρησιμοποιούν το κινητό τους τηλέφωνο κατά την διάρκεια της οδήγησης.
  - **Συνεχείς μεταβλητές (continuous variables):** Οι μεταβλητές αυτές έχουν την δυνατότητα να πάρουν οποιαδήποτε τιμή στο εύρος των πραγματικών αριθμών και

η διαφορά μεταξύ δύο δυνατών τιμών μπορεί να είναι απεριόριστα μικρή.Παράδειγμα συνεχούς μεταβλητής είναι η χρονική διάρκεια χρήσης του κινητού τηλεφώνου από τους οδηγούς κατά την διάρκεια της οδήγησης.

Αξίζει να σημειωθεί ότι οι μεταβλητές διακρίνονται και σε **εξαρτημένες** και **ανεξάρτητες** , όπως θα αναλυθεί στο επόμενο κεφάλαιο.

Για ανάλυση ενός δείγματος με στοιχεία $x_1,x_2,x_3,\ldots,x_n$ όρους ορίζονται στον παρακάτω πίνακα τα εξής:

Πίνακας 3.1:Μέτρα κεντρικής τάσης, μέτρα διασποράς και μεταβλητότητας

Μέτρα κεντρικής τάσης	Μέση Τιμή x	$\frac{x1+x2+x3+\cdots+xn}{\nu}=\frac{1}{\nu}\sum_{i=1}^{\nu} xi$
Μέτρα διασποράς και μεταβλητότητας	Διακύμανση δεδομένων s^2	$s^2=\frac{1}{\nu-1}\sum_{i=1}^{\nu}(xi-x)$
	Τυπική απόκλιση s	$s=\sqrt{s^2}$

Σε συμμετρικά κατανεμημένο δείγμα δεδομένων, σύμφωνα με εμπειρικό κανόνα προκύπτει ότι το διάστημα (-s, +s) περιέχει περίπου το 68% των δεδομένων, το διάστημα (-2s,+2s) περιέχει περίπου το 95% των δεδομένων, ενώ στο διάστημα (-3s,+3s) περιέχεται περίπου το 99% των δεδομένων.

Η **συνδιακύμανση** (covariance) αποτελεί ένα μέτρο της σχέσης μεταξύ δύο περιοχών δεδομένων και δίνεται από τη σχέση:

$$\mathbf{cov(X,Y)}=\frac{1}{\nu-1}\sum_{i=1}^{\nu}[(xi-x)*(yi-y)]$$

όπου X και Y οι περιοχές δεδομένων.

Στα μέτρα αξιοπιστίας εντάσσονται το **επίπεδο εμπιστοσύνης**, το οποίο αποτελεί ένα διάστημα αριθμών που εκτιμάται ότι εμπεριέχει μια άγνωστη παράμετρο του πληθυσμού, όπως ο μέσος όρος και η τυπική απόκλιση,καθώς και το **επίπεδο σημαντικότητας**, δηλαδή η αναλογία των περιπτώσεων που ένα συμπέρασμα είναι εσφαλμένο.

3.3 Συντελεστής συσχέτισης μεταβλητών

Για τον προσδιορισμό της συσχέτισης μεταξύ ανεξάρτητων μεταβλητών υπολογίζεται ο **συντελεστής συσχέτισης Pearson (**Pearson Correlation) και συμβολίζεται με r. Ο συντελεστής αυτός ορίζεται από την παρακάτω εξίσωση:

$$\mathbf{r}=\frac{\sum_{i=1}^{\nu}[(xi-x)*(yi-y)]}{\sqrt{\sum_{i=1}^{\nu}(xi-x)2}\sqrt{\sum_{i=1}^{\nu}(yi-y)2}}$$

όπου $-1\leq r\leq 1$.Για την ακραία τιμή -1 η συσχέτιση είναι πλήρως αρνητική, ενώ για την τιμή 1 είναι πλήρως θετική. Η τιμή 0 αντιστοιχεί σε μηδενική συσχέτιση. Η θετική συσχέτιση μεταξύ 2 μεταβλητών παρατηρείται όταν είναι ανάλογες μεταξύ τους, δηλαδή αύξηση της μίας συνεπάγεται αύξηση της άλλης, ενώ η αρνητική συσχέτιση αντιστοιχεί σε αντιστρόφως ανάλογες μεταβλητές.

3.4 Μαθηματικά πρότυπα

3.4.1 Αλγόριθμοι Ταξινόμησης και Παλινδρόμησης

Οι αλγόριθμοι ταξινόμισης στην Μηχανική Μάθηση αποτελούν 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.

Αντίθετα, οι αλγόριθμοι παλινδρόμησης αξιοποιούνται ως μια μέθοδος μοντελοποίησης μιας τιμής στόχου αντί για κλάση, που βασίζεται σε ανεξάρτητους προγνωστικούς παράγοντες. Η μέθοδος αυτή αξιοποιείται κυρίως για την πρόβλεψη της σχέσης μεταξύ αιτίου και αποτελέσματος μεταξύ των ανεξάρτητων και εξαρτημένων μεταβλητών. Οι αλγόριθμοι παλινδρόμησης διακρίνονται σε γραμμικούς και μη γραμμικούς ανάλογα με τον αριθμό των ανεξάρτητων μεταβλητών και τον τύπο σχέσης που τις συνδέει.

Στην συγκεκριμένη διπλωματική εργασία για την ταξινόμηση από τον συγκριτικό πίνακα μεταξύ των μοντέλων επιλέχθηκαν τα εξής δύο μοντέλα για περαιτέρω αξιοποίηση:

Πίνακας 3.2:Επιλεγμένα μοντέλα ταξινόμησης

Αγγλική ονομασία αλγορίθμου	Ελληνική ονομασία αλγορίθμου	Συμβολισμός
Linear Discriminant Analysis	Γραμμική Διαχωριστική Ανάλυση	LDA
Logistic Regression	Λογιστική Παλινδρόμηση	LR

Για την παλινδρόμηση από τον συγκριτικό πίνακα μεταξύ των μοντέλων επιλέχθηκαν τα εξής δύο μοντέλα για περαιτέρω αξιοποίηση:

Πίνακας 3.3:Επιλεγμένα μοντέλα παλινδρόμησης

Αγγλική ονομασία αλγορίθμου	Ελληνική ονομασία αλγορίθμου	Συμβολισμός
AdaBoost Regressor	Προσαρμοστική Ενδυνάμωση	ADA
Linear Regression	Γραμμική Παλινδρόμηση	LR

3.4.2 Ταξινόμηση ανισόρροπης κατανομής κλάσεων δεδομένων- Class imbalance

Η ταξινόμηση ανισορροπίας της κλάσης των δεδομένων αναφέρεται σε ένα πρόβλημα μοντελοποίησης που εμφανίζεται όταν ο αριθμός των παρατηρήσεων στο σύνολο δεδομένων εκπαίδευσης (training set) δεν είναι ισορροπημένος. Με άλλα λόγια, η κατανομή των κλάσεων δεν είναι ίση ή κοντινή, με αποτέλεσμα να κυριαρχεί μια κλάση ή αλλιώς ένα χαρακτηριστικό. Αυτό έχει ως άμεση συνέπεια το μοντέλο πρόβλεψης να είναι ακριβές μόνο για την κυρίαρχη κλάση δεδομένων και ουσιαστικά να δίνει λάθος αποτελέσματα πρόβλεψης για το σύνολο των εξεταζόμενων στοιχείων. Τα προβλήματα ανισορροπίας μπορεί να οφείλονται σε σφάλματα μέτρησης, μεροληπτικές μεθόδους δειγματοληψίας ή και μη διαθεσιμότητα ποικιλίας στοιχείων και διαφορετικών κλάσεων (Viloria et al., 2020).

Για την αντιμετώπιση του προβλήματος υπάρχουν τρεις τρόποι:

- Over sampling minority
- Under sampling majority
- Both (Over and under)

Στην συγκεκριμένη περίπτωση επιλέχθηκε η μέθοδος Over sampling για την εξισορρόπηση του δείγματος μέσω της εντολής SMOTE, τα αποτελέσματα της οποίας θα αναλυθούν σε επόμενο κεφάλαιο.

3.5 Κριτήρια αποδοχής μοντέλων

3.5.1 Μήτρα σύγχυσης-Confusion matrix

Η μήτρα σύγχυσης αξιοποιείται ώστε να μετρηθεί η απόδοση ενός συστήματος για δύο κλάσεις ή παραπάνω. Στην συγκεκριμένη διπλωματική εργασία έγινε χρήση της μήτρας σύγχυσης για την αξιολόγηση των μοντέλων γραμμικής διαχωριστικής ανάλυσης και λογιστικής παλινδρόμησης. Μέσω της συγκεκριμένης αναπαράστασης, η οποία απεικονίζεται παρακάτω, τα αποτελέσματα καθίστανται περισσότερο εύληπτα στον μελετητή και τον αναγνώστη.

		Predicted	
		Negative **(N)** -	Positive **(P)** +
Actual	Negative -	True Negative **(TN)**	**False Positive (FP) Type I Error**
	Positive +	**False Negative (FN) Type II Error**	True Positive **(TP)**

Εικόνα 3.1:Μήτρα σύγχυσης-Confusion Matrix

Πηγή:Medium,2022[Available at: https://medium.com/analytics-vidhya/what-is-a-confusion-matrix-d1c0f8feda5]

Τα πιθανά σενάρια μετά την εκπαίδευση του αλγορίθμου είναι τα ακόλουθα:

- Αληθώς θετικά (**True Positives-TP**):Το πλήθος των προβλέψεων για τις οποίες ο ταξινομητής προέβλεψε σωστά την κλάση που ανήκουν.
- Ψευδώς θετικά (**False Positives-FP**): Το πλήθος των προβλέψεων για τις οποίες ο ταξινομητής προέβλεψε την κλάση που ανήκουν, χωρίς όμως να ανήκουν πραγματικά σε αυτή.
- Αληθώς αρνητικά (**True Negatives-TN**):Το πλήθος των προβλέψεων που δεν ανήκουν στην κλάση, αλλά ο ταξινομητής προέβλεψε σωστά.
- Ψευδώς αρνητικά (**False Negatives-FN**): Το πλήθος των προβλέψεων για τις οποίες ο ταξινομητής λανθασμένα προέβλεψε ότι ανήκουν σε άλλη κλάση.

3.5.1.1 Ορθότητα

Η ορθότητα (Accuracy) ορίζεται ως το ποσοστό των σωστών προβλέψεων ενός μοντέλου και προσδιορίζεται μέσω του παρακάτω τύπου:

$$\textbf{Accuracy}=\frac{TP+TN}{TP+TN+FP+FN}$$

Ο βαθμός των εσφαλμένων κατηγοριοποιήσεων του ταξινομητή εκφράζεται μπορεί να χρησιμοποιηθεί αντί της ορθότητας μέσω του μέτρου του λόγου σφάλματος (error rate) ως εξής: **error rate= 1- accuracy**

3.5.1.2 Ευαισθησία και Εξειδιτευτικότητα

Η ευαισθησία (sensitivity ή recall) ορίζεται ως η πιθανότητα το μοντέλο να προβλέψει ένα θετικό αποτέλεσμα για μια παρατήρηση όταν το αποτέλεσμα είναι θετικό (true positive rate).

$$\textbf{Sensitivity/Recall}=\frac{TP}{TP+FN}$$

Η εξειδιτευτικότητα (specifisity) ορίζεται ως η πιθανότητα το μοντέλο να προβλέψει ένα αρνητικό αποτέλεσμα για μια παρατήρηση όταν το αποτέλεσμα είναι αρνητικό (true negative rate).

$$\textbf{Specifisity}=\frac{TN}{TN+FP}$$

Η ευαισθησία και η εξειδιτευτικότητα αποτελούν τους άξονες για την καμπύλη ROC (Receiver Operating Characteristic),από την οποία προκύπτει το AUC, δηλαδή η περιοχή κάτω από την καμπύλη.

3.5.1.3 Μέτρο F1

Το μέτρο F1 (F1-Score) ορίζεται ως εξής:

$$\textbf{F1-Score}=2*\frac{Precision*Recall}{Precision+Recall}$$

,όπου:

Precision-Ακρίβεια: Το μέτρο της ακρίβειας εκφράζει τις σωστές προβλέψεις σε σχέση με τις τις συνολικές σωστές προβλέψεις:

$$\textbf{Precision}=\frac{TP}{TP+FP}$$

3.5.1.4 Στατιστικός Συντελεστής Κάππα

Ο στατιστικός συντελεστής Κάππα (Kappa statistic) είναι το μέτρο αξιολόγησης εξεταζόμενου μοντέλου κατηγοριοποίησης:

$$\textbf{Kappa Statistic}=\frac{Po-Pe}{1-Pe}$$

Όπου:

Po:Παρατηρούμενη σχετική συμφωνία μεταξύ των μοντέλων κατηγοριοποίησης.

Pe: Η υποθετική πιθανότητα η συμφωνία αυτή να οφείλεται σε τυχαίο παράγοντα.

Ο συντελεστής Κάππα λαμβάνει τιμές από μηδέν έως ένα. Για Kappa=0 δεν υπάρχει καμία συμφωνία μεταξύ των δύο κατηγοριών,ενώ το Kappa=1 αντιστοιχεί σε πλήρη συμφωνία.

3.5.1.5 Συντελεστής MCC

Ο στατιστικός συντελεστής MCC (Mattheus correlation coefficient) αποτελεί μια μέτρηση προσδιορισμού της απόδοσης ενός μοντέλου ταξινόμησης και υπολογίζεται μέσω του παρακάτω τύπου:

$$\textbf{MCC}=\frac{TP*TN-FP*FN}{\sqrt{(TP+FP)*(TP+FN)*(TN+FP)*(TN+FN)}}$$

Ο συντελεστής MCC λαμβάνει τιμές στο διάστημα [-1,1] με την τιμή -1 να δηλώνει ασυμφωνία μεταξύ των προβλεπόμενων και των πραγματικών κλάσεων, ενώ τιμή MCC ίση με την μονάδα δείχνει απόλυτη συμφωνία μεταξύ προβλεπόμενων και πραγματικών κλάσεων.Τέλος, τιμή MCC=0 δηλώνει απόλυτη τυχαιότητα.

3.5.2 : Κριτήρια αποδοχής μοντέλων παλινδρόμησης

3.5.2.1:Συντελεστής R^2

Ο συντελεστής προσδιορισμού R^2 (Coefficient of determination) χρησιμοποιείται για την αξιολόγηση της απόδοσης ενός μοντέλου γραμμικής παλινδρόμησης,δηλαδή για τον έλεγχο του πόσο καλά τα παρατηρούμενα αποτελέσματα αναπαράγονται από το μοντέλο, ανάλογα με τον λόγο της συνολικής απόκλισης των αποτελεσμάτων που περιγράφονται από το μοντέλο (Cameron & Windmeijer, 1996).

3.5.2.1:Μέσο Απόλυτο Σφάλμα

Το MAE (Mean Absolute Error) αποτελεί ένα μέσο μέτρησης του σφάλματος σε αναλογική κλίμακα με τα δεδομένα. Οι τιμές του σφάλματος ανήκουν στο διάστημα [0,άπειρο). Όσο πιο κοντά τείνει στο 0 τόσο πιο αξιόπιστο είναι το μοντέλο.

3.5.2.2:Μέσο Τετραγωνικό Σφάλμα

Το **MSE** (Mean Squared Error) ή αλλιώς MSD (Mean Squared Deviation) αποτελεί ένα δυσκολότερο στην ερμηνεία μέσο μέτρησης του σφάλματος σε σχέση με το MAE με τις τιμές του σφάλματος να ανήκουν στο διάστημα [0,άπειρο). Όσο πιο κοντά τείνει στο 0 τόσο πιο αξιόπιστο είναι το μοντέλο.

$$\mathbf{MSE}=\frac{1}{\nu}\sum_{i=1}^{\nu}[(yi-y)^2]$$

3.5.2.3:Μέσο Απόλυτο Εκατοστιαίο σφάλμα

Το **MAPE** (Mean Absolute Percentage Error) μαζί με το MAE έχουν την δυνατότητα να δώσουν μια πιο εύληπτη και κατανοητή εικόνα για την αξιοπιστία των μοντέλων. Το MAPE λαμβάνει τιμές εντός του εύρους [0,100%] με τις πιο χαμηλές τιμές και συγκεκριμένα τιμές κάτω από 10% να δηλώνουν εξαιρετικά αξιόπιστο μοντέλο.

3.5.2.4:Root Mean Square Error

Το **RMSE** δείχνει την μέση απόσταση μεταξύ των προβλεπόμενων τιμών που προέκυψαν από το μοντέλο και των πραγματικών τιμών στην βάση δεδομένων. Όσο χαμηλότερη είναι η τιμή του RMSE τόσο καλύτερα ταιριάζει το μοντέλο στα συγκεκριμένα δεδομένα.

$$\mathbf{RMSE}=\sqrt{\sum_{i=1}^{\nu}[\frac{(Pi-Oi)^2}{\nu}]}$$

,όπου P_i και O_i οι προβλεπόμενες και πραγματικές τιμές αντίστοιχα.

4 : ΣΥΛΛΟΓΗ ΚΑΙ ΕΠΕΞΕΡΓΑΣΙΑ ΣΤΟΙΧΕΙΩΝ

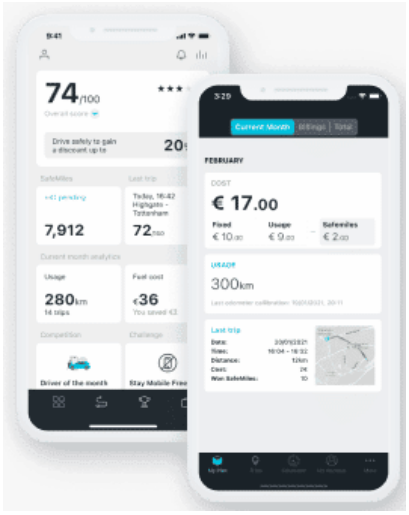
4.1:Εισαγωγή

Όπως αναφέρθηκε και στο 1^{ο} κεφάλαιο, στόχο της παρούσας διπλωματικής εργασίας αποτελεί η διερεύνηση της επιρροής χρήσης του κινητού τηλεφώνου στην συμπριφορά του οδηγού. Σε αυτό το κεφάλαιο θα περιγραφεί με σαφήνεια ο τρόπος με τον οποίο συλλέχτηκαν τα δεδομένα που αξιοποιήθηκαν στην συγκεκριμένη διπλωματική εργασία.

4.2:Συλλογή των στοιχείων

4.2.1 Εφαρμογή OSeven Telematics

Η εταιρία τηλεματικής **OSeven** Telematics (https://oseven.io/) παρείχε τα δεδομένα που αξιοποιήθηκαν για την εκπόνηση της συγκεκριμένης διπλωματικής εργασίας. Η συγκεκριμένη εταιρία εξειδικεύεται πάνω από 25 χρόνια στους τομείς της ανάλυσης της οδηγικής συμπεριφοράς, της ανάλυσης τροχαίων ατυχημάτων, της απόσπασης της προσοχής, της οδικής ασφάλειας, της μηχανικής μεταφορών, της μοντελοποίησης, της ανάλυσης μεγάλων δεδομένων και της μηχανικής μάθησης. Η πλατφόρμα OSeven ακολουθεί αυστηρή ασφάλεια πληροφοριών και πολιτικές απορρήτου σε πλήρη συμμόρφωση με τη γενική νομοθεσία περί Προστασίας Δεδομένων Προσωπικού Χαρακτήρα και τις σχετικές οδηγίες της ΕΕ. Ως εκ τούτου, όλα τα δεδομένα έχουν παρασχεθεί από την OSeven σε **ανώνυμη** μορφή και χωρίς πληροφορίες γεωγραφικού εντοπισμού για τα ταξίδια.



Εικόνα 4.1:Εφαρμογή OSeven Telematics
Πηγή: OSeven (2022)

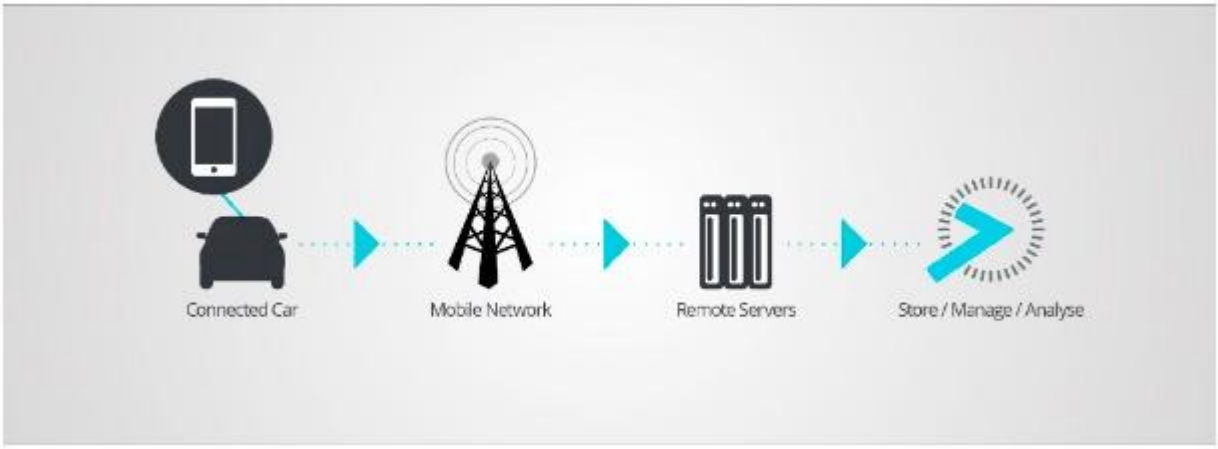
Ειδικότερα, τα δεδομένα αυτά συλλέχθησαν σε πραγματικές οδικές συνθήκες μέσω εφαρμογής ενσωματωμένης στο κινητό τηλέφωνο οδηγών, η οποία κατέγραφε τα χαρακτηριστικά της οδηγικής τους συμπεριφοράς για μια συγκεκριμένη χρονική περίοδο το έτος 2020 είτε χρησιμοποιούσαν το κινητό τους τηλέφωνο εν ώρα οδήγησης είτε όχι. Η ανάπτυξη της εφαρμογής, η οποία παρουσιάστηκε το 2014, στηρίχθηκε στην ανάγκη να συλλεχθούν και να αναλυθούν δεδομένα οδικής συμπεριφοράς στοιχείων αληθινών οδικών συνθηκών, μεγάλης κλίμακας σε πραγματικό χρόνο και με άμεση καταγραφή και αποθήκευσή τους. Απώτερος στόχος της ανάπτυξης αυτής της καινοτόμου εφαρμογής είναι η αξιολόγηση και βελτίωση της οδηγικής συμπεριφοράς και συνεπώς της οδικής ασφάλειας. Στην παρακάτω εικόνα παρουσιάζεται συνοπτικά και απλοϊκά η διεργασία που ακολουθεί η συγκεκριμένη εφαρμογή ώστε να αξιολογήσει την οδηγική συμπεριφορά των οδηγών που την έχουν κατεβάσει στις κινητές τους συσκευές (Katrakazas et al.,2020).

Εικόνα 4.2: Διαδικασία συλλογής και επεξεργασίας δεδομένων από την OSeven Telematics

Πηγή: OSeven (2022)

4.2.2 Τρόπος λειτουργίας εφαρμογής OSeven Telematics

Οι αισθητήρες hardware της συσκευής smartphone αποτελούν την βάση για την λειτουργία της εφαρμογής OSeven Telematics. Αξίζει να σημειωθεί ότι δεν απαιτείται κάποιος επιπλέον εξοπλισμός για την ορθή λειτουργία της εφαρμογής. Ακόμα, αξιοποιείται πληθώρα APIs (Application Programming Interfaces) για την ανάγνωση των δεδομένων προερχόμενων από τους αισθητήρες και την προσωρινή αποθήκευσή τους στην βάση δεδομένων του κινητού τηλεφώνου πριν αυτά σταλούν στην κεντρική βάση δεδομένων (back-end database) της εταιρίας για περαιτέρω ανάλυση και αξιοποίηση. Τα δεδομένα που συλλέγονται καθίστανται χωροχρονικά διακεκριμένα και αφότου αποθηκευτούν στην τελική βάση δεδομένων μετατρέπονται σε δείκτες οδηγικής συμπεριφοράς και ασφάλειας μέσω επεξεργασίας σημάτων, αλγορίθμους Μηχανικής Μάθησης, συγχώνευση δεδομένων (data fusion) και αλγορίθμους Μαζικών Δεδομένων (Big Data algorithms) (Kontaxi et al., 2022).



Εικόνα 4.3: Ροή δεδομένων συστήματος OSeven

Πηγή: Tselentis, D., Vlahogianni, E., Yannis, G. & Kavouras, L. (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 February 25, 2023)

Ειδικότερα, οι αισθητήρες του smartphone που αξιοποιούνται περιλαμβάνουν επιταχυνσιόμετρο, γυροσκόπιο, μαγνητόμετρο και GPS, ενώ οι τεχνικές συγχώνευσης δεδομένων παρέχονται από την Android και την iOS 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. Μετά το πέρας του ταξιδιού, η εφαρμογή διαβιβάζει όλα τα δεδομένα σε μια κεντρική βάση δεδομένων μέσω κατάλληλου καναλιού επικοινωνίας όπως ένα δίκτυο Wi-Fi ή ένα κυψελοειδές δίκτυο (δίκτυο 3G/4G) με

βάση τις ρυθμίσεις του χρήστη. Στη συνέχεια, τα δεδομένα αποθηκεύονται σε διακομιστή cloud για κεντρική επεξεργασία και μείωση των δεδομένων και υποβάλλονται σε επεξεργασία με την βοήθεια της μηχανικής μάθησης (Tselentis et al,2020).

4.2.3 Στοιχεία που συλλέχθηκαν από την εφαρμογή OSeven Telematics

Στην συγκεκριμένη διπλωματική εργασία συλλέχθηκαν δεδομένα τα οποία αφορούσαν 356.162 οδικές διαδρομές το έτος 2020 στην Ελλάδα. Η συλλογή των δεδομένων έγινε την περίοδο έξαρσης της πανδημίας SARS-CoV-2. Αφαιρέθηκαν δείκτες που αφορούσαν στην αξιολόγηση των οδηγών (Stars), όπως επίσης και δείκτες σχετικοί με τα scores που πέτυχαν οι οδηγοί κατά την διάρκεια της οδικής διαδρομής (total_score, speeding_score, mu_score, hb_score, ha_score), με σκοπό την προστασία των προσωπικών δεδομένων. Τέλος, αφαιρέθηκαν και δείκτες, οι οποίοι έχουν άμεση εξάρτηση μεταξύ τους και επομένως μπορεί να οδηγούσαν σε μη αξιόπιστη μετέπειτα ανάλυση. Στην προκειμένη περίπτωση τέτοιοι δείκτες αποτελούν αυτοί της απότομης επιτάχυνσης και επιβράδυνσης με την απότομη επιτάχυνση ανά 100km και την απότομη επιβράδυνση ανά 100 km αντίστοιχα.

Συνεπώς, παρέμειναν μόλις οι 13 από τους 21 δείκτες για περαιτέρω επεξεργασία και ανάλυση.

4.3:Επεξεργασία των στοιχείων

Η επεξεργασία των δεδομένων έγινε μέσω της γλώσσας προγραμματισμού Python (πακέτο Pycaret) σε περιβάλλον Jupyter Notebook με χρήση των βιβλιοθηκών ανάλυσης δεδομένων numpy, pandas και seaborn για την παρουσίασή τους.

4.3.1:Περιγραφή των στοιχείων

Σε αυτό το υποκεφάλαιο παρατίθενται μέσω του παρακάτω πίνακα (Πίνακας 4.1) οι συμβολισμοί της κάθε μεταβλητής, η μονάδα μέτρησης που της αντιστοιχεί, καθώς και μια συνοπτική περιγραφή του περιεχομένου της.

Πίνακας 4.1:Περιγραφή των υπό εξέταση μεταβλητών

Ονομασία Μεταβλητής	Μονάδα μέτρησης	Περιγραφή Μεταβλητής
Duration	sec	Συνολική διάρκεια διαδρομής
total_distance	km	Συνολική διανυθείσα απόσταση
risky_hours	km	Οδηγηθείσα απόσταση στις κρίσιμες ώρες (00:00-05:00)

ha/100km	-	Απότομες επιταχύνσεις στα 100 χιλιόμετρα
hb/100km	-	Απότομες επιβραδύνσεις στα 100 χιλιόμετρα
sum_speeding	Sec	Συνολική διάρκεια οδήγησης με υπέρβαση ορίου ταχύτητας και ανοχής
av_speeding_kmh	km/h	Μέση ταχύτητα οδήγησης με υπέρβαση ορίου ταχύτητας και ανοχής σε μια διαδρομή
avg speed	km/h	Μέση ταχύτητα διαδρομής
avg_driving_speed	km/h	Μέση ταχύτητα οδήγησης
time_mobile_usage	Sec	Συνολική διάρκεια χρήσης κινητού τηλεφώνου σε μια διαδορμή
driving_duration	Sec	Συνολική διάρκεια οδήγησης (χωρίς την διάρκεια των στάσεων)
time_mobile_usage/driving_duration	sec/sec	Διάρκεια χρήσης κινητού τηλεφώνου ανά μονάδα συνολικής διάρκειας οδήγησης
sum_speeding/driving_duration	sec/sec	Διάρκεια οδήγησης με υπέρβαση ορίου ταχύτητας και ανοχής ανά μονάδα συνολικής διάρκειας οδήγησης

4.3.2:Περιγραφική Στατιστική των στοιχείων

Στην ακόλουθη υποενότητα γίνεται η περιγραφική στατιστική ανάλυση των δεδομένων που παρασχέθηκαν από την OSeven Telematics. Πιο συγκεκριμένα, στον Πίνακα 4.2 παρουσιάζονται για κάθε μεταβλητή βασικά στατιστικά μεγέθη που προέκυψαν μέσω της εντολής describe()round() σε περιβάλλον Jupyter Notebook, όπως η μέση τιμή, η τυπική απόκλιση, καθώς και οι ελάχιστες και μέγιστες τιμές τους.

Πίνακας 4.2:Περιγραφική στατιστική των μεταβλητών

Μεταβλητή	Μέση Τιμή	Τυπική Απόκλιση	Ελάχιστη Τιμή	Μέγιστη Τιμή
duration (sec)	962.04	1093.98	61.00	25549.00
total_distance (km)	11.60	22.31	0.50	648.69
risky_hours (km)	0.37	4.01	0.00	427.70
ha/100km (-)	11.95	27.86	0.00	597.01
hb/100km (-)	16.39	29.76	0.00	819.67
sum_speeding (sec)	65.63	194.31	0.00	7697.00
av_speeding_kmh (km/h)	3.99	6.03	0.00	314.16
avg speed (km/h)	35.13	18.89	1.97	262.52
avg_driving_speed (km/h)	42.57	17.58	5.57	323.91
time_mobile_usage (sec)	39.11	159.27	0.00	9901.00
driving_duration (sec)	769.97	967.15	61.00	23900.00
time_mobile_usage/driving_duration (sec/sec)	0.05	0.14	0.00	1.00
sum_speeding/driving_duration (sec/sec)	0.06	0.11	0.00	1.00

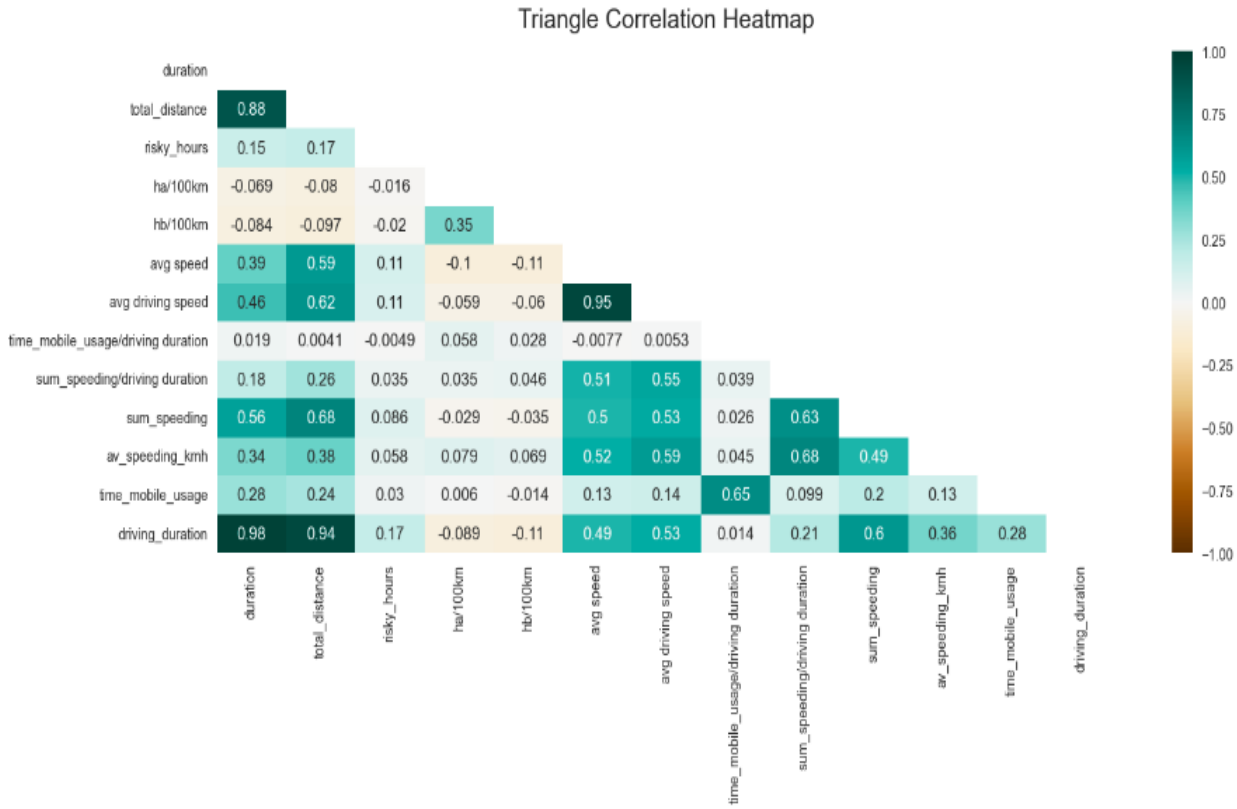
Από την παραπάνω περιγραφική στατιστική των στοιχείων γίνεται αντιληπτό ότι η συνολική **διάρκεια διαδρομής** έχει **μέση τιμή** 962.04 δευτερόλεπτα, δηλαδή **16.03 λεπτά**. Επίσης, η μέση διάρκεια διαδρομής χωρίς ενδιάμεσες στάσεις έχει μέση τιμή 769.97 δευτερόλεπτα, δηλαδή υπολείπεται κατά 3.2 λεπτά της συνολικής διάρκειας διαδρομής. Αυτό ενδεχομένως οφείλεται σε τυχόν αυξημένους κυκλοφοριακούς φόρτους και στάσεις πριν από κόκκινη ένδειξη σηματοδότη. Αξιοσημείωτο είναι το γεγονός ότι η μέση τιμή διάρκειας χρήσης κινητού τηλεφώνου κυμαίνεται στα 0.65 λεπτά και στην πραγματικότητα το νούμερο αυτό είναι αρκετά μεγαλύτερο για τους οδηγούς που χρησιμοποιούν κινητό τηλέφωνο εν ώρα οδήγησης, καθώς σε αυτή την μέση τιμή συμπεριλαμβάνονται και πάνω από τους μισούς οδηγούς, οι οποίοι δεν έκαναν χρήση της κινητής τους συσκευής ενώ οδηγούσαν. Στο επόμενο κεφάλαιο θα αναλυθεί περαιτέρω το πώς η χρήση κινητού τηλεφώνου επηρεάζει την οδηγική τους συμπεριφορά.

4.3.3:Συσχέτιση Pearson

Για την ανάπτυξη τόσο των μοντέλων ταξινόμησης όσο και των μοντέλων παλινδρόμησης οφείλει να διερευνηθεί η σχέση μεταξύ ανεξάρτητων μεταβλητών. Για την εξακρίβωση της μεταξύ τους σχέσης χρησιμοποιείται ο συντελεστής Pearson, του οποίου οι τιμές κυμαίνονται στο διάστημα [-1,1] και ισχύουν τα ακόλουθα:

- | 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

Χρησιμοποιώντας την διαθέσιμη βιβλιοθήκη seaborn σε προγραμματιστικό περιβάλλον Python συντάχθηκε ο παρακάτω τριγωνικός χάρτης θερμότητας. Η θετική συσχέτιση μεταξύ των ανεξάρτητων μεταβλητών συμβολίζεται με αποχρώσεις του μπλε χρώματος, ενώ η αρνητική συσχέτιση με αποχρώσεις του καφέ χρώματος.



Γράφημα 4.1:Τριγωνικός χάρτης συσχέτισης μεταβλητών

Από το Γράφημα 4.1 απορρέουν τα εξής συμπεράσματα σχετικά με την συσχέτιση μεταξύ των μεταβλητών:

- Μεταξύ των διαφορετικών περιγραφικών στατιστικών της **ίδιας μεταβλητής** παρατηρείται **υψηλή συσχέτιση**. Η παραπάνω υψηλή συσχέτιση είναι λογική δεδομένου ότι αφορά στη σχέση μεταξύ διαφορετικών εκφάνσεων του ίδιου στοιχείου. Παραδείγματος χάριν η ανεξάρτητη μεταβλητή driving_duration με την ανεξάρτητη

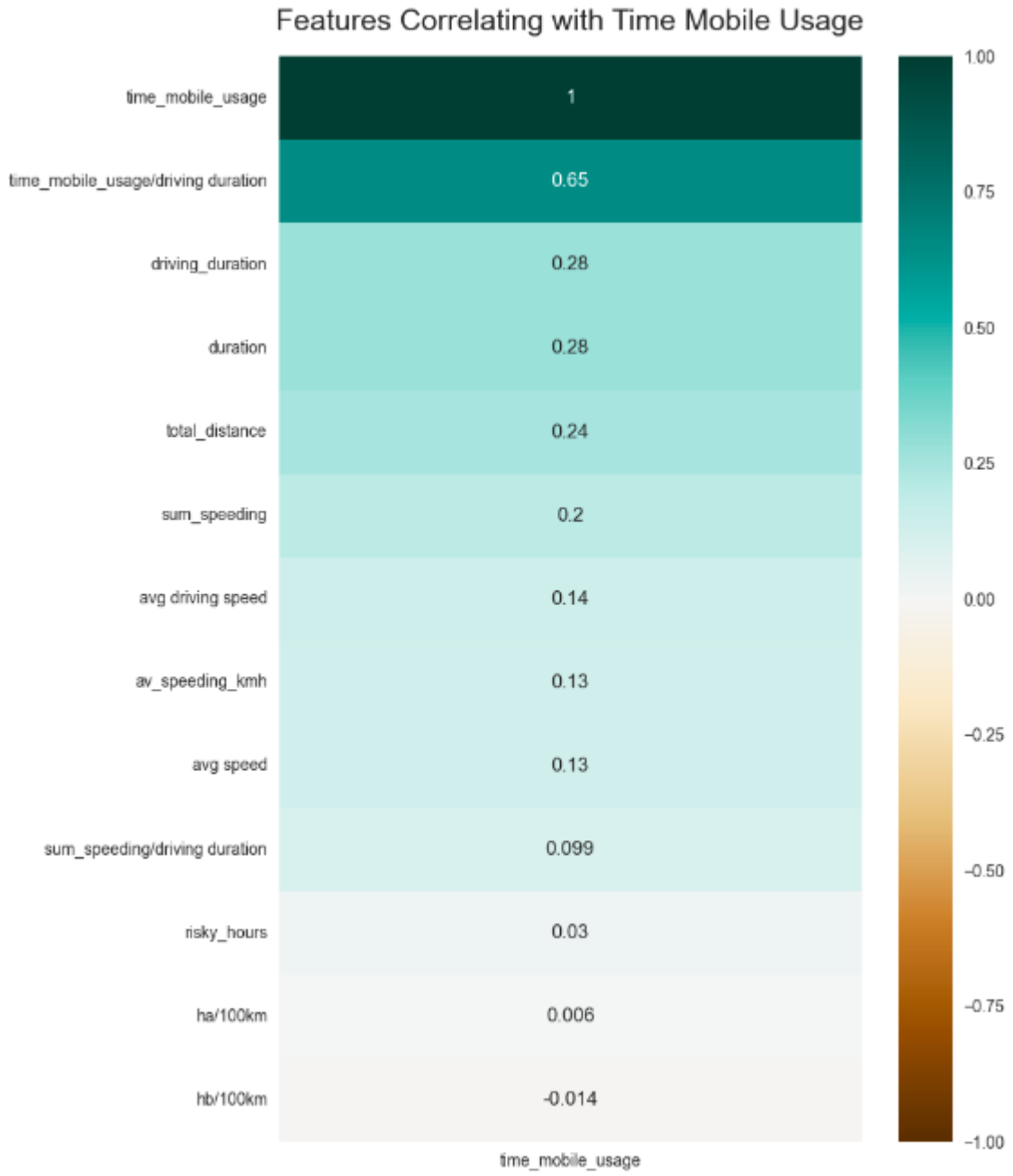
μεταβλητή duration παρουσιάζουν πολύ ισχυρή συσχέτιση (r=0.98), διότι και οι δύο αποτελούν εκφάνσεις του μεγέθους της διάρκειας οδήγησης.

- Η **συνολική διανυθείσα απόσταση** (total_distance) παρουσιάζει **ισχυρή συσχέτιση** με την μεταβλητή της **συνολικής διάρκειας οδήγησης** είτε με είτε χωρίς στάσεις (duration και driving_duration αντίστοιχα), γεγονός που απορρέει λογικά από την εξίσωση του ορισμού της ταχύτητας, όπου χρονική διάρκεια οδήγησης και συνολική διανυθείσα απόσταση είναι ανάλογα ποσά.

- Στο παραπάνω γράφημα παρατηρείται **πολύ μικρή αρνητική συσχέτιση** μεταξύ των ανεξάρτητων μεταβλητών, γεγονός που αποδεικνύει ότι η αύξηση μίας ανεξάρτητης μεταβλητής δεν μειώνει το μέγεθος μίας άλλης.

- Η χρήση του κινητού τηλεφώνου εν ώρα οδήγησης **αυξάνει** εν γένει την **διάρκεια οδήγησης**, ενώ παράλληλα αυξάνει ελάχιστα και την **μέση ταχύτητα οδήγησης**. Αυτό οφείλεται στην απόσπαση της προσοχής του οδηγού όταν χρησιμοποιεί το κινητό του τηλέφωνο, η οποία οδηγεί σε μεγαλύτερους χρόνους αντιδράσεως και μεγαλύτερες αυξομειώσεις της ταχύτητας. Βέβαια, σύμφωνα με το Γράφημα 4.1 η χρήση κινητού τηλεφώνου δεν επηρεάζει σχεδόν καθόλου τις απότομες επιταχύνσεις και επιβραδύνσεις ανά 100km.

Αξίζει να σημειωθεί, όπως αναφέρθηκε και παραπάνω, ότι αφαιρέθηκαν μεταβλητές που αφορούσαν σε scores ταχύτητας και απότομης επιτάχυνσης και επιβράδυνσης, τα οποία δύναται να δώσουν υψηλή αρνητική συσχέτιση με τα μεγέθη της ταχύτητας και απότομων επιταχύνσεων και επιβραδύνσεων αντίστοιχα.



Γράφημα 4.2: Συσχέτιση Pearson ανεξάρτητων μεταβλητών με διάρκεια χρήσης κινητού τηλεφώνου (time_mobile_usage)

Στο παραπάνω Γράφημα 4.2 απεικονίζεται πιο εύληπτα η συσχέτιση της χρήσης κινητού τηλεφώνου (time_mobile_usage) που αποτελεί και την εξαρτημένη μεταβλητή της στατιστικής ανάλυσης με τις επιλεγείσες ανεξάρτητες μεταβλητές. Προφανώς, η χρήση κινητής συσκευής συσχετίζεται περισσότερο με την χρήση κινητού τηλεφώνου ανά διάρκεια οδήγησης χωρίς στάσεις, καθώς η τελευταία αποτελεί εναλλακτική έκφραση της διάρκειας χρήσης κινητού τηλεφώνου. Για αυτόν τον λόγο η μεταβλητή αυτή αφαιρείται από το δείγμα πριν την στατιστική ανάλυση. Στο επόμενο κεφάλαιο θα αναλυθεί διεξοδικά μέσω της μηχανικής μάθησης το πώς οι παραπάνω ανεξάρτητες μεταβλητές επηρεάζουν την διάρκεια χρήσης κινητού τηλεφώνου.

5 : ΕΠΕΞΕΡΓΑΣΙΑ ΔΕΔΟΜΕΝΩΝ ΚΑΙ ΑΝΑΛΥΣΕΙΣ

5.1 Εισαγωγή

Σε αυτό το κεφάλαιο περιλαμβάνεται η εφαρμογή της μεθοδολογίας, η οποία περιγράφηκε με σαφήνεια στο 3^{ο} κεφάλαιο ως το θεωρητικό υπόβαθρο της εργασίας, καθώς και τα αποτελέσματα που προέκυψαν στο πλαίσιο της μελέτης. Για την επίτευξη του στόχου του παρόντος ερευνητικού έργου αξιοποιήθηκε η βιβλιογραφική ανασκόπηση που αναλύθηκε στο 2^{ο} κεφάλαιο.

Πιο συγκεκριμένα, το παρόν κεφάλαιο διαρθρώνεται σε 2 υποκεφάλαια, στα οποία **αναπτύσσονται κατάλληλοι αλγόριθμοι ταξινόμησης και παλινδρόμησης** με την συμβολή της μηχανικής μάθησης.

Το πρώτο υποκεφάλαιο αφορά στην εκπαίδευση και **ανάλυση μοντέλων ταξινόμησης** (classification models) μέσω της μηχανικής μάθησης, με σκοπό τον διαχωρισμό των οδηγών, οι οποίοι χρησιμοποιούν το κινητό τους τηλέφωνο κατά την οδήγηση από εκείνους που δεν το χρησιμοποιούν. Τα δεδομένα που συλλέχθηκαν από την εφαρμογή της εταιρίας OSeven Telematics αποτελούν τις ενδογενείς μεταβλητές, ενώ η δυαδική μεταβλητή της χρήσης κινητού τηλεφώνου αποτελεί την εξωγενή μεταβλητή. Από τις δοκιμές που έγιναν με διάφορα μοντέλα ταξινόμησης επιλέχθηκαν δύο από αυτά για περαιτέρω ανάλυση και εξαγωγή αποτελεσμάτων. Επιπροσθέτως, αξιολογήθηκε η σημαντικότητα της κάθε ανεξάρτητης μεταβλητής και επιλέχθηκαν οι πέντε σημαντικότερες ανεξάρτητες μεταβλητές (Feature Importance) και στην συνέχεια επαναλήφθηκε η ανάλυση των ίδιων δύο μοντέλων, ώστε να αυξηθεί η αποδοτικότητα της ανάλυσης και να εξεταστεί αν θα προκύψουν πιο αξιόπιστα μοντέλα.

Στο δεύτερο υποκεφάλαιο αναλύονται τα **μοντέλα παλινδρόμησης** (regression models)**,** έχοντας στόχο την πρόβλεψη της διάρκειας χρήσης κινητού τηλεφώνου την ώρα της οδήγησης. Ομοίως με την ταξινόμηση, τα δεδομένα που προέκυψαν από την εφαρμογή της OSeven Telematics αποτελούν τις ανεξάρτητες μεταβλητές των μοντέλων παλινδρόμησης, ενώ η διάρκεια χρήσης κινητού τηλεφώνου αποτελεί την εξαρτημένη μεταβλητή. Όπως και στην ταξινόμηση, έτσι και στην παλινδρόμηση από τις δοκιμές που έγιναν επιλέχθηκαν τα δύο πιο κατάλληλα μοντέλα για την περαιτέρω επεξεργασία, ανάλυση και εξαγωγή συμπερασμάτων και με βάση τις σημαντικότερες ανεξάρτητες μεταβλητές επαναλήφθηκε η ανάλυση των ίδιων δύο μοντέλων, συγκρίνοντας τα με τα αρχικά.

Η αξιολόγηση της προγνωστικής ικανότητας των μοντέλων θα πραγματοποιηθεί αξιοποιώντας μερικές **μετρικές αξιολόγησης**, όπως αυτές ορίστηκαν στα κριτήρια αποδοχής μοντέλων εντός του 3^{ου} κεφαλαίου της συγκεκριμένης διπλωματικής εργασίας.

Η ανάλυση πραγματοποιείται μέσω του πακέτου Pycaret της γλώσσας προγραμματισμού Python σε προγραμματιστικό περιβάλλον Jupyter Notebook με χρήση των διαθέσιμων βιβλιοθηκών, όπως η βιβλιοθήκη numpy για τους υπολογισμούς, pandas για την ανάλυση και τον χειρισμό των δεδομένων, Scikit-learn για την μηχανική εκμάθηση, Matplotlib και Seaborn για την γραφική απεικόνιση και του εργαλείου για τον χειρισμό ανομοιογενών δεδομένων Imbalanced Learn.

5.2 Διαδικασία Ταξινόμησης

5.2.1 Δημιουργία δυαδικής μεταβλητής

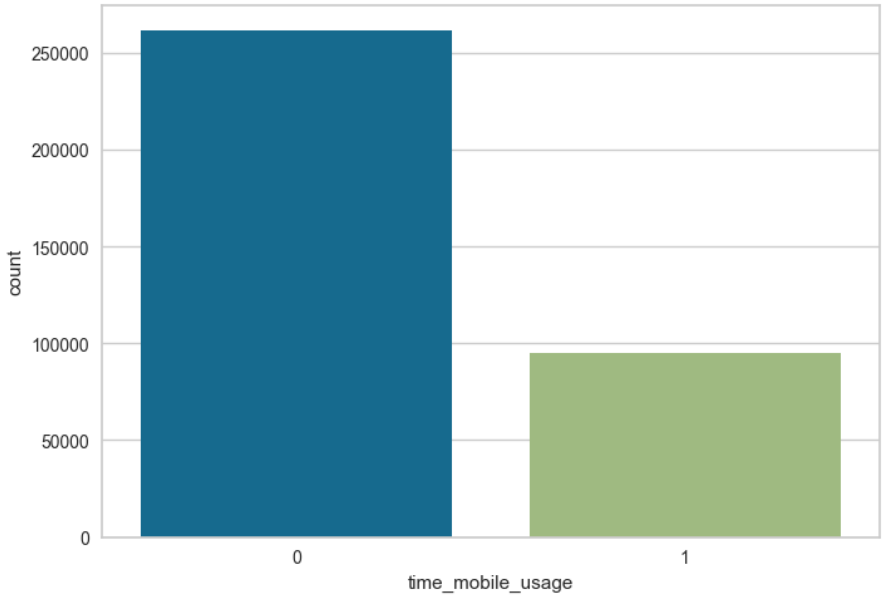
Στην συγκεκριμένη υποενότητα εκτελείται ανάλυση δεδομένων με εξαρτημένη μεταβλητή την χρήση του κινητού τηλεφώνου. Πιο συγκεκριμένα, από τα αρχικά δεδομένα που παρασχέθηκαν από την εταιρία OSeven Telematics προέκυψε η απαίτηση να δημιουργηθεί μια δυαδική μεταβλητή, η οποία θα αποτελέσει την εξαρτημένη μεταβλητή για την ταξινόμηση. Η συγκεκριμένη συνεχής μεταβλητή είναι η διάρκεια χρήσης του κινητού τηλεφώνου (time_mobile_usage), η οποία μέσω αλγορίθμου στην γλώσσα προγραμματισμού Python (Εικόνα 5.1) μετατράπηκε στην δυαδική μεταβλητή της χρήσης κινητού τηλεφώνου με τις εξής τιμές:

- Τιμή 0: Για μη χρήση κινητού τηλεφώνου κατά την διάρκεια της οδήγησης
- Τιμή 1: Για χρήση κινητού τηλεφώνου κατά την διάρκεια της οδήγησης

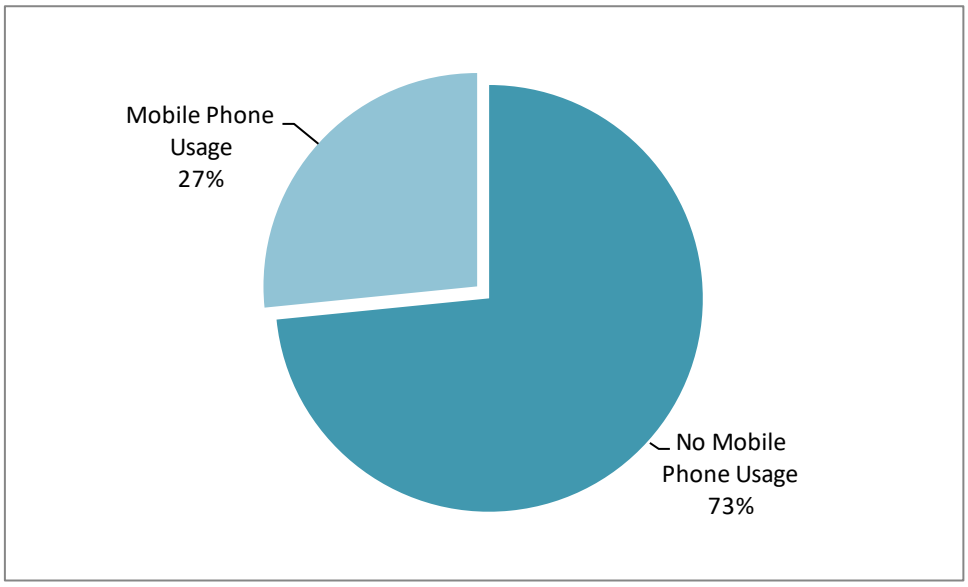
```
#Convert to binary variable
df.loc[df['time_mobile_usage'] ==0 , 'time_mobile_usage'] = 0
df.loc[df['time_mobile_usage'] > 0, 'time_mobile_usage'] = 1
df
```

Εικόνα 5.1 Κώδικας μετατροπής συνεχούς μεταβλητής σε δυαδική

Επισημαίνεται ότι οι τιμές 0 και 1 είναι καθαρά συμβολικές, καθώς συνεισφέρουν στην ταξινόμηση της μη χρήσης και χρήσης κινητού τηλεφώνου αντίστοιχα εν ώρα οδήγησης.



Γράφημα 5.1 Πλήθος οδηγών που έκαναν χρήση και μη κινητού τηλεφώνου εν ώρα οδήγησης



Γράφημα 5.2 Ποσοστό κατανομών χρήσης και μη χρήσης κινητού στα αρχικά δεδομένα

Στα Γραφήματα 5.1 και 5.2 απεικονίζεται στην τάξη 0 το πλήθος και το ποσοστό αντίστοιχα των οδηγών που δεν χρησιμοποίησαν καθόλου το κινητό τους τηλέφωνο την ώρα της οδήγησης (261374 οδηγοί), ενώ στην τάξη 1 το πλήθος και το ποσοστό εκείνων των οδηγών που έκαναν χρήση κινητού (94788 οδηγοί). Γίνεται αντιληπτό ότι οι οδηγοί που δεν χρησιμοποίησαν κινητό τηλέφωνο είναι πολύ περισσότεροι σε σχέση με εκείνους που χρησιμοποίησαν και πιο συγκεκριμένα η αναλογία μεταξύ των δύο κλάσεων είναι περίπου 1:4, δηλαδή 26,6% έκαναν χρήση κινητού και 73,4% δεν χρησιμοποίησαν κινητό.

Συνεπώς, από την παραπάνω κατανομή των δύο τάξεων δεδομένων προκύπτει ότι τα δεδομένα καθίστανται ανισόρροπα, επομένως πρόκειται για ένα πρόβλημα μη ισορροπημένης μάθησης, το οποίο πρέπει να αντιμετωπιστεί. Στο ακόλουθο υποκεφάλαιο αναλύεται η διαδικασία εξισορρόπησης των δεδομένων.

5.2.2 Μη ισορροπημένη μάθηση

5.2.2.1 Διαχωρισμός σε δεδομένα εκπαίδευσης και εξέτασης

Τα δεδομένα της συγκεκριμένης διπλωματικής εργασίας συνιστούν πρόβλημα Μη Ισορροπημένης Μάθησης (Imbalanced learn), όπως αποδείχθηκε στην προηγούμενη παράγραφο. Ειδικότερα, τα δείγματα που ανήκουν στην μειονοτική τάξη είναι αυτά της χρήσης κινητού τηλεφώνου, εξαιτίας της μη ισορροπημένης φύσης τους. Για να επιτευχθεί η διαδικασία της ταξινόμησης θα πρέπει τα δεδομένα να **διαχωριστούν** σε σύνολα **δεδομένων εκπαίδευσης** (training dataset) και **εξέτασης** (testing dataset). Η συγκεκριμένη διεργασία πραγματοποιήθηκε μέσω της τεχνικής train_test_split της βιβλιοθήκης επεξεργασίας

sklearn.model_selection, στόχος της οποίας είναι να διαχωρίσει σε πίνακες ένα υπερσύνολο δεδομένων από τυχαία κατανεμημένα στοιχεία εκπαίδευσης και εξέτασης. Η αξιολόγηση της αξιοπιστίας των μοντέλων ταξινόμησης για ικανές προβλέψεις γίνεται με βάση τα δεδομένα εξέτασης, ενώ τα δεδομένα εκπαίδευσης αξιοποιούνται για την εκμάθηση των αλγορίθμων Μη Επιβλεπόμενης Μάθησης (Unsupervised Learning).

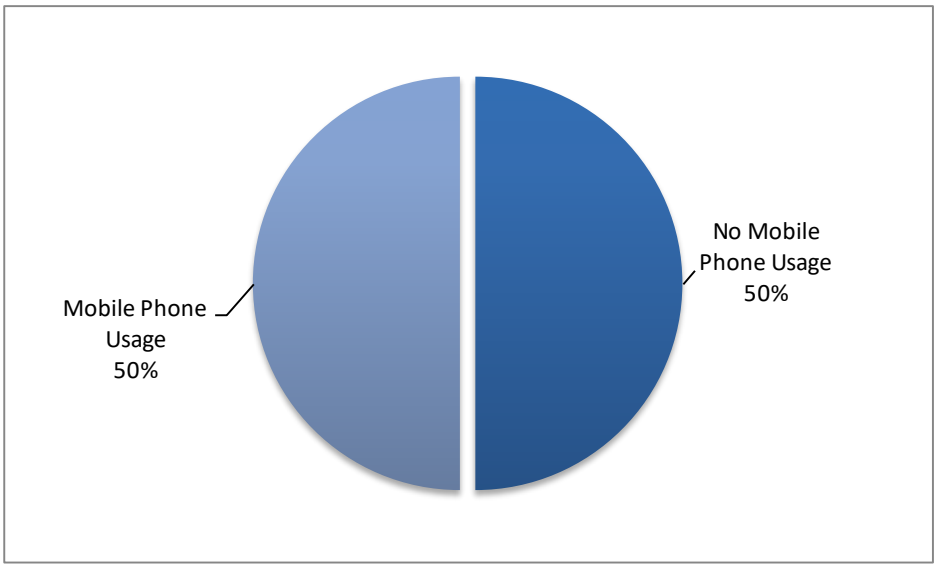
Έπειτα από δοκιμές σε 70% - 30% και 75% - 25% προέκυψε ως αναλογία που θα έδινε τα πιο αξιόπιστα αποτελέσματα για τα μοντέλα η αναλογία **80%** δεδομένα εκπαίδευσης και **20%** δεδομένα εξέτασης και η διαδικασία ταξινόμησης συνεχίστηκε με βάση αυτή την αναλογία.

5.2.2.2 Μέθοδος υπερδειγματοληψίας

Γενικά, τα μοντέλα **ταξινόμησης** στηρίζουν την λειτουργία τους σε προβλήματα **Ισορροπημένης Μάθησης**, ενώ σε προβλήματα **άνισων** κατανομών καθίστανται ασυνεπή. Το πιο σημαντικό πρόβλημα με τα ανισόρροπα δεδομένα είναι το υψηλό ποσοστό λανθασμένης ταξινόμησης για την μειονοτική κλάση, επειδή ο ταξινομητής ευνοεί την πλειοψηφούσα κλάση (Katrakazas et al.,2019). Για αυτό τον λόγο απαιτείται η εξισορρόπησή τους, ώστε συνεχίσει η διαδικασία της ταξινόμησης.

Ειδικότερα, στο συγκεκριμένο ερευνητικό έργο τα δεδομένα που ανήκουν στην μη χρήση κινητού τηλεφώνου είναι πολύ περισσότερα σε σχέση με εκείνα τα δεδομένα της χρήσης, όπως αναφέρθηκε σε προηγούμενο υποκεφάλαιο. Από τις **τρεις** μεθόδους επαναδειγματοληψίας και εξισορρόπησης των δεδομένων, όπως αυτές περιγράφηκαν στο κεφάλαιο 3, επιλέχθηκε τελικά η μέθοδος **Υπερδειγματοληψίας (Oversampling**) μέσω της τεχνικής της Συνθετικής Μειονοτικής Υπερδειγματοληψίας (Synthetic Minority Oversampling – SMOTE). Η συγκεκριμένη επιλογή έγκειται στο γεγονός ότι οι υπόλοιπες μέθοδοι παρουσίαζαν αδυναμίες. Πιο συγκεκριμένα, η μέθοδος Υποδειγματοληψίας (Undersampling) δεν προκρίθηκε στο παρόν ερευνητικό έργο, υπό τον φόβο της απώλειας σημαντικής πληροφορίας, καθώς η διακύμανση των τιμών μεταβλητών είναι ισχυρή. Επίσης, εναλλακτικές τεχνικές όπως αυτή της Προσθετικής Ανάλυσης (ADASYN) δεν επιλέχθηκαν λόγω ζητημάτων στάθμισης που απαιτούταν στα δείγματα, καθώς οι διακυμάνσεις ήταν σχετικώς ισχυρές. Αξιοσημείωτο 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, γεγονός που θα καθιστούσε τα μοντέλα ασυνεπή.

Αφού εφαρμόστηκε η τεχνική SMOTE στα δύο σύνολα δεδομένων εκπαίδευσης για τις δύο ξεχωριστές εξαρτημένες μεταβλητές, η αναλογία της χρήσης και μη κινητού τηλεφώνου κατέστη 1:1 και παρουσιάζεται στο Γράφημα 5.3. υπό μορφή πίτας.



Γράφημα 5.3 Ποσοστό κατανομών χρήσης και μη χρήσης κινητού μετά το Oversampling

5.2.3 Ταξινόμηση δεδομένων με όλες τις ανεξάρτητες μεταβλητές

Για την διαδικασία της ταξινόμησης δημιουργήθηκε η εξαρτημένη δυαδική μεταβλητή της χρήσης κινητού τηλεφώνου, όπως αναλύθηκε στο υποκεφάλαιο 5.2.1. Με βάση την συγκεκριμένη μεταβλητή συντάχθηκε στην γλώσσα προγραμματισμού **Python** μέσω της εντολής σύγκρισης των μοντέλων (compare_models()) ένας συγκριτικός πίνακας μετρικών αξιολόγησης μοντέλων ταξινόμησης, του οποίου τα αποτελέσματα παρουσιάζονται παρακάτω:

Πίνακας 5.1:Συγκριτικός πίνακας μετρικών αξιολόγησης μοντέλων ταξινόμησης

	Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC	TT (Sec)
dt	Decision Tree Classifier	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.3340
rf	Random Forest Classifier	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	17.6180
ada	Ada Boost Classifier	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.4100
gbc	Gradient Boosting Classifier	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	38.8850
lightgbm	Light Gradient Boosting Machine	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.9170
et	Extra Trees Classifier	0.9917	0.9999	0.9716	0.9973	0.9843	0.9787	0.9788	44.1410
lda	Linear Discriminant Analysis	0.8354	0.9183	0.4220	0.9167	0.5779	0.4924	0.5497	2.2430
ridge	Ridge Classifier	0.8299	0.0000	0.3946	0.9261	0.5534	0.4686	0.5338	0.2890
knn	K Neighbors Classifier	0.7443	0.7116	0.3783	0.5300	0.4415	0.2816	0.2884	6.1880
lr	Logistic Regression	0.7416	0.6433	0.1319	0.5706	0.2140	0.1266	0.1759	17.4660
dummy	Dummy Classifier	0.7329	0.5000	0.0000	0.0000	0.0000	0.0000	0.0000	0.1580
nb	Naive Bayes	0.7302	0.6960	0.2438	0.4901	0.3256	0.1801	0.1972	0.2880
svm	SVM - Linear Kernel	0.7076	0.0000	0.2069	0.5370	0.2150	0.1084	0.1558	18.8580
qda	Quadratic Discriminant Analysis	0.6078	0.7325	1.0000	0.4507	0.6061	0.3574	0.4513	1.5940

Από τον Πίνακα 5.1 αποφεύχθηκαν μοντέλα που παρουσίαζαν **AUC** ίσο με την **μονάδα**, καθώς παραπέμπουν σε **υπερπροσαρμογή** (overfitting), που αποτελεί κρίσιμο πρόβλημα σε μοντέλα μηχανικής μάθησης. Στην συνέχεια, καταβλήθηκε προσπάθεια να γίνει ανάλυση με το μοντέλο ταξινόμησης δέντρων ταξινόμησης (Extra Trees), το οποίο παρουσίαζε την αμέσως καλύτερη επίδοση σύμφωνα με τις μετρικές αξιολόγησης του Πίνακα 5.1 και θα επέτρεπε να πραγματοποιηθεί μετέπειτα ανάλυση ευαισθησίας. Όμως, αφότου έγινε βελτίωση (tune_model) και πρόβλεψη της χρήσης κινητού τηλεφώνου με αυτό το μοντέλο παρουσιάστηκε συντελεστής περιοχής κάτω από την καμπύλη ίσος με την μονάδα (AUC_{ET}=1), επομένως αποκλείστηκε και το συγκεκριμένο μοντέλο.

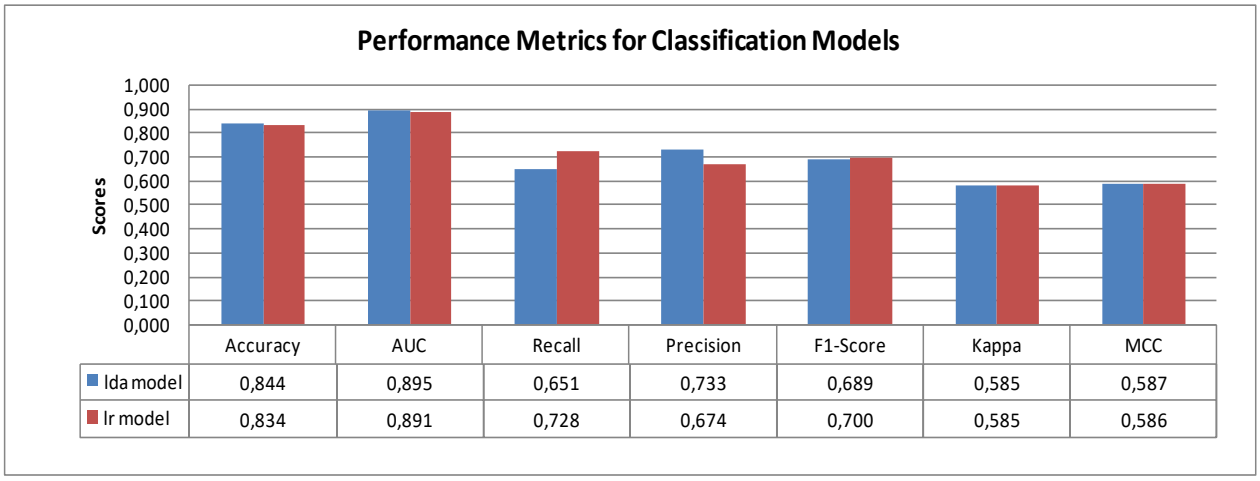
Έπειτα από δοκιμές και σε άλλα μοντέλα επιλέχθηκαν τελικά τα δύο μοντέλα που παρουσιάζονται στον Πίνακα 5.2

Πίνακας 5.2 Επιλεγμένα μοντέλα ταξινόμησης

Αγγλική ονομασία αλγορίθμου	Ελληνική ονομασία αλγορίθμου	Συμβολισμός
Linear Discriminant Analysis	Γραμμική Διαχωριστική Ανάλυση	LDA
Logistic Regression	Λογιστική Παλινδρόμηση	LR

Αξίζει να σημειωθεί ότι θα μπορούσε να γίνει ταξινόμηση και με την εκπαίδευση του αλγόριθμου ταξινόμησης Κ-Πλησιέστερων Γειτόνων (K-Nearest Neighbours), αλλά επιλέχθηκε το μοντέλο της Λογιστικής Παλινδρόμησης.

Στην συνέχεια, βελτιστοποιήθηκαν τα παραπάνω μοντέλα (tune_model) και έγινε πρόβλεψη της χρήσης κινητού τηλεφώνου με βάση τα **δεδομένα εξέτασης** (testing dataset). Στο Γράφημα 5.4 φαίνονται τα αποτελέσματα των μετρικών αξιολόγησης, οι οποίες προέκυψαν από την στατιστική ανάλυση των δύο παραπάνω αλγορίθμων.



Γράφημα 5.4 Μετρικές αξιολόγησης μοντέλων ταξινόμησης με όλες τις ανεξάρτητες μεταβλητές

Από το Γράφημα 5.4 γίνεται αντιληπτό ότι οι μέσες τιμές των μετρικών αξιολόγησης ταξινόμησης για τα δύο μοντέλα δίνουν παρόμοιες τιμές και πιο συγκεκριμένα το μοντέλο της Γραμμικής Διαχωριστικής Ανάλυσης παρουσιάζει ελαφρώς καλύτερη προβλεπτική ικανότητα, όπως άλλωστε αναμενόταν σύμφωνα με τον αρχικό συγκριτικό Πίνακα 5.1. Τόσο η ορθότητα (Accuracy) που διαθέτουν τα δύο μοντέλα όσο και ο δείκτης της περιοχής κάτω από την καμπύλη (AUC) καθίστανται ικανοποιητικά αποτελέσματα.

Για πληρέστερη παρουσίαση των αποτελεσμάτων παρατίθεται ο πίνακας επίδοσης του αλγορίθμου Linear Discriminant Analysis ξεχωριστά τόσο για τις περιπτώσεις χρήσης κινητού τηλεφώνου όσο και για τις περιπτώσεις μη χρήσης.

Πίνακας 5.3 Επίδοση μοντέλου Linear Discriminant Analysis για την χρήση κινητού τηλεφώνου

Οδική Συμπεριφορά	Ακρίβεια	Ανάκληση	F1-Score	Σύνολο δεδομένων εξέτασης
Μη χρήση κινητού (Μη επικίνδυνη)	0.878	0.914	0.896	52113
Χρήση κινητού (Επικίνδυνη)	0.737	0.654	0.693	19120

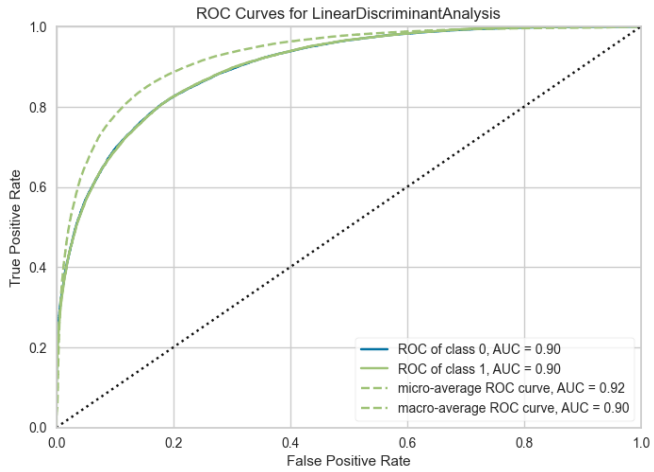
Σύμφωνα με τον Πίνακα 5.3, η τάξη μη χρήσης κινητού τηλεφώνου παρουσιάζει αξιόλογη προβλεπτική ικανότητα της τάξης του 87,8%, όπως επίσης και η τάξη της χρήσης κινητού τηλεφώνου, η οποία παρουσιάζει ακρίβεια σε ποσοστό 73,7%. Επιπλέον, η ανάκληση στην ταξινόμηση της μη Επικίνδυνης τάξης (μη χρήση κινητού) είναι 91,4%, ένα ποσοστό που κρίνεται επίσης αξιόλογο, ενώ για την χρήση κινητού τηλεφώνου η ανάκληση κυμαίνεται στο 65,4%, δηλαδή προβλέπεται λανθασμένα ένα αξιόλογο ποσοστό οδηγών που δεν χρησιμοποιούσαν κινητό (Ψευδώς Αρνητικό), ενώ στην πραγματικότητα χρησιμοποιούσαν κινητό εν ώρα οδήγησης, καθιστώντας το μοντέλο όχι τόσο αξιόπιστο.

Αντίστοιχα, συντάσσεται ο Πίνακας 5.4 για την επίδοση του μοντέλου Logistic Regression.

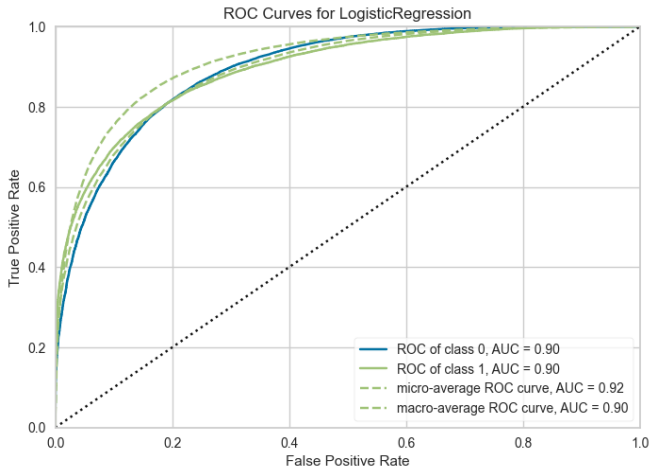
Πίνακας 5.4 Επίδοση μοντέλου Logistic Regression για την χρήση κινητού τηλεφώνου

Οδική Συμπεριφορά	Ακρίβεια	Ανάκληση	F1-Score	Σύνολο δεδομένων εξέτασης
Μη χρήση κινητού (Μη επικίνδυνη)	0.902	0.874	0.888	52394
Χρήση κινητού (Επικίνδυνη)	0.677	0.736	0.705	18839

Η τάξη της μη χρήσης κινητού τηλεφώνου (Μη επικίνδυνη οδηγική συμπεριφορά) παρουσιάζει ακρίβεια σε ένα εξαιρετικό ποσοστό της τάξης του 90,2% με ανάκληση ίση με 87,4%, συνεπώς το μοντέλο της Λογιστικής Παλινδρόμησης προβλέπει ικανοποιητικά την μη χρήση κινητού τηλεφώνου. Όμως, όσον αφορά στην χρήση κινητού το μοντέλο παρουσιάζει χαμηλότερες επιδόσεις ως προς την ακρίβεια και την ανάκληση. Η μειωμένη ακρίβεια δείχνει ότι είναι πιο συντηρητικό το μοντέλο λόγω της αύξησης των Ψευδώς Θετικών αποτελεσμάτων.



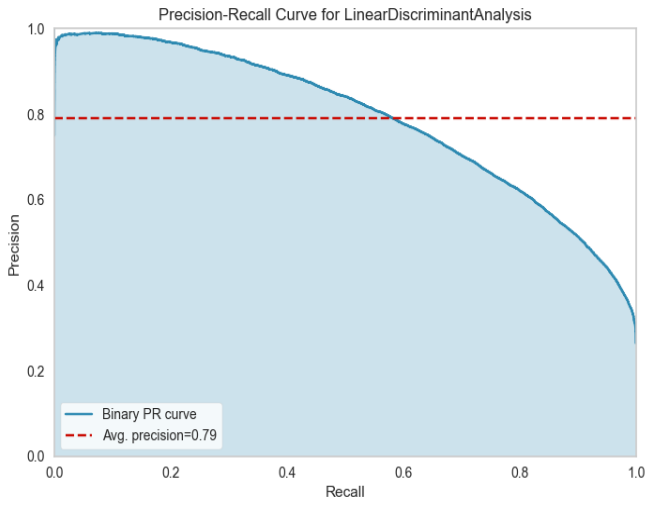
Γράφημα 5.5 Καμπύλη ROC αλγορίθμου Linear Discriminant Analysis για την μεταβλητή της χρήσης κινητού τηλεφώνου.



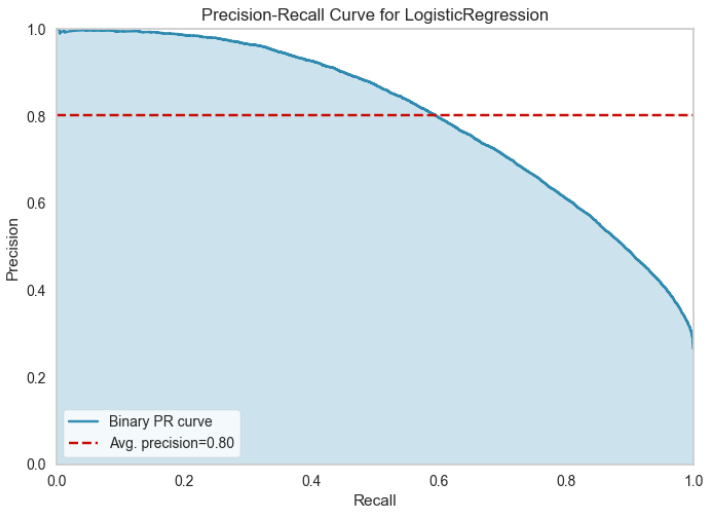
Γράφημα 5.6 Καμπύλη ROC αλγορίθμου Logistic Regression για την μεταβλητή της χρήσης κινητού τηλεφώνου.

Στα Γραφήματα 5.5 και 5.6 αναπαρίστανται οι Καμπύλες ROC για τα 2 μοντέλα. Τα σκορ Περιοχής κάτω από την Καμπύλη (AUC score) υπολογίστηκαν στο 89,5% για το μοντέλο Linear Discriminant Analysis και 89,1% για το Logistic Regression και κρίνονται ικανοποιητικά.

Τα Γραφήματα 5.7 και 5.8 αναπαριστούν τις Καμπύλες Ακρίβειας-Ανάκλησης για τα δύο



Γράφημα5.7 Καμπύλη Ακρίβειας-Ανάκλησης αλγορίθμου Linear Discriminant Analysis για την χρήση κινητού τηλεφώνου

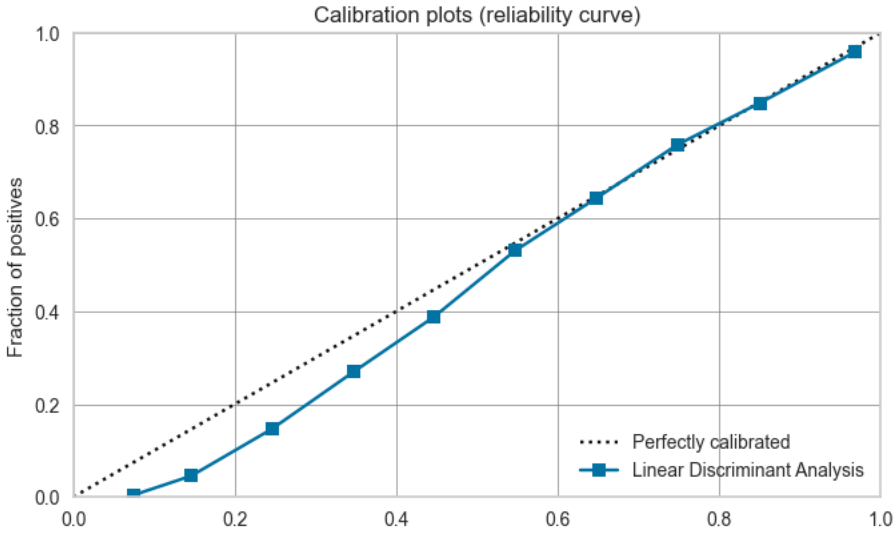


Γράφημα 5.8 Καμπύλη Ακρίβειας-Ανάκλησης αλγορίθμου Logistic Regression για την χρήση κινητού

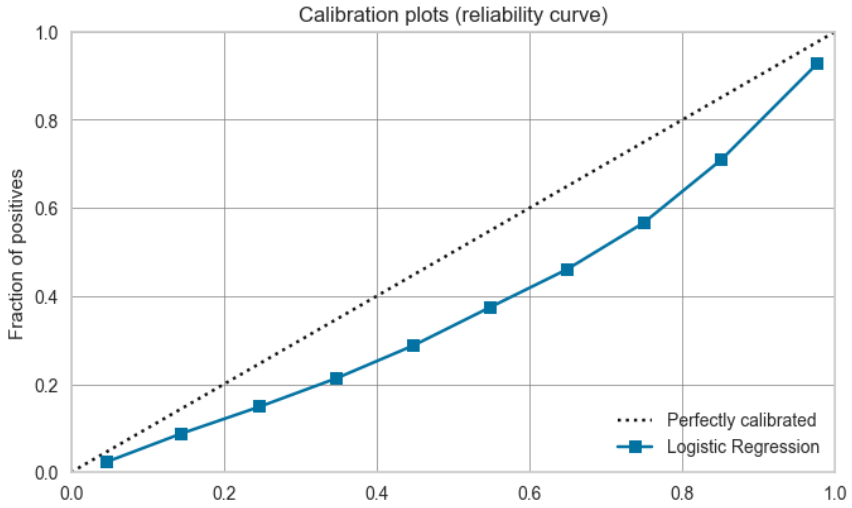
μοντέλα.

Και τα δύο μοντέλα κρίνονται ικανοποιητικά με παρεμφερή αποτελέσματα μέσης ακρίβειας.

Στα Γραφήματα 5.9 και 5.10 παρατίθενται οι καμπύλες βαθμονόμησης των δύο μοντέλων, οι οποίες αποτελούν οπτικό εργαλείο για την αξιολόγηση της συμφωνίας μεταξύ των προβλέψεων και των παρατηρήσεων.



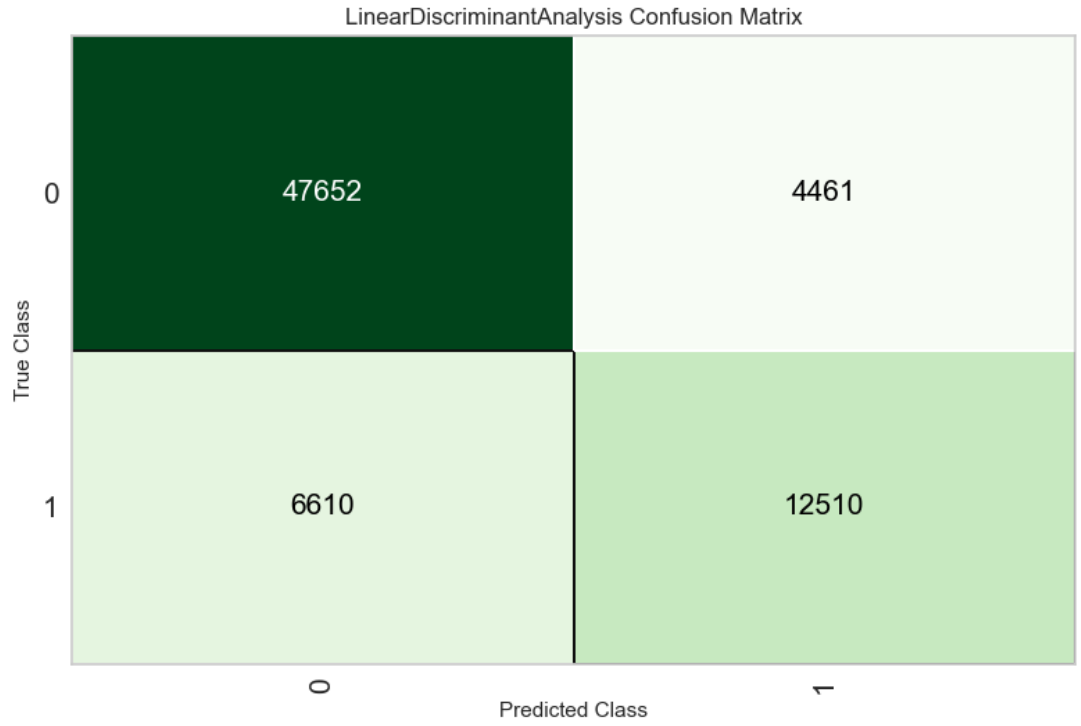
Γράφημα 5.9 Καμπύλη βαθμονόμησης αλγορίθμου Linear Discriminant Analysis για την χρήση κινητού τηλεφώνου



Γράφημα 5.10 Καμπύλη βαθμονόμησης αλγορίθμου Logistic Regression για την χρήση κινητού τηλεφώνου

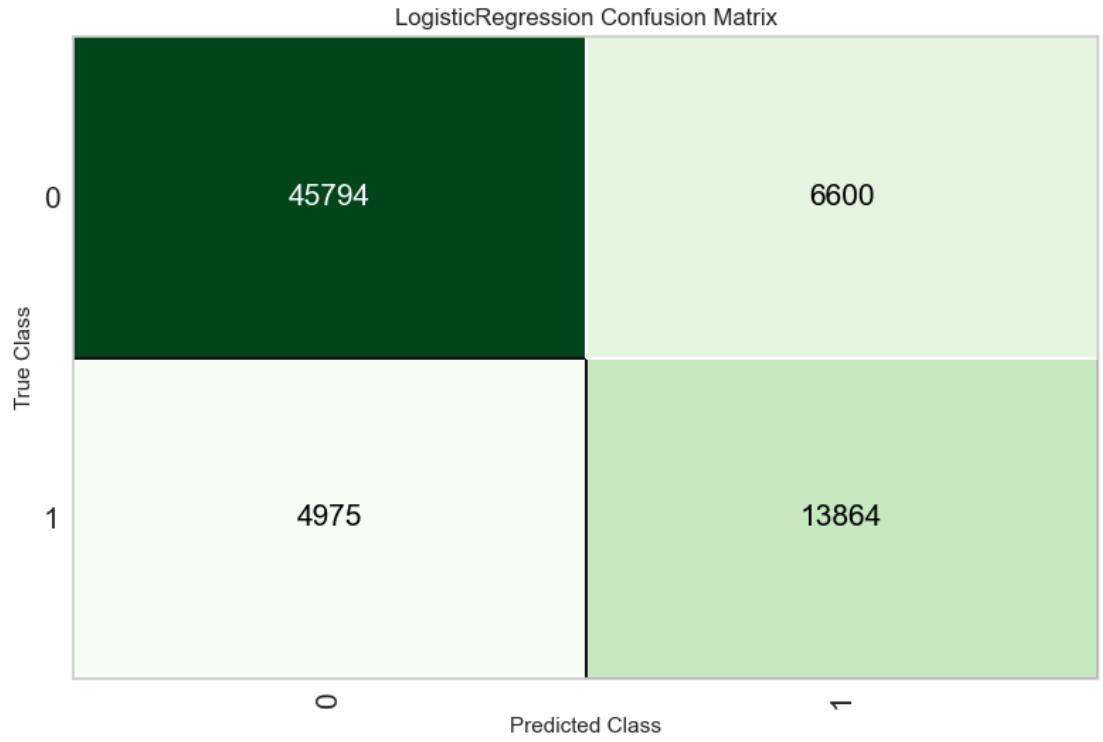
Γίνεται αντιληπτό από τα παραπάνω διαγράμματα ότι ο αλγόριθμος Linear Discriminant Analysis καθίσταται πιο αξιόπιστος σε σχέση με τον Logistic Regression, διότι υφίσταται μεγαλύτερη συμφωνία μεταξύ προβλεπόμενων και πραγματικών τιμών.

Ο αλγόριθμος Linear Discriminant Analysis κατάφερε να ταξινομήσει ορθά μεγάλο ποσοστό των δειγμάτων (84,4% ορθότητα) τόσο για την χρήση κινητού τηλεφώνου όσο και για την μη χρήση. Επιπροσθέτως, για την τάξη της μη χρήσης κινητού το ποσοστό λανθασμένων ταξινομήσεων ανέρχεται σε μόλις 8,5%, ενώ για την τάξη της χρήσης κινητού τηλεφώνου το ποσοστό αυτό έχει τιμή 34,6%, όπως απορρέει από την παρακάτω μήτρα σύγχυσης (Confusion Matrix).



Γράφημα 5.11 Μήτρα σύγχυσης μοντέλου Linear Discriminant Analysis για την χρήση κινητού τηλεφώνου

Ο αλγόριθμος Logistic Regression παρουσιάζει εξίσου υψηλή ορθότητα σε ποσοστό 83,4% εκ του συνόλου, επομένως αποτελεί ένα αξιόπιστο μοντέλο ταξινόμησης. Επίσης, για την τάξη της μη χρήσης κινητού το ποσοστό λανθασμένων ταξινομήσεων ανέρχεται σε μόλις 12,6%, ενώ για την τάξη της χρήσης κινητού τηλεφώνου το ποσοστό αυτό έχει τιμή 26,4%, όπως φαίνεται στην παρακάτω μήτρα σύγχυσης (Confusion Matrix).



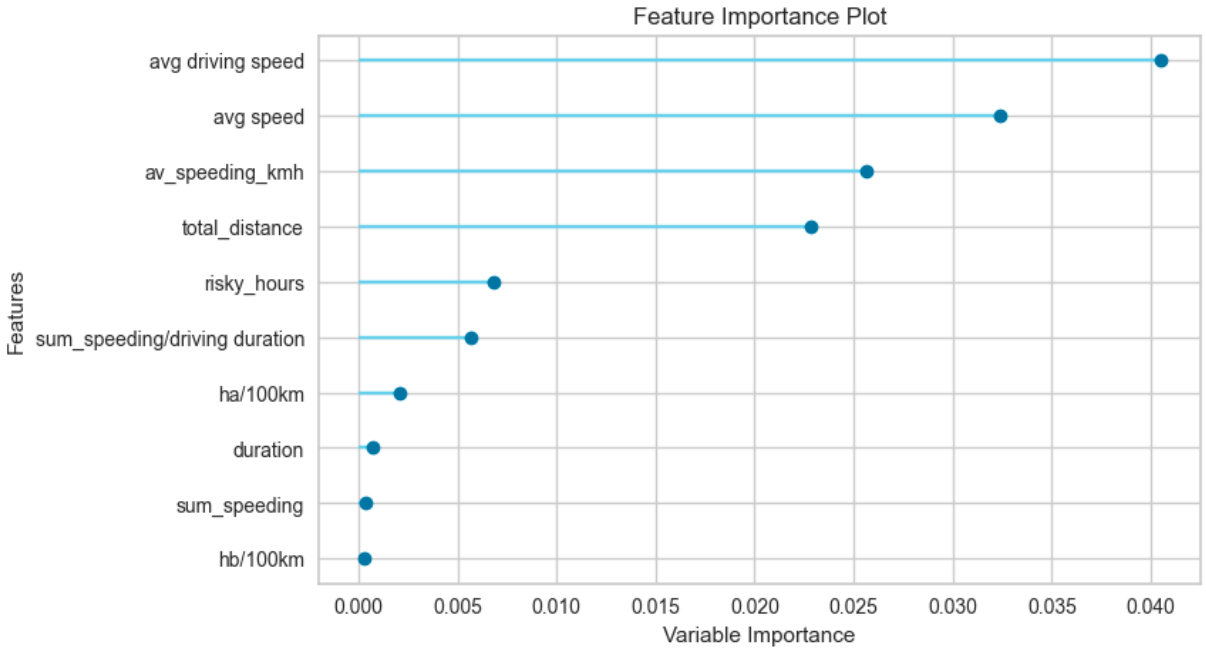
Γράφημα 5.12 Μήτρα σύγχυσης μοντέλου Logistic Regression για την χρήση κινητού τηλεφώνου

Από το Γράφημα 5.12 παρατηρούνται περισσότερα Ψευδώς Θετικά στοιχεία σε σχέση με τα Ψευδώς Αρνητικά στοιχεία σε αντίθεση με το μοντέλο Linear Discriminant Analysis, καθιστώντας το μοντέλο της λογιστικής παλινδρόμησης πιο συντηρητικό.

5.2.2 Σημαντικότητα μεταβλητών (Feature Importance)

Αφότου αναλύθηκαν τα δύο αναφερθέντα μοντέλα επιλέχθηκαν από το σύνολο των υπό εξέταση μεταβλητών εκείνες που συσχετίζονται περισσότερο με την εξαρτημένη μεταβλητή, δηλαδή με την χρήση κινητού τηλεφώνου. Η συγκεκριμένη επιλογή επετεύχθη μέσω της **τεχνικής της Σημαντικότητας Χαρακτηριστικών**. Στόχος της συγκεκριμένης διαδικασίας είναι η ελαχιστοποίηση του υπολογιστικού κόστους και η βελτίωση της προγνωστικής απόδοσης του μοντέλου, μειώνοντας τον αριθμό των μεταβλητών εισόδου. Επιπροσθέτως, με την διαδικασία αυτή μειώνεται η πιθανότητα υπερπροσαρμογής (over-fitting) του μοντέλου, όπως επιβεβαιώνεται και από την διεθνή βιβλιογραφία.

Αρχικά, εκτελείται η παραπάνω διαδικασία για το μοντέλο Linear Discriminant Analysis και προκύπτει το Γράφημα 5.13.



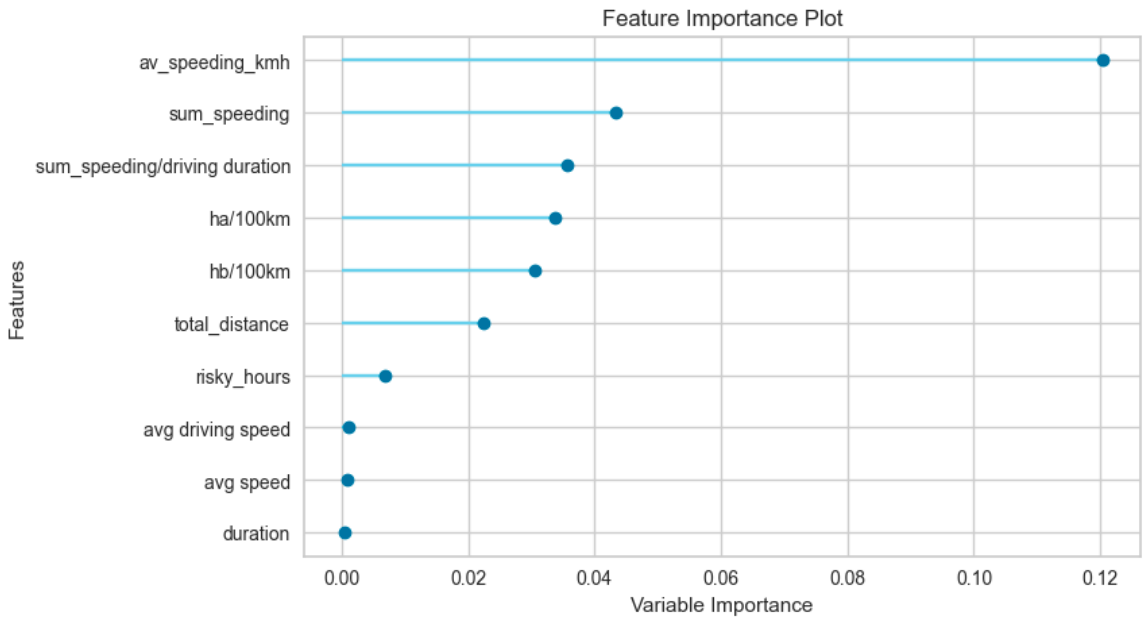
Γράφημα 5.13 Σημαντικότητα μεταβλητών με εξαρτημένη μεταβλητή την χρήση κινητού τηλεφώνου με Linear Discriminant Analysis πριν την εφαρμογή της μεθοδολογίας SMOTE.

Από το παραπάνω διάγραμμα προκρίθηκαν οι ακόλουθες **πέντε** μεταβλητές για περαιτέρω επεξεργασία και ανάλυση.

- Μέση ταχύτητα οδήγησης (km/h).
- Μέση ταχύτητα διαδρομής (km/h).
- Μέση ταχύτητα οδήγησης με υπέρβαση ορίου ταχύτητας και ανοχής σε μια διαδρομή (km/h).
- Συνολική διανυθείσα απόσταση (km).
- Οδηγηθείσα απόσταση στις κρίσιμες ώρες (00:00-05:00) (km).

Σε αυτό το σημείο επισημαίνεται ότι το διάγραμμα της σημαντικότητας των ανεξάρτητων μεταβλητών ως προς την εξαρτημένη μεταβλήθηκε με την εξισορρόπηση των δεδομένων σε σχέση με πριν την υπερδειγματοληψία (Oversampling), γεγονός που απορρέει από το ότι ο αλγόριθμος SMOTE δημιουργεί συνθετικά δείγματα της μειονοτικής κατηγορίας με παρεμβολή μεταξύ των υπαρχόντων δειγμάτων, γεγονός που μπορεί να αλλάξει την κατανομή και τις σχέσεις μεταξύ των χαρακτηριστικών στο σύνολο δεδομένων. Αυτό μπορεί δυνητικά να επηρεάσει τον τρόπο με τον οποίο το μοντέλο μηχανικής μάθησης σταθμίζει και επιλέγει τα χαρακτηριστικά και, επομένως, να αλλάξει το διάγραμμα σπουδαιότητας των χαρακτηριστικών. Επιπλέον, ο αλγόριθμος SMOTE μπορεί να αυξήσει τον αριθμό των δειγμάτων στο σύνολο δεδομένων, γεγονός που μπορεί δυνητικά να αυξήσει την απόδοση του μοντέλου μηχανικής μάθησης και να οδηγήσει σε διαφορετικό διάγραμμα σπουδαιότητας χαρακτηριστικών από ό,τι αν το μοντέλο είχε εκπαιδευτεί στο αρχικό ανισόρροπο σύνολο δεδομένων.

Αντίστοιχα, έγινε παρόμοια διαδικασία για το μοντέλο Logistic Regression και προέκυψε το Γράφημα 5.14.



Γράφημα 5.14 Σημαντικότητα μεταβλητών εξαρτημένης μεταβλητής χρήσης κινητού με Logistic Regression πριν την εφαρμογή της μεθοδολογίας SMOTE.

Από το παραπάνω διάγραμμα προκρίθηκαν οι ακόλουθες **πέντε** μεταβλητές για περαιτέρω επεξεργασία και ανάλυση.

- Μέση ταχύτητα οδήγησης με υπέρβαση ορίου ταχύτητας και ανοχής σε μια διαδρομή (km/h).
- Διάρκεια οδήγησης με υπέρβαση ορίου ταχύτητας και ανοχής (sec).
- Διάρκεια οδήγησης με υπέρβαση ορίου ταχύτητας και ανοχής ανά μονάδα συνολικής διάρκειας οδήγησης χωρίς στάσεις (sec/sec).
- Απότομες επιταχύνσεις ανά 100km (-).
- Απότομες επιβραδύνσεις ανά 100km (-).

5.2.3 Ταξινόμηση δεδομένων με τις πέντε σημαντικότερες ανεξάρτητες μεταβλητές

Στην ακόλουθη υποενότητα τέθηκαν σε στατιστική επεξεργασία και ανάλυση με τις σημαντικότερές τους μεταβλητές τα ίδια δύο επιλεγμένα μοντέλα, δηλαδή αυτό της Γραμμικής Διαχωριστικής Ανάλυσης (Linear Discriminant Analysis) και εκείνο της Λογιστικής Παλινδρόμησης (Logistic Regression), ώστε να αποτελέσουν αντικείμενο σύγκρισης με τα αρχικά μοντέλα που περιλαμβάνουν όλες τις μεταβλητές.

Στον Πίνακα 5.5, ο οποίος συντάχθηκε σε περιβάλλον Jupyter Notebook, παρουσιάζονται οι σημαντικότερες ανεξάρτητες μεταβλητές που επηρεάζουν την εξαρτημένη μεταβλητή χρήσης κινητού τηλεφώνου σύμφωνα με το μοντέλο Linear Discriminant Analysis.

Πίνακας 5.5 Απόσπασμα πίνακα δεδομένων με τις σημαντικότερες μεταβλητές σύμφωνα με το μοντέλο Linear Discriminant Analysis.

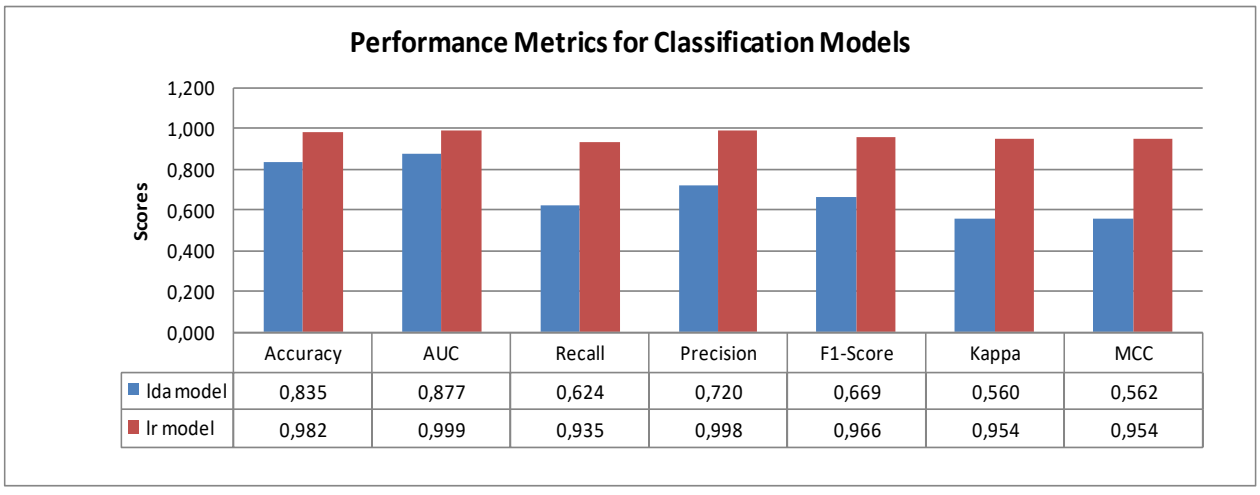
	total_distance	risky_hours	avg speed	avg driving speed	av_speeding_kmh	time_mobile_usage
0	7.818	0.0	60.656897	62.963758	2.767	0
1	2.472	0.0	24.858101	30.064865	0.000	0
2	6.181	0.0	36.358824	44.503200	7.653	0
3	8.085	0.0	26.801105	40.090909	1.978	0
4	18.292	0.0	33.961423	54.693688	4.029	0

Αντίστοιχα, στον Πίνακα 5.6 παρουσιάζονται οι πέντε καταλληλότερες μεταβλητές σύμφωνα με το μοντέλο Logistic Regression.

Πίνακας 5.6 Απόσπασμα πίνακα δεδομένων με τις σημαντικότερες μεταβλητές σύμφωνα με το μοντέλο Logistic Regression.

	ha/100km	hb/100km	sum_speeding/driving duration	sum_speeding	av_speeding_kmh	time_mobile_usage
0	0.000000	12.790995	0.111857	50	2.767	0
1	80.906149	40.453074	0.000000	0	0.000	0
2	48.535836	32.357224	0.174000	87	7.653	0
3	24.737168	0.000000	0.015152	11	1.978	0
4	10.933742	27.334354	0.137874	166	4.029	0

Στην συνέχεια, βελτιστοποιήθηκαν τα παραπάνω μοντέλα (tune_model) και έγινε πρόβλεψη της χρήσης κινητού τηλεφώνου με βάση τα **δεδομένα εξέτασης** (testing dataset). Στο Γράφημα 5.15 φαίνονται τα αποτελέσματα των μετρικών αξιολόγησης, οι οποίες προέκυψαν από την στατιστική ανάλυση των δύο παραπάνω αλγορίθμων.



Γράφημα 5.15 Μετρικές αξιολόγησης μοντέλων ταξινόμησης με τις πέντε σημαντικότερες ανεξάρτητες μεταβλητές

Από το Γράφημα 5.15 προκύπτει ότι το μοντέλο Λογιστικής Παλινδρόμησης με τις πέντε σπουδαιότερες ανεξάρτητες μεταβλητές καθίσταται πιο αξιόπιστο σε σχέση με το μοντέλο της

Γραμμικής Διαχωριστικής Ανάλυσης. Πιο συγκεκριμένα, και τα δύο μοντέλα παρουσιάζουν πολύ υψηλή ορθότητα και σκορ κάτω από την καμπύλη. Επιπροσθέτως, αξιοσημείωτο αποτελεί το γεγονός ότι οι μετρικές αξιολόγησης του μοντέλου της Γραμμικής Διαχωριστικής Ανάλυσης μειώθηκαν ελαφρώς, ενώ αντιθέτως εκείνες του μοντέλου της Λογιστικής Παλινδρόμησης αυξήθηκαν αρκετά σε σχέση με τις μετρικές αξιολόγησης, λαμβάνοντας υπόψη όλες τις ανεξάρτητες μεταβλητές.

Για πληρέστερη παρουσίαση των αποτελεσμάτων παρατίθεται ο πίνακας επίδοσης του αλγορίθμου Linear Discriminant Analysis ξεχωριστά τόσο για τις περιπτώσεις χρήσης κινητού τηλεφώνου όσο και για τις περιπτώσεις μη χρήσης.

Πίνακας 5.7 Επίδοση μοντέλου Linear Discriminant Analysis για την χρήση κινητού τηλεφώνου

Οδική Συμπεριφορά	Ακρίβεια	Ανάκληση	F1-Score	Σύνολο δεδομένων εξέτασης
Μη χρήση κινητού (Μη επικίνδυνη)	0.872	0.911	0.891	52361
Χρήση κινητού (Επικίνδυνη)	0.719	0.628	0.670	18872

Ο αλγόριθμος Linear Discriminant ανάλυσης σύμφωνα τον Πίνακα 5.7 παρουσιάζει επιδόσεις σε παρόμοια επίπεδα και για τις δύο τάξεις με την ανάλυση με όλες τις υπό εξέταση ανεξάρτητες μεταβλητές και μάλιστα ελαφρώς χαμηλότερες. Όμως, τέτοιες μικρές μεταβολές δεν επηρεάζουν πρακτικά την επίδοση του αλγορίθμου.

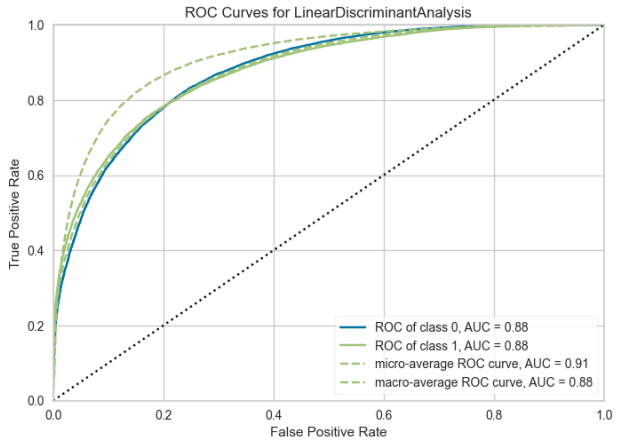
Αντίστοιχα, συντάσσεται ο Πίνακας 5.8 για την επίδοση του μοντέλου Logistic Regression.

Πίνακας 5.8 Επίδοση μοντέλου Logistic Regression για την χρήση κινητού τηλεφώνου

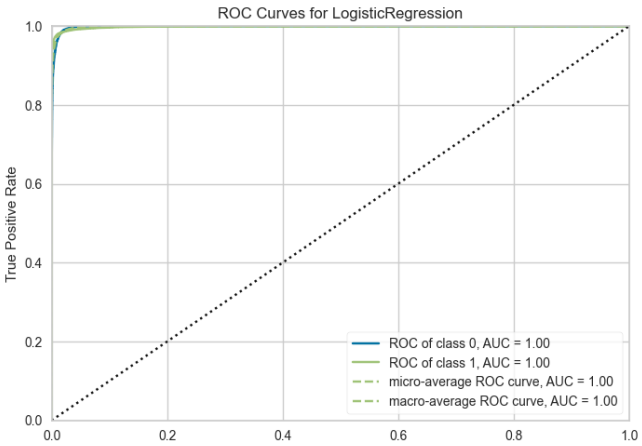
Οδική Συμπεριφορά	Ακρίβεια	Ανάκληση	F1-Score	Σύνολο δεδομένων εξέτασης
Μη χρήση κινητού (Μη επικίνδυνη)	0.977	1.000	0.988	52160
Χρήση κινητού (Επικίνδυνη)	0.999	0.936	0.966	19073

Εδώ παρατηρείται σημαντική βελτίωση της επίδοσης του αλγορίθμου Λογιστικής Παλινδρόμησης με αρκετά μεγάλη ακρίβεια πρόβλεψης της χρήσης κινητού τηλεφώνου σε ποσοστό σχεδόν 100%, δηλαδή με πολύ λίγα Ψευδώς Θετικά στοιχεία. Σε συνδυασμό με την πολύ υψηλή ευαισθησία (ανάκληση), συμπεραίνεται μικρή αναλογία και Ψευδώς αρνητικών στοιχείων, παρέχοντας ασφάλεια για τα αποτελέσματα του μοντέλου.

Στα Γραφήματα 5.16 Και 5.17 αναπαρίστανται οι Καμπύλες ROC για τα 2 μοντέλα. Τα σκορ Περιοχής κάτω από την Καμπύλη (AUC score) υπολογίστηκαν στο 87,7% για το μοντέλο Linear Discriminant Analysis και 99,9% για το Logistic Regression και κρίνονται ικανοποιητικά.

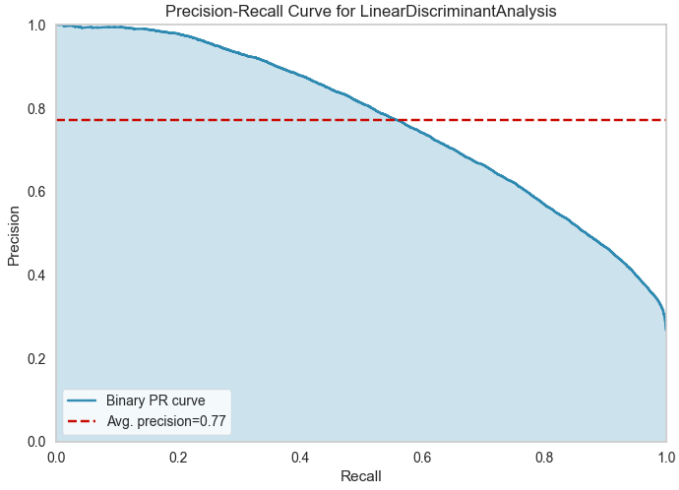


Γράφημα 5.16 Καμπύλη ROC αλγορίθμου Linear Discriminant Analysis για την χρήση κινητού τηλεφώνου.

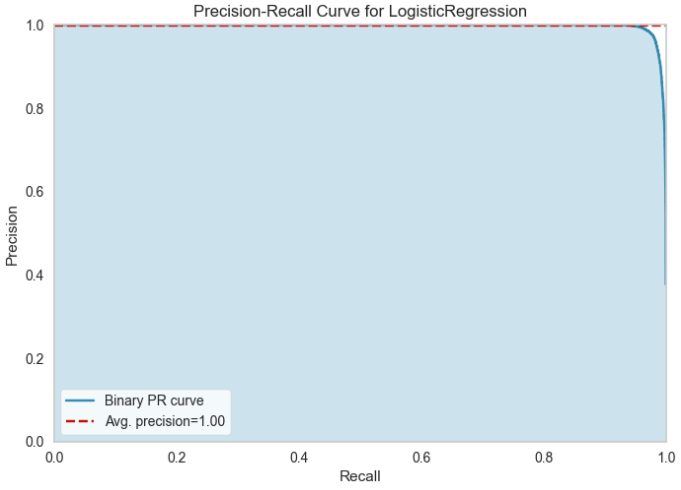


Γράφημα 5.17: Καμπύλη ROC αλγορίθμου Logistic Regression για την χρήση κινητού τηλεφώνου.

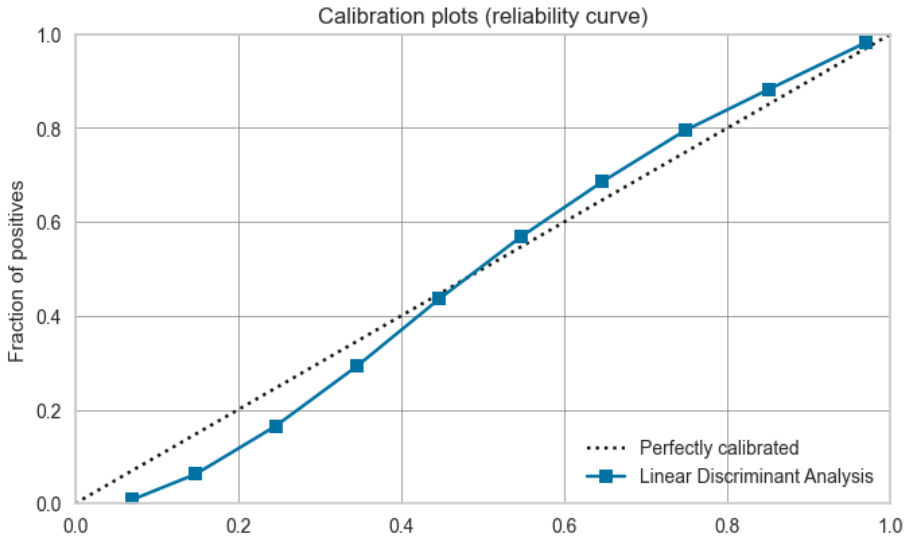
Επιπλέον, παρατίθενται και τα γραφήματα Ακρίβειας-Ανάκλησης των δύο μοντέλων.



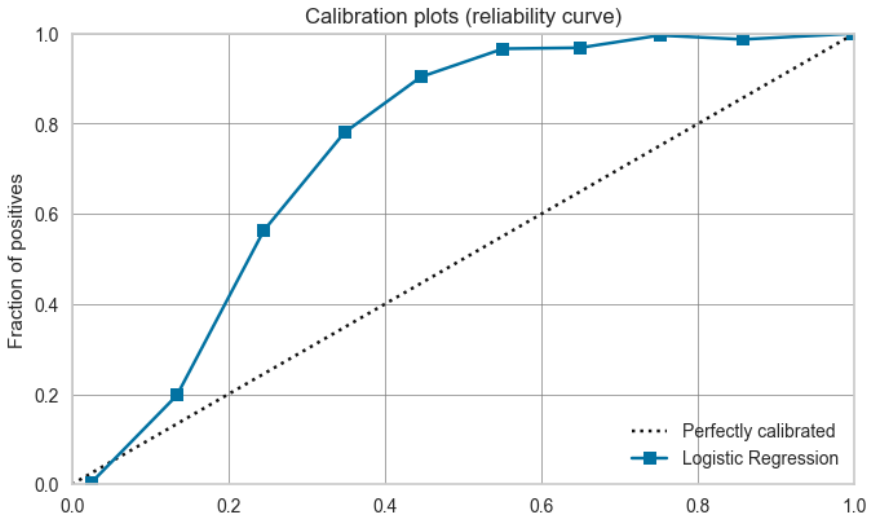
Γράφημα 5.18 Καμπύλη Ακρίβειας-Ανάκλησης αλγορίθμου Linear Discriminant Analysis για την χρήση κινητού τηλεφώνου



Γράφημα 5.19 Καμπύλη Ακρίβειας-Ανάκλησης αλγορίθμου Logistic Regression για την χρήση κινητού τηλεφώνου



Γράφημα 5.20. Καμπύλη βαθμονόμησης αλγορίθμου Linear Discriminant Analysis για χρήση κινητού

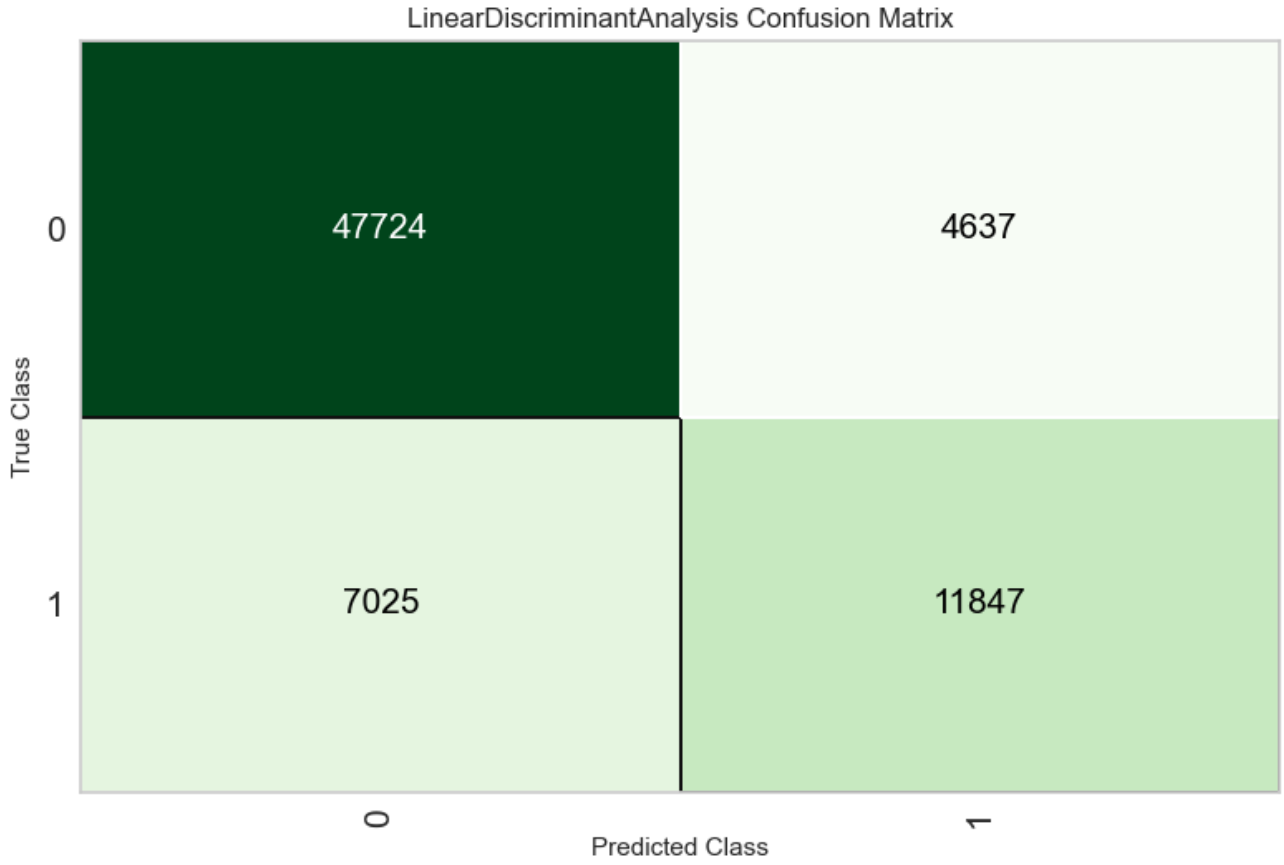


Γράφημα 5.21 Καμπύλη βαθμονόμησης αλγορίθμου Logistic Regression για χρήση κινητού

Συγκρίνοντας τα γραφήματα 5.20 Και 5.21 εξάγεται το συμπέρασμα ότι ο αλγόριθμος Logistic Regression δεν παρουσιάζει υψηλή αξιοπιστία, καθώς οι προβλεπόμενες τιμές αποκλίνουν σε σχέση με τις πραγματικές. Οι υψηλές τιμές στις μετρικές αξιολογήσεις αυτού του μοντέλου ενδεχομένως να οφείλονται σε υπερπροσαρμογή (Overfitting). Αντίθετα το μοντέλο Linear Discriminant Analysis παρουσιάζεται ακόμα πιο βελτιωμένο σε σχέση με το ίδιο μοντέλο λαμβάνοντας υπόψη όλες τις υπό εξέταση ανεξάρτητες μεταβλητές.

Ο αλγόριθμος Linear Discriminant Analysis κατάφερε να ταξινομήσει ορθά μεγάλο ποσοστό των δειγμάτων (83,5% ορθότητα) τόσο για την χρήση κινητού τηλεφώνου όσο και για την μη χρήση.

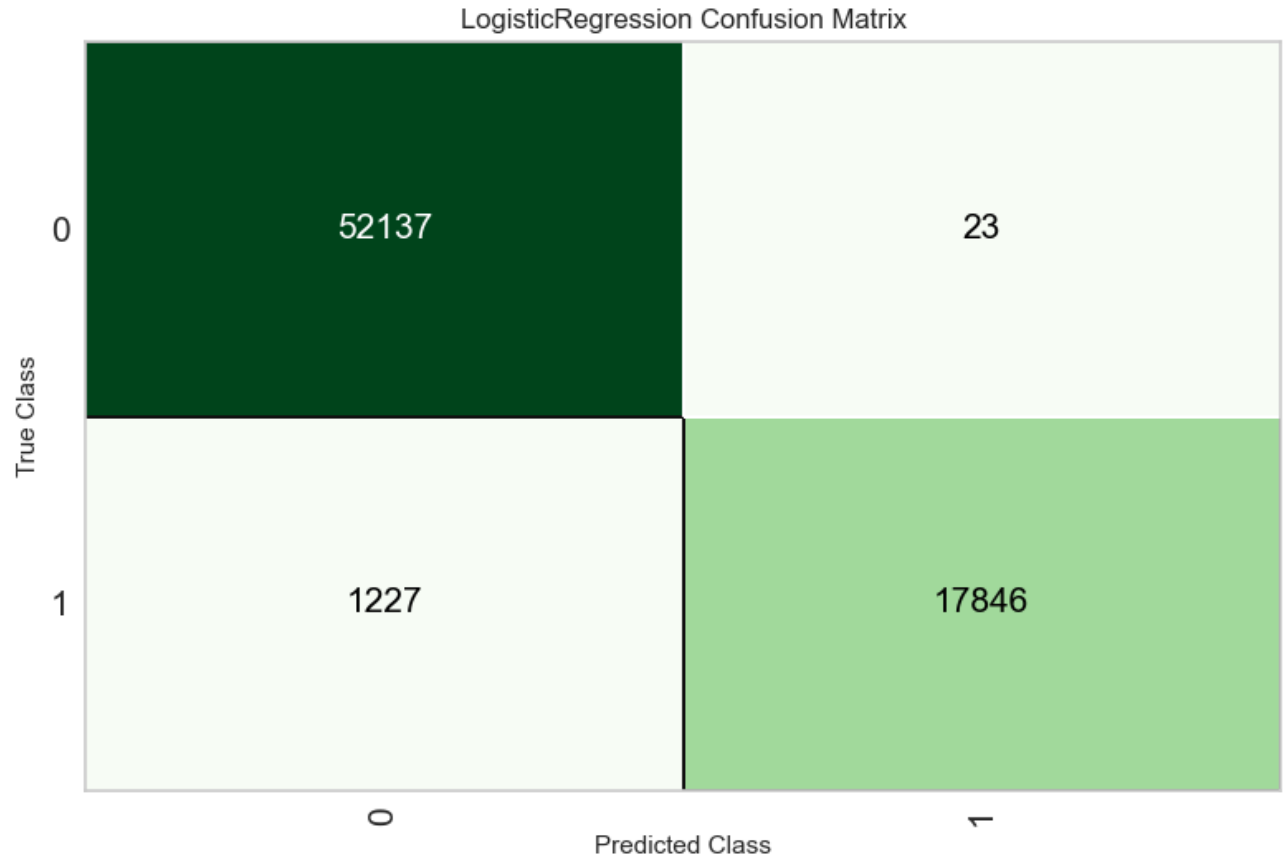
Επιπροσθέτως, για την τάξη της μη χρήσης κινητού το ποσοστό λανθασμένων ταξινομήσεων ανέρχεται σε μόλις 8,9%, ενώ για την τάξη της χρήσης κινητού τηλεφώνου το ποσοστό αυτό έχει τιμή 37,2%, όπως απορρέει από την παρακάτω μήτρα σύγχυσης (Confusion Matrix).



Γράφημα5.22 Μήτρα σύγχυσης μοντέλου Linear Discriminant Analysis για την χρήση κινητού τηλεφώνου

Αξίζει να σημειωθεί ότι παρατηρείται μεγάλη αναλογία Ψευδώς Αρνητικών στοιχείων, επομένως ο ταξινομητής ενδεχομένως να υποτιμά το πλήθος οδηγών που χρησιμοποιούν κινητό τηλέφωνο και να κατατάσσει πολλούς από αυτούς στους μη χρήστες.

Ο αλγόριθμος Logistic Regression φαίνεται να έχει εξαιρετική ορθότητα της τάξης του 98,2%, ιδιαίτερα αυξημένη σε σχέση με την ορθότητα όταν εξετάστηκε το μοντέλο με όλες τις ανεξάρτητες μεταβλητές, με πολύ ικανή προβλεπτική ικανότητα για την τάξη της μη χρήσης κινητού τηλεφώνου. Το μοντέλο καθίσταται αξιόπιστο και για την τάξη της χρήσης κινητού τηλεφώνου, η οποία όπως επιβεβαιώνεται από έρευνες οδηγεί σε επικίνδυνη συμπεριφορά (Caird et al., 2014).



Γράφημα 5.23 Μήτρα σύγχυσης μοντέλου Logistic Regression για την χρήση κινητού τηλεφώνου

Η συγκεκριμένη μήτρα σύγχυσης δίνει πολύ καλύτερα αποτελέσματα από την αντίστοιχη των Linear Discriminative Analysis, με αρκετά μειωμένο ποσοστό FNR και FPR.

5.2.4 Ανάλυση ευαισθησίας

Η ανάλυση ευαισθησίας είναι μια στατιστική τεχνική που χρησιμοποιείται για την αξιολόγηση της ευαισθησίας των αποτελεσμάτων σε αλλαγές στις παραδοχές ή στα δεδομένα εισόδου του μοντέλου. Περιλαμβάνει τη συστηματική μεταβολή των τιμών των παραμέτρων εισόδου για την αξιολόγηση των επιπτώσεων στην έξοδο (Stevens et al, 2014).

Στην περίπτωση του μοντέλου Linear Discriminant Analysis δεν μπορεί να γίνει ανάλυση ευαισθησίας, επειδή οι παραδοχές του μοντέλου είναι σταθερές και δεν μπορούν να μεταβληθούν. Η Γραμμική Διαχωριστική Παλινδρόμηση υποθέτει ότι οι προγνωστικοί παράγοντες είναι κανονικά κατανεμημένοι και ότι ο πίνακας συνδιακύμανσης είναι ο ίδιος σε όλες τις κλάσεις. Αυτές οι υποθέσεις δεν μπορούν να αλλάξουν ή να μεταβληθούν στο μοντέλο. Ωστόσο, εξακολουθεί να είναι δυνατή η αξιολόγηση της απόδοσης με τη χρήση διαφόρων μετρικών αξιολόγησης, όπως η ακρίβεια, η ακρίβεια, η ανάκληση, το F1-score, η καμπύλη ROC, όπως παρουσιάστηκαν σε προηγούμενη υποενότητα.

Για το μοντέλο Logistic Regression καθίσταται δυνατή η ανάλυση ευαισθησίας, η οποία όμως δεν μπορεί να αναπαρασταθεί γραφικά, καθώς η εντολή ανάλυσης ευαισθησίας της Python (interpret()) ικανοποιεί μόνο μοντέλα ταξινόμησης που στηρίζονται σε δέντρα αποφάσεων. Με βάση το διάγραμμα σημαντικότητας των μεταβλητών παρατηρείται ότι η **ταχύτητα** (km/h)

αποτελεί με διαφορά την σπουδαιότερη μεταβλητή σε σχέση με τις υπόλοιπες μεταβλητές. Επίσης, το μοντέλο αυτό παρουσιάζει μεγάλη ευαισθησία/ανάκληση ειδικάμε τις 5 καλύτερες μεταβλητές.

5.3 Διαδικασία Παλινδρόμησης

5.3.1 Παλινδρόμηση δεδομένων με όλες τις ανεξάρτητες μεταβλητές

Για την διαδικασία της παλινδρόμησης έγινε μια αρχική προσπάθεια να γίνει στατιστική ανάλυση μοντέλων με εξαρτημένη μεταβλητή την διάρκεια χρήσης κινητού τηλεφώνου ανά χρονική διάρκεια διαδρομής χωρίς στάσεις (time_mobile_usage/driving_duration). Πιο συγκεκριμένα, αφού αφαιρέθηκαν από το δείγμα ανεξάρτητες μεταβλητές που αποτελούν έκφανση της ίδιας μεταβλητής και άρα έχουν άμεση εξάρτηση με την εξαρτημένη, όπως η διάρκεια χρήσης κινητού τηλεφώνου (time_mobile_usage) και η συνολική διάρκεια διαδρομής χωρίς στάσεις (driving_duration) συντάχθηκε στην γλώσσα προγραμματισμού Python μέσω της εντολής σύγκρισης των μοντέλων (compare_models()) ένας συγκριτικός πίνακας μετρικών αξιολόγησης μοντέλων παλινδόμησης, του οποίου τα αποτελέσματα παρουσιάζονται παρακάτω:

Πίνακας 5.9 :Συγκριτικός Πίνακας μετρικών αξιολόγησης μοντέλων παλινδρόμησης με εξαρτημένη μεταβλητή 'time_mobile_usage/driving_duration'

	Model	MAE	MSE	RMSE	R2	RMSLE	MAPE	TT (Sec)
lightgbm	Light Gradient Boosting Machine	0.0749	0.0203	0.1423	0.0252	0.1100	2.5683	1.1910
gbr	Gradient Boosting Regressor	0.0752	0.0205	0.1431	0.0143	0.1105	2.5941	31.7340
ridge	Ridge Regression	0.0757	0.0206	0.1436	0.0078	0.1109	2.5391	0.0850
lar	Least Angle Regression	0.0757	0.0206	0.1436	0.0078	0.1109	2.5391	0.0770
br	Bayesian Ridge	0.0757	0.0206	0.1436	0.0078	0.1109	2.5394	0.2140
lr	Linear Regression	0.0757	0.0206	0.1436	0.0078	0.1109	2.5391	1.1260
omp	Orthogonal Matching Pursuit	0.0760	0.0207	0.1438	0.0051	0.1110	2.6755	0.0830
en	Elastic Net	0.0765	0.0208	0.1441	0.0003	0.1113	2.4154	0.0930
llar	Lasso Least Angle Regression	0.0765	0.0208	0.1441	-0.0000	0.1114	2.3727	0.0750
lasso	Lasso Regression	0.0765	0.0208	0.1441	-0.0000	0.1114	2.3727	0.0910
dummy	Dummy Regressor	0.0765	0.0208	0.1441	-0.0000	0.1114	2.3727	0.0720
rf	Random Forest Regressor	0.0798	0.0213	0.1461	-0.0278	0.1146	3.0558	135.3440
et	Extra Trees Regressor	0.0821	0.0222	0.1488	-0.0664	0.1174	3.1848	67.2970
huber	Huber Regressor	0.0489	0.0230	0.1515	-0.1056	0.1176	1.0057	5.5430
knn	K Neighbors Regressor	0.0774	0.0242	0.1555	-0.1637	0.1235	2.8322	97.2180
ada	AdaBoost Regressor	0.1460	0.0299	0.1728	-0.4380	0.1454	6.7198	4.9080
dt	Decision Tree Regressor	0.0892	0.0431	0.2076	-1.0748	0.1599	3.7935	2.2000
par	Passive Aggressive Regressor	0.2938	0.1585	0.3753	-6.6616	0.2726	24.2228	0.2150

Με βάση τον Πίνακα 5.9 παρατηρούνται εξαιρετικά χαμηλές τιμές του δείκτη προσδιορισμού R^2 για όλα τα μοντέλα και μάλιστα αρνητικές τιμές, γεγονός που υποδηλώνει ότι οι ανεξάρτητες μεταβλητές δεν συσχετίζονται με την εξαρτημένη και αυτό θα οδηγήσει σε εντελώς αναξιόπιστα μοντέλα. Αρνητικός συντελεστής προσδιορισμού R^2 υποδηλώνει ότι οι προβλέψεις έχουν χειρότερη ερμηνεία και από τυχηματικά γεγονότα.

Συνεπώς, εγκαταλείφθηκε η συγκεκριμένη αρχική προσπάθεια και στην συνέχεια εξετάστηκε η επιρροή των ανεξάρτητων μεταβλητών στην **εξαρτημένη μεταβλητή** της **διάρκειας χρήσης κινητού τηλεφώνου** (time_mobile_usage).

Σε προγραμματιστικό φύλλο Jupyter Notebook συντάχθηκε ο Πίνακας 5.10 , ο οποίος αποτελεί τον συγκριτικό πίνακα μετρικών αξιολόγησης για τα μοντέλα παλινδρόμησης με εξαρτημένη την συνεχή μεταβλητή διάρκειας χρήσης κινητού τηλεφώνου (time_mobile_usage).

Πίνακας 5.10 Συγκριτικός πίνακας μετρικών αξιολόγησης μοντέλων παλινδρόμησης

	Model	MAE	MSE	RMSE	R2	RMSLE	MAPE	TT (Sec)
et	Extra Trees Regressor	0.3123	38.5291	5.4327	0.9986	0.0054	0.0053	29.0870
rf	Random Forest Regressor	0.3618	53.9818	6.8025	0.9980	0.0055	0.0054	60.0150
lightgbm	Light Gradient Boosting Machine	1.2870	112.7824	10.1854	0.9957	0.1095	0.0797	1.3140
gbr	Gradient Boosting Regressor	4.0636	111.0895	10.4536	0.9956	0.8167	0.2684	38.1490
dt	Decision Tree Regressor	0.9873	145.4560	11.8645	0.9943	0.0143	0.0142	0.9120
ada	AdaBoost Regressor	63.3461	9379.4487	96.3039	0.6275	3.2195	3.3666	17.0730
lr	Linear Regression	40.1748	12227.7045	110.3907	0.5138	2.5889	1.7137	1.1530
ridge	Ridge Regression	40.1721	12227.7051	110.3907	0.5138	2.5888	1.7136	0.0740
br	Bayesian Ridge	40.1741	12227.7036	110.3907	0.5138	2.5889	1.7136	0.2840
lasso	Lasso Regression	39.1998	12277.4618	110.6064	0.5120	2.5510	1.6886	8.2780
lar	Least Angle Regression	40.7294	12293.1050	110.6855	0.5112	2.6044	1.7454	0.0820
omp	Orthogonal Matching Pursuit	29.4625	14142.7866	118.7358	0.4375	1.4762	1.0754	0.0780
par	Passive Aggressive Regressor	43.3368	20626.6727	142.8514	0.1842	2.2288	1.6530	1.6510
en	Elastic Net	55.2459	22040.3843	148.3241	0.1215	2.7501	2.4495	7.7830
llar	Lasso Least Angle Regression	62.4061	25086.2039	158.2467	-0.0000	3.2552	2.0155	0.0810
dummy	Dummy Regressor	62.4061	25086.2039	158.2467	-0.0000	3.2552	2.0155	0.0450
huber	Huber Regressor	39.2397	26441.9989	162.4792	-0.0544	1.9456	0.9598	5.5240
knn	K Neighbors Regressor	59.9063	26911.6422	163.9491	-0.0744	2.6434	2.7920	2.5940

Από τα παραπάνω μοντέλα παλινδρόμησης επιλέχθηκαν αρχικά μοντέλα που παρουσίασαν πολύ υψηλό συντελεστή R^2. Ειδικότερα, επιλέχθηκαν τα μοντέλα παλινδρόμησης Ενίσχυσης Κλίσης (Gradient Boosting) και Δέντρων Απόφασης (Decision Trees), τα οποία όμως με την εντολή της βελτίωσης των μοντέλων (tune_model) και την πρόβλεψη για τα testing data (80% training data και 20% testing data) παρουσίασαν συντελεστή προσδιορισμού $R^2_{GBR}=0.999$ και $R^2_{DT}=1.000$ αντίστοιχα. Οι συγκεκριμένες τιμές του συντελεστή προσδιορισμού R^2 δηλώνουν

υπερπροσαρμογή μοντέλου (overfitting), γεγονός που αποτελεί σημαντική πρόκληση σε προβλήματα μηχανικής μάθησης.

Συνεπώς, ύστερα από αρκετές δοκιμές επιλέχθηκαν δύο μοντέλα παλινδρόμησης για περαιτέρω στατιστική επεξεργασία και ανάλυση, τα οποία παρουσίασαν ενδιάμεσες τιμές στον συντελεστή προσδιορισμού R^2. Αυτά καθίστανται τα μοντέλα Προσαρμοστικής Ενδυνάμωσης (ada) και Γραμμικής Παλινδρόμησης (lr), όπως παρουσιάζονται στον Πίνακα 5.11.

Πίνακας 5.11 Επιλεγμένα μοντέλα παλινδρόμησης

Αγγλική ονομασία αλγορίθμου	Ελληνική ονομασία αλγορίθμου	Συμβολισμός
AdaBoost Regressor	Προσαρμοστική Ενδυνάμωση	ADA
Linear Regression	Γραμμική Παλινδρόμηση	LR

Στην συνέχεια, βελτιστοποιήθηκαν τα παραπάνω μοντέλα (tune_model) και έγινε πρόβλεψη της διάρκειας χρήσης κινητού τηλεφώνου με βάση τα **δεδομένα εξέτασης** (testing dataset). Στον Πίνακα 5.12 φαίνονται οι συντελεστές προσδιορισμού R^2, οι οποίοι προέκυψαν από την στατιστική ανάλυση των δύο παραπάνω μοντέλων.

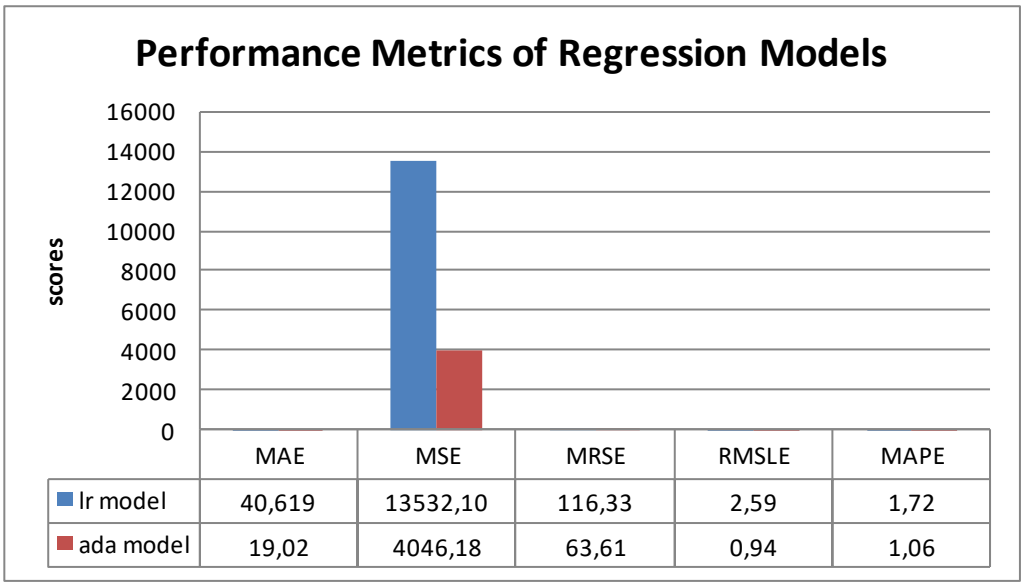
Πίνακας 5.12 Συντελεστής προσδιορισμού R^2

Μοντέλα Παλινδρόμησης	Συντελεστής προσδιορισμού R^2	Χαρακτηρισμός Μοντέλου
AdaBoost Regressor	R^2=0.842	Πολύ Ικανοποιητικό
Linear Regression	R^2=0.497	Μη ικανοποιητικό

Γενικά, ικανοποιητικές τιμές του $\mathbf{R^2}$ θεωρούνται εκείνες που κυμαίνονται στο διάστημα από **0.8 έως 0.9**, διότι μεγαλύτερες τιμές ενδέχεται να δηλώνουν υπερπροσαρμογή (overfitting) και για μικρότερες τιμές από 0.8 το μοντέλο δεν καθίσταται τόσο αξιόπιστο. Άξιο αναφοράς αποτελεί το γεγονός ότι ο συντελεστής προσδιορισμού R^2 καθίσταται η πιο κατατοπιστική απλή μετρική στην αξιολόγηση αναλύσεων παλινδρόμησης (Chicco et al., 2021).

Συνεπώς, για το μοντέλο παλινδρόμησης AdaBoost οι ανεξάρτητες μεταβλητές έχουν υψηλή ικανότητα ερμηνείας της εξαρτημένης μεταβλητής. Αντίθετα, για το μοντέλο παλινδρόμησης Linear Regression οι ανεξάρτητες μεταβλητές μπορούν να ερμηνεύσουν μόλις το 50% (49.7%) της διακύμανσης της διάρκειας χρήσης κινητού τηλεφώνου που αποτελεί την εξαρτημένη μεταβλητή, καθιστώντας το μοντέλο μη ικανοποιητικό.

Στο παρακάτω συγκριτικό Γράφημα 5.24 απεικονίζονται οι τιμές για τις υπόλοιπες μετρικές αξιολόγησης των μοντέλων παλινδρόμησης για τα 2 προαναφερθέντα μοντέλα, όπως ορίστηκαν και στο κεφάλαιο 3.

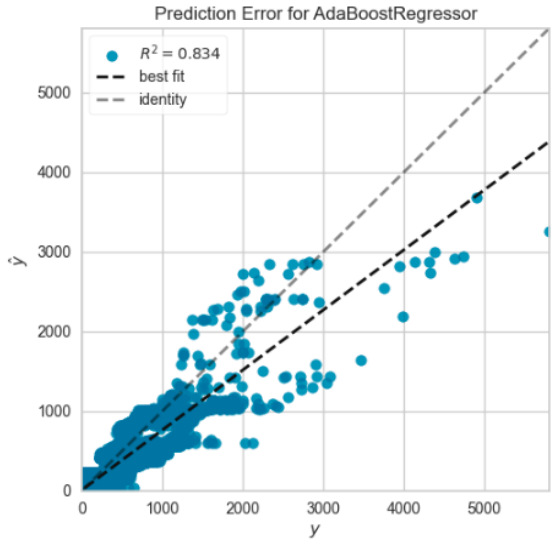


	MAE	MSE	MRSE	RMSLE	MAPE
lr model	40,619	13532,10	116,33	2,59	1,72
ada model	19,02	4046,18	63,61	0,94	1,06

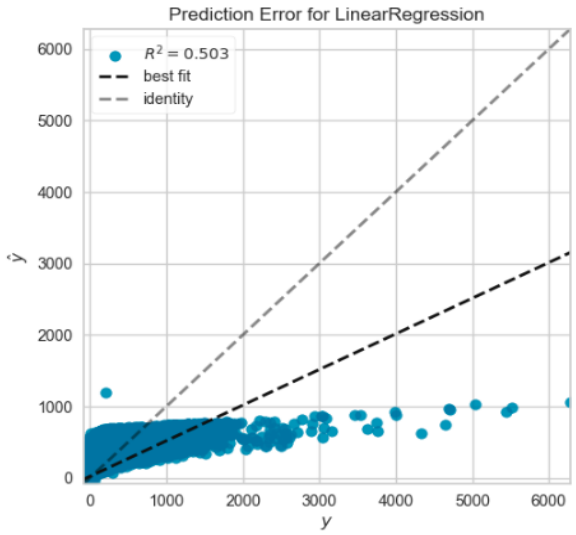
Γράφημα 5.24 Μετρικές αξιολόγησης μοντέλων παλινδρόμησης

Γίνεται αντιληπτό ότι οι τιμές των μετρικών αξιολόγησης είναι ανάλογες της τάξης μεγέθους των τιμών του δείγματος. Ειδικά η τιμή του μέσου τετραγωνικού σφάλματος είναι εξαιρετικά μεγάλη για το μοντέλο της Γραμμικής Παλινδρόμησης (Linear Regression), γεγονός που παράλληλα με το χαμηλό R^2 σε αποδεικνύει την αναξιοπιστία του. Όμως, παρατηρείται ότι η τιμή του μέσου απόλυτου εκατοστιαίου σφάλματος είναι κάτω από 10% και για τα δύο μοντέλα, επομένως σύμφωνα και με το 3^{ο} κεφάλαιο οι τιμές αυτές δηλώνουν εξαιρετικά αξιόπιστα μοντέλα.

Στα Γραφήματα 5.25 Και 5.26 απεικονίζονται οι αποκλίσεις μεταξύ πραγματικών και προβλεπόμενων τιμών για τα μοντέλα AdaBoost και Linear Regression.



Γράφημα 5.25 Σφάλμα πρόβλεψεων για μοντέλο AdaBoost



Γράφημα 5.26 Σφάλμα προβλέψεων για μοντέλο Linear Regression

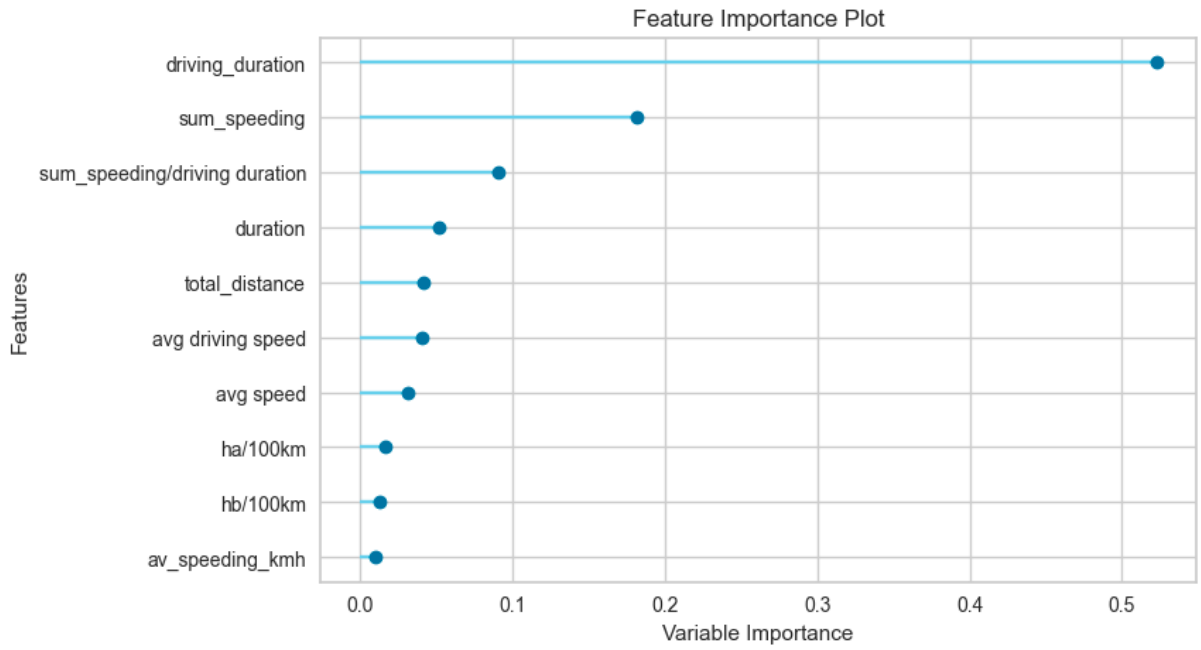
Με βάση τα παραπάνω διαγράμματα γίνεται πιο εύληπτο ότι οι τιμές του R2 ακολουθούν την ευθεία των προβλέψεων στο διάγραμμα 5.25 σε αντίθεση με το διάγραμμα 5.26, καθιστώντας

το AdaBoost το πιο αξιόπιστο μοντέλο εκ των δύο. Επιπλέον, παρατηρείται αυξημένη απόκλιση σε μεγάλες τιμές.

5.3.2 Σημαντικότητα Μεταβλητών (Feature Importance)

Αφότου αναλύθηκαν τα δύο αναφερθέντα μοντέλα επιλέχθηκαν από το σύνολο των υπό εξέταση μεταβλητών εκείνες που συσχετίζονται περισσότερο με την εξαρτημένη μεταβλητή, δηλαδή με την διάρκεια χρήσης κινητού τηλεφώνου. Ομοίως με την ταξινόμηση, η συγκεκριμένη επιλογή επετεύχθη μέσω της **τεχνικής της Σημαντικότητας Χαρακτηριστικών**. Στόχος της συγκεκριμένης διαδικασίας είναι η ελαχιστοποίηση του υπολογιστικού κόστους και η βελτίωση της προγνωστικής απόδοσης του μοντέλου, μειώνοντας τον αριθμό των μεταβλητών εισόδου. Επιπροσθέτως, με την διαδικασία αυτή μειώνεται η πιθανότητα υπερπροσαρμογής (over-fitting) του μοντέλου, όπως επιβεβαιώνεται και από την διεθνή βιβλιογραφία.

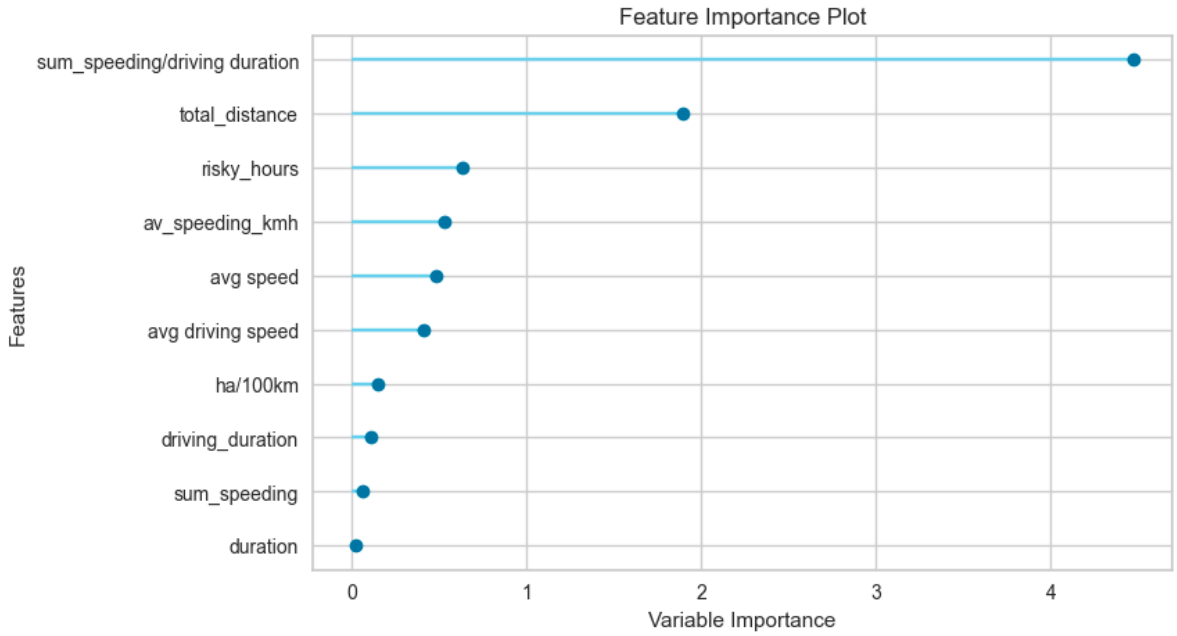
Αρχικά, εκτελείται η παραπάνω διαδικασία για το μοντέλο AdaBoost και προκύπτει το Γράφημα 5.27.



Γράφημα 5.27 Σημαντικότητα μεταβλητών για την εξαρτημένη μεταβλητή time_mobile_usage με AdaBoost Regressor

Από το παραπάνω διάγραμμα εξυπακούεται ότι η συνολική διάρκεια οδήγησης χωρίς στάσεις, η συνολική διάρκεια υπέρβασης του ορίου ταχύτητας και ανοχής, η συνολική διάρκεια διαδρομής και η συνολική διανυθείσα απόσταση αποτελούν τις μεταβλητές που συσχετίζονται περισσότερο με την εξαρτημένη μεταβλητή, επομένως προκρίθηκαν οι συγκεκριμένες μεταβλητές για περαιτέρω επεξεργασία και ανάλυση.

Αντίστοιχα, έγινε παρόμοια διαδικασία για το μοντέλο Linear Regression και προέκυψε το Γράφημα 5.28.



Γράφημα 5.28 Σημαντικότητα μεταβλητών για την εξαρτημένη μεταβλητή time_mobile_usage με Linear Regression

Από το παραπάνω διάγραμμα προκρίθηκαν οι ακόλουθες μεταβλητές για περαιτέρω επεξεργασία και ανάλυση.

- Διάρκεια οδήγησης με υπέρβαση ορίου ταχύτητας και ανοχής ανά μονάδα συνολικής διάρκειας οδήγησης χωρίς στάσεις (sec/sec).
- Συνολική διανυθείσα απόσταση (km).
- Οδηγηθείσα απόσταση στις κρίσιμες ώρες (00:00-05:00) (km).
- Μέση ταχύτητα οδήγησης με υπέρβαση ορίου ταχύτητας και ανοχής σε μια διαδρομή (km/h).
- Μέση ταχύτητα διαδρομής (km/h).

5.3.3 Παλινδρόμηση δεδομένων με τις καταλληλότερες μεταβλητές

Στην ακόλουθη υποενότητα τέθηκαν σε στατιστική επεξεργασία και ανάλυση με τις σημαντικότερές τους μεταβλητές τα ίδια δύο επιλεγμένα μοντέλα, δηλαδή αυτό της Προσαρμοστικής Ενδυνάμωσης (AdaBoost Regressor) και εκείνο της Γραμμικής Παλινδρόμησης (Linear Regression), ώστε να αποτελέσουν αντικείμενο σύγκρισης με τα αρχικά μοντέλα που περιλαμβάνουν όλες τις μεταβλητές.

Στον Πίνακα 5.13 , ο οποίος συντάχθηκε σε περιβάλλον Jupyter Notebook, παρουσιάζονται οι σημαντικότερες ανεξάρτητες μεταβλητές που επηρεάζουν την εξαρτημένη μεταβλητή διάρκειας χρήσης κινητού τηλεφώνου σύμφωνα με το μοντέλο AdaBoost.

Πίνακας 5.13 Απόσπασμα πίνακα δεδομένων με τις σημαντικότερες μεταβλητές σύμφωνα με το μοντέλο AdaBoost

	duration	total_distance	sum_speeding/driving duration	sum_speeding	time_mobile_usage	driving_duration
0	464	7.818	0.111857	50	0	447
1	358	2.472	0.000000	0	0	296
2	612	6.181	0.174000	87	0	500
3	1086	8.085	0.015152	11	0	726
4	1939	18.292	0.137874	166	0	1204

Στον Πίνακα 5.14, ο οποίος επίσης συντάχθηκε σε περιβάλλον Jupyter Notebook, παρουσιάζονται οι σημαντικότερες ανεξάρτητες μεταβλητές που επηρεάζουν την εξαρτημένη μεταβλητή διάρκειας χρήσης κινητού τηλεφώνου σύμφωνα με το μοντέλο Linear Regression.

Πίνακας 5.14 Απόσπασμα πίνακα δεδομένων με τις σημαντικότερες μεταβλητές σύμφωνα με το μοντέλο Linear Regression

	total_distance	risky_hours	avg speed	sum_speeding/driving duration	av_speeding_kmh	time_mobile_usage
0	7.818	0.0	60.656897	0.111857	2.767	0
1	2.472	0.0	24.858101	0.000000	0.000	0
2	6.181	0.0	36.358824	0.174000	7.653	0
3	8.085	0.0	26.801105	0.015152	1.978	0
4	18.292	0.0	33.961423	0.137874	4.029	0

Στην συνέχεια, βελτιστοποιήθηκαν τα παραπάνω μοντέλα (tune_model) και έγινε πρόβλεψη της διάρκειας χρήσης κινητού τηλεφώνου με βάση τα **δεδομένα εξέτασης** (testing dataset). Η αναλογία **δεδομένων εκπαίδευσης** (training dataset) και **δεδομένων εξέτασης** (testing dataset) ήταν **80% και 20%** αντίστοιχα. Στον Πίνακα 5.15 φαίνονται οι συντελεστές προσδιορισμού R^2, οι οποίοι προέκυψαν από την στατιστική ανάλυσή των δύο παραπάνω μοντέλων.

Πίνακας 5.15 Συντελεστής προσδιορισμού R^2 για τα μοντέλα παλινδρόμησης (με όλες τις καταλληλότερες μεταβλητές)

Μοντέλα Παλινδρόμησης	Συντελεστής προσδιορισμού R^2	Χαρακτηρισμός Μοντέλου
AdaBoost Regressor	R^2=0.840	Πολύ Ικανοποιητικό
Linear Regression	R^2=0.422	Μη ικανοποιητικό

Συνεπώς, για το μοντέλο παλινδρόμησης AdaBoost οι ανεξάρτητες μεταβλητές έχουν υψηλή ικανότητα ερμηνείας της διακύμανσης της εξαρτημένης μεταβλητής. Αντίθετα, για το μοντέλο

παλινδρόμησης Linear Regression οι ανεξάρτητες μεταβλητές μπορούν να ερμηνεύσουν μόλις το 42,2% (R^2=0.422) της διακύμανσης της διάρκειας χρήσης κινητού τηλεφώνου που αποτελεί την εξαρτημένη μεταβλητή, καθιστώντας το μοντέλο μη ικανοποιητικό.

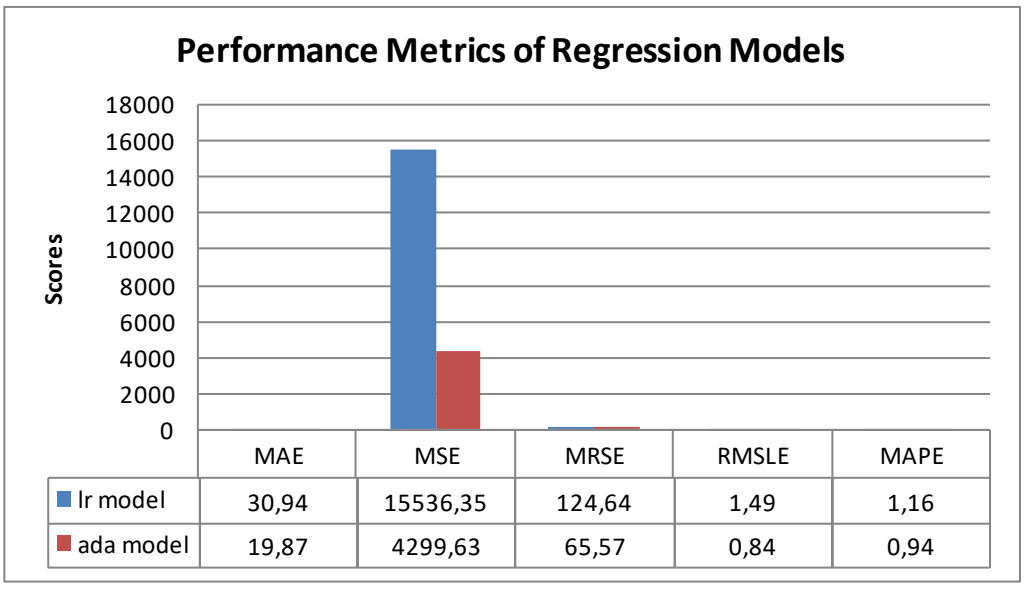
Παρακάτω παρατίθεται ο συγκριτικός Πίνακας 5.16 μεταξύ των συντελεστών προσδιορισμού για τα δύο μοντέλα.

Πίνακας 5.16 Συγκριτικός Πίνακας συντελεστών R2 πριν και μετά την μείωση των ανεξάρτητων μεταβλητών

Μοντέλα Παλινδρόμησης	Συντελεστής προσδιορισμού R^2 με όλες τις υπό εξέταση μεταβλητές	Συντελεστής προσδιορισμού R^2 με τις σημαντικότερες μεταβλητές
AdaBoost Regressor	R^2=0.842	R^2=0.840
Linear Regression	R^2=0.497	R^2=0.422

Από τον Πίνακα 5.16 αξίζει να σημειωθεί ότι τα μοντέλα με τις σημαντικότερες μεταβλητές παρουσιάζουν ελαφρώς χαμηλότερο συντελεστή προσδιορισμού R^2 σε σχέση με τα αρχικά και αυτό απορρέει από το γεγονός ότι ο συντελεστής προσδιορισμού R^2 λαμβάνει υπόψη του τον συνδυασμό της καλύτερης ερμηνείας της εξαρτημένης μεταβλητής και της ταυτόχρονης αξιοποίησης όσο το δυνατόν περισσότερων μεταβλητών.

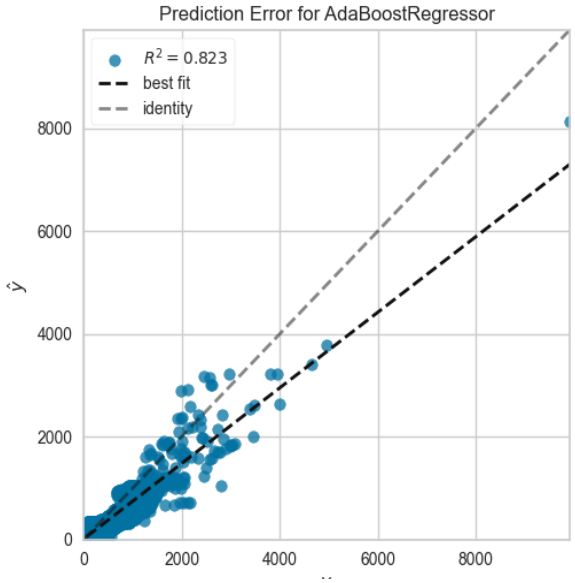
Στο παρακάτω συγκριτικό Γράφημα 5.27 απεικονίζονται για λόγους πληρότητας της συγκεκριμένης διπλωματικής εργασίας οι τιμές για τις υπόλοιπες μετρικές αξιολόγησης των μοντέλων παλινδρόμησης για τα δύο προαναφερθέντα μοντέλα με τις επιλεγμένες ανεξάρτητες μεταβλητές, όπως ορίστηκαν και στο κεφάλαιο 3.



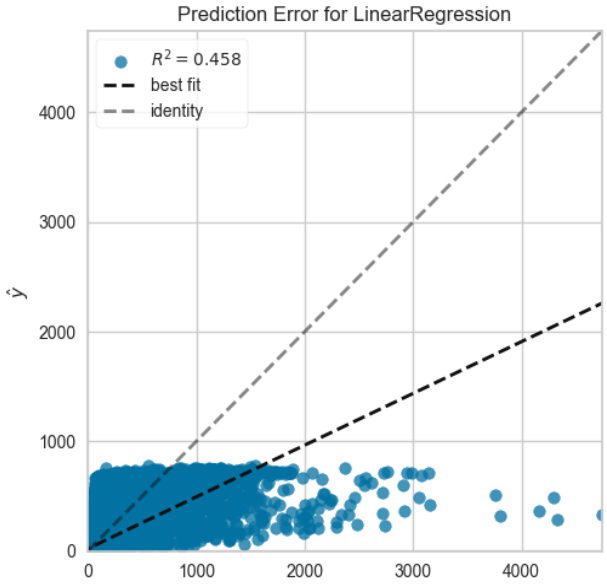
Γράφημα 5.27. Μετρικές αξιολόγησης μοντέλων παλινδρόμησης

Παρατηρούνται ελαφρώς προσαυξημένα σφάλματα στα μοντέλα με τις σημαντικότερες μεταβλητές σε σχέση με εκείνα των ίδιων μοντέλων με όλες τις υπό εξέταση μεταβλητές (Γράφημα 5.24).

Τέλος, παρουσιάζονται υπό μορφή γραφημάτων οι αποκλίσεις μεταξύ πραγματικών και προβλεπόμενων τιμών.



Γράφημα 5.28 Σφάλμα προβλέψεων μοντέλου AdaBoost



Γράφημα 5.29 Σφάλμα προβλέψεων μοντέλου Linear Regression

5.4 Σύνοψη

Σε αυτό το σημείο επιχειρείται μια επιγραμματική σύνοψη των προηγούμενων υποκεφαλαίων, για την οποία θα γίνει εκτενέστερη αναφορά στο επόμενο κεφάλαιο.

Για την διαδικασία της **ταξινόμησης** προέκυψε η ανεξάρτητη μεταβλητή της **ταχύτητας (km/h)** ως σημαντικότερη μεταβλητή και από τις αναλύσεις των μοντέλων επιλέχθηκε τελικά το μοντέλο της Γραμμικής Διαχωριστικής Ανάλυσης με όλες τις υπό εξέταση ανεξάρτητες μεταβλητές, καθώς εκείνο παρουσίαζε πιο αξιόπιστες και ρεαλιστικές μετρικές αξιολόγησης σε σχέση με τον αλγόριθμο της Λογιστικής Παλινδρόμησης, παρά το μέτριο ποσοστό Ψευδώς Αρνητικών Ταξινομήσεων.

Κατά την **παλινδρόμηση**, σπουδαιότερη ανεξάρτητη μεταβλητή αναδείχθηκε η **διάρκεια οδικής διαδρομής (sec)** με ή χωρίς υπέρβαση του ορίου ταχύτητας. Σύμφωνα με τα αποτελέσματα των δύο μοντέλων που εξετάστηκαν με όλες και με τις πέντε σημαντικότερες μεταβλητές επιλέχθηκε ως καταλληλότερο το μοντέλο της Προσαρμοστικής Ενδυνάμωσης με όλες τις ανεξάρτητες μεταβλητές, διότι παρουσίαζε υψηλότερο συντελεστή R^2 και μικρότερα σφάλματα, με αποτέλεσμα να προβλέπει καλύτερα την διάρκεια χρήσης κινητού τηλεφώνου σε σχέση με την Γραμμική Παλινδρόμηση.

6 : ΣΥΜΠΕΡΑΣΜΑΤΑ

Στο παρόν κεφάλαιο πραγματοποιείται μια **γενική ανασκόπηση** του παρόντος ερευνητικού έργου, με στόχο την εξαγωγή χρήσιμων **συμπερασμάτων** μέσω των αποτελεσμάτων, τα οποία προέκυψαν από την στατιστική ανάλυση με την βοήθεια της Μηχανικής Μάθησης. Επιπροσθέτως, παρατίθενται προτάσεις για περαιτέρω αξιοποίηση των αποτελεσμάτων και στο τέλος μέσω της ανάπτυξης προτάσεων για περαιτέρω έρευνα δίνεται το έναυσμα για την δημιουργία νέων ερευνητικών έργων.

6.1: Σύνοψη Αποτελεσμάτων

Η συγκεκριμένη διπλωματική εργασία έχει στόχο να **διερευνήσει την επιρροή της χρήσης του κινητού τηλεφώνου στην συμπεριφορά του οδηγού** και πιο συγκεκριμένα στα μεγέθη της ταχύτητας, της επιτάχυνσης και του χρόνου οδήγησης μέσω της μηχανικής μάθησης ανισόρροπων δεδομένων και αλγορίθμων ταξινόμησης και παλινδρόμησης. Η βιβλιογραφική ανασκόπηση οδήγησε στην ανάγκη για περαιτέρω στατιστικές αναλύσεις οδικών δεδομένων και ταξινόμησης της οδηγικής συμπεριφοράς ανάλογα με την χρήση ή όχι κινητού τηλεφώνου. Κρίσιμος δείκτης της επικίνδυνης οδικής συμπεριφοράς είναι και η χρήση κινητού τηλεφώνου, που αποτελεί έναν από τους κυριότερους παράγοντες απόσπασης της προσοχής του οδηγού. Τα δεδομένα συλλέχθηκαν από την βάση δεδομένων της εταιρείας OSeven Telematics.

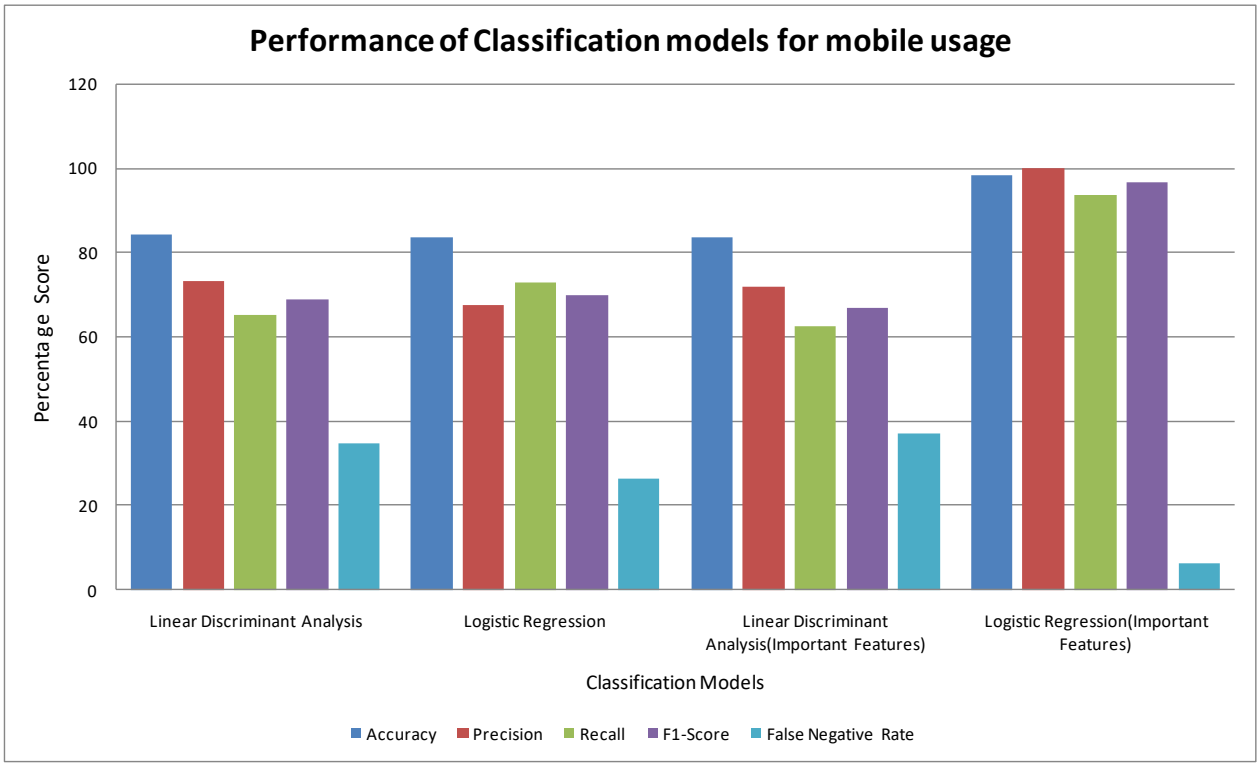
Στο στάδιο της πρώτης ανασκόπησης των δεδομένων πριν την επεξεργασία τους συντάχθηκε ο πίνακας συσχέτισης **Pearson**, ο οποίος σε συνεργασία με την διαδικασία της **Σημαντικότητας Χαρακτηριστικών** (Feature Importance), έδωσε μια σαφέστερη εικόνα σχετικά με τις σπουδαιότερες ανεξάρτητες μεταβλητές που επηρεάζουν την εξαρτημένη μεταβλητή. Η διαδικασία της Σημαντικότητας Χαρακτηριστικών εκτελέστηκε τόσο για την διαδικασία της **ταξινόμησης** όσο και για την διαδικασία της **παλινδρόμησης.**

Κατά την διαδικασία της **ταξινόμησης** έπειτα από δοκιμές και σε άλλους αλγορίθμους κρίθηκαν καταλληλότερα τα μοντέλα της Γραμμικής Διαχωριστικής Ανάλυσης και της Λογιστικής Παλινδρόμησης. Η παραπάνω διαδικασία ανέδειξε ως σημαντικότερες για την δυαδική μεταβλητή της **χρήσης κινητού τηλεφώνου** τις μεταβλητές που σχετίζονται με την **ταχύτητα** και δευτερευόντως μεταβλητές που σχετίζονται με την συνολική διανυθείσα απόσταση και τα απότομα περιστατικά. Τα δύο μοντέλα υπέστησαν ταξινόμηση τόσο για όλες τις υπό εξέταση μεταβλητές, όσο και για τις πέντε καλύτερες μεταβλητές, αφότου είχε προηγηθεί όμως εξισορρόπηση των εκ φύσεως ανισόρροπων δεδομένων με την βοήθεια της τεχνικής Συνθετικής Μειονοτικής Υπερδειγματοληψίας (SMOTE). Η ταξινόμηση αφορούσε δύο τάξεις, την τάξη της χρήσης κινητού τηλεφώνου που καθιστά την επικίνδυνη οδική συμπεριφορά και την τάξη της μη χρήσης κινητού που οδηγεί σε μη επικίνδυνη οδηγική συμπεριφορά. Τα αποτελέσματα της διαδικασίας αυτής αναπαρίστανται γραφικά στο Γράφημα 6.1.

Κατά την διαδικασία της **παλινδρόμησης** επιλέχθηκαν ύστερα από δοκιμές τα μοντέλα της Προσαρμοστικής Ενδυνάμωσης και της Γραμμικής Παλινδρόμησης. Για την εξαρτημένη συνεχή μεταβλητή της **διάρκειας χρήσης κινητού τηλεφώνου** κρίθηκε διαφορετική σημαντικότερη ανεξάρτητη μεταβλητή για καθένα από τα δύο μοντέλα. Πιο συγκεκριμένα, για την Προσαρμοστική Ενδυνάμωση την σημαντικότερη μεταβλητή αποτέλεσε η συνολική διάρκεια διαδρομής χωρίς στάσεις, ενώ για την Γραμμική Παλινδρόμηση η διάρκεια υπέρβασης του

ορίου ταχύτητας και ανοχής ανά συνολική διάρκεια διαδρομής εν κινήσει (sec/sec). Επιπροσθέτως, λοιπές σπουδαιότερες μεταβλητές που ταιριάζουν καλύτερα με τα παραπάνω μοντέλα αποδείχθηκαν για το μοντέλο της Προσαρμοστικής ενδυνάμωσης η συνολική διάρκεια οδήγησης χωρίς στάσεις, η συνολική διάρκεια υπέρβασης του ορίου ταχύτητας και ανοχής και η συνολική διανυθείσα απόσταση, ενώ για το μοντέλο της Γραμμικής Παλινδρόμησης αναδείχθηκαν η διάρκεια οδήγησης με υπέρβαση ορίου ταχύτητας και ανοχής ανά μονάδα συνολικής διάρκειας οδήγησης χωρίς στάσεις (sec/sec), μεταβλητές σχετικές με την ταχύτητα και η συνολική διανυθείσα απόσταση.

Παρακάτω απεικονίζονται συγκετρωτικά τα αποτελέσματα των μοντέλων ταξινόμησης και παλινδρόμησης.



Γράφημα 6.1 Συγκεντρωτικές τιμές μετρικών αξιολόγησης μοντέλων ταξινόμησης

Στους Πίνακες 6.1 και 6.2 παρατίθενται συγκεντρωτικά τα αποτελέσματα των αλγορίθμων ταξινόμησης και παλινδρόμης που προέκυψαν μέσω της στατιστικής ανάλυσης των στοιχείων.

Πίνακας 6.1 Συγκεντρωτικός Πίνακας μοντέλων Ταξινόμησης

ΤΑΞΙΝΟΜΗΣΗ	Αλγόριθμοι Ταξινόμησης											
	Linear Discriminant Analysis						Logistic Regression					
	Ορθότητα	Ακρίβεια	Ανάκληση	FNR	F1-Score	AUC Score	Ορθότητα	Ακρίβεια	Ανάκληση	FNR	F1-Score	AUC Score
Με όλες τις μεταβλητές	84,4%	73,3%	65,1%	34,6%	68,8%	89,5%	83,4%	67,4%	72,4%	26,4%	70,0%	89,1%
Με τις σημαντικότερες μεταβλητές	83,5%	72,0%	62,4%	37,2%	66,9%	87,7%	98,2%	99,8%	93,5%	6,4%	96,6%	99,9%

Σύμφωνα με τον Πίνακα 6.1 καταλληλότεροι αλγόριθμοι για την πρόβλεψη της οδηγικής συμπεριφοράς κρίθηκε το μοντέλο της Γραμμικής Διαχωριστικής Ανάλυσης με όλες τις υπό εξέταση μεταβλητές, το οποίο παρουσίασε υψηλό και ρεαλιστικό ποσοστό ορθότητας (ρεαλιστικές τιμές ορθότητας κυμαίνονται από 70% έως 90% σύμφωνα με την διεθνή βιβλιογραφία), καθώς και μέτριο ποσοστό Ψευδώς Αρνητικών στοιχείων.

Πίνακας 6.2 Συγκεντρωτικός Πίνακας μοντέλων Παλινδρόμησης

ΠΑΛΙΝΔΡΟΜΗΣΗ	Αλγόριθμοι Παλινδρόμησης	
	AdaBoost Regressor	Linear Regression
	R2	R2
Με όλες τις μεταβλητές	0,842	0,497
Με τις σημαντικότερες μεταβλητές	0,840	0,422

Από τον Πίνακα 6.2 γίνεται αντιληπτό ότι το μοντέλο Προσαρμοστικής Ενδυνάμωσης αποτελεί το βέλτιστο μοντέλο παλνδρόμησης, καθώς παρουσιάζει πολύ ικανοποιητική τιμή συντελεστή προσδιορισμού R^2 τόσο με όλες τις υπό εξέταση μεταβλητές όσο και με τις σημαντικότερες εξ αυτών.

6.2: Σύνοψη Συμπερασμάτων

Κατά την διάρκεια εκπόνησης της Διπλωματικής Εργασίας και της παρατήρησης των αποτελεσμάτων εξήχθησαν ορισμένα αξιοσημείωτα συμπεράσματα τόσο για την ανάλυση της οδικής συμπεριφοράς όσο και για το γενικότερο ερευνητικό πεδίο της Οδικής Ασφάλειας.

1. **Βασικότερη παράμετρος επιρροής** της χρήσης κινητού τηλεφώνου σύμφωνα με τα μοντέλα **ταξινόμησης** αποδείχθηκε **η ταχύτητα (km/h)**. Το γεγονός αυτό καθίσταται λογικό, καθώς η διάρκεια χρήσης κινητού τηλεφώνου που μετατράπηκε στην δυαδική μεταβλητή της χρήσης κινητού τηλεφώνου οδηγεί στην απόσπαση της προσοχής του οδηγού και έμμεσα επηρεάζει και την ταχύτητά του είτε αυξάνοντάς την με υπέρβαση του ορίου της είτε μειώνοντας την, επιβεβαιώνοντας την διεθνή βιβλιογραφία.

2. **Βασικότερες παράμετροι επιρροής** της διάρκειας χρήσης κινητού σύμφωνα με τα μοντέλα **παλινδρόμησης** αναδείχθηκαν τόσο **η συνολική διάρκεια οδήγησης με υπέρβαση του ορίου ταχύτητας και ανοχής ανά μονάδα διάρκειας οδικής διαδρομής χωρίς στάσεις** (sec/sec) όσο και η **διάρκεια οδικής διαδρομής χωρίς στάσεις** (sec). Η διάρκεια χρήσης κινητού τηλεφώνου αυξάνει όσο αυξάνει και μια οδική διαδρομή. Επιπλέον, υπάρχει εξάρτηση μεταξύ της διάρκειας χρήσης κινητού και της διάρκειας οδήγησης με υπέρβαση του ορίου ταχύτητας, καθώς έχει παρατηρηθεί ότι σε οδηγούς, των οποίων η προσοχή αποσπάται με το κινητό τηλέφωνο, αυξάνεται ο χρόνος αντίδρασής τους και οδηγούνται σε πιο απότομη οδηγική συμπεριφορά, άρα και σε υπερβάσεις του ορίου ταχύτητας.

3. Από το παραπάνω συμπέρασμα σε συνδυασμό με τον τριγωνικό πίνακα Pearson του κεφαλαίου 4 μπορεί να εξαχθεί έμμεσα το γεγονός ότι όπως και στην ταξινόμηση η **διάρκεια χρήσης κινητού τηλεφώνου** σχετίζεται και με την υπέρβαση των ορίων ταχύτητας, συνεπώς και με την **ταχύτητα**, η οποία αποτελεί τον κρισιμότερο παράγοντα πρόκλησης ατυχημάτων σύμφωνα με την εγχώρια και διεθνή βιβλιογραφία.

4. Στην παρούσα διπλωματική εργασία εκπαιδεύτηκαν **δύο αλγόριθμοι ταξινόμησης** και **δύο αλγόριθμοι παλινδρόμησης** τόσο με **όλες** τις υπό εξέταση μεταβλητές όσο και με τις **σπουδαιότερες** εξ' αυτών. Καλύτερος αλγόριθμος για τα μοντέλα ταξινόμησης αναδείχθηκε το μοντέλο Linear Discriminant Analysis, ενώ για την παλινδρόμηση το μοντέλο AdaBoost Regressor έδινε πιο αξιόπιστα αποτελέσματα.

5. Η **συνολική οδηγηθείσα απόσταση** επηρεάζει την διάρκεια χρήσης κινητού τηλεφώνου, όπως αποδεικνύεται από τον τριγωνικό πίνακα Pearson σε συνδυασμό με την σημαντικότητα των μεταβλητών και αυτό εξηγείται από το γεγονός ότι όσο μεγαλύτερη είναι η οδική απόσταση που διανύει ο οδηγός τόσο περισσότερη ώρα διαθέτει για να χρησιμοποιήσει το κινητό του τηλέφωνο. Επιπροσθέτως, με την απόσπαση της προσοχής του οδηγού μέσω του κινητού, μπορεί να ακολουθήσει μια μακρύτερη διαδρομή για να φτάσει στον προορισμό του.

6. Παραδόξως, παρατηρήθηκε ότι η χρήση κινητού τηλεφώνου δεν επηρεάζει σχεδόν καθόλου **απότομα περιστατικά** και πιο συγκεκριμένα απότομες επιταχύνσεις και επιβραδύνσεις

στους οδηγούς που χρησιμοποιούσαν κινητό. Το γεγονός αυτό ενδεχομένως να οφείλεται στο ότι η ενασχόληση με το κινητό προκαλεί κάποιες φορές μείωση στην ταχύτητα, όπως επιβεβαιώνεται και από την διεθνή βιβλιογραφία (Gazder & Assi, 2021).

7. Σημαντικό ποσοστό των οδηγών, περίπου το **25% εξ' αυτών** που καταγράφηκαν στην βαση δεδομένων της OSeven Telematics, **έκανε χρήση κινητού τηλεφώνου** εν ώρα οδήγησης, μη γνωρίζοντας όμως την δραστηριότητα για την οποία χρησιμοποιούσαν το κινητό τους τηλέφωνο. Αξιοσημείωτο, όμως, παραμένει το γεγονός ότι το ποσοστό είναι αρκετά μικρότερο σε σχέση με τον Ευρωπαϊκό μέσο όρο που κυμαίνεται περίπου στο 33% σύμφωνα με το κεφάλαιο 1 (Γράφημα 1.2).

8. Οι **επικίνδυνες ώρες οδήγησης**, δηλαδή το χρονικό διάστημα από 00:00 μέχρι 05:00, φαίνεται να **μην επηρεάζουν** ιδιαίτερα την **χρήση κινητού** τηλεφώνου. Παρά την γενικότερη εμφάνιση συμπτωμάτων επικίνδυνης οδικής συμπεριφοράς στο συγκεκριμένο χρονικό διάστημα, οι οδηγοί δεν φαίνεται να ασχολούνται ιδιαίτερα με το κινητό τους τηλέφωνο, επομένως πρακτικά δεν οφείλεται εκείνο για την πρόκληση σοβαρών ατυχημάτων εκείνο το χρονικό διάστημα.

9. Παρατηρήθηκε ότι οι αλγόριθμοι **παλινδρόμησης** με τις **καταλληλότερες μεταβλητές** παρουσίασαν **χαμηλότερο** συντελεστή προσδιορισμού $\mathbf{R^2}$ σε σχέση με τους ίδιους αλγόριθμους λαμβάνοντας υπόψη όλες τις υπό εξέταση ανεξάρτητες μεταβλητές. Το συμπέρασμα αυτό έγκειται στο γεγονός ότι ο συντελεστής R^2 δηλώνει την προβλεπτική ικανότητα του μοντέλου με όσο περισσότερες ανεξάρτητες μεταβλητές.

10. 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 (μη χρήσης κινητού τηλεφώνου) πολύ μεγάλη. Η SMOTE αποδείχθηκε πιο αποτελεσματική μέθοδος σε καταστάσεις μεγάλων και πολυεπίπεδων δεδομένων, επιβεβαιώνοντας την διεθνή βιβλιογραφία. 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) και του πλήθους των Ψευδώς Αρνητικών στοιχείων στις μήτρες σύγχυσης των μοντέλων.

11. Τα διαγράμματα **Σημαντικότητας Χαρακτηριστικών** (Feature Importance) **μεταβλήθηκαν** όταν εκτελέστηκε η τεχνική Υπερδειγματοληψίας SMOTE σε σχέση με τα ανισόρροπα δεδομένα. Αυτό συνέβη, καθώς η συγκεκριμένη τεχνική δημιουργεί συνθετικά δείγματα της μειονοτικής κατηγορίας με παρεμβολή μεταξύ των υπαρχόντων δειγμάτων, γεγονός που μπορεί να αλλάξει την κατανομή και τις σχέσεις μεταξύ των χαρακτηριστικών στο σύνολο δεδομένων. Όμως, η **επιλογή σπουδαιότερων μεταβλητών** έγινε με βάση το **διάγραμμα** που αντιστοιχεί στο **ανισόρροπο** σύνολο δεδομένων που αποτελεί την ρεαλιστική στατιστική σημαντικότητα των στοιχείων.

6.3: Προτάσεις για αξιοποίηση των αποτελεσμάτων

Στο παρόν υποκεφάλαιο επιχειρείται η παράθεση μιας σειράς προτάσεων για περαιτέρω ανάλυση και **αξιοποίηση των ευρημάτων** που προέκυψαν από την εκπόνηση του παρόντος ερευνητικού έργου.

1. Περαιτέρω **διερεύνηση** των κρίσιμων **παραγόντων** που επιδρούν έμμεσα στην αναγνώριση της επικίνδυνης οδηγικής συμπεριφοράς μέσω της χρήσης κινητού τηλεφώνου.

2. **Επιπλέον αξιοποίηση των μοντέλων ταξινόμησης**, τα οποία παρουσίασαν ικανή προβλεπτική ικανότητα, και διαχωρισμός της τάξης χρήσης του κινητού τηλεφώνου σε επιμέρους τάξεις ανάλογα με την διάρκεια χρήσης κινητού την ώρα της οδήγησης και συναρτήσει του χρόνου διαδρομής. Κατά αυτόν τον τρόπο, θα δημιουργηθούν επιπλέον επίπεδα ασφαλείας που θα μπορέσουν να προβλέψουν καλύτερα την οδηγική συμπεριφορά και να την κατατάξουν σε περισσότερες τάξεις επικινδυνότητας.

3. **Αναβάθμιση της υπάρχουσας εφαρμογής** και δημιουργία νέων εφαρμογών καταγραφής της χρήσης κινητού τηλεφώνου και πιο συγκεκριμένα της δραστηριότητας του οδηγού κατά την χρήση κινητού, καθώς κάθε δραστηριότητα στο κινητό τηλέφωνο ενέχει διαφορετικό συντελεστή επικινδυνότητας για πρόκληση ατυχήματος (Πίνακας 1.1).

4. **Αναγγελία οπτικοακουστικού μηνύματος** στο κινητό την ώρα χρήσης του εν ώρα οδήγησης, το οποίο θα παροτρύνει τον οδηγό να μην χρησιμοποιεί το κινητό του τηλέφωνο.

5. **Αξιοποίηση** των αποτελεσμάτων ταξινόμησης και παλινδρόμησης από τα **πανεπιστημιακά ιδρύματα** και λοιπούς φορείς και σύγκρισή τους με τα αντίστοιχα αποτελέσματα που δίνει ο προσομοιωτής οδήγησης.

6. Περαιτέρω **αξιοποίηση** των ευρημάτων από **δημοτικούς**, **περιφερειακούς** και **κρατικούς** φορείς, ώστε να εντοπιστούν σημεία στο οδικό δίκτυο, τα οποία χρήζουν βελτίωσης προς όφελος της Οδικής Ασφάλειας.

6.4: Προτάσεις για περαιτέρω έρευνα

Η **Οδική Ασφάλεια** οφείλει να αποτελεί στην σύγχρονη εποχή στόχος και προτεραιότητα όλων των κοινωνιών για την αποφυγή σοβαρών ατυχημάτων. Στην παρούσα διπλωματική εργασία χρησιμοποιήθηκαν τεχνικές μηχανικής μάθησης, που αποτελούν την πιο σύγχρονη τάση του κλάδου της μηχανικής, με συλλογή, επεξεργασία και ανάλυση οδικών δεδομένων για την εξαγωγή συμπερασμάτων σχετικά με την οδηγική συμπεριφορά. Συνεπώς, εκτελέστηκε ανάλυση σε πολυεπίπεδα δεδομένα, η οποία αποσκοπούσε στην εξέλιξη του γνωσιακού υποβάθρου του τομέα και δίνει περαιτέρω περιθώρια εξέλιξης. Στην συγκεκριμένη υποενότητα διατυπώνονται εύστοχες **προτάσεις**, οι οποίες μπορούν να αποτελέσουν τον θεμέλιο λίθο για

περαιτέρω έρευνα, με απώτερο σκοπό τόσο την βελτίωση τόσο της συμπεριφοράς των οδηγών όσο και την ενίσχυση της Οδικής Ασφάλειας γενικότερα.

1. **Διαχωρισμός σε περισσότερες τάξεις ταξινόμησης**. Σύμφωνα με την διεθνή βιβλιογραφία, βέλτιστος θεωρείται ο διαχωρισμός σε τέσσερις τάξεις ταξινόμησης. Συνεπώς, προτείνεται για τον εντοπισμό των επιπέδων επικινδυνότητας ο διαχωρισμός της χρήσης κινητού τηλεφώνου (Class 1) σε επιμέρους τάξεις τόσο ανάλογα με την χρονική διάρκεια χρήσης του κινητού όσο και ανάλογα με την δραστηριότητα για την οποία ο οδηγός χρησιμοποιεί την κινητή του συσκευή.

2. Η **αύξηση του όγκου δεδομένων** στην υφιστάμενη βάση δεδομένων. Προτείνεται να εισαχθούν δεδομένα που αφορούν σε δημογραφικά στοιχεία, όπως το φύλο, η ηλικία, το μορφωτικό επίπεδο και η οδηγική εμπειρία του οδηγού, στοιχεία της οδού και της κυκλοφορίας, όπως η οριζόντια και κατακόρυφη σήμανση, η σηματοδότηση και οι κυκλοφοριακοί φόρτοι, καθώς και στοιχεία που αφορούν στις επιδόσεις των οδηγών με την συγκατάθεσή τους, τα οποία αφαιρέθηκαν από την παρούσα ερευνητική εργασία με σκοπό την προστασία των προσωπικών δεδομένων. Με αυτόν τον τρόπο θα υπάρξει πρόσθετη ακρίβεια και εγκυρότητα στα αποτελέσματα των μοντέλων ταξινόμησης και παλινδρόμησης.

3. **Εισαγωγή δεδομένων** που αφορούν σε **οδικά ατυχήματα** και διερεύνηση της επιρροής της χρήσης κινητού τηλεφώνου για την πρόκληση ατυχημάτων.

4. **Ανάπτυξη περισσότερων μοντέλων ταξινόμησης και παλινδρόμησης** που παρουσιάζουν υψηλές επιδόσεις σύμφωνα με τον συγκριτικό πίνακα των μοντέλων**,** όπως το μοντέλο Κ Πλησιέστερων Γειτόνων, και σύγκριση των αποτελεσμάτων τους με τα αποτελέσματα των επιλεγμένων αλγορίθμων του παρόντος ερευνητικού έργου.

5. **Ανάπτυξη κατάλληλων μοντέλων βαθιάς εκμάθησης (deep learning)**, τα οποία έχουν σχεδιαστεί για τη μοντελοποίηση πολύπλοκων προτύπων σε μεγάλα σύνολα δεδομένων με τη χρήση τεχνητών νευρωνικών δικτύων διαμορφωμένα με βάση τον ανθρώπινο εγκέφαλο. Στην παρούσα φάση από τους κύριους αλγορίθμους βαθιάς εκμάθησης προτείνεται η ανάπτυξη Επαναλαμβανομένων Νευρωνικών Δικτύων (Recurrent Neural Networks-RNNs) και τύπων τους, όπως τα Δίκτυα μακράς βραχυπρόθεσμης μνήμης (Long Short-Term Memory Networks-LSTM) που διαθέτουν βρόχους ανατροφοδότησης, γεγονός που τους επιτρέπει να θυμούνται προηγούμενες εισόδους και να χρησιμοποιούν αυτές τις πληροφορίες για να κάνουν προβλέψεις. Η βαθιά εκμάθηση αφαιρεί την χειροκίνητη αναγνώριση χαρακτηριστικών των δεδομένων, επιταχύνοντας την συνολική διαδικασία στατιστικής ανάλυσης.

ΒΙΒΛΙΟΓΡΑΦΙΑ

1. Caird, J., Johnston, A., Willness, C., Asbridge, M. & Steel, P. (2014). A meta-analysis of the effects of texting on driving. Available at: https://doi.org/10.1016/j.aap.2014.06.005 (Accessed January 4, 2023).

2. Cameron C. & Windmeijer F. (1996). An *R*-squared measure of goodness of fit for some common nonlinear regression models. Available at: https://doi.org/10.1016/S0304-4076(96)01818-0 (Accessed February 23, 2023).

3. Chicco,D. ,Warrens, M. J. & Jurman, G. (2021). The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation. Available at: https://peerj.com/articles/cs-623/ (Accessed February 23, 2023).

4. Choudhary, P., Gupta, A. & Velaga, N.R. (2022) Perceived risk vs actual driving performance during distracted driving: A comparative analysis of phone use and other secondary distractions, Transportation Research Part F: Traffic Psychology and Behaviour. Available at: https://doi.org/10.1016/j.trf.2022.03.001 (Accessed: March 1, 2023).

5. Datagy: Learn Python Programming[WWW Document] (2022).Available at: https://datagy.io/

6. European Road Safety Observatory (2022) Road Safety Thematic Report – Driver distraction. European Commission. Available at: https://road-safety.transport.ec.europa.eu/system/files/2022-04/Road_Safety_Thematic_Report_Driver_Distraction_2022.pdf (Accessed: February 19, 2023).

7. Gazder U. & Assi J. K. (2022),Determining driver perceptions about distractions and modeling their effects on driving behavior at different age groups, Available at: https://doi.org/10.1016/j.jtte.2020.12.005

8. GeeksforGreeks[WWW Document] (2022). Available at: https://www.geeksforgeeks.org/

9. Ghandour, R., Potams, A.J., Boulkaibet, I., Neji, B. &l Barakeh, Z. (2021). Driver Behavior Classification System Analysis Using Machine Learning Methods. Available at: https://doi.org/10.3390/app112210562 (Accessed December 11, 2022).

10. Github: Let's build from here [WWW Document] (2022). Available at: https://github.com/ (Accessed November 19,2022).

11. 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. Available at: https://doi.org/10.1016/j.trip.2020.100186

12. 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, Available at: https://www.nrso.ntua.gr/geyannis/wp-content/uploads/geyannis-pc327.pdf (Accessed January 22, 2023)

13. Katrakazas C., Eva Michelaraki E., Marios Sekadakis M., Apostolos Ziakopoulos A., Armira Kontaxi A. & Yannis G. (2021). Identifying the impact of the COVID-19 pandemic on driving behavior using naturalistic driving data and time series forecasting. Available at:https://doi.org/10.1016/j.jsr.2021.04.007 (Accessed March 3,2022).

14. Khan, I., Rizvi, S. S., Khusro, S., Ali, S., & Chung, T. S. (2021). Analyzing drivers' distractions due to smartphone usage: evidence from AutoLog dataset. Mobile Information Systems.Available at: https://www.hindawi.com/journals/misy/2021/5802658/ (Accessed February 16, 2023)

15. Kontaxi ,A., Ziakopoulos, A., Katrakazas, C. & Yannis, G. (2022). Measuring the impact of driver behavior telematics in road safety. Available at: https://fersi.org/wp-content/uploads/2022/10/Armira-Kontaxi-et-al.pdf (Accessed November 19, 2022)

16. Kulkarni, A., Chong, D. & Batarseh, F. (2020). 5-Foundations of data imbalance and solutions for a data democracy. Available at: https://doi.org/10.1016/B978-0-12-818366-3.00005-8 (Accessed November 19, 2022)

17. Michelaraki, E., Sekadakis, M., Katrakazas, C., Ziakopoulos, A. & Yannis, G. (2023). One year of COVID-19: Impacts on safe driving behavior and policy recommendations. Available at: https://doi.org/10.1016/j.jsr.2022.10.007 (Accessed November 19, 2022)

18. OSeven Telematics [WWW Document] (2022). Availabe at: https://oseven.io/

19. Papadakaki, M. ,Tzamalouka G. & Gnardellis C. (2016) Driving performance while using mobile phone: A simulation study of greek professional drivers, Transportation Research Part F: Traffic Psychology and Behaviour. Available at: https://doi.org/10.1016/j.trf.2016.02.006 (Accessed: March 3, 2023).

20. Phuksuksakul, N., Kanitpong K. & Chantranuwathana, S. (2021),Factors affecting behavior of mobile phone use while driving and effect of mobile phone use on driving performance, Available at: https://doi.org/10.1016/j.aap.2020.105945 (Accessed: March 3, 2023).

21. Pires, C., Areal, A., & Trigoso, J. (2019). Distraction (mobile phone use). ESRA2 Thematic report Nr. 3. ESRA project (E-Survey of Road users' Attitudes) (Issue 3). Lisbon, Portugal: Portuguese Road Safety Association. Available at: https://www.researchgate.net/publication/341089527

22. PyCaret[WWW Document] (2022). Available at: https://pycaret.org/

23. Rekkala J.R. (2021), Mobile usage detection of driver Using CNN (Convolutional Neural Network), Available at: https://dspace.calstate.edu/bitstream/handle/10211.3/222082/Rekkala-Jaswanth%20Reddy-thesis-2022.pdf?sequence=1

24. Scikit-learn: Machine Learning in Python[WWW Document] (2022). Available at: https://scikit-learn.org/

25. Seaborn: statistical data visualization [WWW Document] (2022). Available at: https://seaborn.pydata.org/ (Accessed December 14,2022).

26. Stevens, M., Sunseri I. & Alexanderian, A. (2022). Hyper-differential sensitivity analysis for inverse problems governed by ODEs with application to COVID-19 modeling. Available at: https://doi.org/10.1016/j.mbs.2022.108887 (Accessed January 13, 2023)

27. Tran, L. D. (2017). Data Fusion with 9 degrees of freedom Inertial Measurement Unit to determine object's orientation. Available at: https://digitalcommons.calpoly.edu/eesp/400 (Accessed February 23, 2023)

28. Tselentis, D. & Vlahogianni, E.& Yannis, G. & Kavouras, L. (2020). Hybrid Data Envelopment Analysis for Large-Scale Smartphone Data Modeling. Transportation Research Procedia . Available at: https://doi.org/10.1016/j.trpro.2020.08.126 (Accessed February 23, 2023)

29. Viloria, A., Lezama, O. & Mercado-Caruzo, N. (2020). Unbalanced data processing using oversampling: Machine Learning. Available at: https://doi.org/10.1016/j.procs.2020.07.018 (Accessed February 28, 2023)

30. Wang, R., Huang, L. & Wang, C. (2021). Distracted driving detection by sensing the hand gripping of the phone. In: Proceedings of the 27th Annual International Conference on Mobile Computing and Networking. Available at: https://www.researchgate.net/publication/355589102 (Accessed February 19, 2023)

31. World Health Organization, 2021. Global Plan: Decade of Action for Road Safety 2021- 2030 [WWW Document]. World Health Organization. Available at: https://www.who.int/news-room/fact-sheets/detail/road-traffic-injuries (Accessed December 12, 2022).

32. WOO, H., Ph.D., & LIN, J., Ph.D. (2014), INFLUENCE OF MOBILE PHONE USE WHILE DRIVING: The Experience in Taiwan, Available at: https://doi.org/10.1016/S0386-1112 (14)60066-2 (Accessed: February 23, 2023).

33. Wu, M., Zhang, S. & Dong, Y. (2016). A Novel Model-Based Driving Behavior Recognition System Using Motion Sensors. Sensors 16. Available at: https://doi.org/10.3390/s16101746 (Accessed March 3, 2023)

34. Yannis G., Laiou A., Papantoniou P., & Christoforou C. (2014). Impact of texting on young drivers' behavior and safety on urban and rural roads through a simulation experiment. Available at: https://doi.org/10.1016/j.jsr.2014.02.008 (Accessed: March 3, 2023).

35. Zhang, Y., Chen, Y. &Gao, C. (2020). Deep unsupervised multi-modal fusion network for detecting driver distraction, Available at: https://doi.org/10.1016/j.neucom.2020.09.023 (Accessed February 20, 2023)

36. Ziakopoulos, A., Kontaxi, A. & Yannis, G. (2023). Analysis of mobile phone use engagement during naturalistic driving through explainable imbalanced machine learning, Available at: https://doi.org/10.1016/j.aap.2022.106936 (Accessed February 20, 2023)

37. Ziakopoulos, A., Vlahogianni E., Antoniou, C. & Yannis, G. (2022). Spatial predictions of harsh driving events using statistical and machine learning methods. Available at: https://doi.org/10.1016/j.ssci.2022.105722 (Accessed February 22, 2023)

38. Ziakopoulos, A., Theofilatos, A., Papadimitriou, E., & Yannis, G. (2017). Cell Phone Use – Texting. European Road Safety Decision Support System, developed by the H2020 project SafetyCube. Available at: https://www.roadsafety-dss.eu/#/ (Accesed October 4, 2022).

39. Ziakopoulos, A., Theofilatos, A., Papadimitriou, E., & Yannis, G. (2018). Distraction - Cell Phones - Hands Free, European Road Safety Decision Support System, developed by the H2020 project SafetyCube. Available at: https://www.roadsafety-dss.eu/#/ (Accesed October 4, 2022).

40. Γιαννής, Γ. (2018).Τροχαία Ατυχήματα Οδική Συμπεριφορά και Ασφάλεια. Διαθέσιμο σε: https://www.nrso.ntua.gr/geyannis/wp-content/uploads/geyannis-cp329.pdf (Accessed November 11, 2022).

41. Γναρδέλλης Χ. (2018).Βιοστατιστική:Εισαγωγικές έννοιες στη Στατιστική. Διαθέσιμο σε: https://eclass.upatras.gr/ (Accessed November 19, 2022).

42. Κατάκης, Ι. (2009): ΔΙΔΑΚΤΟΡΙΚΗ ΔΙΑΤΡΙΒΗ: Μέθοδοι Μηχανικής Μάθησης για Αυτόματη Ταξινόμηση Κειμένων. Διαθέσιμο σε: http://ikee.lib.auth.gr/record/113419/files/Katakis_Phd_Thesis.pdf (Accessed November 30, 2022)

43. Φραντζεσκάκης, Ι. & Γκόλιας, Ι. Οδική Ασφάλεια, Αθήνα (1994)

44. Φραντζεσκάκης, Ι., Γκόλιας, Ι. & Πιτσιάβα-Λατινοπούλου. Κυκλοφοριακή Τεχνική, Αθήνα (2009).

ΠΑΡΑΡΤΗΜΑΤΑ

Παρακάτω παρατίθεται ο κώδικας που χρησιμοποιήθηκε σε γλώσσα προγραμματισμού Python μέσω του πακέτου PyCaret σε προγραμματιστικό περιβάλλον Jupyter Notebook.

#Install Pycaret

!pip install pycaret

 # Importing necessary libraries

import pandas as pd

import numpy as np

import warnings

import seaborn as sns

import matplotlib.pyplot as plt

warnings.filterwarnings ("ignore")

This command will basically import all the modules from pycaret that are necessary for classification tasks

df=pd.read_csv ('./gr_trip_data_o7_5.csv')

df.head ()

#dropping the columns wich we dont need

df=df.drop
(["M1_Wildcard","StringencyIndexForDisplay","StringencyLegacyIndex","StringencyLegacyIndexForDisplay","GovernmentResponseIndex","GovernmentResponseIndexForDisplay","ContainmentHealthIndex","ContainmentHealthIndexForDisplay","EconomicSupportIndex","EconomicSupportIndexForDisplay","H1_Public.information.campaigns","H1_Flag","H2_Testing.policy","H3_Contact.tracing","H4_Emergency.investment.in.healthcare","H5_Investment.in.vaccines","H6_Facial.Coverings","H6_Flag","H7_Vaccination.policy","H7_Flag","C6_Stay.at.home.requirements","C6_Flag","C7_Restrictions.on.internal.movement","C7_Flag","C8_International.travel.controls","E1_Income.support","E1_Flag","E2_Debt.contract.relief","E3_Fiscal.measures","E4_International.support","C1_School.closing","C1_Flag","C2_Workplace.closing","C2_Flag","C3_Cancel.public.events","C3_Flag","C4_Restrictions.on.gatherings","C4_Flag","C5_Close.public.transport","C5_Flag","Stringency
Categorical","GRdriving","GRwalking","GRTotalCases","GRTotalDeaths","GRNewCases","GRNewDeaths","TotalCasesPerMillion","TotalDeathsPerMillion","ReproductionRate","av_speeding_km h_no_changer","StringencyIndex","stars","start_country_code","Unnamed:
0","Date","time_mobile_usage/driving duration","ha","hb"],axis=1)

df.head ()

df=df.drop
(["total_score","speeding_score","mu_score","hb_score","ha_score","driver_id"],axis=1)

df.head ()

#Προεπεξεργασία Δεδομένων (Δημιουργία τριγωνικού θερμικού χάρτη Pearson)

plt.figure (figsize= (16, 6))

```
mask = np.triu (np.ones_like (df.corr (), dtype=np.bool))
heatmap = sns.heatmap (df.corr (), mask=mask, vmin=-1, vmax=1, annot=True, cmap='BrBG')
heatmap.set_title ('Triangle Correlation Heatmap', fontdict={'fontsize':18}, pad=16);
```

#Classification

```
from pycaret.classification import *
```

#Μετατροπή από συνεχή σε δυαδική μεταβλητή (Binary)

```
df.loc[df['time_mobile_usage'] ==0 , 'time_mobile_usage'] = 0
df.loc[df['time_mobile_usage'] > 0, 'time_mobile_usage'] = 1
df
```

#train_test_split

```
data = df.sample (frac=0.8, random_state=786)
data_unseen = df.drop (data.index)
data.reset_index (inplace=True, drop=True)
data_unseen.reset_index (inplace=True, drop=True)
print ('Data for Modeling: ' + str (data.shape))
print ('Unseen Data For Predictions: ' + str (data_unseen.shape))
# Setting up the classifier
# Pass the complete dataset as data and the featured to be predicted as target
clf=setup (data=df,target='time_mobile_usage',transformation=False , fix_imbalance=True
,train_size=0.8 )
# This model will be used to compare all the model along with the cross validation
compare_models ()
```

#Υπερδειγματοληψία-Oversampling (για Classification)

```
from imblearn.over_sampling import SMOTE
sm = SMOTE (random_state=42)
X_resampled, y_resampled = sm.fit_resample (X, y)
```

#Classification and Regression models (Ενδεικτικός κώδικας για το μοντέλο Linear Discriminant Analysis-όμοια διαδικασία και για τα υπόλοιπα)

```
lda=create_model ('lda')
tuned_lda=tune_model (lda)
predict_model (tuned_lda);
final_lda = finalize_model (tuned_lda)
predict_model (final_lda);
unseen_predictions = predict_model (final_lda, data=data_unseen)
```

```
unseen_predictions.head ()
#Πίνακας επίδοσης και μήτρα σύγχυσης
plot_model (final_lda,plot='class_report')
plot_model (final_lda,plot='confusion_matrix')
#Σημαντικότητα μεταβλητών, γραφήματα ROC curve, precision-recall και calibration
plot_model (final_lda, plot="feature",)
plot_model (final_lda, plot="auc","pr","calibration")
#Regression Error (Για τον αλγόριθμο AdaBoost)
plot_model (final_ada, plot = 'error')
```