ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΠΟΛΙΤΙΚΩΝ ΜΗΧΑΝΙΚΩΝ
ΤΟΜΕΑΣ ΜΕΤΑΦΟΡΩΝ ΚΑΙ ΣΥΓΚΟΙΝΩΝΙΑΚΗΣ ΥΠΟΔΟΜΗΣ

ΕΝΤΟΠΙΣΜΟΣ ΣΥΜΒΑΝΤΩΝ ΜΕ ΒΑΣΗ ΤΑ ΟΔΗΓΙΚΑ ΧΑΡΑΚΤΗΡΙΣΤΙΚΑ ΣΕ ΥΠΕΡΑΣΤΙΚΕΣ ΟΔΟΥΣ

Διπλωματική Εργασία

Ακριβή Βαρελά

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

Αθήνα, Ιούλιος 2020

ΕΥΧΑΡΙΣΤΙΕΣ

Η παρούσα διπλωματική εργασία σηματοδοτεί το πέρας των προπτυχιακών μου σπουδών στη Σχολή Πολιτικών Μηχανικών του Εθνικού Μετσόβιου Πολυτεχνείου.

Θα ήθελα πρωτίστως να ευχαριστήσω τον κύριο Γεώργιο Γιαννή, Καθηγητή του Εθνικού Μετσόβιου Πολυτεχνείου, για την ανάθεση του θέματος της διπλωματικής αυτής εργασίας, καθώς και για τις συμβουλές του σε καίρια ζητήματα.

Επίσης θα ήθελα να ευχαριστήσω θερμά το Δρ. Χρήστο Κατρακάζα για την πολύτιμη βοήθεια που μου προσέφερε και για το εξαιρετικό κλίμα συνεργασίας, συμβάλλοντας ουσιαστικά στην υλοποίηση της παρούσας διπλωματικής εργασίας.

Τέλος ευχαριστώ πολύ τον αδερφό μου, την οικογένειά μου και τους φίλους μου για την υποστήριξη που μου πρόσφεραν καθ' όλη τη διάρκεια των σπουδών μου.

Αθήνα, Ιούλιος 2020
Ακριβή Βαρελά

ΕΝΤΟΠΙΣΜΟΣ ΣΥΜΒΑΝΤΩΝ ΜΕ ΒΑΣΗ ΤΑ ΟΔΗΓΙΚΑ ΧΑΡΑΚΤΗΡΙΣΤΙΚΑ ΣΕ ΥΠΕΡΑΣΤΙΚΕΣ ΟΔΟΥΣ

Ακριβή Βαρελά

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

Σύνοψη

Στόχος της παρούσας διπλωματικής εργασίας είναι ο εντοπισμός των συμβάντων με βάση τα οδηγικά χαρακτηριστικά σε υπεραστικές οδούς, καθώς και ο προσδιορισμός των κυριότερων παραγόντων που αφορούν στην κατάσταση πριν και κατά τη διάρκεια ενός συμβάντος. Η συλλογή των απαιτούμενων στοιχείων έγινε μέσω πειράματος σε προσομοιωτή οδήγησης σε υπεραστικές οδούς. Για την ανάλυση των δεδομένων αναπτύχθηκαν μοντέλα διωνυμικής λογιστικής παλινδρόμησης και τυχαίων δασών, καθώς και ανάλυσης παραγόντων, με εξαρτημένη μεταβλητή την ύπαρξη ή όχι συμβάντος. Στη συνέχεια, η ανάλυση παραγόντων πραγματοποιήθηκε για την ομαδοποίηση των ανεξάρτητων μεταβλητών για την κατάσταση που περιγράφει ένα λεπτό πριν από κάθε συμβάν, τη διάρκεια των συμβάντων και το συνδυασμό των δύο προηγούμενων καταστάσεων. Τα αποτελέσματα έδειξαν πως το μοντέλο τυχαίων δασών αποδίδει πολύ καλύτερα από εκείνο της διωνυμικής λογιστικής παλινδρόμησης στον έγκαιρο εντοπισμό της ύπαρξης ή όχι συμβάντος με πολύ λίγα σφάλματα. Παράλληλα, η ταχύτητα και η διαμήκης επιτάχυνση μαζί με τη συνολική διανυόμενη απόσταση, προέκυψε ότι μπορούν να περιγράψουν επαρκώς τα δεδομένα που αφορούν στη χρονική διάρκεια του ένος λεπτού πριν το συμβάν. Τέλος, προέκυψε ότι η οδήγηση κατά τη διάρκεια συμβάντος μπορεί να περιγραφεί καλύτερα μέσω της ταχύτητας, της απόκλισης του οχήματος από το μέσο της οδού και το χρόνο μέχρι τη σύγκρουση από το προπορευόμενο όχημα.

Λέξεις κλειδιά: εντοπισμός συμβάντων, διάρκεια συμβάντος, διωνυμική λογιστική παλινδρόμηση, μοντέλο τυχαίων δασών, μοντέλο ταξινόμησης, ανάλυση παραγόντων

EVENT IDENTIFCATION BASED ON DRIVING CHARACTERISTICS ON RURAL ROADS

Akrivi Varela

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

Abstract

The aim of this Diploma Thesis is to identify events based on driving characteristics on rural roads and determine the main factors that can describe the situation before and during an event. The data used were collected from a driving simulator experiment in rural roads. Fort the data analysis, binomial logistic regression, random forests, as well as factor analysis statistical models were developed, with dependent variable the occurrence of an event. Furthermore, factor analysis aimed at identifying groups of independent variables for the data describing the situation one minute before each event, the duration of the events and the combination of the two aforementioned situations. Results showed that random forest model performs much better than the binomial logistic regression in identifying event occurrence with very few false alarms. Moreover, speed and longitudinal acceleration along with total distance driven from the beginning of the driving session, turned out to better describe the case of driving one minute prior to an event. Finally, driving during an event can be sufficiently described through speed, the deviation of the vehicle from the middle of the road as well as the time headway with the vehicle ahead.

Keywords: event detection, event duration, binary logistic regression, random forest model, classification, factor analysis

ΠΕΡΙΛΗΨΗ

Η εκπόνηση της παρούσας διπλωματικής εργασίας έχει **στόχο τη διερεύνηση της ικανότητας εντοπισμού ύπαρξης συμβάντων με βάση τα οδηγικά χαρακτηριστικά σε υπεραστικές οδούς.** Για το σκοπό αυτό χρησιμοποιήθηκαν στοιχεία από πείραμα σε προσομοιωτή οδήγησης σε υπεραστικές οδούς.

Τα στοιχεία αυτά επεξεργάστηκαν κατάλληλα προκειμένου να ετοιμαστούν οι συγκεντρωτικοί πίνακες δεδομένων που περιλάμβαναν στοιχεία που αφορούσαν στη χρονική διάρκεια του ενός λεπτού πριν και κατά τη διάρκεια κάθε συμβάντος. Οι πίνακες αυτοί, που χρησιμοποιήθηκαν στη στατιστική ανάλυση, αποτελούνταν από την εξαρτημένη μεταβλητή που εκφράζει την ύπαρξη ή όχι συμβάντος και από ένα σύνολο από ανεξάρτητες, μη συσχετισμένες μεταξύ τους μεταβλητές, όπως η ταχύτητα του οχήματος, η διαμήκης επιτάχυνση, ο χρόνος μέχρι τη σύγκρουση από το προπορευόμενο όχημα, η συνολική διανυόμενη απόσταση, η απόκλιση του οχήματος από τη μέση του δρόμου, η γωνία του τιμονιού και η οδηγική εμπειρία.

Για τη στατιστική ανάλυση των δεδομένων χρησιμοποιήθηκαν τα μοντέλα διωνυμικής λογιστικής παλινδρόμησης και τυχαίων δασών, για τα οποία πραγματοποιήθηκαν δύο παραλλαγές, Α και Β, ανάλογα με το πλήθος των ανεξάρτητων μεταβλητών που χρησιμοποιήθηκαν στην καθεμία. Η παραλλαγή Α περιελάμβανε όλες τις μεταβλητές που έπειτα από δοκιμές βρέθηκαν στατιστικά σημαντικές, ενώ για την επιλογή των μεταβλητών της παραλλαγής Β, έγινε προσδιορισμός της σημαντικότητάς τους και ανάλογα με αυτή και την αναγκαιότητα εξαγωγής χρήσιμων αποτελεσμάτων επιλέχθηκαν οι 4 πιο σημαντικές ανεξάρτητες μεταβλητές. Συγκεκριμένα κάθε παραλλαγή περιελάμβανε τις εξής μεταβλητές:

- **Παραλλαγή Α**: την εξαρτημένη διακριτή μεταβλητή **Event** και τις ανεξάρτητες **Speed, AccLon, Thead, rdist, rspur και Driving Experience.**
- **Παραλλαγή Β**: την εξαρτημένη διακριτή μεταβλητή **Event** και τις ανεξάρτητες **Speed, AccLon, rdist και Driving Experience.**

Τα αποτελέσματα της παραλλαγής Β για το μοντέλο της διωνυμικής λογιστικής παλινδρόμησης δεν ήταν ικανοποιητικά, οπότε αξιοποιήθηκαν μόνο για το μοντέλο τυχαίων δασών.

Επίσης πραγματοποιήθηκε η μέθοδος ανάλυσης παραγόντων για κάθε μία από τις περιπτώσεις, ένα λεπτό πριν από κάθε συμβάν, κατά τη διάρκεια καθώς και για το άθροισμα αυτών, για τη διερεύνηση της ικανότητας παραγόντων να εκφράσουν μια ομάδα μεταβλητών.

Πίνακας 1: Μοντέλο διωνυμικής λογιστικής παλινδρόμησης

	Παραλλαγή Α
Μέτρηση	**Τιμή**
Ορθότητα **(Accuracy)**	83.80%
Στατιστικός συντελεστής Κάππα **(Kappa Statistic)**	31.60%

Ευαισθησία/Ανάκληση **(Sensitivity/Recall)**	27.90%
Εξειδικευτικότητα **(Specificity)**	96.60%
Ακρίβεια **(Precision)**	65.70%
Μέτρο F **(F-measure)**	39.20%
Δείκτης λάθος συναγερμού **(False alarm rate)**	3.30%
Εμβαδόν κάτω από την καμπύλη ROC **(AUC)**	80%

Πίνακας 2: Μοντέλο τυχαίων δασών

	Παραλλαγή Α	**Παραλλαγή Β**
Μέτρηση	**Τιμή**	**Τιμή**
Ορθότητα **(Accuracy)**	99.80%	99.20%
Στατιστικός συντελεστής Κάππα **(Kappa Statistic)**	99.40%	97.30%
Ευαισθησία/Ανάκληση **(Sensitivity/Recall)**	99.50%	96.60%
Εξειδικευτικότητα **(Specificity)**	99.90%	99.80%
Ακρίβεια **(Precision)**	99.60%	98.90%
Μέτρο F **(F-measure)**	99.50%	97.80%
Δείκτης λάθος συναγερμού **(False alarm rate)**	0.09%	0.23%
Εμβαδόν κάτω από την καμπύλη ROC **(AUC)**	99.99%	99.94%

Πίνακας 3: Ανάλυση παραγόντων

Πίνακας	**Παράγοντας 1**	**Παράγοντας 2**	**Παράγοντας 3**
1΄πριν κάθε συμβάν	Speed	Rdist	
	AccLon		
διάρκεια συμβάντος	Speed	Rspur	THead
1΄πριν/κατά τη διάρκεια κάθε συμβάντος	Speed	rspur	rdist
	THead		

Τα αποτελέσματα που προέκυψαν από τη στατιστική ανάλυση οδήγησαν στη **διατύπωση των παρακάτω συμπερασμάτων** για την παρούσα διπλωματική εργασία.

- Οι μεταβλητές που περιγράφουν την ταχύτητα του οχήματος, τη διαμήκη επιτάχυνση, τη συνολική διανυόμενη απόσταση καθώς και την εμπειρία του οδηγού στην οδήγηση, προέκυψε πως αποτελούν τις **μεταβλητές με τη μεγαλύτερη σημαντικότητα** για τον εντοπισμό της ύπαρξης ενός συμβάντος. Το γεγονός αυτό επιβεβαιώνεται και από τη διεθνή βιβλιογραφία.

- Βασικό συμπέρασμα είναι ότι το **μοντέλο τυχαίων δασών ήταν αρκετά πιο αποδοτικό από τη διωνυμική λογιστική παλινδρόμηση** στη στατιστική ανάλυση που πραγματοποιήθηκε. Αυτό είναι πιθανό να οφείλεται στο γεγονός ότι το μοντέλο τυχαίων δασών περιγράφει καλύτερα την ύπαρξη συμβάντος, καθώς ζυγίζει ορισμένα χαρακτηριστικά ως πιο σημαντικά από άλλα, δεν υποθέτει ότι το μοντέλο έχει γραμμική σχέση όπως τα μοντέλα παλινδρόμησης, και επεξεργάζεται τυχαία δείγματα έτσι ώστε να καταλήξει στο βέλτιστο μοντέλο.

- Η εφαρμογή του **μοντέλου διωνυμικής λογιστικής παλινδρόμησης** παρατηρήθηκε πως δεν λειτουργεί αποτελεσματικά στον εντοπισμό ενός συμβάντος με μικρό αριθμό ανεξάρτητων μεταβλητών (μικρή ανάκληση, σχετικά υψηλός δείκτης ποσοστού πιθανότητας λάθους ταξινόμησης, χαμηλές τιμές ακρίβειας και μέτρο αξιολόγησης). Στην περίπτωση που έχουν ληφθεί υπόψη όλες οι στατιστικά σημαντικές μεταβλητές (παραλλαγή Α), η **ορθότητα** του μοντέλου διωνυμικής λογιστικής παλινδρόμησης, δηλαδή η ακρίβεια που προσφέρει για τις σωστές προβλέψεις στο σύνολό τους, τη σωστή πρόβλεψη για την ύπαρξη ή όχι συμβάντος, είναι ικανοποιητική.

- Τα αποτελέσματα που εξήχθηκαν από το **μοντέλο τυχαίων δασών** και για τις δύο παραλλαγές που αναφέρθηκαν παραπάνω ήταν στο σύνολό τους πολύ καλύτερα από το μοντέλο διωνυμικής λογιστικής παλινδρόμησης. Η **ορθότητα** στην πρόβλεψη ύπαρξης συμβάντος ή όχι ήταν πολύ υψηλή και για τις 2 παραλλαγές και η **ανάκληση** του μοντέλου (αυξημένη ικανότητα εύρεσης των στιγμιότυπων) βρέθηκε πολύ υψηλή. Ο δείκτης της **εξειδικευτικότητας**, προέκυψε επίσης πολύ υψηλός, όπως άλλωστε και η **ακρίβεια** (βαθμός πιστότητας) και το **μέτρο αξιολόγησης**, που καθιστούν το μοντέλο πολύ αξιόπιστο για τον εντοπισμό συμβάντων με βάση τα οδηγικά χαρακτηριστικά.

- **Συγκρίνοντας τις δύο παραλλαγές** που πραγματοποιήθηκαν με το μοντέλο τυχαίων δασών για τη στατιστική ανάλυση, προέκυψε πως παρόλο που και οι δύο παραλλαγές έξαγουν χρήσιμα και αξιόπιστα αποτελέσματα και είναι αποδεκτές, εκείνη που περιείχε το μεγαλύτερο πλήθος μεταβλητών έδινε καλύτερους δείκτες.

- Τα αποτελέσματα της **παραγοντικής ανάλυσης** έδειξαν πως για κάθε σύνολο στοιχείων, i) εκείνων που περιγράφουν τη χρονική διάρκεια του ενός λεπτού πριν από κάθε συμβάν, ii) εκείνων που περιγράφουν τη χρονική διάρκεια του κάθε συμβάντος, και iii) το σύνολο αυτών, υπάρχει ένα πλήθος παραγόντων που εκφράζουν ένα σύνολο μεταβλητών, το οποίο σε κάθε περίπτωση είναι διαφορετικό.

- Τα δεδομένα που περιγράφουν την κατάσταση **ένα λεπτό πριν από κάθε συμβάν** βρέθηκε πως μπορούν να εκφραστούν με δύο παράγοντες, έναν που περιγράφει την επιρροή της ταχύτητας και της διαμήκους επιτάχυνσης και έναν που περιγράφει την επιρροή της συνολικής διανυόμενης απόστασης του

οχήματος. Αυτοί οι δύο παράγοντες πιθανόν να προκύπτουν καθώς, όπως έχει παρατηρηθεί και στη διεθνή βιβλιογραφία, η ταχύτητα και η επιτάχυνση παίζουν καθοριστικό ρόλο στην πιθανότητα εμπλοκής σε κάποιο απρόοπτο περιστατικό στην οδό. Επίσης το αίσθημα κούρασης που μπορεί να έχει δημιουργηθεί στον οδηγό έπειτα από μια μεγάλη διαδρομή, ενδέχεται να αυξήσει τη πιθανότητα εμπλοκής σε συμβάν.

- Σε αντίθεση με τη χρονική διάρκεια του ενός λεπτού πριν, στη **διάρκεια ενός συμβάντος**, εκτός από την ταχύτητα που κυριαρχεί και σε αυτή την περίπτωση, οι μεταβλητές που έχουν τη μεγαλύτερη επιρροή είναι η απόκλιση του οχήματος από το μέσο της οδού και ο χρόνος μέχρι τη σύγκρουση. Η απόκλιση από το μέσο της οδού μπορεί να συμβαίνει διότι ο οδηγός κατά τη διάρκεια ενός περιστατικού ενδέχεται να χάσει τον πλήρη έλεγχο του οχήματος και ο χρόνος μέχρι τη σύγκρουση συνδέεται άμεσα με την ταχύτητα.

- Όσον αφορά στα δεδομένα που περιγράφουν την κατάσταση **ένα λεπτό πριν και κατά τη διάρκεια κάθε συμβάντος**, αυτά μπορούν να εκφραστούν με τρεις παράγοντες: έναν για την επιρροή της ταχύτητας και του χρόνου μέχρι τη σύγκρουση από το προπορευόμενο όχημα, έναν που περιγράφει την επιρροή της απόκλισης του οχήματος από το μέσο της οδού και τον παράγοντα της επιρροής της συνολικής διανυόμενης απόστασης του οχήματος στην ύπαρξη συμβάντος. Τα αποτελέσματα αυτά μπορούν να εξηγηθούν από τις επιμέρους περιπτώσεις χρονικών διαστημάτων που αναφέρθηκαν παραπάνω, καθώς αποτελούν το σύνολό τους.

ΠΕΡΙΕΧΟΜΕΝΑ

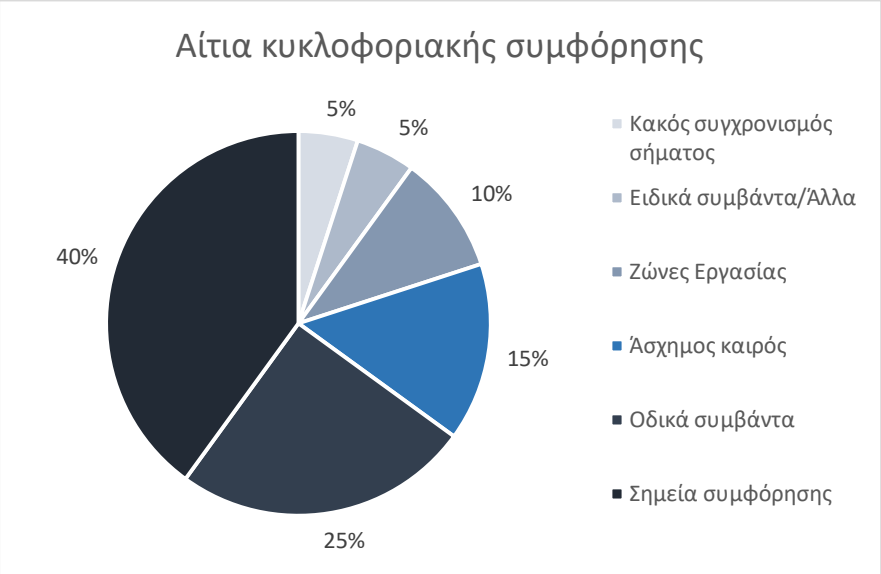
1. ΕΙΣΑΓΩΓΗ

1.1. Γενική Ανασκόπηση

Ο αυξημένος φόρτος εργασιών και υποχρεώσεων των ανθρώπων στη σύγχρονη εποχή καθιστά τις **μεταφορές** ως ένα πολύ σημαντικό και μεγάλο κομμάτι της ζωής τους. Για τις διάφορες καθημερινές τους δραστηριότητες, από την εργασία έως την αναψυχή, οι μεταφορές παίζουν κυρίαρχο ρόλο και αποτελούν σημαντικό κλάδο απασχόλησης εργασίας.

Η **διαχείριση της κυκλοφορίας** αναφέρεται στο συνδυασμό των μέτρων που συμβάλουν στη διατήρηση της αποτελεσματικότητας της κυκλοφορίας και στη βελτίωση της ασφάλειάς της. Της ασφάλειας και της αξιοπιστίας του συνολικού συστήματος οδικών μεταφορών.

Πριν από μερικά χρόνια, μια μελέτη στις Ηνωμένες Πολιτείες της Αμερικής (Η.Π.Α.) ξεκίνησε να ποσοτικοποιεί τις κύριες αιτίες κυκλοφοριακής συμφόρησης. Τα αποτελέσματα, κατά μέσο όρο για όλους τους δρόμους, φαίνονται στο Γράφημα 1.1 που ακολουθεί.



Γράφημα 1.1: Αίτια κυκλοφοριακής συμφόρησης

Αν και τα δεδομένα που φαίνονται στο παραπάνω γράφημα είναι ειδικά για τις Η.Π.Α., η υψηλή συχνότητα συμφόρησης δείχνει ζητήματα διαχείρισης δικτύου και ανεπάρκειας υποδομής. Αυτό συμβαίνει ιδίως σε διασταυρώσεις, και μπορεί να φανεί σε ολόκληρο τον κόσμο. Στο σχήμα φαίνεται συγκεκριμένα πως τα οδικά και ειδικά συμβάντα αποτελούν ένα ποσοστό της τάξης του 30% των συνολικών οδικών προβλημάτων που προκαλούν κυκλοφοριακή συμφόρηση, γεγονός που επιτάσσει την ανάγκη για μελέτη των συμβάντων αυτών.

Εκτός από τον ρόλο του στη διαχείριση κυκλοφορίας, ο εντοπισμός συμβάντων παίζει καθοριστικό ρόλο και στην ύπαρξη και διασφάλιση της **οδικής ασφάλειας**.

Οι τραυματισμοί στην οδική κυκλοφορία παραμένουν σημαντικό πρόβλημα δημόσιας υγείας. Όταν τα προϊόντα οδικής ασφάλειας χρησιμοποιούνται σωστά και αποτελεσματικά, μπορούν να βοηθήσουν στη διάσωση ζωών, στην πρόληψη ατυχημάτων και τραυματισμών και στη διατήρηση της λειτουργίας της κοινωνίας με τον βέλτιστο τρόπο.

Σημαντικό ρόλο στα οδικά συμβάντα και ατυχήματα παίζουν οι εξής παράγοντες (Φρατζεσκάκης και Γκόλιας 1994) :

- οι χρήστες της οδού
- η οδός και το περιβάλλον
- το όχημα

Ως **χρήστες της οδού** αναφέρονται όχι μόνο οι οδηγοί, αλλά και οι πεζοί. Όλοι είναι πιθανόν να εμπλακούν σε κάποιο συμβάν ή ατύχημα, η έκβαση του οποίου συνήθως εξαρτάται από την αντίδραση και τη συμπεριφορά του οδηγού.

Η οδός και το περιβάλλον μπορεί να αποτελούν καθοριστικό παράγοντα για την ύπαρξη ή μη συμβάντος, καθώς και για την εξέλιξη και έκβασή του. Είναι προφανές πως όσο καλύτερα είναι τα γεωμετρικά χαρακτηριστικά της οδού όπως η σήμανση , οι λωρίδες κυκλοφορίας, ο φωτισμός και το πλάτος ερεισμάτων, τόσο μικρότερη θα είναι η πιθανότητα εμπλοκής.

Αναμφίβολα ο πιο κρίσιμος παράγοντας για την πρόκληση ατυχήματος ή συμβάντος είναι ο **άνθρωπος**. Πολλοί παράγοντες επηρεάζουν τα οδηγικά χαρακτηριστικά του και σε πολλές περιπτώσεις μειώνουν την αποτελεσματικότητα της αντίδρασής του. Έρευνες έχουν δείξει πως η απόσπαση προσοχής όπως η χρήση κινητού τηλεφώνου, η συνομιλία με συνεπιβάτες, η κατανάλωση φαγητού και ποτού, είναι κάποιοι από τους παράγοντες που συμβάλλουν αρνητικά στη συμπεριφορά του οδηγού.

Η **απόσπαση προσοχής** μπορεί να επιφέρει τις εξής συνέπειες:

- αύξηση χρόνου αντίδρασης του οδηγού σε απρόβλεπτα συμβάντα
- μείωση αποστάσεων ασφαλείας
- μείωση της ικανότητας διατήρησης σωστής θέσης λωρίδας
- μείωση αντίληψης του οδηγού για το τι συμβαίνει γύρω του
- αγνόηση σημάτων ή αργή αντίδραση σε αυτά

Στο βαθμό απόσπασης προσοχής του οδηγού παίζουν ρόλο και **τα χαρακτηριστικά του οδηγού**, όπως το φύλο, η ηλικία και η οδηγική εμπειρία.

Γίνεται συνεπώς κατανοητό ότι η **διαχείριση περιστατικών κυκλοφορίας**, είναι μια σημαντική λύση σε περιπτώσεις τροχαίων ατυχημάτων, περιστατικών και άλλων μη προγραμματισμένων συμβάντων, όπως είναι η ξαφνική είσοδος ενός παρκαρισμένου οχήματος στην οδό ή ενός πεζού, που συμβαίνουν στο οδικό δίκτυο, και καταλήγουν συχνά σε δυνητικά επικίνδυνες καταστάσεις. Αυτό τονίζει την ανάγκη εντοπισμού των συμβάντων αυτών, προκειμένου να υπάρξει άμεση αντιμετώπισή τους και να αποφευχθούν όσο το δυνατόν περισσότερα από αυτά. Σε αυτό θα συμβάλλει η παρούσα

διπλωματική εργασία, με στόχο τον εντοπισμό συμβάντων ανάλογα με τα οδηγικά χαρακτηριστικά του οδηγού, εστιάζοντας στις επαρχιακές οδούς.

1.2. Στόχος

Ο στόχος της παρούσας διπλωματικής εργασίας **είναι ο εντοπισμός των συμβάντων με βάση τη μεταβολή των οδηγικών χαρακτηριστικών, σε υπεραστικές οδούς.**

Σύμφωνα με τα παραπάνω στοιχεία γίνεται σαφές ότι τα απρόβλεπτα συμβάντα και η συμπεριφορά του οδηγού πριν και κατά τη διάρκεια αυτών, σε περιπτώσεις απόσπασης προσοχής και μη, αποτελούν σημαντικό παράγοντα για τη διαχείριση της κυκλοφορίας, αλλά και για την οδική ασφάλεια. Τα οδηγικά χαρακτηριστικά θα μελετηθούν και θα αναλυθούν στην παρούσα διπλωματική, εστιάζοντας στην κατηγορία των υπεραστικών οδών. Η χρήση κατάλληλων στοιχείων από μια μεγάλη βάση δεδομένων, που προέκυψαν από έρευνα που πραγματοποιήθηκε με πείραμα σε προσομοιωτή και η κατάλληλη μεθοδολογία, θα συμβάλλουν στην εύρεση χρήσιμων αποτελεσμάτων.

Για την επίτευξη αυτού του στόχου είναι απαραίτητη η επιλογή κατάλληλης μεθόδου ανάλυσης και η εφαρμογή της με σκοπό την απόσπαση ορθών και σημαντικών αποτελεσμάτων. Σημαντική επίσης είναι η εφαρμογή του θεωρητικού υπόβαθρου των μαθηματικών μοντέλων που θα χρησιμοποιηθούν. Τα μοντέλα αυτά θα προβλέπουν αν ο οδηγός, σύμφωνα με τα χαρακτηριστικά του, εμπλέκεται σε συμβάν ή βρίσκεται λίγο πριν από αυτό.

Τελικός στόχος είναι τα αποτελέσματα αυτής της διπλωματικής εργασίας να αποτελέσουν πηγή για περαιτέρω έρευνα.

1.3. Μεθοδολογία διπλωματικής εργασίας

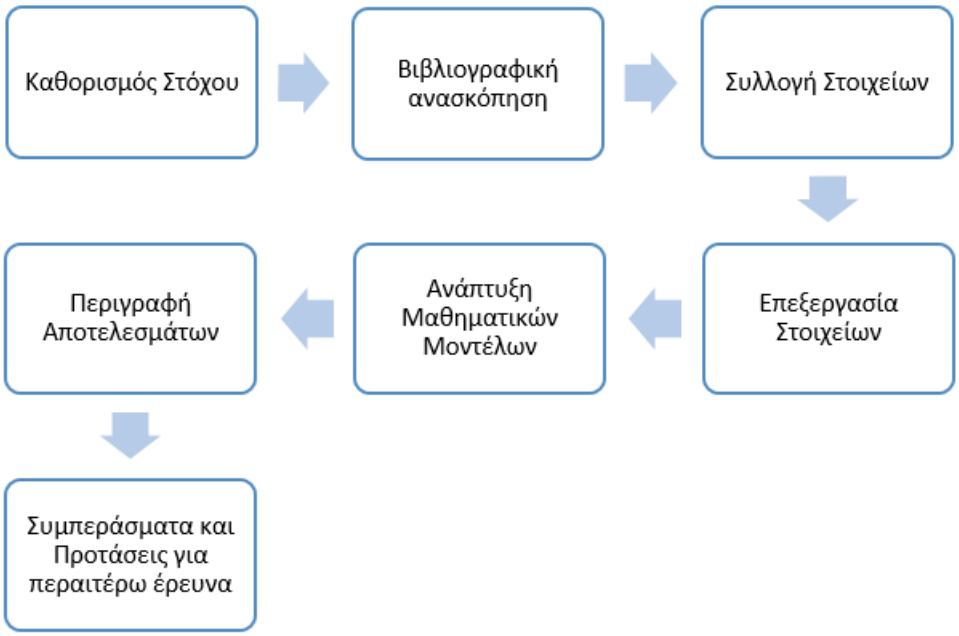
Για την επίτευξη του στόχου της διπλωματικής εργασίας ακολουθήθηκε η **μεθοδολογία** που περιγράφεται στη συγκεκριμένη ενότητα.

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

Ακολουθεί η συλλογή στοιχείων από μια μεγάλη βάση δεδομένων. Για την επεξεργασία των στοιχείων που συλλέχθηκαν, στην παρούσα διπλωματική, χρησιμοποιήθηκαν διάφορα μοντέλα στατιστικής ανάλυσης, προκειμένου να προκύψουν ορθά και χρήσιμα αποτελέσματα, για την πρόβλεψη της κατάστασης στην οποία βρίσκεται ο οδηγός, πριν ή κατά τη διάρκεια συμβάντος.

Τα παραπάνω βήματα οδηγούν στην επίτευξη των στόχων της διπλωματικής εργασίας και στην περιγραφή των συμπερασμάτων τους, καθώς δημιουργούν και ερωτήματα που χρήζουν περαιτέρω έρευνας.

Στο παρακάτω διάγραμμα (Γράφημα 1.2) παρουσιάζονται συνοπτικά τα βήματα της μεθοδολογίας εκπόνησης της συγκεκριμένης διπλωματικής εργασίας.



Γράφημα 1.2: Βήματα μεθοδολογίας διπλωματικής εργασίας

1.4. Δομή διπλωματικής εργασίας

Η δομή της διπλωματικής εργασίας καθορίζεται σύμφωνα με τη μεθοδολογία που αναφέρθηκε παραπάνω. Στην ενότητα αυτή παρουσιάζεται η σύνοψη των κεφαλαίων που την απαρτίζουν.

Το παρόν κεφάλαιο, που αποτελεί το ***κεφάλαιο 1***, είναι εισαγωγικό και αποτελεί μια σύντομη παρουσίαση της σχέσης μεταξύ της διαχείρισης κυκλοφορίας και οδικής ασφάλειας και των απρόβλεπτων συμβάντων και χαρακτηριστικών του οδηγού. Παρουσιάζονται παράγοντες που επιδρούν στην οδηγική συμπεριφορά και μεταβάλλονται κατά τη διάρκεια συμβάντων. Επίσης περιγράφεται ο στόχος της εργασίας και η μεθοδολογία που απαιτείται για την εκπόνησή της.

Στο ***κεφάλαιο 2*** παρατίθενται συναφείς έρευνες και μεθοδολογίες, που είναι χρήσιμες για την εκπόνηση της εργασίας, και αποτελούν τη βιβλιογραφική ανασκόπηση. Με βάση τα αποτελέσματα των ερευνών αυτών και των ζητημάτων που προκύπτουν για μελέτη, οριστικοποιείται το αντικείμενο της διπλωματικής εργασίας και των παραγόντων που θα εξεταστούν.

Στο επόμενο κεφάλαιο, ***κεφάλαιο 3***, παρουσιάζεται το θεωρητικό υπόβαθρο και οι μέθοδοι που απαιτούνται για τη στατιστική ανάλυση των στοιχείων. Περιγράφονται οι μαθηματικοί τύποι των μοντέλων, καθώς και οι έλεγχοι που πρέπει να ικανοποιούνται για την αποδοχή τους.

Στο ***κεφάλαιο 4***, περιγράφεται η διαδικασία συλλογής και επεξεργασίας των απαραίτητων, για αυτή τη διπλωματική εργασία, στοιχείων από μια βάση δεδομένων προϋπάρχουσας έρευνας.

Στο ***κεφάλαιο 5***, αναλύονται τα βήματα εκτέλεσης των μαθηματικών μοντέλων που ακολουθήθηκαν για τη στατιστική ανάλυση και πληρούν τα κριτήρια αποδοχής. Χρησιμοποιήθηκαν το μοντέλο λογιστικής παλινδρόμησης (logistic regression) και το μοντέλο τυχαίων δασών (random forest) για την πρόβλεψη της κατάστασης στην οποία βρίσκεται ο οδηγός, πριν ή κατά τη διάρκεια συμβάντος, και η ανάλυση παραγόντων (factor analysis) για την αναζήτηση ύπαρξης κοινών παραγόντων ανάμεσα σε μια ομάδα μεταβλητών.

Τα συμπεράσματα που προκύπτουν παρουσιάζονται στο ***κεφάλαιο 6***, τα οποία αποτελούν την ερμηνεία των μαθηματικών μοντέλων. Στο κεφάλαιο αυτό αναφέρονται και προτάσεις για περαιτέρω έρευνα σχετική με το αντικείμενο της παρούσας διπλωματικής εργασίας.

Η βιβλιογραφία παρουσιάζεται στο ***κεφάλαιο 7*** σε μορφή καταλόγου και περιλαμβάνει όλες τις πηγές που χρησιμοποιήθηκαν για την πραγματοποίηση της διπλωματικής εργασίας.

2. ΒΙΒΛΙΟΓΡΑΦΙΚΗ ΑΝΑΣΚΟΠΗΣΗ

2.1. Εισαγωγή

Στο κεφάλαιο αυτό παρουσιάζονται **συναφείς έρευνες** που έχουν δημοσιευτεί και είναι σχετικές με το αντικείμενο που πραγματεύεται η παρούσα διπλωματική εργασία, τον εντοπισμό συμβάντων με βάση τα οδηγικά χαρακτηριστικά. Η αναζήτηση επικεντρώνεται στη διερεύνηση παραγόντων που οδηγούν σε συμβάν ή ατύχημα, καθώς και στη μελέτη της μεταβολής των χαρακτηριστικών του οδηγού πριν και κατά τη διάρκεια ενός συμβάντος.

Στη συνέχεια θα παρουσιαστούν οι σχετικές έρευνες, οι μέθοδοι που χρησιμοποιήθηκαν για την ανάλυσή τους, καθώς και τα αποτελέσματα που προέκυψαν. Στο τέλος αυτής της ενότητας θα προσδιοριστούν πιθανές ελλείψεις που παρατηρήθηκαν, πάνω στις οποίες θα βασιστεί ο προσδιορισμός του στόχου και οι μέθοδοι ανάλυσης της διπλωματικής εργασίας.

2.2. Συναφείς έρευνες και μεθοδολογίες

Οι έρευνες που αναφέρονται παρακάτω εξέτασαν τους παράγοντες που οδηγούν σε λάθη κατά την οδήγηση, με έμφαση στην επιρροή της απόσπασης προσοχής στα οδηγικά συμβάντα και τον τρόπο με τον οποίο οι οδηγοί αντιδρούν σε αυτούς τους παράγοντες και προσπαθούν να αποφύγουν τις συγκρούσεις. Εξετάστηκε επίσης ο τρόπος με τον οποίο τα απρόβλεπτα συμβάντα επηρεάζουν τα οδηγικά χαρακτηριστικά,οι παράγοντες που επηρεάζουν την πιθανότητα ατυχήματος καθώς και τα οδηγικά χαρακτηριστικά πριν το ατύχημα.

2.2.1. Παράγοντες που οδηγούν σε λάθη κατά την οδήγηση και ο τρόπος με τον οποίο οι οδηγοί αποφεύγουν τις συγκρούσεις

Σε έρευνα των Papantoniou et al. (2019_a) προσδιορίστηκαν οι παράγοντες που οδηγούν σε οδηγικά λάθη μέσω πειράματος σε προσομοιωτή οδήγησης, σε δείγμα 95 οδηγών όλων των ηλικιών. Συγκεκριμένα, χρησιμοποιώντας για την ανάλυση των στοιχείων μοντέλα δομικών εξισώσεων, βρέθηκε ότι ο αντίκτυπος των χαρακτηριστικών του οδηγού και του τύπου της οδού είναι οι μόνοι στατιστικά σημαντικοί παράγοντες που επηρεάζουν την πιθανότητα λαθών στην οδήγηση. Είναι ενδιαφέρον, πως ούτε η συνομιλία με συνεπιβάτη ούτε η συζήτηση στο κινητό έχει στατιστικά σημαντικό αντίκτυπο στη λανθασμένη οδηγική συμπεριφορά, που τονίζει τη σημασία της παρούσας ανάλυσης και πιο συγκεκριμένα την ανάπτυξη ενός μέτρου που αντιπροσωπεύει συνολικά τη λανθασμένη οδηγική συμπεριφορά αντί για μεμονωμένες μεταβλητές σφαλμάτων οδήγησης. Ειδικότερα προέκυψε πως γυναίκες και μεγαλύτεροι ηλικιακά οδηγοί έχουν μεγαλύτερη πιθανότητα να εμφανίσουν λάθη στην οδηγική τους συμπεριφορά. Οι νέοι έχουν καλύτερα νοητικά και φυσικά χαρακτηριστικά με αποτέλεσμα να μειώνουν αυτήν την πιθανότητα και οι πιο μορφωμένοι και πιο έμπειροι εμφανίζουν μικρότερη πιθανότητα λάθους κατά την

οδήγηση. Όσον αφορά το περιβάλλον οδήγησης, στις επαρχιακές περιοχές υπάρχει μεγαλύτερη πιθανότητα οδηγικού λάθους από τις αστικές, καθώς επιτυγχάνονται μεγάλες ταχύτητες και ενδεχομένως η συγκέντρωση να είναι μικρότερη λόγω της συνήθως μεγάλης διάρκειας της διαδρομής. Συμπερασματικά, μεγαλύτερη πιθανότητα λάθους στην οδήγηση έχουν οι ηλικιωμένες γυναίκες, με χαμηλό επίπεδο εκπαίδευσης και μικρή εμπειρία, σε επαρχιακή οδό.

Όσον αφορά τον τρόπο με τον οποίο οι οδηγοί αποφεύγουν τις συγκρούσεις, έρευνα των Li et al. (2019), έδειξε ότι το φρένο είναι η πιο συχνή απόκριση των οδηγών σε τρία ενδεχόμενα συγκρούσεων. Τα τρία αυτά ενδεχόμενα ήταν η πλευρική και μετωπική σύγκρουση οχημάτων και η σύγκρουση με πεζό. Η έρευνα πραγματοποιήθηκε σε σύνολο 45 οδηγών και η ανάλυση έγινε με τη χρήση μικτών γραμμικών μοντέλων. Το σενάριο της πλευρικής σύγκρουσης είχε το μεγαλύτερο χρόνο μέχρι τη σύγκρουση και οι οδηγοί με καθυστερημένη αντίδραση έτειναν να αντισταθμίσουν τον κίνδυνο σύγκρουσης λαμβάνοντας μεγαλύτερο ρυθμό επιβράδυνσης. Συμπερασματικά, ο μεγάλος χρόνος αντίδρασης φρένων και η λανθασμένη κατεύθυνση ελιγμού ήταν οι κύριοι παράγοντες που οδήγησαν σε υψηλή πιθανότητα σύγκρουσης. Ο ελιγμός προκάλεσε καθυστέρηση στο φρενάρισμα και το φαινόμενο αυτό έδειξε ότι οι οδηγοί τείνουν να χρησιμοποιούν ένα διαισθητικό τρόπο λήψης αποφάσεων σε κρίσιμες καταστάσεις κυκλοφορίας.

2.2.1.1. Επιρροή της απόσπασης προσοχής στα οδηγικά συμβάντα

Ο όρος απόσπαση προσοχής περιλαμβάνει όλες τις δευτερεύουσες ενέργειες του οδηγού. Οι δευτερεύουσες ενέργειες ορίζονται ως οποιοσδήποτε περισπασμός που δεν σχετίζεται με την οδήγηση, όπως η ομιλία, το τραγούδι, το φαγητό ή η απόσπαση στο κινητό κάποιου (Monselise et al. 2019) .

Την επίδραση στην ταχύτητα και την προσαρμογή της οδηγικής συμπεριφοράς για τη χρήση κινητού τηλεφώνου κατά την οδήγηση εξέτασαν οι Choudhary και Velaga (2017). Στην έρευνα χρησιμοποιήθηκε δείγμα 100 οδηγών τριών διαφορετικών ηλικιακών ομάδων. Τα αποτελέσματα του μοντέλου γραμμικής παλινδρόμησης που χρησιμοποιήθηκε για την ανάλυση της ταχύτητας, έδειξαν ότι οι οδηγοί προσάρμοσαν σημαντικά τη συμπεριφορά τους στις αυξημένες ενέργειές τους, μειώνοντας τη **μέση ταχύτητά** τους κατά 2,62 m/s και 5,29 m/s κατά την ομιλία και και την αποστολή γραπτών μηνυμάτων αντίστοιχα. Έγινε επίσης σαφές ότι η απόσπαση προσοχής των οδηγών λόγω χρήσης κινητού τηλεφώνου, πρέπει να αντιμετωπιστεί μειώνοντας την ταχύτητά τους (κατά 30% ή και περισσότερο) για να διατηρηθεί το κατάλληλο επίπεδο ασφάλειας. Το 80% όμως των οδηγών δεν προσαρμόστηκε ικανοποιητικά στις νέες συνθήκες λόγω υποτίμησης του κινδύνου χρήσης κινητού τηλεφώνου σε συνδυασμό με τα ξαφνικά συμβάντα.

Οι Choudhary και Velaga (2019_a) σε άλλη έρευνα εξέτασαν πιο ειδικά την επίδραση της χρήσης του κινητού τηλεφώνου, σε δύο κατηγορίες οδηγών, νέων και επαγγελματιών, 25 και 24 οδηγών αντίστοιχα εξετάζοντας την ταχύτητα και την επιβράδυνση των οδηγών, καθώς και τη θέση του οχήματος και την κίνηση του τιμονιού με χρήση μικτών γραμμικών μοντέλων και αρνητικής διωνυμικής παλινδρόμησης. Τα αποτελέσματα έδειξαν ότι οι νέοι οδηγοί προσαρμόζονται λιγότερο στις απαιτήσεις που επιβάλλει η απόσπαση προσοχής και έχουν μικρότερο έλεγχο της ταχύτητας και της επιτάχυνσής τους, έχουν μικρότερες αποκλίσεις στο τιμόνι, αλλά

μεταβάλλεται περισσότερο η θέση τους στη λωρίδα. Ειδικότερα, η μέση ταχύτητα κατά τη χρήση κινητού τηλεφώνου μειώνεται περισσότερο στους έμπειρους οδηγούς, καθώς οι νέοι χαρακτηρίζονται από άγνοια κινδύνου. Επίσης η διακύμανση της επιβράδυνσης ήταν μεγαλύτερη στους νέους οδηγούς, οι οποίοι φαίνεται να είχαν μικρότερο έλεγχο του αυτοκινήτου.

Μια ακόμα έρευνα εκπονήθηκε από τους Yannis et al. (2014), στην οποία συγκεκριμένα εξετάστηκε η επίδραση της αποστολής γραπτών μηνυμάτων στην οδηγική συμπεριφορά των νέων και στην ασφάλεια σε αστικές και επαρχιακές οδούς, σε διάφορες καιρικές συνθήκες. Χρησιμοποιήθηκε δείγμα 34 οδηγών και προέκυψε με χρήση γραμμικής παλινδρόμησης πως η αποστολή γραπτών μηνυμάτων κατά την οδήγηση μειώνει τη μέση ταχύτητα και με χρήση διωνυμικής λογιστικής παλινδρόμησης πως αυξάνει το χρόνο αντίδρασης και σε αστικές και σε υπεραστικές οδούς.

Έρευνα για την επίδραση της αποστολής γραπτών μηνυμάτων, σύντομων και εκτεταμένων, στον **κίνδυνο ατυχήματος** κατά τη διάρκεια ενός ξαφνικού συμβάντος, σε προγενέστερη και μεταγενέστερη φάση, έκαναν οι Choudhary και Velaga (2018). Με χρήση γενικευμένων μικτών γραμμικών μοντέλων και σε δείγμα 49 οδηγών, και για τους δύο τύπους μηνυμάτων παρατηρήθηκε αυξημένος κίνδυνος ατυχήματος, ιδιαίτερα για τους άπειρους νέους οδηγούς. Ο αυξημένος χρόνος αντίδρασης για τη μείωση της ταχύτητας αύξησε σημαντικά τις πιθανότητες ατυχήματος. Ένα ακόμα συμπέρασμα στο οποίο οδηγήθηκαν είναι ότι οι μεγαλύτεροι, επαγγελματίες, οδηγοί δεν είναι σε θέση να μετριάσουν τον αυξημένο κίνδυνο ατυχήματος που σχετίζεται με την αποστολή γραπτών μηνυμάτων, λόγω της καθυστερημένης ανίχνευσης του συμβάντος κατά τη διάρκεια των ενεργειών τους.

Στη σύγκριση των περιπτώσεων, ομιλίας στο κινητό τηλέφωνο και αποστολής γραπτού μηνύματος με αυτό, προέκυψε πως η πιθανότητα ατυχήματος αυξάνεται 3 φορές με ομιλία και 4 φορές με γραπτά μηνύματα κατά την οδήγηση (Choudhary και Velaga 2017). Η έρευνα των Yannis et al. (2014) σε αστικές και επαρχιακές οδούς, έδειξε πως η πιθανότητα ατυχήματος αυξάνεται λόγω της απόσπασης προσοχής και της καθυστερημένης αντίδρασης. Οι οδηγοί με κινητό αφής αντιδρούν διαφορετικά όσον αφορά την ταχύτητα τους , έχουν όμως και αυτοί αυξημένη πιθανότητα ατυχήματος. Σε επαρχιακές οδούς η πιθανότητα ατυχήματος είναι 1.4 φορές μεγαλύτερη κατά την ανάγνωση γραπτών μηνυμάτων και 1.5 φορές μεγαλύτερη κατά τη γραφή μηνυμάτων σε σύγκριση με τις αστικές οδούς. Επίσης η γραφή μηνυμάτων αυξάνει περισσότερο την πιθανότητα ατυχήματος από την ανάγνωσή τους. Ειδικά σε αστικές οδούς είναι 2.8 φορές μεγαλύτερη, ενώ σε επαρχιακές οδούς 1.1 φορές μεγαλύτερη. Η πιθανότητα ατυχήματος επηρεάζεται επίσης από την αναλογία ταχύτητας του οδηγού προς τη μέση ταχύτητα. Συγκεκριμένα, οι οδηγοί με ταχύτητες υψηλότερες από τη μέση ταχύτητα, έχουν αυξημένη πιθανότητα ατυχήματος τόσο σε αστικό όσο και σε επαρχιακό περιβάλλον.

Σε έρευνα των Osman et al. (2019) για τον εντοπισμό δευτερευουσών ενεργειών του οδηγού και τον τύπο αυτών (κλήση στο κινητό τηλέφωνο,αποστολή γραπτού μηνύματος και συνομιλία με συνεπιβάτη), από δείγμα 373 οδηγών, προέκυψε ότι η συνολική ακρίβεια ανίχνευσης δευτερεύουσας ενέργειας κυμαινόταν από 66% έως 96% από τις διάφορες μεθόδους ανάλυσης, ενώ η μέθοδος Decision Tree επέφερε ακρίβεια 99,8%. Για τον προσδιορισμό του τύπου δευτερεύουσας ενέργειας η συνολική

ακρίβεια κυμαινόταν από 55% έως79%, με υψηλότερη ακρίβεια 82.2% που επιτεύχθηκε από τη μέθοδο τυχαίων δασών. Στην έρευνα αυτή εξετάστηκαν παράμετροι όπως η ταχύτητα, η πλευρική επιτάχυνση, η διαμήκης επιτάχυνση, η θέση του πεντάλ και ο ρυθμός εκτροπής, μαζί με τις αποκλίσεις τους για την εύρεση των αποτελεσμάτων. Καταλήγει στο συμπέρασμα ότι οι δευτερεύουσες αυτές ενέργειες λειτουργούν ανασταλτικά στην προσπάθεια πρόληψης ατυχημάτων και υπενθυμίζει τη σημασία της συγκέντρωσης κατά την οδήγηση ιδιαίτερα όταν πραγματοποιούνται αλλαγές στη συμπεριφορά των οδηγών.

Οι Choudhary και Velaga (2019_b) εκπόνησαν έρευνα για μια ακόμη περίπτωση απόσπασης προσοχής και συγκεκριμένα για την ανάλυση του κινδύνου που σχετίζεται με την κατανάλωση φαγητού, ποτού και την αποστολή γραπτών μηνυμάτων κατά την οδήγηση σε μη σηματοδοτημένες διασταυρώσεις, σε δείγμα 89 οδηγών. Από τη χρήση του μοντέλου γενικευμένων εκτιμήσεων εξισώσεων για την ανάλυση, παρατηρήθηκε πως η απόσταση αντίδρασης του οδηγού σε ενδεχόμενο συμβάν κατά την αποστολή γραπτών μηνυμάτων είναι μεγαλύτερη στη δευτερεύουσα οδό. Αντίθετα η κατανάλωση φαγητού και ποτού επιδρά πιο αρνητικά στην κύρια οδό. Ο κίνδυνος ατυχήματος στην κύρια οδό είναι αυξημένος με την κατανάλωση φαγητού και ποτού και παρατηρήθηκε μεγαλύτερη πιθανότητα ατυχήματος κατά την είσοδο σε αυτή για τους οδηγούς που προσέγγιζαν τη διασταύρωση με αυξημένη ταχύτητα. Συγκεκριμένα αύξηση 1m/s στην ταχύτητα προσέγγισης επιφέρει αύξηση 26% στον κίνδυνο ατυχήματος. Επίσης οι έμπειροι οδηγοί είχαν 74% μικρότερη πιθανότητα ατυχήματος συγκριτικά με τους άπειρους. Όσον αφορά τη χρονοαπόσταση από το μπροστινό όχημα παρατηρήθηκε πως η επιλογή μικρών κενών ($\leq$3s) αύξανε την πιθανότητα ατυχήματος συγκριτικά με την επιλογή μεγαλύτερων κενών ($\geq$4s). Συμπερασματικά η αύξηση της ταχύτητας επιφέρει και αύξηση των ατυχημάτων.

2.2.2. Ο τρόπος με τον οποίο ένα απρόβλεπτο συμβάν επηρεάζει τα οδηγικά χαρακτηριστικά και οι παράγοντες που επηρεάζουν την πιθανότητα ατυχήματος

Σε έρευνα των Papantoniou et al. (2018) εξετάστηκε ο τρόπος με τον οποίο ένα απρόβλεπτο συμβάν επηρεάζει τα χαρακτηριστικά της οδήγησης που σχετίζονται με την ταχύτητα. Για δείγμα 95 οδηγών και με χρήση μοντέλων γραμμικής παλινδρόμησης για την ανάλυση, βρέθηκε πως έπειτα από ένα τέτοιο συμβάν η οδήγηση αλλάζει από τα φυσιολογικά επίπεδα σε πιο προσεκτικές ενέργειες. Η αλλαγή αυτή οφείλεται στις πολλαπλές ενέργειες που καλείται να εκτελέσει ο οδηγός, όπως η λήψη αποφάσεων και η προσαρμογή της οδηγικής συμπεριφοράς στις εκάστοτε συνθήκες. Ακόμα το στυλ οδήγησης, το φύλο, η ηλικία μπορούν να επηρεάσουν τον τρόπο με τον οποίο οι οδηγοί αντιλαμβάνονται τον κόσμο τους, το καθήκον τους και τον κίνδυνο τον ενεργειών τους. Οι οδηγοί μιλώντας στο κινητό τηλέφωνο προσέχουν πολύ περισσότερο μετά από ένα απροσδόκητο συμβάν, επειδή η χρήση του ενός χεριού στο τηλέφωνο λειτουργεί ως υπενθύμιση στον οδηγό της πιθανής απειλής για την ασφάλεια, που δημιουργεί η χρήση του τηλεφώνου. Αφετέρου ενώ συνομιλεί με τον επιβάτη, ο οδηγός έχει χαμηλότερο επίπεδο αντισταθμιστικής συμπεριφοράς, ωστόσο η προσοχή του πιο συχνά εκτρέπεται από το δρόμο. Επιπλέον, ποσοτικοποιήθηκε η επίδραση πολλών χαρακτηριστικών του οδηγού και του οδικού περιβάλλοντος για τις διαφορετικές στρατηγικές επιτάχυνσης μετά από ένα απροσδόκητο συμβάν.

Οι Papantoniou et al. (2019_b) εξέτασαν τους παράγοντες που επηρεάζουν την πιθανότητα ατυχήματος στα απροσδόκητα συμβάντα, είτε υπάρχει απόσπαση προσοχής, συνομιλία με συνεπιβάτη και χρήση κινητού τηλεφώνου, είτε όχι, μέσω πειράματος σε προσομοιωτή οδήγησης, σε δείγμα 95 οδηγών, σε αστικές και επαρχιακές οδούς. Όσον αφορά την απόσπαση της προσοχής του οδηγού, τα αποτελέσματα των μοντέλων δομικής εξίσωσης δείχνουν ότι η χρήση κινητού τηλεφώνου έχει αρνητική επίδραση στον κίνδυνο ατυχήματος επιβεβαιώνοντας την αρχική υπόθεση, ότι όταν μιλούν στο κινητό τηλέφωνο οι οδηγοί είναι δύσκολο να αντιμετωπίσουν ένα απροσδόκητο συμβάν και κατά συνέπεια είναι πιο πιθανό να οδηγηθεί σε ατύχημα. Εστιάζοντας στην οδηγική απόδοση, η συνομιλία με τον επιβάτη δε βρέθηκε να έχει στατιστικά σημαντική επίδραση, που δείχνει ότι τα οδηγικά χαρακτηριστικά δεν αλλάζουν την οδηγική τους απόδοση, ενώ ο οδηγός συνομιλεί με έναν επιβάτη, σε σύγκριση με τους μη αποσπασμένους οδηγούς.

Έρευνα για τους παράγοντες που σχετίζονται με τον κίνδυνο ατυχήματος εκπόνησαν οι Monselise et al. (2019) χρησιμοποιώντας δεδομένα οδηγικής συμπεριφοράς από 7707 διαδρομές, και προγνωστική ανάλυση. Σύμφωνα με το μοντέλο ανάλυσης, ο πιο σημαντικός παράγοντας που προκαλεί τα ατυχήματα με όχημα είναι η συμπεριφορά του οδηγού. Η μελέτη επικεντρώνεται κυρίως σε ατυχήματα σε διασταυρώσεις, και δείχνει ότι ο τύπος ελιγμών έχει αντίκτυπο στην εμφάνιση ατυχήματος. Ο πιο συχνός ελιγμός που κατέληξε σε ατύχημα είναι μια δεξιά στροφή ακολουθούμενη από είσοδο σε θέση στάθμευσης. Επίσης βρέθηκε ότι ορισμένες δευτερεύουσες ενέργειες είναι πιο πιθανό να προκαλέσουν ατυχήματα από άλλες. Σε περισσότερα από τα μισά ατυχήματα που αφορούσαν τη χρήση κινητού τηλεφώνου, το τηλέφωνο ήταν αυτό που συνέβαλε στο ατύχημα. Επίσης η διάρκεια της δευτερεύουσας ενέργειας έπαιξε καθοριστικό ρόλο στην πρόκληση ατυχήματος ή όχι. Συγκεκριμένα παρατηρήθηκε πως αν μια δευτερεύουσα ενέργεια διαρκούσε παραπάνω από 6 δευτερόλεπτα οδηγούσε σε ατύχημα.

2.2.3. Οδηγικά χαρακτηριστικά πριν το ατύχημα

Ανάλυση δεδομένων οδηγικής συμπεριφοράς, από δείγμα 987 οδηγών με 3604 συμβάντα συνολικά, για την εξέταση των οδηγικών χαρακτηριστικών πριν το ατύχημα έκαναν σε έρευνά τους οι Papazikou et al. (2019). Εξετάστηκε ολόκληρη η ακολουθία της σύγκρουσης, από μια κανονική κατάσταση οδήγησης μέχρι ένα συμβάν σύγκρουσης ή παραλίγο σύγκρουσης. Οι στόχοι ήταν να εξερευνήσουν την κινηματική του οχήματος πριν από το συμβάν, να διερευνήσουν τους δεικτες κινδύνου σύγκρουσης, να ανιχνεύσουν τα πρώτα στάδια της ανάπτυξης ατυχημάτων και να εξετάσουν περαιτέρω τους παράγοντες που επηρεάζουν το χρονικό διάστημα μέχρι τη σύγκρουση. Τα αποτελέσματα της ιεραρχικής γραμμικής μοντελοποίησης που χρησιμοποιήθηκε αποκάλυψαν ότι η διαμήκης επιτάχυνση, η πλευρική επιτάχυνση και ο ρυθμός εκτροπής μπορεί να είναι αξιόπιστοι δείκτες για την ανίχνευση αποκλίσεων από την κανονική οδήγηση. Επίσης οι τιμές του χρόνου μέχρι τη σύγκρουση επηρεάζονται από τον τύπο οχήματος, την ταχύτητα του οχήματος, τη διαμήκη επιτάχυνση και το χρόνο κατά το ατύχημα. Κατά τη διάρκεια ολόκληρου του ατυχήματος, ο δείκτης επιβράδυνσης, πλευρικής επιτάχυνσης και ρυθμού εκτροπής τείνουν να μειώνονται απότομα περίπου 10-20 δευτερόλεπτα πριν από το συμβάν. Η κρίσιμη χρονική στιγμή, όπου η τιμή του χρόνου μέχρι τη σύγκρουση αρχισε να πέφτει,

ήταν 1.62 λεπτά πριν το συμβάν. Αυτά τα ευρήματα θα μπορούσαν να είναι χρήσιμα για τον αποτελεσματικότερο και έγκαιρο εντοπισμό και την αποφυγή συντριβής.

2.3. Σύνοψη

Στον παρακάτω πίνακα (Πίνακας 2.1) παρουσιάζονται συνοπτικά τα στοιχεία των ερευνών που αναφέρθηκαν προηγουμένως, με σκοπό να συμβάλει στον εντοπισμό των χρήσιμων συμπερασμάτων που προκύπτουν και των ελλείψεων, βάση των οποίων θα γίνει κατανοητός ο στόχος της παρούσας διπλωματικής εργασίας.

Πίνακας 2.1: Συνοπτικά στοιχεία και αποτελέσματα

Έρευνα	Τύπος αποσπάσεως	Κατηγορία δείγματος	Πλήθος Δείγματος	Μέθοδος Ανάλυσης	Αποτελέσματα
Choudhary, Velaga (2017)	ομιλία και αποστολή γραπτών μηνυμάτων μέσω κινητού τηλεφώνου	σύνολο οδηγών	100	1) Linear regression 2) Binary logistic regression	1) μείωση μέσης ταχύτητας 2) αύξηση πιθανότητας ατυχήματος
Choudhary, Velaga (2019_a)	χρήση κινητού τηλεφώνου	νέοι και επαγγελματίες οδηγοί	25+24=49	1) Multiple linear regression 2) Negative binomial regression	1) μεγαλύτερη μείωση ταχύτητας στους έμπειρους οδηγούς 2) μεγαλύτερη διακύμανση επιτάχυνσης στους νέους οδηγούς
Yannis et al. (2014)	αποστολή γραπτών μηνυμάτων	νέοι οδηγοί σε αστικές και υπεραστικές οδούς	34	1) Log-normal linear regression 2) Binary logistic regression	1) μείωση μέσης ταχύτητας 2) αύξηση χρόνου αντίδρασης
Choudhary, Velaga (2018)	αποστολή γραπτών μηνυμάτων κατά τη διάρκεια ξαφνικού συμβάντος	σύνολο οδηγών	49	Generalized linear mixed models (GLMM)	1) αύξηση πιθανότητας ατυχήματος 2) αδυναμία των μεγαλύτερων οδηγών να μετριάσουν την

					πιθανότητα ατυχήματος
Osman et al. (2019)	δευτερεύουσες ενέργειες (εντοπισμός)	σύνολο οδηγών	373	1) Decision Tree 2) Random Forest	ανασταλτική επίδραση των δευτερευουσών ενεργειών στην πρόληψη ατυχήματος
Choudhary, Velaga (2019b)	φαγητό-ποτό-αποστολή γραπτων μηνυμάτων	μη σηματοδοτημένες διασταυρώσεις	89	Generalized Estimating Equations (GEE)	1) αύξηση πιθανότητας ατυχήματος 2) μικρότερη πιθανότητα ατυχήματος οι έμπειροι οδηγοί
Έρευνα	**Αντικείμενο Έρευνας**	**Κατηγορία Δείγματος**	**Πλήθος Δείγματος**	**Μέθοδος Ανάλυσης**	**Αποτελέσματα**
Papantoniou et al. (2019$_a$)	παράγοντες που οδηγούν σε οδηγικά λάθη	σύνολο οδηγών	95	Structural equation models	1)χαρακτηριστικά οδηγού 2)τύπος της οδού
Li et al. (2019)	τρόπος αποφυγής συγκρούσεων	σύνολο οδηγών	45	Linear mixed models	1) φρένο η πιο συχνή απόκριση 2) ο μεγάλος χρόνος αντίδρασης φρεναρίσματος και ο λανθασμένος ελιγμός αυξάνουν την πιθανότητα σύγκρουσης
Papantoniou et al. (2018)	επιρροή απρόσμενου συμβάντος στα χαρακτηριστικά της οδήγησης που σχετίζονται με την ταχύτητα	σύνολο οδηγών σε αστική και επαρχιακή οδό	95	Linear regression	έπειτα από απρόσμενο συμβάν η οδήγηση γίνεται πιο προσεκτική

Papantoniou et al. (2019b)	παράγοντες που επηρεάζουν την πιθανότητα ατυχήματος στα απροσδόκητα συμβάντα	σύνολο οδηγών σε αστική και επαρχιακή οδό	95	Structural equation models	1) η χρήση κινητού τηλεφώνου αυξάνει την πιθανότητα ατυχήματος 2) η συνομιλία με επιβάτη δεν επιδρά τόσο αρνητικά
Monselise et al.	παράγοντες που σχετίζονται με τον κίνδυνο ατυχήματος	διαδρομές	7707	Gradient Boosted Classification with Grid Search	συμπεριφορά του οδηγού ο πιο σημαντικός παράγοντας πρόκλησης ατυχημάτων
Papazikou et al. (2019)	οδηγικά χαρακτηριστικά πριν το ατύχημα	σύνολο οδηγών	3604 συμβάντα από 987 οδηγούς	Hierarchical Linear Modelling	μείωση διαμήκους και πλευρικής επιτάχυνσης και ρυθμού εκτροπής

Από την παραπάνω σύνοψη προκύπτει το συμπέρασμα πως έχουν αναλυθεί εκτενώς τα οδηγικά χαρακτηριστικά που επηρεάζει η απόσπαση προσοχής, για την πιθανότητα ατυχήματος, ιδιαίτερα η χρήση κινητού τηλεφώνου αλλά και άλλες δευτερεύουσες ενέργειες. Έχει εξεταστεί η επιρροή απρόσμενων συμβάντων στα οδηγικά χαρακτηριστικά, ο τρόπος αποφυγής των συγκρούσεων καθώς και τα χαρακτηριστικά αυτά πριν από ένα ατύχημα, με τη χρήση διάφορων μοντέλων στατιστικής ανάλυσης. Έχει επομένως πραγματοποιηθεί περιγραφική στατιστική των οδηγικών χαρακτηριστικών, που μεταβάλλονται λόγω απόσπασης προσοχής, λόγω ενός απρόσμενου συμβάντος αλλά και πριν από ένα ατύχημα.

Παρατηρήθηκε όμως, πως δεν υπάρχει επαρκής έρευνα όσον αφορά στα χαρακτηριστικών οδήγησης πριν και κατά τη διάρκεια ενός συμβάντος, συγκριτικά, έρευνα η οποία μπορεί να οδηγήσει στον **εντοπισμό συμβάντων σύμφωνα με τα οδηγικά χαρακτηριστικά**, με αποτέλεσμα να χρειάζεται περαιτέρω ανάλυση. Συλλέγοντας στοιχεία από πείραμα που έχει διεξαχθεί σε προσομοιωτή, εστιάζοντας

στις επαρχιακές οδούς, την επεξεργασία τους και με τα κατάλληλα μοντέλα ανάλυσης, σκοπός της παρούσας διπλωματικής εργασίας είναι να προβλέψει με βάση τα οδηγικά χαρακτηριστικά αν ο οδηγός βρίσκεται ή όχι στη διάρκεια ενός συμβάντος.

3. ΘΕΩΡΗΤΙΚΟ ΥΠΟΒΑΘΡΟ

3.1. Εισαγωγή

Σε αυτό το κεφάλαιο θα αναλυθεί το **θεωρητικό υπόβαθρο** που χρησιμοποιήθηκε **για τη στατιστική ανάλυση** αυτής της διπλωματικής εργασίας. Με σκοπό την εύρεση κατάλληλων και χρήσιμων αποτελεσμάτων χρησιμοποιήθηκαν τα μοντέλα διωνυμικής λογιστικής παλινδρόμησης και τυχαίων δασών για την επεξεργασία των μεταβλητών με στόχο την πρόβλεψη των συμβάντων, και η ανάλυση παραγόντων για την αναζήτηση ύπαρξης κοινών παραγόντων ανάμεσα σε μια ομάδα μεταβλητών. Χρησιμοποιήθηκε επίσης η μέθοδος Boruta για την εκτίμηση της σημαντικότητας των κάποιων μεταβλητών. Οι μεταβλητές που χρησιμοποιήθηκαν αποτελούν οδηγικά στοιχεία που απομονώθηκαν από μεγάλη βάση δεδομένων που είχε συλλεχθεί από προσομοιωτή οδήγησης σε προηγούμενη έρευνα. Εκτός από το θεωρητικό υπόβαθρο των μαθηματικών προτύπων γίνεται και αναφορά στον τρόπο αξιολόγησης και αποδοχής τους.

3.2. Μαθηματικά Πρότυπα

3.2.1. Διωνυμικό λογιστικό μοντέλο (binomial logistic regression)

Τα διωνυμικά μοντέλα λογιστικής παλινδρόμησης (DR Cox 1958) χρησιμοποιούνται για την αναζήτηση της σχέσης μεταξύ μίας διακριτής εξαρτημένης μεταβλητής και μίας ή περισσότερων ανεξάρτητων μεταβλητών, η οποία περιγράφεται από την εξίσωση του μοντέλου. Εξαρτημένη είναι η μεταβλητή της οποίας γίνεται η πρόβλεψη και ανεξάρτητη είναι η μεταβλητή η οποία έχει δεδομένη τιμή και χρησιμοποιείται για την πρόβλεψη της εξαρτημένης.

Η μορφή της εξίσωσης είναι η εξής:

$$\mathbf{y_i = logit\ (P_i)} = ln\frac{Pi}{1-Pi} = \mathbf{\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta 3 x 3 i + \ldots + \beta_n x_{ni}}$$

όπου:

n: πλήθος ανεξάρτητων μεταβλητών

$\beta_1, \beta_2, \ldots, \beta_n$: συντελεστές παλινδρόμησης των ανεξάρτητων μεταβλητών $x_1, x_2, \ldots, x_n$

β_0: σταθερός όρος εξίσωσης

P_i: προβλεπόμενη πιθανότητα (0 ή 1)

Η εξαρτημένη μεταβλητή, και αντίστοιχα το μοντέλο, καλείται δυαδική (Binary) ή διωνυμική (Binomial) καθώς λαμβάνει είτε τιμή 1 με πιθανότητα επιτυχίας P είτε τιμή 0 με πιθανότητα αποτυχίας 1-P.

Στη στατιστική ανάλυση των στοιχείων που έγινε σε αυτή την εργασία για τον εντοπισμό της κατάστασης στην οποία βρίσκεται ο οδηγός, εντός ή εκτός συμβάντος, η εξαρτημένη μεταβλητή ήταν το συμβάν. Οι πιθανές τιμές που μπορούσε να πάρει ήταν 1 ή 0, για την ύπαρξη ή όχι συμβάντος αντίστοιχα. Για το λόγο αυτό ένα από τα μοντέλα που εξετάστηκαν ήταν αυτό της διωνυμικής λογιστικής παλινδρόμησης.

3.2.2 Μοντέλο τυχαίων δασών (Random Forests)

Το μοντέλο τυχαίων δασών (Breiman 2001) δημιουργεί τυχαία δάση που αποτελούνται από πολλά δέντρα αποφάσεων. Ο αλγόριθμος που εφαρμόζεται για τη δημιουργία ενός τυχαίου δάσους είναι ο εξής:

1. Δημιουργία κάθε δέντρου από ένα ανεξάρτητο δείγμα εκκίνησης (bootstrap sample), το οποίο επιλέγεται από τα δεδομένα εκπαίδευσης (training data) με τη μέθοδο της αντικατάστασης.
   Σε κάθε κόμβο γίνεται:
   - Επιλογή m μεταβλητών τυχαία από όλες τις δυνατές M μεταβλητές
   - Εύρεση βέλτιστου διαχωρισμού των m μεταβλητών
2. Ανάπτυξη των δέντρων κατηγοριοποιώντας (ταξινομώντας) τα δεδομένα στο μέγιστο βαθμό.
3. Κατάταξη των δέντρων ώστε να προβλεφθεί η συμπεριφορά των καινούργιων δέντρων.
4. Κατάταξη των εναπομεινάντων (M-m) δεδομένων, τα οποία ονομάζονται " out of bag" ή "oob", σε κλάσεις των δέντρων του δάσους.
5. Εξετάζεται αν η κλάση που υπερέχει ως επιλογή από τα περισσότερα δέντρα απόφασης είναι η πραγματική κλάση του κάθε δεδομένου εισαγωγής. Ο ρυθμός σφάλματος αυτής της κατηγοριοποίησης (oob error rate) συνιστά το ρυθμό σφάλματος πρόγνωσης του δάσους.

Η επιλογή χρήσης του συγκεκριμένου μοντέλου έγινε διότι προσφέρει πολλά πλεονεκτήματα, κρίσιμα για την ανάλυση της παρούσας διπλωματικής εργασίας:

1. Μπορεί να χρησιμοποιηθεί τόσο για ταξινομήσεις όσο και για διαδικασίες παλινδρόμησης.
2. Λειτουργεί καλά τόσο με κατηγορικές όσο και με συνεχείς μεταβλητές.
3. Παρέχει υψηλή ακρίβεια καθώς
4. Το σφάλμα γενίκευσης είναι αρκετά περιορισμένο από τη στιγμή που αναπτύσσεται ένας πολύ μεγάλος αριθμός δέντρων με αποτέλεσμα να είναι απίθανο να εμφανιστεί πρόβλημα υπέρ-εκπαίδευσης (over fitting).
5. Η τυχαία επιλογή των μεταβλητών πρόβλεψης μειώνει τη σχέση των μεγάλων και un-pruned δέντρων, κάτι που κάνει τη μέθοδο αρκετά αμερόληπτη.
6. Δεν απαιτεί κλιμάκωση χαρακτηριστικών (τυποποίηση και κανονικοποίηση) καθώς χρησιμοποιεί προσέγγιση κατά κανόνα για τον υπολογισμό της απόστασης.

7. Αντιμετωπίζει αποτελεσματικά τις γραμμικές παραμέτρους.
8. Είναι ανθεκτικό στα ακραία σημεία και μπορεί να τα χειριστεί αυτόματα.
9. Έχει τη δυνατότητα να χειριστεί τιμές που ενδεχομένως να λείπουν και να διατηρήσει την ακρίβεια ενός μεγάλου ποσοστού δεδομένων.
10. Έχει τη δυνατότητα να χειρίζεται μεγάλο σύνολο δεδομένων.
11. Είναι πολύ σταθερό. Ακόμα και αν εισαχθεί ένα νέο στοιχείο στη βάση δεδομένων, ο συνολικός αλγόριθμος δεν επηρεάζεται πολύ, καθώς το νέο δεδομένο ενδέχεται να επηρεάσει ένα δέντρο, αλλά είναι πολύ δύσκολο να επηρεάσει όλα τα δέντρα.

3.3. Κριτήρια αποδοχής μοντέλων

3.3.1. Βασικά κριτήρια ελέγχου λογιστικού μοντέλου

Παρακάτω αναφέρονται τα βασικά κριτήρια ελέγχου για την αξιολόγηση και την αποδοχή των μοντέλων. Απαραίτητη προϋπόθεση είναι ο έλεγχος της συσχέτισης μεταξύ των μεταβλητών, δηλαδή οι ανεξάρτητες μεταβλητές πρέπει να είναι γραμμικώς ανεξάρτητες μεταξύ τους.

Λογική εξήγηση συντελεστών μοντέλου
Στην εξίσωση που θα προκύψει από τα μοντέλα εξετάζεται αν τα πρόσημα των συντελεστών παλινδρόμησης ($\boldsymbol{\beta_i}$) έχουν λογική ερμηνεία. Γίνεται, δηλαδή, έλεγχος βάσει του πρόσημου των για το αν η εξαρτημένη μεταβλητή αναμένεται να αυξηθεί ή να μειωθεί αν το πρόσημο των συντελεστών είναι θετικό ή αρνητικό αντίστοιχα. Σε περίπτωση που τα πρόσημα αυτά δεν έχουν λογική ερμηνεία, η αντίστοιχη μεταβλητή θα απορριφθεί.

Στατιστική σημαντικότητα:
Για την επιλογή ενός μοντέλου προσδιορίζεται το επίπεδο εμπιστοσύνης, το οποίο πρέπει να έχει υψηλή τιμή.
Για τα λογιστικά μοντέλα γίνεται ο έλεγχος Wald test (z-test), με τον εξής τύπο:

$$\mathbf{z_i} = \frac{\beta i}{s\beta i}$$

όπου:
β_i: οι συντελεστές παλινδρόμησης των ανεξάρτητων μεταβλητών x_i
$s_{\beta i}$: το τυπικό σφάλμα των συντελεστών παλινδρόμησης β_i

Ενδεικτικές τιμές του συντελεστή z είναι για 95% επίπεδο εμπιστοσύνης 1.7 και για 90% επίπεδο εμπιστοσύνης 1.3.

3.3.2. Μήτρα σύγχυσης (Confusion Matrix)

Η μήτρα σύγχυσης χρησιμοποιείται για τη μέτρηση της απόδοσης ενός συστήματος για δύο κλάσεις ή παραπάνω και στην παρούσα διπλωματική χρησιμοποιήθηκε για την αξιολόγηση των μοντέλων διωνυμικής λογιστικής παλινδρόμησης και τυχαίων δασών.

Η αξιολόγηση ενός μοντέλου κατηγοριοποίησης περιλαμβάνει την εξέταση της απόδοσής του σε ένα σύνολο δεδομένων που συνήθως είναι διαφορετικό από αυτό της εκπαίδευσής του. Για καθένα από τα στιγμιότυπα της βάσης εξέτασης (test), το μοντέλο κατηγοριοποίησης θα είναι είτε σωστό (θα του αναθέσει την προβλεπόμενη-επισημειωμένη κλάση) είτε εσφαλμένο. Βάσει αυτού, συμπεραίνεται πως διακρίνοντας το σύνολο των στιγμιοτύπων που ο ταξινομητής κατηγοριοποίησε σωστά ή εσφαλμένα, μπορεί να γίνει μια πρώτη εκτίμηση σχετικά με την απόδοση του εξεταζόμενου ταξινομητή.

Σε μια βάση δεδομένων με δύο μόνο κλάσεις ορίζονται τέσσερις περιπτώσεις κατηγοριοποίησης των πλειάδων της βάσης. Στις παρούσες αξιολογήσεις των μοντέλων τα στοιχεία που έδειχναν ύπαρξη συμβάντος ορίστηκαν ως θετικά, ενώ τα στοιχεία που έδειχναν μη ύπαρξη συμβάντος ως αρνητικά. Οι τέσσερις περιπτώσεις κατηγοριοποίησης των πλειάδων της βάσης είναι οι εξής:

1. ***Αληθώς Θετικά (True Positives – TP):*** Το πλήθος των στιγμιοτύπων της βάσης (+), ύπαρξη συμβάντος, που κατηγοριοποιήθηκαν ως (+) από τον ταξινομητή.
2. ***Αληθώς Αρνητικά (True Negative – TN):*** Το πλήθος των στιγμιοτύπων που ανήκουν στην κλάση (-), μη ύπαρξη συμβάντος, και ο ταξινομητής κατηγοριοποίησε ως (-).
3. ***Ψευδώς Θετικά (False Positive – FP):*** Είναι το πλήθος των παραδειγμάτων της κλάσης (-), μη ύπαρξη συμβάντος, που εσφαλμένα ο ταξινομητής κατηγοριοποίησε ως (+), ύπαρξη συμβάντος.
4. ***Ψευδώς Αρνητικά (False Negative – FN):*** Είναι το πλήθος των παραδειγμάτων της κλάσης (+), ύπαρξη συμβάντος, που εσφαλμένα κατηγοριοποιήθηκαν από τον ταξινομητή ως (-), μη συμβάντος.

Με βάση τις προαναφερθείσες περιπτώσεις, η αντίστοιχη μήτρα σύγχυσης θα είχε τη μορφή του Πίνακα 3.1

Πίνακας 3.1: Πιθανά αποτελέσματα για την πρόβλεψη ύπαρξης (+) ή μη (-) συμβάντος

	Κατηγοριοποίηση Ταξινομητή (Πρόβλεψη)	
Πραγματική Κλάση	**Συμβάν (+)**	**Όχι συμβάν (-)**
Συμβάν (+)	*TP*	*FN*
Όχι συμβάν (-)	*FP*	*TN*

Με βάση των πίνακα σύγχυσης οι μετρήσεις που χρησιμοποιούνται ευρέως περιλαμβάνουν (Catrakazas et. al. 2019):

3.3.2.1. Ορθότητα (accuracy)

Ορίζεται ως η συνολική ακρίβεια ή το ποσοστό των σωστών προβλέψεων του μοντέλου:

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

Αντί της ορθότητας μπορεί να χρησιμοποιηθεί το **μέτρο του λόγου σφάλματος (error rate)** ή λόγος εσφαλμένων κατηγοριοποιήσεων (misclassification rate), το οποιο κατ' αντιστοιχία εκφράζει το βαθμό εσφαλμένων κατηγοριοποιήσεων του **ταξινομητή:**

$$error\ rate = 1 - accuracy$$

3.3.2.2. Στατιστικός Συντελεστής Κάππα (Kappa Statistic)

Ο στατιστικός συντελεστής κάππα αποτελεί το μέτρο αξιολόγησης εξεταζόμενου μοντέλου κατηγοριοποίησης:

$$Kappa\ Statistic = \frac{P(A)-P(E)}{1-P(E)}$$

όπου:

P(A): η παρατηρούμενη σχετική συμφωνία μεταξύ των μοντέλων κατηγοριοποίησης

και

P(E): η πιθανότητα η συμφωνία αυτή να οφείλεται σε τυχαίο παράγοντα

3.3.2.3. Ευαισθησία και Εξειδικευτικότητα (Sensitivity and Specificity)

Η ευαισθησία ή ανάκληση (sensitivity ή recall) εκτιμά την ικανότητα του ταξινομητή να κατηγοριοποιήσει σωστά τα θετικά στιγμιότυπα της βάσης εξέτασης, ενώ η εξειδικευτικότητα (specificity) αξιολογεί την απόδοση του μοντέλου για την κατηγοριοποίηση αρνητικών στιγμιοτύπων:

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

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

3.3.2.4. Ακρίβεια (Precision)

Το μέτρο της ακρίβειας (precision) εκφράζει το βαθμό πιστότητας της διαδικασίας κατηγοριοποίησης*:*

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

3.3.2.5. Μέτρο F (F-measure)

Το μέτρο F εκφράζει τον αρμονικό μέσο της ακρίβειας και της ανάκλησης συνδυάζοντας τα δύο αυτά μέτρα με τη σχέση:

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

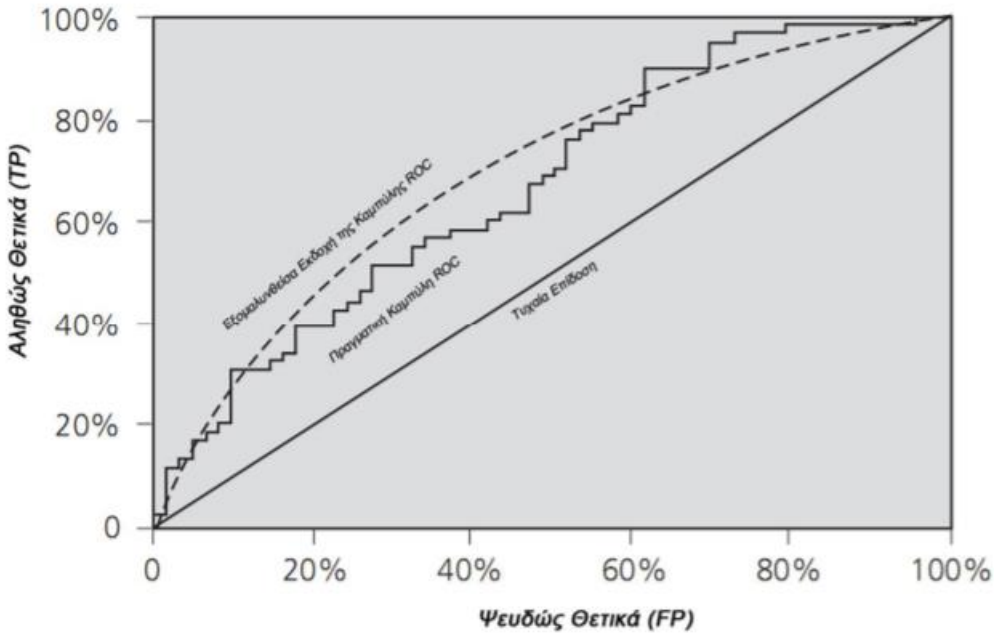
3.3.2.6. Δείκτης λάθος συναγερμού (False alarm rate)

Ο δείκτης αυτός εκφράζει την πιθανότητα λάθους στην ταξινόμηση των θετικών στιγμιότυπων και δίνεται από τη σχέση:

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

3.3.2.7. Καμπύλη Receiver Operating Characteristic (ROC Curve)

Η καμπύλη ROC είναι μια γραφική παράσταση που απεικονίζει τη διαγνωστική ικανότητα ενός δυαδικού συστήματος ταξινόμησης καθώς το όριο διάκρισής του ποικίλει. Η καμπύλη δημιουργείται από την απεικόνιση του TP (κατακόρυφος άξονας) προς το FP (οριζόντιος άξονας) σε διάφορες τιμές στιγμιοτύπων (Σχήμα 3.1). Η ιδανική περίπτωση θα ήταν μια καμπύλη ROC που θα ταυτίζεται με τον κατακόρυφο άξονα (TP).



Σχήμα 3.1: Παράδειγμα χαρακτηριστικής καμπύλης λειτουργίας δέκτη (ROC curve)

Ένα ακόμα κριτήριο επιλογής κατάλληλου ταξινομητή με βάση τις χαρακτηριστικές καμπύλες λειτουργίας είναι το εμβαδόν κάτω από την καμπύλη ROC (Area Under the ROC Curve – AUC). Στην ιδανική περίπτωση AUC=1 και στην περίπτωση τυχαίας κατηγοριοποίησης (καμπύλη τυχαίας επίδοσης του σχήματος 3.1) ισχύει ότι AUC=0.5.

3.4. Ανάλυση παραγόντων (factor analysis)

Σκοπός της ανάλυσης παραγόντων (Comrey 1978) είναι να συνοψίσει τις σχέσεις ανάμεσα σε ένα μεγάλο αριθμό ανεξάρτητων μεταβλητών με έναν περιεκτικό και ακριβή τρόπο, ώστε να βοηθήσει να γίνει αντιληπτή μία έννοια.

Η ανάλυση παραγόντων ακολουθεί τα εξής βήματα:

1. Επιλογή και μέτρηση μιας ομάδας μεταβλητών
2. Δημιουργία πίνακα ενδοσυναφειών (correlation matrix)
3. Επιλογή μεθόδου εξαγωγής παραγόντων
4. Επιλογή μεθόδου περιστροφής παραγόντων
5. Ερμηνεία των παραγόντων που προκύπτουν

Η ανάλυση παραγόντων επηρεάζεται σε σημαντικό βαθμό από την ποιότητα των δεδομένων που χρησιμοποιούνται:

- Οι μεταβλητές πρέπει να έχουν επαρκή συσχέτιση (correlation) μεταξύ τους (correlation > 0.2), αλλά όχι υπερβολική (correlation < 0.80) και να είναι ευθύγραμμες, να μην υπάρχουν δηλαδή ακραίες τιμές.
- Οι μεταβλητές πρέπει να έχουν μετρηθεί τουλάχιστον σε κλίμακα ίσων διαστημάτων
- Ο συνολικός αριθμός των μεταβλητών που αναλύονται πρέπει να είναι 3 με 5 φορές μεγαλύτερος από τους υποτιθέμενους παράγοντες
- Ο συνολικός αριθμός των στοιχείων πρέπει να είναι σημαντικός (τουλάχιστον > 300)
- Πρέπει να υπάρχει μια αναλογία ανάμεσα στον αριθμό των μεταβλητών και των στοιχείων που χρησιμοποιούνται (10:1, ή 5:1)

Για τον έλεγχο ποιότητας των δεδομένων χρησιμοποιείται ο **δείκτης Kaiser-Meyer-Olkin (KMO)** για την αξιολόγηση της επάρκειας του δείγματος (πρέπει KMO>0.60) και ο δείκτης **Bartlett's Test of Sphericity (p)** για την αξιολόγηση της συσχέτισης μεταξύ των μεταβλητών που επιτρέπει την ανάλυση παραγόντων (πρέπει p<0.05).

Η εξαγωγή των παραγόντων γίνεται είτε με τη μέθοδο ανάλυσης παραγόντων (Factor Analysis) είτε με τη μέθοδο ανάλυσης κύριων συνιστωσών (Principal Components Analysis).

Η ανάλυση κύριων συνιστωσών έχει στόχο τη μελέτη όλης της υπάρχουσας διακύμανσης (κοινή, μοναδική και σφάλμα) ώστε να εξαχθεί το μεγαλύτερο ποσοστό της διακύμανσης από τους λιγότερους δυνατούς συντελεστές. Η μέθοδος αυτή παράγει συνιστώσεις και αποτελέι την καλύτερη μέθοδο όταν θέλουμε να μειώσουμε τον αριθμό των μεταβλητών.

Αντίθετα στόχος της ανάλυσης παραγόντων είναι η μελέτη μόνο του ποσοστού της διακύμανσης που έχουν κοινό οι μεταβλητές που μελετώνται. Η μέθοδος αυτή παράγει παράγοντες και είναι κατάλληλη για την κατασκευή παραγόντων.

3.5. Σημαντικότητα ανεξάρτητων μεταβλητών (Μέθοδος Boruta)

Η μέθοδος Boruta (2010) προσπαθεί να εντοπίσει όλες τις σημαντικές μεταβλητές που μπορεί να έχει ένα σύνολο δεδομένων σε σχέση με μια εξαρτημένη μεταβλητή και ακολουθεί τα παρακάτω βήματα:

- Αντιγράφει το σύνολο δεδομένων και ανακατεύει τις τιμές σε κάθε στήλη. Αυτές οι τιμές ονομάζονται χαρακτηριστικά σκιάς.
- Εκπαιδεύει έναν ταξινομητή, όπως ένας ταξινομητής τυχαίων δασών, στο σύνολο δεδομένων. Με αυτόν τον τρόπο, προκύπτει η δυνατότητα μιας πρώτης αξιολόγησης, μέσω της μέσης μείωσης της ακρίβειας ή της μέσης μείωσης

λάθους, για καθένα από τα χαρακτηριστικά του συνόλου δεδομένων. Όσο υψηλότερη είναι η βαθμολογία, τόσο καλύτερη ή πιο σημαντική η αντίστοιχη μεταβλητή.

- Ο αλγόριθμος της μεθόδου ελέγχει για καθεμία από τις πραγματικές τιμές των μεταβλητών, εάν έχει μεγάλη σημασία. Δηλαδή, εάν μια μεταβλητή έχει υψηλότερη βαθμολογία από τη μέγιστη βαθμολογία των σκιαγραφικών χαρακτηριστικών της. Εάν αυτό συμβαίνει, καταγράφεται σε ένα διάνυσμα. Έπειτα εκτελείται ένα σύνολο τέτοιων επαναλήψεων.
- Σε κάθε επανάληψη, ο αλγόριθμος συγκρίνει τις βαθμολογίες των ανακατεμένων αντιγράφων των χαρακτηριστικών σκιας και των πραγματικών χαρακτηριστικών, για να δει αν η τελευταία είχε καλύτερη απόδοση από την πρώτη. Εάν συμβαίνει αυτό, ο αλγόριθμος θα επισημάνει τη μεταβλητή ως σημαντική.

Ουσιαστικά, ο αλγόριθμος προσπαθεί να επικυρώσει τη σημασία της μεταβλητής συγκρίνοντας με τυχαία αντίγραφα, γεγονός που αυξάνει την αξιοπιστία των αποτελεσμάτων.

4. ΣΥΛΛΟΓΗ ΚΑΙ ΕΠΕΞΕΡΓΑΣΙΑ ΣΤΟΙΧΕΙΩΝ

4.1. Εισαγωγή

Σύμφωνα με αυτά που έχουν αναφερθεί ως τώρα, στόχος της παρούσας διπλωματικής εργασίας είναι να εντοπίσει, με βάση τα οδηγικά χαρακτηριστικά, αν ο οδηγός βρίσκεται κατά τη διάρκεια ενός συμβάντος ή πριν από αυτό, σε υπεραστικές οδούς.

Σε αυτό το κεφάλαιο θα περιγραφεί η **συλλογή και ο τρόπος επεξεργασίας των στοιχείων** που χρησιμοποιήθηκαν για την εκπόνηση της διπλωματικής εργασίας, που έγιναν με στόχο να σχηματιστεί η τελική βάση δεδομένων για ανάλυση.

4.2. Συλλογή στοιχείων

Η βάση δεδομένων που χρησιμοποιήθηκε για τη συλλογή των απαραίτητων στοιχείων ήταν εκείνη που προέκυψε από πείραμα σε προσομοιωτή οδήγησης για ένα ελληνικό ερευνητικό πρόγραμμα.

Το πείραμα αυτό πραγματοποιήθηκε στον **προσομοιωτή οδήγησης** (Driving Simulator FPF) του Εργαστηρίου Μεταφορών και Συγκοινωνιακής Υποδομής της Σχολής Πολιτικών Μηχανικών του Εθνικού Μετσόβιου Πολυτεχνείου. Ο συγκεκριμένος προσομοιωτής έχει κατασκευαστεί από τη γερμανική εταιρεία FOERST ώστε να εξυπηρετεί ερευνητικούς σκοπούς. Η Εικόνα 4.1 παρουσιάζει τον προσομοιωτή που αποτελείται από τρεις οθόνες LCD40”, θέση οδήγησης και βάση υποστήριξης. Οι διαστάσεις σε πλήρη ανάπτυξη είναι 230 X 180 cm., ενώ το πλάτος βάσης 78cm.

Εικόνα 4.1: Προσομοιωτής οδήγησης

Οι Εικόνες 4.2 και 4.3 απεικονίζουν τη θέση οδήγησης του προσομοιωτή οδήγησης, ο οποίος διαθέτει ρυθμιζόμενο κάθισμα, τιμόνι διαμέτρου 27 cm, ποδόπληκτρα χειρισμού (συμπλέκτης, γκάζι, φρένο), πίνακα οργάνων οχήματος (ταχογράφος, στροφόμετρο) καθώς και δύο εξωτερικούς και έναν κεντρικό καθρέπτη που εμφανίζονται στις πλάγιες και την κεντρική οθόνη αντίστοιχα και απεικονίζουν σε πραγματικό χρόνο αντικείμενα και συμβάντα που συμβαίνουν πίσω από το «όχημα». Τα χειριστήρια που έχει στη διάθεσή του ο οδηγός είναι μοχλός 5 ταχυτήτων και όπισθεν, φλας, υαλοκαθαριστήρες, φώτα, κόρνα, χειρόφρενο και μίζα.

Εικόνες 4.2, 4.3: Θέση οδήγησης του προσομοιωτή οδήγησης

Ο προσομοιωτής οδήγησης καταγράφει μετρήσεις σε χρονικά διατήματα των 16-17ms, που σημαίνει ότι οι μετρήσεις είναι περίπου 60 το δευτερόλεπτο, και δίνει πληροφορίες, μεταξύ άλλων, για τις παρακάτω μεταβλητές που παρουσιάζονται στον Πίνακα 4.1.

Πίνακας 4.1: Μεταβλητές προσομοιωτή οδήγησης

Μεταβλητή	Επεξήγηση
Time	Πραγματικός χρόνος από την έναρξη της οδήγησης σε m.
x-pos	x-θέση του οχήματος σε m.
y-pos	y-θέση του οχήματος σε m.
z-pos	z-θέση του οχήματος σε m.
road	νούμερο οδού που βρίσκεται το όχημα [int].
richt	κατεύθυνση του οχήματος στην οδό σε [BOOL] (0/1).
rdist	διανυόμενη απόσταση οχήματος από την αρχή της οδήγησης σε m.

rspur	απόκλιση οχήματος από το μέσο της οδού σε m.
ralpha	κατεύθυνση του οχήματος σε σχέση με την κατεύθυνση της οδού σε μοίρες.
Dist	οδηγημένη πορεία από την έναρξη της πορείας σε m.
Speed	πραγματική ταχύτητα σε km/h.
Brk	θέση πεντάλ φρένου σε ποσοστό επί τοις εκατό.
Acc	θέση πεντάλ γκαζιού σε ποσοστό επί τοις εκατό.
Clutch	θέση συμπλέκτη σε ποσοστό επί τοις εκατό.
Gear	επιλεγμένη ταχύτητα (0 = αδράνεια, 6 = όπισθεν).
RPM	περιστροφή κινητήρα σε 1/min.
HWay	απόσταση από το προπορευόμενο όχημα σε m.
Dleft	απόσταση από την αριστερή οριογραμμή σε m.
DRight	απόσταση από τη δεξιά οριογραμμή σε m.
Wheel	θέση τιμονιού σε μοίρες.
THead	χρόνος μέχρι τη σύγκρουση με το προπορευόμενο όχημα σε sec.
TTL	χρόνος προς τη διασταύρωση έως ότου ξεπεραστεί η γραμμή του ορίου σε sec.
TTC	χρόνος μέχρι τη σύγκρουση (όλα τα εμπόδια) σε sec.
AccLat	πλευρική επιτάχυνση σε m/s^2.
AccLon	διαμήκης επιτάχυνση σε m/s^2.
EvVis	συμβάν-ορατό-σημαία/συμβάν-ένδειξη, 0 = όχι συμβάν, 1 = συμβάν.
EvDist	συμβάν-απόσταση σε m.
Err1No	αριθμός του πιο σημαντικού οδηγικού λάθους από το τελευταίο σύνολο δεδομένων.
Err1Val	στιγμή του λάθους, το περιεχόμενο διαφέρει ανάλογα με τον τύπο του λάθους.
Err2No	αριθμός του επόμενου οδηγικού λάθους (πιθανόν κενό).
Err2Val	πρόσθετη στιγμή του λάθους 2.
Err3No	αριθμός περαιτέρω λάθους οδήγησης (ίσως κενό)
Err3Val	πρόσθετη στιγμή του λάθους 3.

Για αυτές τις μεταβλητές έγινε ένας αρχικός διαχωρισμός για τις μεταβλητές που είναι χρήσιμες και πρόκειται να συμβάλλουν στην ανάλυση και τον προσδιορισμό του στόχου. Εκείνες που θεωρήθηκαν πιο χρήσιμες παρουσιάζονται στον Πίνακα 4.1 με έντονα γράμματα.

Εκτός από αυτές τις μεταβλητές μέσω ερωτηματολογίων προσδιορίστηκαν οι μεταβλητές που αφορούν τα χαρακτηριστικά κάθε οδηγού. Οι μεταβλητές αυτές παρουσιάζονται στον Πίνακα 4.2 παρακάτω.

Πίνακας 4.2: Μεταβλητές χαρακτηριστικών οδηγού

Variables	Μεταβλητές	Πιθανές τιμές ή μονάδες
PersonID	κωδικός οδηγού	D0i
Age	ηλικία	έτη ζωής (αριθμός)
AgeGroup	ηλικιακή ομάδα	νέος/ μεσήλικας/ηλικιωμένος
Gender	φύλο	γυναίκα/άνδρας
Education	εκπαίδευση	έτη εκπαίδευσης (αριθμός)

Driving Experience	οδηγική εμπειρία	έτη οδηγικής εμπειρίας (αριθμός)
Disease	ασθένεια	υγιής

Τέλος στον Πίκανα 4.3 αναφέρονται οι μεταβλητές που προέκυψαν από την εκτέλεση του πειράματος και τα σενάρια στα οποία κλήθηκαν να οδηγήσουν οι οδηγοί. Τα σενάρια αυτά αφορούσαν τον κυκλοφοριακό φόρτο (υψηλός, χαμηλός), την απόσπασης προσοχής (όχι απόσπαση, χρήση κινητού τηλεφώνου, συνομιλία με συνεπιβάτη) καθώς και ξαφνικά συμβάντα στη διάρκεια της οδήγησης.

Πίνακας 4.3: Μεταβλητές σεναρίων πειράματος

Variables	Μεταβλητές	Πιθανές τιμές
Trial	αριθμός διαδρομής	1, 2, ...
Traffic	κυκλοφοριακός φόρτος κατα x	υψηλός/χαμηλός
Distractor	απόσπαση	όχι/κινητό/συνομιλία
Event	συμβάν (ανάλογα με το είδος)	0, 4 ,22
State	συνθήκη	μηδενική ταχύτητα/όχι συμβάν/συμβάν

4.3. Βάση δεδομένων

Με βάση όλα τα παραπάνω δημιουργήθηκε η **αρχική βάση δεδομένων RuralControl** με 46,233,642 παρατηρήσεις συνολικά από 26 μεταβλητές, απόσπασμα του οποίου απεικονίζεται στον Πίνακα 4.4 και αποτελούταν από 39 υγιείς οδηγούς και οδήγηση σε επαρχιακή μόνο οδό.

Πίνακας 4.4: Απόσπασμα πίνακα RuralControl (14/1778217 σειρές και 10/26 στήλες)

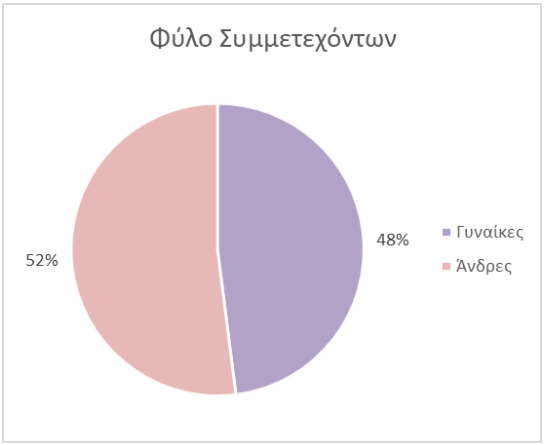
RuralControl										
	X1	PersonID	Trial	Time	Traffic_x	Distractor	Speed	AccLat	AccLon	Event
1	105788	D003	1	00:00.0	QL	NO	0	0	0	0
2	105790	D003	1	00:00.1	QL	NO	0	0	0	0
3	105792	D003	1	00:00.1	QL	NO	0	0	0	0
4	105794	D003	1	00:00.1	QL	NO	0	0	0	0
5	105796	D003	1	00:00.2	QL	NO	0	0	0	0
6	105798	D003	1	00:00.2	QL	NO	0	0	0	0
7	105800	D003	1	00:00.2	QL	NO	0	0	0	0
8	105802	D003	1	00:00.3	QL	NO	0	0	0	0
9	105804	D003	1	00:00.3	QL	NO	0	0	0	0
10	105806	D003	1	00:00.3	QL	NO	0	0	0	0
11	105808	D003	1	00:00.4	QL	NO	0	0	0	0
12	105810	D003	1	00:00.4	QL	NO	0	0	0	0
13	105812	D003	1	00:00.4	QL	NO	0	0	0	0
14	105814	D003	1	00:00.5	QL	NO	0	0	0	0

4.4. Επεξεργασία Στοιχείων

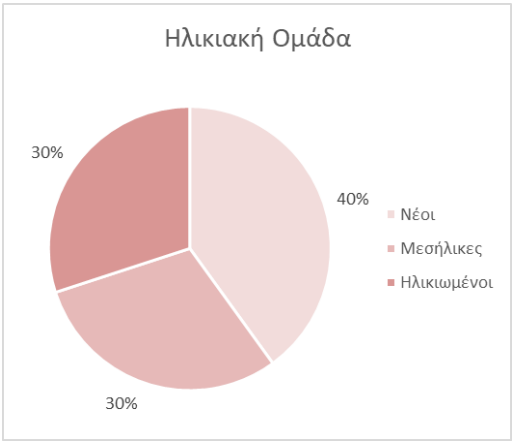
Στόχος της παρούσας διπλωματικής είναι ο εντοπισμός συμβάντων με βάση τα οδηγικά χαρακτηριστικά. Για το λόγο αυτό ήταν απαραίτητος ο **διαχωρισμός της βάσης δεδομένων** σε τρεις επιμέρους πίνακες δεδομένων οι οποίοι περιείχαν:

1. Στοιχεία κατά τη διάρκεια των συμβάντων.
2. Στοιχεία ένα λεπτό πριν από κάθε συμβάν.
3. Σύνολο των δύο παραπάνω περιπτώσεων.

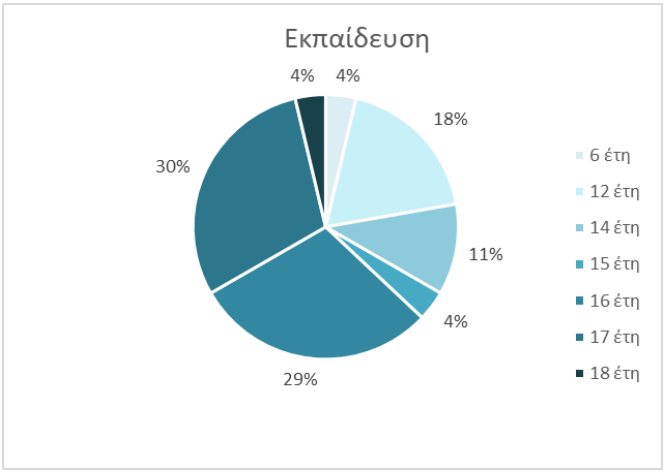
Παρατηρήθηκε στη στήλη Driving Experience της βάσης δεδομένων πως υπάρχει απώλεια κάποιων στοιχείων (ύπαρξη NA) σε κάποιους από τους οδηγούς. Για να είναι επεξεργάσιμα τα στοιχεία της στήλης αποφασίστηκε να διαγραφούν τα στοιχεία αυτά. Με την αφαίρεση αυτών των στοιχείων από τους 39 απέμειναν οι **27 οδηγοί**, των οποίων η κατανομή φύλου, ηλικιακής ομάδας και μορφωτικού επιπέδου φαίνονται στην Εικόνα 4.4.



(α)



(β)



(γ)

Εικόνα 4.4: Κατανομή (α) φύλου (β) ηλικιακής ομάδας και (γ) μορφωτικού επιπέδου συμμετεχόντων

Για το διαχωρισμό στους πίνακες δεδομένων που αναφέρθηκε παραπάνω, σχηματίστηκε ο κατάλληλος κώδικας (Παράρτημα). Το **κριτήριο για το διαχωρισμό** αυτό ήταν η τιμή της μεταβλητής **Event**, η οποία όπου έχει την τιμή 0 συμβολίζει πως δεν υπάρχει συμβάν, και όπου έχει τιμή διαφορετική του μηδενός, συμβολίζει πως υπάρχει συμβάν.

Για το σκοπό αυτό έπρεπε να προσδιοριστούν η χρονική στιγμή έναρξης και η χρονική στιγμή λήξης κάθε συμβάντος, για κάθε οδηγό. Έτσι δημιουργήθηκε ο **πίνακας index** (Πίνακας 4.5), ο οποίος αποτελείται από τρεις στήλες και γραμμές τόσες, όσες το σύνολο των συμβάντων. Η πρώτη στήλη περιέχει τον κωδικό κάθε οδηγού (PersonID), η δεύτερη την έναρξη του συμβάντος (begEvent Tme) σε h:m:s.ml και η τρίτη τη λήξη του (endEvent Time).

Πίνακας 4.5: Απόσπασμα πίνακα index (20/428 γραμμές)

index			
	PersonID	**begEvent-Time**	**endEvent-Time**
1	D003	00:01:57.787	00:02:12.622
2	D003	00:03:00.125	00:03:12.990
3	D003	00:06:01.355	00:06:14.189
4	D003	00:07:56.098	00:08:07.665
5	D003	00:10:55.128	00:11:06.562
6	D003	00:11:47.197	00:11:59.399
7	D003	00:14:50.961	00:15:04.864
8	D003	00:15:36.667	00:15:49.368
9	D006	00:01:12.849	00:01:24.085
10	D006	00:01:48.453	00:01:59.521
11	D006	00:03:43.595	00:03:55.295
12	D006	00:04:53.365	00:05:03.165
13	D006	00:07:22.327	00:07:31.827
14	D006	00:07:54.696	00:08:08.031
15	D006	00:10:08.254	00:10:19.456
16	D006	00:10:35.758	00:10:47.057
17	D010	00:01:33.390	00:01:46.125
18	D010	00:02:26.561	00:02:40.194
19	D010	00:05:22.156	00:05:34.891
20	D010	00:07:08.230	00:07:20.098

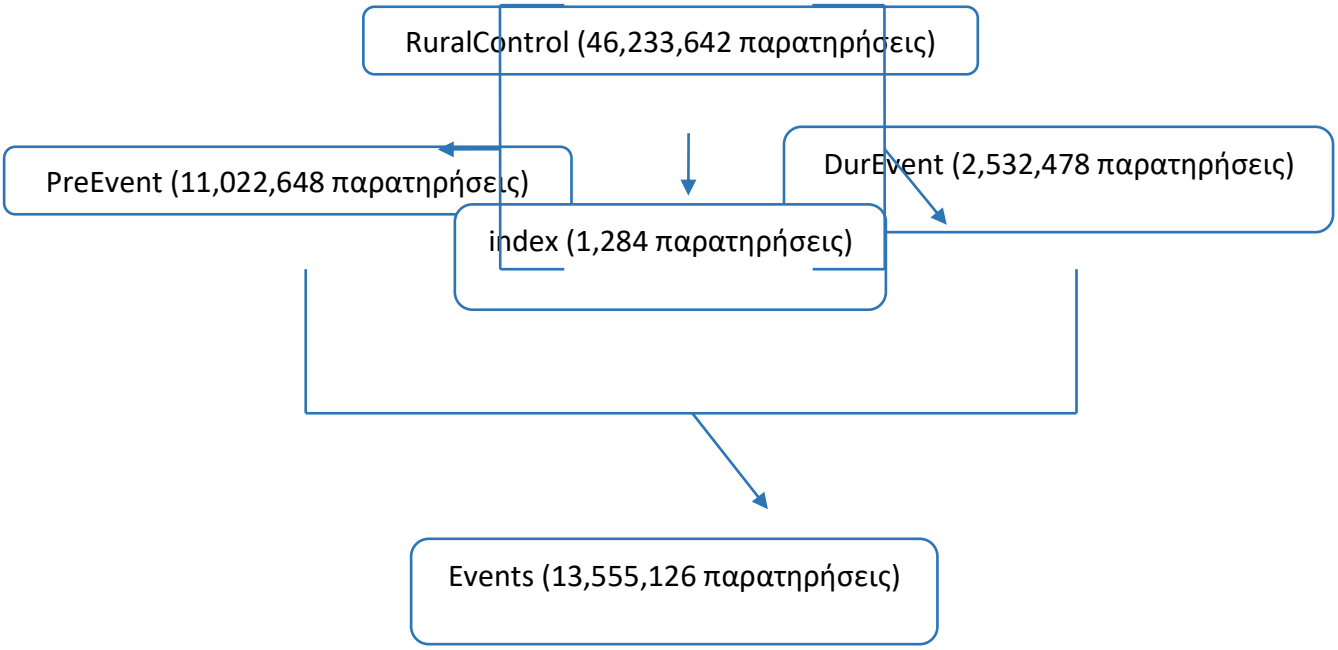
Σύμφωνα με τον πίνακα index δημιουργήθηκαν οι 3 επιμέρους πίνακες από την αρχική βάση δεδομένων. Καθένας από αυτούς τους πίνακες σχεδιάστηκε σύμφωνα με τις εξής συνθήκες:

1. Πίνακας PreEvent: αποτελείται από όλες τις στήλες του RuralControl και από τις γραμμές που αντιστοιχούν στις χρονικές στιγμές μεταξύ του χρόνου έναρξης κάθε συμβάντος μείον 60 δευτερόλεπτα (begEventTime-60) και της έναρξης του αντίστοιχου συμβάντος(begEventTime), για κάθε ένα από αυτά.
2. Πίνακας DurEvent: αποτελείται από όλες τις στήλες του RuralControl και από τις γραμμές που αντιστοιχούν στις χρονικές στιγμές μεταξύ του χρόνου έναρξης

(begEventTime) και της χρονικής στιγμής λήξης (endEventTime), για κάθε συμβάν.

3. Πίνακας Events: αποτελείται από όλες τις στήλες του RuralControl και από τις γραμμές που αντιστοιχούν στις χρονικές στιγμές μεταξύ του χρόνου έναρξης κάθε συμβάντος μείον 60 δευτερόλεπτα (begEventTime-60) και της χρονικής στιγμής λήξης του αντίστοιχου συμβάντος(begEventTime), για κάθε ένα από αυτά. Αποτελεί ουσιαστικά το άθροισμα των δύο παραπάνω πινάκων.

Η διαδικασία αυτή περιγράφεται συνοπτικά στο Γράφημα 4.1 όπου εμφανίζονται οι πίνακες με τον αντίστοιχο αριθμό γραμμών σε καθέναν από αυτούς.



Γράφημα 4.1: Δημιουργία πινάκων DurEvent, PreEvent και Events από RuralControl

Στον πίνακα Events κάποιες γραμμές είχαν τις ίδιες τιμές, και αυτό συνέβαινε γιατί το 1 λεπτό πριν από κάποιο συμβάν, στην αρχή του, συνέπεφτε χρονικά με τη διάρκεια του προηγούμενου. Για το λόγο αυτό οι διπλές γραμμές διαγράφηκαν. Στον Πίνακα 4.6 παρουσιάζεται απόσπασμα του πίνακα Events. Αντίστοιχη μόρφή έχουν και οι πίνακες PreEvent και DurEvent.

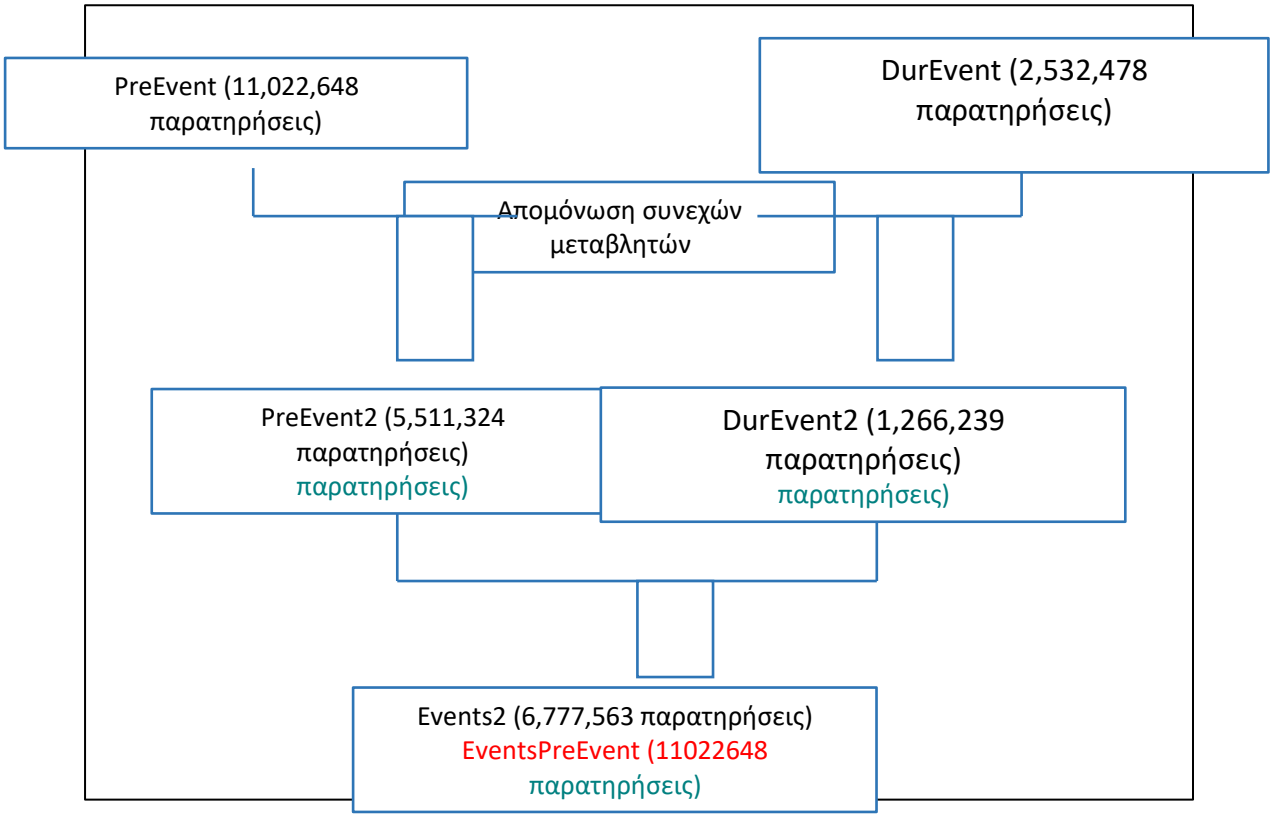
Πίνακας 4.6: Απόσπασμα πίνακα Events (10/521,351 σειρές και 10/26 στήλες)

Events										
	X1	PersonID	Trial	Time	Traffic_x	Distractor	Speed	AccLat	AccLon	Event
1	109254	D003	1	00:00:57.799	QL	NO	41.1	0.019	1.004	0
2	109256	D003	1	00:00:57.833	QL	NO	41.2	0.014	0.728	0
3	109258	D003	1	00:00:57.866	QL	NO	41.2	0.010	0.509	0
4	109260	D003	1	00:00:57.899	QL	NO	41.3	0.008	0.423	0
5	109262	D003	1	00:00:57.933	QL	NO	41.3	0.009	0.504	0

6	109264	D003	1	00:00:57.966	QL	NO	41.4	0.012	0.636	0
7	109266	D003	1	00:00:57.999	QL	NO	41.5	0.012	0.646	0
8	109268	D003	1	00:00:58.033	QL	NO	41.6	0.011	0.618	0
9	109270	D003	1	00:00:58.066	QL	NO	41.6	0.005	0.293	0
10	109272	D003	1	00:00:58.099	QL	NO	41.6	0.000	-0.024	0

4.5. Περιγραφική στατιστική

Στη συνέχεια **απομονώθηκαν οι συνεχείς μεταβλητές** των πινάκων PreEvent και DurEvent (**Speed, AccLat, AccLon, HWay, THead, TTC, DLeft, DRight, rdist, rspur, Wheel, Age, Driving Experience**) σε δύο νέους πίνακες PreEvent2 και DurEvent2 για να πραγματοποιηθεί η περιγραφική στατιστική που παρουσιάζεται παρακάτω. Το άθροισμα αυτών των 2 πινάκων είναι ο Events2. Η διαδικασία αυτή παρουσιάζεται διάγραμμα ροής του Γραφήματος 4.2.



Γράφημα 4.2: Δημιουργία PreEvent2, DurEvent2, Events2

Αποσπάσματα των πινάκων που προέκυψαν εμφανίζονται στους Πίνακες 4.7 και 4.8.

Πίνακας 4.7: Απόσπασμα πίνακα PreEvent2 (10/423,948 γραμμές)

PreEvent2													
	Speed	AccLat	AccLon	Hway	THead	TTC	Dleft	Dright	rdist	rspur	Wheel	Age	Driving Experience
1	41.1	0.019	1.004	348.4	30.5	9999.9	0.99	0.51	559.38	1.81	14	72	46
2	41.2	0.014	0.728	348.7	30.5	9999.9	0.99	0.51	559.77	1.81	14	72	46
3	41.2	0.010	0.509	349.0	30.5	9999.9	0.99	0.51	560.14	1.81	16	72	46
4	41.3	0.008	0.423	349.2	30.5	9999.9	0.99	0.51	560.52	1.81	17	72	46
5	41.3	0.009	0.504	349.5	30.4	9999.9	0.99	0.51	560.91	1.81	17	72	46
6	41.4	0.012	0.636	349.8	30.4	9999.9	0.99	0.51	561.28	1.81	17	72	46
7	41.5	0.012	0.646	350.0	30.4	9999.9	0.99	0.51	561.66	1.81	17	72	46
8	41.6	0.011	0.618	350.3	30.3	9999.9	0.99	0.51	562.05	1.81	17	72	46
9	41.6	0.005	0.293	350.6	30.3	9999.9	0.99	0.51	562.43	1.81	17	72	46
10	41.6	0.000	-0.024	350.8	30.4	9999.9	0.99	0.51	562.81	1.81	17	72	46

Πίνακας 4.8: Απόσπασμα πίνακα DurEvent2 (10/97,403 γραμμές)

PreEvent2													
	Speed	AccLat	AccLon	Hway	THead	TTC	Dleft	Dright	rdist	rspur	Wheel	Age	Driving Experience
1	27.6	-0.001	-0.576	923.0	120.3	9999.9	1.08	0.47	1150.37	1.88	14	72	46
2	27.5	-0.001	-0.613	923.3	120.7	9999.9	1.08	0.47	1150.63	1.88	14	72	46
3	27.5	-0.001	-0.564	923.7	121.0	9999.9	1.08	0.47	1150.88	1.88	16	72	46
4	27.4	-0.001	-0.477	924.1	121.3	9999.9	1.08	0.47	1151.14	1.88	17	72	46
5	27.3	-0.001	-0.378	924.5	121.6	9999.9	1.08	0.47	1151.39	1.88	17	72	46
6	27.3	0.000	-0.271	924.9	121.8	9999.9	1.08	0.47	1151.65	1.88	17	72	46
7	27.3	0.000	-0.201	925.3	121.9	9999.9	1.08	0.47	1151.90	1.88	17	72	46
8	27.3	0.000	-0.156	925.7	122.1	9999.9	1.08	0.47	1152.15	1.88	17	72	46
9	27.3	0.000	-0.083	926.1	122.2	9999.9	1.08	0.47	1152.40	1.88	17	72	46
10	27.3	0.000	-0.022	926.4	122.2	9999.9	1.08	0.47	1152.65	1.88	17	72	46

Στους παραπάνω πίνακες πραγματοποιήθηκε περιγραφική στατιστική των μεταβλητών, με την εκτέλεση κώδικα (Παράρτημα) και προέκυψαν οι Πίνακες 4.9 και 4.10 αντίστοιχα.

Πίνακας 4.9: Περιγραφική στατιστική μεταβλητών του PreEvent2

Variable	Μεταβλητή	Ελάχιστη Τιμή	Μέση Τιμή	Μέγιστη τιμή	Διακύμανση	Τυπική Απόκλιση
Speed (km/h)	Ταχύτητα	0.00	45.68	103.60	351.013	18.735
AccLat (m/s^2)	Πλευρική επιτάχυνση	-0.724	0.00	0.377	0.00	0.017
AccLon (m/s^2)	Διαμήκης επιτάχυνση	-9523.00	13.433	6399.00	37251.45	193.006
HWay (m)	Απόσταση από προπ. όχημα	1.20	956.10	1241.90*	6368759.00	2523.64

THead (s)	Χρ. μέχρι σύγκρουση με όχημα	0.20	1304.36	9993.7*	11093839.00	3330.742
TTC (s)	Χρ. μέχρι σύγκρουση	0.40	6436.66	29.90*	22907929.00	4786.223
DLeft (m)	Απόσταση από αριστερή οριογρ.	-1.28	0.713	2.17	0.077	0.277
DRight (m)	Απόσταση από δεξιά οριογρ.	-0.66	0.80	2.76	0.078	0.279
rdist (m)	Διανυόμενη απόσταση	5.00	1142.392	2510.90	262085.20	511.943
rspur (m)	Απόκλιση οχήματος από το μέσο της οδού	-0.44	1.532	2.98	0.077	0.278
Wheel (degrees)	Γωνία τιμονιού	-156.00	-4.126	109.00	232.217	15.239
Age (years)	Ηλικία	22	41.322	78	270.506	16.447
Driving Experience (years)	Οδηγική εμπειρία	3	19.898	46	182.114	13.495

Πίνακας 4.10: Περιγραφική στατιστική μεταβλητών του DurEvent2

Variable	Μεταβλητή	Ελάχιστη Τιμή	Μέση Τιμή	Μέγιστη τιμή	Διακύμανση	Τυπική Απόκλιση
Speed (km/h)	Ταχύτητα	0.00	29.28	104.80	620.617	24.912
AccLat (m/s^2)	Πλευρική επιτάχυνση	-11.328	0.108	7048.00	619.669	24.893
AccLon (m/s^2)	Διαμήκης επιτάχυνση	-487799.00	-49.50	3561.00	3597746.00	1896.773
HWay (m)	Απόσταση από προπ. όχημα	0.00	2018.80	1135.60*	14169430.00	3764.230
THead (s)	Χρ. μέχρι σύγκρουση με όχημα	0.00	3389.50	9969.10*	21670860.00	4655.197
TTC (s)	Χρ. μέχρι σύγκρουση	0.00	5783.60	29.90*	24347950.00	4934.364
DLeft (m)	Απόσταση από	-1.37	0.672	2.07	0.094	0.307

	αριστερή οριογρ.					
DRight (m)	Απόσταση από δεξιά οριογρ.	-0.53	0.838	2.79	0.093	0.305
rdist (m)	Διανυόμενη απόσταση	850.20	1551.80	2047.70	138608.90	372.302
rspur (m)	Απόσταση οχήματος από τη μέση της οδού	-0.49	1.492	2.87	0.093	0.305
Wheel (degrees)	Γωνία τιμονιού	-125.00	-4.094	137.00	222.189	14.906
Age (years)	Ηλικία	22	41.41	78	272.435	16.506
Driving Experience (years)	Οδηγική εμπειρία	3	20.01	46	184.354	13.578

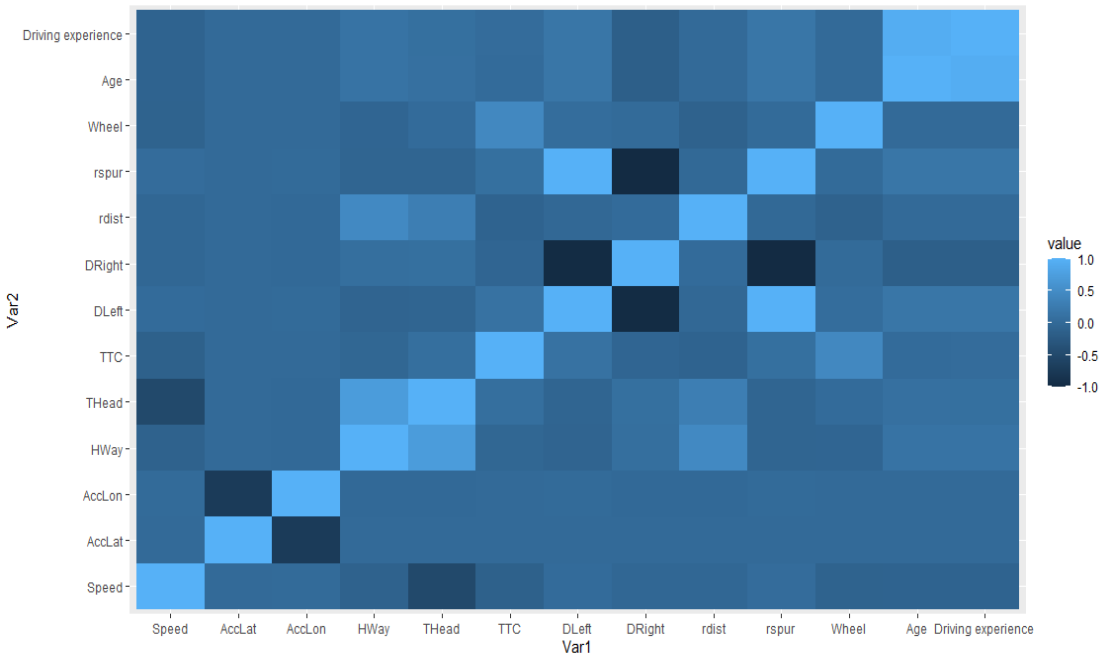
*οι τιμές με αστερίσκο είναι οι δεύτερες μεγαλύτερες μετά το 9999.99.

Από τους 2 τελευταίους πίνακες παρατηρείται πως αν συγκριθούν οι τιμές των μεταβλητών του Πίκακα 4.5 (DurEvent2) με τον 4.5 (PreEvent2) προκύπτουν οι εξής **βασικές παρατηρήσεις**:

- Η μέση τιμή της ταχύτητας (Speed) μειώθηκε.
- Η διακύμανση και η τυπική απόκλιση αυξήθηκαν για την πλευρική επιτάχυνση (AccLat).
- Η μέση τιμή της διαμήκους επιτάχυνσης (AccLon) μειώθηκε σημαντικά (μεγάλη επιβράδυνση), ενώ διακύμανση και τυπική απόκλιση αυξήθηκαν.
- Η μέση τιμή της απόστασης από το προπορευόμενο όχημα (HWay) αυξήθηκε, όπως και της χρονικής απόστασης από το προπορευόμενο όχημα (THead).

4.6. Συσχέτιση ανεξάρτητων μεταβλητών

Τέλος στον συνολικό πίνακα Events2 προσδιορίστηκε η συσχέτιση (correlation) μεταξύ των ανεξάρτητων μεταβλητών, ώστε να βρεθούν οι τελικές μεταβλητές, που δε θα έχουν πολύ μεγάλη συσχέτιση μεταξύ τους (<0.8) και που χρησιμοποιήθηκαν στη συνέχεια στην ανάπτυξη των μοντέλων στατιστικής ανάλυσης. Τα αποτελέσματα της συσχέτισης των μεταβλητών περιγράφονται από το χάρτη θερμότητας της Εικόνας 4.5.



Εικόνα 4.5: Χάρτης θερμότητας συσχέτισης ανεξάρτητων μεταβλητών

Από την Εικόνα 4.2 προκύπτει πως υψηλή συσχέτιση έχουν οι εξής μεταβλητές μεταξύ τους:

- Η πλευρική επιτάχυνση (AccLat) με τη διαμήκη επιτάχυνση (AccLon).
- Η απόσταση από το προπορευόμενο όχημα (HWay) με τη χρονική απόσταση από το προπορευόμενο όχημα (THead).
- Η διανυόμενη απόσταση (rdist) με την απόσταση από τη δεξιά (DRight) και από την αριστερά (DLeft) οριογραμμή.
- Η οδηγική εμπειρία (Driving Experience) με την ηλικία του οδηγού (Age).

Λόγω των παραπάνω οι μεταβλητές που χρησιμοποιήθηκαν στα μοντέλα παρουσιάζονται στον Πίνακα 4.6 και αυτές που αφαιρέθηκαν στον Πίνακα 4.7.

Πίνακας 4.11

Τελικές Μεταβλητές
Speed (km/h)
AccLon (m/s^2)
THead (s)
rdist (m)
rspur (m)
Wheel (degrees)
Driving Experience (years)

Η μεταβλητή TTC αφαιρέθηκε, καθώς έχει περίπου την ίδια χρήση με τη μεταβλητή THead.

Πίνακας 4.12

Μεταβλητές που αφαιρέθηκαν
AccLat (m/s^2)
HWay (m)
TTC (s)
DLeft (m)
DRight (m)
Age (years)

4.7. Πίνακες δεδομένων για τα μοντέλα διωνυμικής λογιστικής παλινδρόμησης και τυχαίων δασών

Η τελική βάση δεδομένων, η οποία χρησιμοποιήθηκε στα μοντέλα στατιστικής ανάλυσης, ονομάστηκε **DATA** με σύνολο 4,170,808 παρατηρήσεις (Πίνακας 4.13) και προέκυψε απομονώνοντας από τον πίνακα **Events** τις στήλες:

- Event
- Speed
- AccLon
- THead
- Rdist
- Rspur
- Wheel
- Driving Experience

Πίνακας 4.13: Απόσπασμα πίνακα DATA (20/521,351 γραμμές)

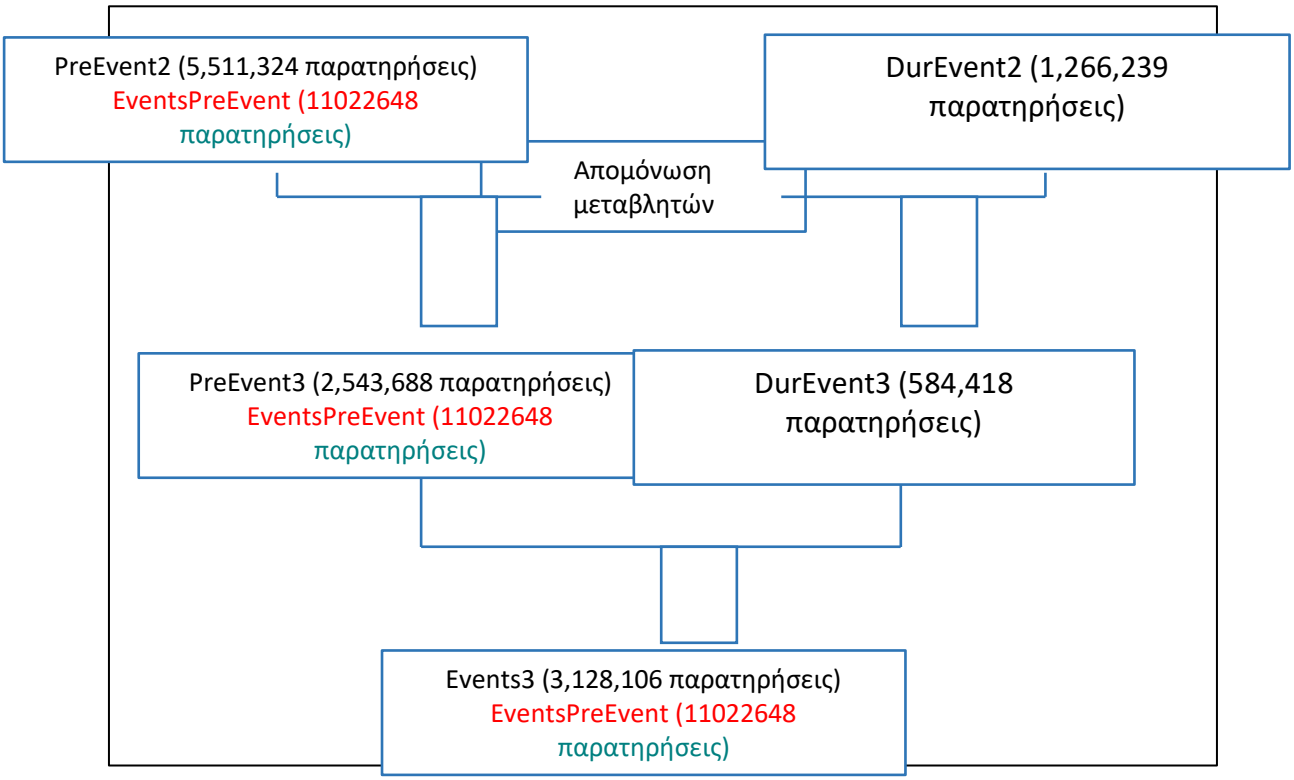
DATA								
	Event	**Speed**	**AccLon**	**Thead**	**rdist**	**rspur**	**Wheel**	**Driving Experience**
1	0	41.1	1.004	30.5	559.38	1.81	14	46
2	0	41.2	0.728	30.5	559.77	1.81	14	46
3	0	41.2	0.509	30.5	560.14	1.81	16	46
4	0	41.3	0.423	30.5	560.52	1.81	17	46
5	0	41.3	0.504	30.4	560.91	1.81	17	46
6	0	41.4	0.636	30.4	561.28	1.81	17	46
7	0	41.5	0.646	30.4	561.66	1.81	17	46
8	0	41.6	0.618	30.3	562.05	1.81	17	46
9	0	41.6	0.293	30.3	562.43	1.81	17	46
10	0	41.6	-0.024	30.4	562.81	1.81	17	46

Η στήλη Event λαμβάνει τιμές ανάλογα με το αν υπάρχει συμβάν (τιμή διάφορη του 0) ή όχι (τιμή 0). Ανάλογα με το είδος του συμβάντος, όταν υπάρχει, λαμβάνει και αντίστοιχες τιμές, για παράδειγμα ο αριθμός 22 υποδηλώνει ελάφι. Επειδή στην παρούσα διπλωματική εργασία μας ενδιαφέρει μόνο η ύπαρξη ή όχι συμβάντος και όχι το είδος του, οι τιμές του Event που ήταν διάφορες του 0 αντικαταστάθηκαν με την τιμή 1, ώστε το 0 να υποδηλώνει μη ύπαρξη και το 1 ύπαρξη συμβάντος.

Για το μοντέλο διωνυμικής λογιστικής παλινδρόμησης, για την παραλλαγή Α δημιουργήθηκε ο πίνακας DATA1 όπου χρησιμοποιήθηκαν όλες οι παρατηρήσεις των μεταβλητών Event (εξαρτημένη), Speed, AccLon, THead, rdist, rspur, Driving Experience από τον πίνακα DATA, εκτός από αυτές της μεταβλητής Wheel, η οποία έπειτα από δοκιμή κρίθηκε ως στατιστικά μη σημαντική.

4.8. Πίνακες δεδομένων για τη μέθοδο παραγοντικής ανάλυσης

Στην παρούσα διπλωματική πραγματοποιήθηκε και η μέθοδος ανάλυσης παραγόντων για τις ανεξάρτητες μεταβλητές της βάσης δεδομένων που χρησιμοποιήθηκε, με σκοπό την εξέταση ύπαρξης κοινών παραγόντων ανάμεσά τους. Η μέθοδος αυτή εφαρμόστηκε σε τρεις πίνακες, PreEvent3, DurEvent3 και Events3 οι οποίοι αποτελούνταν από τις μεταβλητές **Speed, AccLon, THead, rdist, rspur** και **Wheel** των αντίστοιχων πινάκων PreEvent2, DurEvent2 και Events2. Η μεταβλητή Driving Experience αφαιρέθηκε. Τα βήματα αυτού του διαχωρισμού περιγράφονται στο Γράφημα 4.3, και αποσπάσματα των πινάκων που χρησιμοποιήθηκαν γι' αυτή τη μέθοδο παρουσιάζονται στους Πίνακες 4.14, 4.15 και 4.16.



Γράφημα 4.3: Δημιουργία PreEvent3, DurEvent3, Events3

Πίνακας 5.14: Απόσπασμα πίνακα PreEvent3 (10/423,948 γραμμές)

PreEvent3						
	Speed	**AccLon**	**Thead**	**rdist**	**rspur**	**Wheel**
1	41.1	1.004	30.5	559.38	1.81	14
2	41.2	0.728	30.5	559.77	1.81	14
3	41.2	0.509	30.5	560.14	1.81	16
4	41.3	0.423	30.5	560.52	1.81	17
5	41.3	0.504	30.4	560.91	1.81	17
6	41.4	0.636	30.4	561.28	1.81	17
7	41.5	0.646	30.4	561.66	1.81	17
8	41.6	0.618	30.3	562.05	1.81	17

9	41.6	0.293	30.3	562.43	1.81	17
10	41.6	-0.024	30.4	562.81	1.81	17

Πίνακας 5.15: Απόσπασμα πίνακα DurEvent3 (10/97,403 γραμμές)

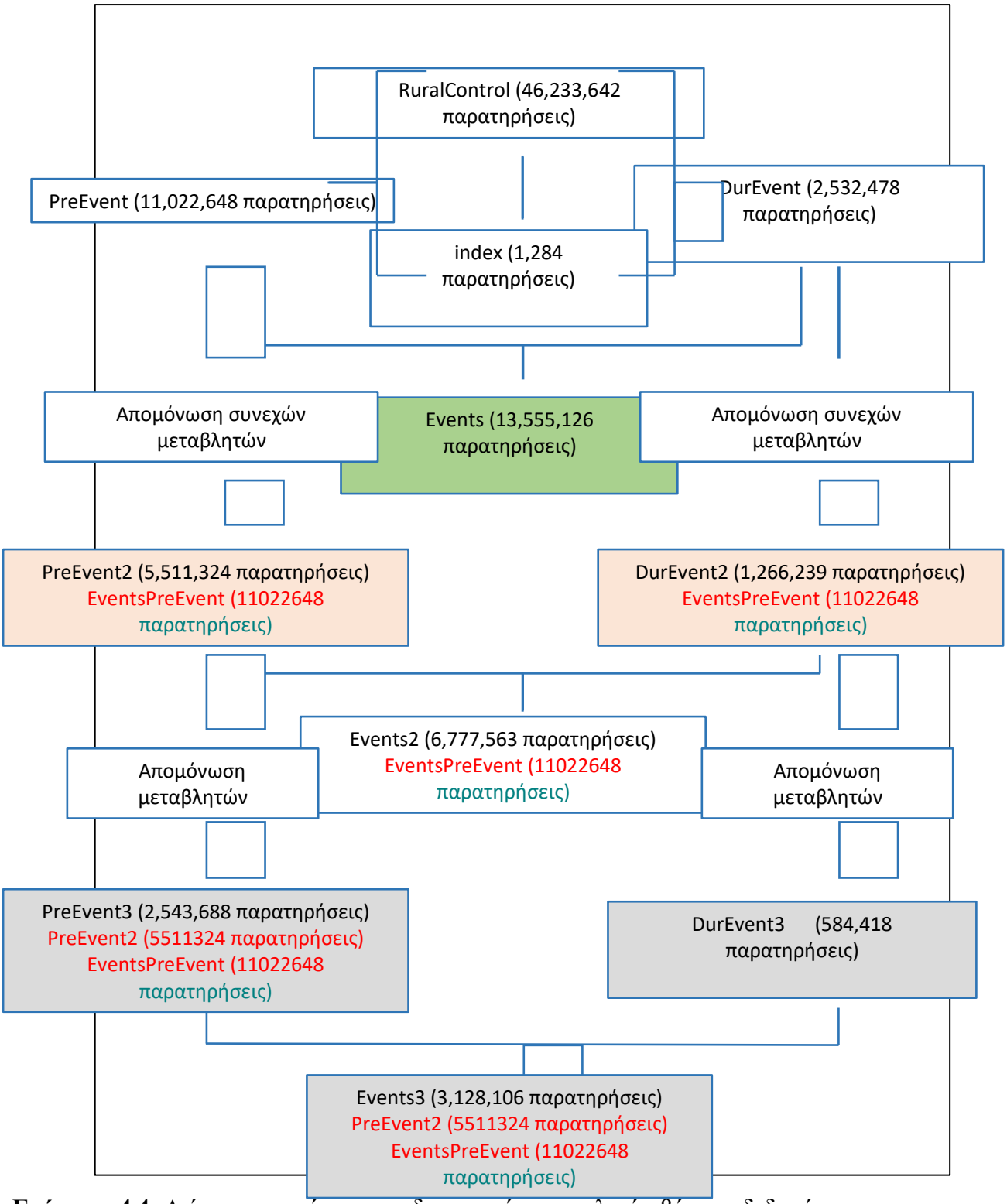
DurEvent3						
	Speed	**AccLon**	**Thead**	**rdist**	**rspur**	**Wheel**
1	27.6	-0.576	120.3	1150.37	1.88	0
2	27.5	-0.613	120.7	1150.63	1.88	0
3	27.5	-0.564	121.0	1150.88	1.88	0
4	27.4	-0.477	121.3	1151.14	1.88	0
5	27.4	-0.378	121.6	1151.39	1.88	0
6	27.3	-0.271	121.8	1151.65	1.88	0
7	27.3	-0.201	121.9	1151.90	1.88	0
8	27.3	-0.156	122.1	1152.15	1.88	0
9	27.3	-0.083	122.2	1152.40	1.88	0
10	27.3	-0.022	122.2	1152.65	1.88	0

Πίνακας 5.16: Απόσπασμα πίνακα Events3 (10/521,351 γραμμές)

Events3						
	Speed	**AccLon**	**Thead**	**rdist**	**rspur**	**Wheel**
1	41.1	1.004	30.5	559.38	1.81	14
2	41.2	0.728	30.5	559.77	1.81	14
3	41.2	0.509	30.5	560.14	1.81	16
4	41.3	0.423	30.5	560.52	1.81	17
5	41.3	0.504	30.4	560.91	1.81	17
6	41.4	0.636	30.4	561.28	1.81	17
7	41.5	0.646	30.4	561.66	1.81	17
8	41.6	0.618	30.3	562.05	1.81	17
9	41.6	0.293	30.3	562.43	1.81	17
10	41.6	-0.024	30.4	562.81	1.81	17

4.9. Διάγραμμα ροής για τη δημιουργία των τελικών βάσεων δεδομένων

Στο διάγραμμα ροής του Γραφήματος 4.4 παρουσιάζονται συγκεντρωτικά όλα τα βήματα για το διαχωρισμό της αρχικής βάσης δεδομένων RuralControl στις επιμέρους βάσεις δεδομένων που χρησιμοποιήθηκαν για τη στατιστική και την ανάλυση παραγόντων.



Γράφημα 4.4: Διάγραμμα ροής για την δημιουργία των τελικών βάσεων δεδομένων

Στο παραπάνω διάγραμμα ροής με πορτοκαλί χρώμα απεικονίζονται οι πίνακες που χρησιμοποιήθηκαν για την περιγραφική στατιστική, με πράσινο ο πίνακας των μοντέλων στατιστικής ανάλυσης και με γκρι χρώμα οι πίνακες για τους οποίους πραγματοποιήθηκε η ανάλυση παραγόντων.

5. ΕΦΑΡΜΟΓΗ ΜΕΘΟΔΟΛΟΓΙΑΣ – ΑΠΟΤΕΛΕΣΜΑΤΑ

5.1. Εισαγωγή

Στο κεφάλαιο αυτό παρουσιάζεται η **εφαρμογή της μεθοδολογίας των μοντέλων στατιστικής ανάλυσης και τα αποτελέσματα αυτών**. Στόχος της ανάλυσης αυτής ήταν ο εντοπισμός συμβάντος. Για το σκοπό αυτό, όπως έχει ήδη αναφερθεί χρησιμοποιήθηκε το μοντέλο διωνυμικής λογιστικής παλινδρόμησης και το μοντέλο τυχαίων δασών. Με τα μοντέλα αυτά πραγματοποιήθηκαν 2 παραλλαγές. Η **παραλλαγή Α** περιελάμβανε όλες τις στατιστικά σημαντικές για την ανάλυση, συγκεκριμένα την **ταχύτητα**, τη **διαμήκη επιτάχυνση**, τη **συνολική διανυόμενη απόσταση**, την **απόκλιση του οχήματος από το μέσο της οδού**, το **χρόνο μέχρι τη σύγκρουση** από το προπορευόμενο όχημα και την **οδηγική εμπειρία**, ενώ η **παραλλαγή Β** περιελάμβανε μόνο 4 ανεξάρτητες μεταβλητές, την **ταχύτητα,** τη **διαμήκη επιτάχυνση**, τη **συνολική διανυόμενη απόσταση** και την **οδηγική εμπειρία**. Έγινε επίσης ανάλυση παραγόντων για την ομαδοποίηση των μεταβλητών.

Για την επιλογή των μεταβλητών που χρησιμοποιήθηκαν στην παραλλαγή Β έγινε αρχικά εκτίμηση της σημαντικότητας των ανεξάρτητων μεταβλητών (feature selection) με τη μέθοδο Boruta και προέκυψε η κατάταξή τους, από την περισσότερο στη λιγότερο σημαντική όπως φαίνεται στον Πίνακα 5.1.

Πίνακας 5.1: Σημαντικότητα ανεξάρτητων μεταβλητών

Μεταβλητή	Μέση Σημαντικότητα
Συνολική διανυόμενη απόσταση: **rdist (m)**	289.68
Ταχύτητα: **Speed (km/h)**	216.67
Διαμήκης επιτάχυνση: **AccLon (m/s^2)**	143.54
Οδηγική εμπειρία**: Driving Experience (years)**	130.75
Απόκλιση οχήματος από το μέσο της οδού**: rspur (m)**	111.56
Χρόνος μέχρι τη σύγκρουση: **THead (s)**	100.7

Με κριτήριο αυτόν τον πίνακα έγιναν διάφορες δοκιμές και προέκυψε μια που είχε μικρό αριθμό μεταβλητών, αλλά έδινε και ικανοποιητικά αποτελέσματα. Για την **παραλλαγή Β** τελικά χρησιμοποιήθηκαν 4 μεταβλητές και συγκεκριμένα οι **Speed, AccLon, rdist και Driving Experience.**

Παρακάτω περιγράφονται **αναλυτικά τα βήματα** που ακολουθήθηκαν στο περιβάλλον του RStudio, τα αποτελέσματα που προέκυψαν και οι στατιστικοί έλεγχοι που έγιναν για την αποδοχή ή όχι των μοντέλων.

Για τα μοντέλα στατιστικής ανάλυσης χρησιμοποιήθηκαν οι παρακάτω 8 μεταβλητές:

1. Event, συμβάν που λαμβάνει τιμές 0 ή 1.

2. Speed, ταχύτητα σε χλμ/ώρα (km/h).
3. AccLon, διαμήκης επιτάχυνση σε μέτρα/δευτερόλεπτο^2 (m/s^2).
4. THead, χρόνος μέχρι τη σύγκρουση σε δευτερόλεπτα (s).
5. rdist, διανυόμενη απόσταση σε μέτρα (m).
6. rspur, απόκλιση του οχήματος από το μέσο της οδού σε μέτρα (m).
7. Wheel, γωνία τιμονιού σε μοίρες (degrees).
8. Driving Experience, εμπειρία οδήγησης σε χρόνια (years).

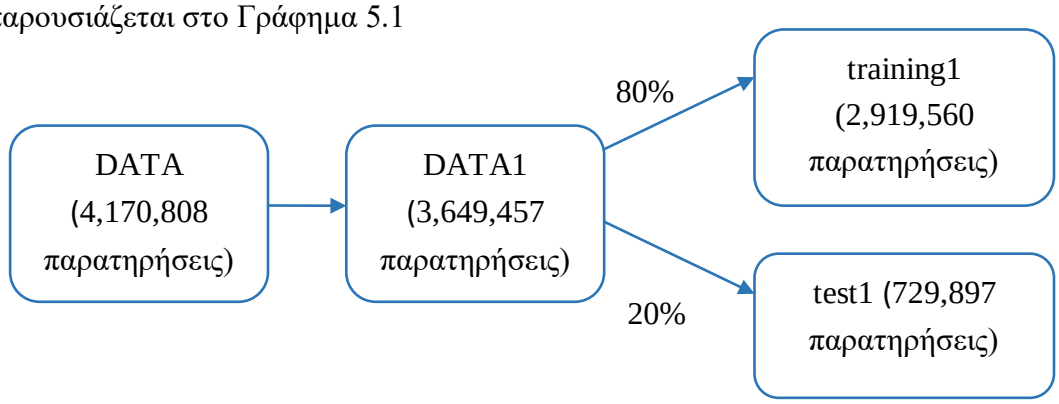
Η μεταβλητή Event αποτελεί την **εξαρτημένη μεταβλητή** και είναι διακριτή (factor), ενώ οι υπόλοιπες αποτελούν τις **ανεξάρτητες μεταβλητές** και είναι συνεχείς (numeric).

5.2. Ανάπτυξη μοντέλου διωνυμικής λογιστικής παλινδρόμησης

5.2.1. Παραλλαγή Α

Για το μοντέλο διωνυμικής λογιστικής παλινδρόμησης, για την παραλλαγή Α δημιουργήθηκε ο πίνακας **DATA1** όπου χρησιμοποιήθηκαν όλες οι παρατηρήσεις των μεταβλητών **Event (εξαρτημένη), Speed, AccLon, THead, rdist, rspur, Driving Experience** από τον πίνακα DATA, εκτός από αυτές της μεταβλητής Wheel, η οποία έπειτα από δοκιμή κρίθηκε ως στατιστικά μη σημαντική. Ολόκληρος ο κώδικας που χρησιμοποιήθηκε παρουσιάζεται στο Παράρτημα.

Αρχικά η μεταβλητή «Event» μετατράπηκε σε διακριτή και πραγματοποιήθηκε διαχωρισμός του πίνακα σε 2 επιμέρους, έναν για εκπαίδευση (training1), που αποτελούσε το 80% του αρχικού πίνακα (2,919,560 παρατηρήσεις), και έναν για έλεγχο (test1) με το υπόλοιπο 20% (729,897 παρατηρήσεις). Η παραπάνω διαδικασία παρουσιάζεται στο Γράφημα 5.1



Γράφημα 5.1: Πίνακες παραλλαγής Α

Αποσπάσματα των επιμέρους πινάκων που χρησιμοποιήθηκαν εμφανίζονται στους Πίνακες 5.2 και 5.3 αντίστοιχα.

Πίνακας 5.2: Απόσπασμα πίνακα εκπαίδευσης training1 (10/417,080 γραμμές)

training1							
	Event	Speed	AccLon	Thead	rdist	rspur	Driving Experience
1	0	41.1	1.004	30.5	559.38	1.81	46
2	0	41.2	0.728	30.5	559.77	1.81	46
3	0	41.2	0.509	30.5	560.14	1.81	46
4	0	41.3	0.423	30.5	560.52	1.81	46
5	0	41.6	0.618	30.3	562.05	1.81	46
6	0	41.6	0.293	30.3	562.43	1.81	46
7	0	41.6	-0.024	30.4	562.81	1.81	46
8	0	41.6	-0.216	30.4	563.20	1.81	46
9	0	41.5	-0.253	30.4	563.57	1.81	46
10	0	41.5	-0.237	30.5	563.96	1.81	46

Πίνακας 5.3: Απόσπασμα πίνακα εκπαίδευσης test1 (10/104,271 γραμμές)

test1							
	Event	Speed	AccLon	Thead	rdist	rspur	Driving Experience
1	0	41.3	0.504	30.4	560.91	1.81	46
2	0	41.4	0.636	30.4	561.28	1.81	46
3	0	41.5	0.646	30.4	561.66	1.81	46
4	0	41.5	-0.200	30.5	564.34	1.81	46
5	0	41.4	-0.371	30.6	564.72	1.81	46
6	0	40.7	-1.106	31.3	567.37	1.81	46
7	0	40.5	-0.920	31.5	568.11	1.82	46
8	0	40.1	-0.553	31.9	569.59	1.82	46
9	0	40.0	-0.474	32.0	570.32	1.83	46
10	0	40.0	-0.450	32.1	570.69	1.83	46

5.2.1.1. Εκπαίδευση μοντέλου

Η εκπαίδευση του μοντέλου πραγματοποιήθηκε στον πίνακα εκπαίδευσης (**training1**) που αναφέρθηκε παραπάνω.
Βρέθηκαν οι συντελεστές των μεταβλητών με τους οποίους διαμορφώθηκε η συνάρτηση χρησιμότητας για την ύπαρξη συμβάντος και έγινε ο έλεγχος ποιότητάς τους. Τα αποτελέσματα που προέκυψαν παρουσιάζονται στον παρακάτω Πίνακα 5.4.

Πίνακας 5.4: Έλεγχος συντελεστών

Μεταβλητή	Estimate	Std.Error	z value	Pr(>|z|)
Σταθερός όρος	-1.45E+00	2.89E-02	-49.921	< 2e-16
Speed	-3.69E-02	2.34E-04	-157.593	< 2e-16
AccLon	-3.27E-04	1.52E-05	-21.587	< 2e-16
Thead	-2.45E-05	1.22E-06	-20.077	< 2e-16

rdist	1.57E-03	9.19E-06	170.476	< 2e-16
rspur	-3.96E-01	1.61E-02	-24.615	< 2e-16
DrExp	-2.39E-03	3.29E-04	-7.256	3.99E-13

Η συνάρτηση χρησιμότητας που προέκυψε είναι η εξής:

Event = -1.445 – 0.037*Speed – 0.00033*AccLon – 0.00002*THead + 0.0016*rdist - 0.396*rspur – 0.00224*Driving Experience

Τα **πρόσημα των συντελεστών** έχουν λογική εξήγηση καθώς το αρνητικό πρόσημο στην εξίσωση, της ταχύτητας (Speed), της διαμήμους επιτάχυνσης (AccLon) και του χρόνου μέχρι τη σύγκρουση (THead), δείχνουν πως όσο μειώνονται αυξάνεται η πιθανότητα ύπαρξης συμβάντος. Ακόμα η απόκλιση του οχήματος από τη μέση του δρόμου (rspur) είναι λογικό να έχει αρνητικό πρόσημο στην ύπαρξη συμβάντος,και η εμπειρία στην οδήγηση (Driving Experience). Η μόνη μεταβλητή με θετικό συντελεστή είναι η συνολική διανυόμενη απόσταση (rdist).

Όσον αφορά τη **στατιστική σημαντικότητα** των μεταβλητών, σύμφωνα και με την υποενότητα 3.3.1., μια μεταβλητή θεωρείται στασιστικά σημαντική όταν Pr(>|z|)<0.05. Από τον παραπάνω πίνακα φαίνεται πως όλες οι μεταβλητές είναι στατιστικά σημαντικές για τον υπολογισμό της εξαρτημένης μεταβλητής.

5.2.1.2. Έλεγχος μοντέλου

Έπειτα από την εκπαίδευση του μοντέλου έγινε έλεγχος για το πόσο καλά αυτό προβλέπει τις τιμές της εξαρτημένης μεταβλητής στον πίνακα για έλεγχο του μοντέλου (**test1)**. Ο έλεγχος αυτός έγινε με τη μέθοδο της μήτρας σύγχυσης (confusion matrix) που παρουσιάστηκε στην υποενότητα 3.3.2. Ως **θετική κλάση** ορίστηκε ο αριθμός **1** που συμβολίζει την **ύπαρξη συμβάντος**. Τα αποτελέσματα της μήτρας σύγχυσης (σε αντιστοιχία με τον Πίνακα 3.1) δίνονται στον παρακάτω Πίνακα 5.5.

Πίνακας 5.5: Μήτρα Σύγχυσης

	Κατηγοριοποίηση Ταξινομητή (Πρόβλεψη)	
Πραγματική Κλάση	**Συμβάν**	**Όχι Συμβάν**
Συμβάν	*5448*	*14033*
Όχι Συμβάν	*2847*	*81943*

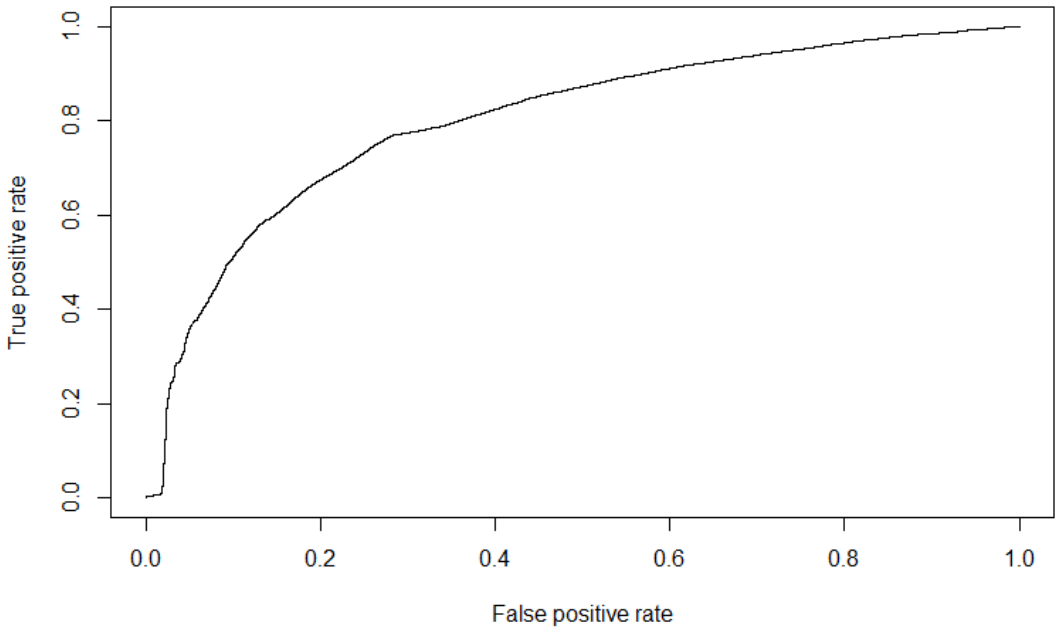
Στη συνέχεια υπολογίστηκαν οι απαραίτητες για την αξιολόγηση του μοντέλου μετρήσεις που ορίστηκαν στην υποενότητα 3.3.2 και εμφανίζονται στον Πίνακα 5.6.

Πίνακας 5.6: Μετρήσεις αξιολόγησης μοντέλου

ΜΕΤΡΗΣΗ	**ΤΙΜΗ**
Ορθότητα **(Accuracy)**	83.80 %
Στατιστικός συντελεστής Κάππα **(Kappa Statistic)**	31.60%

Ευαισθησία/Ανάκληση **(Sensitivity/Recall)**	27.90%
Εξειδικευτικότητα **(Specificity)**	96.60%
Ακρίβεια **(Precision)**	65.70%
Μέτρο F **(F-measure)**	39.20%
Δείκτης λάθος συναγερμού **(False alarm rate)**	3.30%
Εμβαδόν κάτω από την καμπύλη ROC **(AUC)**	80%

Το εμβαδόν κάτω από την καμπύλη (AUC), το οποίο στην ιδανική περίπτωση είναι ίσο με τη μονάδα, υπολογίστηκε από την καμπύλη ROC (Εικόνα 5.1) που απεικονίζει τη διαγνωστική ικανότητα του μοντέλου ταξινόμησης.



Εικόνα 5.1: Καμπύλη ROC

5.2.1.3. Αξιολόγηση μοντέλου

Για την 1η Παραλλαγή με το μοντέλο διωνυμικής λογιστικής παλινδρόμησης, σύμφωνα με τις μετρήσεις του Πίνακα 5.6, προέκυψαν τα εξής συμπεράσματα για την ικανότητα του μοντέλου να προβλέψει την ύπαρξη συμβάντος:

- Η συνολική ακρίβεια (accuracy) είναι ικανοποιητική που σημαίνει πως το μοντέλο είναι αρκετά αξιόπιστο για τις σωστές προβλέψεις, δηλαδή την ύπαρξη ή όχι συμβάντος.

- Το μοντέλο έχει μικρό μέτρο αξιολόγησης (Kappa Statistic), το οποίο είναι αρνητικό.
- Προβλέπει καθόλου ικανοποιητικά τα θετικά στιγμιότυπα (πολύ μικρό sensitivity/recall), επομένως δεν είναι αξιόπιστο όσον αφορά τις προβλέψεις για την ύπαρξη συμβάντος,
- ενώ έχει την ικανότητα να προβέψει αρκετά καλά τα αρνητικά στιγμιότυπα (αρκετά μεγάλο specificity), δηλαδή τα στιγμιότυπα που συμβολίζουν τη μη ύπαρξη συμβάντος.
- Το μέτρο της ακρίβειας (precision) δεν είναι ικανοποιητικό, με αποτέλεσμα ο βαθμός πιστότητας της διαδικασίας να είναι μικρός.
- Η πιθανότητα λάθους ταξινόμησης των θετικών στιγμιότυπων (false alarm rate), δηλαδή της ύπαρξη συμβάντος, δεν είναι ικανοποιητική γεγονός που καθιστά μη αξιόπιστο το μοντέλο.
- Το μέτρο F (F-measure) που εκφράζει τον αρμονικό μέσο της ακρίβειας και της ανάκλησης είναι επίσης πολύ χαμηλό.
- Το εμβαδόν κάτω από την καμπύλη ROC (AUC) είναι ικανοποιητικό.

Συνοψίζοντας τα παραπάνω προκύπτει το συμπέρασμα πως το μοντέλο της διωνυμικής λογιστικής παλινδρόμησης για την παραλλαγή Α δε λειτουργεί ικανοποιητικά στον εντοπισμό συμβάντος.

5.2.2. Παραλλαγή Β

Η Παραλλαγή Β που αναφέρθηκε στην εισαγωγή του κεφαλαίου δεν έδωσε ικανοποιητικά αποτελέσματα, ούτε για μεμονωμένες μετρήσεις, οπότε δε θεωρήθηκε άξιο ανάλυσης.

5.3. Ανάπτυξη μοντέλου τυχαίων δασών

5.3.1. Παραλλαγή Α

5.3.1.1. Εκπαίδευση μοντέλου

Για αυτήν την παραλλαγή χρησιμοποιήθηκαν οι ίδιοι πίνακες με την προηγούμενη μέθοδο για εκπαίδευση (training1) και έλεγχο (test1).

Αφού πραγματοποιήθηκε η εκπαίδευση στον πίνακα εκπαίδευσης (**training1**), προέκυψαν τα αποτελέσματα του Πίνακα 5.7.

Πίνακα 5.7: Αποτελέσματα εκπαίδευσης

Τύπος μεθόδου τυχαίων δασών: **ταξινόμηση**			
Αριθμός δέντρων:**100**			
Αριθμός μεταβλητών που δοκιμάστηκαν σε κάθε διαχωρισμό: **2**			
Ρυθμός σφάλματος κατηγοριοποίησης (OOB error rate**): 0.2%**			
Μήτρα Σύγχυσης			
	0	**1**	Λάθος ταξινόμησης

0	338812	346	0.00102
1	479	77443	0.00615

Από τα αποτελέσματα αυτά παρατηρείται πως ο ρυθμός σφάλματος της κατηγοριοποίησης (OOB error rate) που αναφέρθηκε στην υποενότητα 3.2.2. είναι μικρός. Αυτό αποτελεί μια πρώτη ένδειξη πως το μοντέλο υπάρχει μεγάλη πιθανότητα να είναι αξιόπιστο.

5.3.1.2. Έλεγχος μοντέλου

Για τον έλεγχο του μοντέλου εξετάστηκε η δυνατότητά του να προβλέψει την ύπαρξη ή όχι συμβάντος στον πίνακα test1. Αυτό πραγματοποιήθηκε με τη μήτρα σύγχυσης, θέτοντας ως **θετική κλάση** τον αριθμό **1**, δηλαδή την ύπαρξη συμβάντος. Τα αποτελέσματα της μήτρας σύγχυσης δίνονται στον παρακάτω Πίνακα 5.8.

Πίνακας 5.8: Μήτρα Σύγχυσης για την Παραλλαγή Α

	Κατηγοριοποίηση Ταξινομητή (Πρόβλεψη)	
Πραγματική Κλάση	**Συμβάν**	**Όχι Συμβάν**
Συμβάν	*19384*	*97*
Όχι Συμβάν	*79*	*84711*

Στη συνέχεια υπολογίστηκαν οι απαραίτητες για την αξιολόγηση του μοντέλου μετρήσεις και τα αποτελέσματά τους παρουσιάζονται στον Πίνακα 5.9.

Πίνακας 5.9: Μετρήσεις αξιολόγησης μοντέλου

ΜΕΤΡΗΣΗ	**ΤΙΜΗ**
Ορθότητα **(Accuracy)**	99.80%
Στατιστικός συντελεστής Κάππα **(Kappa Statistic)**	99.40%
Ευαισθησία/Ανάκληση **(Sensitivity/Recall)**	99.50%
Εξειδικευτικότητα **(Specificity)**	99.90%
Ακρίβεια **(Precision)**	99.60%
Μέτρο F **(F-measure)**	99.50%
Δείκτης λάθος συναγερμού **(False alarm rate)**	0.09%
Εμβαδόν κάτω από την καμπύλη **ROC (AUC)**	99.99%

5.3.1.3. Αξιολόγηση μοντέλου

Με βάση τις παραπάνω μετρήσεις προέκυψαν τα παρακάτω συμπεράσματα για αυτή την παραλλαγή εκτέλεσης του μοντέλου. Συγκεκριμένα παρατηρήθηκε πως το μοντέλο:

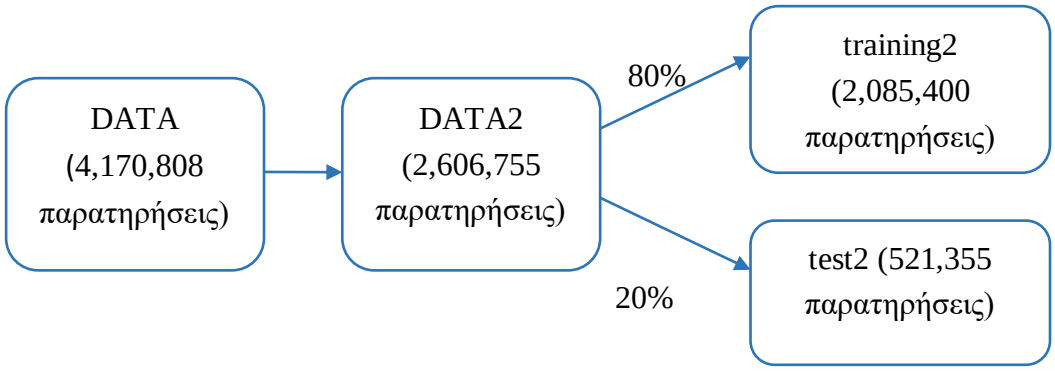
- Έχει πολύ μεγάλη ακρίβεια, δηλαδή είναι πολύ αξιόπιστο όσον αφορά τις σωστές προβλέψεις για την ύπαρξη ή όχι συμβάντος.
- Έχει μεγάλο μέτρο αξιολόγησης ως προς την ικανότητα κατηγοριοποίησής του.
- Προβλέπει πολύ ικανοποιητικά τα θετικά στιγμιότυπα, δηλαδή την ύπαρξη συμβάντος.
- Προβλέπει εξίσου καλά και τα αρνητικά στιγμιότυπα που αντιστοιχούν στη μη ύπαρξη συμβάντος.
- Έχει μεγάλη συνολική ακρίβεια, οπότε και υψηλό βαθμό πιστότητας για τη διαδικασία κατηγοριοποίησης.
- Έχει πολύ μικρή πιθανότητα λάθους ταξινόμησης των θετικών στιγμιότυπων, δηλαδή της ύπαρξη συμβάντος, γεγονός που καθιστά πολύ αξιόπιστο το μοντέλο και
- μεγάλο μέτρο F που εκφράζει τον αρμονικό μέσο της ακρίβειας και της ανάκλησης.
- Έχει πολύ μεγάλο εμβαδόν κάτω από την καμπύλη.

Από τα παραπάνω συμπεραίνεται πως το μοντέλο τυχαίων δασών για την παραλλαγή Α λειτουργεί πολύ καλά και δίνει ικανοποιητικά αποτελέσματα.

5.3.2. Παραλλαγή Β

Για να πραγματοποιηθεί παραλλαγή Β χρησιμοποιήθηκε ο πίκανας **DATA2** που αποτελούταν από 2606755 παρατηρήσεις σύνολο των μεταβλητών Event (εξαρτημένη), Speed, AccLon, rdist και Driving Experience του πίνακα DATA.

Ο πίνακας DATA2 χωρίστηκε επίσης σε 2 επιμέρους, έναν για εκπαίδευση (training2) που αποτελείται από το 80% (2,085,400 παρατηρήσεις) του και έναν για έλεγχο (test2) με το υπόλοιπο 20% (521,355 παρατηρήσεις). Τα βήματα αυτά περιγράφονται στο Γράφημα 5.2.



Γράφημα 5.2: Πίνακες παραλλαγής Β

Αποσπάσματα των πινάκων αυτών παρουσιάζονται στους Πίνακες 5.10 και 5.11 αντίστοιχα.

Πίνακας 5.10: Απόσπασμα πίνακα εκπαίδευσης training2 (10/417,080 γραμμές)

training2					
	Event	Speed	AccLon	rdist	Driving Experience
1	0	41.1	1.004	559.38	46
2	0	41.2	0.728	559.77	46
3	0	41.2	0.509	560.14	46
4	0	41.3	0.504	560.91	46
5	0	41.5	0.646	561.66	46
6	0	41.6	0.618	562.05	46
7	0	41.5	-0.253	563.57	46
8	0	41.5	-0.237	563.95	46
9	0	41.5	-0.200	564.34	46
10	0	41.4	-0.371	564.72	46

Πίνακας 5.11: Απόσπασμα πίνακα εκπαίδευσης test2 (10/104,271 γραμμές)

test2					
	Event	Speed	AccLon	rdist	Driving Experience
1	0	41.3	0.423	560.52	46
2	0	41.4	0.636	561.28	46
3	0	41.6	0.293	562.43	46
4	0	41.6	-0.024	562.81	46
5	0	41.6	-0.216	563.20	46
6	0	39.8	-0.408	571.79	46
7	0	39.6	-0.389	573.25	46
8	0	38.7	-0.374	580.81	46
9	0	38.5	-0.373	581.87	46
10	0	38.4	-0.373	582.93	46

5.3.2.1. Εκπαίδευση μοντέλου

Για την εκπαίδευση του μοντέλου σε αυτή την παραλλαγή χρησιμοποιήθηκε ο πίνακας εκπαίδευσης (training2) και τα αποτελέσματα που προέκυψαν φαίνονται στον Πίνακα 5.12.

Πίνακας 5.12

Τύπος μεθόδου τυχαίων δασών: ταξινόμηση			
Αριθμός δέντρων:100			
Αριθμός μεταβλητών που δοκιμάστηκαν σε κάθε διαχωρισμό: 2			
Ρυθμός σφάλματος κατηγοριοποίησης (OOB error rate): 0.87%			
Μήτρα Σύγχυσης			
	0	**1**	Λάθος ταξινόμησης
0	338302	856	0.00252
1	2848	75074	0.03655

Και σε αυτή την παραλλαγή παρατηρείται πως ο ρυθμός σφάλματος της κατηγοριοποίησης (OOB error rate) είναι μικρός, μικρότερος του 1%, που παραπέμπει σε αξιόπιστο μοντέλο. Παρακάτω γίνεται αναλυτικός έλεγχος.

5.3.2.2. Έλεγχος μοντέλου

Ο έλεγχος του μοντέλου έγινε στον πίνακα test2, για την ικανότητά του να προβλέψει την ανεξάρτητη μεταβλητή. Πραγματοποιήθηκε με τη μήτρα σύγχυσης, θέτοντας ως **θετική κλάση** τον αριθμό **1**, δηλαδή την ύπαρξη συμβάντος. Τα αποτελέσματα της μήτρας σύγχυσης δίνονται στον παρακάτω Πίνακα 5.13.

Πίνακας 5.13: Μήτρα Σύγχυσης για την παραλλαγή Β

	Κατηγοριοποίηση Ταξινομητή (Πρόβλεψη)	
Πραγματική Κλάση	**Συμβάν**	**Όχι συμβάν**
Συμβάν	*18823*	*658*
Όχι συμβάν	*198*	*84592*

Στη συνέχεια υπολογίστηκαν οι απαραίτητες για την αξιολόγηση του μοντέλου μετρήσεις που απεικονίζονται στον Πίνακα 5.14

Πίνακας 5.14: Μετρήσεις αξιολόγησης μοντέλου

ΜΕΤΡΗΣΗ	**ΤΙΜΗ**
Ορθότητα **(Accuracy)**	99.20%
Στατιστικός συντελεστής Κάππα **(Kappa Statistic)**	97.30%
Ευαισθησία/Ανάκληση **(Sensitivity/Recall)**	96.60%
Εξειδικευτικότητα **(Specificity)**	99.80%
Ακρίβεια **(Precision)**	98.90%
Μέτρο F **(F-measure)**	97.80%
Δείκτης λάθος συναγερμού **(False alarm rate)**	0.23%
Εμβαδόν κάτω από την καμπύλη ROC **(AUC)**	99.94%

5.3.2.3. Αξιολόγηση μοντέλου

Από τις παραπάνω μετρήσεις προέκυψαν τα παρακάτω συμπεράσματα για αυτή την παραλλαγή εκτέλεσης του μοντέλου. Το μοντέλο παρατηρήθηκε πως σε αυτή την περίπτωση:

- Έχει πολύ μεγάλη ακρίβεια, επομένως προβλέπει πολύ καλά την ύπαρξη ή όχι συμβάντος.
- Έχει μεγάλο μέτρο αξιολόγησης της ικανότητάς του για κατηγοριοποίηση.
- Προβλέπει αρκετά ικανοποιητικά τα θετικά στιγμιότυπα, σηλαδή την ύπαρξη συμβάντος.
- Προβλέπει εξίσου ικανοποιητικά και τα αρνητικά στιγμιότυπα, τη μη ύπαρξη συμβάντος.

- Έχει αρκετά μεγάλη συνολική ακρίβεια, επομένως έχει μεγάλη πιστότητα στη διαδικασία κατηγοριοποίησης.
- Έχει πολύ μικρή πιθανότητα λάθους ταξινόμησης των θετικών στιγμιότυπων, δηλαδή της ύπαρξη συμβάντος, γεγονός που καθιστά πολύ αξιόπιστο το μοντέλο και
- μεγάλο μέτρο F που εκφράζει τον αρμονικό μέσο της ακρίβειας και της ανάκλησης.
- Έχει πολύ μεγάλο εμβαδόν κάτω από την καμπύλη.

Σύμφωνα με τα παραπάνω, προκύπτει το συμπέρασμα πως το μοντέλο τυχαίων δασών λειτουργεί πολύ ικανοποιητικά και για την παραλλαγή Β.

5.4. Ανάλυση Παραγόντων

5.4.1. Μέθοδος Παραγοντικής Ανάλυσης στον πίνακα PreEvent3

5.4.1.1. Επιλογή αριθμού παραγόντων με τη μέθοδο κύριων συνιστωσών

Πρώτο βήμα ήταν η δημιουργία του πίνακα συσχέτισης των ανεξάρτητων μεταβλητών που φαίνεται στον Πίνακα 5.18.

Πίνακας 5.18: Πίνακας συσχέτισης ανεξάρτητων μεταβλητών

	Speed	Acclon	Thead	rdist	rspur	Wheel
Speed	1.0000					
Acclon	-0.0066	1.0000				
Thead	-0.4827	-0.0210	1.0000			
rdist	0.0896	-0.0218	0.2330	1.0000		
rspur	0.0025	0.0089	-0.0564	-0.0086	1.0000	
Wheel	-0.1112	0.0073	-0.0179	0.1605	-0.0281	1.0000

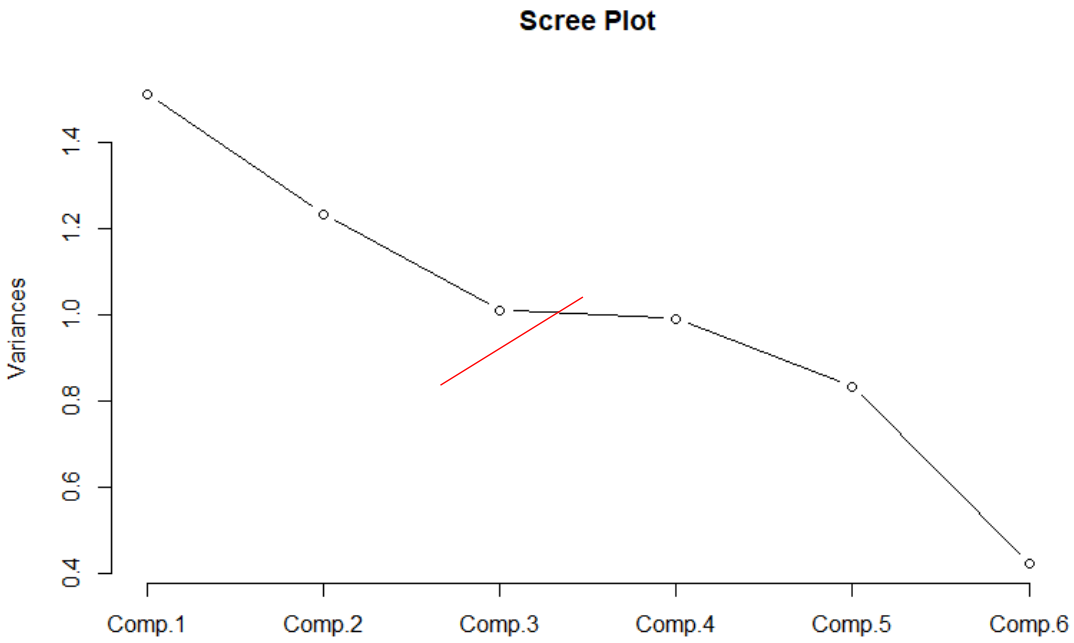
Παρατηρείται πως οι μεταβλητές δεν έχουν μεγάλη συσχέτιση μεταξύ τους (< 0.8 παντού), όμως οι περισσότερες από αυτές έχουν σχεδόν καθόλου συσχέτιση. Παρ' όλα αυτά η μέθοδος εκτελέστηκε και αξιολογήθηκαν τα αποτελέσματα, προκειμένου να εξεταστεί αν υπάρχουν παράγοντες μέσα στα δεδομένα που θα βοηθήσουν περαιτέρω την ανάλυση.

Στη συνέχεια για την επιλογή του αριθμού των παραγόντων χρησιμοποιήθηκε η μέθοδος κύριων συνιστωσών, η οποία μελετά όλη την υπάρχουσα διακύμανση (κοινή, μοναδική και σφάλμα), με στόχο να εξαχθεί το μεγαλύτερο ποσοστό της διακύμανσης από τις λιγότερες δυνατές συνιστώσες. Σημαντικό βήμα για την επιλογή του αριθμού των παραγόντων ήταν η εξέταση της σημαντικότητας των συνιστωσών (Πίνακας 5.19).

Πίνακας 5.19: Σημαντικότητα συνιστωσών

	Συν.1	**Συν.2**	**Συν.3**	**Συν.4**	**Συν.5**	**Συν.6**
Τυπική Απόκλιση	1.2294	1.1104	1.0051	0.9948	0.9131	0.6496
Ποσοστό Διακύμανσης	0.2519	0.2055	0.1684	0.1649	0.1390	0.0703
Αθροιστικό Ποσοστό	0.2519	0.4574	0.6258	0.7907	0.9297	1.0000

Δημιουργήθηκε ακόμα το διάγραμμα της Εικόνας 5.1 που απεικονίζει τη διακύμανση ανά συντελεστή. Από το διάγραμμα αυτό προέκυψε ο αριθμός των παραγόντων που χρησιμοποιήθηκαν για την ανάλυση παραγόντων. Επιλέχθηκαν 3 παράγοντες, καθώς για το πλήθος των ανεξάρτητων μεταβλητών αυτός είναι ο μεγαλύτερος αριθμός παραγόντων που μπορεί να χρησιμοποιηθεί και το συνολικό ποσοστό διακύμανσης ήταν μεγαλύτερο από το 0.6 το οποίο είναι ικανοποιητικό.



Εικόνα 5.1: Διάγραμμα Διακύμανσης- Συνιστωσών

5.4.1.2. Ανάλυση Παραγόντων

Στη συνέχεια έγινε ανάλυση παραγόντων για αριθμό παραγόντων ίσο με 3. Προέκυψε η μοναδικότητα των ανεξάρτητων μεταβλητών που απεικονίζεται στον Πίνακα 5.20, η οποία αποτελεί ένα πρώτο δείγμα για το αν μια μεταβλητή μπορεί να εκφραστεί μέσο ενός παράγοντα ή όχι. Μεταβλητές με μεγάλη μοναδικότητα συνήθως δεν εκφράζονται μέσω παραγόντων.

Πίνακας 5.20: Μοναδικότητα μεταβλητών

Speed	Acclon	Thead	rdist	rspur	Wheel
0.564	0.998	0.054	0.695	0.927	0.891

Από τον παραπάνω πίνακα παρατηρείται πως οι μεταβλητές AccLon, rspur, Wheel είναι πολύ πιθανό να μην εκφράζονται μέσω παραγόντων. Τα αποτελέσματα της παραγοντικής ανάλυσης, δηλαδή της συσχέτισης μεταξύ παραγόντων και μεταβλητών, παρουσιάζονται στον Πίνακα 5.21.

Πίνακας 5.21: Ανάλυση Παραγόντων

	Παρ.1	Παρ.2	Παρ.3
Speed	-0.616	0.231	
Acclon			
Thead	0.908	0.285	0.202
rdist		0.539	0.1
rspur			-0.267
Wheel		-0.317	
	Παρ.1	**Παρ.2**	**Παρ.3**
SS loadings	1.212	0.528	0.13
Ποσοστό Διακύμανσης	0.202	0.088	0.022
Αθροιστικό Ποσοστό	0.202	0.29	0.312

Μια μεταβλητή είναι εφικτό να εκφραστεί μέσω ενός παράγοντα όταν ο βαθμός ικανότητας του παράγοντα να την εκφράσει (Πίνακας 5.21) είναι μεγαλύτερος του 50%. Έτσι με βάση των παραπάνω πίνακα προκύπτουν οι εξής ομάδες παραγόντων:

Πίνακας 5.22: Ομάδες παραγόντων

Παράγοντας 1	Παράγοντας 2
Speed	rdist
AccLon	

Ο πίνακας 5.22 μας δείχνει πως τα στοιχεία που περιγράφουν την κατάσταση πριν από κάθε συμβάν μπορούν να εκφραστούν με δύο παράγοντες1) τον παράγοντα που περιγράφει την επιρροή της ταχύτητας και της επιμήκους επιτάχυνσης και ιι) τον παράγοντα που περιγράφει την επιρροή της συνολικής διανυόμενης απόστασης του οχήματος.

5.4.1.3. Έλεγχος ποιότητας δεδομένων δείγματος

Για τον έλεγχο του δείγματος έγιναν 2 βασικοί έλεγχοι που αναφέρονται στην υποενότητα 3.4, έλεγχος για την επάρκεια του δείγματος (KMO) και έλεγχος για τη σφαιρικότητα του δείγματος (Bartlett Test).

1ος Έλεγχος: KMO

Από τον έλεγχο αυτόν προέκυψε συντελεστής KMO = 0.4282 < 0.6, οπότε δεν ικανοποιείται ο έλεγχος επάρκειας δείγματος.

2ος Έλεγχος: Bartlett Test

Αποτέλεσμα αυτού του ελέγχου ήταν η τιμή p-value < 0.05, οπότε ικανοποιείται ο έλεγχος σφαιρικότητας του δείγματος.

5.4.2. Μέθοδος Παραγοντικής Ανάλυσης στον πίνακα DurEvent3

5.4.2.1. Επιλογή αριθμού παραγόντων με τη μέθοδο κύριων συνιστωσών

Αρχικά βρέθηκε η συσχέτιση μεταξύ των ανεξάρτητων μεταβλητών (Πίνακας 5.23).

Πίνακας 5.23: Πίνακας συσχέτισης ανεξάρτητων μεταβλητών

	Speed	Acclon	Thead	rdist	rspur	Wheel
Speed	1.0000					
Acclon	0.0048	1.0000				
Thead	-0.4648	0.0002	1.0000			
rdist	-0.0628	-0.0017	0.2887	1.0000		
rspur	0.0428	0.0195	-0.1154	0.0022	1.0000	
Wheel	-0.1151	-0.0097	0.1261	0.0651	0.1552	1.0000

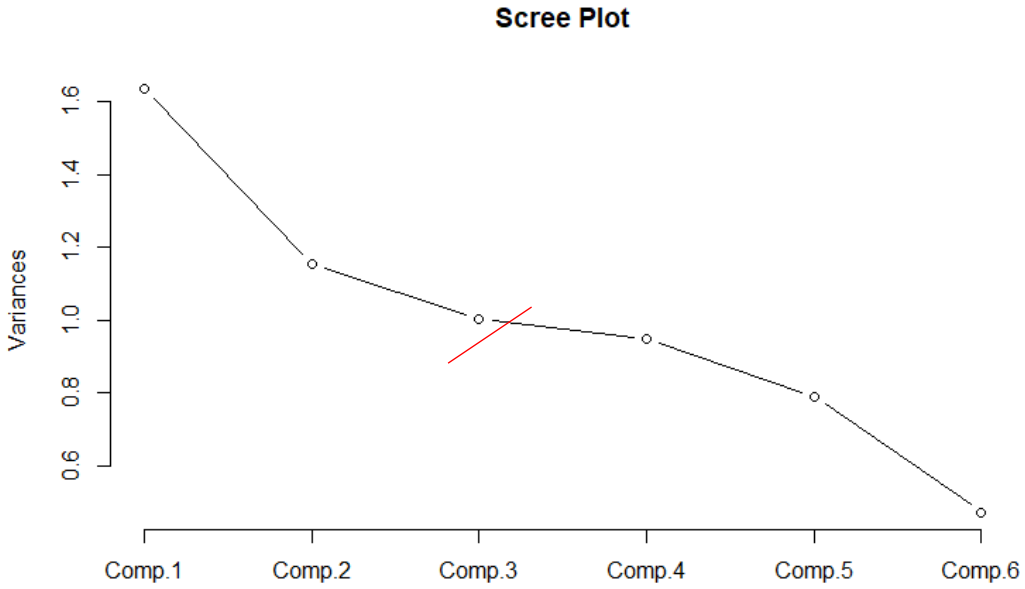
Οι ίδιες παρατηρήσεις για τη συσχέτιση των μεταβλητών που έγιναν για τον πίνακα PreEvent3 ισχύουν κι εδώ.

Για την επιλογή του αριθμού των παραγόντων χρησιμοποιήθηκε η μέθοδος των κύριων συνιστωσών, τα αποτελέσματα της σημαντικότητας των συνιστωσών παρουσιάζονται στον Πίνακα 5.24.

Πίνακας 5.24: Σημαντικότητα συνιστωσών

	Συν.1	Συν.2	Συν.3	Συν.4	Συν.5	Συν.6
Τυπική Απόκλιση	1.2787	1.0743	1.0006	0.9734	0.8888	0.6869
Ποσοστό Διακύμανσης	0.2725	0.1924	0.1669	0.1579	0.1317	0.0786
Αθροιστικό Ποσοστό	0.2725	0.4649	0.6318	0.7897	0.9213	1.0000

Δημιουργήθηκε και σε αυτή την περίπτωση το διάγραμμα της Εικόνας 5.2 που απεικονίζει τη διακύμανση ανά συντελεστή. Από το διάγραμμα αυτό προέκυψε ο αριθμός των παραγόντων που χρησιμοποιήθηκαν για την ανάλυση παραγόντων. Επιλέχθηκαν 3 παράγοντες, καθώς για το πλήθος των ανεξάρτητων μεταβλητών αυτός είναι ο μεγαλύτερος αριθμός παραγόντων που μπορεί να χρησιμοποιηθεί και το συνολικό ποσοστό διακύμανσης ήταν μεγαλύτερο από το 0.6 το οποίο είναι ικανοποιητικό.



Εικόνα 5.2: Διάγραμμα Διακύμανσης-Συνιστωσών

5.4.2.2. Ανάλυση Παραγόντων

Όπως και στην προηγούμενη περίπτωση έγινε ανάλυση παραγόντων για αριθμό παραγόντων ίσο με 3 και προέκυψε η μοναδικότητα των ανεξάρτητων μεταβλητών που απεικονίζεται στον Πίνακα 5.25.

Πίνακας 5.25: Μοναδικότητα μεταβλητών

Speed	Acclon	Thead	rdist	rspur	Wheel
0.005	1	0.373	0.828	0.005	0.942

Από τον παραπάνω πίνακα παρατηρείται πως οι μεταβλητές AccLon, rdist, Wheel είναι πολύ πιθανό να μην εκφράζονται μέσω παραγόντων. Τα αποτελέσματα της παραγοντικής ανάλυσης παρουσιάζονται στον Πίνακα 5.26.

Πίνακας 5.26: Ανάλυση Παραγόντων

	Παρ.1	**Παρ.2**	**Παρ.3**
Speed	0.98		-0.174
Acclon			
Thead	-0.343		0.710
rdist			0.414
rspur		0.994	
Wheel	-0.102	0.164	0.145
	Παρ.1	**Παρ.2**	**Παρ.3**
SS loadings	1.09	1.026	0.733
Ποσοστό Διακύμανσης	0.182	0.171	0.122
Αθροιστικό Ποσοστό	0.182	0.353	0.475

Με βάση των παραπάνω πίνακα προκύπτουν οι εξής ομάδες παραγόντων:

Πίνακας 5.27: Ομάδες παραγόντων

Παράγοντας 1	**Παράγοντας 2**	**Παράγοντας 3**
Speed	rspur	THead

Από τον πίνακα 5.27 συμπεραίνεται πως τα στοιχεία που περιγράφουν την κατάσταση κατά τη διάρκεια κάθε συμβάντος μπορούν να εκφραστούν με τρεις παράγοντες: 1) τον παράγοντα που περιγράφει την επιρροή της ταχύτητας, 2) τον παράγοντα που περιγράφει την επιρροή της απόκλιση του οχήματος από το μέσο της οδού και 3) τον παράγοντα που περιγράφει την επιρροή του χρόνου μέχρι τη σύγκρουση από το προπορευόμενο όχημα.

5.4.2.3. Έλεγχος ποιότητας δεδομένων δείγματος

1ος Έλεγχος: KMO

Από τον έλεγχο αυτόν προέκυψε συντελεστής KMO = 0.5095 < 0.6, οπότε δεν ικανοποιείται ο έλεγχος επάρκειας δείγματος.

2ος Έλεγχος: Bartlett Test

Αποτέλεσμα αυτού του ελέγχου ήταν η τιμή p-value < 0.05, οπότε ικανοποιείται ο έλεγχος σφαιρικότητας του δείγματος.

5.4.3. Μέθοδος Παραγοντικής Ανάλυσης στον πίνακα Events3

5.4.3.1. Επιλογή αριθμού παραγόντων με τη μέθοδο κύριων συνιστωσών

Αρχικά βρέθηκε η συσχέτιση μεταξύ των ανεξάρτητων μεταβλητών (Πίνακας 5.28).

Πίνακας 5.28: Πίνακας συσχέτισης ανεξάρτητων μεταβλητών

	Speed	**Acclon**	**Thead**	**rdist**	**rspur**	**Wheel**
Speed	1.0000					
Acclon	0.0102	1.0000				
Thead	-0.5102	-0.0099	1.0000			
rdist	-0.0398	-0.0136	0.2867	1.0000		
rspur	0.0286	0.0121	-0.0815	-0.0235	1.0000	
Wheel	-0.1061	-0.0027	0.0162	-0.1216	0.0081	1.0000

Οι ίδιες παρατηρήσεις για τη συσχέτιση των μεταβλητών που έγιναν για τον πίνακα PreEvent3 και DurEvent3 ισχύουν κι εδώ.

Για την επιλογή του αριθμού των παραγόντων χρησιμοποιήθηκε η μέθοδος των κύριων συνιστωσών, τα αποτελέσματα της σημαντικότητας των συνιστωσών παρουσιάζονται στον Πίνακα 5.29.

Πίνακας 5.29: Σημαντικότητα συνιστωσών

	Συν.1	**Συν.2**	**Συν.3**	**Συν.4**	**Συν.5**	**Συν.6**
Τυπική Απόκλιση	1.271	1.069	1.002	0.992	0.907	0.655
Ποσοστό Διακύμανσης	0.269	0.190	0.167	0.164	0.137	0.071
Αθροιστικό Ποσοστό	0.269	0.460	0.627	0.791	0.929	1.000

Δημιουργήθηκε και σε αυτή την περίπτωση το διάγραμμα της Εικόνας 5.3 που απεικονίζει τη διακύμανση ανά συντελεστή. Πάλι επιλέχθηκαν 3 παράγοντες, καθώς για το πλήθος των ανεξάρτητων μεταβλητών αυτός είναι ο μεγαλύτερος αριθμός παραγόντων που μπορεί να χρησιμοποιηθεί και το συνολικό ποσοστό διακύμανσης ήταν μεγαλύτερο από το 0.6 το οποίο είναι ικανοποιητικό.

Εικόνα 5.3: Διάγραμμα Διακύμανσης-Συνιστωσών

5.4.3.2. Ανάλυση Παραγόντων

Όπως και στην προηγούμενη περίπτωση έγινε ανάλυση παραγόντων για αριθμό παραγόντων ίσο με 3 και προέκυψε η μοναδικότητα των ανεξάρτητων μεταβλητών που απεικονίζεται στον Πίνακα 5.30.

Πίνακας 5.30: Μοναδικότητα μεταβλητών

Speed	Acclon	Thead	rdist	rspur	Wheel
0,005	1	0.655	0.109	0.005	0.971

Από τον παραπάνω πίνακα παρατηρείται πως οι μεταβλητές AccLon, rdist, Wheel είναι πολύ πιθανό να μην εκφράζονται μέσω παραγόντων. Τα αποτελέσματα της παραγοντικής ανάλυσης παρουσιάζονται στον Πίνακα 5.31.

Πίνακας 5.31: Ανάλυση Παραγόντων

	Παρ.1	Παρ.2	Παρ.3
Speed	0.953		0.295
Acclon			
Thead	-0.572		0.117
rdist	-0.317		0.889
rspur		0.997	
Wheel			-0.159
	Παρ.1	**Παρ.2**	**Παρ.3**
SS loadings	1.341	0.998	0.916

Ποσοστό Διακύμανσης	0.224	0.166	0.153
Αθροιστικό Ποσοστό	0.224	0.39	0.543

Με βάση των παραπάνω πίνακα προκύπτουν οι εξής ομάδες παραγόντων:

Πίνακας 5.32: Ομάδες παραγόντων

Παράγοντας 1	**Παράγοντας 2**	**Παράγοντας 3**
Speed	rspur	rdist
THead		

Από τον πίνακα 5.32 συμπεραίνεται πως τα στοιχεία που περιγράφουν την κατάσταση ένα λεπτό πριν και κατά τη διάρκεια κάθε συμβάντος μπορούν να εκφραστούν με τρεις παράγοντες: 1) τον παράγοντα που περιγράφει την επιρροή της ταχύτητας και του χρόνου μέχρι τη σύγκρουση από το προπορευόμενο όχημα, 2) τον παράγοντα που περιγράφει την επιρροή της απόκλισης του οχήματος από το μέσο της οδού και 3) τον παράγοντα που περιγράφει την επιρροή της συνολικής διανυόμενης απόστασης του οχήματος.

5.4.2.3. Έλεγχος ποιότητας δεδομένων δείγματος

1ος Έλεγχος: KMO

Από τον έλεγχο αυτόν προέκυψε συντελεστής KMO = 0.4826 < 0.6, οπότε δεν ικανοποιείται ο έλεγχος επάρκειας δείγματος.

2ος Έλεγχος: Bartlett Test

Αποτέλεσμα αυτού του ελέγχου ήταν η τιμή p-value < 0.05, οπότε ικανοποιείται ο έλεγχος σφαιρικότητας του δείγματος.

5.5. Σύνοψη και σχολιασμός αποτελεσμάτων του κεφαλαίου

Στο κεφάλαιο αυτό πραγματοποιήθηκε η στατιστική ανάλυση των δεδομένων της διπλωματικής εργασίας, με σκοπό τον εντοπισμό συμβάντων με βάση τα οδηγικά χαρακτηριστικά. Χρησιμοποιήθηκαν 2 μοντέλα στατιστικής ανάλυσης, το μοντέλο διωνυμικής λογιστικής παλινδρόμησης και το μοντέλο τυχαίων δασών, καθώς και η μέθοδος παραγοντικής ανάλυσης προκειμένου να διερευνηθεί η δυνατότητα ομοδοποίησης των μεταβλητών.

Για τα **μοντέλα λογιστικής παλινδρόμησης και τυχαίων δασών** πραγματοποιήθηκαν 2 παραλλαγές, Α και Β, σε δεδομένα που αφορούσαν τα σύνολο των στοιχείων για

1΄πριν και κατά τη διάρκεια συμβάντων, προκειμένου να εξεταστεί η δυνατότητα πρόβλεψης του μοντέλου για την ύπαρξη ή όχι συμβάντος. Τα αποτελέσματα που προέκυψαν έδειξαν πως το μοντέλο τυχαίων δασών λειτουργεί καλύτερα για το συγκεκριμένο σύνολο δεδομένων. Πιο συγκεκριμένα για την παραλλαγή Α προέκυψε πως το μοντέλο λογιστικής παλινδρόμησης δεν είχε την ικανότητα να προβλέψει την ύπαρξη συμβάντος, ενώ ήταν αρκετά αποτελεσματικό για την πρόβλεψη της μη ύπαρξης συμβάντος. Η παραλλαγή Β κατά τις δοκιμές δεν έδωσε ικανοποιητικά αποτελέσματα ούτε για μεμονωμένες μετρήσεις και δεν αναφέρθηκε αναλυτικά στην παρούσα διπλωματική. Αντίθετα το μοντέλο τυχαίων δασών έδωσε πολύ ικανοποιητικά αποτελέσματα και στις 2 παραλλαγές, τόσο για την πρόβλεψη ύπαρξης συμβάντος όσο και μη, με πολύ μικρό ποσοστό πιθανότητας λάθους κατά την ταξινόμησης της ύπαρξης συμβάντος. Βασικοί δείκτες για την αξιολόγηση των μοντέλων είναι ο δείκτης recall, που δείχνει πόσο καλά προβλέπει το μοντέλο την ύπαρξη συμβάντος, ο δείκτης false alarm rate που εκφράζει την πιθανότητα λάθους αυτής της πρόβλεψης, ο δείκτης specificity για την πρόβλεψη της κατάστασης μη ύπαρξης συμβάντος και το εμβαδόν κάτω από την καμπύλη (AUC) για τις σωστές προβλέψεις στο σύνολό τους. Στο σύνολό τους αυτοί οι δείκτες ορίζουν την αξιοπιστία των μοντέλων ως προς τη σωστή πρόβλεψη της ύπαρξης συμβάντων. Για την παραλλαγή κάθε μοντέλου, τα αποτελέσματα παρουσιάζονται στους Πίνακες 5.33 και 5.34.

Πίνακας 5.33: Διωνυμική λογιστική παλινδρόμηση

	Δείκτης			
Παραλλαγή	Recall	Specificity	False alarm rate	AUC
A	27.90%	96.60%	3.30%	80%

Πίνακας 5.34: Τυχαία Δάση

	Δείκτης			
Παραλλαγή	Recall	Specificity	False alarm rate	AUC
A	99.50%	99.90%	0.09%	99.99%
B	96.60%	99.80%	0.23%	99.94%

Συγκρίνοντας τις δύο παραλλαγές του μοντέλου τυχαίων δασών φαίνεται πως δίνει καλύτερα αποτελέσματα στην παραλλαγή που χρησιμοποιείται το σύνολο των ανεξάρτητων μεταβλητών, Α.

Η **ανάλυση παραγόντων** πραγματοποιήθηκε ξεχωριστά για τα δεδομένα που αφορούσαν το χρόνο πριν το συμβάν, κατά τη διάρκεια του συμβάντος και το σύνολο αυτών. Παρατηρήθηκε πως τα αποτελέσματα στις 3 αυτές περιπτώσεις διέφεραν αρκετά, τόσο στον αριθμό των παραγόντων όσο και των μεταβλητών που εξέφραζε καθένας από αυτούς.

Τα στοιχεία που αφορούσαν τη χρονική διάρκεια του **ενός λεπτού πριν το συμβάν** προέκυψε ότι μπορούν να εκφραστούν μέσω της ταχύτητας, της επιτάχυνσης και της συνολικής απόστασης που έχει διανύσει ο οδηγός. Το γεγονός αυτό θα μπορούσε να εξηγηθεί λογικά καθώς, όπως έχει προκύψει σε έρευνες, η ταχύτητα και η επιτάχυνση

μεταβάλλονται έντονα πριν από ένα συμβάν. Επίσης το αίσθημα κούρασης που ενδέχεται να έχει δημιουργηθεί στον οδηγό λόγω της μεγάλης απόστασης που έχει διανύσει είναι πιθανόν να αυξήση την πιθανότητα εμπλοκής του σε κάποιο συμβάν.

Βρέθηκε πως η ταχύτητα αποτελεί σημαντικό παράγοντα και στην περίπτωση εξέτασης των δεδομένων που αφορούν τα **συμβάντα στη διάρκειά τους**, ενώ η επιτάχυνση και η διανυόμενη απόσταση παύουν να παίζουν κυρίαρχο ρόλο και τη θέση τους παίρνουν η απόκλιση του οχήματος από το μέσο της οδού και ο χρόνος μέχρι τη σύγκρουση, που μεταβάλλονται έντονα κατά τη διάρκεια ενός περιστατικού στην οδό.

6. ΣΥΜΠΕΡΑΣΜΑΤΑ

6.1. Σύνοψη αποτελεσμάτων

Η εκπόνηση της παρούσας διπλωματικής εργασίας έχει **στόχο τη διερεύνηση της ικανότητας εντοπισμού ύπαρξης συμβάντων με βάση τα οδηγικά χαρακτηριστικά σε υπεραστικές οδούς.** Για το σκοπό αυτό χρησιμοποιήθηκαν στοιχεία από πείραμα σε προσομοιωτή οδήγησης σε υπεραστικές οδούς.

Τα στοιχεία αυτά επεξεργάστηκαν κατάλληλα προκειμένου να ετοιμαστούν οι συγκεντρωτικοί πίνακες δεδομένων που περιλάμβαναν στοιχεία που αφορούσαν στη χρονική διάρκεια του ενός λεπτού πριν και κατά τη διάρκεια κάθε συμβάντος. Οι πίνακες αυτοί, που χρησιμοποιήθηκαν στη στατιστική ανάλυση, αποτελούνταν από την εξαρτημένη μεταβλητή που εκφράζει την ύπαρξη ή όχι συμβάντος και από ένα σύνολο από ανεξάρτητες, μη συσχετισμένες μεταξύ τους μεταβλητές, όπως η ταχύτητα του οχήματος, η διαμήκης επιτάχυνση, ο χρόνος μέχρι τη σύγκρουση από το προπορευόμενο όχημα, η συνολική διανυόμενη απόσταση, η απόκλιση του οχήματος από τη μέση του δρόμου, η γωνία του τιμονιού και η οδηγική εμπειρία.

Για τη στατιστική ανάλυση των δεδομένων χρησιμοποιήθηκαν τα μοντέλα διωνυμικής λογιστικής παλινδρόμησης και τυχαίων δασών, για τα οποία πραγματοποιήθηκαν δύο παραλλαγές, Α και Β, ανάλογα με το πλήθος των ανεξάρτητων μεταβλητών που χρησιμοποιήθηκαν στην καθεμία. Η παραλλαγή Α περιελάμβανε όλες τις μεταβλητές που έπειτα από δοκιμές βρέθηκαν στατιστικά σημαντικές, ενώ για την επιλογή των μεταβλητών της παραλλαγής Β, έγινε προσδιορισμός της σημαντικότητάς τους και ανάλογα με αυτή και την αναγκαιότητα εξαγωγής χρήσιμων αποτελεσμάτων επιλέχθηκαν οι 4 πιο σημαντικές ανεξάρτητες μεταβλητές. Συγκεκριμένα κάθε παραλλαγή περιελάμβανε τις εξής μεταβλητές:

- **Παραλλαγή Α**: την εξαρτημένη διακριτή μεταβλητή **Event** και τις ανεξάρτητες **Speed, AccLon, Thead, rdist, rspur και Driving Experience.**
- **Παραλλαγή Β**: την εξαρτημένη διακριτή μεταβλητή **Event** και τις ανεξάρτητες **Speed, AccLon, rdist και Driving Experience.**

Τα αποτελέσματα της παραλλαγής Β για το μοντέλο της διωνυμικής λογιστικής παλινδρόμησης δεν ήταν ικανοποιητικά, οπότε αξιοποιήθηκαν μόνο για το μοντέλο τυχαίων δασών.

Επίσης πραγματοποιήθηκε η μέθοδος ανάλυσης παραγόντων για κάθε μία από τις περιπτώσεις, ένα λεπτό πριν από κάθε συμβάν, κατά τη διάρκεια καθώς και για το άθροισμα αυτών, για τη διερεύνηση της ικανότητας παραγόντων να εκφράσουν μια ομάδα μεταβλητών.

Πίνακας 1: Μοντέλο διωνυμικής λογιστικής παλινδρόμησης

	Παραλλαγή Α
Μέτρηση	**Τιμή**
Ορθότητα **(Accuracy)**	83.80%
Στατιστικός συντελεστής Κάππα **(Kappa Statistic)**	31.60%
Ευαισθησία/Ανάκληση **(Sensitivity/Recall)**	27.90%
Εξειδικευτικότητα **(Specificity)**	96.60%
Ακρίβεια **(Precision)**	65.70%
Μέτρο F **(F-measure)**	39.20%
Δείκτης λάθος συναγερμού **(False alarm rate)**	3.30%
Εμβαδόν κάτω από την καμπύλη ROC **(AUC)**	80%

Πίνακας 2: Μοντέλο τυχαίων δασών

	Παραλλαγή Α	**Παραλλαγή Β**
Μέτρηση	**Τιμή**	**Τιμή**
Ορθότητα **(Accuracy)**	99.80%	99.20%
Στατιστικός συντελεστής Κάππα **(Kappa Statistic)**	99.40%	97.30%
Ευαισθησία/Ανάκληση **(Sensitivity/Recall)**	99.50%	96.60%
Εξειδικευτικότητα **(Specificity)**	99.90%	99.80%
Ακρίβεια **(Precision)**	99.60%	98.90%
Μέτρο F **(F-measure)**	99.50%	97.80%
Δείκτης λάθος συναγερμού **(False alarm rate)**	0.09%	0.23%
Εμβαδόν κάτω από την καμπύλη ROC **(AUC)**	99.99%	99.94%

Πίνακας 3: Ανάλυση Παραγόντων

Πίνακας	**Παράγοντας 1**	**Παράγοντας 2**	**Παράγοντας 3**
1΄πριν κάθε συμβάν	Speed	Rdist	
	AccLon		
διάρκεια συμβάντος	Speed	Rspur	THead
1΄πριν/κατά τη διάρκεια κάθε συμβάντος	Speed	rspur	rdist
	THead		

6.2. Συνολικά συμπεράσματα

Τα αποτελέσματα που προέκυψαν από τη στατιστική ανάλυση οδήγησαν στη **διατύπωση των παρακάτω συμπερασμάτων** για την παρούσα διπλωματική εργασία.

- Οι μεταβλητές που περιγράφουν την ταχύτητα του οχήματος, τη διαμήκη επιτάχυνση, τη συνολική διανυόμενη απόσταση καθώς και την εμπειρία του οδηγού στην οδήγηση, προέκυψε πως αποτελούν τις **μεταβλητές με τη μεγαλύτερη σημαντικότητα** για τον εντοπισμό της ύπαρξης ενός συμβάντος. Το γεγονός αυτό επιβεβαιώνεται και από τη διεθνή βιβλιογραφία.

- Βασικό συμπέρασμα είναι ότι το **μοντέλο τυχαίων δασών ήταν αρκετά πιο αποδοτικό από τη διωνυμική λογιστική παλινδρόμηση** στη στατιστική ανάλυση που πραγματοποιήθηκε. Αυτό είναι πιθανό να οφείλεται στο γεγονός ότι το μοντέλο τυχαίων δασών περιγράφει καλύτερα την ύπαρξη συμβάντος, καθώς ζυγίζει ορισμένα χαρακτηριστικά ως πιο σημαντικά από άλλα, δεν υποθέτει ότι το μοντέλο έχει γραμμική σχέση όπως τα μοντέλα παλινδρόμησης, και επεξεργάζεται τυχαία δείγματα έτσι ώστε να καταλήξει στο βέλτιστο μοντέλο.

- Η εφαρμογή του **μοντέλου διωνυμικής λογιστικής παλινδρόμησης** παρατηρήθηκε πως δεν λειτουργεί αποτελεσματικά στον εντοπισμό ενός συμβάντος με μικρό αριθμό ανεξάρτητων μεταβλητών (μικρή ανάκληση, σχετικά υψηλός δείκτης ποσοστού πιθανότητας λάθους ταξινόμησης, χαμηλές τιμές ακρίβειας και μέτρο αξιολόγησης). Στην περίπτωση που έχουν ληφθεί υπόψη όλες οι στατιστικά σημαντικές μεταβλητές (παραλλαγή Α), η **ορθότητα** του μοντέλου διωνυμικής λογιστικής παλινδρόμησης, δηλαδή η ακρίβεια που προσφέρει για τις σωστές προβλέψεις στο σύνολό τους, τη σωστή πρόβλεψη για την ύπαρξη ή όχι συμβάντος, είναι ικανοποιητική.

- Τα αποτελέσματα που εξήχθηκαν από το **μοντέλο τυχαίων δασών** και για τις δύο παραλλαγές που αναφέρθηκαν παραπάνω ήταν στο σύνολό τους πολύ καλύτερα από το μοντέλο διωνυμικής λογιστικής παλινδρόμησης. Η **ορθότητα** στην πρόβλεψη ύπαρξης συμβάντος ή όχι ήταν πολύ υψηλή και για τις 2 παραλλαγές και η **ανάκληση** του μοντέλου (αυξημένη ικανότητα εύρεσης των στιγμιότυπων) βρέθηκε πολύ υψηλή. Ο δείκτης της **εξειδικευτικότητας**, προέκυψε επίσης πολύ υψηλός, όπως άλλωστε και η **ακρίβεια** (βαθμός πιστότητας) και το **μέτρο αξιολόγησης**, που καθιστούν το μοντέλο πολύ αξιόπιστο για τον εντοπισμό συμβάντων με βάση τα οδηγικά χαρακτηριστικά.

- **Συγκρίνοντας τις δύο παραλλαγές** που πραγματοποιήθηκαν με το μοντέλο τυχαίων δασών για τη στατιστική ανάλυση, προέκυψε πως παρόλο που και οι δύο παραλλαγές έξαγουν χρήσιμα και αξιόπιστα αποτελέσματα και είναι αποδεκτές, εκείνη που περιείχε το μεγαλύτερο πλήθος μεταβλητών έδινε καλύτερους δείκτες.

- Τα αποτελέσματα της **παραγοντικής ανάλυσης** έδειξαν πως για κάθε σύνολο στοιχείων, i) εκείνων που περιγράφουν τη χρονική διάρκεια του ενός λεπτού

πριν από κάθε συμβάν, ii) εκείνων που περιγράφουν τη χρονική διάρκεια του κάθε συμβάντος, και iii) το σύνολο αυτών, υπάρχει ένα πλήθος παραγόντων που εκφράζουν ένα σύνολο μεταβλητών, το οποίο σε κάθε περίπτωση είναι διαφορετικό.

- Τα δεδομένα που περιγράφουν την κατάσταση **ένα λεπτό πριν από κάθε συμβάν** βρέθηκε πως μπορούν να εκφραστούν με δύο παράγοντες, έναν που περιγράφει την επιρροή της ταχύτητας και της διαμήκους επιτάχυνσης και έναν που περιγράφει την επιρροή της συνολικής διανυόμενης απόστασης του οχήματος. Αυτοί οι δύο παράγοντες πιθανόν να προκύπτουν καθώς, όπως έχει παρατηρηθεί και στη διεθνή βιβλιογραφία, η ταχύτητα και η επιτάχυνση παίζουν καθοριστικό ρόλο στην πιθανότητα εμπλοκής σε κάποιο απρόοπτο περιστατικό στην οδό. Επίσης το αίσθημα κούρασης που μπορεί να έχει δημιουργηθεί στον οδηγό έπειτα από μια μεγάλη διαδρομή, ενδέχεται να αυξήσει τη πιθανότητα εμπλοκής σε συμβάν.

- Σε αντίθεση με τη χρονική διάρκεια του ενός λεπτού πριν, στη **διάρκεια ενός συμβάντος**, εκτός από την ταχύτητα που κυριαρχεί και σε αυτή την περίπτωση, οι μεταβλητές που έχουν τη μεγαλύτερη επιρροή είναι η απόκλιση του οχήματος από το μέσο της οδού και ο χρόνος μέχρι τη σύγκρουση. Η απόκλιση από το μέσο της οδού μπορεί να συμβαίνει διότι ο οδηγός κατά τη διάρκεια ενός περιστατικού ενδέχεται να χάσει τον πλήρη έλεγχο του οχήματος και ο χρόνος μέχρι τη σύγκρουση συνδέεται άμεσα με την ταχύτητα.

- Όσον αφορά στα δεδομένα που περιγράφουν την κατάσταση **ένα λεπτό πριν και κατά τη διάρκεια κάθε συμβάντος**, αυτά μπορούν να εκφραστούν με τρεις παράγοντες: έναν για την επιρροή της ταχύτητας και του χρόνου μέχρι τη σύγκρουση από το προπορευόμενο όχημα, έναν που περιγράφει την επιρροή της απόκλισης του οχήματος από το μέσο της οδού και τον παράγοντα της επιρροής της συνολικής διανυόμενης απόστασης του οχήματος στην ύπαρξη συμβάντος. Τα αποτελέσματα αυτά μπορούν να εξηγηθούν από τις επιμέρους περιπτώσεις χρονικών διαστημάτων που αναφέρθηκαν παραπάνω, καθώς αποτελούν το σύνολό τους.

6.3. Προτάσεις για βελτίωση της οδικής ασφάλειας

Ακολουθούν **προτάσεις** οι οποίες με βάση τα αποτελέσματα θα μπορούσαν να συμβάλουν στη βελτίωση της οδικής ασφάλειας.

- Χρήση του αλγόριθμου εντοπισμού συμβάντος με βάση τα οδηγικά χαρακτηριστικά ενός λεπτού πριν και κατά τη διάρκεια αυτού, από **Κέντρο Διαχείρισης Κυκλοφορίας**, όπου ο εντοπισμός συμβάντων θα ήταν αρκετά χρήσιμος για την αντιμετώπισή τους.

- Αξιολογώντας τα οδηγικά χαρακτηριστικά, να εμφανίζεται **προειδοποιητικό μήνυμα** στον οδηγό μέσα από την οθόνη διεπαφής οδηγού-οχήματος (HMI),

στην περίπτωση που γίνει αντιληπτή οδηγική συμπεριφορά που οδηγεί σε συμβάν , ώστε ο οδηγός να κάνει τις απαραίτητες διορθωτικές ενέργειες για την αποφυγή του.

- Αντίστοιχη χρήση μπορεί να έχει και ως **εφαρμογή σε έξυπνο κινητό τηλέφωνο**, που θα προειδοποιεί τον οδηγό για την ανάγκη εκτέλεσης διορθωτικών κινήσεων ώστε να αποφευχθεί το προβλεπόμενο συμβάν.

- Αξιοποίηση του αλγόριθμου σε **έξυπνα κινητά τηλέφωνα**, όπου σε περίπτωση εντοπισμού οδηγικών χαρακτηριστικών αντίστοιχων με αυτών που παρατηρήθηκαν ένα λεπτό πριν και κατά τη διάρκεια συμβάντων, να απενεργοποιούνται αυτόματα εφαρμογές ώστε να μην υπάρχει ενδεχόμενο απόσπασης προσοχής του οδηγού.

- Αξιοποίηση του αλγόριθμου σε **αυτόματα ή ημι-αυτόματα οχήματα**, όπου με κατάλληλες προσαρμογές θα επιτρέπουν τον καλύτερο εντοπισμό των πιθανών συμβάντων και την ανάληψη σχετικής δράσης.

- **Εκστρατείες ενημέρωσης** και ευαισθητοποίησης των οδηγών για τη σημαντική επιρροή των υψηλών ταχυτήτων και επιταχύνσεων καθώς και των μεγάλων διανυόμενων αποστάσεων στον κίνδυνο πρόκλησης ατυχημάτων.

6.4. Προτάσεις για περαιτέρω έρευνα

Αντικείμενο αυτής της διπλωματικής εργασίας ήταν ο εντοπισμός συμβάντων με βάση τα οδηγικά χαρακτηριστικά, σε υπεραστικές οδούς . Από αυτό και σε συνδυασμό με τη βιβλιογραφική ανασκόπηση εντοπίστηκαν ελλείψεις και απουσία εξέτασης ορισμένων θεμάτων, τα οποία προτείνονται παρακάτω για **περαιτέρω έρευνα,** που θα συνεισφέρουν στην πληρέστερη αντιμετώπιση του αντικειμένου της παρούσας Διπλωματικής Εργασίας.

- Ανάλυση οδηγικών χαρακτηριστικών σε χρονικό διάστημα **5 λεπτών πριν** από ένα συμβάν.

- Εξέταση του εντοπισμού συμβάντων χρησιμοποιώντας **δεδομένα οδήγησης σε πραγματικές συνθήκες** (naturalistic driving), αντί για στοιχεία από προσομοιωτή.

- Εξέταση **επιπλέον τύπων οδού**, εκτός από το υπεραστικό περιβάλλον που εξετάστηκε στην παρούσα εργασία, όπως είναι οι αστικές οδοί ή οι αυτοκινητόδρομοι.

- Διερεύνηση **μεγαλύτερου δείγματος** συμμετεχόντων από εκείνο που εξετάστηκε σε αυτή τη Διπλωματική Εργασία, καθώς πιθανόν η επιρροή κάποιων μεταβλητών για μεγαλύτερο δείγμα να είναι διαφορετική στη στατιστική ανάλυση.

- Δοκιμή **διαφορετικών μεθόδων στατιστικής ανάλυσης**, από εκείνες που χρησιμοποιήθηκαν, για την ανάπτυξη μοντέλων που πιθανό να επιφέρουν εξίσου σημαντικά αποτελέσματα.

- Εξέταση του εντοπισμού συμβάντων με πείραμα σε προσομοιωτή οδήγησης και σε σενάρια που περιλαμβάνουν πληθώρα **κυκλοφοριακών συνθηκών**, για παράδειγμα σε συνθήκες βροχής, ομίχλης ή χιονιού, με υψηλό ή χαμηλό φόρτο κυκλοφορίας, με απόσπαση προσοχής ή χωρίς.

- **Ανάλυση της μεταβολής των οδηγικών χαρακτηριστικών** από την κατάσταση πριν από απρόσμενο συμβάν στην κατάσταση κατά τη διάρκεια του συμβάντος.

ΒΙΒΛΙΟΓΡΑΦΙΑ

[1] Φραντζεσκάκης, Ι. Μ., Ι. Κ. Γκόλιας, "Οδική Ασφάλεια" Εκδόσεις Παπασωτηρίου, Αθήνα, 1994

[2] Monselise M., Liang S. O., Yang C. C. "Identifying Important Risk Factors Associated with Vehicle Injuries Using Driving Behavior Data and Predictive Analytics"

[3] Choudhary P., Velaga N. R., "Mobile phone use during driving: Effects on speed and effectiveness of driver compensatory behavior", Accident Analysis and Prevention 106 Indian Institute of Technology Mumbai, India, 2017

[4] Choudhary P., Velaga N. R., "Effects of phone use on driving performance: A comparative analysis of young and professional drivers", Safety Science 111 Indian Institute of Technology Mumbai, India, 2019

[5] Yannis G., Laiou A., Papantoniou P., Christoforou C., "Impact of texting on young drivers' behavior and safety on urban and rural roads through a simulation experiment", Journal of Safety Research 49 National Technical University of Athens, Greece, 2014

[6] Choudhary P., Velaga N. R., "Effects of texting on accident risk during a sudden hazardous event: Analysis of predetection and postdetection phases", Traffic Injury Prevention Indian Institute of Technology Mumbai, India, 2018

[7] Osman O., Hajij M., Karbalaieali S., Ishak S., "A hierarchical machine learning classification approach for secondary task identification from observed driving behavior data", Accident Analysis and Prevention 123 Louisiana State University LA, United States, 2019

[8] Choudhary P., Velaga N. R., "A comparative analysis of risk associated with eating, drinking and texting during driving at unsignalised intersections", Transportation Research Part F 63 Indian Institute of Technology Mumbai, India, 2019

[9] Papantoniou P., Yannis G., Christofa E., "Which factors lead to driving errors? A structural equation model analysis through a driving simulator experiment", IATSS Research 43 National Technical University of Athens, Greece, 2019

[10] Li x., Rakotonirainy A., Yan X., "How do drivers avoid collisions? A driving simulator-based study", Journal of Safety Research 70 School of Traffic and Transportation Beijing, China, 2019

[11] Papantoniou P., Pavlou D., Yannis G., Vlachogianni E., "How an unexpected incident affects speed related driving performance measures", Transportation Research Procedia 41 National Technical University of Athens, Greece, 2018

[12] Papantoniou P., Antoniou C., Yannis G., Pavlou D., "Which factors affect accident probability at unexpected incidents? A structural equation model approach", Journal of

Transportation Safety & Security National Technical University of Athens, Greece, 2019

ΠΑΡΑΡΤΗΜΑ- Κώδικας ανάλυσης στο Rstudio

```
# Φόρτωση πακέτου “dplyr” για τη χρήση της εντολής filter
install.packages("dplyr")
library(dplyr)
# Δημιουργία index (έναρξη και λήξη κάθε συμβάντος)
index<-data.frame()
k<-0
for(i in 1:nrow(RuralControl)){
   if(RuralControl[i,10]==0 & RuralControl[i+1,10]!=0){
      k<-k+1
      index[k,1]<-RuralControl[i,2]
      index[k,2]<-RuralControl[i+1,4]
   }
   if(RuralControl[i,10]!=0 & RuralControl[i+1,10]==0){
         index[k,3]<-RuralControl[i,4]
   }
}
# Δημιουργία PreEvent
PreEvent<-data.frame()
EVENTS<-data.frame()
x<-unique(index[,1])
ID<-data.frame()
f<-0
y<-0
k<-0
l<-0
for(y in x){
   a<-0
   b<-0
   ID<-filter(index,PersonID==y)
   f<-nrow(ID)
```

```
   k<-k+f
   l<-k-f+1
   for(i in l:k){
      a<-index[i,2]-60
      b<-index[i,2]
      EVENTS<-filter(RuralControl,Time>=a & Time<=b &
PersonID==y)
      PreEvent<-rbind(PreEvent,EVENTS)
   }
}
# Δημιουργία Events
Events<-data.frame()
EVENTS<-data.frame()
x<-unique(index[,1])
ID<-data.frame()
f<-0
y<-0
k<-0
l<-0
for(y in x){
   a<-0
   b<-0
   ID<-filter(index,PersonID==y)
   f<-nrow(ID)
   k<-k+f
   l<-k-f+1
   for(i in l:k){
      a<-index[i,2]-60
      b<-index[i,3]
      EVENTS<-filter(RuralControl,Time>=a & Time<=b &
PersonID==y)
      Events<-rbind(Events,EVENTS)
   }
}
```

```
# Δημιουργία DurEvent
DurEvent<-filter(Events, Event!=0)
# Drop NA της στήλης Driving Experience
events<-data.frame()
events<-Events[complete.cases(Events$`Driving
experience`), ]
Events<-events
preEvent<-data.frame()
preEvent<-PreEvent[complete.cases(PreEvent$`Driving
experience`), ]
PreEvent<-preEvent
durEvent<-data.frame()
durEvent<-DurEvent[complete.cases(DurEvent$`Driving
experience`), ]
DurEvent<-durEvent
# Αφαιρεση κοινών γραμμών
ID<-data.frame()
ID2<-data.frame()
events<-data.frame()
for(y in x){
   ID<-filter(Events,PersonID==y)
   ID2<-ID %>% distinct(Time, .keep_all = TRUE)
   events<-rbind(events,ID2)
}
Events<-events
PreEvent<-filter(Events,Event==0)
# το DurEvent παρέμεινε ίδιο
# Απομόνωση συνεχών μεταβλητών
PreEvent2<-data.frame()
PreEvent2<-
PreEvent[,c(7,8,9,12,13,14,15,16,17,18,19,20,24)]
DurEvent2<-data.frame()
DurEvent2<-
DurEvent[,c(7,8,9,12,13,14,15,16,17,18,19,20,24)]
```

```
Events2<-data.frame()
Events2<-Events[,c(7,8,9,12,13,14,15,16,17,18,19,20,24)]
#περιγραφική στατιστική
summary(PreEvent2)
sapply(PreEvent2, var, na.rm=TRUE)
sapply(PreEvent2, sd, na.rm=TRUE)
summary(DurEvent2)
sapply(DurEvent2, var, na.rm=TRUE)
sapply(DurEvent2, sd, na.rm=TRUE)
#correlation
cor(Events2)
#LOGISTIC REGRESSION
# Παραλλαγή Α
# αφαίρεση μεταβλητής Wheel από τον πίνακα DATA και
δημιουργία DATA1
DATA1<-DATA[,c(1,2,3,4,5,6,8)]
# διαχωρισμός DATA σε training (80%) και test (20%)
install.packages("caTools")
library(caTools)
split<-sample.split(DATA1$Event, SplitRatio = 0.8)
training1<-subset(DATA1, split == TRUE)
test1<- subset(DATA1, split == FALSE)
# μετατροπή της μεταβλητής Event σε διακριτή
training1$Event<-as.factor(training1$Event)
test1$Event<-as.factor(test1$Event)
# διωνυμική λογιστική παλινδρόμηση
model<-data.frame()
model<-
glm(Event~.,family=binomial(link='logit'),data=training1)
# έλεγχος συντελεστών παλινδρόμησης
summary(model)
# πρόβλεψη στον πίνακα test
fitted.results <- predict(model,test1,type='response')
fitted.results <- ifelse(fitted.results > 0.5,1,0)
```

```
misClasificError <- mean(fitted.results != test$Event)
print(paste('Accuracy',1-misClasificError))
# confusion matrix
install.packages("caret")
library(caret)
prediction<-as.data.frame(fitted.results)
prediction$fitted.results<-
factor(prediction$fitted.results)
actual<-test[,1]
actual$Event<-factor(actual$Event)
confusionMatrix(prediction$fitted.results,actual$Event,po
sitive = "1")
# δημιουργία καμπύλης ROC
install.packages("ROCR")
library(ROCR)
p <- predict(model,test1, type="response")
pr <- prediction(p, test1$Event)
prf <- performance(pr, measure = "tpr", x.measure =
"fpr")
plot(prf)
# εμβαδόν κάτω από την καμπύλη ROC
auc <- performance(pr, measure = "auc")
auc <- auc@y.values[[1]]
# FEATURE IMPORTANCE
install.packages("Boruta")
library(Boruta)
boruta_output <- Boruta(Event ~ .,
data=na.omit(training1), doTrace=0)
names(boruta_output)
boruta_signif <- getSelectedAttributes(boruta_output,
withTentative = TRUE)
print(boruta_signif)
roughFixMod <- TentativeRoughFix(boruta_output)
boruta_signif <- getSelectedAttributes(roughFixMod)
print(boruta_signif)
```

```
imps <- attStats(roughFixMod)

imps2 = imps[imps$decision != 'Rejected', c('meanImp',
'decision')]

head(imps2[order(-imps2$meanImp), ])

plot(boruta_output, cex.axis=.7, las=2, xlab="",
main="Variable Importance")

# RANDOM FOREST

install.packages("randomForest")

library(randomForest)

memory.limit()

memory.limit(100000)

# Παραλλαγή Α Random Forest

# αλλαγή του ονόματος της στήλης Driving Experience σε
DrExp

names(training1)[7] <- "DrExp"

names(test1)[7] <- "DrExp"

# τυχαία δάση – εκπαίδευση στον training1

set.seed(222)

rf <- randomForest(

   Event ~ .,

   data=training1,

   ntree=100

)

# πρόβλεψη στον πίνακα test1

p1<-predict(rf,test1)

confusionMatrix(p1,testing$Event,positive = "1")

# τέλος παραλλαγής Α

# Παραλλαγή Β Random Forest

# δημιουργία DATA2 από DATA

DATA2<-DATA[,c(1,2,3,5,8)]

# διαχωρισμός DATA σε training (80%) και test (20%)

split<-sample.split(DATA2$Event, SplitRatio = 0.8)

training2<-subset(DATA2, split == TRUE)

test2<- subset(DATA2, split == FALSE)
```

```
# μετατροπή της μεταβλητής Event σε διακριτή
training2$Event<-as.factor(training2$Event)
test2$Event<-as.factor(test2$Event)
# αλλαγή του ονόματος της στήλης Driving Experience σε
DrExp
names(training2)[5] <- "DrExp"
names(test2)[5] <- "DrExp"
# τυχαία δάση – εκπαίδευση στον training2
set.seed(222)
rf <- randomForest(
   Event ~ .,
   data=training2,
   ntree=100
)
print(rf)
attributes(rf)
plot(rf)
# πρόβλεψη στον πίνακα test2
p2<-predict(rf,test2)
confusionMatrix(p2,test2$Event,positive = "1")
# τέλος παραλλαγής A
# FACTOR ANALYSIS
install.packages("REdaS")
library(REdaS)
install.packages("grid")
# PRE-EVENT
PreEvent3<-PreEvent2[,c(1,3,5,9,10,11)]
summary(PreEvent3)
cor(PreEvent3)
# κυριες συνιστώσες για την εκτίμηση του αριθμού των
παραγόντων PCA
pca1<-princomp(PreEvent3,scores = TRUE,cor = TRUE)
summary(pca1)
loadings(pca1)
```

```
plot(pca1)
screeplot(pca1,type="line",main = "Scree Plot")
pca1$scores[1:10,]
#FACTOR ANALYSIS
fa1<-factanal(PreEvent3,factors = 3)
fa1
#KMO Test
KMOS(PreEvent3)
# Bartlett Test
bart_spher(PreEvent3, use = c("everything", "all.obs",
"complete.obs",
                          "na.or.complete",
"pairwise.complete.obs"))
# DUR-EVENT
DurEvent3<-DurEvent2[,c(1,3,5,9,10,11)]
summary(DurEvent3)
cor(DurEvent3)
# κυριες συνιστώσες για την εκτίμηση του αριθμού των
παραγόντων PCA
pca2<-princomp(DurEvent3,scores = TRUE,cor = TRUE)
summary(pca2)
loadings(pca2)
plot(pca2)
screeplot(pca2,type="line",main = "Scree Plot")
pca2$scores[1:10,]
#FACTOR ANALYSIS
fa2<-factanal(DurEvent3,factors = 3,rotation =
"varimax",scores = "regression")
fa2
#KMO Test
KMOS(DurEvent3)
# Bartlett Test
bart_spher(DurEvent3, use = c("everything", "all.obs",
"complete.obs",
                          "na.or.complete",
"pairwise.complete.obs"))
```

```
# EVENTS
Events3<-Events2[,c(1,3,5,9,10,11)]
summary(Events3)
cor(Events3)
# κυριες συνιστώσες για την εκτίμηση του αριθμού των παραγόντων PCA
pca3<-princomp(Events3,scores = TRUE,cor = TRUE)
summary(pca3)
loadings(pca3)
plot(pca3)
screeplot(pca3,type="line",main = "Scree Plot")
pca3$scores[1:10,]
#FACTOR ANALYSIS
fa3<-factanal(Events3,factors = 3)
fa3
#KMO Test
KMOS(Events3)
# Bartlett Test
bart_spher(Events3, use = c("everything", "all.obs", "complete.obs",
                        "na.or.complete", "pairwise.complete.obs"))
```