



ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
Σχολή Πολιτικών Μηχανικών
Τομέας Μεταφορών και Συγκοινωνιακής Υποδομής

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

Κωστόπουλος Αντώνης

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

Αθήνα, Νοέμβριος 2022

Στόχος Διπλωματικής Εργασίας

Ο εντοπισμός, ταξινόμηση και πρόβλεψη της **επικίνδυνης** οδικής συμπεριφοράς και των **απότομων** περιστατικών μέσω Μάθησης ανισόρροπων δεδομένων και αλγορίθμων ταξινόμησης.

Για την επίτευξη του **στόχου** πραγματοποιείται:

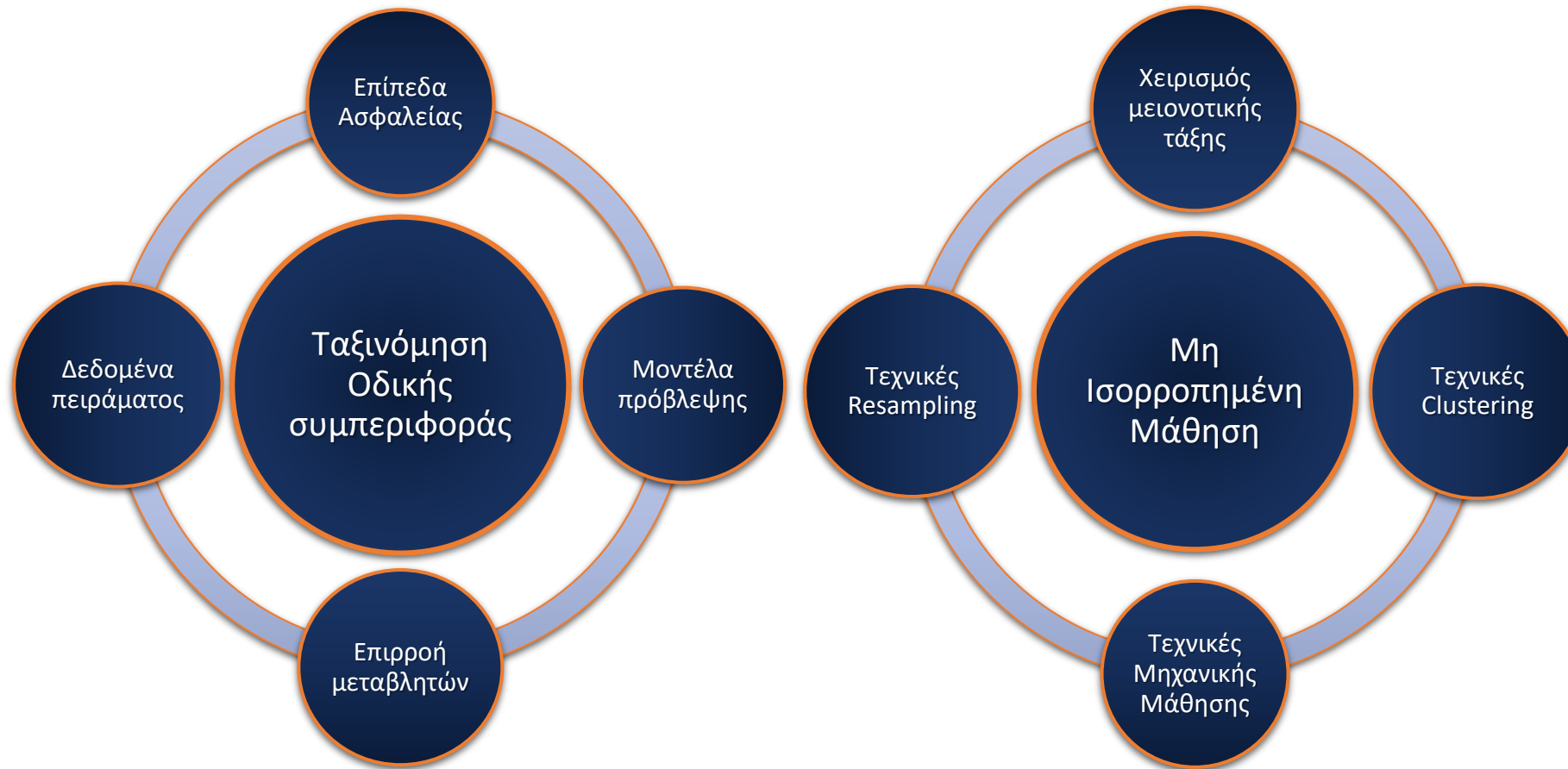
- Ανάπτυξη αλγορίθμων για την εύρεση του **βαθμού επιρροής** των οδικών δεδομένων στην επικίνδυνη συμπεριφορά
- Ανάπτυξη αλγορίθμων ομαδοποίησης για τον **διαχωρισμό** των επιπέδων ασφαλείας.
- Ανάπτυξη αλγορίθμων ταξινόμησης, για τον εντοπισμό του **επιπέδου ασφαλείας** που ανήκουν τα δεδομένα



Μεθοδολογία



Βιβλιογραφική Ανασκόπηση



Ελλείψεις ερευνών:

- Περιορισμένος **όγκος** δεδομένων
- Δεν αναλύθηκαν μεταβλητές **ορίων ταχύτητας**
- **Πειραματικά** δεδομένα
- Περιορισμένος **αριθμός** μοντέλων

Θεωρητικό Υπόβαθρο (1/4)





Θεωρητικό Υπόβαθρο (2/4)

Αλγόριθμοι Παλινδρόμησης και Ταξινόμησης

- Γραμμική Παλινδρόμηση
- Decision Trees
- Gradient Boosting
- XGBoost
- Random Forests
- AdaBoost
- K-nearest Neighbors
- Supported Vector Machines
- Multilayered Perceptrons NN

Τεχνικές Μεταχείρισης Δεδομένων

- Κανονικοποίηση
- Ομαδοποίηση με αλγόριθμο K-μέσου
- Τεχνικές Oversampling
- GridSearch



Θεωρητικό Υπόβαθρο (3/4)



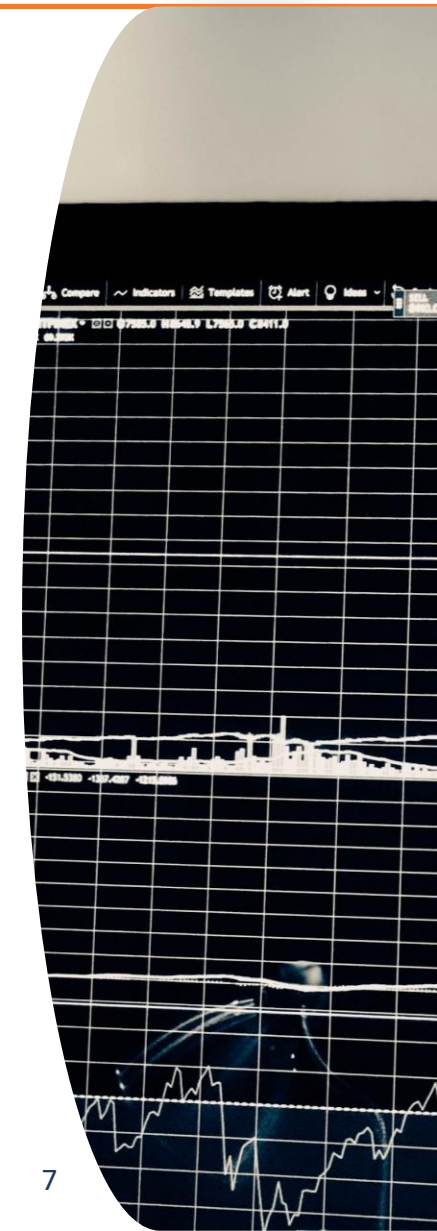
Αξιολόγηση μοντέλων
Μηχανικής Μάθησης

Κριτήριο αποδοχής
μοντέλων
παλινδρόμησης

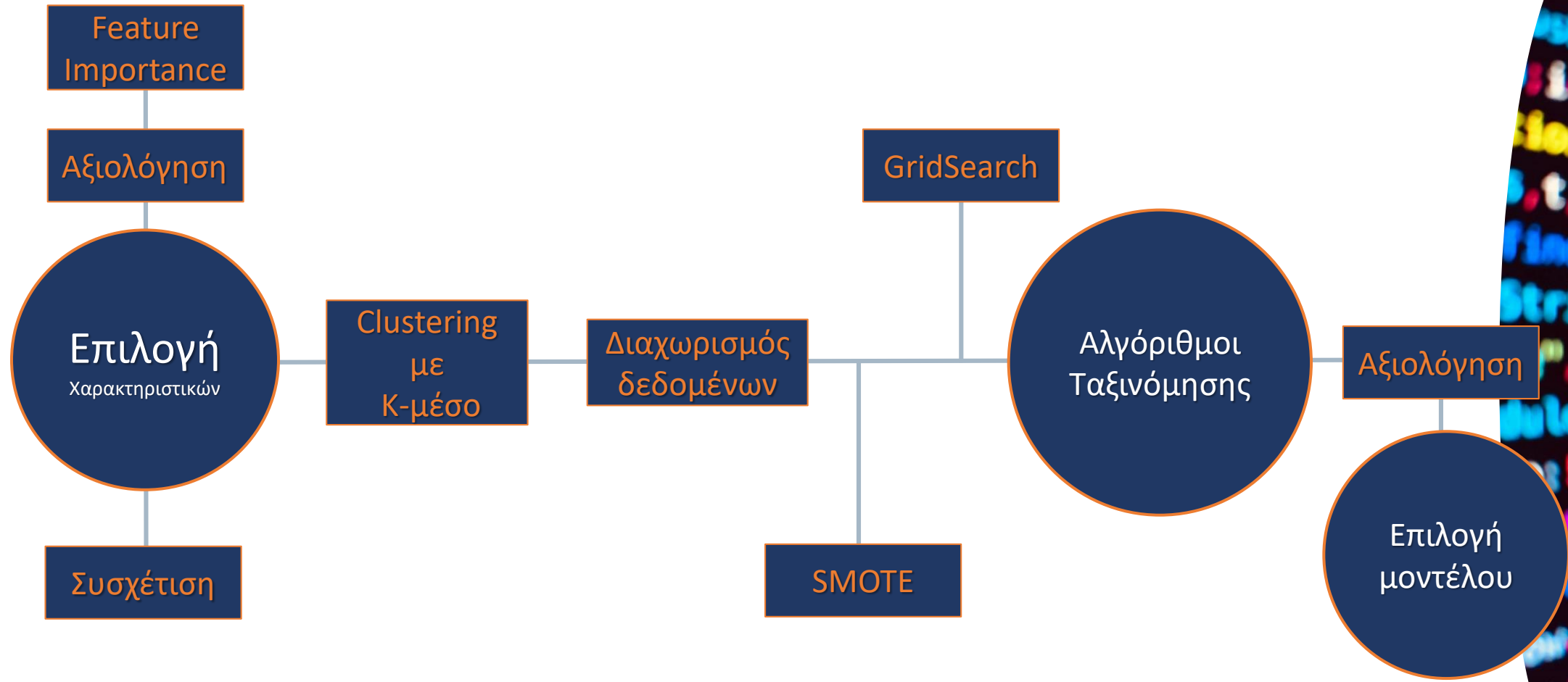
➤ R^2

Μετρικές αξιολόγησης
μοντέλων ταξινόμησης

- Ορθότητα
- Ακρίβεια
- Ανάκληση
- f-1 score
- False Negative Rate
- AUC score



Θεωρητικό Υπόβαθρο (4/4)



Συλλογή των δεδομένων



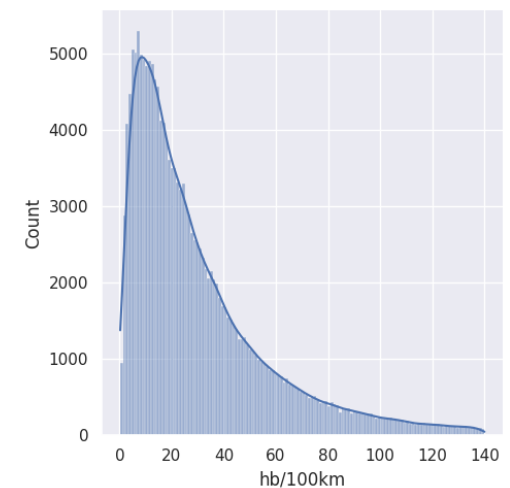
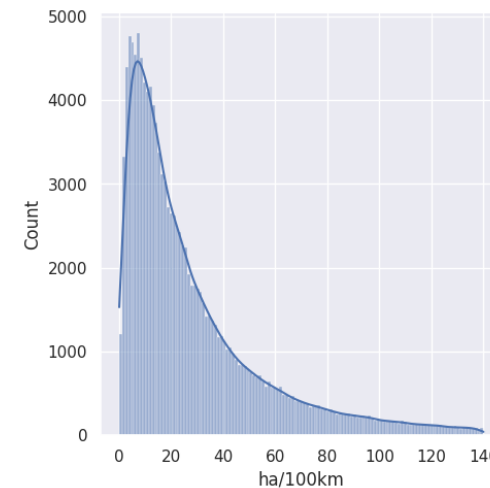
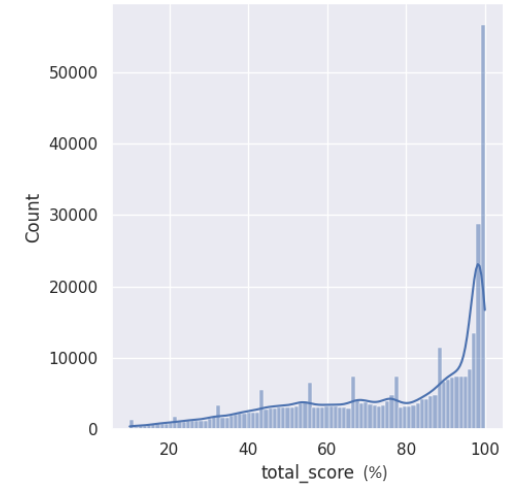
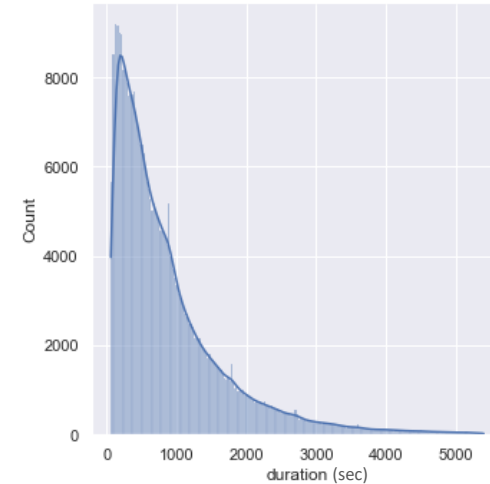
- Τα δεδομένα συλλέχθηκαν από τις βάσεις δεδομένων της εταιρίας **OSeven**®.
- Δεδομένα **τηλεματικής** μέσω ειδικά σχεδιασμένης εφαρμογής smartphone.
- **356.162** διαφορετικές οδικές διαδρομές.
- Η συλλογή πραγματοποιήθηκε κατά τη διάρκεια της **πανδημίας** SARS CoV-2.
- Από τα στοιχεία παράχθηκαν 75 **δείκτες** οδικής συμπεριφοράς, φόρτου κυκλοφορίας και περιγραφικών στοιχείων διαδρομής.



Περιγραφική Στατιστική



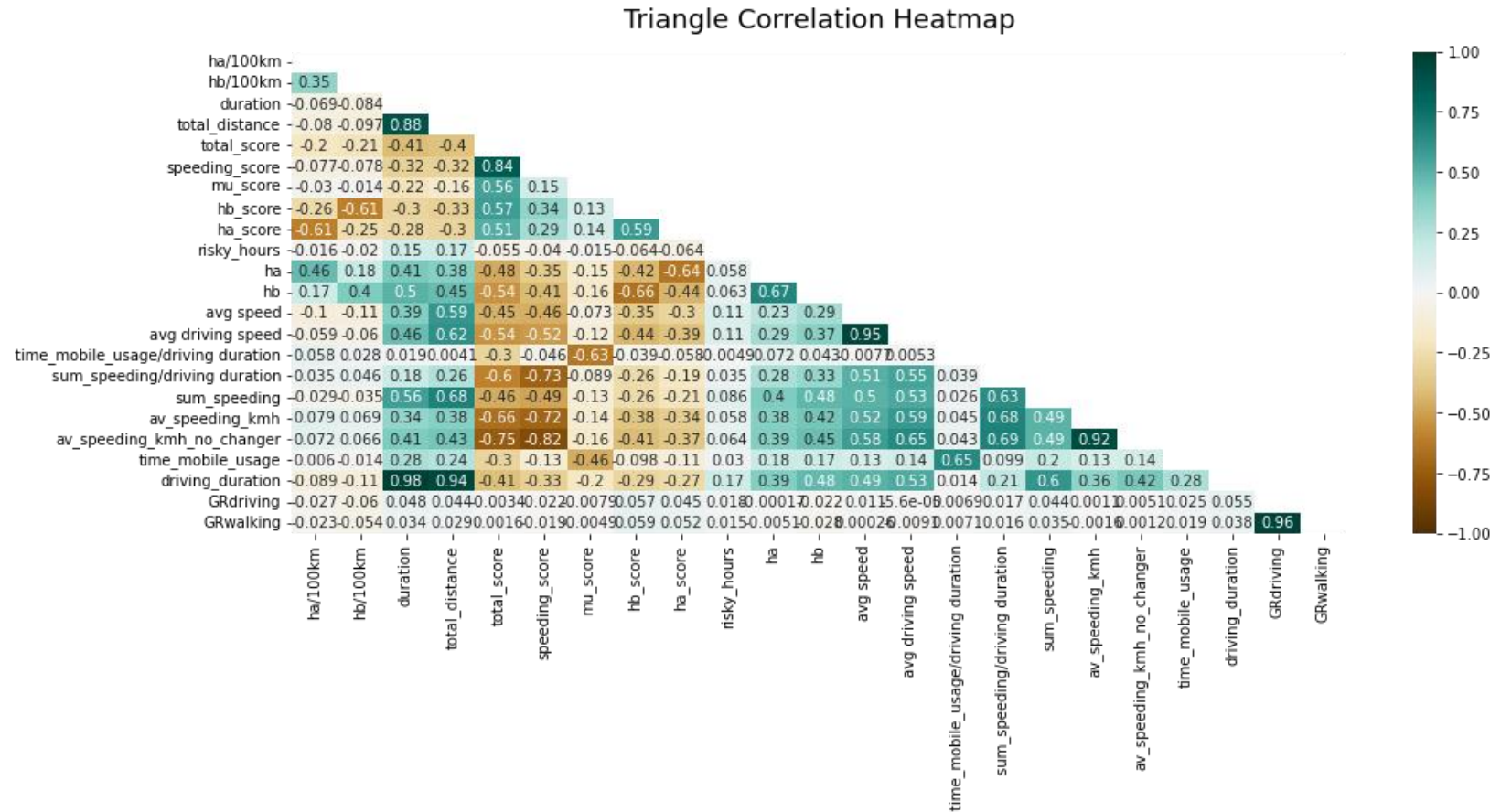
- Μέση **διάρκεια** διαδρομής ήταν 16.03 λεπτά.
- Οι διάμεσοι των απότομων περιστατικών ήταν **μηδενικοί**.
- **Ασύμμετρη** κατανομή των απότομων περιστατικών λόγω υψηλού μέτρου κύρτωσης.
- Ο δείκτης **συνολικού** σκορ διαδρομής έχει διάμεσο 84.00/100.



Συσχέτιση Pearson



- **Πολύ υψηλή** συσχέτιση μεταξύ απότομων περιστατικών και συνολικής διάρκειας διαδρομής και διανυθείσας απόστασης.
- **Υψηλή** συσχέτιση μεταξύ απότομων περιστατικών και χρήσης κινητού τηλεφώνου στην διαδρομή.
- **Μέτρια** συσχέτιση με φόρτους οδηγών Ι.Χ. και πεζών.
- **Χαμηλή** συσχέτιση με την οδήγηση σε επικίνδυνη ζώνη ώρας (00:00-05:00)



Σημαντικότητα Χαρακτηριστικών



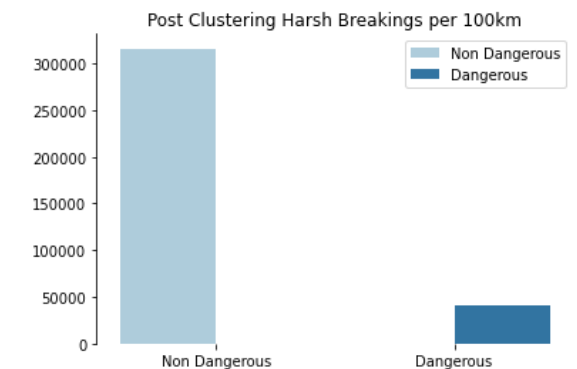
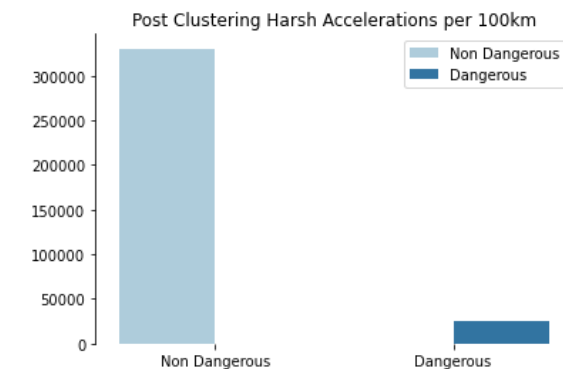
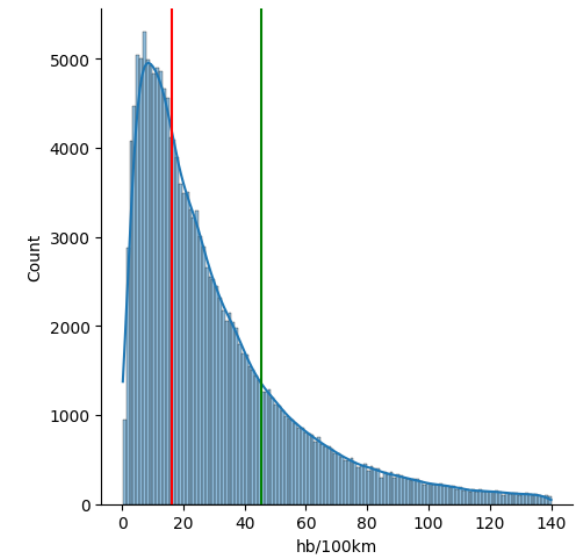
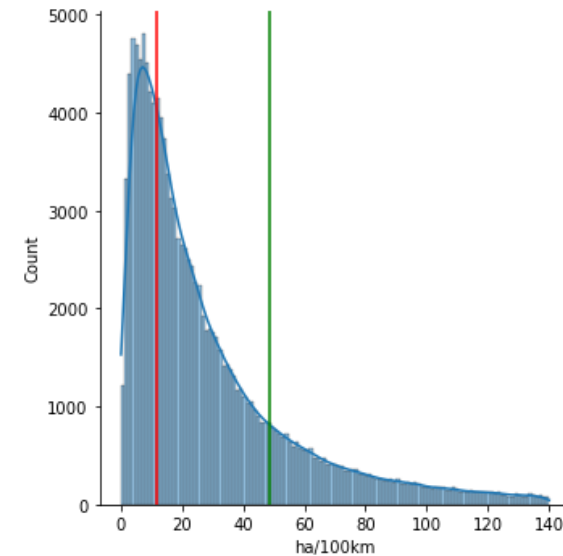
- Δημιουργία **τριών** υποσυνόλων: ενός με τις **ανεξάρτητες** μεταβλητές και δύο με τις **εξαρτημένες**
- Ανάπτυξη μοντέλων **Παλινδρόμησης** για κάθε εξαρτημένη μεταβλητή: Γραμμική Παλινδρόμηση, Decision Trees, XGBoost, Random Forests, Linear SVR
- Υπολογισμός **R^2** κάθε Παλινδρόμησης → Decision Trees = 0.999, Linear Regression: $h_a/100 = 0.073$, $h_b/100 = 0.077$
- **Σημαντικότερες** μεταβλητές για την εμφάνιση απότομων περιστατικών βάσει **συσχέτισης** και **σημαντικότητας**:
 - (1) Συνολική διανυθείσα απόσταση
 - (2) Διάρκεια οδήγησης εν κινήσει
 - (3) Μέση ταχύτητα εν κινήσει
 - (4) Σκορ υπερβολικής ταχύτητας
 - (5) Σκορ χρήσης κινητού τηλεφώνου



Προεπεξεργασία για Ταξινόμηση



- Κανονικοποίηση μεταβλητών εισόδου
- Δύο κλάσεις ταξινόμησης: **Επικίνδυνη** συμπεριφορά και **Μη** Επικίνδυνη συμπεριφορά
- Για την εύρεση άνω και κάτω ορίων ταξινόμησης εφαρμόστηκε ο αλγόριθμος **K-μέσου**
- Επίπεδο “**Dangerous**”: Απότομες επιταχύνσεις ανά 100χλμ. > **48.82**
- Επίπεδο “**Dangerous**”: Απότομες επιβραδύνσεις ανά 100χλμ. > **45.40**



Διαχείριση Μη Ισορροπημένης Μάθησης

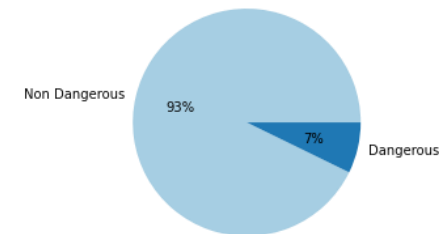


- Διαχωρισμός των δεδομένων σε δεδομένα **εκπαίδευσης** και δεδομένα **εξέτασης** με αναλογία 75%-25%
- **Ανισορροπία** δεδομένων εκπαίδευσης της μειονοτικής τάξης
- Επιλογή μεθόδου **Υπερδειγματοληψίας**, για να μην χαθεί πολύτιμη πληροφορία λόγω υψηλής διακύμανσης
- Μέθοδος Υπερδειγματοληψίας **SMOTE**

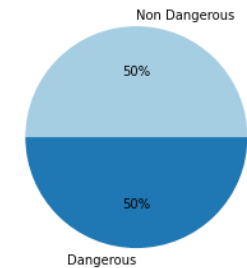
Synthetic Minority Oversampling Technique



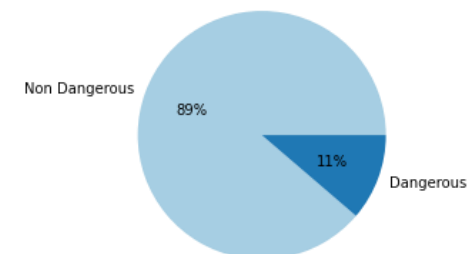
Post clustering percentage of Harsh Accelerations per 100km



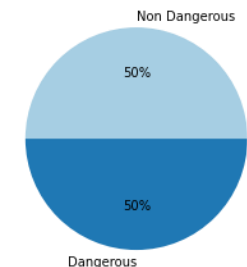
Post Oversampling Class percentage of Harsh Accelerations per 100km



Post clustering percentage of Harsh Breakings per 100km



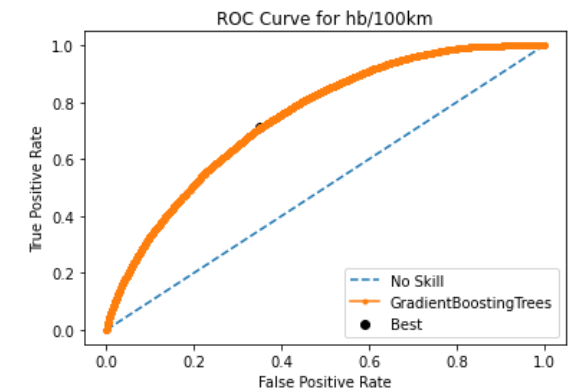
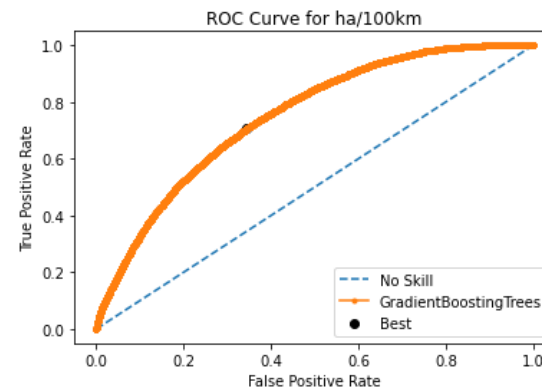
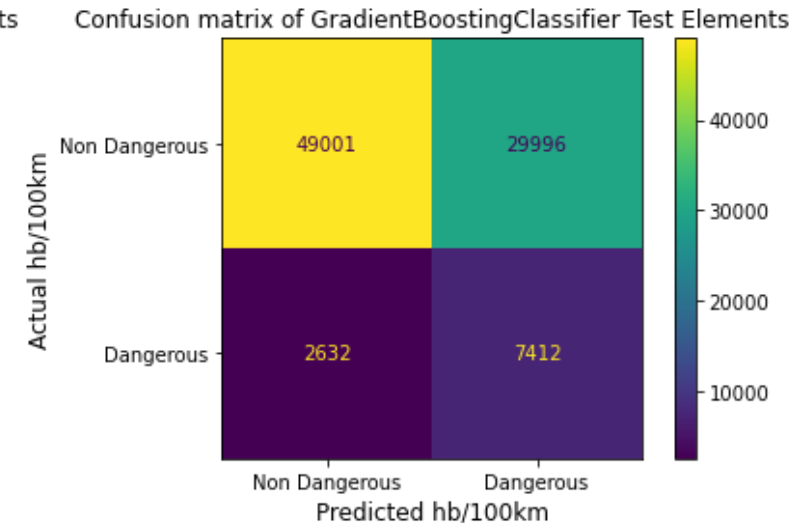
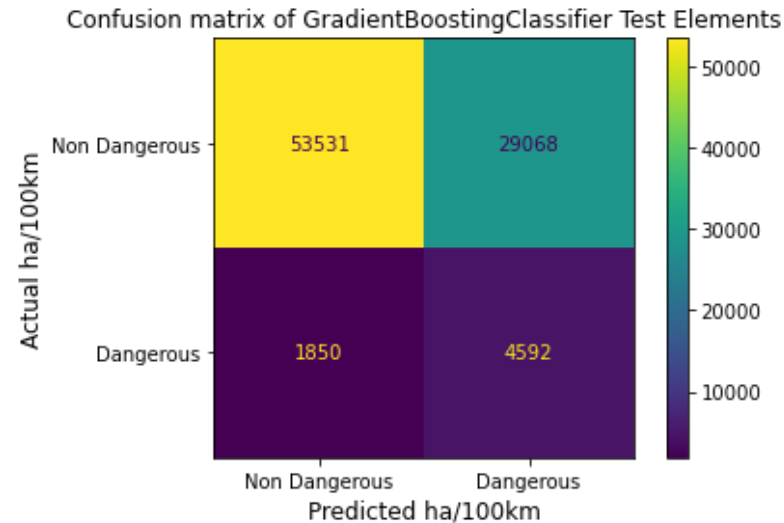
Post Oversampling Class percentage of Harsh Breakings per 100km



Διαδικασία Ταξινόμησης (1/2)



- Ανάπτυξη 8 αλγορίθμων ταξινόμησης και για τις δύο μεταβλητές: Decision Trees, Gradient Boosting, XGBoost, Random Forests, AdaBoost, K-nearest Neighbors, SVM, Multilayered Perceptrons NN
- Αξιοποίηση των μετρικών δεικτών για την αξιολόγηση της προβλεπτικής ικανότητας
- Καλύτερες επιδόσεις τα Gradient Boosting και μετά τα MLPs

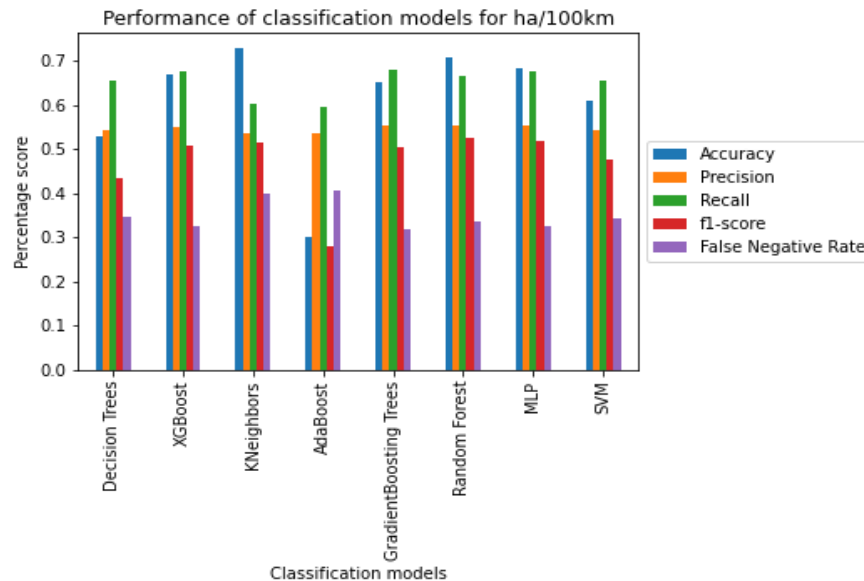


Διαδικασία Ταξινόμησης (2/2)



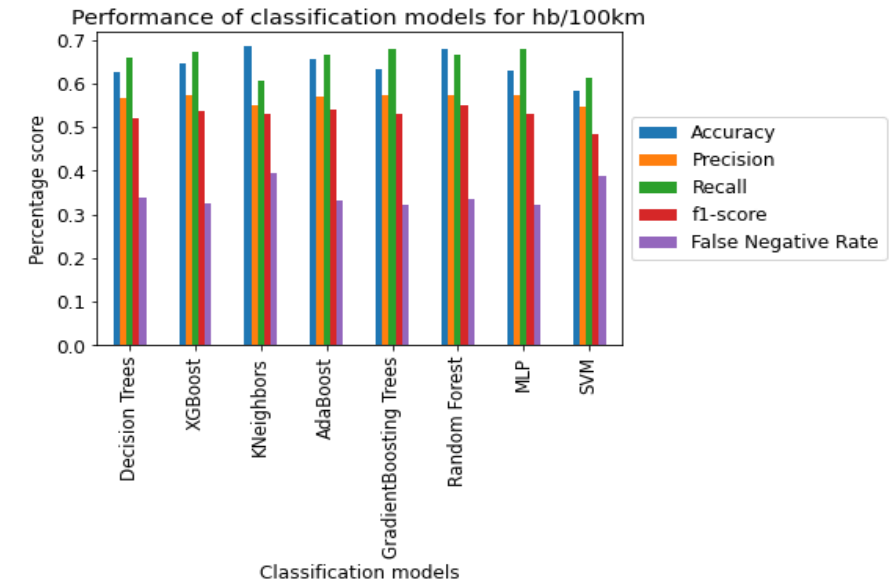
Απότομες **επιταχύνσεις** ανά 100χλμ.

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



Απότομες **επιβραδύνσεις** ανά 100χλμ.

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



Συμπεράσματα (1/2)



- Η **συνολική διανυθείσα απόσταση** είναι η πιο **σημαντική** μεταβλητή για την εμφάνιση Επικίνδυνης συμπεριφοράς του οδηγού.
- Η **μέση ταχύτητα** της οδήγησης και η διάρκεια της κρίνονται **πολύ** επιδραστικές μεταβλητές.
- Οι οδικές συμπεριφορές **δεν** είναι πολύ επιθετικές και τα απότομα περιστατικά εμφανίζονται σε διαδρομές **επιβαρυσμένες** σε χρόνο και διαδρομή.
- Η οδήγηση την νυχτερινή ζώνη ώρας **δεν** επιδρά αρνητικά στην συμπεριφορά του οδηγού.



Συμπεράσματα (2/2)

- Η μέθοδος Υπερδειγματοληψίας **SMOTE** αποδείχτηκε η πιο αποτελεσματική τεχνική, λόγω του έντονου φαινομένου **υπερπροσαρμογής** (Overfitting) χωρίς αυτή.
- Οι αλγόριθμοι **Gradient Boosting** και **Multilayered Perceptrons NN** έχουν την αποδοτικότερη προβλεπτική ικανότητα.
- Η μέθοδος ομαδοποίησης με αλγόριθμο **K-means** κατέστησε τις κλάσεις **πολύ** καλά διακριτές, καταδεικνύοντας επικίνδυνη συμπεριφορά για απότομες **επιταχύνσεις** άνω των **48** και **επιβραδύνσεις** άνω των **45** ανά 100χλμ.
- Οι Παλινδρομήσεις μορφής **Δέντρου** (**CART**) ήταν πολύ ανώτερες από τις **Γραμμικές**, επιβεβαιώνοντας την **απουσία** γραμμικότητας των δεδομένων.



Προτάσεις για αξιοποίηση των αποτελεσμάτων



- **Αξιοποίηση** των διακριθέντων μοντέλων ταξινόμησης και πρόβλεψης για την **αναγνώριση** της συμπεριφοράς του οδηγού από δεδομένα **φυσικής** οδήγησης.
- Η βελτίωση εφαρμογής κινητών τηλεφώνων για τον **εντοπισμό** απότομων περιστατικών σε πραγματικό χρόνο, βάσει της οδικής συμπεριφοράς.
- Η ενημέρωση των χρηστών της οδού σε πραγματικό χρόνο, εάν εντοπιστεί **υπέρβαση** του ορίου απότομων περιστατικών.
- Ανάπτυξη συστημάτων δημιουργίας **προφίλ οδηγών**, ανάλογα με τον αριθμό απότομων περιστατικών.

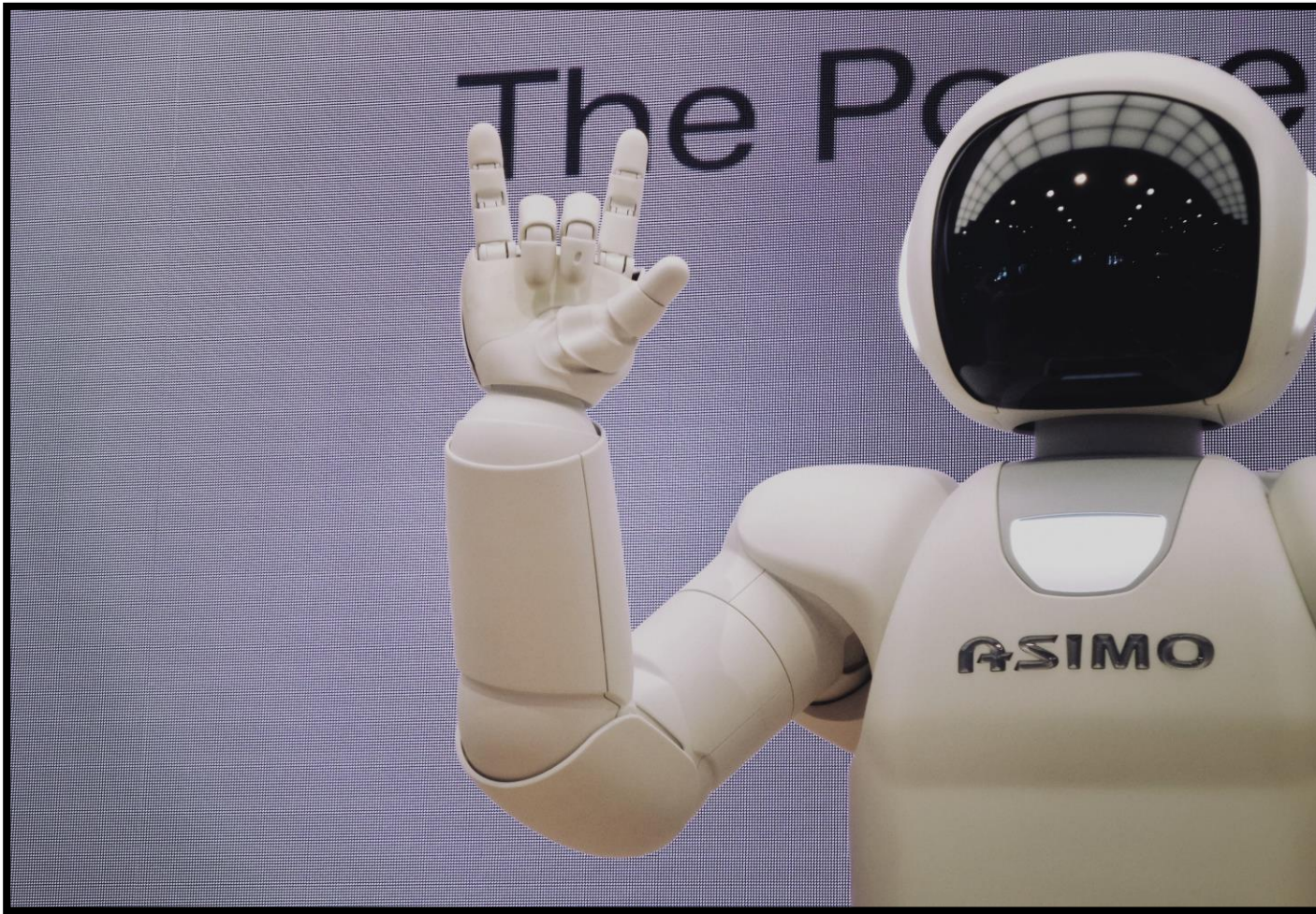


Προτάσεις για περαιτέρω έρευνα



- Εξέταση δημογραφικών, ψυχογραφικών, γεωμετρικών και παραπάνω οδικών δεδομένων.
- Έλεγχος επάρκειας δείγματος στοιχείων για την ομαδοποίηση των δεικτών κάθε οδηγού και την κατασκευή ολοκληρωμένων οδικών προφίλ.
- Αξιολόγηση σημαντικότητας με μεθόδους πολυπαραγοντικών κριτηρίων, όπως Bayes I.C. και Akaike I.C.
- Διάκριση σε περισσότερες κλάσεις ταξινόμησης.
- Εναλλακτικός τρόπος clustering, όπως Gaussian Mixture Models.
- Εξέταση περισσότερων μοντέλων Deep Learning, όπως LSTM.





Ευχαριστώ 😊 ❤️
για την προσοχή σας!



ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
Σχολή Πολιτικών Μηχανικών
Τομέας Μεταφορών και Συγκοινωνιακής Υποδομής

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

- Κωστόπουλος Αντώνης

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

Αθήνα, Νοέμβριος 2022