ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

ΤΟΜΕΑΣ ΤΕΧΝΟΛΟΓΙΑΣ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΥΠΟΛΟΓΙΣΤΩΝ

Διάγνωση ασθενών με νευρολογικές παθήσεις μέσω οδηγικής συμπεριφοράς χρησιμοποιώντας τεχνικές μηχανικής μάθησης

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

του

ΙΑΤΡΟΠΟΥΛΟΥ ΠΕΤΡΟΥ

Επιβλέπων : Κοζύρης Νεκτάριος
Καθηγητής Ε.Μ.Π.

Αθήνα, Οκτώβριος 2017

ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ
ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ
ΤΟΜΕΑΣ ΤΕΧΝΟΛΟΓΙΑΣ ΠΛΗΡΟΦΟΡΙΚΗΣ
ΚΑΙ ΥΠΟΛΟΓΙΣΤΩΝ

Διάγνωση ασθενών με νευρολογικές παθήσεις μέσω οδηγικής συμπεριφοράς χρησιμοποιώντας τεχνικές μηχανικής μάθησης

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

του

ΙΑΤΡΟΠΟΥΛΟΥ ΠΕΤΡΟΥ

Επιβλέπων : Κοζύρης Νεκτάριος
Καθηγητής Ε.Μ.Π.

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 31η Οκτωβρίου 2017.

.....	.....	.....
Κοζύρης Νεκτάριος, Καθηγητής Ε.Μ.Π.	Παπασπύρου Νικόλαος, Αναπληρωτής Καθηγητής Ε.Μ.Π.	Γκούμας Γεώργιος, Επίκουρος Καθηγητής Ε.Μ.Π.

Αθήνα, Οκτώβριος 2017

.....

ΙΑΤΡΟΠΟΥΛΟΣ ΠΕΤΡΟΣ

Διπλωματούχος Ηλεκτρολόγος Μηχανικός και Μηχανικός Υπολογιστών Ε.Μ.Π.

Περίληψη

Η οδήγηση είναι μια πολύπλοκη διαδικασία η οποία πραγματοποιείται από εκατομμύρια ανθρώπους κάθε μέρα. Ο χαρακτηρισμός των οδηγικών συμπεριφορών από δεδομένα καταγεγραμμένα από αισθητήρες δεν είναι μόνο μια ενδιαφέρουσα επιστημονική έρευνα αλλά και μια απαίτηση του πραγματικού κόσμου. Ο σκοπός της παρούσας διπλωματικής είναι η σχεδίαση και πρόταση ενός συστήματος μηχανικής μάθησης, το οποίο εντοπίζει την οδηγική συμπεριφορά ανθρώπων που πάσχουν από νευροεκφυλιστικές ασθένειες.

Συγκεκριμένα, το πρόβλημα διάγνωσης ενός οδηγού ανάγεται στο πρόβλημα ταξινόμησης χρονοσειρών. Αφού συλλέχθηκαν τα δεδομένα από τον προσομοιωτή και καθαρίστηκαν, τρία μοντέλα εκπαιδεύτηκαν και αξιολογήθηκαν. Το πρώτο μοντέλο ήταν k-nearest neighbors με αλγόριθμο δυναμικής χρονικής περιδίνησης (dynamic time warping) υπολογισμού απόστασης, το δεύτερο ήταν ένα multilayer perceptron και το τρίτο ένα decision tree.

Τα αποτελέσματα έδειξαν ότι είναι εφικτό να διαγνωστεί ένα άτομο από την οδηγική του συμπεριφορά φτιάχνοντας το κατάλληλο μοντέλο. Οι προτεινόμενες τεχνικές μπορούν να γενικευτούν σε δεδομένα από τον πραγματικό κόσμο.

Λέξεις Κλειδιά: Ανάλυση Δεδομένων, Μηχανική Μάθηση, Κατηγοριοποίηση Χρονοσειρών, Επιβλεπόμενη μάθηση, Συμπεριφορική Ανάλυση, k-Κοντινότεροι Γείτονες, Νευρωνικά Δίκτυα, Δέντρο Απόφασης, Δυναμική Στρέβλωση Χρόνου, Κινούμενος Μέσος Όρος, Ανάλυση Κυρίων Συνιστωσών

Abstract

Driving is a complex action performed by millions of people every day. Characterizing driving styles from sensory data is not only an interesting scientific research but also a real world requirement. The scope of this thesis is to design and propose a machine learning system architecture which analyzes and detects the driving behavior of people suffering from neurodegenerative diseases.

Specifically, the problem of a driver's diagnosis is reduced to time series classification problem. After collecting the data from a simulator and cleaning them, three models were trained and evaluated. The first model is a k-nearest neighbors classifier with dynamic time warping distance algorithm, the second is a multilayer perceptron and the third is a decision tree.

The results showed that it is feasible to diagnose a driver from his/her driving behavior by building the appropriate model. The proposed techniques could be generalized for data from the real world.

Keywords: Data Analysis, Machine Learning, Time Series Classification, Supervised Learning, Behavioral Analysis, k-Nearest Neighbors, Neural Networks, Decision Tree, Dynamic Time Warping, Moving Average, Principal Component Analysis

Contents

1

Εισαγωγή

1.1 Επιστήμη των Δεδομένων

1.1.1 Επιστήμη Δεδομένων και Συμπεριφορική Ανάλυση

Τα τελευταία έτη παρατηρείται μια εκθετική αύξηση στην ποσότητα των παραγόμενων ψηφιακών δεδομένων. Η αύξηση αυτή είναι τόσο μεγάλη, που πλέον οι συμβατικοί τρόποι επεξεργασίας και ανάλυσης αποδεικνύονται ανεπαρκείς. Ως λογικό επακόλουθο, εμφανίζονται ανάγκες στη διαχείριση τους σε τομείς όπως η καταγραφή, η αποθήκευση, ο διαμοιρασμός, η ανάλυση, η μεταφορά, η επεξεργασία τους, καθώς και η εμπιστευτικότητα των δεδομένων αυτών.

Παράλληλα με την αυξητική τάση που παρουσιάζει η ποσότητα των δεδομένων, αντίστοιχη τάση έχει και η πολυπλοκότητα τους αλλά και η ποικιλομορφία τους. Η κατάσταση αυτή επιβαρύνει ακόμα περισσότερο τα υπολογιστικά συστήματα επεξεργασίας και ανάλυσης, τα οποία καλούνται να μεγιστοποιήσουν την απόδοση τους στην αξιοπιστία των αποτελεσμάτων αλλά και στην εκμετάλλευση των διαθέσιμων πόρων.

Ο όρος *«Μεγάλα Δεδομένα»* (*Big Data)* χρησιμοποιήθηκε και χρησιμοποιείται για να περιγράψει τα δεδομένα – δομημένα και αδόμητα – των οποίων ο όγκος ξεπερνά τις δυνατότητας των κοινά

χρησιμοποιούμενων λογισμικών, να τα καταγράψουν, να τα διαχειριστούν και να τα επεξεργαστούν σε 'βιώσιμο' χρόνο. Η συλλογή δεδομένων γίνεται με καταιγιστικό ρυθμό πλέον από σχετικά φθηνές συσκευές οι οποίες έχουν μόνιμη πρόσβαση στο διαδίκτυο *(Internet of Things – IoT)* όπως smartphones, αισθητήρες, software logs και κάμερες. Πλέον, οι συμβατικές σχεσιακές βάσεις δεδομένων αδυνατούν να διαχειριστούν τα δεδομένα αυτά καθώς είναι απαραίτητη η μαζικά παράλληλη επεξεργασία σε μεγάλο αριθμό από servers [3].

Αναδεικνύεται, επομένως, μια τεράστια πρόκληση όσον αφορά στην αντιμετώπιση της κατάστασης αυτής και στον εντοπισμό και επίλυση των πιθανών προβλημάτων που αυτή προκαλεί. Στόχος είναι η όσο το δυνατόν εντατικότερη και πιο εποικοδομητική επεξεργασία των δεδομένων αυτών, για την εξαγωγή νέας γνώσης και πληροφορίας. Με τη διαδικασία αυτή, μπορεί με τη σειρά τους να παραχθούν νέα δεδομένα προς επεξεργασία, οπότε μπορεί να επαναλαμβάνεται και κυκλικά.

Η υπολογιστική διαδικασία της ανακάλυψης προτύπων σε μεγάλα σύνολα δεδομένων χρησιμοποιώντας μεθόδους μηχανικής μάθησης, στατιστικής και βάσεων δεδομένων αναφέρεται στη βιβλιογραφία ως *Data Mining* [1][2]. Ο όρος αυτός, ωστόσο, είναι εν μέρει παραπλανητικός και μπορεί να αποπροσανατολίσει από την πραγματική ερμηνεία, η οποία είναι η παραγωγή γνώσης από τα υπάρχοντα δεδομένα και όχι η εξόρυξη δεδομένων αυτή καθ' αυτή. Παράλληλα αποτελεί και όρο εντυπωσιασμού (buzzword) που χρησιμοποιείται για να περιγράψει οποιοδήποτε είδος επεξεργασίας πληροφορίας μεγάλης κλίμακας, καθώς και οποιαδήποτε εφαρμογή βοηθητικού συστήματος υπολογιστικών αποφάσεων, συμπεριλαμβανομένης της τεχνικής νοημοσύνης, της μηχανικής μάθησης και του business intelligence.

1.1.2 Μηχανική Μάθηση

Η ανάγκη για *Εξόρυξη Δεδομένων* και εξαγωγή πληροφορίας οδήγησε στην έντονη προσπάθεια για ανάπτυξη μεθόδων, τεχνικών και αλγορίθμων, οι οποίοι θα επεξεργάζονται τα δεδομένα αυτά και θα παράγουν νέα γνώση.

Η προσπάθεια αυτή γέννησε ένα νέο κλάδο της επιστήμης των υπολογιστών, ο οποίος είναι ευρέως γνωστός με το όνομα *Μηχανική Μάθηση (Machine Learning)* [4][5]. Ως μηχανική μάθηση ορίζεται η δυνατότητα των υπολογιστικών μηχανών να έχουν την δυνατότητα να μάθουν χωρίς να προγραμματιστούν ρητά. Η θεωρία αποτελεί εξέλιξη των ερευνών πάνω σε αναγνώριση προτύπων και υπολογιστική θεωρία μάθησης στην τεχνητή νοημοσύνη.

Η μηχανική μάθηση είναι ένας τομέας ο οποίος ασχολείται με την μελέτη και την κατασκευή αλγορίθμων που μπορούν να μάθουν και τελικά, να κάνουν προβλέψεις πάνω σε δεδομένα. Αυτοί οι αλγόριθμοι δε βασίζονται σε στατικά γραμμένες εντολές κώδικα αλλά παρουσιάζουν μια δεδομενοκεντρική συμπεριφορά για τη λήψη αποφάσεων και την εκτέλεση προβλέψεων, δημιουργώντας κατάλληλα μοντέλα από δειγματικά δεδομένα.

Όπως αναφέρθηκε προηγουμένως, τα σύνολα δεδομένων τα οποία υπόκεινται σε επεξεργασία τείνουν να γίνονται όλο και πιο πολύπλοκα και να περιέχουν υψηλό αριθμό χαρακτηριστικών. Είναι πολύ σημαντικό να καθοριστεί πριν την προσπάθεια κατασκευής ενός μοντέλου μηχανική μάθησης ποια χαρακτηριστικά θα χρησιμοποιηθούν. Η σωστή επιλογή χαρακτηριστικών είναι κομβική για την αποτελεσματικότητα του μοντέλου αφού μπορεί να επηρεάσει την χρονική διάρκεια εκπαίδευσης του, να αποτραπεί η *«κατάρα της διαστατικότητας»* (*curse of dimensionality)* και να αποφευχθεί το υπερπροσαρμογή (overfitting) (στα δεδομένα εκπαίδευσης και απώλεια ικανότητας γενίκευσης). Η επιλογή αυτή μπορεί να γίνει είτε με χειροκίνητο τρόπο, αναλύοντας το πρόβλημα είτε με αυτόματο κάνοντας χρήση ειδικών αλγορίθμων.

Η κατάρα της διαστατικότητας (*curse of dimensionality*) αναφέρεται στα προβλήματα που ανακύπτουν όταν αναλύονται δεδομένα σε χώρους υψηλής διάστασης (δεκάδων ή ακόμα και εκατοντάδων διαστάσεων). Σε αυτές τις περιπτώσεις ο όγκος του χώρου αναζήτησης γίνεται τεράστιος και κατά συνέπεια η πυκνότητα των δεδομένων μειώνεται. Έτσι, ο όγκος των δεδομένων που χρειάζονται για να οδηγήσουν σε κάποιο αξιόπιστο γενικό

συμπέρασμα, αυξάνεται με εκθετικό ρυθμό σε σχέση με την αύξηση της διάστασης.

Τα τελευταία έτη είναι φανερή η ολοένα και αυξανόμενη χρησιμοποίηση της μηχανικής μάθησης για την ανάλυση δεδομένων και την εξαγωγή συμπερασμάτων από αυτά. Το γεγονός αυτό, οφείλεται στον τεράστιο παραγόμενο όγκο δεδομένων αλλά και στο ευρύ φάσμα προβλημάτων και προκλήσεων που μπορούν να αντιμετωπιστούν από την αλγοριθμική επεξεργασία δεδομένων. Ένα από τα γενικότερα προβλήματα στον τομέα της μηχανικής μάθησης είναι η ανάλυση χρονοσειρών, δηλαδή σημείων δεδομένων με διατεταγμένη χρονική σειρά *(sequential data)* [8]. Η επεξεργασία και ανάλυση τέτοιου είδους συνόλου δεδομένων είναι ιδιαίτερα απαιτητική, καθώς ο παράγοντας του χρόνου εισάγει εξαρτήσεις μεταξύ μετρήσεων οι οποίες πρέπει να αναγνωριστούν από τους αλγορίθμους μάθησης, ώστε να εξάγουν ικανοποιητικά αποτελέσματα.

1.1.3 Βαθιά Μάθηση

Η *«Βαθιά Μάθηση» Deep Learning* [7] (γνωστή και ως βαθιά δομημένη μάθηση (deep structured learning) ή ιεραρχική μάθηση (hierarchical learning)) είναι μέρος μιας ευρύτερης κατηγορίας μεθόδων *μηχανικής μάθησης* βασιζόμενων σε αναπαραστάσεις δεδομένων μάθησης παρά σε αλγορίθμους συγκεκριμένων εργασιών. Η μάθηση μπορεί να είναι επιβλεπόμενη μη επιβλεπόμενη (βλ. Κεφ. 2.1).

Μερικές αναπαραστάσεις είναι βασισμένες στην ερμηνεία της επεξεργασίας πληροφορίας και των μοτίβων επικοινωνίας σε ένα βιολογικό νευρικό σύστημα, όπως για παράδειγμα η νευρική κωδικοποίηση δηλαδή ο ορισμός μια σχέσης μεταξύ διαφορετικών ερεθισμάτων και τις συσχετισμένες νευρικές αποκρίσεις του εγκεφάλου. Η έρευνα προσπαθεί να δημιουργήσει αποδοτικά συστήματα τα οποία μπορούν να μάθουν αυτές τις αναπαραστάσεις από μη επισημασμένα (*unlabeled*) δεδομένα μεγάλης κλίμακας.

Οι αρχιτεκτονικές της βαθιάς μάθησης όπως τα νευρωνικά δίκτυα βαθιάς μάθησης και τα αναδρομικά νευρωνικά δίκτυα (*recurrent neural*

networks) έχουν εφαρμοστεί σε τομείς όπως η όραση υπολογιστών (*computer vision*), αναγνώριση φωνής, επεξεργασία φυσικής γλώσσας (*natural language processing*), αναγνώρισης ήχου, διήθηση κοινωνικών δικτύων, μηχανική μετάφραση (*machine translation*) και βιοπληροφορικής όπου παρήχθησαν αποτελέσματα τα οποία ήταν συγκρίσιμα και σε ορισμένες περιπτώσεις καλύτερα από τους αντίστοιχους ειδικούς.

Η βαθιά μάθηση αποτελεί μια ομάδα αλγορίθμων μηχανικής μάθησης οι οποίοι:

- Χρησιμοποιούν μια αλληλουχία πολλαπλών επιπέδων από μη γραμμικές μονάδες επεξεργασίας για εξαγωγή και μετασχηματισμό χαρακτηριστικών. Κάθε επίπεδο έχει σαν είσοδο την έξοδο του προηγούμενου επιπέδου.
- Μαθαίνουν με επιβλεπόμενο και/ή με μη επιβλεπόμενο τρόπο.
- Είναι μέρος της ευρύτερης περιοχής της μηχανικής μάθησης.
- Μαθαίνουν πολλαπλά επίπεδα αναπαραστάσεων που αντιστοιχούν σε διαφορετικά επίπεδα αφαίρεσης. Τα επίπεδα σχηματίζουν μια ιεραρχία εννοιών.
- Χρησιμοποιούν ένα είδος *ομαλής κατάβασης (gradient descent)* για εκπαίδευση μέσω *οπισθοδιάδοσης* (*backpropagation).*

1.2 Αντικείμενο διπλωματικής

Όπως αναφέρθηκε και προηγουμένως παρατηρείται μια μεγάλη αύξηση των καταγραφόμενων δεδομένων. Στην πλειοψηφία τους τα δεδομένα αυτά καταγράφονται από αισθητήρες, κάμερες και logs επομένως παρουσιάζουν έντονη χρονική εξάρτηση. Τα δεδομένα τέτοιου τύπου είναι ιδιαίτερα 'δύστροπα' και απαιτητικά στην επεξεργασία τους καθώς πρέπει αν δεν συμπεριληφθεί ο παράγοντας του χρόνου ίσως τα μοντέλα πρόβλεψης να μην έχουν την αναμενόμενη απόδοση

Μία σχετικά 'ανεξερεύνητη' κατηγορία χρονοεξαρτώμενων δεδομένων είναι τα δεδομένα που συλλέγονται από τις διαδρομές των αυτοκινήτων μέσω GPS ή μέσω καταγραφών από αισθητήρες που υπάρχουν στα αυτοκίνητα

κυρίως νέας τεχνολογίας. Η παρούσα δουλειά είναι μια προσέγγιση για την αποτελεσματική επεξεργασία δεδομένων που έχουν συλλεχθεί από έναν προσομοιωτή οδήγησης με σκοπό την αναγνώριση και διάγνωση ασθενών που πάσχουν από νευροεκφυλιστικές ασθένειες. Τα δεδομένα καταγράφηκαν από αισθητήρες του προσομοιωτή σε διαφορετικές διαδρομές και με διαφορετικές συνθήκες.

Έγινε προσπάθεια για δημιουργία ενός μοντέλου μηχανικής μάθησης το οποίο θα μπορούσε να εντοπίσει τα χαρακτηριστικά εκείνα τα οποία μπορούν να διακρίνουν έναν υγιή οδηγό από έναν ασθενή. Ως επόμενος στόχος τέθηκε η διάκριση και ο εντοπισμός της συγκεκριμένης νευροεκφυλιστικής ασθένειας και όχι ο απλός εντοπισμός των ασθενών. Τα μοντέλα που αναπτύχθηκαν εκπαιδεύτηκαν πρώτα με δεδομένα που είχαν επισημανθεί με την ανάλογη ετικέτα και ύστερα αξιολογήθηκαν πάνω σε δεδομένα τα οποία δεν είχαν χρησιμοποιηθεί στη φάση εκπαίδευσης.

1.2.1 Συνεισφορά

Όπως αναφέρθηκε στην προηγούμενη ενότητα η παρούσα εργασία εξετάζει και τελικά καταλήγει σε ένα αξιόπιστο μοντέλο διάγνωσης οδηγών βασισμένο πάνω σε δεδομένα που έχουν καταγραφεί και συλλεχθεί από έναν προσομοιωτή οδήγησης. Η ανάλυση εστίασε πάνω σε τρία διαφορετικά μοντέλα μηχανικής μάθησης. Αρχικά, υλοποιήθηκε και δοκιμάστηκε ένας αλγόριθμος k εγγύτερων γειτόνων (*k-nearest neighbors)* ο οποίος εκτελούσε κατηγοριοποίηση (classification) σύμφωνα με την 'απόσταση' που είχαν μεταξύ τους οι χρονοσειρές των ταχυτήτων κάθε οδηγού σε σχέση με τους υπόλοιπους. Έπειτα, δοκιμάστηκε ένα νευρωνικό δίκτυο τύπου perceptron πολλαπλών επιπέδων (*multilayer perceptron)* με δύο επίπεδα 20 κρυφών νευρώνων αφού τα δεδομένα υποβλήθηκαν στη διαδικασία ανάλυσης κύριων συνιστωσών (*principle component analysis – PCA)* για απομείωση χαρακτηριστικών *(feature reduction)* ώστε να μειωθεί ο χρόνος εκπαίδευσης αφού συρρικνώνεται ο χώρος αναζήτησης συνάρτησης κατηγοριοποίησης. Το τρίτο μοντέλο ήταν

ένας κατηγοριοποιητής δέντρου απόφασης *(decision tree classifier)* το οποίο ορίστηκε να έχει μέγιστο βάθος 10 επιπέδων και χρησιμοποιεί την μετρική *Gini impurity* (βλ. Κεφ. 2).

Έγινε επίσης προσπάθεια για βελτίωση της απόδοσης των μοντέλων αυτών αντικαθιστώντας τις αρχικές τιμές δεδομένων με τον κινούμενο μέσο όρο ανά συγκεκριμένο αριθμό εγγραφών. Τα μοντέλα αξιολογήθηκαν τόσο σε δυαδική κατηγοριοποίηση *(binary classification)* δηλαδή στον απλή διαπίστωση αν ο οδηγός είναι ασθενής, όσο και σε κατηγοριοποίηση πολλαπλών κατηγοριών *(multiclass classification)* δηλαδή στη διάγνωση της συγκεκριμένης ασθένειας από την οποία πάσχει.

Κρίνοντας από τα αποτελέσματα των μετρήσεων ο πιο αποτελεσματικός και σταθερός αλγόριθμος φαίνεται να είναι αυτός του κατηγοριοποιητή δέντρου απόφασης δίνοντας πάνω από **93**% ακρίβεια στη δυαδική κατηγοριοποίηση και πάνω από **81**% στην αναγνώριση της συγκεκριμένης κατηγορίας που ανήκει ο οδηγός (υγιής, νόσος Alzheimer, νόσος Parkinson, ήπια γνωστική ανεπάρκεια (MCI)).

1.3 Σχετικές εργασίες

Στην αναζήτηση σχετικών εργασιών για την εκπόνηση της παρούσας διπλωματικής διαπιστώθηκε η απουσία παρόμοιας δουλειάς υπό την έννοια ότι δε βρέθηκε άλλη ερευνητική εργασία που να έχει ως θέμα τη διάγνωση οδηγών συλλέγοντας δεδομένα κατά τη διάρκεια της οδήγησης. Ωστόσο, υπάρχουν αξιοσημείωτες εργασίες πάνω στην ευρύτερη περιοχή της διάγνωσης μέσω μηχανικής μάθησης αλλά και στην αποτελεσματική επεξεργασία οδηγικών δεδομένων.

1.3.1 Ανάλυση Οδηγικής Συμπεριφοράς

Η συλλογή δεδομένων οδηγικών χαρακτηριστικών τα τελευταία χρόνια γίνεται από δορυφόρους GPS και από «Διαδίκτυο Πραγμάτων» (Internet of Things – IoT) συσκευές π.χ. αισθητήρες που υπάρχουν κυρίως σε αυτοκίνητα τελευταίας τεχνολογίας. Η ανάλυση, επεξεργασία και η εξαγωγή συμπερασμάτων από τα δεδομένα αυτά αποτελεί μια μεγάλη

πρόκληση και παράλληλα δημιουργεί έναν ευρύ χώρο ενδιαφέροντος. Σε σχετικές εργασίες που προτείνουν λύσεις για την αντιμετώπιση αυτής της κατάστασης έχουν επιστρατευτεί κυρίως μέθοδοι βαθιάς μάθησης λόγω αυξημένης πολυπλοκότητας.

Τα αποτελέσματα των προτεινόμενων μεθόδων αφορούν κυρίως στην αναγνώριση και ταυτοποίηση του οδηγού αλλά και στην πρόβλεψη συμπεριφοράς κατά τη διάρκεια της οδήγησης. Σε αυτήν την κατηγορία ανάλυσης δεδομένων η επιλογή των κατάλληλων χαρακτηριστικών χειροκίνητα είναι ιδιαίτερα δύσκολη καθώς συνήθως απαιτείται η εξαγωγή νέων χαρακτηριστικών από τα υπάρχοντα.

Σε αυτό το πλαίσιο έχουν εκπονηθεί εργασίες / προτάσεις για την αποδοτική επεξεργασία των δεδομένων χρησιμοποιώντας διάφορα μοντέλα. Στην εργασία [19], δορυφορικά δεδομένα GPS τα οποία έπειτα από δύο στάδια προεπεξεργασίας χρησιμοποιούνται για την εκπαίδευση μοντέλων βασισμένα σε Δέντρα Απόφασης, Συνελικτικών Νευρωνικών Δικτύων (*Convolutional Neural Networks – CNN*) και Αναδρομικών Νευρωνικών Δικτύων (*Recurrent Neural Networks – RNN*). Το μοντέλο αυτό προτείνεται για την αναγνώριση και ταυτοποίηση του οδηγού. Ένα άλλο μοντέλο [18] χρησιμοποιεί κι αυτό δεδομένα από δορυφόρο GPS και *RNN* με σκοπό να προβλέψει τις αποφάσεις των οδηγών και να τις μοντελοποιήσει ως μοντέλο μείξης Γκαουσιανών κατανομών *(Gaussian mixture model)*. Υπάρχει επίσης μοντελοποίηση η οποία ενσωματώνει *ασαφή λογική (fuzzy logic)* σε νευρωνικά δίκτυα ώστε να προβλέψει τις κινήσεις των οδηγών όπως π.χ. αλλαγές λωρίδας σε συνθήκες κίνησης.

1.3.2 Διάγνωση με τεχνικές μηχανικής μάθησης

Η μηχανική μάθηση εφαρμόζεται εδώ και κάποια χρόνια στον τομέα των ιατρικών διαγνώσεων [10]. Στην εργασία [9] γίνεται μια επισκόπηση των πιο συνηθισμένων εφαρμοζόμενων τεχνικών στις οποίες περιλαμβάνονται ο *ταξινομητής Bayes*, τα *νευρωνικά δίκτυα* και τα *δέντρα απόφασης*. Οι τρεις αυτές τεχνικές αναφέρεται ότι έχουν παρόμοια απόδοση αλλά τα αποτελέσματα τους αντιμετωπίζονται με επιφύλαξη από την ιατρική κοινότητα καθώς τα πρότυπα που αναγνώριζαν δεν ήταν απόλυτα

κατανοητά σε βιολογικό πλαίσιο. Καθώς περνούν τα χρόνια όμως και οι τεχνικές της μηχανικής μάθησης ωριμάζουν τα αποτελέσματα που δίνουν και τα συμπεράσματα που εξάγουν θα τείνουν να γίνονται όλο και πιο αποδεκτά [12][16].

Στον τομέα της ιατρικής, η μηχανική μάθηση έχει επιστρατευτεί για τη διάγνωση ασθενών μέσω της ταξινόμησης εικόνων (image classification) και *αναγνώρισης προτύπων (pattern recognition)* [11][13]. Τέτοιες μέθοδοι έχουν προταθεί για την αναγνώριση *Alzheimer* και *Mild Cognitive Impairment (MCI)* με αναγνώριση εικόνων από *μαγνητικές τομογραφίες (Magnetic Resonance Imaging - MRI)* χρησιμοποιώντας αλγορίθμους Μηχανών Διανυσμάτων Υποστήριξης *(Support Vector Machine - SVM).* [14]

1.3.3 Οργάνωση κειμένου

Εργασίες σχετικές με το αντικείμενο της διπλωματικής παρουσιάζονται στο Κεφάλαιο 2, καθώς και θεωρητικά θέματα και προαπαιτούμενης γνώσης για την κατανόηση της εργασίας. Στο Κεφάλαιο 3 αναλύονται τα προς επεξεργασία δεδομένα. Παρουσιάζουμε στο κεφάλαιο 5 τις υπό δοκιμή μεθόδους. Στο Κεφάλαιο 4 γίνεται παρουσίαση αποτελεσμάτων, αξιολόγηση και εξαγωγή συμπερασμάτων. Τα τεχνικά θέματα εξηγούνται στο Κεφάλαιο 5. Προτάσεις για μελλοντική επέκταση και συνέχεια βρίσκονται στο Κεφάλαιο 6 ενώ η Βιβλιογραφία είναι στο Κεφάλαιο 7.

2

Θεωρητικό υπόβαθρο

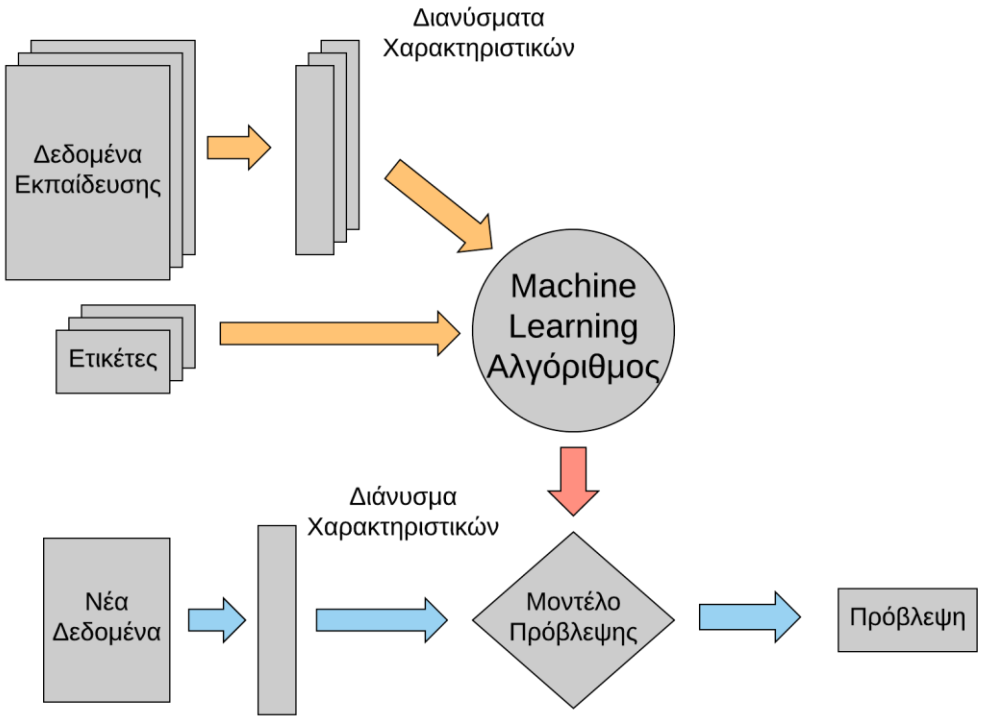
Εδώ σημειώνονται οι αλγόριθμοι, οι τεχνικές, οι μεθοδολογίες και τα μοντέλα που θα χρησιμοποιήσει η διπλωματική και είναι αναγκαία η κατανόησή τους από τον αναγνώστη πριν από την παρουσίαση της ανάλυσης και σχεδίασης του συστήματος.

Πρόκειται για τεχνικές, μεθοδολογίες και μοντέλα που έχουν προταθεί από άλλους και δεν είναι πρωτότυπη δουλειά της διπλωματικής.

2.1 Διαδικασίες Μάθησης

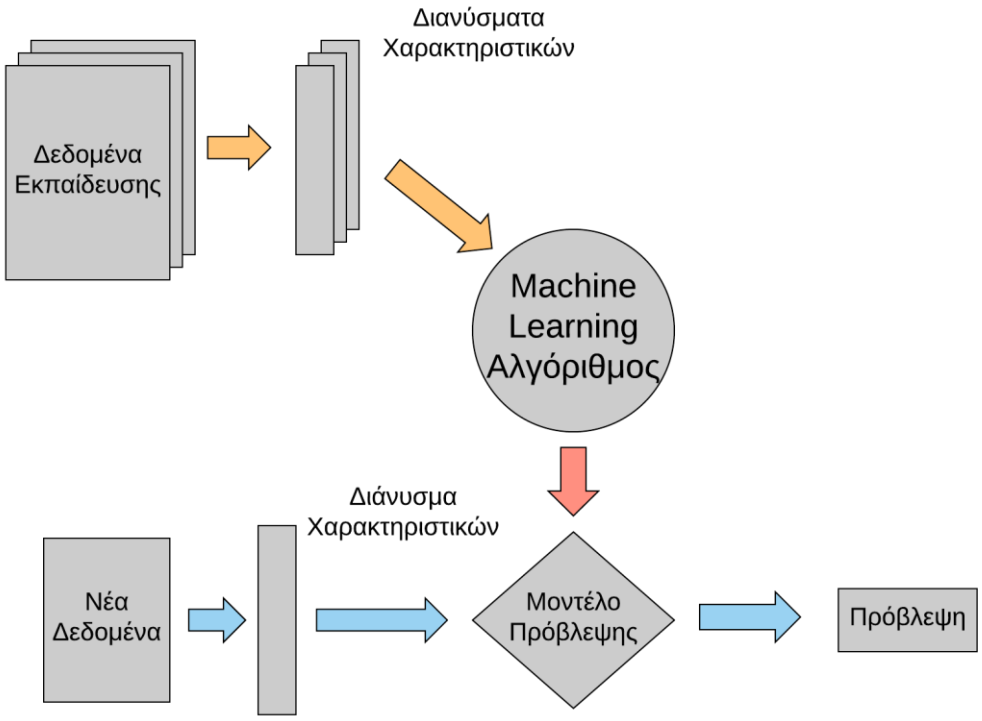
Οι διαδικασίες μάθησης μέσω των οποίων λειτουργούν και μαθαίνουν τα νευρωνικά δίκτυα μπορούν να κατηγοριοποιηθούν στις εξής δύο ευρείες κατηγορίες: επιβλεπόμενη μάθηση και μη επιβλεπόμενη μάθηση [6]. Η σχηματική τους αναπαράσταση φαίνεται στα διαγράμματα των αντίστοιχων ενοτήτων παρακάτω.

2.1.1 Επιβλεπόμενη μάθηση



Η επιβλεπόμενη μάθηση είναι η διαδικασία μηχανικής μάθησης κατά την οποία γίνεται προσέγγιση μιας άγνωστης συνάρτησης με επισημασμένα δεδομένα. Στη διάρκεια της εκπαίδευσης είναι γνωστή η κατηγορία στην οποία ανήκουν τα δείγματα εκπαίδευσης. Τα δείγματα εισόδου αποτελούνται από ένα ζεύγος διανύσματος χαρακτηριστικών και την επιθυμητή έξοδο. Ο αλγόριθμος επιβλεπόμενης μάθησης παράγει μια συνάρτηση αναλύοντας τα δεδομένα εκπαίδευσης η οποία μπορεί να χρησιμοποιηθεί για την αντιστοίχιση νέων δειγμάτων. Ιδανικά ο αλγόριθμος μετά την εκπαίδευση θα μπορούσε να προβλέψει την κατηγορία δειγμάτων που δεν του είχαν ξαναπαρουσιαστεί. Αυτό προϋποθέτει ότι αλγόριθμος έχει την ικανότητα να γενικεύσει από τα δεδομένα εκπαίδευσης σε δεδομένα που δεν είχε αντιμετωπίσει ξανά.

2.1.2 Μη επιβλεπόμενη μάθηση



Η μη επιβλεπόμενη μάθηση είναι η διαδικασία μηχανικής μάθησης κατά την οποία γίνεται προσέγγιση μιας συνάρτησης ώστε να περιγράψει την κρυφή δομή σε μη επισημασμένα δεδομένα (πχ. μια κατηγοριοποίηση ή μια ταξινόμηση η οποία δεν περιλαμβάνεται στα αρχικά δεδομένα). Εφόσον τα δείγματα που δίνονται στο μοντέλο μάθησης δεν είναι επισημασμένα, δεν υπάρχει αξιολόγηση της δομής που ανακαλύπτει τελικά το μοντέλο. Μια συνηθισμένη εφαρμογή της μη επιβλεπόμενης μάθησης είναι η εκτίμηση της συνάρτησης πυκνότητας πιθανότητας σε κατανομές αλλά μπορεί να περιλαμβάνει και άλλα προβλήματα όπως η επεξήγηση των βασικών χαρακτηριστικών των δεδομένων.

Τα μοντέλα και οι τεχνικές που υλοποιήθηκαν βασίζονται σε διαδικασίες επιβλεπόμενης μάθησης καθώς υπήρχε ήδη η γνώση για την κατηγορία που ανήκαν τα εκάστοτε δεδομένα.

2.2 k-Εγγύτεροι Γείτονες

Ο αλγόριθμος k-Εγγύτερων Γειτόνων *(k-nearest neighbors)* είναι μία μη παραμετρική μέθοδος η οποία χρησιμοποιείται για κατηγοριοποίηση και παλινδρόμηση. Και στις δύο περιπτώσεις η είσοδος αποτελείται από τα k κοντινότερα παραδείγματα του χώρου χαρακτηριστικών. Η έξοδος, όταν ο αλγόριθμος χρησιμοποιείται για κατηγοριοποίηση, αποτελείται από την κατηγορία στην οποία ανήκει το αντικείμενο που θέλουμε να κατηγοριοποιηθεί. Το αντικείμενο κατηγοριοποιείται σύμφωνα με την πλειοψηφία των γειτόνων του, δηλαδή το αντικείμενο 'παίρνει' την κατηγορία που ανήκουν οι περισσότεροι από τους k γείτονές του. Εάν $k = 1$, τότε το αντικείμενο παίρνει την κατηγορία του κοντινότερου γείτονα.

Ο *k-nearest neighbors* είναι παράδειγμα αλγορίθμου «μάθησης βασισμένης σε στιγμιότυπα» (*instance-based* learning), ή ράθυμης μάθησης (*lazy* learning), όπου η συνάρτηση προσεγγίζεται μόνο τοπικά και όλοι οι υπολογισμοί αναβάλλονται μέχρι την κατηγοριοποίηση. Ο *k-nearest neighbors* αλγόριθμος είναι από τους πιο απλούς αλγόριθμους που χρησιμοποιούνται στην μηχανική μάθηση.

Οι γείτονες λαμβάνονται από ένα σύνολο αντικειμένων των οποίων η κατηγορία είναι γνωστή. Αυτό μπορεί να θεωρηθεί το σύνολο εκπαίδευσης του αλγορίθμου αν και δεν λαμβάνει χώρα καμία ρητή εκπαίδευση.

2.2.1 Αλγόριθμος

Τα δείγματα εκπαίδευσης είναι διανύσματα σε έναν πολυδιάστατο χώρο χαρακτηριστικών, καθένα εκ των οποίων έχει μία ετικέτα που σηματοδοτεί την κατηγορία στην οποία ανήκει. Η φάση εκπαίδευσης του αλγορίθμου αποτελείται απλά από την αποθήκευση αυτών των διανυσμάτων χαρακτηριστικών και των ετικετών τους.

Στη φάση της κατηγοριοποίησης ένα μη κατηγοριοποιημένο διάνυσμα κατηγοριοποιείται αναθέτοντας του την ετικέτα που είναι πιο συχνά εμφανιζόμενη στα k κοντινότερα δείγματα εκπαίδευσης, όπου το k είναι μία οριζόμενη από το χρήστη σταθερά.

Η επιλογή για τους κοντινότερους γείτονες πρέπει να γίνει αφού οριστεί μια κατάλληλη μετρική/συνάρτηση για τον υπολογισμό της απόστασής τους. Μια συχνά χρησιμοποιούμενη τεχνική για συνεχείς μεταβλητές είναι η Ευκλείδεια απόσταση. Για διακριτές μεταβλητές μπορούν να χρησιμοποιηθούν άλλες μετρικές όπως η απόσταση Hamming. Στο πλαίσιο της κατηγοριοποίησης μικροσυστοιχιών γονιδιακής έκφρασης, ο *k-nearest neighbors* έχει επίσης χρησιμοποιηθεί με συντελεστές συσχέτισης όπως Pearson και Spearman.

Ένα μειονέκτημα της βασικής πλειοψηφικής διαδικασίας είναι ότι η κατανομή των κατηγοριών είναι στρεβλή. Αυτό σημαίνει ότι παραδείγματα μιας πιο συχνά εμφανιζόμενης κατηγορίας τείνουν να κυριαρχήσουν στην πρόβλεψη των νέων παραδειγμάτων γιατί βρίσκονται συχνά στους κοντινότερους k γείτονες λόγω του μεγάλου πλήθους τους. Ένας τρόπος να υπερπηδήσουμε αυτό το εμπόδιο είναι να σταθμίσουμε την κατηγοριοποίηση, λαμβάνοντας υπ' όψιν την απόσταση από καθέναν από τους k κοντινότερους γείτονές του. Η κατηγορία καθενός από τα k κοντινότερα 'σημεία' πολλαπλασιάζεται με ένα βάρος ανάλογο του αντιστρόφου της απόστασης από το εξεταζόμενο δείγμα.

2.3 Δυναμική Χρονική Στρέβλωση

Στην ανάλυση χρονοσειρών ο αλγόριθμος δυναμικής χρονικής στρέβλωσης *(dynamic time warping)* χρησιμοποιείται για την μέτρηση της απόστασης δύο χρονικών σειρών οι οποίες μπορεί να διαφέρουν σε αριθμό στοιχείων [20]. Για παράδειγμα, οι ομοιότητες στο περπάτημα θα μπορούσαν να εντοπιστούν, ακόμα και αν ένας άνθρωπος περπατάει γρηγορότερα από κάποιον άλλο ή υπήρχαν επιταχύνσεις και επιβραδύνσεις στη διάρκεια της διαδρομής. Ο αλγόριθμος δυναμικής χρονικής στρέβλωσης έχει εφαρμοστεί σε χρονικές σειρές βίντεο, ήχου, γραφικών και γενικά σε οποιοδήποτε είδος δεδομένων που μπορεί να μετατραπεί σε διανυσματική σειρά. Μία πολύ γνωστή εφαρμογή είναι η αναγνώριση ομιλίας ώστε να αντιμετωπιστούν οι διαφορετικές ταχύτητες

ομιλίας. Άλλες εφαρμογές είναι η ταυτοποίηση ομιλητή και η αναγνώριση υπογραφής.

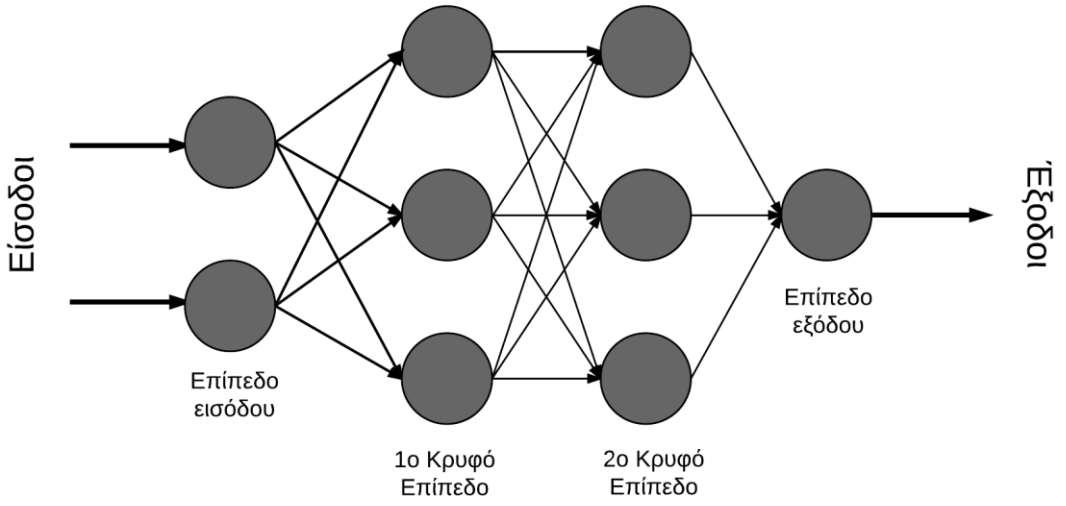
Γενικά, ο αλγόριθμος είναι μια μέθοδος υπολογισμού του καλύτερου ταιριάσματος μεταξύ δύο ακολουθιών (π.χ. χρονοσειρών) υπό ορισμένους περιορισμούς. Οι ακολουθίες 'στρεβλώνονται' μη γραμμικά στη διάσταση του χρόνου για να καθορίσουν ένα μέτρο ομοιότητας ανεξάρτητα από ορισμένες μη γραμμικές διακυμάνσεις στον άξονα του χρόνου. Αυτή η μέθοδος ευθυγράμμισης της ακολουθίας χρησιμοποιείται συχνά στην ταξινόμηση χρονοσειρών. Παρόλο που το *dynamic time warping* αποτελεί μετρική της απόστασης δεν εξασφαλίζει ότι ισχύει η τριγωνική ανισότητα.

Επιπλέον του μέτρου ομοιότητας που μεταξύ των δύο ακολουθιών παράγεται και το ονομαζόμενο *'μονοπάτι στρέβλωσης'* – *'warping path'*. Το μονοπάτι αυτό δείχνει την ακολουθία στρέβλωσης στο χρόνο σύμφωνα με την οποία οι δύο χρονοσειρές ευθυγραμμίζονται στο χρόνο. Το σήμα με αρχικό σύνολο σημείων *X(original), Y(original)* μετασχηματίζεται σε *X(warped), Y(original).* Αυτό βρίσκει εφαρμογές και σε γενετικές ακολουθίες και συγχρονισμό ήχου.

Στην παρούσα εργασία το *dynamic time warping* είναι η μετρική που θα χρησιμοποιήσει ο αλγόριθμος *k-nearest neighbors* για να αποφασίσει τα πλησιέστερα ταιριάσματα κάθε ακολουθίας.

2.4 Πολυεπίπεδο Perceptron

Το πολυεπίπεδο perceptron *(multilayer perceptron - MLP)* είναι ένα είδος τεχνητού νευρωνικού δικτύου εμπρόσθιας τροφοδότησης και αποτελείται από τουλάχιστον τρία επίπεδα νευρώνων όπως φαίνεται στο παρακάτω σχήμα. Εκτός από τους κόμβους εισόδου, κάθε κόμβος, ο οποίος είναι και νευρώνας, χρησιμοποιεί μία μη γραμμική συνάρτηση ενεργοποίησης. Το πολυεπίπεδο perceptron κάνει χρήση μιας τεχνικής επιβλεπόμενης μάθησης η οποία ονομάζεται backpropagation ώστε να εκπαιδευτεί. Τα πολλαπλά επίπεδα *(layers)* και η μη γραμμική ενεργοποίηση διαφοροποιεί το πολυεπίπεδο perceptron από τα γραμμικά perceptron.



2.4.1 Συνάρτηση ενεργοποίησης

Εάν ένα MLP έχει μια γραμμική συνάρτηση ενεργοποίησης σε όλους τους νευρώνες, αποδεικνύεται με γραμμική άλγεβρα ότι οποιοσδήποτε αριθμός επιπέδων μπορεί να μετασχηματιστεί σε ένα μοντέλο εισόδου-εξόδου δύο επιπέδων. Οι νευρώνες των πολυεπίπεδων perceptron χρησιμοποιούν μη γραμμικές συναρτήσεις ενεργοποίησης ώστε να μπορούν να μοντελοποιήσουν δυναμικά δράσης ή πυροδότησης βιολογικών νευρώνων. Οι πιο σύνηθες συναρτήσεις είναι οι σιγμοειδείς:

$$y(u_i) = \tanh(u_i) \text{ ή } y(u_i) = (1 + e^{-u_i})^{-1}$$

Η πρώτη είναι η υπερβολική εφαπτομένη η οποία κυμαίνεται από -1 έως 1 ενώ η δεύτερη ονομάζεται λογιστική συνάρτηση και μοιάζει με την πρώτη αλλά κυμαίνεται από 0 έως 1. Εδώ το y είναι η έξοδος του i-οστού νευρώνα ενώ το u_i το σταθμισμένο άθροισμα των εισόδων του.

2.4.2 Επίπεδα

Ένα *multilayer perceptron* αποτελείται από τρία ή παραπάνω επίπεδα (ένα επίπεδο εισόδου, ένα επίπεδο εξόδου και ένα η περισσότερα κρυμμένα επίπεδα) από μη γραμμικά ενεργοποιούμενους κόμβους κάνοντας το ένα "βαθύ" νευρωνικό δίκτυο. Δεδομένου ότι όλοι οι νευρώνες είναι πλήρως συνδεδεμένοι μεταξύ τους, κάθε κόμβος i ενός επιπέδου συνδέεται με τον

κόμβο j του επόμενου επιπέδου με έναν κόμβο με ένα συγκεκριμένο βάρος w_{ij}.

2.4.3 Μάθηση

Η μάθηση στο *multilayer perceptron* λαμβάνει χώρα μεταβάλλοντας τα βάρη σύνδεσης αφού γίνεται επεξεργασία του εκάστοτε κομματιού δεδομένων βασίζοντας την αλλαγή στο σφάλμα εξόδου συγκρινόμενο με το αναμενόμενο αποτέλεσμα. Αυτό είναι ένα παράδειγμα επιβλεπόμενης μάθησης και πραγματοποιείται μέσω οπισθοδιάδοσης (*backpropagation*), το οποίο αποτελεί γενίκευση του αλγορίθμου ελάχιστου μέσου τετραγωνικού σφάλματος του γραμμικού perceptron.

Ως σφάλμα στον κόμβο εξόδου j στο n-οστό σημείο δεδομένων θεωρούμε την διαφορά $e_j(n) = d_j(n) - y_j(n)$ όπου d είναι η πραγματική τιμή και y η τιμή που παρήγαγε το perceptron. Τα βάρη των κόμβων προσαρμόζονται βασιζόμενα σε διορθώσεις το σφάλμα σε ολόκληρη την έξοδο το οποίο δίνεται από την εξίσωση $\mathrm{E}(n) = \frac{1}{2}\sum_j e_j^2(n)$

Χρησιμοποιώντας βαθμωτή κατάβαση (*gradient descent*), η αλλαγή στο βάρος είναι

$$\Delta w_{ij}(n) = -\eta \frac{\partial E(n)}{\partial u_j(n)} y_i(n)$$

όπου το y_i είναι η έξοδος του προηγούμενου νευρώνα και η είναι ο ρυθμός μάθησης, οποίος επιλέγεται για να εξασφαλίσει ότι τα βάρη σύντομα συγκλίνουν σε μια απόκριση χωρίς ταλαντώσεις.

Η παράγωγος στην εξίσωση που πρέπει να υπολογιστεί εξαρτάται από το επαγόμενο τοπικό πεδίο u_j το οποίο με τη σειρά του ποικίλει. Για τους κόμβους εξόδου αποδεικνύεται ότι

$$-\frac{\partial E(n)}{\partial u_j(n)} = e_j(n)\varphi'(u_j(n))$$

όπου η φ' είναι η παράγωγος της συνάρτησης ενεργοποίησης. Η ανάλυση είναι δυσκολότερη για την αλλαγή των βαρών στους νευρώνες των κρυφών

επιπέδων αλλά μπορεί να δειχτεί ότι η μερική παράγωγος δίνεται από τον τύπο

$$-\frac{\partial E(n)}{\partial u_j(n)} = \varphi'(u_j(n)) \sum_k -\frac{\partial E(n)}{\partial u_k(n)} w_{kj}(n)$$

Η αλλαγή των βαρών βλέπουμε ότι γίνεται με βάση την αλλαγή των βαρών των k-οστών κόμβων οι οποίοι αντιπροσωπεύουν το επίπεδο εξόδου. Έτσι για να αλλάξουν τα βάρη των κρυφών επιπέδων, τα βάρη των κόμβων εξόδου μειώνονται σύμφωνα με την παράγωγο της συνάρτησης ενεργοποίησης και έτσι αυτός ο αλγόριθμος αντιπροσωπεύει μια οπισθοδιάδοση της παραγώγου της συνάρτησης ενεργοποίησης.

2.5 Ανάλυση Κύριων Συνιστωσών

Η ανάλυση κυρίων συνιστωσών (*principal component analysis*) είναι μια στατιστική διαδικασία η οποία χρησιμοποιεί έναν ορθογώνιο μετασχηματισμό για να μετατρέψει ένα σύνολο παρατηρήσεων πιθανά συσχετισμένων μεταβλητών σε ένα σύνολο από γραμμικά ανεξάρτητες μεταβλητές οι οποίες ονομάζονται κύριες συνιστώσες. Ο αριθμός των κύριων συνιστωσών είναι μικρότερος ή ίσος από τον αριθμό των αρχικό αριθμό των μεταβλητών των παρατηρήσεων. Ο μετασχηματισμός είναι ορισμένος με τέτοιον τρόπο ώστε η πρώτη συνιστώσα να έχει τη μέγιστη δυνατή διασπορά και κάθε επόμενη συνιστώσα έχει επίσης με τη σειρά της τη μέγιστη δυνατή διασπορά υπό τον περιορισμό ότι είναι ορθογώνια με την προηγούμενη της. Τα παραγόμενα διανύσματα αποτελούν ασυσχέτιστο ορθογώνιο σύνολο βάσης. Η ανάλυση κυρίων συνιστωσών είναι ευαίσθητη στην σχετική κλιμάκωση των αρχικών μεταβλητών.

Πριν τα δεδομένα παρουσιαστούν στο *multilayer perceptron* μοντέλο που θα εκπαιδευτεί, τα δεδομένα υφίστανται μέσω ανάλυση κύριων μετασχηματισμό σε 10 χαρακτηριστικά. Η επιλογή αυτή έγινε για να μειωθεί ο χώρος αναζήτησης του *multilayer perceptron* και κατά συνέπεια και ο χρόνος εκπαίδευσής του αλλά και για όσο το δυνατόν καλύτερη απόδοση του μοντέλου.

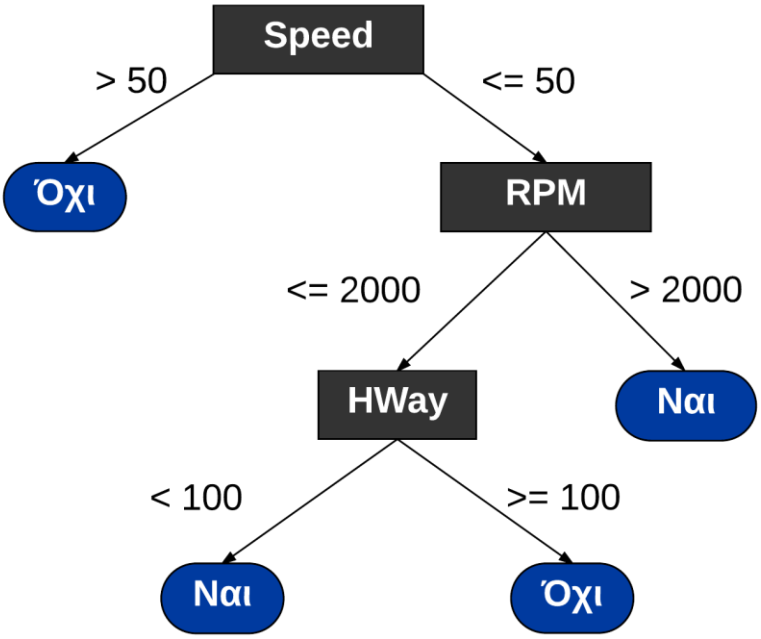
2.6 Δέντρο Απόφασης

Τα decision trees χρησιμοποιούνται ως μοντέλα πρόβλεψης ξεκινώντας από τις παρατηρήσεις για ένα αντικείμενο (οι οποίες αναπαρίστανται στα κλαδιά) μπορούν να εξάγουν συμπεράσματα για κάποια άλλα χαρακτηριστικά του αντικειμένου (τα οποία αναπαρίστανται στα φύλλα). Τα μοντέλα των δέντρων όπου η μεταβλητή που θέλουμε να προσδιορίσουμε παίρνει διακριτές τιμές ονομάζονται *classification trees* ενώ αν η μεταβλητή που θέλουμε να προσδιορίσουμε παίρνει συνεχείς τιμές ονομάζονται *regression trees*.

Στο πεδίο της εξόρυξης δεδομένων τα *decision trees* χρησιμοποιούνται για την περιγραφή δεδομένων και την ταξινόμηση τους σε κλάσεις.

2.6.1 Περιγραφή λειτουργίας

Η μάθηση μέσω *decision trees* χρησιμοποιείται ευρέως στην εξόρυξη δεδομένων. Ο στόχος είναι η δημιουργία ενός μοντέλου το οποίο προβλέπει το χαρακτηριστικό που επιθυμούμε βασισμένο σε τιμές των υπόλοιπων χαρακτηριστικών του αντικειμένου που μας ενδιαφέρει.



Κάθε εσωτερικός κόμβος αντιπροσωπεύει μία από τις μεταβλητές/χαρακτηριστικά εισόδου. Για κάθε πιθανή τιμή της μεταβλητής υπάρχει μια αντίστοιχη ακμή προς κόμβους παιδιά. Κάθε φύλλο αντιπροσωπεύει την πρόβλεψη της τιμής του επιθυμητού χαρακτηριστικού δεδομένων των τιμών των χαρακτηριστικών τα οποία βρίσκονται στο μονοπάτι που ξεκινάει από τη ρίζα και καταλήγει στο συγκεκριμένο φύλλο.

Το decision tree είναι μία απλή αναπαράσταση για την ταξινόμηση αντικειμένων. Η εσωτερικές διακλαδώσεις είναι οι διαφορετικές τιμές που έχουν τα χαρακτηριστικά του προς ταξινόμηση αντικειμένου και η μεταβλητή της οποίας την τιμή προβλέπουν τα φύλλα είναι το id της ομάδας που ταξινομείται το αντικείμενο.

Το δέντρο μαθαίνει χωρίζοντας το αρχικό σύνολο σε υποσύνολα βασιζόμενο σε δοκιμές τιμών των χαρακτηριστικών. Αυτή η διαδικασία επαναλαμβάνεται σε κάθε παραγόμενο υποσύνολο με αναδρομικό τρόπο και ονομάζεται αναδρομική διαμέριση (*recursive partitioning*).

Η αναδρομή σταματάει όταν το υποσύνολο ενός κόμβου έχει όλο την ίδια τιμή για την προσδιοριστέα μεταβλητή, δηλαδή την μεταβλητή που

υποδεικνύει την κλάση που ανήκει το αντικείμενο, ή όταν η διάσπαση δεν προσθέτει επιπλέον αξία στις προβλέψεις.

Αυτή η διαδικασία από την κορυφή προς τα κάτω επαγωγής των δέντρων απόφασης (top-down induction of decision trees) είναι ένας greedy αλγόριθμος και είναι ο πιο συνηθισμένος τρόπος μάθησης για decision trees.

2.6.2 Μετρικές

2.6.2.1 *Gini impurity*

Είναι ένα μέτρο για το πόσο συχνά ένα τυχαία επιλεγμένο χαρακτηριστικό από το σύνολο θα ήταν λάθος ταξινομημένο αν η ταξινόμηση γινόταν τυχαία σύμφωνα με την κατανομή των ταξινομήσεων στο υποσύνολο. Μπορεί να υπολογιστεί αθροίζοντας την πιθανότητα p_i για ένα αντικείμενο με ετικέτα i να επιλεγεί επί την πιθανότητα $1 - p_i$ να έχει γίνει λάθος στην ετικέτα. Μηδενίζεται όταν όλες οι περιπτώσεις σε αυτόν τον κόμβο είναι σε μια συγκεκριμένη κατηγορία.

Η Gini impurity για ένα σύνολο με J κλάσεις και p_i το ποσοστό των αντικειμένων με κλάση i στο σύνολο όπου $i = \{1, 2, \dots, J\}$ υπολογίζεται με τον τύπο

$$I_G(p) = \sum_{i=1}^{J} p_i(1 - p_i) = \sum_{i \neq k}^{J} p_i p_k$$

2.7 Κινούμενος Μέσος

Ο κινούμενος μέσος είναι ένας υπολογισμός για ανάλυση των σημείων δεδομένων δημιουργώντας μια σειρά από μέσους όρους από διαφορετικά υποσύνολα του συνολικού συνόλου δεδομένων. Δεδομένης μιας σειράς αριθμών και ενός σταθερού μεγέθους υποσυνόλου, το πρώτο στοιχείο του κινούμενου μέσου υπολογίζεται λαμβάνοντας τον μέσο όρο του αρχικού υποσυνόλου της σειράς αριθμών. Έπειτα το υποσύνολο μεταβάλλεται

μετατοπιζόμενο προς τα μπροστά, δηλαδή εξαιρώντας τον πρώτο αριθμό της σειράς και προσθέτοντας την επόμενη τιμή της σειράς.

Ο κινούμενος μέσος χρησιμοποιείται κοινώς με δεδομένα από χρονοσειρές *(time series)* για να ομαλοποιηθούν οι βραχείες διακυμάνσεις και να δώσει έμφαση στις μακροπρόθεσμες τάσεις. Το κατώφλι μεταξύ βραχυπρόθεσμου και μακροπρόθεσμου εξαρτάται από την εφαρμογή και οι παράμετροι του κινούμενου μέσου όρου θα τεθούν ανάλογα. Χρησιμοποιείται στην τεχνική ανάλυση χρηματοοικονομικών δεδομένων, όπως μετοχές, αποδόσεις επενδύσεων ή όγκους συναλλαγών. Χρησιμοποιείται επίσης στα οικονομικά για να εξετάσει το ακαθάριστο εθνικό προϊόν, εργασιακή απασχόληση ή άλλες μακροοικονομικές χρονοσειρές. Μαθηματικά, ο κινούμενος μέσος όρος είναι ένα είδος συνέλιξης και συνεπώς μπορεί να θεωρηθεί σαν βαθυπερατό φίλτρο το οποίο χρησιμοποιείται στην επεξεργασία σήματος. Όταν χρησιμοποιείται σε ανεξάρτητες από χρόνο σειρές δεδομένων, ο κινούμενος μέσος φιλτράρει τις συνιστώσες υψηλής συχνότητας χωρίς κάποια συγκεκριμένη σύνδεση με το χρόνο, παρόλο που ένα είδος ταξινόμησης υπονοείται. Από στατιστική σκοπιά μπορεί να θεωρηθεί ότι εξομαλύνει τα δεδομένα (βλ. διάγραμμα στο τέλος της ενότητας).

Η πιο απλή περίπτωση κινούμενου μέσου όρου είναι ο απλός κινούμενος μέσος *(simple moving average)*. Στις οικονομικές εφαρμογές ο απλός κινούμενος μέσος όρος είναι ο αστάθμητος μέσος όρος των προηγούμενων n δεδομένων. Ωστόσο, στην επιστήμη και στη μηχανική ο μέσος όρος συνήθως λαμβάνεται από ίσο αριθμό δεδομένων από κάθε πλευρά ενός κεντρικού σημείου. Αυτό εξασφαλίζει ότι οι διακυμάνσεις του μέσου όρου θα ακολουθούν τις διακυμάνσεις των δεδομένων και δεν θα μετατοπιστούν χρονικά. Έστω ότι οι n τελευταίες τιμές είναι $p_M, p_{M-1}, \cdots, p_{M-(n-1)}$. Ο μέσος όρος είναι:

$$p_M^* = \frac{p_M + p_{M-1} + \cdots + p_{M-(n-1)}}{n}$$

Όταν υπολογίζουμε διαδοχικές τιμές επειδή ένας όρος εισέρχεται και κάποιος άλλος εξέρχεται είναι άσκοπο να αθροίζουμε κάθε φορά καθώς μπορούμε να υπολογίσουμε κάθε μέσο όρο από τον προηγούμενο ως εξής:

$$p_M^* = p_{M-1}^* + \frac{p_M}{n} - \frac{p_{M-n}}{n}$$

Η επιλογή του n εξαρτάται από το είδος του ενδιαφέροντος δηλαδή βραχύ, μέσο ή σε μάκρος.

Για κάποιες εφαρμογές, είναι επωφελές να αποφευχθεί η μετακίνηση χρησιμοποιώντας μόνο παρελθοντικά δεδομένα. Κατά συνέπεια, ένας κεντρικά κινούμενος μέσος όρος μπορεί να υπολογιστεί, χρησιμοποιώντας δεδομένα ισόποσα μοιρασμένα σε κάθε 'πλευρά' του σημείου της σειράς του οποίου ο μέσος όρος υπολογίζεται. Αυτό προϋποθέτει τη χρήση περιττού αριθμού σημείων δεδομένων στο υποσύνολο υπολογισμού.

3

Αρχιτεκτονική συστήματος

Στο παρακάτω διάγραμμα φαίνεται η αρχιτεκτονική του συστήματος σύμφωνα με την οποία δομήθηκε. Στη συνέχεια του παρόντος κεφαλαίου γίνεται περεταίρω περιγραφή και ανάλυση των σταδίων 1-5.

3.1 Συλλογή δεδομένων

Τα δεδομένα προέρχονται από έναν *driving simulator* ο οποίος χρησιμοποιήθηκε στο project [17] για την στατιστική μελέτη των ασθενών οδηγών και για την εξαγωγή συμπερασμάτων για την οδηγική τους συμπεριφορά. Συγκεκριμένα οδήγησαν συνολικά 118 άτομα εκ των οποίων τα 49 ήταν υγιή και τα 69 είχαν διαγνωστεί με ασθένειες νευρολογικού τύπου όπως *Alzheimer, Ήπια Γνωστική Ανεπάρκεια (Mild Cognitive Impairment)* και *νόσο του Parkinson*.

Ο προσομοιωτής διέθετε διαφορετικά οδηγικά σενάρια συμπεριλαμβανομένου της οδήγησης σε επαρχιακή οδό και με διαφορετικές συνθήκες κίνησης. Τα οδηγικά πειράματα ξεκίνησαν με διαδρομή εξάσκησης έτσι ώστε να υπάρχει η απαραίτητη εξοικείωση πριν ξεκινήσει το κυρίως πείραμα.

Το κυρίως πείραμα περιλάμβανε οδήγηση σε δύο διαφορετικά sessions περίπου 20 λεπτά το κάθε ένα. Οι δύο διαδρομές που ακολουθήθηκαν ήταν οι εξής:

- Μία διαδρομή σε επαρχιακό δρόμο 2.1 χλμ με μία λωρίδα ανά κατεύθυνση πλάτους 3 μέτρων, μηδενικής κλίσης και ήπιες στροφές
- Μια διαδρομή σε αστικό περιβάλλον 1.7 χλμ με δύο λωρίδες ανά κατεύθυνση στο μεγαλύτερο μέρος του πλάτους 3.5 μέτρων και μπαριέρα. Επιπλέον, υπήρχαν στενά πεζοδρόμια, πινακίδες και χώροι parking υπήρχαν στα δεξιά περιθώρια του δρόμου.

Σε κάθε μία από τις διαδρομές υπήρχε η δυνατότητα για προσομοίωση διαφορετικών συνθηκών κίνησης. Αυτά τα σενάρια ήταν:

- Low traffic: Ήπιες συνθήκες κίνησης – οι αφίξεις των περιβάλλοντων οχημάτων προέρχονταν από μία κατανομή Γάμμα με μέσο μ=12 sec, and variance σ^2=6 sec, αντιστοιχώντας σε ένα μέσο όγκο κίνησης Q=300 οχήματα/ώρα.
- High traffic: Συνθήκες αυξημένης κίνησης – οι αφίξεις των περιβάλλοντων οχημάτων προέρχονταν από μία κατανομή Γάμμα

με μέσο μ=6 sec, and variance σ^2=12 sec, αντιστοιχώντας σε ένα μέσο όγκο κίνησης Q=600 οχήματα/ώρα.

Κατά τη διάρκεια της διαδρομής ο προσομοιωτής κατέγραφε τα εξής χαρακτηριστικά:

ID	**Variable**	**Explanation**
1	Time	current real-time in milliseconds since start of the drive.
2	x-pos	x-position of the vehicle in m.
3	y-pos	y-position of the vehicle in m.
4	z-pos	z-position of the vehicle in m.
5	Road	road number of the vehicle in [int].
6	Richt	direction of the vehicle on the road in [BOOL] (0/1).
7	Rdist	distance of the vehicle from the beginning of the drive in m.
8	rspur	track of the vehicle from the middle of the road in m.
9	ralpha	direction of the vehicle compared to the road direction in degrees.
10	Dist	driven course in meters since begin of the drive.
11	Speed	actual speed in km/h.
12	Brk	brake pedal position in percent.
13	Acc	gas pedal position in percent.
14	Clutch	clutch pedal position in percent.
15	Gear	chosen gear (0 = idle, 6 = reverse).
16	RPM	motor revolvation in 1/min.
17	HWay	headway, distance to the ahead driving vehicle in m.
18	DLeft	distance to the left road board in meter.
19	DRight	distance to the right road board in meter.
20	Wheel	steering wheel position in degrees.
21	THead	time to headway, i. e. to collision with the ahead driving vehicle, in seconds.

22	TTL	time to line crossing, time until the road border line is exceeded, in seconds.
23	TTC	time to collision (all obstacles), in seconds.
24	AccLat	acceleration lateral, in m/s^2
25	AccLon	acceleration longitudinal, in m/s^2
26	EvVis	event-visible-flag/event-indication, 0 = no event, 1 = event.
27	EvDist	event-distance in m.
28	Err1No	number of the most important driving failure since the last data set
29	Err1Val	state date belonging to the failure, content varies according to type of failure.
30	Err2No	number of the next driving failure (maybe empty).
31	Err2Val	additional date to failure 2.
32	Err3No	number of a further driving failure (maybe empty).
33	Err3Val	additional date to failure 3.

Η καταγραφή των δεδομένων γινόταν περίπου ανά 33-34 ms.

3.2 Καθαρισμός δεδομένων

Όπως φαίνεται από τον πίνακα χαρακτηριστικών της ενότητας 3.1, τα χαρακτηριστικά *Err1No, Err1Val, Err2No, Err2Val, Err3No, Err3Val* αναφέρονται στην ύπαρξη ή απουσία σφάλματος. Πριν προχωρήσουμε σε περαιτέρω επεξεργασία των δεδομένων είναι χρήσιμο να αφαιρέσουμε τις εγγραφές που κάποια από τις προαναφερθείσες μεταβλητές υποδεικνύει ότι έχει γίνει σφάλμα. Κατόπιν αφαιρούμε τις μεταβλητές αυτές από το χώρο χαρακτηριστικών αφού δεν περιέχουν κάποια πληροφορία για την οδηγική συμπεριφορά.

Επιπλέον, υπάρχουν χαρακτηριστικά τα οποία δεν καταγράφονταν συνεχώς από τον προσομοιωτή παρά μόνον όποτε συνέβαινε κάποιο

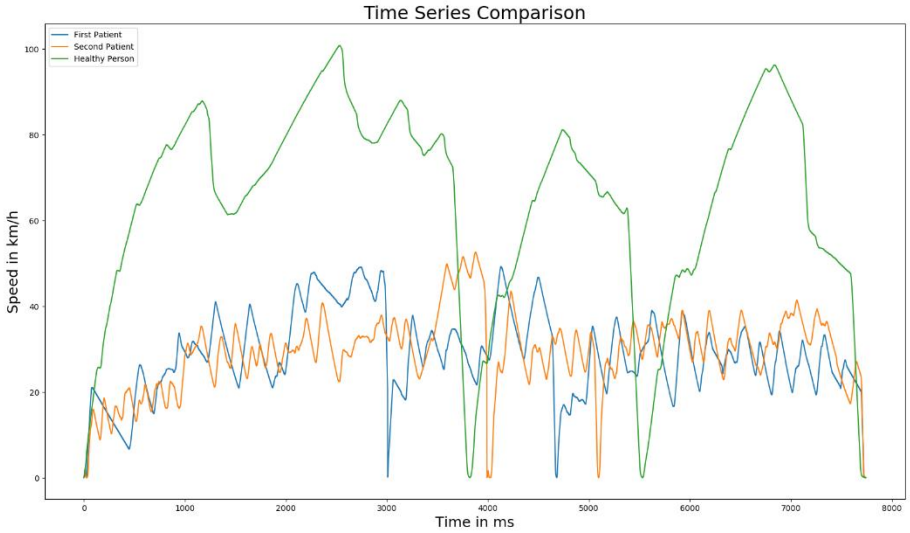
γεγονός. Τέτοιου είδους χαρακτηριστικά αποτελούν οι μεταβλητές *EvVis, EvDist* που φαίνονται στον πίνακα αλλά και οι μεταβλητές *ReactionTime* και *Crash.* Τα χαρακτηριστικά αυτά αφαιρέθηκαν καθώς υπήρχαν μόνο σε ορισμένες εγγραφές από τις χιλιάδες που ανήκουν σε κάθε διαδρομή και κατά συνέπεια αφαιρέθηκαν από την ανάλυση.

Τέλος, κάθε εγγραφή έχει και δύο στήλες *Time* και *Dist* οι οποίες σηματοδοτούν τη χρονική στιγμή που γίνεται η καταγραφή και την διανυθείσα μέχρι εκείνη τη στιγμή απόσταση αντίστοιχα. Ακόμα υπάρχουν και οι μεταβλητές *x-pos, y-pos, z-pos* και *Rdist* οι οποίες είναι ουσιαστικά οι συντεταγμένες του αυτοκινήτου κάθε χρονική στιγμή. Οι προαναφερθείσες μεταβλητές λειτουργούν σαν αναγνωριστικά του χωρικού και χρονικού σημείου της διαδρομής. Δεν παρέχουν κάποια πληροφορία για την οδηγική κατάσταση αλλά ορίζουν μία διάταξη για τις εγγραφές ανά συνεδρία. Επομένως μπορούμε να παραλείψουμε την συμμετοχή τους ως χαρακτηριστικά στους αλγορίθμους *multilayer perceptron* και *decision tree*.

3.3 Προσδιορισμός βέλτιστης τεχνικής

Η ενότητα που ακολουθεί περιγράφει την προσπάθεια εύρεσης μίας όσο το δυνατόν περισσότερο αξιόπιστης μεθόδου για την αναγνώριση των ασθενειών. Κάθε μέθοδος απαιτεί ανάλυση δεδομένων και επιλογή κάποιου αλγορίθμου μηχανικής μάθησης ώστε να επιτύχουμε την αυτόματη συμπερασματολογία. Τα μοντέλα μάθησης που δοκιμάστηκαν βασίζονται όλα σε επιβλεπόμενη μάθηση και ο διαχωρισμός δεδομένων εκπαίδευσης και άγνωστων δεδομένων δοκιμής ήταν 70-30 αντίστοιχα.

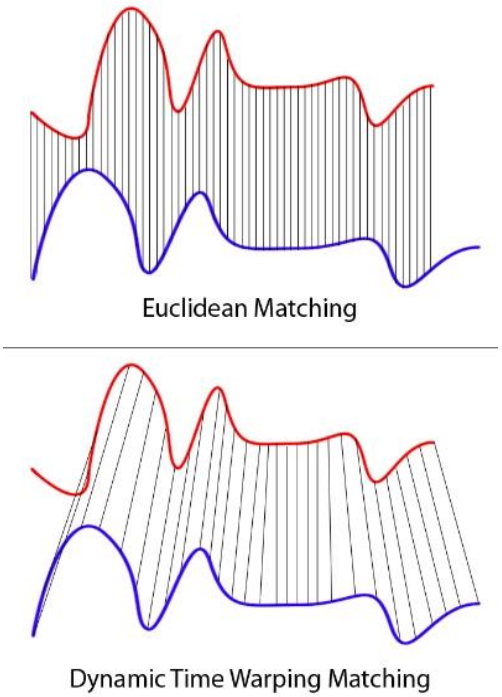
3.3.1 Επιβλεπόμενη μάθηση ανά χρονοσειρά



Σύγκριση χρονοσειρών μεταξύ ενός υγιούς οδηγού και δύο ασθενών

Σε αυτή τη μέθοδο κάθε άτομο το οποίο οδήγησε στον προσομοιωτή αντιπροσωπεύεται από μια χρονοσειρά. Ως χρονοσειρά θεωρούμε την καταγραφή των τιμών τις ταχύτητας με αύξουσα χρονική σειρά. Οι χρονοσειρές αυτές είναι ανισομήκεις καθώς κάποια άτομα χρειάστηκαν περισσότερο ή λιγότερο χρόνο από κάποια άλλα για την ολοκλήρωση της διαδρομής.

Στο παραπάνω διάγραμμα φαίνονται γραφικά οι χρονοσειρές από τρεις οδηγούς. Γίνεται και οπτικά αντιληπτό ότι υπάρχει ‘απόσταση’ μεταξύ των γραφικών απεικονίσεων των διανυσμάτων ταχύτητας ανάμεσα στον υγιή οδηγό σε σχέση με τους ασθενείς. Παράλληλα, φαίνεται να υπάρχει μεγαλύτερη συγγένεια μεταξύ των χρονοσειρών δύο ασθενών οδηγών.



Είναι απαραίτητο για τον αλγόριθμο *k-nearest neighbors* να καθοριστεί ένα μέτρο απόστασης μεταξύ των διανυσμάτων ταχύτητας. Η μετρική που επιλέχθηκε για να γίνει η σύγκριση των χρονοσειρών είναι ο αλγόριθμος *dynamic time warping*. Η επιλογή αυτή έγινε με βάση το γεγονός ότι το μήκος των χρονοσειρών είναι άνισο επομένως δεν θα μπορούσε να χρησιμοποιηθεί η Ευκλείδεια απόσταση αλλά και την ευρεία χρήση της συγκεκριμένης μετρικής σε άλλα πεδία όπως η αναγνώριση φωνής.

3.3.2 Επιβλεπόμενη μάθηση ανά εγγραφή

Σε αυτή την εκδοχή τα δεδομένα παρουσιάζονταν ανά καταγραφή του προσομοιωτή στους αλγορίθμους για να γίνει η απαραίτητη εκπαίδευση τους. Κάθε διάνυσμα δεδομένων είχε επισημανθεί με μία ετικέτα η οποία υποδείκνυε τον τύπο της ασθένειας του οδηγού στον οποίο ανήκει η συγκεκριμένη εγγραφή.

Για να αποφευχθεί η μείωση της απόδοσης των αλγορίθμων τα χαρακτηριστικά που λειτουργούσαν ουσιαστικά σαν *id* όπως το *timestamp* της εγγραφής και η ένδειξη της διανυθείσας απόστασης. Στο διάνυσμα των χαρακτηριστικών που χρησιμοποιήθηκε για την εκπαίδευση υπήρχαν μόνο τα χρονομεταβαλλόμενα χαρακτηριστικά.

Ειδικότερα στον αλγόριθμο του multilayer perceptron έγινε κανονικοποίηση των δεδομένων ώστε να έχουν $\mu=0$ και $\sigma=1$. Κατόπιν, εκτελέστηκε ανάλυση κυρίων συνιστωσών έτσι ώστε να γίνει μείωση της διαστατικότητας στα 10 χαρακτηριστικά.

3.3.3 Βελτίωση μοντέλου με κινούμενο μέσο

Για την παρούσα εργασία επιλέγεται να χρησιμοποιηθεί απλός μέσος αλλά υπολογίστηκε με βάση παρελθοντικές και μελλοντικές τιμές ισόποσα, με σκοπό να ομαλοποιήσουμε τα δεδομένα ώστε τελικά να αυξηθεί η απόδοση των μοντέλων στην ακρίβεια των προβλέψεων. Επιλέγουμε το εύρος του παραθύρου, έστω $n = 2d + 1 > 0$. Επαναπροσδιόρίζουμε κάθε τιμή μιας εγραφής ως εξής:

Έστω x_t η εγγραφή τη χρονική στιγμή t. Θεωρούμε την τιμή

$$x_t^* = \frac{x_{t-d} + x_{t-d+1} + \cdots + x_{t-1} + x_t + x_{t+1} + \cdots + x_{t+d-1} + x_{t+d}}{n}$$

Αντικαθιστούμε την τιμή x_t με την τιμή x_t^*. Χρησιμοποιώντας την ανανεωμένη τιμή επιτυγχάνουμε να ενσωματώσουμε κάποιο μέρος της πληροφορίας -αν και όχι μεγάλο- για τις 'χρονικά γειτονικές' τιμές των μετρήσιμων χαρακτηριστικών. Η ενσωμάτωση αυτή μπορεί να αποδειχτεί ιδιαίτερα βοηθητική καθώς υπάρχει μεγάλη πιθανότητα να υπάρχει συσχέτιση μεταξύ της τρέχουσας τιμής και των χρονικά κοντινών της. Το n ορίστηκε στην τιμή 31 έτσι ώστε να υπολογίζεται ο μέσος όρος από τις 15 προηγούμενες και 15 επόμενες τιμές του κάθε χαρακτηριστικού.

4

Αξιολόγηση

Στην παρούσα ενότητα παρουσιάζονται και επεξηγούνται οι παράμετροι οι οποίες μετρήθηκαν και πάνω στις οποίες βασίστηκε η αξιολόγηση των μοντέλων.

4.1 Παράμετροι αξιολόγησης

4.1.1 Δυαδική Ταξινόμηση

Ως πρώτος στόχος των αλγορίθμων τέθηκε η αναγνώριση της ασθένειας ανεξάρτητα από το είδος της. Αυτό σημαίνει ότι η απόκριση του συστήματος ήταν η πρόβλεψη για το εάν το τρέχων δείγμα ανήκει σε ασθενή ή υγιή οδηγό. Εφ' όσον ο στόχος όμως δεν ήταν η απλή αναγνώριση των εγγραφών αλλά η απόφαση εάν κάποιος συγκεκριμένος οδηγός του δείγματος εξέτασης, η τάξη στην οποία ανατέθηκε ο οδηγός ήταν η ίδια με αυτή που αποφασίστηκε για τις περισσότερες εγγραφές του δείγματος του.

Πριν ορίσουμε μετρικές για την αξιολόγηση των δυαδικών ταξινομήσεων είναι χρήσιμο να ορίσουμε τις εξής μεταβλητές:

- ***N*** (*Negative)* : Ο αριθμός των υγιών ατόμων στο δείγμα εξέτασης
- ***P*** *(Positive)* : Ο αριθμός των ασθενών ατόμων στο δείγμα εξέτασης
- ***TN*** *(True Negative)* : Ο αριθμός των υγιών ατόμων που αναγνωρίστηκαν

— $\boldsymbol{TP}$ *(True Positive)* : Ο αριθμός των ασθενών ατόμων που αναγνωρίστηκαν

— $\boldsymbol{FN}$ *(False Negative)* : Ο αριθμός των ασθενών ατόμων που δεν αναγνωρίστηκαν

— $\boldsymbol{FP}$ *(False Positive)* : Ο αριθμός των υγιών ατόμων που διαγνώστηκαν ως ασθενή

Ορισμένες μετρικές για την αξιολόγηση της δυαδικής ταξινόμησης είναι οι εξής:

- $\boldsymbol{Accuracy} = \frac{\boldsymbol{TP+TN}}{\boldsymbol{P+N}}$: ποσοστό σωστών προβλέψεων
- $\boldsymbol{Precision} = \frac{\boldsymbol{TP}}{\boldsymbol{TP+FP}}$: ποσοστό των ατόμων που διαγνώστηκαν και ήταν όντως ασθενείς
- $\boldsymbol{Recall} = \frac{\boldsymbol{TP}}{\boldsymbol{TP+FN}} = \frac{\boldsymbol{TP}}{\boldsymbol{P}}$: ποσοστό ασθενών που αναγνωρίστηκαν
- $False\ Negative\ Rate\ (\boldsymbol{FNR}) = \frac{FN}{FN+TP} = \frac{\boldsymbol{FN}}{\boldsymbol{P}} = \boldsymbol{1 - Recall}$: ποσοστό των ασθενών που δεν κατάφεραν να εντοπιστούν από τον αλγόριθμο
- $False\ Positive\ Rate\ (\boldsymbol{FPR}) = \frac{FP}{FP+TN} = \frac{\boldsymbol{FP}}{\boldsymbol{N}}$: ποσοστό των υγιών ατόμων που επισημάνθηκαν ως ασθενή
- $\boldsymbol{F1\ Score} = \frac{\boldsymbol{2 \cdot Precision \cdot Recall}}{\boldsymbol{2 \cdot Precision + Recall}}$: μετρική η οποία αποτελεί ουσιαστικά τον αρμονικό μέσο του $Precision$ και του $Recall$.

Οι μετρήσεις για την αξιολόγηση των αλγορίθμων έγιναν με βάση τις μετρικές $\boldsymbol{Precision}$, η οποία ουσιαστικά αναφέρεται στο πόσο αξιόπιστη είναι η διάγνωση ασθένειας από το μοντέλο, $\boldsymbol{Recall}$, η οποία δείχνει την ικανότητα του μοντέλου να εντοπίζει τους ασθενείς και $\boldsymbol{Accuracy}$ και $\boldsymbol{F1}$ ως πιο γενικές μετρικές.

4.1.2 Ταξινόμηση πολλαπλών κατηγοριών

Οι μετρικές στην ταξινόμηση πολλαπλών κατηγοριών είναι όμοιες με αυτές της δυαδικής ταξινόμησης με τη διαφορά ότι εδώ αναφέρονται σε κάθε διαφορετική κατηγορία ξεχωριστά. Εξαιρείται η μετρική της ακρίβειας (*accuracy*) η οποία δείχνει την επίδοση του μοντέλου ανεξάρτητα από την κατηγορία.

4.2 Σύστημα αξιολόγησης

Σε κάθε αλγόριθμο αρχικά παρουσιάζονταν το 70% των δεδομένων βάσει των οποίων γινόταν η εκπαίδευση. Το υπόλοιπο 30% παρέμενε ως σύνολο εξέτασης του μοντέλου που είχε δημιουργηθεί μετά το πέρας της διαδικασία μάθησης.

Το αρχικό στάδιο ήταν η αξιολόγηση των μοντέλων στη δυαδική ταξινόμηση με βάση τις προαναφερθείσες μετρικές. Ως επόμενο βήμα εξετάσαμε την συμπεριφορά των ίδιων αλγορίθμων εκπαιδεύοντας τους ώστε να δημιουργήσουν μοντέλα όχι μόνο για τον εντοπισμό ή μη ασθένειας αλλά και για την ακριβή διάγνωση της ασθένειας από την οποία πάσχει ο εκάστοτε οδηγός.

Αντίστοιχα βήματα έγιναν και για την αξιολόγηση των αλγορίθμων πάνω στα ίδια δεδομένα τα οποία είχαν υποστεί επεξεργασία με την μέθοδο του *moving average*. Τα αρχικά δεδομένα που επιλέξαμε να παρουσιαστούν στους αλγορίθμους ήταν αυτά των πειραμάτων που έγιναν στον επαρχιακό δρόμο, με ήπια κίνηση και χωρίς απόσπαση της προσοχής των οδηγών.

Τέλος, αφού επιλέξαμε το μοντέλο με την καλύτερη επίδοση για *binary* και *multiclass classification* προσπαθήσαμε να εξάγουμε συμπεράσματα για την επίδραση των παραμέτρων του περιβάλλοντος –περιοχή, κίνηση, απόσπαση προσοχής– στην αποτελεσματική διάγνωση.

5

Αποτελέσματα

Στο παρόν κεφάλαιο περιγράφονται και αναλύονται ορισμένα τεχνικά χαρακτηριστικά της διπλωματικής εργασίας καθώς και τα αποτελέσματα των πειραμάτων μας.

5.1 Πλατφόρμες και λογισμικό

Εδώ περιγράφονται τα χαρακτηριστικά της συγκεκριμένης υλοποίησης, όπως η πλατφόρμα ανάπτυξης και εκτέλεσης, τα προγραμματιστικά εργαλεία, οι απαιτήσεις της εφαρμογής σε hardware.

Η ανάπτυξη του κώδικα έγινε στο λειτουργικό σύστημα *Linux Ubuntu 16.04* το οποίο είναι open source και διανέμεται από την *Canonical*. Η γλώσσα που επιλέχθηκε για την εργασία ήταν η *Python* στην έκδοση 3.5 σε συνδυασμό με το framework του *Apache Spark* 2.1.0. Από το Spark χρησιμοποιήθηκε η βιβλιοθήκη *ML,* η οποία αποτελεί συνέχεια και εξέλιξη της *MLlib*, η οποία παρείχε τους αλγορίθμους για τα μοντέλα των *multilayer perceptron* και *decision tree*. Από την *MLlib* χρησιμοποιήθηκαν ορισμένα εργαλεία για την εξαγωγή των μετρικών επί των αποτελεσμάτων καθώς κάποιες από αυτές δεν ήταν διαθέσιμες στη νεότερη *ML*. Ο αλγόριθμος *k-nearest neighbors* υλοποιήθηκε εξ ολοκλήρου καθώς δεν υπήρχε έτοιμη υλοποίηση από τη βιβλιοθήκη. Απαραίτητη προϋπόθεση για την σωστή λειτουργία του *Spark* ήταν και η εγκατάσταση του *Apache Hadoop* 2.7.3. Όσον αφορά στο hardware, η εκτέλεση έγινε σε *4πύρηνο*

Intel επεξεργαστή *Nehalem* με δυνατότητα *multithreading (2 threads/core)* επομένως ήταν διαθέσιμοι 8 λογικοί πυρήνες επεξεργασίας. Η μνήμη που είχαμε στην διάθεσή μας ήταν *8GB*. Όλα τα απαραίτητα προγραμματιστικά εργαλεία εγκαταστάθηκαν σε ένα image στον *OpenStack* του εργαστηρίου. Η εκτέλεση των scripts στο engine του *Spark* γινόταν μέσω του pyspark.

5.2 Λεπτομέρειες υλοποίησης

5.2.1 Δυναμική Χρονική Στρέβλωση

Ο αλγόριθμος δυναμικής χρονικής στρέβλωσης είναι ο εξής:

Έστω δύο ακολουθίες *s, t* με μήκος *n, m* αντίστοιχα.

```
int DTWDistance(s, t){
        DTW = int[n, m]

        for i = 1 to n
                DTW[i, 0] = inf
        for i = 1 to m
                DTW[0, i] = inf
        DTW[0, 0] = 0

        for i = 1 to n
                for j = 1 to m
                        cost = d(s[i], t[j])
                        DTW[i, j] = cost + min(DTW[i-1, j],
                                               DTW[i, j-1],
                                               DTW[i-1, j-1])
        return DTW[n, m]
}
```

Επειδή η πολυπλοκότητα του αλγορίθμου είναι *O(nm)* για τον υπολογισμό της μετρικής χρησιμοποιήσαμε μια βιβλιοθήκη της python η οποία παρείχε μια βελτιστοποίηση του αλγορίθμου (fastdtw) [].

5.2.2 Κινούμενος Μέσος

Για τον μετασχηματισμό των δεδομένων με την τεχνική του κινούμενου μέσου χρησιμοποιήθηκε το API του *Spark*. Το παρακάτω κομμάτι κώδικα δείχνει την υλοποίηση

```
windowSpec = Window.partitionBy('personId').orderBy('time').rowsBetween(-15,15)
attributes = ['rspur', 'ralpha', 'speed', 'brk', 'acc', 'clutch', 'gear', 'rpm', 'hway', 'dleft', \
        'dright', 'wheel', 'thead', 'ttl', 'ttc', 'accLat', 'accLon']
augmData = data.select('personId', 'disease', \
                *(func.avg[attr].over(windowSpec).alias(attr) \
                        for attr in attributes))
```

5.2.3 k-Εγγύτεροι Γείτονες

Όπως αναφέρθηκε και προηγουμένως το API του *Spark* δεν προσέφερε έτοιμη υλοποίηση του αλγορίθμου *k-εγγύτερων γειτόνων* επομένως υλοποιήθηκε σύμφωνα με τον παρακάτω ψευδοκώδικα:

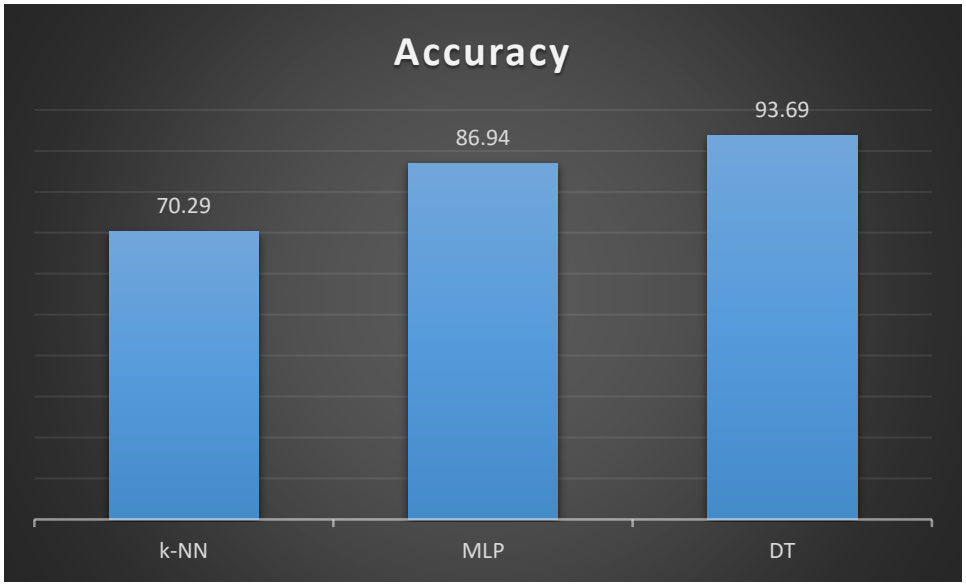
```
results = empty
predictions = empty
for t in test testData:
        for s in trainData:
                dist = fastdtw(s, t)
                add (s, dist) to results
        sortedResults = results.sortByDistance()
        if (k == 1) then:
                prediction = sortedResults.first().label
        else:
                count[1…numberOfLabels] = 0
                for t in sortedResults.take(k)
                        count[t.label]++
                prediction = argmax(count)
add (t, prediction) to predictions
```

5.3 Αποτελέσματα

Παρακάτω φαίνονται τα αποτελέσματα των μετρήσεων που έγιναν στα μοντέλα:

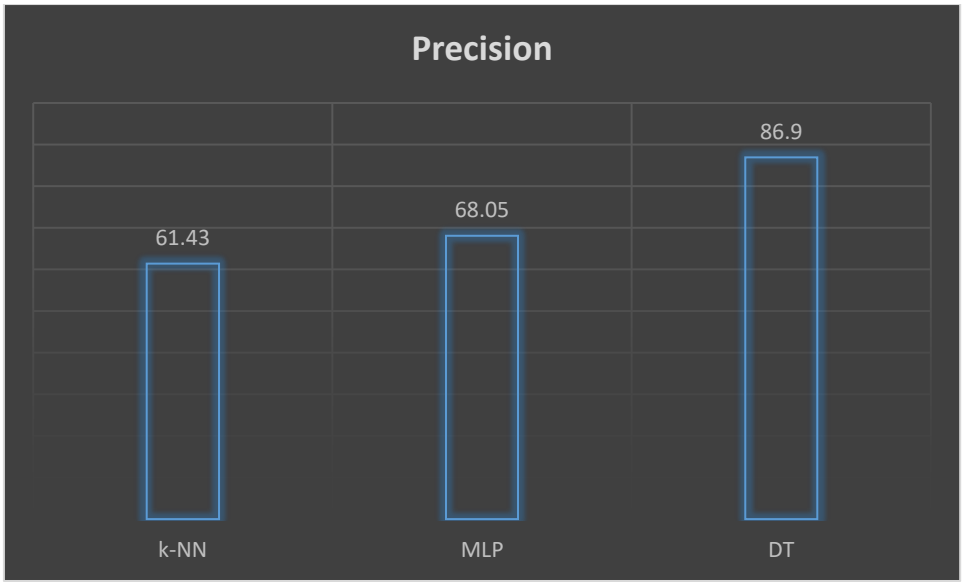
5.3.1 Δυαδική Ταξινόμηση

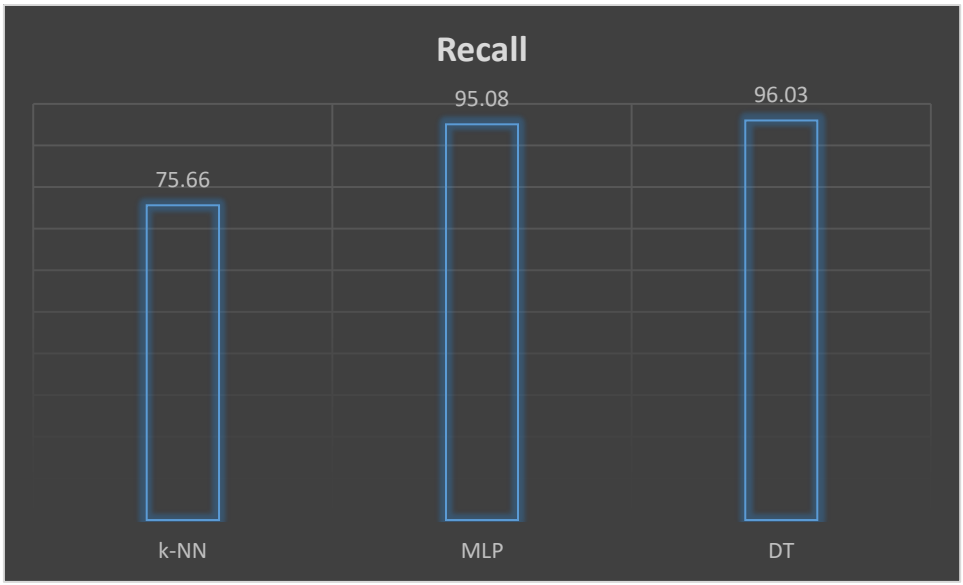
Αρχικά γίνονται μετρήσεις και συγκρίνεται η απόδοση σε *δυαδική ταξινόμηση:*

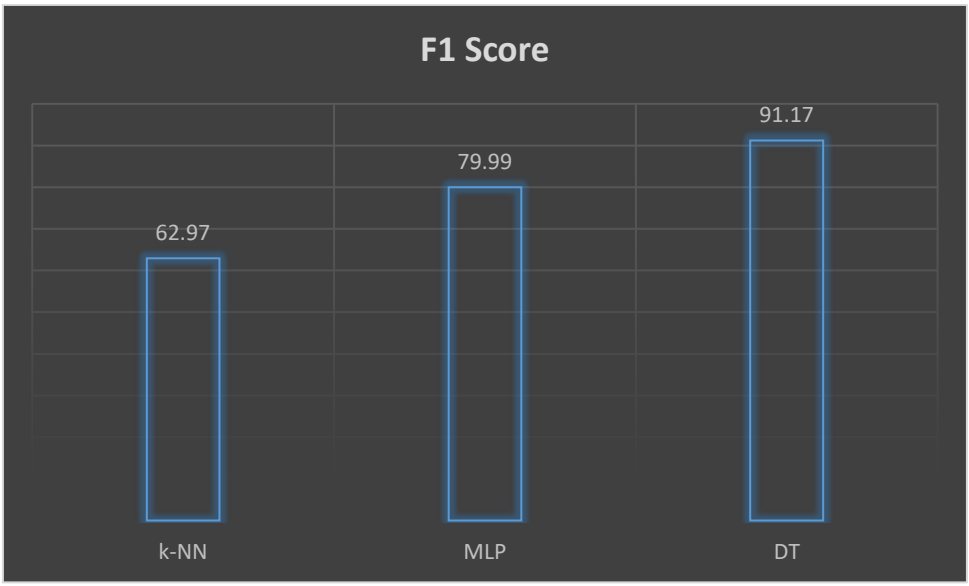


Παρατηρούμε ότι την καλύτερη απόδοση στη γενικότερη μετρική *accuracy* την έχει το *decision tree classifier* με ποσοστό σωστής αναγνώρισης **93.69***%*. Επειδή μας ενδιαφέρει όμως και πιο συγκεκριμένα η απόδοση των μοντέλων στην αναγνώριση ασθενών παραθέτουμε και τα παρακάτω διαγράμματα των μετρικών που αφορούν στην απόδοση των μοντέλων για την αναγνώριση των ασθενών.

Η υπεροχή του *decision tree* είναι εμφανής και στα ακόλουθα διαγράμματα. Έχει πολύ καλή επίδοση τόσο στην αξιοπιστία της διάγνωσης *(precision* – **86.9***%)* όσο και στην δυνατότητα αναγνώρισης τους *(recall* – **96.03***%).* Παρατηρούμε επίσης ότι και το *multilayer perceptron* αποδίδει αρκετά καλά – **95.08***%* – στη δυνατότητα εντοπισμού της ασθένειας.

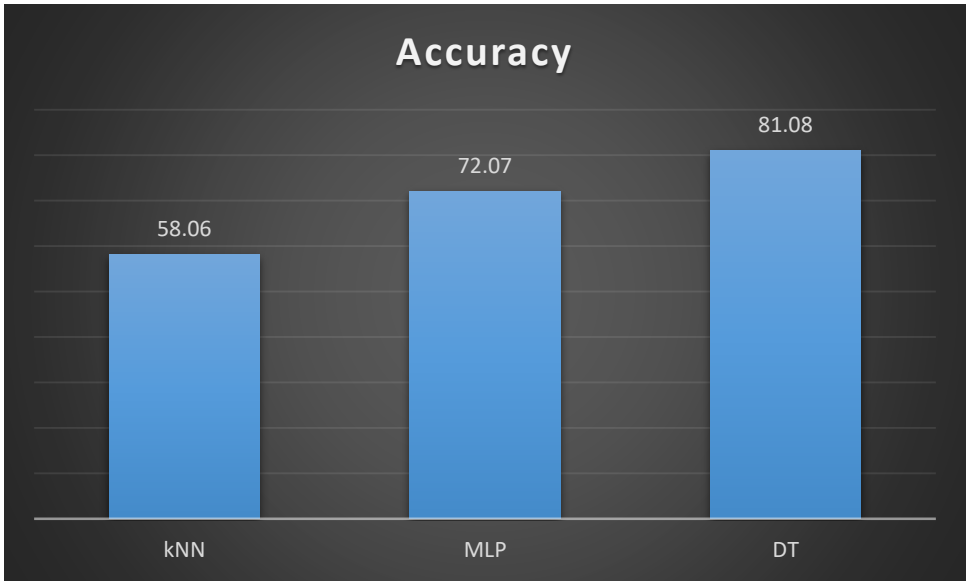

Precision
61.43
68.05
86.9
k-NN
MLP
DT


Recall
75.66
95.08
96.03
k-NN
MLP
DT


F1 Score
62.97
79.99
91.17
k-NN
MLP
DT

5.3.2 Ταξινόμηση πολλαπλών κατηγοριών

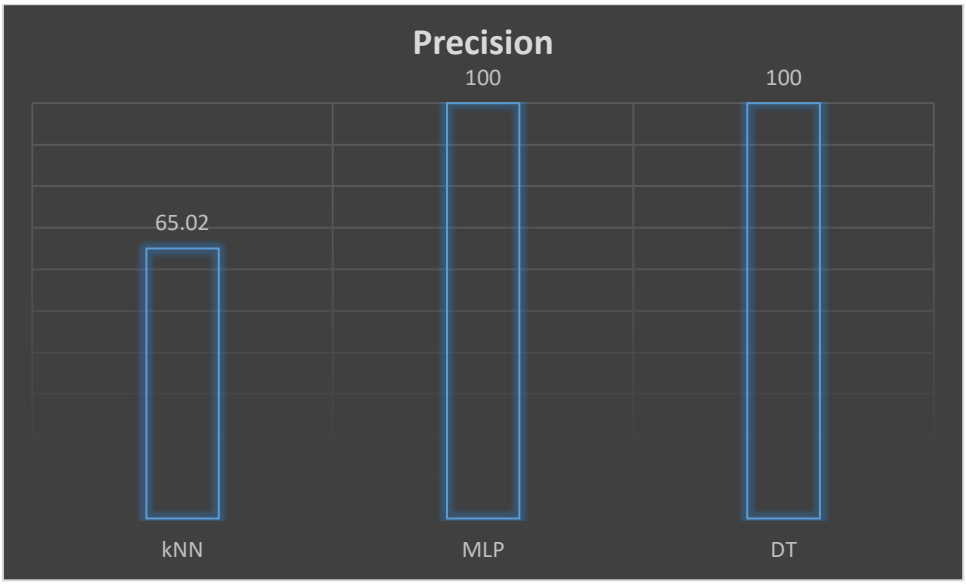
Σε δεύτερο στάδιο εξετάζουμε και παραθέτουμε την απόδοση στην σωστή κατηγοριοποίηση των δειγμάτων σε πολλαπλές κατηγορίες.

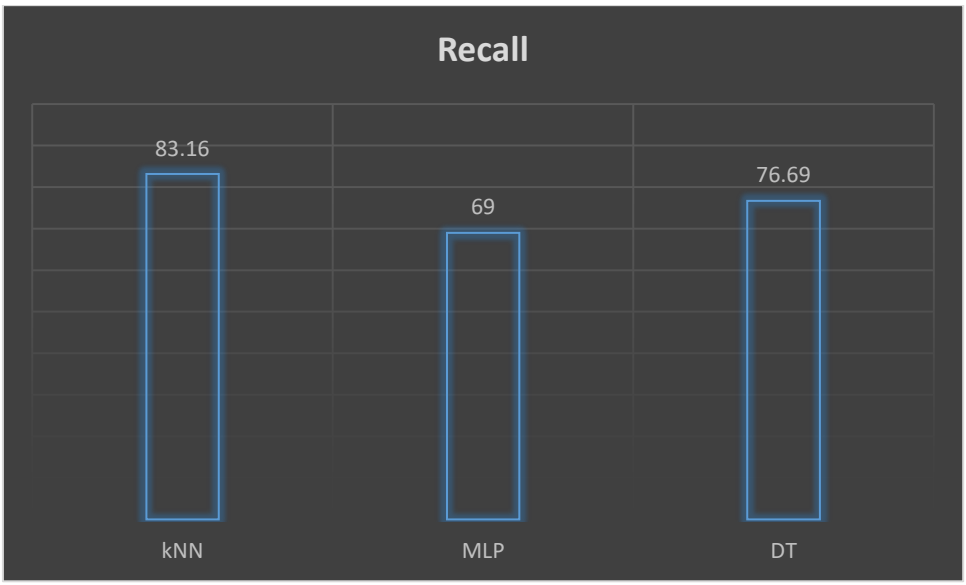


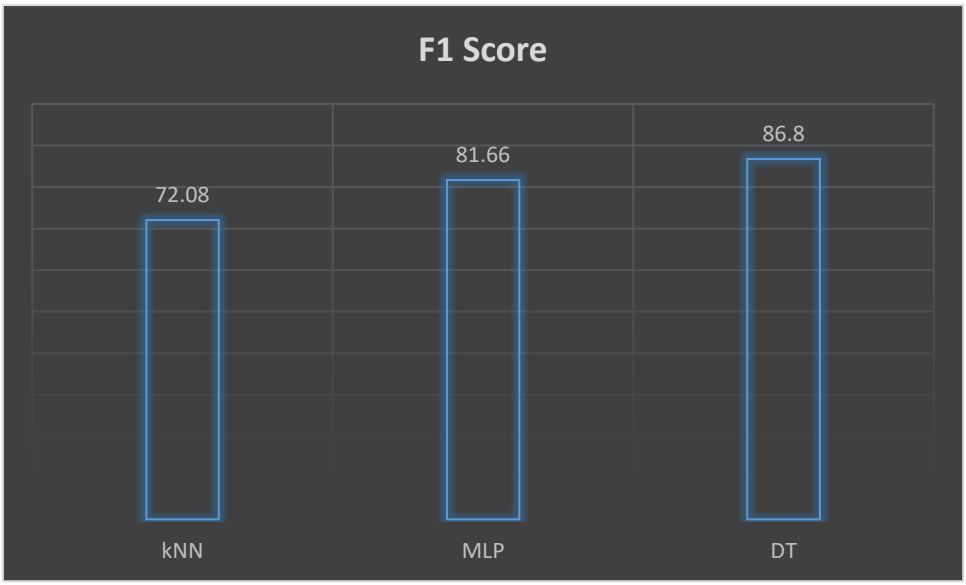
Όπως προηγουμένως, η συνολική αναγνωριστική δυνατότητα του *decision tree* υπερέχει των άλλων δύο ταξινομητών. Επιπρόσθετα, η επίδοση των αντίστοιχων μοντέλων που κατασκευάστηκαν για *multiclass classification* είναι μικρότερη σε κάθε περίπτωση το οποίο είναι αναμενόμενο αφού είναι πιο περίπλοκη διεργασία η οποία απαιτεί καλύτερη προσέγγιση της επιθυμητής συνάρτησης ταξινόμησης.

Στα διαγράμματα που ακολουθούν στις επόμενες σελίδες παρουσιάζεται η επιμέρους απόδοση κάθε μοντέλου σε κάθε κατηγορία που μπορούσε να αναγνωριστεί. Είναι εμφανές από το *F1 Score* ότι η πιο δύσκολα αναγνωρίσιμη κατηγορία ήταν η *Ήπια Γνωστική Ανεπάρκεια (MCI)* ενώ πιο εύκολα εντοπίζονται τα υγιή άτομα. Οι αλγόριθμοι *multilayer peceptron* και *decision tree* φαίνονται άψογοι στο *precision* εκτίμησης υγιών ατόμων και στο *recall* των ασθενειών. Ο αλγόριθμος *k-nearest neighbors* παρόλο που είναι πίσω στις περισσότερες μετρήσεις εμφανίζει την καλύτερη επίδοση στο *precision* αναγνώρισης οδηγών πάσχοντες από *Ήπια Γνωστική Ανεπάρκεια (MCI).*

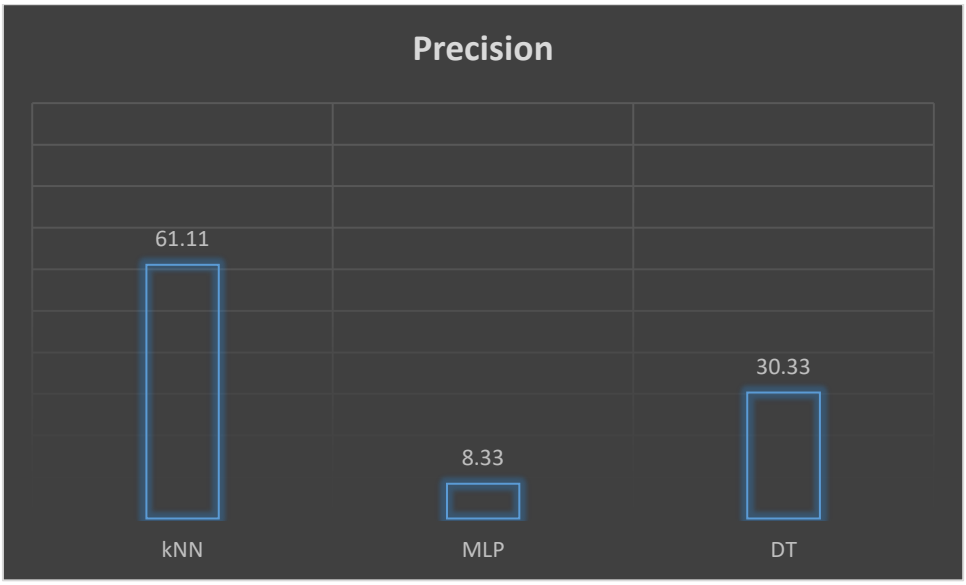
5.3.2.1 *Υγιή δείγματα*

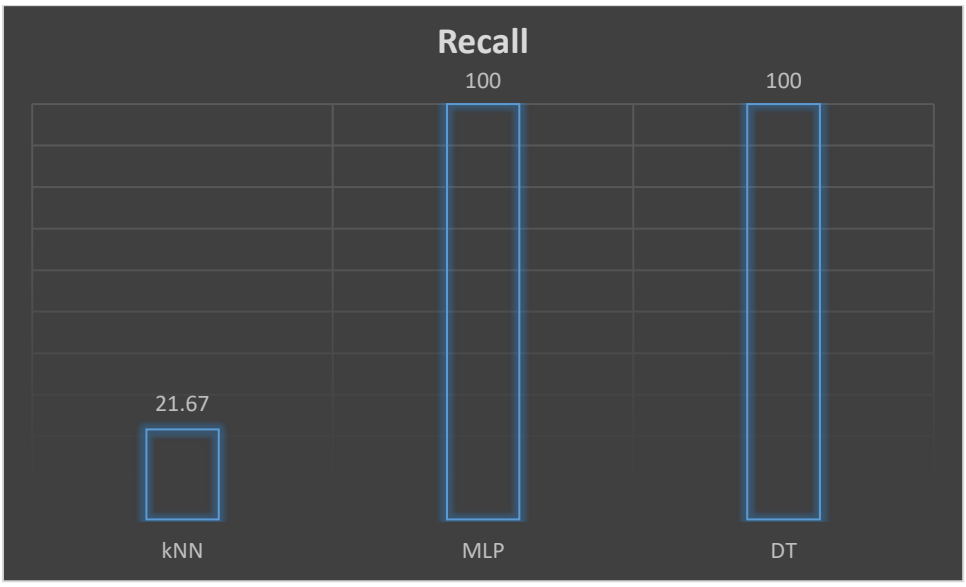


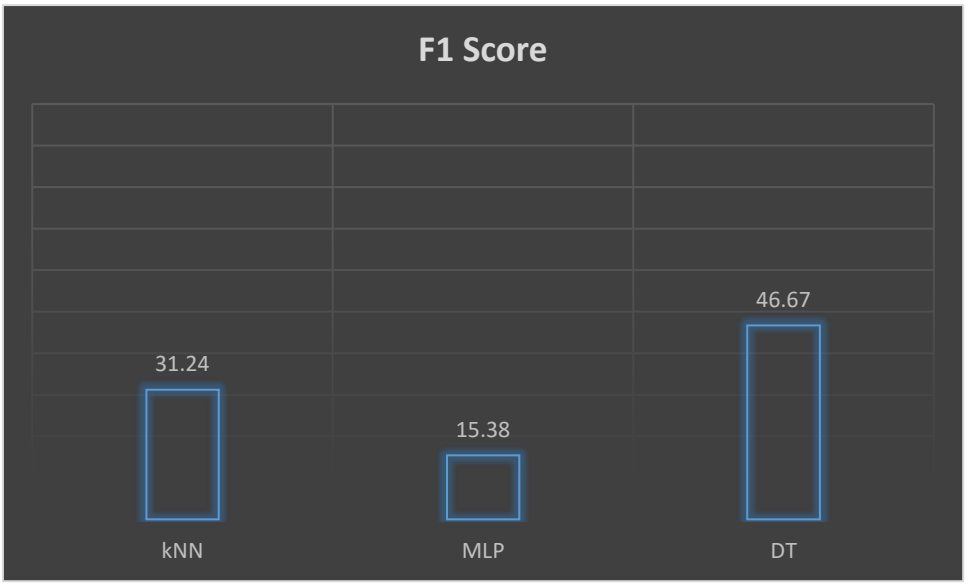




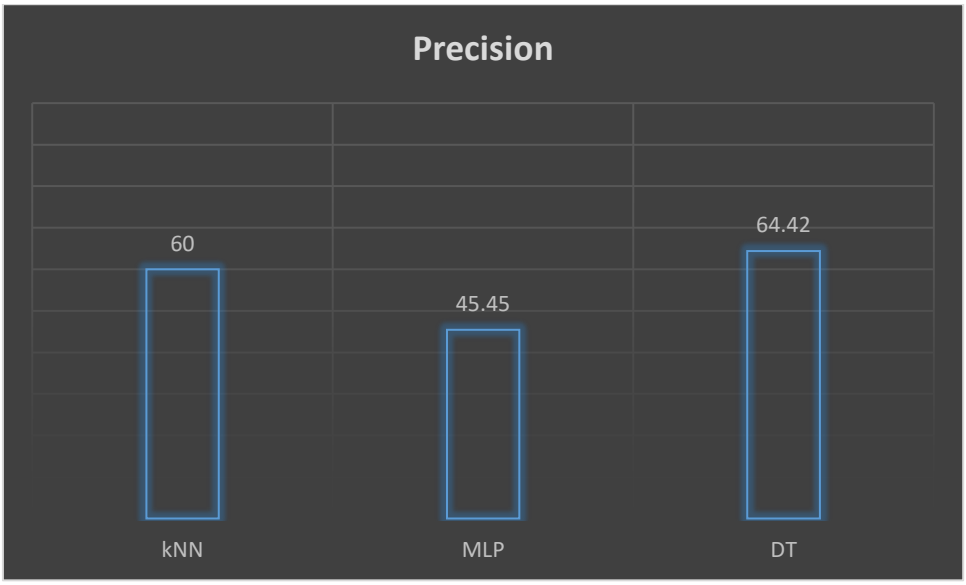
5.3.2.2 *Ήπια Γνωστική Ανεπάρκεια (Mild Cognitive Impairment – MCI)*

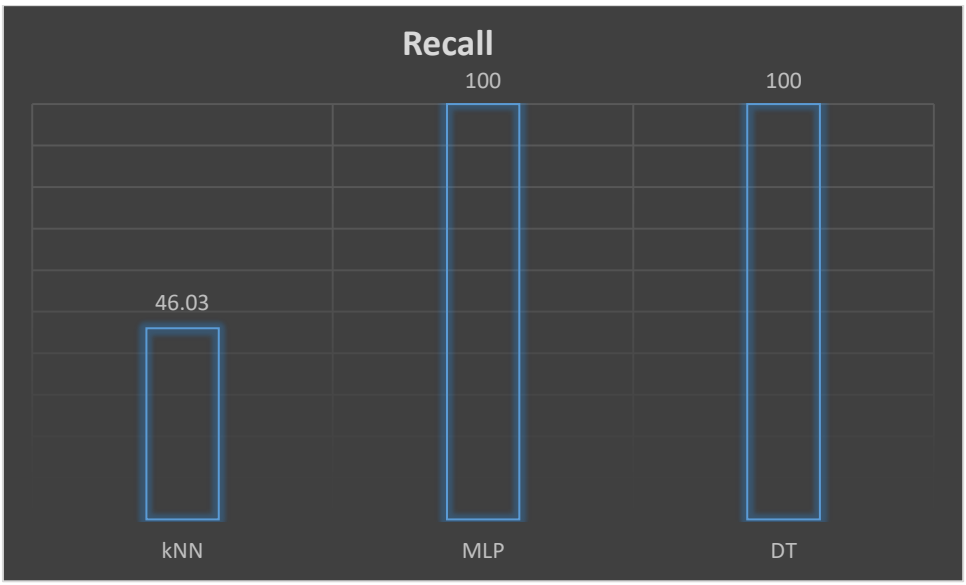


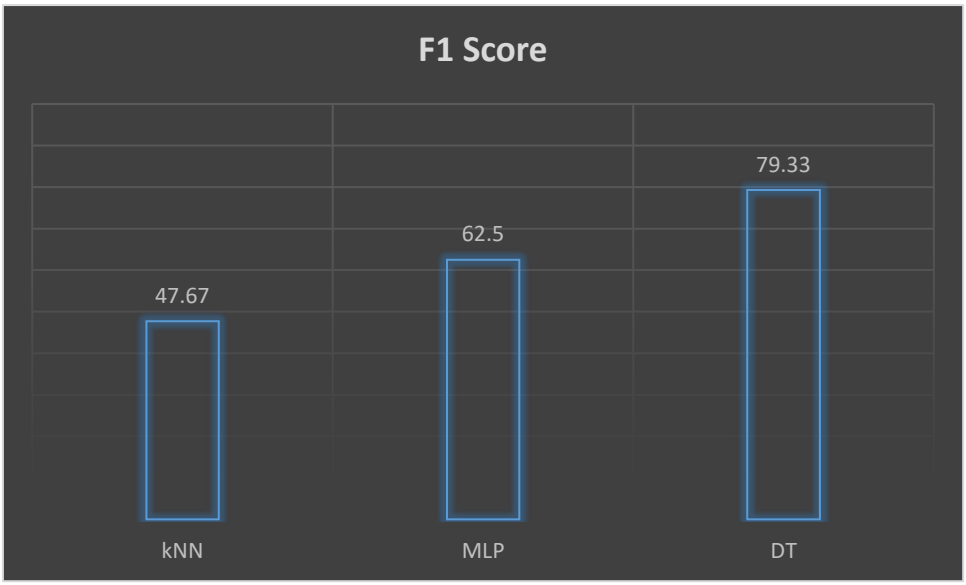




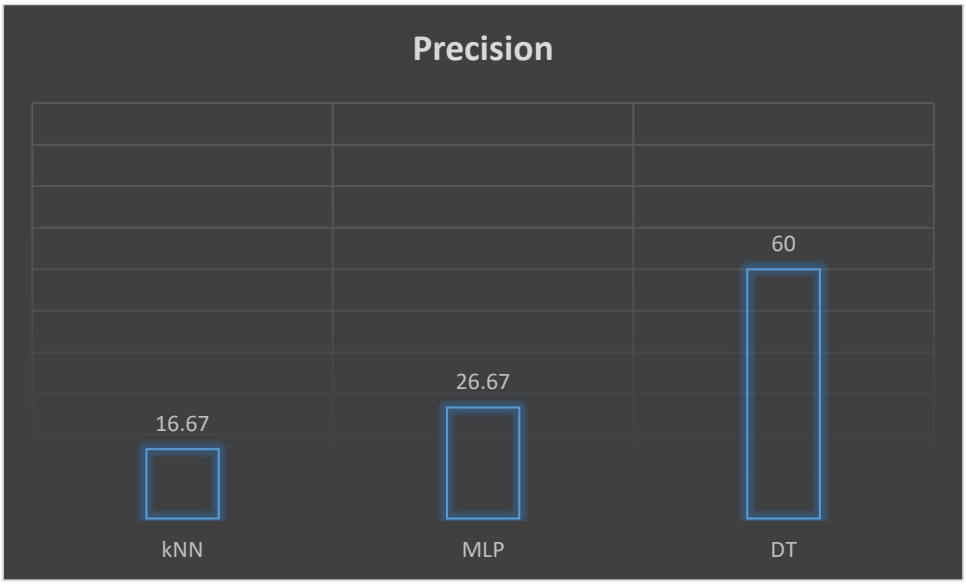
5.3.2.3 *Νόσος Parkinson*

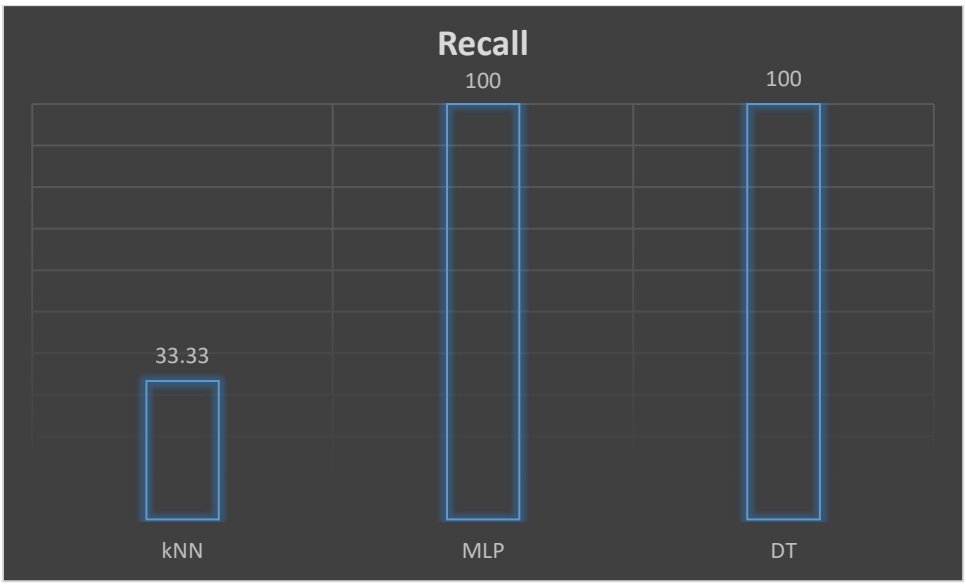


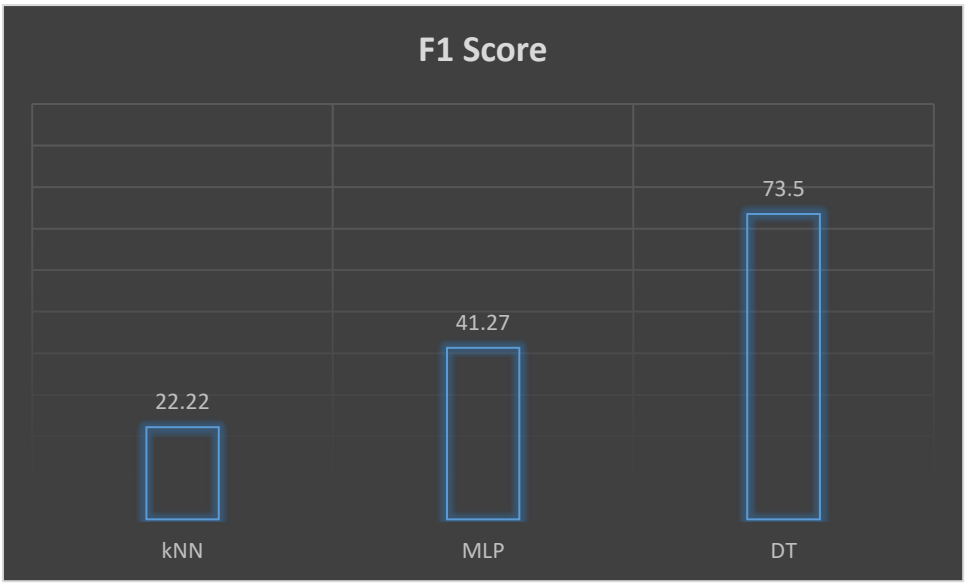




5.3.2.4 *Νόσος Alzheimer*

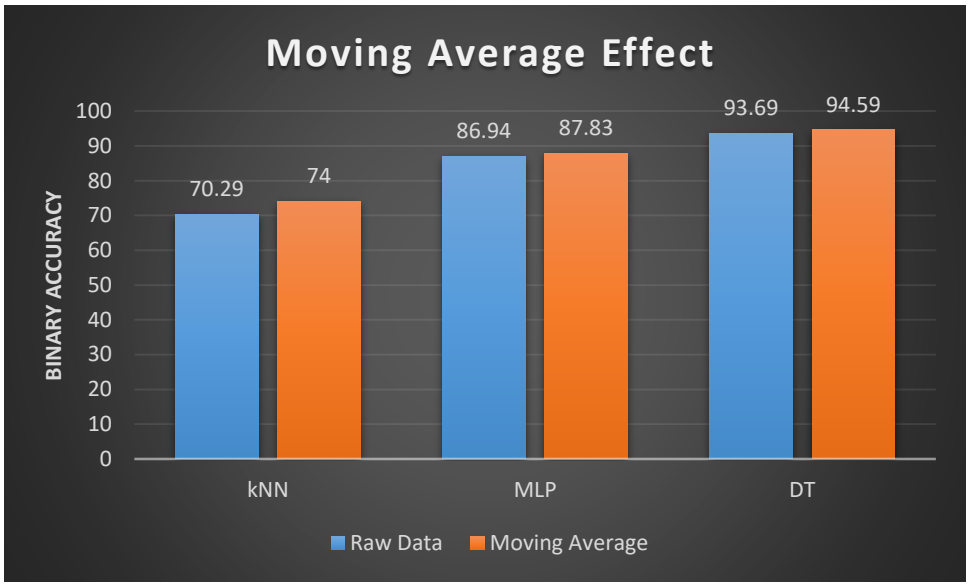


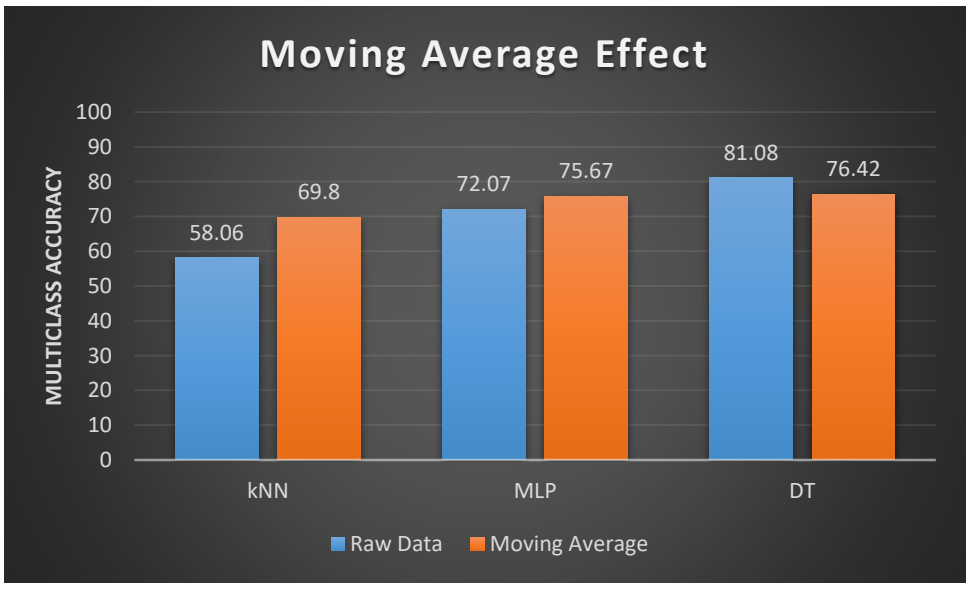




5.3.3 *Επαύξηση Χαρακτηριστικών*

Σε αυτήν την ενότητα παρατίθενται οι μετρήσεις που έγιναν στην ακρίβεια *(accuracy)* των ίδιων μοντέλων αλλά μετά από εκπαίδευση που έγινε σε επαυξημένα δεδομένα *(augmented features)*. Όπως εξηγήθηκε στην ενότητα 3.3.3 οι τιμές των χαρακτηριστικών σε μια χρονική στιγμή αντικαταστάθηκαν από τον μέσο όρο των 15 προηγούμενων και 15 επόμενων τιμών των αντίστοιχων χαρακτηριστικών. Στα παρακάτω διαγράμματα φαίνεται η επίδραση που είχε η προεπεξεργασία των δεδομένων με την τεχνική του *κινούμενου μέσου*.





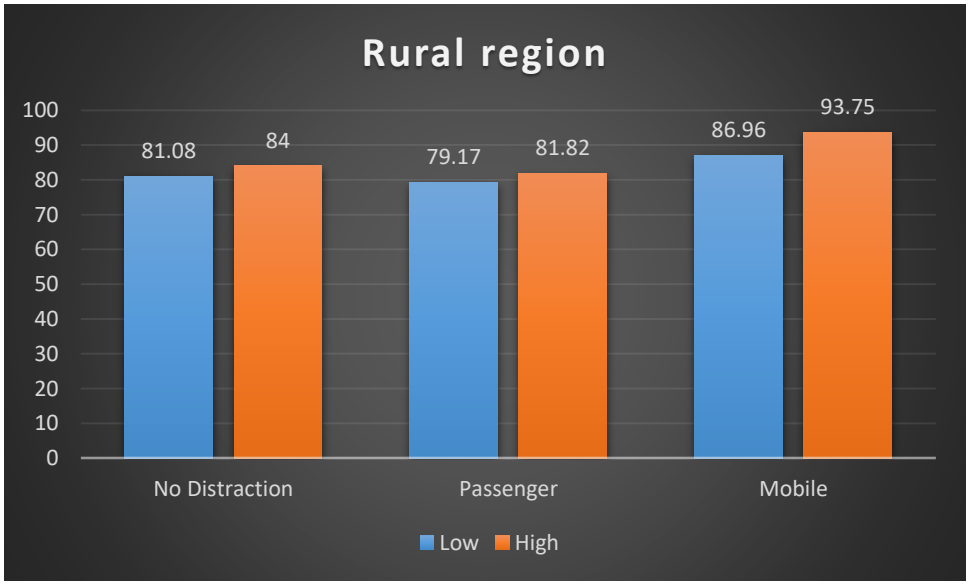
Από τα διαγράμματα φαίνεται ότι εκπαιδεύοντας τα μοντέλα με τα δεδομένα που προέκυψαν μετά την εφαρμογή του *moving average*, έχουν καλύτερη συμπεριφορά στην ακρίβεια των προβλέψεων τους. Η βελτίωση αυτή είναι ιδιαίτερα εμφανής στο μοντέλο των *k-nearest neighbors* και μικρότερη στο *multilayer perceptron*.

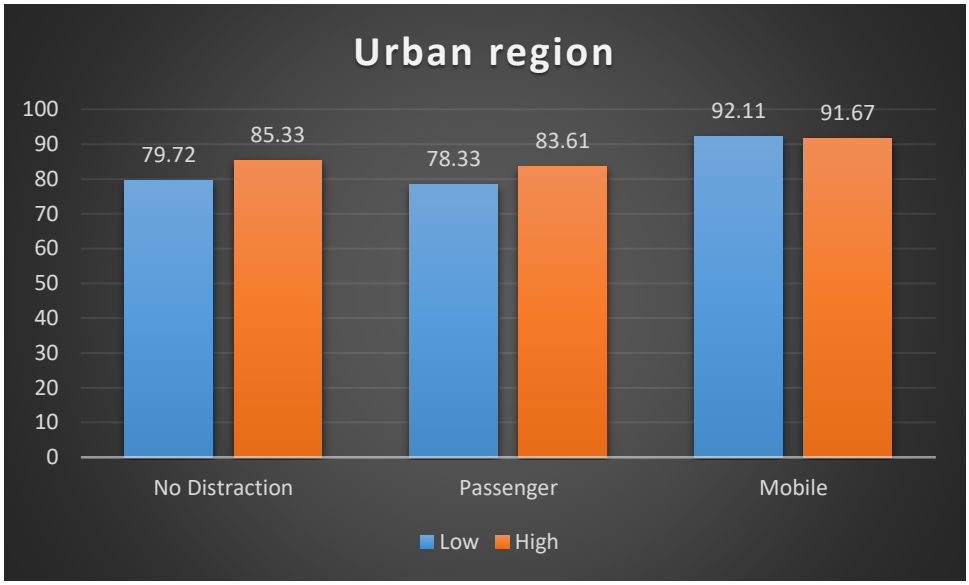
Το *decision tree* παρουσιάζει μια ιδιομορφία καθώς, ενώ στο *binary classification* υπάρχει μια μικρή αύξηση στην ακρίβεια των προβλέψεων, στο *multiclass* παρουσιάζει μείωση η οποία ίσως οφείλεται σε υπερπροσαρμογή πάνω στα δεδομένα *(overfitting)*.

5.3.4 Επίδραση των παραμέτρων του περιβάλλοντος

Στα πλαίσια της εργασίας κρίθηκε επίσης σκόπιμο να μελετηθεί η επίδραση που είχαν οι παράγοντες του περιβάλλοντος, όπως αν η προσομοίωση οδήγησης έγινε σε αστικό ή επαρχιακό περιβάλλον, αν υπήρχε χαμηλή ή υψηλή κίνηση και τι είδους απόσπαση είχε ο οδηγός κατά τη διάρκεια της προσομοίωσης, στην αναγνώριση των ασθενών. Για να είναι συγκρίσιμα τα αποτελέσματα επιλέχθηκε ο ταξινομητής με την καλύτερα επίδοση ο οποίος ήταν ο *decision tree* ταξινομητής και μετρήθηκε η ακρίβεια του στο *multiclass classification* χωρίς να χρησιμοποιείται ο *moving average* για προεπεξεργασία.

Τα αποτελέσματα των μετρήσεων φαίνονται στα παρακάτω διαγράμματα.

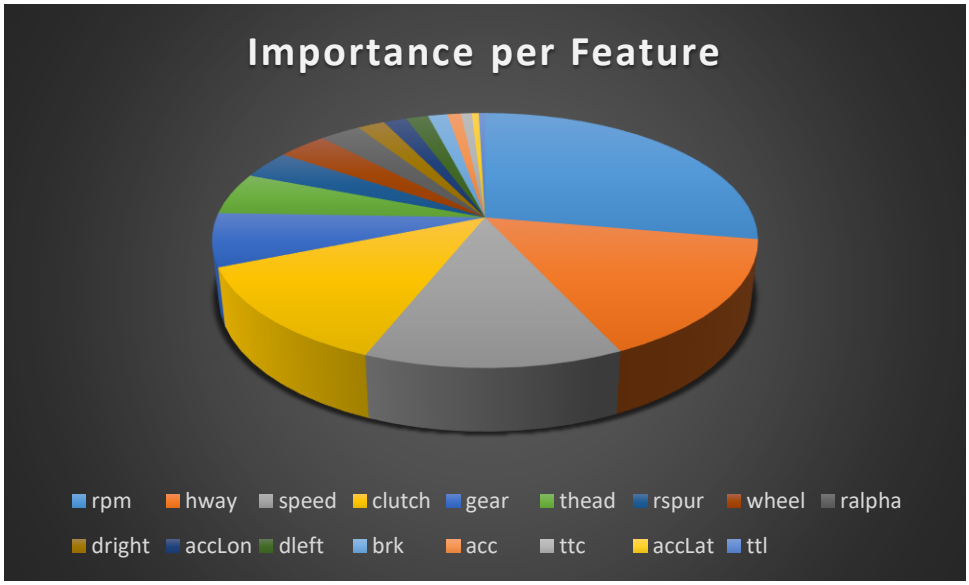




Οι μετρήσεις έδειξαν ότι η περιοχή η οποία προσομοιώθηκε δεν έχει ουσιαστικά κάποια επίδραση στην απόδοση. Ωστόσο, η ποσότητα της κίνησης φαίνεται να επηρεάζει την οδηγική συμπεριφορά και συγκεκριμένα το μοντέλο είχε καλύτερη συμπεριφορά αφού είχε εκπαιδευτεί με δεδομένα από προσομοίωση οδήγησης σε συνθήκες υψηλής κίνησης. Την μεγαλύτερη επίδραση είχε η απόσπαση του οδηγού κατά τη διάρκεια της προσομοίωσης και συγκεκριμένα στην οδήγηση με ταυτόχρονη ομιλία στο κινητό τα αποτελέσματα ήταν πολύ καλύτερα. Η καλύτερη απόδοση σημειώθηκε στην εκπαίδευση του μοντέλου με δεδομένα από την επαρχιακή διαδρομή σε συνθήκες υψηλής κίνησης και με απόσπαση του οδηγού σε συνομιλία στο κινητό.

5.3.5 Βαρύτητα χαρακτηριστικών

Ως τελικό βήμα στη διαδικασία ήταν η επισκόπηση της σημαντικότητας των χαρακτηριστικών όπως αποτυπώθηκε από το μοντέλο του *decision tree* ταξινομητή.



Στο παραπάνω γράφημα φαίνεται η σημαντικότητα κάθε χαρακτηριστικό όπως αποτυπώθηκε μετά την εκπαίδευση του *δέντρου απόφασης* για την κατηγοριοποίηση πολλαπλών κλάσεων. Το **75**% της βαρύτητας τον αποφάσεων βρίσκεται στις μεταβλητές *rpm, hway, speed, clutch* και *gear*. Οι μεταβλητές *hway* και *speed* είναι αναμενόμενο να έχουν μεγάλη βαρύτητα στο μοντέλο καθώς αντιστοιχούν στην απόσταση που διατηρεί ο οδηγός από το προπορευόμενο όχημα και στην ταχύτητα που διατηρεί αντίστοιχα. Τα στοιχεία *rpm, clutch, gear* αντιστοιχούν στην μηχανολογική κατάσταση του αυτοκινήτου κατά τη διάρκεια της διαδρομής.

5.4 Σύνοψη συμπερασμάτων αξιολόγησης

Τα αποτελέσματα των μετρήσεων υποδεικνύουν ως καλύτερο μοντέλο ταξινόμησης το *δέντρο απόφασης*. Ο συγκεκριμένος ταξινομητής έδωσε τα καλύτερα αποτελέσματα από πλευράς ακρίβειας τόσο σε *binary classification* (**93.69%**) όσο και σε *multiclass classification* (**81.08%**). Σε επίπεδο *binary classification* παρουσίασε υψηλή αξιοπιστία διάγνωσης (**86.09%**). Σε επίπεδο *multiclass classification* ήταν άψογο στην αξιοπιστία αναγνώρισης των υγιών οδηγών (**100%**) και στην ικανότητα εντοπισμού (**100%**) κάθε κατηγορίας ξεχωριστά. Το *multilayer perceptron* ακολουθούσε τη συμπεριφορά του *decision tree* αλλά με μικρότερη απόδοση σε ορισμένες κατηγορίες και μετρικές. Και τα δύο μοντέλα είχαν

πολύ κακή απόδοση στην κατηγορία του precision για την *Ήπια Γνωστική Ανεπάρκεια* (MLP – **8.33%**, DT – **30.33%**). Η άσχημη επίδοση αυτή είναι δικαιολογημένη καθώς η *Ήπια Γνωστική Ανεπάρκεια* αποτελεί ένα στάδιο πριν την πιθανή εκδήλωση της νόσου *Alzheimer* η οποία έχει πιο έντονα νευρολογικά συμπτώματα και κατά συνέπεια θα έχει μεγαλύτερη επίδραση στην οδηγική συμπεριφορά. Ο αλγόριθμός των *k-nearest neighbors* φαίνεται πιο αδύναμος σε όλες τις παραπάνω κατηγορίες το οποίο δικαιολογείται ίσως από τη χρήση της χρονοσειράς μόνο ενός εκ των δοθέντων χαρακτηριστικών (speed) για την μέτρηση της 'απόστασης' μεταξύ των δειγμάτων. Παρουσιάζει ωστόσο ένα ενδιαφέρον αποτέλεσμα όσον αφορά στην αξιοπιστία αναγνώρισης οδηγών με *Ήπια Γνωστική Ανεπάρκεια* (**61.11%**). Προσθέτοντας ακόμα ένα στάδιο προεπεξεργασίας πάνω στα δεδομένα, εφαρμόζοντας δηλαδή moving average στο χαρακτηριστικό speed για τον *k-nearest neighbors* αλγόριθμο και σε όλα τα χαρακτηριστικά στο *multilayer perceptron* και στο *decision tree*, υπήρξε μία μικρή βελτίωση στη ακρίβεια πρόβλεψης των μοντέλων με εξαίρεση μία περίπτωση. Η μεγαλύτερη βελτίωση έγινε αισθητή στην περίπτωση του *k-nearest neighbors* ταξινομητή (*binary*: **70.29%** → **74%**, *multiclass*: **58.06%** → **69.8%**). Μια πιθανή εξήγηση για την συγκριτικά μεγαλύτερη βελτίωση είναι ότι ο αλγόριθμος *k-nearest neighbors* βασίζεται στην καθολική εικόνα της χρονοσειράς και ο *moving average* βοηθάει στην ομαλοποίηση της καμπύλης. Επιπλέον, εξετάστηκε η επιρροή που έχουν οι συνθήκες οδήγησης στην ικανότητα εντοπισμού της ασθένειας. Η περιοχή που γίνεται η οδήγηση δεν επηρεάζει ιδιαίτερα αλλά οι συνθήκες κίνησης και η απόσπαση του οδηγού οξύνουν τα χαρακτηριστικά που βοηθούν στη διάγνωση και τον εντοπισμό της ασθένειας. Σε συνθήκες με υψηλή κίνηση στην επαρχιακή περιοχή και με απόσπαση προσοχής σε συνομιλία το κινητό το *decision tree* έδειξε **12.76%** αύξηση (**81.08%** → **93.75%**) στο *multiclass classification* . Τέλος, μέσω του *decision tree* κατόπιν της εκπαίδευσης υπάρχει μια ένδειξη για την βαρύτητα κάθε χαρακτηριστικού στο διαχωρισμό των οδηγών σε κατηγορίες όπως φαίνεται στον παρακάτω πίνακα.

rpm	27.74%
hway	15.33%
speed	12.98%
clutch	12.77%
gear	6.75%
thead	5.31%
rspur	3.59%
wheel	3.50%
ralpha	2.88%
dright	1.93%
accLon	1.62%
dleft	1.57%
brk	1.37%
acc	0.96%
ttc	0.75%
accLat	0.50%
ttl	0.44%

6

Επίλογος

Ακολουθεί μια σύνοψη της παρούσας διπλωματικής εργασίας καθώς και ιδέες για μελλοντική επέκταση και πιθανές εφαρμογές.

6.1 Σύνοψη και συμπεράσματα

Στην παρούσα εργασία προτείνεται μια προσέγγιση μηχανικής μάθησης για διάγνωση ασθενών με δεδομένα που συλλέχθηκαν μέσω λογισμικού κατά τη διάρκεια της οδήγησης τους σε προσομοιωμένο περιβάλλον. Τα μοντέλα που εξετάσαμε και αξιολογήσαμε είχαν εκπαιδευτεί με επιβλεπόμενο τρόπο καθώς υπήρχε από πριν η πληροφορία για την κατηγορία στην οποία ανήκαν τα δεδομένα, δηλαδή αν ήταν υγιής ή ασθενής οδηγός.

Αρχικά, υλοποιήθηκε ένα μοντέλο *k-εγγύτερων γειτόνων (k-nearest neighbors)* το οποίο προέβλεπε την κατάσταση των οδηγών βασιζόμενο στην καθολική εικόνα της χρονοσειράς της ταχύτητας υπολογίζοντας την απόσταση με τον αλγόριθμο *δυναμικής χρονικής στρέβλωσης (dynamic time warping).* Υλοποιήθηκαν επίσης ένα μοντέλο *πολυεπίπεδου perceptron) multilayer perceptron* με δύο επίπεδα 20 κρυφών νευρώνων και ένα *δέντρο απόφασης (decision tree)* μέγιστου βάθους 10 τα οποία έκαναν επίπεδο πρόβλεψης ανά εγγραφή η οποία καθόριζε και την συνολική απόφαση για το δείγμα του εκάστοτε οδηγού. Δοκιμάστηκε επίσης η προσθήκη ενός σταδίου προεπεξεργασίας με *απλό κινούμενο*

μέσο για βελτίωση των δεδομένων πριν την εκπαίδευση των μοντέλων το οποίο απέδωσε κυρίως στην περίπτωση του *k-NN*. Τέλος από το μοντέλο του δέντρου απόφασης, το οποίο μας έδωσε τα καλύτερα αποτελέσματα, εξαγάγαμε ορισμένες μετρικές που επισημαίνουν την βαρύτητα των χαρακτηριστικών στις διακλαδώσεις του δέντρου.

Τα αποτελέσματα των μετρήσεων της απόδοσης των μοντέλων μας επιβεβαιώνουν την υπόθεση ότι υπάρχει η δυνατότητα αναγνώρισης της ασθένειας ενός ατόμου μέσω ανάλυσης της οδηγικής του συμπεριφοράς με αυτόματο τρόπο. Ακόμα και με μοναδικό χαρακτηριστικό τις μετρήσεις της ταχύτητας σε συγκεκριμένη διαδρομή μπορούμε να έχουμε ικανοποιητικά αποτελέσματα με *δυναμική χρονική στρέβλωση* και *k-NN*.

6.2 Μελλοντικές επεκτάσεις

Σε αυτή την εργασία οι παράμετροι των αλγορίθμων και των μοντέλων είναι σταθερές και βασίζονται κυρίως σε εμπειρικά δείγματα. Θα ήταν ενδιαφέρουσα προσθήκη η μελέτη της επίδρασης των διάφορων παραμέτρων που προαναφέρθηκαν για την επίδραση τους στις επιδόσεις των μοντέλων. Επιπρόσθετα, θα μπορούσε να γίνει περεταίρω μελέτη για περισσότερα μοντέλα και τεχνικές βαθιάς μάθησης ώστε να βελτιωθεί ακόμα περισσότερο η ακρίβεια και η δυνατότητα διάγνωσης και εντοπισμού ασθένειας.

Για να ενισχυθούν και να πιστοποιηθούν τα αποτελέσματα είναι απαραίτητο να δοκιμαστούν και σε πραγματικά δεδομένα μεγάλης κλίμακας είτε από GPS είτε από αισθητήρες επί των οχημάτων, ίσως και με τροποποιημένα μοντέλα. Τα μοντέλα αυτά θα μπορούσαν να είναι και βασιζόμενα σε μη επιβλεπόμενη τεχνική μάθησης καθώς σε πραγματικά δεδομένα ίσως να μην υπάρχει η πληροφορία για το εάν ο οδηγός είναι ασθενής ή όχι.

Το σύστημα το οποίο αναπτύχθηκε μπορεί να αποκτήσει και πρακτικές εφαρμογές. Για παράδειγμα, θα μπορούσε να χρησιμοποιηθεί για την αξιολόγηση της ικανότητας των οδηγών να συνεχίσουν να έχουν στην κατοχή τους δίπλωμα. Μια άλλη πιθανή εφαρμογή είναι η παροχή μιας

ένδειξης, αν όχι έγκυρης διάγνωσης, έπειτα από ‘εξέταση΄ οδηγών σε προσομοιωτή οδήγησης.

7

Βιβλιογραφία

[1]	I. Witten, E. Frank, M. A. Hall, C. J. Pal. Data Mining: Practical Machine Learning Tools and Techniques. *The Morgan Kaufmann Series in Data Management Systems.* Morgan Kaufmann, 2016
[2]	Christopher Clifton. Data Mining. *Encyclopaedia Brittanica*, 2009. URL: https://www.britannica.com/technology/data-mining
[3]	A. Jacobs. The Pathologies of Big Data. *Communications of the ACM*. 52(8): 36-44 (2009)
[4]	N. J. Nilsson. Introduction to Machine Learning. *Manuscript*, 2015. URL: https://ai.stanford.edu/~nilsson/mlbook.html
[5]	C. M. Bishop. Neural Networks for Pattern Recognition. *Advanced Texts in Econometrics*, Oxford University Press, 1996
[6]	S. S. Haykin. Neural Networks and Learning Machines. Prentice Hall, 2009
[7]	I. Goodfellow, Y. Bengio, A. Courville. Deep Learning. *Adaptive computation and machine learning*, MIT Press, 2016.

[8]	T. G. Dietterich. Machine Learning for Sequential Data: A Review. In Proceedings of the *Joint International Workshops on Structural, Syntactic, and Statistical Pattern Recognition (SSPR/SPR)*, pages 15-30. Springer, 2002.
[9]	I. Kononenko. Machine Learning for Medical Diagnosis: History, State of the Art and Perspective. *Artificial Intelligence in Medicine* 23(1): 89-109 (2001)
[10]	G. D. Magoulas, A. Prentza. Machine Learning in Medical Applications. In: G. Paliouras, V. Karkaletsis, C. D. Spyropoulos, editors, Machine Learning and Its Applications, *Advanced Lectures*, volume 2049 of LNCS, pages 300-307. Springer, 2001
[11]	E. B. Reategui, J. A. Campbell, B. F. Leao. Combining a neural network with casebased reasoning in a diagnostic system. *Artificial Intelligence in Medicine*, 9(1): 5-27 (1997)
[12]	J. Ridderikhoff, B. van Herk, Who is afraid of the system? Doctors' attitude towards diagnostic systems. *International Journal of Medical Informatics* 53(1): 91-100 (1999)
[13]	S. Dreiseitl, L. Ohno-Machado, H. Kittler, S. Vinterbo, H. Billhardt, M. Binder. A Comparison of Machine Learning Methods for the Diagnosis of Pigmented Skin Lesions. *Journal of Biomedical Informatics* 34(1): 28-36 (2001)
[14]	Yudong Zhang, Zhengchao Dong, Lenan Wu, Shuihua Wang: A hybrid method for MRI brain image classification. *Expert Syst. Appl.* 38(8): 10049-10053 (2011)
[15]	G. Palacios, C. Roberto. Optimal Data Distributions in Machine Learning. Dissertation (Ph.D.), California

	Institute of Technology. Doi:10.7907/Z9DR2SD5. http://resolver.caltech.edu/CaltechTHESIS:05262015-094933189
[16]	L. G. Biesecker. Hypothesis-generating research and predictive medicine. *Genome Res.* 23(7): 1051–1053 (2013)
[17]	Dimosthenis Pavlou, Panagiotis Papantoniou, Eleonora Papadimitriou, Sophia Vardaki, George Yannis, Constantinos Antoniou, John Golias, Sokratis Papageorgiou. Which are the effects of driver distraction and brain pathologies on reaction time and accident risk?. *Advances in Transportation Studies*. Special Issue, 2016
[18]	K. Leung, E. Schmerling, M. Pavone. Distributional Prediction of Human Driving Behaviours using Mixture Density Networks, 2016
[19]	W. Dong, J. Li, R. Yao, C. Li, T. Yuan & L. Wang. Characterizing driving styles with deep learning. arXiv preprint arXiv:1607.03611, 2016
[20]	D. J. Berndt, J. Clifford. Using dynamic time warping to find patterns in time series. In *KDD workshop* 10(16): 359-370 (1994)