ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ

Σχολή Χρηματοοικονομικής και Στατιστικής

Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης

ΜΕΤΑΠΤΥΧΙΑΚΟ ΠΡΟΓΡΑΜΜΑ ΣΠΟΥΔΩΝ
ΣΤΗΝ **ΕΦΑΡΜΟΣΜΕΝΗ ΣΤΑΤΙΣΤΙΚΗ**

ΕΦΑΡΜΟΓΕΣ ΣΤΑΤΙΣΤΙΚΗΣ ΜΗΧΑΝΙΚΗΣ ΜΑΘΗΣΗΣ ΣΤΟ ΜΗΧΑΝΟΚΙΝΗΤΟ ΑΘΛΗΤΙΣΜΟ

Χρύσα Ιωάννα Μ. Αιγινίτη

Διπλωματική Εργασία

που υποβλήθηκε στο Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης του Πανεπιστημίου Πειραιώς ως μέρος των απαιτήσεων για την απόκτηση του Μεταπτυχιακού Διπλώματος Ειδίκευσης στην *Εφαρμοσμένη Στατιστική*

Πειραιάς
Σεπτέμβριος 2024

Π Α Ν Ε Π Ι Σ Τ Η Μ Ι Ο Π Ε Ι Ρ Α Ι Ω Σ

Σχολή Χρηματοοικονομικής και Στατιστικής

Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης

ΜΕΤΑΠΤΥΧΙΑΚΟ ΠΡΟΓΡΑΜΜΑ ΣΠΟΥΔΩΝ
ΣΤΗΝ **ΕΦΑΡΜΟΣΜΕΝΗ ΣΤΑΤΙΣΤΙΚΗ**

ΕΦΑΡΜΟΓΕΣ ΣΤΑΤΙΣΤΙΚΗΣ ΜΗΧΑΝΙΚΗΣ ΜΑΘΗΣΗΣ ΣΤΟ ΜΗΧΑΝΟΚΙΝΗΤΟ ΑΘΛΗΤΙΣΜΟ

Χρύσα Ιωάννα Μ. Αιγινίτη

Διπλωματική Εργασία

που υποβλήθηκε στο Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης του Πανεπιστημίου Πειραιώς ως μέρος των απαιτήσεων για την απόκτηση του Μεταπτυχιακού Διπλώματος Ειδίκευσης στην *Εφαρμοσμένη Στατιστική*

Πειραιάς
Σεπτέμβριος 2024

Η παρούσα Διπλωματική Εργασία εγκρίθηκε ομόφωνα από την Τριμελή Εξεταστική Επιτροπή που ορίσθηκε από τη ΓΣΕΣ του Τμήματος Στατιστικής και Ασφαλιστικής Επιστήμης του Πανεπιστημίου Πειραιώς στην υπ' αριθμ. ……. συνεδρίασή του σύμφωνα με τον Εσωτερικό Κανονισμό Λειτουργίας του Προγράμματος Μεταπτυχιακών Σπουδών στην Εφαρμοσμένη Στατιστική

Τα μέλη της Επιτροπής ήταν:

- Σ. Μπερσίμης, Καθηγητής (Επιβλέπων)
- Κ. Πολίτης, Αναπληρωτής Καθηγητής
- Σ. Τασουλής, Επίκουρος Καθηγητής

Η έγκριση της Διπλωματικής Εργασίας από το Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης του Πανεπιστημίου Πειραιώς δεν υποδηλώνει αποδοχή των γνωμών του συγγραφέα.

UNIVERSITY OF PIRAEUS

School of Finance and Statistics

Department of Statistics and Insurance Science

POSTGRADUATE PROGRAM IN
APPLIED STATISTICS

APPLICATIONS OF MACHINE LEARNING METHODS IN MOTORSPORT

By

Chrysa Ioanna M. Aiginiti

MSc Dissertation

submitted to the Department of Statistics and Insurance Science of the University of Piraeus in partial fulfilment of the requirements for the degree of Master of Science in Applied Statistics

Piraeus, Greece
September 2024

Στους γονείς μου
Μιχάλη και Σοφία

Ευχαριστίες

Με την ολοκλήρωση της παρούσας διπλωματικής εργασίας ολοκληρώνεται και ένα εξαιρετικά σημαντικό ταξίδι της ζωής μου που ξεκίνησε εφτά χρόνια πριν με την εισαγωγή μου στο τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης του Πανεπιστημίου Πειραιώς. Αν και έκτη στη σειρά του μηχανογραφικού, η σχολή αυτή αποτελεί σήμερα την πιο σωστή επιλογή μου. Είχα την τύχη να γνωρίσω εξαιρετικούς καθηγητές πρόθυμους όχι μόνο να μου μεταβιβάσουν τις γνώσεις τους και την αγάπη τους για την επιστήμη μας αλλά και να μου δώσουν πολύτιμες συμβουλές για τη μετέπειτα επαγγελματική και προσωπική μου πορεία γι' αυτό και τους ευχαριστώ όλους.

Θα ήθελα να ευχαριστήσω τον επιβλέποντα καθηγητή μου κ. Σωτήριο Μπερσίμη, ο οποίος δέχτηκε χωρίς δεύτερη σκέψη το θέμα που του πρότεινα και μου έδωσε την ευκαιρία να ασχοληθώ με ένα χώρο που παρακολουθώ από τότε που θυμάμαι τον εαυτό μου.

Ευχαριστώ, επίσης, τους γονείς μου για την οικονομική και ψυχολογική τους στήριξη όλα αυτά τα χρόνια. Θα ήθελα, να πω ξεχωριστά ένα τεράστιο ευχαριστώ στον πατέρα μου, Μιχάλη, που με έβαλε στο μαγικό κόσμο του μηχανοκίνητου αθλητισμού και κυρίως της Formula 1 και που, μάλλον, άθελά του μού έβαλε το μικρόβιο και την ιδέα για την παρούσα διπλωματική μετά από μία κουβέντα μας όταν ήμουν στο τρίτο έτος του προπτυχιακού. Ίσως πάλι και ηθελημένα… Τον ευχαριστώ, λοιπόν, για τις ατελείωτες ώρες συζητήσεων σχετικά με την F1 και όχι μόνο.

Τέλος, ευχαριστώ τις φίλες μου Κασσιανή, Γεωργία και Γεωργία, που με συντροφεύουν από το πρώτο εξάμηνο του προπτυχιακού μέχρι σήμερα, για τη στήριξη και την υπομονή τους στις κατά τ' άλλα γλυκές γκρινίτσες μου.

Περίληψη

Η Formula 1 από την πρώτη κιόλας στιγμή είναι άρρηκτα συνδεδεμένη με την ταχύτητα για πολλούς λόγους. Για την ταχύτητα με την οποία τρέχουν τα μονοθέσια με τους οδηγούς να δοκιμάζουν κάθε φορά τόσο τα όρια του μονοθέσιου όσο και του ίδιου τους του εαυτού. Για την ταχύτητα με την οποία εμφανίζονται και εφαρμόζονται καινοτόμες ιδέες που βελτιώνουν την απόδοση του μονοθέσιου και παράλληλα το καθιστούν εξαιρετικά ασφαλές. Τέτοιες καινοτομίες χρησιμοποιούνται πλέον στα μη αγωνιστικά αυτοκίνητα κάνοντας την καθημερινή μετακίνηση εύκολη και συνάμα ασφαλή διαδικασία. Οι ραγδαίες εξελίξεις έχουν γεννήσει πληθώρα δεδομένων που καλούνται οι ομάδες να επεξεργαστούν και να ερμηνεύσουν ούτως ώστε να εξορύξουν την κρυφή πληροφορία και να βελτιώσουν τόσο το μονοθέσιο όσο και τη στρατηγική που θα ακολουθήσουν στοχεύοντας στη νίκη. Για την πλήρη και αποτελεσματική αξιοποίηση των δεδομένων είναι αναγκαία η χρήση μαθηματικών και στατιστικών μεθόδων και πληροφορικής. Γι’ αυτό πλέον η χρήση της στατιστικής μηχανικής μάθησης, που συνδυάζει επιτυχώς αυτές τις τρεις επιστήμες είναι αναπόφευκτη και μείζονος σημασίας για τη λήψη στρατηγικών αποφάσεων σε κλάσματα δευτερολέπτου κατά τη διάρκεια του αγώνα, την πρόβλεψη αποτελεσμάτων και τη εξαγωγή συμπερασμάτων. Στην παρούσα διπλωματική εργασία γίνεται εκτενής ανάλυση των διαφόρων κατηγοριών που χαρακτηρίζουν τον κλάδο της μηχανικής μάθησης, των αλγορίθμων και των μοντέλων που χρησιμοποιούνται και των βασικών μεθόδων προπαρασκευής των δεδομένων πριν την ανάλυση. Τέλος, γίνεται εφαρμογή μεθόδων στατιστικής μηχανικής μάθησης σε δεδομένα από αγώνες της F1 με στόχο την πρόβλεψη της θέσης τερματισμού των οδηγών στον αγώνα.

Abstract

Formula 1 is a domain inextricably linked with speed from the very beginning for a variety of reasons. Speed is embodied in the way the cars race, with drivers constantly testing the limits of both the vehicle and themselves. It is also seen in the rapid emergence and application of innovative ideas that enhance the performance of the car while simultaneously making it exceptionally safe. These innovations are now used in non-racing cars, making everyday transportation both easy and safe. The rapid developments have generated a wealth of data that teams are tasked with processing and interpreting in order to extract hidden information and improve both the car and the strategy they will follow, aiming for victory. To fully and effectively utilize the data, the use of mathematical and statistical methods, as well as computer science, is essential. As a result, the use of statistical machine learning, which successfully combines these three scientific fields, has become inevitable and crucially important for making strategic decisions in fractions of a second during the race, predicting outcomes, and drawing conclusions. This thesis presents an extensive analysis of the various categories that characterize the field of machine learning, the algorithms and models used, and the key methods of data preprocessing before analysis. Finally, statistical machine learning methods are applied to data from F1 races with the aim of predicting the drivers' finishing positions in the race.

Περιεχόμενα

ΚΕΦΑΛΑΙΟ 1ο

Εισαγωγή

1.1 Formula 1 και όγκος δεδομένων

Ο μηχανοκίνητος αθλητισμός και κατ' επέκταση η Formula 1 αποτελεί χώρο συνύπαρξης και εξέλιξης πολλών επιστημονικών κλάδων επομένως απαιτεί την άριστη συνεργασία ανθρώπων από διαφορετικά επαγγέλματα. Όπως για την κατασκευή ενός μονοθέσιου ή ενός κινητήρα είναι απαραίτητη η συνεργασία μηχανολόγων μηχανικών, ηλεκτρολόγων μηχανικών, αεροδυναμιστών και προγραμματιστών, έτσι και για τη διαχείριση του τεράστιου όγκου των δεδομένων που παράγονται τόσο πίσω στο εργοστάσιο όσο και κατά τη διάρκεια του αγώνα είναι αναγκαία η συνδρομή των στατιστικών και των αναλυτών δεδομένων. Τα δεδομένα που παράγονται στο εργοστάσιο αφορούν κυρίως τα στάδια κατασκευής του κινητήρα, του μονοθέσιου και του τρόπου με τον οποίο θα στηθεί κάθε φορά για τον αγώνα. Τέτοιου είδους δεδομένα είναι απόρρητα και τα διαχειρίζονται μόνο οι αναλυτές της ομάδας με στόχο την εξόρυξη πληροφοριών χρήσιμων για την επίτευξη της μέγιστης απόδοσης που μπορεί να δώσει το μονοθέσιο στη διάρκεια του αγώνα. Πληθώρα δεδομένων παράγεται και σε κάθε τριήμερο των αγώνων (ελεύθερες δοκιμές, κατατακτήριες δοκιμές, αγώνας). Τα δεδομένα αυτά μπορούν να χρησιμοποιηθούν για την επιλογή της στρατηγικής που θα αποφέρει τα μέγιστα αποτελέσματα, ειδικά αν τα δεδομένα χρησιμοποιούνται κατά τη διάρκεια κάθε αγώνα σε πραγματικό χρόνο γεγονός υψίστης σημασίας μιας και την ώρα του αγώνα κάθε κλάσμα δευτερολέπτου είναι πολύτιμο και οι αποφάσεις πρέπει να λαμβάνονται ταχύτατα.

1.2 Επεξεργασία και αξιοποίηση των παραγόμενων δεδομένων

Για τη συλλογή, επεξεργασία, ανάλυση των δεδομένων και παρουσίαση της ανακαλυφθείσας γνώσης με μορφή τέτοια ώστε να την χρησιμοποιούν αποτελεσματικά άτομα διαφορετικών επαγγελμάτων απαιτείται ο συνδυασμός γνώσεων μαθηματικών, στατιστικής και πληροφορικής. Τους τρεις παραπάνω τομείς συνδυάζει η στατιστική μηχανική μάθηση, ένας σύγχρονος επιστημονικός κλάδος με τη βοήθεια του οποίου μπορούμε να κάνουμε προβλέψεις χρησιμοποιώντας είτε ιστορικά δεδομένα είτε δεδομένα πραγματικού χρόνου σε συνδυασμό κάθε φορά με το κατάλληλο μοντέλο. Με χρήση της μηχανικής μάθησης έχουν κατά καιρούς αποφευχθεί πολλά πιθανά ατυχήματα τα οποία προβλέφθηκαν εγκαίρως. Ακόμη, οι αγώνες της F1 έχουν μετατραπεί σε σκληρό παιχνίδι στρατηγικής καθώς μία λάθος κίνηση, μία όχι τόσο ακριβής πρόβλεψη μπορεί να στερήσει την πρώτη θέση ακόμη και από το πιο ισχυρό κατασκευαστικά μονοθέσιο ή τον πιο έμπειρο και ικανό οδηγό.

ΚΕΦΑΛΑΙΟ 2ο

Formula 1

2.1 Εισαγωγή στη Formula 1

2.1.1 Σύντομη ιστορία της Formula One

Η Formula 1 είναι ένα διεθνές πρωτάθλημα αγώνων αυτοκινήτου το οποίο βρίσκεται στο υψηλότερο επίπεδο στην κατηγορία των μονοθέσιων με ανοιχτή ρόδα (open-wheel)[1] και ανοιχτό πιλοτήριο (cockpit) στο χώρο του μηχανοκίνητου αθλητισμού. Η λέξη formula (Ελληνικά φόρμουλα) χρησιμοποιείται για να αποδώσει το σύνολο των κανόνων στους οποίους συμμορφώνονται οδηγοί και κατασκευαστές ανεξαιρέτως. Αρχικά ήταν γνωστή ως Formula A. Εμφανίστηκε για πρώτη φορά στην Ευρώπη και στα τέλη της δεκαετίας του 1930 άρχισαν οι συζητήσεις για διεξαγωγή πρωταθλήματος οι οποίες πάγωσαν λόγω του 2ου Παγκοσμίου Πολέμου.

Το 1946, όμως, αναβιώνει η ιδέα για πρωτάθλημα όπου και αποφασίζεται το 1947. Μετά το πέρας των απαραίτητων διαδικασιών, το πρώτο πρωτάθλημα της Formula 1 πραγματοποιείται το 1950 με τον πρώτο αγώνα να λαμβάνει χώρα στο Pau της Γαλλίας. Ο πρώτος πρωταθλητής στην ιστορία της Formula 1 είναι ο Giuseppe "Nino" Farina σε μονοθέσιο της Alfa Romeo.

Κυρίαρχες ομάδες της δεκαετίας του 1950 ήταν οι Ferrari, Alfa Romeo, Maserati, Mercedes-Benz και κυρίαρχη φιγούρα ο Juan Manuel Fangio, κάτοχος 6 πρωταθλημάτων σε μονοθέσια 5 διαφορετικών εταιρειών. Η έλλειψη συμμετεχόντων εξαιτίας του υψηλού κόστους οδήγησε στην υιοθέτηση των κανονισμών της Formula 2 (χαμηλότερη κατηγορία) τις χρονιές 1952 και 1953. Από τις πρώτες 20 εταιρείες που συμμετείχαν στο πρωτάθλημα όλες σχεδόν αναγκάστηκαν σύντομα να αποχωρήσουν εξαιτίας του υψηλού κόστους με τη Ferrari να είναι η μόνη που συμμετέχει από την αρχή της διοργάνωσης μέχρι σήμερα.

Η δεκαετία 1958-1968 χαρακτηρίζεται από πολλούς και British era αφού 6 από τα 10 πρωταθλήματα πηγαίνουν σε βρετανούς οδηγούς και στο grid βρίσκονται την ίδια περίοδο 7 Βρετανοί οδηγοί στους συνολικά 20.

Σταθμό στην ιστορία του αθλήματος όσον αφορά την ασφάλεια των οδηγών αποτελεί αρχικά το ατύχημα του Niki Lauda στο γερμανικό Grand Prix στο Νίρμπουργκρινγκ το 1976. Πιο συγκεκριμένα, η Ferrari που οδηγούσε χτύπησε στα προστατευτικά κιγκλιδώματα και τυλίχτηκε στις φλόγες με αποτέλεσμα να κινδυνέψει θανάσιμα η ζωή του εξαιτίας της εισπνοής τοξικών αερίων και των εξαιρετικά σοβαρών εγκαυμάτων στο κεφάλι. Κόντρα σε όλα τα προγνωστικά επέστρεψε μέσα σε έξι εβδομάδες στους αγώνες και υπήρξε ένας από τους πιο ενεργούς οδηγούς σχετικά με την αύξηση της ασφάλειας των

[1] Αυτοκίνητο ανοιχτής ρόδας (open-wheel car) χαρακτηρίζεται το αυτοκίνητο του οποίου οι ρόδες βρίσκονται εκτός του κύριου σώματος του αυτοκινήτου. Διαφέρουν από τα αυτοκίνητα δρόμου, τα σπορ αυτοκίνητα, τα αυτοκίνητα τύπου stock (NASCAR) και τα αγωνιστικά αυτοκίνητα τύπου touring, τα οποία έχουν τις ρόδες τους κάτω από το κυρίως σώμα του αυτοκινήτου ή στο εσωτερικό των φτερών. Συνήθως δεν έχουν νόμιμη άδεια για οδήγηση σε δρόμο εκτός πίστας αφού δεν είναι καθόλου πρακτικά για καθημερινή χρήση.

οδηγών μέσα στα μονοθέσια. Δεύτερο και εξίσου σημαντικό γεγονός αποτελεί το θανατηφόρο ατύχημα του Ayrton Senna, ενός από τους ικανότερους οδηγούς που υπήρξαν ποτέ, στο Grand Prix του Σαν Μαρίνο το 1994. Την προηγούμενη μέρα είχε σκοτωθεί και ο αυστριακός οδηγός Roland Ratzenberger

2.1.2 FIA – κανονισμοί

Το πρωτάθλημα της F1 διέπεται και επικυρώνεται από τη Διεθνή Ομοσπονδία Αυτοκινήτου γνωστή, κυρίως, με τα αρχικά FIA (Fédération Internationale de l'Automobile). Η FIA ιδρύθηκε το 1904 στη Γαλλία με την έδρα της να βρίσκεται στο Παρίσι. Είναι ένας μη κερδοσκοπικός οργανισμός που έχει υπό την αιγίδα της 244 διεθνείς μηχανοκίνητους και αγωνιστικούς οργανισμούς από 146 χώρες σε όλο τον κόσμο. Βασικός σκοπός της είναι η διακυβέρνηση και η διασφάλιση του μηχανοκίνητου αθλητισμού.

Μέρος των καθηκόντων της αποτελεί η θέσπιση των κανονισμών για κάθε πρωτάθλημα. Πιο συγκεκριμένα, η FIA καθορίζει όλους τους κανονισμούς στο πρωτάθλημα της Formula 1, από το πώς θα κατασκευαστεί και θα στηθεί ένα μονοθέσιο μέχρι τον αριθμό των αγώνων που θα πραγματοποιηθούν εντός μιας σεζόν, τον αριθμό των ομάδων – κατασκευαστών και των μονοθέσιων που λαμβάνουν μέρος καθώς και τον προϋπολογισμό των ομάδων για όλη τη σεζόν.

Αναλυτικότερα, σύμφωνα με τους κανονισμούς της FIA κατά τη διάρκεια μιας σεζόν τρέχουν δύο πρωταθλήματα, ένα για τους οδηγούς (drivers' championship) και ένα για τους κατασκευαστές (constructors' championship). Κάθε ομάδα έχει δικαίωμα να χρησιμοποιήσει μέχρι τέσσερις διαφορετικούς οδηγούς στους αγώνες. Κατά κύριο λόγο οι ομάδες διαθέτουν δύο βασικούς οδηγούς και έναν τρίτο αναπληρωματικό οδηγό σε περίπτωση που ένας εκ των δύο ασθενήσει ή τραυματιστεί. Επιπλέον, αλλαγές οδηγών για λόγους υψηλότερης απόδοσης λαμβάνονται διαφορετικά υπόψιν.

Το σύνολο των μονοθέσιων που τρέχουν σε κάθε GP είναι 20, 2 για κάθε ομάδα, καθώς συμμετέχουν 10 ομάδες σε κάθε σεζόν. Στο παρελθόν έχουν διεξαχθεί GP με περισσότερα των 20 αυτοκινήτων. Χαρακτηριστικό παράδειγμα αποτελεί το γερμανικό GP του Νίρμπουργκρινγκ, το 1953, στο οποίο συμμετείχαν 35 αυτοκίνητα. Πλέον, σύμφωνα με τους κανονισμούς της FIA το όριο για τον αριθμό συμμετοχών των μονοθέσιων είναι 26.

Τέλος, καίριο ρόλο για τις ομάδες της F1 αποτελούν οι ακαδημίες νέων οδηγών (driver academies). Αυτή τη στιγμή μόνο οι οκτώ από τις δέκα ομάδες του grid διαθέτουν driver academies, αν και έχει ανακοινωθεί πως από το 2024 θα διαθέτουν και οι δέκα ομάδες. Ο ρόλος των ακαδημιών είναι ιδιαίτερα σημαντικός καθώς μέσω αυτών αναγνωρίζονται το ταλέντο και οι ικανότητες νέων οδηγών οι οποίοι τρέχουν σε μικρότερες κατηγορίες και στην πορεία θα κερδίσουν μία θέση στο πρωτάθλημα της F1. Χαρακτηριστικό παράδειγμα αποτελεί το γεγονός ότι πολλές από τις ομάδες χρηματοδοτούν τις ακαδημίες τους και τους νέους οδηγούς αποσκοπώντας σε μία μετέπειτα προσοδοφόρα πορεία τόσο για τον εκάστοτε οδηγό όσο και για την ομάδα.

2.1.3 Circuits

Η FIA, επιπροσθέτως, καθορίζει σε ποιες χώρες και πότε θα διεξαχθούν τα GPs της σεζόν δημιουργώντας έτσι ένα ημερολόγιο αγώνων γνωστό ως calendar («καλεντάρι»). Για να συμμετέχει μία χώρα στο καλεντάρι πρέπει η πίστα της να πληροί συγκεκριμένες προδιαγραφές προκειμένου να δοθεί άδεια διεξαγωγής αγώνα. Υπάρχουν διαφορετικά

επίπεδα άδειας ανάλογα με το είδος του αγώνα που καλούνται οι πίστες να φιλοξενήσουν. Τα επίπεδα αυτά είναι 6 (Grade 1-6) και για να διεξαχθεί αγώνας Formula 1 στο εκάστοτε circuit πρέπει να έχει ληφθεί από τη FIA άδεια τύπου 1 (Grade 1) που είναι το υψηλότερο επίπεδο. Αυτό προϋποθέτει να πληρούνται προδιαγραφές που αφορούν αρχικά τις διαστάσεις της πίστας, μερικές από τις οποίες είναι α) οι ευθείες πρέπει να έχουν μήκος μικρότερο των 2 χλμ, β) οι πίστες πρέπει να έχουν συνολικό μήκος τουλάχιστον 3,5 χλμ – εξαίρεση αποτελεί το Μονακό με 3,337 χλμ μήκος-, γ) η σχάρα εκκίνησης (starting grid) πρέπει να έχει τουλάχιστον 15 μέτρα πλάτος το οποίο θα διατηρηθεί μέχρι την πρώτη στροφή για την αποφυγή του pile-up (στοίβαγμα) των μονοθέσιων δ) οι θέσεις εκκίνησης (grid positions) πρέπει να έχουν 8 μ. απόσταση μεταξύ τους.

Στη συνέχεια ακολουθούν τεχνικές προδιαγραφές που αφορούν τις μπαριέρες (barriers) του circuit. Εκτενέστερα, οι προδιαγραφές σχετίζονται με την διάταξη της πίστας, την τοπογραφία, τις αγωνιστικές γραμμές, τις ζώνες ταχύτητας που διαθέτει καθώς και τη δομή του περιβάλλοντος εκτός πίστας όπως οι κερκίδες και οι εγκαταστάσεις γύρω από αυτήν. Κατά κανόνα, τοποθετούνται είτε μπαριέρες που απορροφούν τη δύναμη του μονοθέσιου το οποίο θα προσκρούσει, πιθανόν, με πολλά χλμ. πάνω τους, είτε περιοχές εκτόνωσης (run-off areas) που επιτρέπουν στους οδηγούς να ανακτήσουν τον έλεγχο του μονοθέσιού τους και να επιστρέψουν στον αγώνα ή απλά να μειώσουν ταχύτητα, ακόμα και να σταματήσουν ένα πιθανό τρακάρισμα.

Τέλος, ακολουθούν προδιαγραφές σχετικά με την αποστράγγιση των circuits σε περίπτωση βροχής καθώς και με το ιατρικό κέντρο το οποίο επιβάλλεται να είναι επανδρωμένο με τουλάχιστον δύο γιατρούς με άριστη γνώση καρδιοαναπνευστικής ανάνηψης και τουλάχιστον δύο χειρουργούς, ένας ειδικό σε εγκαύματα και έναν ικανό να διαχειριστεί τραυματισμούς σπονδυλικής στήλης.

2.2 Οικονομικά στοιχεία

Ο Flavio Briatore, είπε σύμφωνα με τον Nauright (2012), ότι «είναι ξεκάθαρο πως οι αγώνες Grand Prix δεν είναι πλέον σπορ αλλά επιχείρηση. Μιλάμε για ομάδες στις οποίες απασχολούνται πάνω από εκατό εργαζόμενοι και που διαθέτουν τεράστιους προϋπολογισμούς», (Mourao, 2017). Πράγματι, η F1 αποτελεί ένα άθλημα για το οποίο δαπανούνται υπέρογκα χρηματικά ποσά της τάξης των εκατομμυρίων ευρώ τόσο από όλες τις ομάδες που συμμετέχουν όσο και από τα κράτη και τους χορηγούς που τις χρηματοδοτούν. Αυτό δεν ισχύει όμως για τα κέρδη τα οποία διαφέρουν αρκετά και είναι ανάλογα του αποτελέσματος που φέρνει κάθε ομάδα.

2.2.1 Σταθερά & Μεταβλητά Κόστη

Ας αναλύσουμε, εν συντομία, τα διάφορα κόστη που αφορούν την F1. Πιο συγκεκριμένα, τα κόστη μιας επιχείρησης χωρίζονται σε σταθερά και μεταβλητά. Σταθερά κόστη (ή εν μέρει σταθερά) είναι αυτά που δεν αλλάζουν για μια ομάδα είτε αυτή κερδίσει το πρωτάθλημα είτε δεν κερδίσει ούτε ένα πόντο. Θεωρούνται και εν μέρει σταθερά διότι παρά το γεγονός ότι δεν μεταβάλλονται, στην περίπτωση που μια ομάδα θέλει να έχει καλύτερα αποτελέσματα θα πρέπει να αυξήσει ακόμη και τα σταθερά κόστη της. Στα σταθερά κόστη περιλαμβάνονται οι μισθοί των οδηγών και όλων των εργαζομένων μιας ομάδας και θεμελιώδη κόστη που αφορούν την κατασκευή και ανακατασκευή του μονοθέσιου.

Όσον αφορά τους μισθούς των οδηγών, αυτοί καθορίζονται βάσει συμβολαίου πριν ξεκινήσει η νέα αγωνιστική σεζόν και είναι ανεξάρτητοι του αποτελέσματος που θα φέρει ο εκάστοτε οδηγός στην ομάδα. Ο μισθός κάθε οδηγού εξαρτάται από πολλούς παράγοντες (Mourao, 2017). Αρχικά, σε μεγάλο ποσοστό αξιολογούνται οι επιδόσεις των οδηγών την προηγούμενη χρονιά όπως αν κατάφερε ή όχι να κερδίσει το πρωτάθλημα, τους πόντους που κέρδισε και την τελική του κατάταξη στο πρωτάθλημα, τα λάθη αλλά και το πόσο σωστά εκμεταλλεύτηκε τις δυνατότητες του μονοθέσιού του. Δεδομένου ότι τα μονοθέσια δεν είναι ισάξια συνυπολογίζονται στα παραπάνω και οι δυνατότητες του κάθε μονοθέσιου ώστε να αξιολογηθεί η ποιότητα ενός οδηγού. Επιπλέον, στην ανταμοιβή των οδηγών προστίθεται και η επιθυμία μιας ομάδας να κερδίσει το πρωτάθλημα καθώς οι υψηλές απολαβές αποτελούν ένα ακόμη κίνητρο για τελεσφόρα προσπάθεια του οδηγού. Τέλος, σημαντικό ρόλο παίζει και αν ο οδηγός που επιλέγει η ομάδα είναι ήδη κάτοχος πρωταθλήματος.

Όπως είναι αντιληπτό μία ομάδα στην F1 δεν αποτελείται μόνο από οδηγούς. Για να πετύχει τους στόχους της είναι απαραίτητα και άλλα επαγγέλματα εξίσου σημαντικά με τους οδηγούς όπως μηχανικοί, αναλυτές και τεχνικοί. Σύμφωνα με τον Paulo Mourao, The Economics of Sports, The Case of F1, (2017), αν θεωρήσουμε ως μισθολογική βάση το μισθό μιας διοικητικής γραμματέως, όπου ο μέσος ετήσιος μισθός της είναι περίπου 24,000 € τότε ένας αρχάριος μηχανικός θα αμείβεται με 2.08 φορές τη βάση και ο επικεφαλής μηχανικός με 6.25 φορές τη βάση. Όσον αφορά τον αγωνιστικό διευθυντή ο μισθός τείνει να είναι τουλάχιστον 8 φορές επί τη βάση. Συνεπώς, είναι φανερό πως οι ομάδες καταβάλουν ετησίως εξαιρετικά μεγάλα χρηματικά ποσά προκειμένου να καλυφθούν οι μισθοί των εργαζομένων.

Τέλος, δεν μπορούν να παραλειφθούν τα σταθερά κόστη που αφορούν στην κατασκευή του αυτοκινήτου. Σε αυτά περιλαμβάνονται η κατασκευή αλλά και ανακατασκευή του αυτοκινήτου μετά από κάθε αγώνα. Αξιοσημείωτο είναι το ποσό που δαπανάται για τον κινητήρα το οποίο κυμάνθηκε μεταξύ 10 και 25 εκατομμυρίων ευρώ, το 2015, για όλη τη σεζόν. Αξίζει να σημειωθεί πως δεν δύνανται όλες οι ομάδες να κατασκευάσουν και να εξελίξουν οι ίδιες τον κινητήρα και για αυτό προμηθεύονται κινητήρα από άλλες ομάδες. Αυτή τη στιγμή οι ομάδες που αγωνίζονται με κινητήρα δικής τους κατασκευής είναι οι Ferrari και Mercedes οι οποίες παρέχουν κινητήρα και σε κάποιες από τις υπόλοιπες ομάδες.

Μετά τα σταθερά ακολουθούν τα μεταβλητά κόστη τα οποία κατά κύριο λόγο αφορούν ενέργειες με στόχο την ευνοϊκότερη έκβαση κάθε αγώνα. Το βασικότερο μεταβλητό κόστος αφορά έξοδα σχετικά με την έρευνα και την ανάπτυξη του αυτοκινήτου και του κινητήρα. Χαρακτηριστικό παράδειγμα αποτελούν ομάδες όπως η Ferrari και η Mercedes των οποίων οι δαπάνες για τη νίκη στο πρωτάθλημα είναι υψηλότερες από οποιασδήποτε άλλης ομάδας στη διοργάνωση. Σύμφωνα με τη Forbes το 2019 δύο από τις συνολικά τέσσερις εταιρείες παροχής κινητήρα εξέτασαν σοβαρά την απόφαση αποχώρησής τους από την F1 εξαιτίας των αυξημένων εξόδων. Απόρροια αυτού είναι το γεγονός ότι οι δύο παραπάνω ομάδες είναι από τις πλέον πιο ανταγωνιστικές στη διοργάνωση σχεδόν κάθε χρόνο καθώς όπως φαίνεται όσο περισσότερα είναι τα χρήματα που επενδύονται στην εξέλιξη του κινητήρα και του αυτοκινήτου τόσο περισσότεροι είναι και οι κερδισμένοι πόντοι σε κάθε αγώνα. Μεταβλητά κόστη θεωρούνται, επίσης, μεταξύ άλλων η επιπρόσθετη εκπαίδευση των οδηγών και η ιατροφαρμακευτική τους περίθαλψη, η στέγαση του προσωπικού σε κάθε αγώνα και η εξέλιξη του αναγκαίου συμπληρωματικού λογισμικού κατά τη διάρκεια της σεζόν.

Επομένως, αθροίζοντας τα σταθερά και μεταβλητά κόστη παίρνουμε τα συνολικά κόστη τα οποία επιδρούν σημαντικά στον καθορισμό του προϋπολογισμού κάθε ομάδας. Ομάδες, λοιπόν, που πετυχαίνουν τη νίκη και τις υψηλότερες θέσεις στο πρωτάθλημα τείνουν να έχουν και τους υψηλότερους προϋπολογισμούς χωρίς ωστόσο να ισχύει πάντα αυτό. Η McLaren, λόγου χάριν, το 2015 παρ' ότι είχε τον τρίτο υψηλότερο προϋπολογισμό (465 εκ. δολάρια) συγκέντρωσε τους λιγότερους πόντους (27) εν αντιθέσει με τη Ferrari που είχε τον τέταρτο σε σειρά υψηλότερο προϋπολογισμό (418 εκ. δολάρια) και τερμάτισε δεύτερη στο πρωτάθλημα με 428 πόντους (Mourao, 2017). Γενικότερα, όμως, κυρίαρχες ομάδες σε κάθε σεζόν είναι αυτές που δαπανούν τα μεγαλύτερα κεφάλαια. Λόγω αυτού, το 2021 η FIA θέσπισε κανονισμό σύμφωνα με τον οποίο ορίζεται συγκεκριμένο χρηματικό ποσό προϋπολογισμού στην αρχή κάθε σεζόν με στόχο να ανέβει το επίπεδο των μικρότερων ομάδων που μέχρι πρότινος δε μπορούσαν να ανταγωνιστούν τις ομάδες κολοσσούς. Για το 2021 το ποσό θα ανερχόταν στα 175 εκ. δολάρια αλλά λόγω των δυσκολιών που αντιμετώπιζαν οι ομάδες εξαιτίας της πανδημίας COVID-19, το ποσό μειώθηκε στα 145 εκ. δολάρια. Τη σεζόν 2022 μειώθηκε 5 εκ. ακόμη, στα 140 εκ. δολάρια και τέλος το 2023 έπεσε στα 135, ποσό που ορίστηκε για 21 αγώνες με δικαίωμα αύξησής του 1,8 εκ. δολάρια για κάθε πρόσθετο αγώνα.

2.2.2 Άλλες οικονομικές δραστηριότητες και κόστη

Παραπάνω αναλύθηκαν διεξοδικά τα έξοδα, σταθερά και μεταβλητά, που καλείται κάθε ομάδα να καλύψει προκειμένου να είναι ανταγωνιστική σε κάθε αγώνα εξασφαλίζοντας στο τέλος της σεζόν το μεγαλύτερο δυνατό κέρδος. Ωστόσο, υπάρχουν και άλλες οικονομικές δραστηριότητες εξίσου σημαντικές για τη διεξαγωγή ενός αγώνα και κατ' επέκταση του πρωταθλήματος συνολικά.

Αρχικά, όπως προαναφέρθηκε, για να μπορέσει μια χώρα να φιλοξενήσει ένα Grand Prix οφείλει να πληροί συγκεκριμένες αυστηρές προδιαγραφές. Οι προδιαγραφές αυτές αφορούν στην καταλληλότητα των υποδομών εντός και εκτός του αγωνιστικού χώρου, στην οικονομική ρευστότητα της περιοχής, στη δυνατότητα για προσφορά ανθρώπινου δυναμικού, στην ικανότητα στη διοργάνωση και τέλος στην αγάπη των κατοίκων για το μηχανοκίνητο αθλητισμό. Έτσι προκύπτουν τρεις βασικές ομάδες που θα χρηματοδοτήσουν τον αγώνα, οι φορολογούμενοι, οι φίλοι του σπορ, και οι χορηγοί (Mourao, 2017).

Κάθε χώρα είναι υποχρεωμένη να καταβάλει χρηματικό ποσό στη FIA το οποίο ορίζεται μέσω συμβολαίου, μονοετούς ή περισσότερων ετών, πριν τη λήξη της τρέχουσας σεζόν ώστε να ανακοινωθεί έγκαιρα το καλεντάρι της επόμενης χρονιάς. Δεδομένου ότι τα ποσά αυτά είναι της τάξης των εκατομμυρίων δολαρίων, ένας τρόπος συγκέντρωσης του ποσού είναι μέσω της φορολογίας. Αξίζει να αναφερθεί ότι το χαμηλότερο ποσό που καταβλήθηκε για το 2022 ήταν από το Μονακό στα 15 εκ. δολάρια ενώ το υψηλότερο ήταν από το Κατάρ, τη Σαουδική Αραβία και το Αζερμπαϊτζάν στα 55 εκ. δολάρια. Επιπροσθέτως, υψηλά είναι και τα ποσά που καταβάλλονται για τη δημιουργία νέων ή τη συντήρηση των ήδη υπαρχόντων circuit αλλά και της μετατροπής του χώρου και της ενοικίασης εξοπλισμού που απαιτούν τα circuit πόλης. Για να μετατραπούν οι δρόμοι μιας πόλης σε circuit, παραδείγματος χάριν στο Μονακό, χρειάζονται περίπου 16 εκ. δολάρια για στελέχωση όλων των υπηρεσιών από τα οποία τα 6.5 εκ. δολάρια αφορούν τις ομάδες προώθησης, διαφήμισης και οργάνωσης και 14 εκ. δολάρια για ενοικίαση κερκίδων 80,000 θέσεων. Επίσης, το ποσό για περίφραξη των δρόμων με ειδικά κιγκλιδώματα και ζώνες εκτόνωσης ταχύτητας για αποφυγή σφοδρών συγκρούσεων καθώς και για ενοικίαση

κτιρίων που λειτουργούν ως γκαράζ και χώροι pit stop, οχήματα ασφαλείας, γραφεία και άλλες υπηρεσίες ανέρχεται στα 27.5 εκ. δολάρια. Αυτό σημαίνει ότι για τη λειτουργία ενός circuit πόλης δαπανούνται συνολικά περίπου 57.5 εκ. δολάρια (Forbes, 2017). Το κόστος κατασκευής καινούριου μόνιμου circuit με προδιαγραφές F1 ανέρχεται περίπου στα 270 εκ. δολάρια με κόστος συντήρησης και ετήσιας χρήσης περίπου 18.5 εκ. δολάρια, ποσό πολύ χαμηλότερο από αυτό που απαιτείται ετησίως για τα circuit πόλης.

Ένα μέρος των εξόδων καλύπτεται από τους φίλους του μηχανοκίνητου αθλητισμού οι οποίοι συνδράμουν με την αγορά εισιτηρίων και τη χρήση των υπόλοιπων υπηρεσιών που παρέχονται τις μέρες του GP όπως φαγητό, ποτό. Αρκεί να αναφέρουμε ότι το βρετανικό GP του 2022 βρίσκεται στη πρώτη θέση σε προσέλευση θεατών με 440,000 θεατές για όλο το Σαββατοκύριακο ενώ το GP με τους λιγότερους θεατές για το 2022 ήταν αυτό του Μπαχρέιν με 98,000 θεατές. Οι τιμές των εισιτηρίων διαμορφώνονται ανάλογα με τη χώρα που διαδραματίζεται ο αγώνας. Για το 2023 τα φθηνότερα εισιτήρια είχε το GP της Βουδαπέστης με μέση τιμή εισιτηρίου στα $184 δολάρια ενώ τα ακριβότερα ήταν αυτά του Λας Βέγκας με μέση τιμή εισιτηρίου στα $1,667 δολάρια. Τέλος, οι χορηγοί, εν προκειμένω οι τοπικές εταιρείες και επιχειρήσεις χρηματοδοτούν το υπολειπόμενο ποσό αποσκοπώντας στην αυξημένη κίνηση της τοπικής αγοράς από τους θεατές του GP.

Μολονότι τα κέρδη είναι αυξημένα την περίοδο των GP για τις επιχειρήσεις της τοπικής κοινωνίας οι επιχειρηματίες δεν αποσκοπούν μόνο σε αυτό. Οι διαφημίσεις στα περιθώρια της πίστας, σε διάφορα σημεία εντός και εκτός αγωνιστικού χώρου καθώς και πάνω στα μονοθέσια και στις στολές των οδηγών είναι ύψιστης σημασίας για τους επιχειρηματίες και η κυριότερη πηγή εσόδων τόσο για τη διοργάνωση όσο και για τις ομάδες. Οι διαφημίσεις που τοποθετούνται στα περιθώρια της πίστας βρίσκονται αποκλειστικά υπό τη διαχείριση της εταιρείας Allsport Management S.A., μια εταιρεία από την Σουηδική Αραβία με έδρα τη Γενεύη που ανήκει στον όμιλο εταιρειών Formula One Group. Τα έσοδα μόνο των διαφημίσεων που τοποθετούνται στα περιθώρια της πίστας κυμαίνονται ετησίως στα 180 εκ δολάρια, τα οποία στο τέλος του έτους αντιστοιχούν περίπου στο 15% των συνολικών εσόδων (Mourao, 2017). Αντίστοιχη είναι η κατάσταση και για τις ομάδες, για τις οποίες το μεγαλύτερο κέρδος κάθε σεζόν προέρχεται από τους χορηγούς. Οι εταιρείες που σκοπεύουν να διαφημιστούν μέσω του αθλήματος συνάπτουν συμβόλαια με τις ομάδες, από ένα έως τρία έτη προκειμένου να μπει το λογότυπό τους είτε στο μονοθέσιο είτε στη στολή του οδηγού. Τέτοιες εταιρείες είναι οι SHELL, SANTANDER, RAY-BAN, BWT, MICROSOFT, ORACLE κα. Μάλιστα η Santander αποτελεί έναν από τους μεγαλύτερους χορηγούς για τη Ferrari αφού δίνει 60 εκ. δολάρια το χρόνο.

Τέλος, αξιοσημείωτα είναι τα ποσά που καταβάλλονται στην εταιρεία Liberty Media Corporation της οποίας ιδρυτής και πρόεδρος είναι ο John C. Malone. Η Liberty Media Corporation κατέβαλε για την αγορά του Formula One Group 4.4 δις δολάρια. Το αμερικανικό κανάλι ESPN κατέχει τα τηλεοπτικά δικαιώματα της F1 στις ΗΠΑ εδώ και μία πενταετία καταβάλλοντας ετησίως στην Liberty 75 εκ.. Η Liberty, ωστόσο, θέλει να αυξήσει το ποσό στα 100 εκ. δολάρια λόγω της ταχείας ανόδου του αριθμού των τηλεθεατών. Ακόμη, τα βρετανικά συνδρομητικά κανάλια του ομίλου Sky Sports έχουν υπογράψει συμφωνία ύψους 1.8 δις δολάρια (2019-2024), ποσό το οποίο υπολογίζεται ότι αυξήθηκε κατά 200 εκ. δολάρια τη σεζόν, με την επέκταση του συμβολαίου. Για το έτος 2022 σημειώνεται ότι η Liberty είχε συνολικά έσοδα ύψους 9 δις δολαρίων, ποσό μεγαλύτερο της προηγούμενης χρονιάς κατά την οποία είχε έσοδα ύψους 8.69 δις δολάρια. Από τα έσοδα αυτά αφαιρούνται τα κόστη και οι φόροι καθώς και χρήματα τα οποία δίνονται στις ομάδες της F1 στο τέλος κάθε σεζόν σύμφωνα με την κατάταξη στο

πρωτάθλημα των κατασκευαστών με τη μορφή επάθλου. Το 2023 η Red Bull που πήρε την πρώτη θέση στο πρωτάθλημα, πήρε από τη Liberty Media ποσό ύψους 140 εκ. δολαρίων. Οι Mercedes και Ferrari, κατέκτησαν τη δεύτερη και τρίτη θέση λαμβάνοντας χρηματικό έπαθλο ύψους 131 και 122 εκ. δολάρια αντίστοιχα. Τελευταία στην κατάταξη ήταν Haas η οποία και πήρε το μικρότερο έπαθλο στα 60 εκ. δολάρια.

2.3 Καινοτομίες της Formula 1 στα αυτοκίνητα δρόμου

Η Formula 1 αποτελεί πεδίο τεράστιας τεχνολογικής εξέλιξης στον τομέα του αυτοκινήτου με πολλές καινοτομίες που εφαρμόστηκαν πρώτη φορά στα μονοθέσια να βρίσκουν το δρόμο προς τη βιομηχανία των αυτοκινήτων δρόμου καθιστώντας την καθημερινή χρήση των αυτοκινήτων ευκολότερη και ασφαλέστερη για οποιονδήποτε πολίτη. Εδώ να σημειωθεί πως με τον όρο αυτοκίνητο δρόμου (road car) χαρακτηρίζεται το αυτοκίνητο που διαθέτει χώρο για παραπάνω του ενός άτομα και όλες τις νόμιμες προδιαγραφές ώστε να οδηγείται σε συνηθισμένο δρόμο. Αντίθετα, το αυτοκίνητο με προδιαγραφές τέτοιες ώστε να χρησιμοποιείται σε αγώνες καλείται αυτοκίνητο πίστας ή αγωνιστικό αυτοκίνητο (racing car). Μερικές από τις νέες τεχνολογίες που προσέφερε η F1 θα αναλυθούν παρακάτω.

Μια από τις προαναφερθείσες ευκολίες είναι τα κουμπιά που βρίσκονται στο τιμόνι του αυτοκινήτου (steering wheel buttons). Η συγκεκριμένη λειτουργία ξεκίνησε στη Formula 1 τη δεκαετία του 1970 όταν τοποθετήθηκαν στο τιμόνι τα πρώτα δύο κουμπιά για την ενδοεπικοινωνία μεταξύ ομάδων και οδηγού και για αλλαγές στη λειτουργία του κινητήρα (RACV, 2022). Στην πορεία τα κουμπιά καθώς και η πολυπλοκότητα των λειτουργιών τους άρχισαν να αυξάνονται με αποτέλεσμα σήμερα να βρίσκονται πάνω στο τιμόνι του οδηγού πάνω από εικοσιπέντε διαφορετικά κουμπιά. Σταδιακά η λειτουργία αυτή πέρασε και στα αυτοκίνητα δρόμου. Στα σύγχρονα αυτοκίνητα ο οδηγός μπορεί χρησιμοποιώντας μόνο το τιμόνι να ελέγξει τον σύστημα ήχου του αυτοκινήτου, να χρησιμοποιήσει ηλεκτρονικά το γκάζι και το φρένο ακόμη και να απαντήσει τηλεφωνική κλήση αν το κινητό του τηλέφωνο βρίσκεται συνδεδεμένο με το bluetooth του αυτοκινήτου χωρίς να μετακινήσει τα χέρια του από το τιμόνι. Είναι φανερό ότι η χρήση του αυτοκινήτου κατ’ αυτόν τον τρόπο είναι εξαιρετικά βοηθητική και για τα άτομα με ειδικές ανάγκες συμβάλλοντας στην αυτονομία τους αφού τα καθιστά ικανά να οδηγήσουν χωρίς τη βοήθεια τρίτου προσώπου. Στις λειτουργίες του τιμονιού προστίθεται και το πετάλι ταχυτήτων (paddle shifters) που βρίσκεται στο πίσω μέρος του τιμονιού. Η τεχνολογία αυτή χρησιμοποιήθηκε πρώτη φορά στο GP της Βραζιλίας το 1989 από τη Ferrari. Ουσιαστικά τοποθετήθηκε στο μονοθέσιο ένα ημι-αυτόματο κιβώτιο ταχυτήτων που επιτρέπει την αλλαγή ταχυτήτων από τα πετάλια στο πίσω μέρος του τιμονιού μέσω του ηλεκτροϋδραυλικού συστήματος. Το 1997 η Ferrari 355 ήταν το πρώτο αυτοκίνητο παραγωγής που χρησιμοποιήθηκε η παραπάνω λειτουργία. Έπειτα ακολούθησαν η BMW και η Alfa Romeo και τελικά όλοι οι κατασκευαστές (24h-lemans.com, 2011).

Καίρια, επίσης, ήταν η συμβολή της χρήσης ανθρακονήματος (carbon fiber) στη κατασκευή των σασί των μονοθέσιων. Συγκεκριμένα το ανθρακόνημα χρησιμοποιούταν κυρίως για τη μείωση του βάρους του μονοθέσιου χωρίς να είναι χρηστικό για κάποιον άλλο λόγο μέχρι το 1981 που ο τεχνικός διευθυντής της McLaren, John Barnard, το χρησιμοποίησε στη McLaren MP4/1 στο GP της Αργεντινής (RACV, 2022). Ο Barnard παρατήρησε πως αν το ανθρακόνημα τοποθετηθεί σε συγκεκριμένα σημεία του μονοθέσιου τότε η αντοχή του στην κρούση αυξάνεται σημαντικά με αποτέλεσμα την αύξηση της ασφάλειας του οδηγού σε πιθανό τρακάρισμα. Πλέον, όλα τα σασί των μονοθέσιων κατασκευάζονται με ανθρακόνημα προς αποφυγή θανατηφόρων ατυχημάτων. Λόγω αυτού, πολυτελή και ημι-αγωνιστικά

αυτοκίνητα όπως Porsche, BMW και Aston Martin κατασκευάζονται πλέον με ανθρακόνημα. Στα παραπάνω έρχονται να προστεθούν και οι προσαρμοστικές αναρτήσεις (adaptive suspension) που έκαναν την εμφάνιση τους στη F1 στα μέσα της δεκαετίας του ’80 με τις πρώτες εφαρμογές να κρατούν το κέντρο του αυτοκινήτου στη μικρότερη κατακόρυφη απόσταση από το έδαφος μεγιστοποιώντας την πρόσφυση. Το 1992 η Williams χρησιμοποίησε στο μονοθέσιο FW14B αποτελεσματικά τις προσαρμοστικές αναρτήσεις έτσι ώστε να μειωθεί ο κλυδωνισμός και να έχει το μονοθέσιο την καλύτερη δυνατή αεροδυναμική. Πλέον οι προσαρμοστικές αναρτήσεις τοποθετούνται σε πολυτελή και ημι-αγωνιστικά αυτοκίνητα παραγωγής από εταιρείες όπως οι Audi, BMW, Ford, Skoda, Land Rover και Mercedes-Benz. Τέλος, εξαιρετικά σημαντική αποτελεί η τεχνολογία υβριδικού συστήματος κινητήρα (hybrid powertrains). Όλες οι ομάδες της F1 ξεκίνησαν να πειραματίζονται με την τεχνολογία του υβριδικού κινητήρα το 2007 προσπαθώντας να ανακτήσουν κινητική ενέργεια από τη χρήση των φρένων κατά τη διάρκεια του αγώνα. Στην παρουσίαση των μονοθέσιων το 2014 ήταν απαραίτητη η χρήση υβριδικού κινητήρα από όλα τα μονοθέσια. Σήμερα στην F1 έχουμε δυο ξεχωριστά είδη ανάκτησης ενέργειας του υβριδικού κινητήρα. Το πρώτο ονομάζεται Moto Generator Unit – Kinetic (MGU-K) και αφορά στην ανάκτηση ενέργειας από τη χρήση των φρένων εν μέσω αγώνα. Το δεύτερο ονομάζεται Motor Generator Unit – Heat (MGU-H) και αφορά την ανάκτηση ενέργειας μέσω της θερμικής ενέργειας που εκπέμπεται από το τούρμπο του κινητήρα. Και στις δύο περιπτώσεις η ενέργεια αποθηκεύεται σε μία μπαταρία ιόντων λιθίου. Ενδεικτικά, δημοφιλή αυτοκίνητα παραγωγής που φοράνε κινητήρες με την συγκεκριμένη τεχνολογία είναι τα Toyota Urban Cruiser Hyryder, Honda City και Land Rover.

ΚΕΦΑΛΑΙΟ 3ο

Βιβλιογραφική ανασκόπηση σε εφαρμογές στη Formula 1

3.1 Εισαγωγή

Ο στόχος σε έναν αγώνα μηχανοκίνητου αθλητισμού είναι η επίτευξη του καλύτερου δυνατού αποτελέσματος. Αυτό επιτυγχάνεται με την ολοκλήρωση του αγώνα στο λιγότερο δυνατό χρόνο και κατ' επέκταση την εξασφάλιση περισσότερων κερδισμένων πόντων. Όμως, η κατασκευή ενός αξιόπιστου και ανταγωνιστικού μονοθέσιου σε συνδυασμό με έναν ικανό οδηγό που θα το εκμεταλλευτεί στο μέγιστο δεν είναι πάντα αρκετά. Υπάρχουν πολλοί παράγοντες που μπορούν να επηρεάσουν την έκβαση ενός αγώνα, επομένως, η υιοθέτηση στρατηγικής λαμβάνοντας υπόψιν όλα τα πιθανά σενάρια είναι ζήτημα μείζονος σημασίας. Ωστόσο, ο όγκος των παραγόμενων δεδομένων είναι τεράστιος και η ανάλυση τους αποτελεί χρονοβόρα και πολύπλοκη διαδικασία ειδικά όταν χρειάζεται να ληφθούν αποφάσεις σε ελάχιστο χρόνο στη διάρκεια του αγώνα. Πώς αντιμετωπίζεται το ζήτημα αυτό;

Την απάντηση στο ερώτημα δίνει η μηχανική μάθηση. Οι αλγόριθμοι μηχανικής μάθησης μπορούν να εκπαιδευτούν με τη χρήση ιστορικών δεδομένων και να δίνουν προβλέψεις για πολλές παραμέτρους που τυχόν επηρεάζουν την ροή του αγώνα και την έκβαση του αποτελέσματος. Μπορούν, επίσης, να κατασκευαστούν νευρωνικά δίκτυα που θα προβλέπουν την καλύτερη στρατηγική που πρέπει να ακολουθήσει μια ομάδα ώστε να οδηγηθεί στο επιθυμητό αποτέλεσμα εξετάζοντας δεδομένα πραγματικού χρόνου που τροφοδοτούνται σε αυτά την ώρα του αγώνα και λαμβάνοντας υπόψιν τη στρατηγική που πιθανόν να ακολουθήσουν αντίπαλες ομάδες. Ακόμη, με τη χρήση των αλγορίθμων και των ιστορικών δεδομένων μπορούν να προβλεφθούν τυχόν ατυχήματα ή να διαπιστωθούν παράγοντες που αυξάνουν τον κίνδυνο σε έναν αγώνα ώστε αυτοί να περιοριστούν και να βελτιωθεί η ασφάλεια των οδηγών. Τέλος, τέτοιες μέθοδοι χρησιμεύουν για τη διευκόλυνση διαφόρων μελών της ομάδας στην πρόσβαση, ανάγνωση και χρήση των αναγκαίων δεδομένων και γραφικών καθώς και στον αντίκτυπο που έχουν στο κοινό οποιασδήποτε μορφής αλλαγές στον τρόπο διεξαγωγής του αγώνα.

3.2 Σχετική εργασία

Κατά τη διάρκεια της βιβλιογραφικής ανασκόπησης διαπιστώθηκε πως έχουν γίνει πολλές έρευνες και εργασίες σχετικά με τη χρήση της μηχανικής μάθησης σε δεδομένα μηχανοκίνητου αθλητισμού στοχεύοντας στην πρόβλεψη της πιο κερδοφόρας στρατηγικής κατά τη διάρκεια του αγώνα. Επιπρόσθετα, έχουν γίνει μελέτες που αποσκοπούν στη βελτίωση της απόδοση του μονοθέσιου αλλά και στην εξοικονόμηση καυσίμων. Αναλυτικότερα, στην εργασία του L. G. Tajeda (2023) χρησιμοποιήθηκαν οι μέθοδοι τυχαίου δάσους (Random Forest - RF), μηχανών διανυσμάτων υποστήριξης (Support Vector Machines - SVM) και τεχνητών νευρωνικών δικτύων (Artificial Neural Network - ANN) με στόχο την δημιουργία αξιόπιστου συστήματος για την αποτελεσματική πρόβλεψη της ορθής χρονικής

στιγμής του pit stop. Κατασκευάστηκε μοντέλο που προβλέπει τόσο την ορθή χρονική στιγμή για να γίνει pit stop όσο και το πόσο αποτελεσματική ήταν η στρατηγική τού να γίνει pit stop τη συγκεκριμένη χρονική στιγμή. Επομένως, έχουμε δύο μεταβλητές απόκρισης, την «Has pit stop» και την «Good pit stop». Μετά την εφαρμογή των μεθόδων φαίνεται πως βάσει του μέτρου f1-score καλύτερη πρόβλεψη δίνει η μέθοδος SVM και για τις δύο μεταβλητές με το f1-score για την Has pit stop να ισούται με περίπου 62% ενώ το f1-score για την Good pit stop να ισούται με περίπου 44%. Τα ANN βρίσκονται ελάχιστα πίσω σε f1-score από το SVM με το RF να δίνει το χειρότερο f1-score και στις δύο μεταβλητές. Παρατηρείται, επομένως, πως ενώ το SVM (και το ANN) δίνει σχετικά ικανοποιητική πρόβλεψη για τη μεταβλητή Has pit stop, κανένα μοντέλο δε δίνει ικανοποιητική πρόβλεψη για τη μεταβλητή Good pit stop. Αντίστοιχα, οι M. Boettinger και D. Klotz (2023) στην έρευνά τους προτείνουν μια καινοτόμα μέθοδο που βασίζεται σε μεθόδους μηχανικής μάθησης για τη βελτιστοποίηση της διαδικασίας λήψης αποφάσεων για τη στρατηγική σε αγώνες Gran Turismo (GT). Συγκεκριμένα, η προτεινόμενη προσομοίωση υιοθετεί το Nürburgring Langstreckenserie, μια σειρά αγώνων δέκα ετήσιων εκδηλώσεων που λαμβάνουν χώρα στο Νίρμπουργκρινγκ στοχεύοντας στην κατασκευή ενός αυτόματου συστήματος λήψης στρατηγικών αποφάσεων με τη χρήση προσομοιώσεων των αγώνων, νευρωνικών δικτύων και ενισχυτικής μάθησης (reinforcement learning).

Στην έρευνα των C. Zhang et al. (2023) δομείται ένα διαδοχικό μοντέλο ελέγχου λήψης αποφάσεων για προσπέραση, που δίνει συνεχόμενα αποτελέσματα δράσης, βασισμένο στον αλγόριθμο ενισχυτικής μάθησης Deep Deterministic Policy Gradient (DDPG). Με τη χρήση της πλατφόρμας TORCS, ένα λογισμικό ανοιχτού κώδικα για την 3D προσομοίωση αγώνων, κατασκευάστηκε ένα ακριβές σενάριο αγώνα. Εισάγοντας στο λογισμικό τα περιβαλλοντικά δεδομένα του οχήματος - θερμοκρασία αέρα, δεδομένα από αισθητήρες βροχής, κατάσταση υαλοκαθαριστήρα, κατάσταση συστήματος πρόσφυσης και άλλες πληροφορίες που προέρχονται από αισθητήρες τοποθετημένους στο όχημα - για εκπαίδευση, αυτό δίνει αμέσως την απόφαση που πρέπει να παρθεί για τον τρόπο οδήγησης του οχήματος. Από τα πειράματα που έγιναν μέσω προσομοιώσεων φαίνεται πως ο αλγόριθμος DDPG είναι ικανός για την μοντελοποίηση διαδικασίας προσπέρασης με ασφάλεια και σταθερότητα σε πολλαπλές φορές υπό συνθήκη μηδενικών συγκρούσεων και ατυχημάτων. Εν αντιθέσει με το Deep Q Learning μοντέλο διακριτής δράσης, ο DDPG φαίνεται να είναι καταλληλότερος για τη χρήση ελέγχου απόφασης προσπέρασης στον χώρο της μη επανδρωμένης οδήγησης.

Στην έρευνα των W. Villegas-Ch et al. (2023) κατασκευάζεται ένα νευρωνικό δίκτυο μακράς βραχυπρόθεσμης μνήμης (Long Short-Term Memory Network - LSTM) συνδυαστικά με τεχνικές εξήγησης (explainability techniques) με στόχο την πρόβλεψη της θέσης των οδηγών σε αγώνες βάσει ιστορικών δεδομένων. Η συγκεκριμένη μελέτη στοχεύει στην ανάδειξη της σημαντικότητας των τεχνικών εξήγησης καθώς υποστηρίζεται πως μέσω αυτών επιτυγχάνεται όχι μόνο καλύτερη κατανόηση των μοντέλων μηχανικής μάθησης αλλά και ενίσχυση της αμεροληψίας, διαφάνειας και ικανότητας εξήγησης και τεκμηρίωσης των αποτελεσμάτων για διαδικασίες λήψης απόφασης βασισμένες σε μοντέλα τεχνητής νοημοσύνης. Ειδικότερα, χρησιμοποιήθηκαν οι τεχνικές attention και permutation feature importance. Από το αποτέλεσμα της μελέτης φαίνεται πως με τη χρήση αυτών των τεχνικών αποκτήθηκε χρήσιμη πληροφορία για τα χαρακτηριστικά (μεταβλητές) και την επιρροή τους στις προβλέψεις του μοντέλου καθώς τα αποτελέσματα έδειξαν ότι πρέπει να ληφθούν υπόψιν χαρακτηριστικά που αφορούν στην γενικότερη απόδοση του εκάστοτε οδηγού κατά τη διάρκεια της καριέρας του, στα τεχνικά χαρακτηριστικά του αυτοκινήτου καθώς και στις καιρικές συνθήκες αν επιδιώκουμε έγκυρες προβλέψεις για την έκβαση του αγώνα.

Στη μελέτη των A. Patil et al. (2022) γίνεται μια συστηματική ανάλυση των διαφόρων παραγόντων που μπορεί να επηρεάσουν έναν αγώνα καθώς και το αποτέλεσμα αυτού. Για τη διαδικασία της ανάλυσης χρησιμοποιήθηκαν τεχνικές μηχανικής μάθησης αφού πρώτα έγινε προπαρασκευή των δεδομένων. Αναφέρεται πως επιτεύχθηκε εξακρίβωση των σημαντικότερων μεταβλητών που κρίνουν την έκβαση ενός αγώνα αλλά και πως όλες οι μεταβλητές είναι ισχυρά συσχετισμένες μεταξύ τους. Τέλος, εκφράζεται η υπόθεση πως δύναται να μειωθεί ο αριθμός των μεταβλητών του συνόλου δεδομένων χωρίς να χαθεί σημαντική πληροφορία.

Η μελέτη των K. Cheng (2023) στοχεύει στο πρόβλημα εξοικονόμησης καυσίμων βελτιστοποιώντας το πίσω φτερό (rear wing) του μονοθέσιου έτσι ώστε να αυξήσει τη συνολική πρόσφυση του αυτοκινήτου και στην ουσία να ελαχιστοποιήσει το φρενάρισμά του. Το πίσω φτερό επιλέχθηκε γιατί συνεισφέρει κατά 25% στη συνολική πρόσφυση του μονοθέσιου και γιατί είναι υπεύθυνο για το 30% του φρεναρίσματος του αυτοκινήτου πράγμα που το κάνει θέμα καίριας σημασίας για την αεροδυναμική επίδοση του μονοθέσιου. Ωστόσο, δεν είναι ξεκάθαρος ο τρόπος με τον οποίο μπορεί να βελτιωθεί το πίσω φτερό του μονοθέσιου αλλά και το γιατί. Για να απαντηθούν τα παραπάνω χρησιμοποιούνται προσομοιώσεις υπολογιστικής ρευστοδυναμικής (Computational Fluid Aerodynamics – CDF simulations) σε συνδυασμό με έναν αλγόριθμο οπισθοδιάδοσης τεχνητού νευρωνικού δικτύου (Backpropagation Artificial Neural Network algorithm). Από την έρευνα προκύπτει ότι η βέλτιστη αεροτομή στο πίσω φτερό έχει κατά 175% περισσότερο πάχος σχετικά με την προεπιλεγμένη αεροτομή NACA 63-412 και έχει γωνία επίθεσης μικρότερη κατά 6 μοίρες. Τέλος, η χρήση πίσω φτερού με τις άνωθεν προδιαγραφές προσφέρει στο μονοθέσιο κατά 43,04% μείωση φρεναρίσματος και κατά 7,131% πρόσφυση συγκριτικά με ένα μονοθέσιο χωρίς πίσω φτερό.

Στη μελέτη των S. Ju et al. (2023) δομείται ένα μοντέλο ενισχυτικής μάθησης που μπορεί να αντιληφθεί και να μεταφράσει μία δοθείσα κατάσταση (contextual Reinforcement Learning - cRL) με σκοπό την εκπαίδευση μιας αγωνιστικής οδηγικής τακτικής για επαγγελματική υψηλής πιστότητας προσομοίωση αγωνιστικού αυτοκινήτου. Στο μοντέλο προσαρμόστηκαν δεδομένα που συλλέχθηκαν από προσομοιωτή Drive-in-Loop (DiL, πρόγραμμα οδήγησης σε βρόχο). Ο προσομοιωτής DiL παρέχει ένα εξαιρετικά εμβριθές περιβάλλον, όπου οι δράσεις του αυτοκινήτου-ελέγχου και του οδηγού συνενώνονται παρέχοντας κρίσιμες και αξιόπιστες πληροφορίες κατά τη διάρκεια της εξέλιξης και επιβεβαίωσης του πλαισίου του αυτοκινήτου, της εκπαίδευσης του οδηγού και του Προηγμένου Συστήματος Υποβοήθησης Οδηγού (Advanced Driver Assistance System - ADAS). Τα δείγματα από την κατανομή αναφοράς χρησιμοποιούνται ως αναφορά οδηγικών γραμμών (driving lines) κατά τη διάρκεια της διερεύνησης και παρέχονται σχετικές πληροφορίες στο υπολογιστικό σύστημα (ή πράκτορα) ενισχυτικής μάθησης μελετώντας διαφορετικές δράσεις και συναρτήσεις ανταμοιβής βασισμένες στο εκάστοτε πείραμα. Φαίνεται, λοιπόν, ότι το μοντέλο αυτό μπορεί δυνητικά να χρησιμοποιηθεί ως ψηφιακό δίδυμο του προσομοιωτή DiL.

Οι J. von Schleinitz et al. (2022) στην έρευνά τους στοχεύουν στην παροχή μεθόδων αξιολόγησης των οδηγών αγώνων προκειμένου να καταλήξουν σε συμπεράσματα για το οδηγικό στυλ και να επισημάνουν περιοχές που χαίρουν βελτίωσης χωρίς ωστόσο να απαιτείται η χρήση πίστας. Θεωρώντας πως τις καλύτερες προϋποθέσεις για ανάλυση οδηγικού στυλ παρέχει μόνο ο προσομοιωτής DiL οι ερευνητές ανέπτυξαν τις μεθόδους τους βασιζόμενοι στον προσομοιωτή της BMW (BMW Motorsports simulator) ο οποίος βρίσκεται σε λειτουργία από το 2017. Για την εκτίμηση των σκορ της εξέλιξης των οδηγών προτείνεται ένα νευρωνικό δίκτυο μακράς βραχυπρόθεσμης μνήμης (LSTM).

Στην έρευνα των B. D. Evans et al. (2023) αναφέρεται το πρόβλημα εκπαίδευσης πράκτορα βαθιάς ενισχυτικής μάθησης (Deep Reinforcement Learning – DRL agent) ώστε να αγωνίζονται οι οδηγοί σε υψηλές ταχύτητες ενισχύοντας παράλληλα την ασφάλεια κατά τη διάρκεια της διαδικασίας εκπαίδευσης προς αποφυγή συγκρούσεων. Παρουσιάζεται ένας επόπτης βασισμένος σε μία θεωρία βιωσιμότητας για την ενίσχυση της ασφάλειας υψηλής απόδοσης που χρησιμοποιείται για την εκπαίδευση πράκτορα για προσομοιωμένους αγώνες F1 Tenth. Το F1 Tenth είναι μια πλατφόρμα λογισμικού ανοιχτού κώδικα για αγώνες μη επανδρωμένων αυτοκινήτων που οργανώνεται σε εξαμηνιαία βάση σε ρομποτικά συνέδρια όπως το IROS και το ICRA.

Ο I.H.D. den Hartog (2022) στην εργασία του, θεωρώντας πως θα ήταν χρήσιμο, προτείνει την κατασκευή ενός διαδραστικού συστήματος που παρουσιάζει σε εργαζόμενους διαφορετικών τμημάτων μόνο τα δεδομένα και τα γραφικά που χρειάζονται για τη λήψη αποφάσεων, μειώνοντας έτσι τον όγκο δεδομένων που έχουν να αντιμετωπίσουν. Η ιδέα, ουσιαστικά, είναι η κατασκευή ενός αυτόματου συστήματος προσωποποίησης γραφικών και απεικονίσεων με χρήση ενός μοντέλου που εφαρμόζεται μέσω ενός περιβάλλοντος χρήσης γραφικών.

Οι Amsury et al., 2022, στην έρευνά τους εξετάζουν την ανταπόκριση που είχε ο αγώνας της Formula E και αν το κοινό της Ινδονησίας εγκρίνει ή απορρίπτει τη Formula E βασιζόμενοι στα σχόλια που συγκέντρωσαν από τις πλατφόρμες κοινωνικής δικτύωσης και συγκεκριμένα του Twitter (νυν X). Για την επίτευξη του παραπάνω συνέκριναν τα αποτελέσματα δύο αλγορίθμων ταξινόμησης, των μηχανών διανυσμάτων υποστήριξης (SVM) και Naïve Bayes αφού πρώτα εφάρμοσαν την τεχνική Synthetic Minority Oversampling TEchnique – SMOTE. Τα δεδομένα που χρησιμοποιήθηκαν είναι ποιοτικά και λήφθηκαν ως tweet δεδομένα με το ερώτημα Formula E. Από τα αποτελέσματα φαίνεται ότι η ορθότητα του SVM είναι 88,11% ενώ η ορθότητα του Naïve Bayes είναι 87.54%. Το υψηλό ποσοστό ακρίβειας για την ταξινόμηση στην τάξη που υποστηρίζει τις εκδηλώσεις της Formula E τόσο για το SVM (97.86%) όσο και για τον Naïve Bayes (96.17%) υποδηλώνει ότι πολλά tweets ήταν υπέρ της καθιέρωσης εκδηλώσεων Formula E. .

Παρακάτω παρατίθενται σε μεγαλύτερη ανάλυση κάποιες μελέτες που εντοπίστηκαν έπειτα από την παραπάνω διερεύνηση.

3.3 Η χρήση της προσομοίωσης σε συνδυασμό με μηχανική μάθηση και βελτιστοποίηση για ένα ψηφιακό δίδυμο – Μελέτη συστήματος υποστήριξης αποφάσεων στη Formula 1

Στην έρευνα των Greasley et al. (2022) εξετάζεται πως μπορεί μέσω της χρήσης προσομοίωσης, αλγορίθμων μηχανικής μάθησης και βελτιστοποίησης να παραχθεί ένα ψηφιακό δίδυμο (digital twin) ώστε να στηθεί ένα σύστημα υποστήριξης αποφάσεων (Decision Support System - DSS) προκειμένου να λαμβάνονται αποφάσεις για την επιλογή της βέλτιστης στρατηγικής σχετικά με τη διαχείριση των ελαστικών και τον αριθμό των pit stops που πρέπει να γίνουν κατά τη διάρκεια ενός αγώνα. Είναι γνωστό ότι στην Formula 1 εν μέσω αγώνα οι αποφάσεις για τη στρατηγική πρέπει να λαμβάνονται σε κλάσματα δευτερολέπτου και ταυτόχρονα να είναι αποτελεσματικές.

Στην παρούσα μελέτη λοιπόν εξηγείται πώς θα ήταν εφικτό, στήνοντας ένα ψηφιακό δίδυμο, να κατασκευαστεί ένα μοντέλο που θα δίνει σε κάθε αγώνα με βάσει τα χαρακτηριστικά που διαθέτει το αυτοκίνητο, τις ικανότητες του οδηγού, τις επικρατούσες καιρικές συνθήκες και το είδος της πίστας, την καλύτερη δυνατή στρατηγική pit stop

προκειμένου ιδανικά να επέλθει η νίκη. Η ερωτήσεις που πρέπει να απαντηθούν από το μοντέλο είναι οι εξής:

1. Πότε να γίνει το pit stop κατά τη διάρκεια του αγώνα;
2. Ποιο είναι το είδος των ελαστικών που πρέπει να τοποθετηθεί στο εκάστοτε pit stop;
3. Ποια είναι η προβλεπόμενη θέση του μονοθέσιου αμέσως μετά το pit stop και ποια η θέση που πρέπει να βρίσκεται πριν το επόμενο pit stop;

Περιπλέκοντας ακόμη περισσότερο τη φύση του μοντέλου που θέλουμε να κατασκευάσουμε είναι απαραίτητο να λάβουμε υπόψιν στην απόφαση της στρατηγικής και μια ακόμη ερώτηση. «Ποια είναι η πιθανή στρατηγική pit stop που θα ακολουθήσουν οι αντίπαλες ομάδες;»

Μετά από κάθε αγώνα παράγεται τεράστιος όγκος δεδομένων που χρήζουν εκτενούς ανάλυσης για να βρεθούν κατάλληλα προβλεπτικά μοντέλα. Παρότι, λοιπόν, η ανθρώπινη νοημοσύνη είναι απαραίτητη για την προσδοκώμενη έκβαση ενός αγώνα, χωρίς τη χρήση της επιβλεπόμενης μηχανικής μάθησης η λήψη τόσο γρήγορων αποφάσεων από τον υπεύθυνο στρατηγικής θα ήταν πρακτικά αδύνατη. Οι αλγόριθμοι τροφοδοτούνται μετά από κάθε αγώνα με δεδομένα ώστε να παραχθούν μοντέλα που θα δίνουν τις αναγκαίες προβλέψεις σχετικά με τον χρόνο του γύρου, τη φθορά των ελαστικών, την συμπεριφορά των ανταγωνιστών και την πιθανότητα ύπαρξης αυτοκινήτου ασφαλείας ή εικονικού αυτοκινήτου ασφαλείας. Έτσι οι μηχανικοί και οι υπεύθυνοι στρατηγικής αποκτούν σημαντική γνώση στον χειρισμό αντίστοιχων καταστάσεων κατά τη διάρκεια της χρονιάς. Οι διαδικασίες αυτές γίνονται τόσο με τη χρήση ιστορικών δεδομένων όσο και με δεδομένα πραγματικού χρόνου εν μέσω αγώνα. Επομένως, σύμφωνα με τους συγγραφείς, η χρήση προσομοίωσης σε συνδυασμό με αλγορίθμους μηχανικής μάθησης και βελτιστοποίησης μπορεί να δώσει αξιοσημείωτα αποτελέσματα σε σύγχρονη και ασύγχρονη ανάλυση και να προσφέρει στη γρήγορη λήψη αποφάσεων αρκεί να ληφθούν υπόψιν οι επικείμενοι περιορισμοί που αντιμετωπίζει κάθε μέθοδος.

3.4 Αναλυτική και προγνωστική μελέτη για την Formula 1 και την ασφάλεια βασισμένη σε αλγορίθμους μηχανικής μάθησης

Στη μελέτη των Dhanvanth et al. (2022) γίνεται χρήση αλγορίθμων μηχανικής μάθησης και εργαλείων απεικόνισης δεδομένων προκειμένου να προβλεφθούν πιθανά ατυχήματα σε επερχόμενους αγώνες με σκοπό την ενίσχυση της ασφάλειας των οδηγών. Επιπλέον, χρησιμοποιείται ο καταλληλότερος αλγόριθμος για την πρόβλεψη του νικητή ενός αγώνα. Τα παραπάνω παρουσιάζονται σε τρεις φάσεις με την πρώτη να αφορά στην σύγκριση τριών αλγορίθμων ως προς την ακρίβειά τους στην πρόβλεψη ατυχημάτων. Έπειτα ακολουθεί η δεύτερη φάση που αφορά στην εξακρίβωση τριών παρατηρήσεων που είναι οι εξής: (i) πώς έχει αλλάξει με την πάροδο των χρόνων η ταχύτητα των μονοθέσιων και αν αυτή αυξάνεται συνεχώς με τις αναβαθμίσεις που υφίστανται τα μονοθέσια, (ii) κατά πόσο σχετίζεται η θέση του μονοθέσιου στη σχάρα εκκίνησης με τη θέση που αυτό θα τερματίσει τον αγώνα και (iii) σύγκριση μεταξύ δύο εκ των δυνατότερων οδηγών και αντιπάλων της τελευταίας δεκαετίας. Τέλος, στην τρίτη φάση παρουσιάζεται ένα μοντέλο μηχανικής μάθησης για την πρόβλεψη του νικητή ενός αγώνα βασισμένο σε συσχετισμένους παράγοντες.

Τα δεδομένα για την έρευνα λήφθηκαν από το Kaggle και περιέχουν πληροφορίες για τους κατασκευαστές, τους οδηγούς, τους χρόνους των γύρων, τα pit stops, και τα αποτελέσματα των αγώνων την περίοδο 1950-2017. Δημιουργήθηκαν, επίσης, και datasets με πληροφορίες από εφημερίδες και αφορούν πληροφορίες για τις καιρικές συνθήκες στη διάρκεια του αγώνα. Μετά την προετοιμασία των datasets ακολούθησε προπαρασκευή των δεδομένων. Όσον αφορά τα δεδομένα για την πρόβλεψη των ατυχημάτων αυτά μετασχηματίστηκαν ώστε να έχουν συγκεκριμένο εύρος με τη μέθοδο mix-max scaler και στη συνέχεια εφαρμόστηκε η μέθοδος SMOTE (Synthetic Minority Oversampling Technique) για την ισορροπία του μεγέθους των κλάσεων. Οι ελλείπουσες τιμές διαγράφηκαν εντελώς μιας και ήταν ελάχιστες ώστε να αποφευχθεί τυχόν απόκλιση που δημιουργούν οι μέθοδοι αντικατάστασής τους.

Για την πρώτη φάση εφαρμόστηκαν στα προπαρασκευασμένα δεδομένα τρεις διαφορετικές μέθοδοι, ο αλγόριθμος τυχαίου δάσους (random forest), ο αλγόριθμος των Κ κοντινότερων γειτόνων (K Nearest Neighbours - KNN) και η λογιστική παλινδρόμηση (logistic regression). Μετά την εφαρμογή τους φαίνεται πως το μοντέλο με την χαμηλότερη ακρίβεια είναι η λογιστική παλινδρόμηση με accuracy 44.74%, μετά ακολουθεί το τυχαίο δάσος με accuracy 72.20% και πιο κατάλληλος δείχνει να είναι ο αλγόριθμος k κοντινότερων γειτόνων με accuracy 85.93%. Συνεπώς, ο αλγόριθμος KNN θα μπορούσε να χρησιμοποιηθεί για την πρόβλεψη ατυχημάτων με στόχο την αποφυγή τους.

Στη δεύτερη φάση, τα διαγράμματα που κατασκευάστηκαν φανερώνουν ότι η ταχύτητα των αυτοκινήτων δείχνει να μειώνεται στα μέσα της περιόδου που εξετάζεται και αυτό μάλλον οφείλεται στους κανονισμούς που έπρεπε οι κατασκευαστές να εφαρμόσουν για την ενίσχυση των προδιαγραφών ασφαλείας των μονοθέσιων. Όσον αφορά τη σημαντικότητα της θέσης εκκίνησης για το αποτέλεσμα του οδηγού στο τέλος του αγώνα, το διάγραμμα διασποράς που κατασκευάστηκε δείχνει ότι οι οδηγοί που βρίσκονται στις πρώτες θέσεις έχουν πλεονέκτημα έναντι των οδηγών που εκκινούν από τις τελευταίες θέσεις της σχάρας εκκίνησης, αφού συναντούν λιγότερη κίνηση στην έναρξη του αγώνα. Γενικά, φαίνεται πως οι οδηγοί τείνουν να τερματίζουν κοντά στη θέση από τη οποία εκκίνησαν. Τέλος, η σύγκριση που έγινε μεταξύ των δύο οδηγών για τη απόδοση τους σε διαφορετικές πίστες παρουσιάζει διαφορές στο αποτέλεσμα τους και φανερώνει πως ίσως η κάθε πίστα ευνοεί διαφορετικό οδηγό ή μονοθέσιο.

Στην τρίτη φάση εφαρμόστηκε λογιστική παλινδρόμηση αφού η εξαρτημένη μεταβλητή μπορεί να έχει δύο τιμές (κέρδισε: ναι / όχι). Επειδή την τελευταία δεκαετία η ομάδα της Mercedes και ο οδηγός της Lewis Hamilton είχαν μεγάλη επιτυχία με 8 πρωταθλήματα κατασκευαστών και εφτά οδηγών (ένα εκ των οποίων ανήκει στον Nico Rosberg) το μοντέλο τείνει να παρουσιάζει μεροληψία. Για αυτό λαμβάνονται υπόψιν μόνο ο αριθμός των νικών και ο αριθμός των βάθρων του κάθε οδηγού στους τελευταίους πέντε αγώνες και χρησιμοποιούνται στο μοντέλο οι μεταβλητές με τις υψηλότερες συσχετίσεις. Η ορθότητα της πρόβλεψης του μοντέλου ελέγχθηκε με βάσει τα αποτελέσματα των αγώνων της σεζόν 2020 με το μοντέλο να προβλέπει σωστά το νικητή σε 11 από τους 17 αγώνες δίνοντας accuracy 64.7%. Παρατηρήθηκε ότι απρόσμενες καταστάσεις όπως ατυχήματα και ποινές δε μπορούσαν να συμπεριληφθούν.

3.5 Εικονικός Μηχανικός Στρατηγικής: Χρήση τεχνητών νευρωνικών δικτύων για αποφάσεις στρατηγικής αγώνα σε μηχανοκίνητο άθλημα πίστας

Στη μελέτη των Heilmeier et al (2020) γίνεται χρήση μηχανικής μάθησης και συγκεκριμένα νευρωνικών δικτύων (Neural Networks - NN) προκειμένου να δημιουργηθεί ένας εικονικός μηχανικός στρατηγικής (Virtual Strategy Engineer - VSE) για την επίτευξη της βέλτιστης στρατηγικής όσον αφορά τον προγραμματισμό των pit stops κατά τη διάρκεια του αγώνα και την επιλογή του κατάλληλου συνδυασμού ελαστικών.

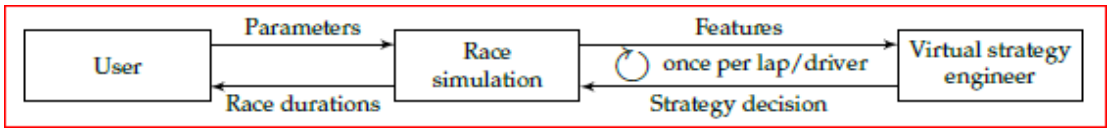
Η αρχική ιδέα που εφαρμόστηκε ήταν να κατασκευαστούν δύο ξεχωριστά νευρωνικά δίκτυα. Το πρώτο θα κατασκευαστεί και θα εκπαιδευτεί με τέτοιο τρόπο ώστε να προβλέπει σε ποιο σημείο του αγώνα χρειάζεται να πραγματοποιηθεί pit stop και το δεύτερο θα κατασκευαστεί για να προβλέπει ποια σύνθεση ελαστικών είναι απαραίτητη για το εκάστοτε σημείο του αγώνα. Τα δεδομένα που χρησιμοποιήθηκαν για τις παραπάνω μεθόδους αφορούν την περίοδο 2014-2019 κατά την οποία δεν υπήρχαν ιδιαίτερα σημαντικές αλλαγές στους κανονισμούς και περιέχουν πληροφορίες για το χρόνο του γύρου, τις θέσεις των οδηγών, τα pit stops και τις FCY φάσεις (Full-Course Yellow). Οι FYC φάσεις αφορούν τις περιπτώσεις που οι οδηγοί είναι υποχρεωμένοι βάσει κανονισμού να μειώσουν ταχύτητα λόγω αγωνιστικού συμβάντος, συνήθως ατυχήματος. Στα δεδομένα οι FCY φάσεις χωρίζονται σε δύο τύπους: εικονικό αυτοκίνητο ασφαλείας (Virtual Safety Car - VSC) και αυτοκίνητο ασφαλείας (Safety Car - SC). Σημειώνεται ότι κατά τη διάρκεια της περιόδου από την οποία αντλούνται τα δεδομένα, υπήρξαν μικρές διαφοροποιήσεις είτε στην ποιότητα και τη διάμετρο των ελαστικών, είτε στη διαδικασία που πρέπει να ακολουθούν οι ομάδες, οι οποίες μπορεί να επηρεάσουν τα αποτελέσματα που θα μας δώσει η μέθοδος αν αυτή εφαρμοστεί σε όλο το dataset.

Έπειτα ακολουθεί η επιλογή των μεταβλητών (features) για καθένα από τα δύο NN καθώς και η προπαρασκευή των δεδομένων. Αναλυτικότερα, αφαιρούνται από το dataset δεδομένα που αφορούν αγώνες με βροχή καθώς υπό τέτοιες συνθήκες οι αποφάσεις λαμβάνονται από τα σχόλια και την κρίση του οδηγού. Επιπλέον, δεν λαμβάνονται υπόψιν οδηγοί που έκαναν παραπάνω από τρία pit stops κατά τη διάρκεια του αγώνα ή αφού είχαν συμπληρώσει το 90% των γύρων αφού αυτά οφείλονται κυρίως σε ατυχήματα και αστοχίες τους αυτοκινήτου και δε βοηθούν στην εκπαίδευση του NN. Τέλος, δε λαμβάνονται υπόψιν οδηγοί με χρόνο γύρου μεγαλύτερο των διακοσίων δευτερολέπτων ή με διάρκεια pit stop μεγαλύτερη των 50 δευτερολέπτων μιας και τέτοιοι χρόνοι υποδηλώνουν πιθανές βλάβες του μονοθέσιου και δεν πρέπει να επηρεάσουν την εκπαίδευση του NN. Όσον αφορά τα μέτρα με τα οποία επιτηρούμε το NN, το σύνηθες είναι η ακρίβεια (accuracy). Στην περίπτωση όμως του πρώτου NN που κατασκευάζεται θα χρησιμοποιηθούν ως μέτρα τα precision και recall μέσω του F score το οποίο συνδυάζει τα δύο προαναφερόμενα μέτρα. Η μαθηματική φόρμουλα του F score είναι: F-score = 2 * (precision * recall) / (precision + recall). Για το δεύτερο NN θα χρησιμοποιηθεί το μέτρο accuracy.

Για την πρόβλεψη της βέλτιστης στρατηγικής πραγματοποίησης pit stop χρησιμοποιήθηκε αρχικά ένας τύπος τεχνητού νευρωνικού δικτύου στο οποίο οι κόμβοι δεν δημιουργούν βρόχο επανάληψης και επομένως η πληροφορία που εισάγεται στο NN μετακινείται μόνο προς τα εμπρός (feed-forward neural network - FFNN). Το FFNN έδωσε F score περίπου 0.35, πολύ χαμηλό σκορ για πρόβλεψη. Για αυτό και έγινε δοκιμή με επαναλαμβανόμενο νευρωνικό δίκτυο (recurrent neural network - RNN) το οποίο χρησιμοποιεί χρονοσειρές ή διαδοχικά δεδομένα. Μολονότι το RNN έδωσε F score περίπου 0.9 τείνει να δίνει ένα αρκετά

ακατανόητο προς τον ερευνητή αποτέλεσμα. Τα παραπάνω οδήγησαν σε προσπάθεια συνδυασμού των πλεονεκτημάτων που παρέχει ο κάθε τύπος νευρωνικού δικτύου δημιουργώντας έτσι ένα υβριδικό NN (hybrid neural network - HNN). Το HNN δίνει χαμηλότερο F score από το RNN, παρέχει όμως αρκετά πιο κατανοητά προς τον ερευνητή αποτελέσματα. Για την απόφαση των ελαστικών που θα χρησιμοποιηθούν δημιουργείται ένα NN με 32 νευρώνες προσαρμοσμένο να αποφασίζει μεταξύ τριών διαφορετικών τύπων γόμας: μαλακή – soft, μεσαία – medium και σκληρή – hard. Στην τελευταία διαμόρφωση του NN επιτυγχάνεται accuracy περίπου 0.77 που σημαίνει ότι η πλειοψηφία των αποφάσεων έχει προβλεφθεί σωστά.

Παρακάτω παρατίθεται ένα σχήμα που συνοψίζει την ενοποίηση του VSE στην προσομοίωση αγώνα. Είναι αντιληπτό ότι το VSE καλείται μία φορά σε κάθε γύρο και ο οδηγός παίρνει απόφαση βάσει αυτού.



Στη συνέχεια γίνεται εφαρμογή του VSE για την πρόβλεψη των θέσεων που θα τερματίσουν οι οδηγοί με χρήση των δεδομένων από το Αυστριακό GP του 2019 μιας και στον συγκεκριμένο αγώνα δεν υπήρχαν ατυχήματα ή εγκαταλείψεις και συνεπώς δεν υπήρχαν FCY φάσεις. Το VSE βασίζεται στα τελικά NN που αναλύθηκαν παραπάνω, τα οποία θα προβλέψουν τη βέλτιστη στρατηγική και κατ' επέκταση την τελική κατάταξη των οδηγών στη λήξη του αγώνα. Αρχικά λήφθηκαν υπόψιν μόνο επιρροές από το χρόνο γύρου, την εκκίνηση των οδηγών και το χρόνο διάρκειας των pit stops χωρίς το VSE να λαμβάνει υπόψιν τυχόν FCY φάσεις. Τα αποτελέσματα δείχνουν ότι παρουσιάζονται αξιοσημείωτες αποκλίσεις στη θέση που τερμάτισαν οι οδηγοί στον αγώνα σε σχέση με τη θέση που πρόβλεψε το VSE. Η παραπάνω διαδικασία επαναλήφθηκε και με την ενεργοποίηση της λειτουργίας πρόβλεψης τυχόν FCY φάσεων. Παρατηρείται ότι με την ενεργοποίηση των FCY φάσεων τα αποτελέσματα που δίνει το VSE είναι ελαφρώς χειρότερα από ότι πριν και αυτό γιατί οι FCY φάσεις ευνοούν τους οδηγούς των χαμηλότερων θέσεων να μειώσουν τη διαφορά που έχουν με τους προπορευόμενους οδηγούς και έτσι να επιχειρήσουν προσπέραση και επομένως αλλαγή θέσης.

Γενικά, φαίνεται πως το VSE συχνά μπορεί να βελτιώσει το αποτέλεσμα ενός αγώνα αρκεί να είναι ταιριαστός ο συνδυασμός των χαρακτηριστικών απόφασης και η παραμετροποίηση των οδηγών.

3.6 Γενικό πλαίσιο μηχανικής μάθησης για πρόβλεψη νικητή αγώνα και κατάταξη πρωταθλήματος στη Formula 1

Στην εργασία του Sicoie (2022) το βασικό ζήτημα που τίθεται είναι κατά πόσο δύνανται οι τεχνικές επιβλεπόμενης μηχανικής μάθησης να προβλέψουν την κατάταξη των οδηγών στο πρωτάθλημα του 2021 χρησιμοποιώντας ιστορικά δεδομένα. Τα δεδομένα που χρησιμοποιήθηκαν αφορούν την περίοδο από 2014 μέχρι τη στιγμή της έρευνας και παρέχουν πληροφορίες σχετικά με τους οδηγούς, τις πίστες, τα αποτελέσματα των αγώνων, τους χρόνους στις κατατακτήριες δοκιμές, τις κατατάξεις των οδηγών και των κατασκευαστών. Ακολούθησε η προπαρασκευή των δεδομένων καθώς και διασταυρούμενη επικύρωση 5-πτυχών (5-Fold Cross Validation) για την ενίσχυση της ακρίβειας των αλγορίθμων. Έγινε,

επίσης, επιλογή των κατάλληλων μεταβλητών προς αποφυγή απώλειας πληροφορίας και κανονικοποίηση των δεδομένων για αποφυγή πλεονασμού.

Τα μοντέλα που ταιριάζουν καλύτερα στη μορφή του dataset που χρησιμοποιείται είναι: παλινδρόμηση με τυχαίο δάσος (Random Forest Regression - RFR), παλινδρόμηση με Gradient Boosting (Gradient Boosting Regression - GBR) και παλινδρόμηση με διάνυσμα υποστήριξης (Support Vector Regression - SVR). Μολονότι η ακρίβεια των αποτελεσμάτων των μεθόδων δεν είναι τόσο απογοητευτική, δεν είναι και ιδιαίτερα υψηλή όπως είχε σχολιαστεί και από τον ερευνητή στην αρχή της εργασίας, αφού είναι εξαιρετικά δύσκολο να προσδιοριστεί ορθά η κατάταξη των οδηγών σε κάθε αγώνα. Χρησιμοποιώντας ως μέτρο συσχέτισης το ρ του Spearman παρατηρείται πως και τα τρία μοντέλα έχουν δώσει υψηλή συσχέτιση με τις πραγματικές κατατάξεις των οδηγών (RFR ρ=0.902, GBR ρ=0.903, SVR ρ=0.883). Παρακάτω παρατίθεται ο πίνακας στον οποίο περιέχονται οι μετρήσεις για την αποτελεσματικότητα των μεθόδων όσον αφορά τις προβλέψεις για τους δέκα πρώτους οδηγούς. Ο λόγος για την αναφορά μόνο των δέκα πρώτων οδηγών της κατάταξης είναι ότι μόνο όσοι τερματίζουν στην πρώτη δεκάδα βαθμολογούνται. Βλέπουμε από τον πίνακα ότι αν είναι επιτρεπτή απόκλιση τουλάχιστον δύο θέσεων τότε και τα τρία μοντέλα αποδίδουν καλύτερα σε σχέση με την απόκλιση μίας θέσης ενώ για απόκλιση μεγαλύτερη των δύο θέσεων τα αποτελέσματα βελτιώνονται στο 64% περίπου και για τις τρεις μεθόδους.

MODEL	TOP 10 ± 1		TOP 10 ± 2		TOP 10 ± 3	
	True	False	True	False	True	False
RFR	0.38	0.61	0.55	0.44	0.63	0.36
GBR	0.35	0.64	0.53	0.46	0.64	0.35
SVR	0.394	0.60	0.54	0.45	0.64	0.35

Συνοψίζοντας, υπήρξε δυσκολία στην παραπάνω εργασία όσον αφορά τη διαθεσιμότητα των δεδομένων και τον χειρισμό τους ώστε να μην χαθεί πληροφορία απαραίτητη για την εκπαίδευση των μοντέλων. Επιπλέον, λόγω της κυριαρχίας της Mercedes την περίοδο που επιλέχθηκαν τα δεδομένα (2019-2021) τα μοντέλα τείνουν να μεροληπτούν προβλέποντας ως νικητές τους οδηγούς της συγκεκριμένης ομάδας. Παρ’ όλα αυτά οι προτεινόμενες μέθοδοι φαίνεται να προβλέπουν την κατάταξη του πρωταθλήματος της F1 για το 2021 σε ικανοποιητικό βαθμό.

ΚΕΦΑΛΑΙΟ 4ο

Μηχανική Μάθηση

4.1 Εισαγωγή

Ο όρος Μηχανική Μάθηση (Machine Learning, ML) δομείται για πρώτη φορά τη δεκαετία του 1950 παράλληλα με τον όρο Τεχνητή Νοημοσύνη (Artificial Intelligence, AI) από πρωτοπόρους επιστήμονες της εποχής όπως οι Alan Turing, Arthur Samuel και John McCarthy οι οποίοι θεωρούνται και «πατέρες» της τεχνητής νοημοσύνης. Η μηχανική μάθηση αποτελεί υποπεδίο της επιστήμης των υπολογιστών και υποσύνολο της τεχνητής νοημοσύνης. Διερευνά τη μελέτη και κατασκευή αλγορίθμων που μπορούν να μαθαίνουν από τα δεδομένα και να κάνουν προβλέψεις σχετικά με αυτά.

Το πρώτο νευρωνικό δίκτυο, που κατασκευάστηκε το 1950, ονομάζεται Τεστ Τούρινγκ (Turing Test) με πρωτότυπη ονομασία «Παιχνίδι Μίμησης» (Imitation Game) και ουσιαστικά αποτελεί μία δοκιμή της μηχανής να επιδείξει ευφυΐα πανομοιότυπη με του ανθρώπου. Ο Arthur Samuel, το 1959, όρισε τη μηχανική μάθηση ως «πεδίο μελέτης που δίνει στους υπολογιστές την ικανότητα να μαθαίνουν από τα δεδομένα χωρίς να έχουν ρητά προγραμματιστεί». Επιπλέον, είναι απόλυτα συνδεδεμένη με την επιστήμη της εφαρμοσμένης (υπολογιστικής) στατιστικής, κλάδο που επίσης στοχεύει στην πρόβλεψη και λήψη αποφάσεων με επεξεργασία δεδομένων μέσω της χρήσης υπολογιστών.

Πλέον, λόγω του καταιγισμού δεδομένων σε οποιονδήποτε τομέα, η χρήση της μηχανικής μάθησης θεωρείται καθοριστική και αναγκαία για την ταχεία επεξεργασία των δεδομένων αυτών αλλά και τη λήψη κρίσιμων αποφάσεων στο λιγότερο δυνατό χρόνο με το ελάχιστο δυνατό κόστος.

Οι τεχνολογίες της μηχανικής μάθησης ταξινομούνται σε τέσσερα κύρια μοντέλα ανάλογα με τη φύση των δεδομένων και το επιθυμητό αποτέλεσμα. Συνεπώς, έχουμε την Εποπτευόμενη Μηχανική Μάθηση (Supervised Machine Learning), την Μη Εποπτευόμενη Μηχανική Μάθηση (Unsupervised Machine Learning), την Ημι-εποπτευόμενη Μηχανική Μάθηση (Semi-supervised Machine Learning) και, τέλος, την Ενισχυτική Μάθηση (Reinforcement Learning).

4.2 Κατηγορίες Μηχανικής Μάθησης

4.2.1 Εποπτευόμενη Μάθηση (Supervised Learning)

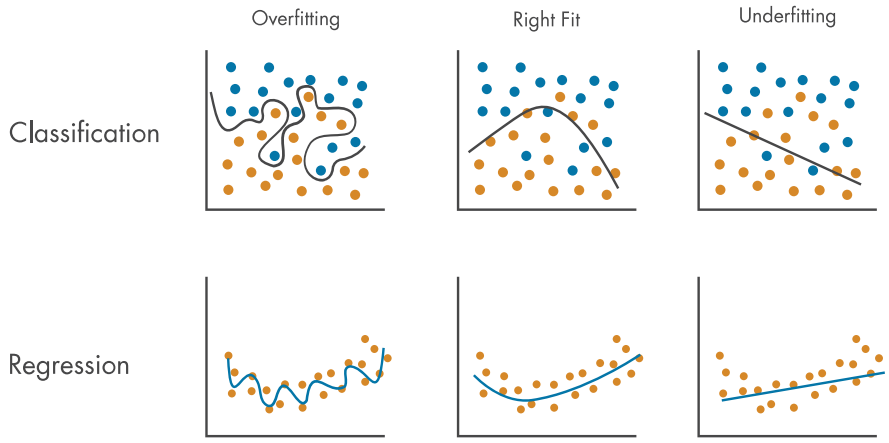
Η εποπτευόμενη μάθηση είναι μία μέθοδος της μηχανικής μάθησης κατά την οποία είναι απαραίτητη η χρήση χαρακτηρισμένων δεδομένων για την εκπαίδευση των αλγορίθμων με στόχο την αναγνώριση μοτίβων και τη πρόβλεψη αποτελεσμάτων. Τα δεδομένα καλούνται χαρακτηρισμένα ή δεδομένα με ετικέτα (labeled data) και το σύνολο δεδομένων που χρησιμοποιείται για την εκπαίδευση του εκάστοτε αλγορίθμου καλείται σύνολο δεδομένων εκπαίδευσης (training dataset). Το όνομα της μεθόδου παραπέμπει στη σχέση καθηγητή-

μαθητή ή γενικότερα στην ύπαρξη ενός επόπτη που καθοδηγεί τον αλγόριθμό βάσει του συνόλου των χαρακτηρισμένων δεδομένων εκπαίδευσης. Οι αλγόριθμοι εποπτευόμενης μάθησης θα αναλύσουν αυτό το σύνολο δεδομένων και θα δημιουργήσουν ένα μοντέλο που έπειτα θα μπορεί να χαρακτηρίσει (να βάλει ετικέτα) σε νέα παραδείγματα. Τα νέα παραδείγματα καλούνται σύνολο δεδομένων ελέγχου (test dataset) τα οποία δεν είναι αρχικώς χαρακτηρισμένα. Σκοπός είναι το μοντέλο να προβλέπει επιτυχώς τις ετικέτες του συνόλου δεδομένων ελέγχου.

Η κατάταξη ενός δανειολήπτη σε φερέγγυο ή αφερέγγυο, η τιμή της μετοχής στη χρηματιστηριακή αγορά αύριο καθώς και η επιλογή αποτελεσματικής στρατηγικής pit stop είναι μερικά από τα προβλήματα που καλούνται να επιλύσουν οι αλγόριθμοι μάθησης με επίβλεψη. Τα παραπάνω προβλήματα χωρίζονται σε δύο βασικές κατηγορίες ως εξής:

- Στα προβλήματα κατηγοριοποίησης χρησιμοποιούνται αλγόριθμοι με στόχο την ταυτοποίηση συγκεκριμένων κατηγοριών για νέες παρατηρήσεις έχοντας ως βάση μία ή περισσότερες ανεξάρτητες μεταβλητές. Πιο συνηθισμένοι αλγόριθμοι κατηγοριοποίησης είναι οι γραμμικοί ταξινομητές (linear classifiers) για δίτιμη μεταβλητή απόκρισης όπως η λογιστική παλινδρόμηση (logistic regression), ο ταξινομητής naïve Bayes και γραμμικός ταξινομητής διανυσμάτων υποστήριξης (SVC) και για πολυδιάστατη μεταβλητή απόκρισης όπως οι μηχανές διανυσμάτων υποστήριξης (SVM). Επιπλέον, μη γραμμική μέθοδος κατηγοριοποίησης είναι η μέθοδος k κοντινότερων γειτόνων (kNN).

- Στα προβλήματα παλινδρόμησης συνεισφέρουν στην πρόβλεψη αποτελεσμάτων για συνεχείς τυχαίες εξαρτημένες μεταβλητές. Ακόμη σκοπεύουν στην χαρτογράφηση μιας προγνωστικής σχέσης μεταξύ της εξαρτημένης και των ανεξάρτητων μεταβλητών. Πιο διαδεδομένη μέθοδος παλινδρόμησης είναι η γραμμική παλινδρόμηση.



Σχήμα 4.1: Διαγραμματική απεικόνιση των μεθόδων κατηγοριοποίησης και παλινδρόμησης (Πηγή: https://www.mathworks.com/discovery/overfitting.html)

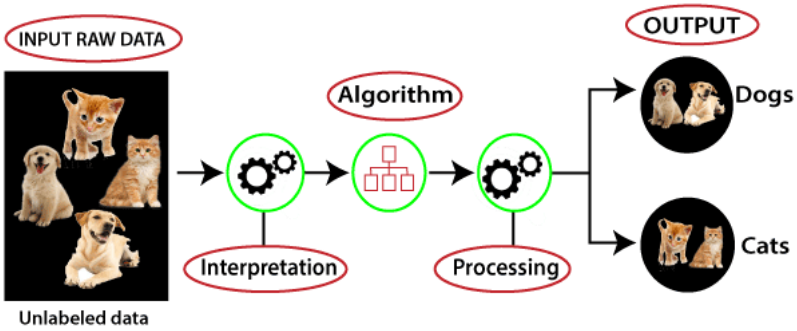
4.2.2 Μη Εποπτευόμενη Μάθηση (Unsupervised Learning)

Η μη εποπτευόμενη μάθηση αποτελεί μέθοδο της μηχανικής μάθησης κατά την οποία δεν πραγματοποιείται επίβλεψη στο μοντέλο με χρήση συνόλου δεδομένων εκπαίδευσης. Αντιθέτως χρησιμοποιούνται αλγόριθμοι ώστε το μοντέλο μόνο του να βρίσκει κρυμμένα μοτίβα και γνώση από τα δοθέντα δεδομένα. Θα μπορούσε να συγκριθεί με τη διαδικασία που ακολουθεί ο ανθρώπινος εγκέφαλος όταν το άτομο εκπαιδεύεται σε νέες δραστηριότητες.

Ουσιαστικά τα μοντέλα εκπαιδεύονται χρησιμοποιώντας μη χαρακτηρισμένα, χωρίς ετικέτα (unlabelled) δεδομένα και μπορούν να δρουν πάνω σε αυτά τα δεδομένα χωρίς την παρουσία επόπτη. Χρησιμοποιείται όταν δεν είναι γνωστό τι αναμένεται να βρεθεί από τα δεδομένα.

Η μη εποπτευόμενη μάθηση εφαρμόζεται στους κάτωθι τύπους προβλημάτων:

- Σε παραγωγή κανόνων συσχέτισης (association rules analysis) η οποία αποτελεί μία εκ των σημαντικότερων και πιο σύγχρονων τεχνικών εξόρυξης γνώσης από μεγάλα σύνολα δεδομένων. Πρόκειται για εύρεση ενδιαφερόντων συσχετίσεων μεταξύ των μεταβλητών μεγάλων βάσεων δεδομένων. Οι κανόνες συσχέτισης διακρίνονται σε τέσσερις τύπους. Υπάρχουν οι Boolean κανόνες συσχέτισης όπου αναζητούνται κανόνες σχετικοί με την ύπαρξη ή απουσία συσχέτισης και προκύπτουν συνήθως κατά την ανάλυση καταναλωτικών προτύπων (market basket analysis). Έπειτα υπάρχουν οι ποσοτικοί (quantitative) κανόνες συσχέτισης που περιγράφουν συσχετίσεις μεταξύ ποσοτικών αντικειμένων ή ιδιοτήτων, δηλαδή δεν έχουν μόνο μία τιμή όπως η ηλικία η οποία μπορεί να ομαδοποιηθεί σε διαστήματα. Ακολουθούν οι κανόνες μονής ή πολλαπλής διάστασης (single or multidimensional) και τέλος οι κανόνες συσχέτισης διαστημάτων (interval association rules) οι οποίοι αφορούν σε διαστήματα τιμών αντί για διακριτά αντικείμενα. Χρησιμοποιείται συχνότερα ο αλγόριθμος apriori, ωστόσο υπάρχουν και δευτερεύοντες αλγόριθμοι όπως οι partition, FP-Growth και Eclat.
- Σε προβλήματα ομαδοποίησης (cluster analysis / clustering) κατά τα οποία γίνεται ταυτοποίηση και ομαδοποίηση δεδομένων από μεγάλα σύνολα δεδομένων που μοιάζουν μεταξύ τους χωρίς κάποιο προβληματισμό για το αποτέλεσμα. Μερικοί από τους δημοφιλέστερους αλγόριθμους είναι ο αλγόριθμος ομαδοποίησης k μέσων (k-means) και ο αλγόριθμος χωρικής ομαδοποίησης εφαρμογών με θόρυβο με βάση την πυκνότητα (DBSCAN).
- Σε προβλήματα μείωσης διαστατικότητας (Dimensionality Reduction) κατά τα οποία χρησιμοποιούνται τεχνικές με στόχο την απομάκρυνση περιττών χαρακτηριστικών (διαστάσεων/μεταβλητών) και τη μείωση του θορύβου. Ο αριθμός των αρχικών μεταβλητών του συνόλου δεδομένων μειώνεται σε ένα διαχειρίσιμο μέγεθος ώστε να γίνουν οι αλγόριθμοι μηχανικής μάθησης περισσότερο αποτελεσματικοί χωρίς ωστόσο να χάνεται μέρος της πληροφορίας που αυτό παρέχει. Αποτελεί μέρος του σταδίου προεπεξεργασίας των δεδομένων με διάφορες τεχνικές όπως η ανάλυση κύριων συνιστωσών (principal component analysis) και η προς τα πίσω μείωση χαρακτηριστικών (backward feature elimination). Ανάλογα με τη φύση των δεδομένων και των αλγορίθμων που πρόκειται να χρησιμοποιηθούν σε μετέπειτα στάδιο επιλέγεται και η κατάλληλη μέθοδος.



Σχήμα 4.2: Σχηματική απεικόνιση επεξήγησης μη εποπτευόμενης μάθησης
(Πηγή: https://www.javatpoint.com/unsupervised-machine-learning)

4.2.3 Ημι-εποπτευόμενη Μάθηση (semi-supervised learning)

Η μάθηση με ημι-επίβλεψη αποτελεί τμήμα της μηχανικής μάθησης που συνδυάζει εποπτευόμενη και μη εποπτευόμενη μάθηση χρησιμοποιώντας ταυτόχρονα χαρακτηρισμένα και μη χαρακτηρισμένα δεδομένα. Ουσιαστικά αποτελεί μια υβριδική τεχνική μεταξύ εποπτευόμενης και μη εποπτευόμενης μάθησης κατά την οποία χρησιμοποιείται ένα μικρό σύνολο επισημασμένων δεδομένων που θα διανθίσει ένα μεγαλύτερο σύνολο μη επισημασμένων δεδομένων.

Από τις πιο απλές εφαρμογές της ημι-εποπτευόμενης μάθησης είναι η αυτό-εκπαίδευση κατά την οποία μπορεί να χρησιμοποιηθεί οποιαδήποτε μέθοδος εποπτευόμενης μάθησης για κατηγοριοποίηση και παλινδρόμηση και να δομηθεί με τέτοιο τρόπο ώστε να εφαρμοστεί με χρήση ημι-εποπτευόμενης μάθησης αξιοποιώντας χαρακτηρισμένα και μη δεδομένα.

Σύνηθες παράδειγμα στο οποία εφαρμόζεται η ημι-εποπτευόμενη μάθηση είναι η κατασκευή ταξινομητή κειμένου. Η μέθοδος είναι αποτελεσματική διότι είναι εξαιρετικά δύσκολη και χρονοβόρα διαδικασία για έναν άνθρωπο να υπομνηματίζει και να ταξινομεί έγγραφα κείμενα. Η ημι-επίβλεψη λύνει αυτό το πρόβλημα αφού ο αλγόριθμος θα εκπαιδευτεί από το μικρό σύνολο επισημασμένων δεδομένων, δηλαδή κάποια κείμενα στα οποία έχει γίνει υπομνημάτιση και κατόπιν κατηγοριοποίηση, και έπειτα θα ταξινομήσει τα μη χαρακτηρισμένα δεδομένα. Αντίστοιχη διαδικασία μπορεί να επιτευχθεί και για την κατηγοριοποίηση εικόνων.

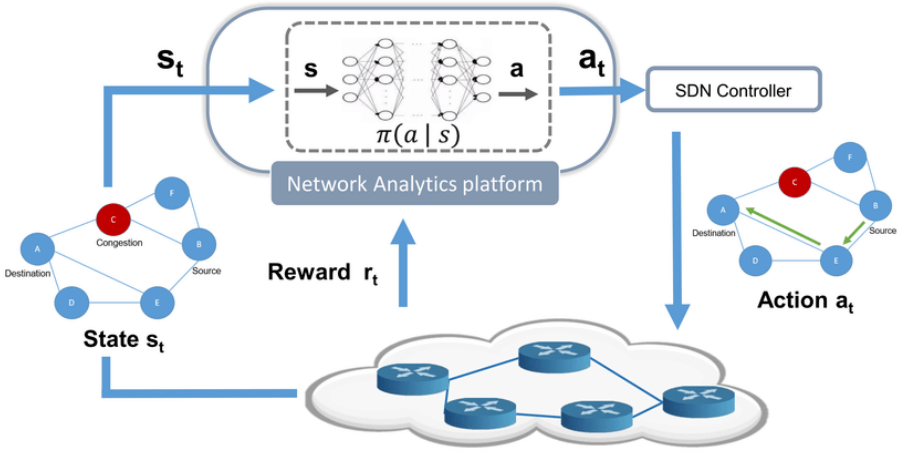
4.2.4 Ενισχυτική Μάθηση (Reinforcement Learning)

Η ενισχυτική μάθηση είναι τεχνική της μηχανικής μάθησης που εκπαιδεύει το λογισμικό (software) ενός υπολογιστικού συστήματος να λαμβάνει αποφάσεις για την επίτευξη του βέλτιστου αποτελέσματος μιμούμενο τη διαδικασία εκμάθησης δοκιμής και σφάλματος (trial and error learning process). Σαν όρος αποδίδεται σε σύνολο μεθόδων στις οποίες το σύστημα εκπαιδεύεται μέσα από την αλληλεπίδραση με το περιβάλλον. Οι παράγοντες ενισχυτικής μάθησης θεωρούνται από τους πιο εξειδικευμένους και ικανούς να επιδείξουν υψηλού επιπέδου νοημοσύνη και ορθολογική συμπεριφορά.

Στην περίπτωση αυτή το σύστημα δεν καθοδηγείται από κάποιον επόπτη για την ενέργεια που πρόκειται να ακολουθήσει. Αντιθέτως, ανακαλύπτει μόνο του ποιες ενέργειες αποφέρουν μεγαλύτερο κέρδος. Αναλυτικότερα, ο ενισχυτικός παράγοντας αλληλοεπιδρά με το

περιβάλλον ώστε να εκπαιδευτεί για να επιχειρεί την καλύτερη δυνατή δράση (action, a_t) υπό τη δεδομένη κατάσταση (state, S_t) που το περιβάλλον βρίσκεται στο δεδομένο βήμα t. Το περιβάλλον από μόνο του μπορεί να επιδείξει πολλαπλές καταστάσεις και ο παράγοντας επιδρά πάνω στο περιβάλλον για να αλλάξει αυτές τις καταστάσεις. Επομένως, η δράση του παράγοντα αλλάζει την κατάσταση του περιβάλλοντος από S_t σε S_{t+1} και παράγει μια επιβράβευση (reward, r_t) για τον παράγοντα. Έπειτα ο παράγοντας λαμβάνει την καλύτερη δυνατή δράση για τη νέα κατάσταση S_{t+1} και κατ' αυτόν τον τρόπο δίνει νέα επιβράβευση r_{t+1} και ούτω καθεξής. Μετά από αρκετές επαναλήψεις ο παράγοντας προσπαθεί να βελτιώσει τις αποφάσεις του σχετικά με το ποια είναι η καλύτερη δράση στη δεδομένη κατάσταση του περιβάλλοντος χρησιμοποιώντας τις επιβραβεύσεις που λαμβάνει κατά τη διάρκεια της εκπαίδευσης.

Ο ρόλος του περιβάλλοντος ουσιαστικά είναι να παρουσιάσει στον ενισχυτικό παράγοντα ποικίλες πιθανές καταστάσεις που μπορεί να προκύψουν σε κάποιο πρόβλημα και που πιθανόν να χρειαστεί η δράση του παράγοντα. Για την διευκόλυνση της διαδικασίας εκπαίδευσης το περιβάλλον δίνει επιβράβευση ή ποινή (αρνητική επιβράβευση) αντίστοιχη με τη δράση κατά την οποία λήφθηκαν αποφάσεις από τον παράγοντα σε δεδομένη κατάσταση. Δηλαδή, η επιβράβευση αποτελεί συνάρτηση τόσο της δράσης όσο και της κατάστασης και επομένως η ίδια δράση δύναται (και ιδανικά θα έπρεπε) να λαμβάνει διαφορετικές επιβραβεύσεις υπό διαφορετικές καταστάσεις.



Σχήμα 4.3: Σχηματική απεικόνιση διαδικασίας ενισχυτικής μάθησης
(Πηγή: https://www.researchgate.net/figure/The-reinforcement-learning-process_fig3_326611368)

Η ενισχυτική μάθηση εφαρμόζεται σε πολλούς τομείς όπως ρομποτική και αυτοματισμοί, χρηματοοικονομικά και συναλλαγές, υγεία, διαχείριση ενέργειας, marketing και συστήματα προώθησης και τέλος μη επανδρωμένα οχήματα που είναι και η πιο σχετική με την παρούσα εργασία εφαρμογή.

Τα μη επανδρωμένα οχήματα αποτελούν μία επαναστατική εφαρμογή της τεχνητής νοημοσύνης σε συνδυασμό με τη ρομποτική στην αυτοκινητοβιομηχανία. Τα μη επανδρωμένα οχήματα έχουν την ικανότητα να προσανατολίζονται και να επιχειρούν αυτόνομα χωρίς ανθρώπινη επίβλεψη. Είναι εξοπλισμένα με αισθητήρες, κάμερες, τεχνική LIDAR (βασίζεται στην εκπομπή παλμικής ακτινοβολίας λέιζερ στην ατμόσφαιρα και κατ'

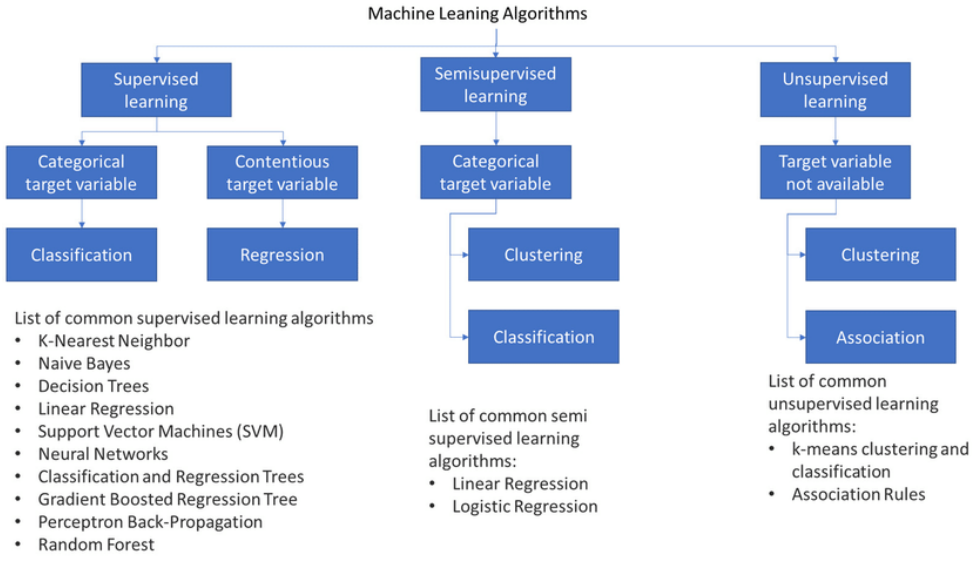
επέκταση στη καταγραφή της οπισθοσκεδαζόμενης ακτινοβολίας λέιζερ ώστε να γίνει μέτρηση των αποστάσεων) και ραντάρ. Όλα τα παραπάνω παράγουν δεδομένα τα οποία επεξεργάζονται με κατάλληλους αλγορίθμους ενισχυτικής αλλά και μηχανικής μάθησης ώστε να παρθούν αποφάσεις σε πραγματικό χρόνο και να ελεγχθούν οι κινήσεις των αυτοκινήτων.

Τα αυτοκίνητα-ρομπότ εκπαιδεύονται ώστε να οδηγούν μέσα σε πίστα λαμβάνοντας επιβράβευση για θεμιτές δράσεις όπως η διατήρηση υψηλής ταχύτητας σε ευθείες ή σωστή πλοήγηση στις στροφές. Λαμβάνουν, επίσης, αρνητική επιβράβευση για λανθασμένες δράσεις όπως κρούσεις ή παρέκκλιση εκτός πίστας.

4.3 Αλγόριθμοι Μηχανικής Μάθησης

Η μηχανική μάθηση περιλαμβάνει ένα ευρύ φάσμα αλγορίθμων οι οποίοι κατασκευάστηκαν για να εξυπηρετούν κάθε είδος προβλήματος αλλά και να προσαρμόζονται σε κάθε περίπτωση στα διαφορετικής φύσης δεδομένα. Βάσει των ερωτημάτων που προκύπτουν και των διαθέσιμων δεδομένων επιλέγεται κάθε φορά το κατάλληλο είδος μάθησης όπως αυτά αναφέρθηκαν παραπάνω καθώς και ο κατάλληλος αλγόριθμος. Επομένως, η κατανόηση και μελέτη των βασικών εννοιών κάθε αλγορίθμου της εποπτευόμενης και της μη εποπτευόμενης μηχανικής μάθησης αποτελεί ζήτημα μείζονος σημασίας ώστε να εξάγονται χρήσιμα και έγκυρα αποτελέσματα.

Στις κάτωθι τρεις ενότητες θα αναφερθούν τα βασικά χαρακτηριστικά των αλγορίθμων διαχωρίζοντάς τους σε αλγορίθμους εποπτευόμενης και μη εποπτευόμενης μάθησης. Η ενότητα 4.3.1 αφορά στις μεθόδους εποπτευόμενης μάθησης που χρησιμοποιούνται σε προβλήματα κατηγοριοποίησης, η 4.3.2 σε μεθόδους εποπτευόμενης μάθησης που χρησιμοποιούνται σε προβλήματα παλινδρόμησης και τέλος η παράγραφος 4.3.3 αφορά σε μεθόδους μη εποπτευόμενης μάθησης.



Σχήμα 4.4: Διαγραμματική απεικόνιση αλγορίθμων εποπτευόμενης και μη εποπτευόμενης μάθησης (Πηγή: https://www.researchgate.net/figure/Machine-learning-algorithms-classification_fig1_349860057)

4.3.1 Μέθοδοι κατηγοριοποίησης (Classification methods)

Η κατηγοριοποίηση αποτελεί μέθοδο της επιβλεπόμενης μηχανικής μάθησης στην οποία το εκάστοτε μοντέλο προσπαθεί να προβλέψει τη σωστή κατηγορία (ή κλάση, ετικέτα) για τα δοθέντα δεδομένα. Είναι στην ουσία μία διαδικασία αναγνώρισης, κατανόησης και ομαδοποίησης ιδεών και αντικειμένων σε προκαθορισμένες κατηγορίες ή υποπληθυσμούς. Οι αλγόριθμοι κατηγοριοποίησης (ταξινομητές - classifiers) χρησιμοποιούν χαρακτηρισμένα δεδομένα (labeled data) και όντας πλήρως εκπαιδευμένοι από το σύνολο δεδομένων εκπαίδευσης, αξιολογούνται στα δεδομένα ελέγχου και έπειτα χρησιμοποιούνται σε νέα άγνωστα μη χαρακτηρισμένα δεδομένα (unlabeled data) ώστε να τα ταξινομήσουν σε μία ή περισσότερες προκαθορισμένες κλάσεις. Η διαδικασία αυτή επιτυγχάνεται με την αναγνώριση κοινών μοτίβων στα δεδομένα εκπαίδευσης όπως παρόμοιες λέξεις ή γνώμη–αίσθημα καταναλωτή-θεατή, ακολουθίες αριθμών, σχήματα και γραφικά εικόνων, κοινά περιγραφικά στατιστικά όπως μέσος, διάμεσος, διακύμανση και κοινά γεωμετρικά χαρακτηριστικά όπως απόσταση στοιχείων μεταξύ τους.

Είναι αντιληπτό ότι οι μέθοδοι κατηγοριοποίησης είναι από τα πιο σημαντικά και χρήσιμα εργαλεία που χρησιμοποιούνται για την επίλυση διαφόρων προβλημάτων σε πολλούς τομείς όπως η ιατρική, τα χρηματοοικονομικά και ο μηχανοκίνητος αθλητισμός.

Στην παρούσα εργασία θα αναλυθούν οι παρακάτω μέθοδοι κατηγοριοποίησης:

- Λογιστική Παλινδρόμηση (Logistic Regression)
- Μηχανές Διανυσμάτων Υποστήριξης (Support Vector Machines)
- Δένδρα Απόφασης (Decision Trees)
- Τυχαία Δάση (Random Forests)
- Κ Κοντινότεροι Γείτονες (K Nearest Neighbours, KNN)
- Extreme Gradient Boosting (XGBoost)
- Γραμμική Διακριτική Ανάλυση (Linear Discriminant Analysis)

4.3.1.1 Λογιστική Παλινδρόμηση (Logistic Regression)

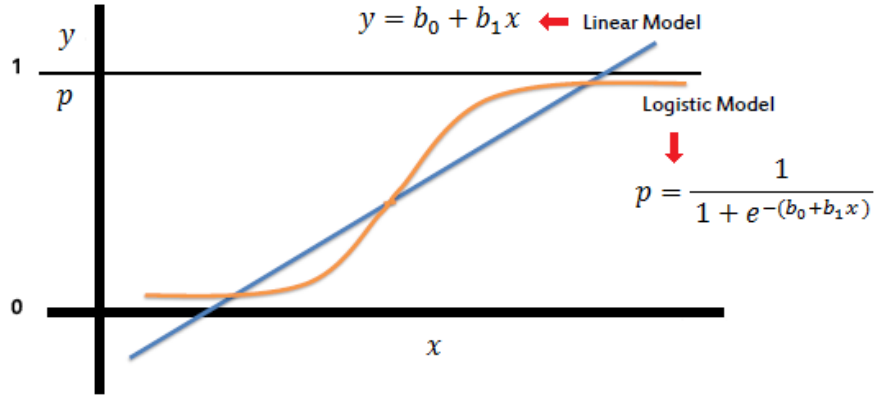
Η λογιστική παλινδρόμηση (Logistic Regression ή logit model) είναι μία από τις σημαντικότερες στατιστικές μεθόδους που χρησιμοποιείται κυρίως για ταξινόμηση και πρόβλεψη αποτελεσμάτων. Έχει, επίσης, στενή σχέση με τα νευρωνικά δίκτυα. Εκτιμά το πόσο πιθανό είναι να προκύψει ένα συμβάν βασιζόμενο στο δοθέν σύνολο δεδομένων των ανεξάρτητων μεταβλητών. Πιο συγκεκριμένα, αναλύει τη σχέση μεταξύ μιας κατηγορικής εξαρτημένης μεταβλητής και μιας ή περισσότερων ανεξάρτητων μεταβλητών.

Η εξαρτημένη μεταβλητή μπορεί να έχει δύο κλάσεις-κατηγορίες, να είναι δηλαδή δυαδική, οι οποίες κωδικοποιούνται συνήθως με τους αριθμούς 0 και 1. Τότε η ταξινόμηση θα είναι δυαδική (binary classification). Η δυαδική ταξινόμηση είναι ειδική περίπτωση της λογιστικής παλινδρόμησης και έχει στόχο την πρόβλεψη της πιθανότητας εμφάνισης κάποιου γεγονότος ή της εμφάνισης θετικού ή αρνητικού αποτελέσματος. Προβλήματα δυαδικής ταξινόμησης είναι η ταξινόμηση ενός δανειολήπτη σε φερέγγυο ή αφερέγγυο, η πρόβλεψη για τον αν το άτομο πάσχει από κάποια ασθένεια ή όχι.

Η πολυωνυμική λογιστική παλινδρόμηση χρησιμοποιείται για την πρόβλεψη της πιθανότητας ενός εκ των τριών τουλάχιστον πιθανών αποτελεσμάτων όπως για παράδειγμα ποιον τύπο προϊόντος θα αγοράσει ο πελάτης ή το πολιτικό κόμμα που υποστηρίζει κάποιος σε μια δημοσκόπηση.

Τέλος, το διατεταγμένο μοντέλο λογιστικής παλινδρόμησης (ordinal logistic regression) χρησιμοποιείται για την πρόβλεψη της πιθανότητας ενός αποτελέσματος που έχει

προκαθορισμένη διάταξη όπως το επίπεδο ευχαρίστησης ενός καταναλωτή ή η θέση τερματισμού ενός οδηγού αγώνων.



Σχήμα 4.4: Διαγραμματική απεικόνιση της διαφοράς Λογιστικής και Γραμμικής Παλινδρόμησης
(Πηγή: https://www.saedsayad.com/logistic_regression.htm)

Το μοντέλο της λογιστικής παλινδρόμησης αποτελεί μια τροποποίηση της γραμμικής παλινδρόμησης στην οποία οποιαδήποτε είσοδος πραγματικής τιμής αντιστοιχίζεται σε μία πιθανότητα μεταξύ 0 και 1. Ουσιαστικά γίνεται μετασχηματισμός του γραμμικού συνδυασμού των ανεξάρτητων μεταβλητών με τη χρήση της σιγμοειδούς (ή λογιστικής) συνάρτησης (sigmoid function). Η σιγμοειδής συνάρτηση ορίζεται ως εξής:

$$S(u) = \frac{1}{1 + e^{-u}} = \frac{e^u}{e^u + 1}$$

όπου S(x) είναι η πιθανότητα να συμβεί το γεγονός, u είναι ο γραμμικός συνδυασμός των ανεξάρτητων μεταβλητών και e η βάση του φυσικού λογαρίθμου.

Η παραπάνω συνάρτηση βοηθάει το λογιστικό μοντέλο να στριμώξει τις τιμές από το διάστημα (-k,k) που δίνει η γραμμική παλινδρόμηση στο (0,1).

Ο γραμμικός συνδυασμός των ανεξάρτητων μεταβλητών δίνεται παρακάτω ως εξής:

$$u = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n$$

όπου β_0 είναι η τεταγμένη και β_1, β_2, …., β_n είναι οι συντελεστές των ανεξάρτητων μεταβλητών x_1, x_2, …, x_n αντίστοιχα.

Για την εκτίμηση των συντελεστών του μοντέλου λογιστικής παλινδρόμησης χρησιμοποιείται η μέθοδος μέγιστης πιθανοφάνειας έχοντας ως στόχο την εύρεση των τιμών που μεγιστοποιούν τη συνάρτηση πιθανοφάνειας. Κατόπιν το μοντέλο μπορεί να χρησιμοποιηθεί για την ταξινόμηση νέων δεδομένων ή την πρόβλεψη εμφάνισης νέου συμβάντος.

Για τη λήψη αποφάσεων ορίζεται ένα κατώφλι (threshold) για το θετικό αποτέλεσμα. Αν δηλαδή η προβλεπόμενη πιθανότητα είναι μεγαλύτερη ή ίση με την προκαθορισμένη τιμή

(συνήθης επιλεχθείσα τιμή 0.5) τότε προβλέπεται ότι θα συμβεί το γεγονός. Αντίθετα, προβλέπεται ότι το γεγονός δε θα συμβεί.

Το μοντέλο της λογιστικής παλινδρόμησης θεωρείται από τα βασικότερα λόγω της απλότητας και της ευκολίας του να εκπαιδευτεί από το σύνολο των δεδομένων εκπαίδευσης καθώς και της ικανότητάς του να διαχειρίζεται σύνολα δεδομένων με κατηγορικές και αριθμητικές ανεξάρτητες μεταβλητές. Παρ' όλα αυτά διαθέτει κάποια μειονεκτήματα όπως η παραδοχή της γραμμικής σχέσης των ανεξάρτητων μεταβλητών με τον λογαριθμικό μετασχηματισμό της εξαρτημένης μεταβλητής, η απαίτηση μεγάλων συνόλων δεδομένων, η επιρρέπεια στην υπερπροσαρμογή του μοντέλου στα δεδομένα (overfitting) και η ευαισθησία στην ύπαρξη ακραίων τιμών.

4.3.1.2 Μηχανές Διανυσμάτων Υποστήριξης (Support Vector Machines, SVMs)

Οι μηχανές διανυσμάτων υποστήριξης (Support Vector Machines, SVMs) είναι ένας αλγόριθμος που αναπτύχθηκε το 1990 και ανήκει στην κατηγορία την εποπτευόμενης μηχανικής μάθησης. Κατηγοριοποιεί τα δεδομένα κάνοντας διαχωρισμό μεταξύ κλάσεων βρίσκοντας το βέλτιστο υπερεπίπεδο το οποίο μεγιστοποιεί το περιθώριο μεταξύ των κοντινότερων σημείων των δεδομένων διαφορετικών κλάσεων. Πιο συγκεκριμένα, ταξινομεί τα δεδομένα βρίσκοντας τη βέλτιστη γραμμή ή το υπερεπίπεδο που μεγιστοποιεί την απόσταση μεταξύ κάθε κλάσης σε έναν n-διάστατο χώρο.

Ο αριθμός των διαστάσεων του υπερεπιπέδου, αν δηλαδή θα είναι γραμμή σε δισδιάστατο χώρο ή επίπεδο σε n-διάστατο χώρο, καθορίζεται κάθε φορά από τον αριθμό των χαρακτηριστικών που εισέρχονται στο μοντέλο. Επειδή μπορούν να βρεθούν περισσότερα του ενός υπερεπίπεδα που ταξινομούν τα δεδομένα σε κλάσεις, ο αλγόριθμος μεγιστοποιώντας το περιθώριο μεταξύ των τιμών των δεδομένων βρίσκει την καλύτερη δυνατή απόφαση διαχωρισμού των κλάσεων. Αυτό βοηθάει στη γενίκευση του αλγορίθμου σε νέα άγνωστα δεδομένα ώστε να δοθούν όσο το δυνατόν καλύτερες προβλέψεις. Οι γραμμές (σημεία των δεδομένων) που βρίσκονται πλησιέστερα στο βέλτιστο υπερεπίπεδο καλούνται διανύσματα υποστήριξης.

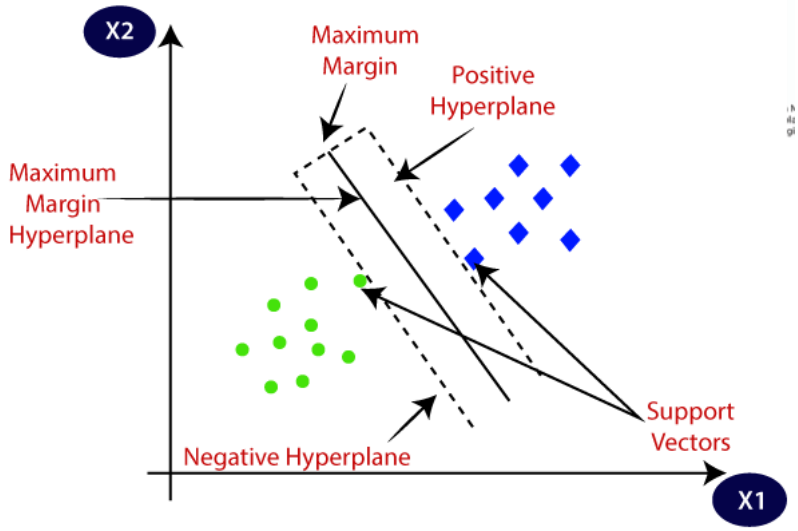
Ο SVM αλγόριθμος είναι εξαιρετικά δημοφιλής καθώς μπορεί να διαχειριστεί τόσο γραμμικά όσο και μη γραμμικά προβλήματα ταξινόμησης. Όταν τα διαθέσιμα δεδομένα δεν είναι γραμμικώς διαχωρίσιμα μετασχηματίζονται, με χρήση συναρτήσεων πυρήνα (kernel functions), σε χώρους χαρακτηριστικών μεγαλύτερων διαστάσεων που επιτρέπουν το γραμμικό διαχωρισμό. Η παραπάνω διαδικασία είναι γνωστή ως τέχνασμα πυρήνα (kernel trick). Οι συναρτήσεις πυρήνα επιτρέπουν στον αλγόριθμο να λειτουργεί σε χώρο υψηλότερης διάστασης χωρίς να υπολογίζει τις συντεταγμένες των δεδομένων αλλά κάνοντας αναπαράσταση των σημείων δεδομένων στον εν λόγω χώρο. Οι συναρτήσεις πυρήνα αποτελούνται από γραμμικούς, πολυωνυμικούς, σιγμοειδής και πυρήνες ακτινωτής βάσης.

Αναλυτικότερα οι συναρτήσεις πυρήνα είναι οι εξής:

- Γραμμική Συνάρτηση Πυρήνα (Linear Kernel Function): Ο γραμμικός πυρήνας είναι η απλούστερη και πιο συνηθισμένη μορφή συνάρτησης πυρήνα. Χρησιμοποιείται όταν τα δεδομένα είναι γραμμικώς διαχωρίσιμα με τον όρο να υποδηλώνει ότι τα σημεία δεδομένων μπορούν να ταξινομηθούν σε δύο τάξεις χρησιμοποιώντας μία ευθεία γραμμή (δισδιάστατο επίπεδο, 2-D).
- Πολυωνυμική Συνάρτηση Πυρήνα (Polynomial Kernel Function): Η πολυωνυμική συνάρτηση πυρήνα μετασχηματίζει τα δεδομένα χρησιμοποιώντας το βαθμωτό γινόμενο ή γινόμενο τελείας (dot product). Έτσι, χρησιμοποιώντας πολυωνυμικές

συναρτήσεις των αρχικών χαρακτηριστικών τα μεταφέρει σε χώρο υψηλότερης διάστασης.

- Συνάρτηση Πυρήνα Ακτινωτής Βάσης (Radial Basis Function, RBF): Η συνάρτηση ακτινωτής βάσης μπορεί να αποδώσει σύνθετες μη γραμμικές σχέσεις των δεδομένων. Παρουσιάζει ομοιότητες με τον KNN αλγόριθμο καθώς μετασχηματίζει τα δεδομένα αποδίδοντάς τα σε υπερεπίπεδο άπειρων διαστάσεων το οποίο θα δώσει μετέπειτα έναν ισχυρό μη γραμμικό ταξινομητή. Έπειτα χρησιμοποιεί τη μέθοδο κοντινότερου γείτονα για την κατηγοριοποίηση. Είναι σημαντική, επίσης η επιλογή σωστής τιμής για την υπερπαράμετρο γάμμα η οποία ελέγχει το πλάτος του πυρήνα. Ανάλογα με την τιμή της υπερπαραμέτρου η ακτινωτή συνάρτηση πυρήνα μπορεί να είναι είτε Gaussian είτε Laplace.
- Σιγμοειδής Συνάρτηση Πυρήνα (Sigmoid Kernel): Αν τα δεδομένα έχουν σιγμοειδή μορφή ή παρουσιάζουν ισχυρή μη γραμμικότητα τότε χρησιμοποιείται η σιγμοειδής συνάρτηση πυρήνα. Μετασχηματίζει και αυτή τα δεδομένα αναπαριστώντας τα σε χώρους υψηλότερης διάστασης με χρήση των σιγμοειδών συναρτήσεων. Βρίσκει μεγαλύτερη εφαρμογή στα νευρωνικά δίκτυα ως συνάρτηση ενεργοποίησης.
- Άλλες Συναρτήσεις Πυρήνα (Other Kernels): Υπάρχουν διάφοροι άλλοι τύποι πυρήνα που δεν εφαρμόζονται συχνά, ωστόσο μπορούν να φανούν χρήσιμοι σε ειδικές περιπτώσεις δεδομένων. Τέτοιοι είναι οι πυρήνες Bessel και ANOVA.



Σχήμα 4.5: Διαγραμματική απεικόνιση του μέγιστου περιθωρίου μεταξύ θετικού και αρνητικού υπερεπιπέδου (Πηγή: https://www.analyticsvidhya.com/blog/2021/10/support-vector-machinessvm-a-complete-guide-for-beginners/)

Ο αλγόριθμος SVM έχει αρκετά πλεονεκτήματα όπως η αποτελεσματικότητά του σε χώρους υψηλών διαστάσεων, η εύκολη χρήση τόσο σε γραμμικά όσο και σε μη γραμμικώς διαχωρίσιμα δεδομένα και η ερμηνευσιμότητα και κατανόηση των αποτελεσμάτων. Επιπλέον, ο SVM αλγόριθμος δεν εμφανίζει τόσο μεγάλη επιρρέπεια στο πρόβλημα υπερπροσαρμογής των δεδομένων. Παρ' όλ' αυτά, εμφανίζει και κάποια αρνητικά χαρακτηριστικά όπως η δυσκολία διαχείρισης μεγάλων συνόλων δεδομένων και η δυσκολία απόδοσης ικανοποιητικών αποτελεσμάτων όταν υπάρχει θόρυβος στα δεδομένα, δηλαδή όταν οι κλάσεις-στόχοι είναι αλληλεπικαλυπτόμενες. Ακόμη, παρουσιάζει ευαισθησία στην επιλογή των κατάλληλων συναρτήσεων πυρήνα καθώς και στην επιλογή κατάλληλων τιμών για τις

υπερπαραμέτρους. Τέλος, η χρήση του SVM αλγορίθμου μπορεί να φανεί υπολογιστικά δαπανηρή και χρονοβόρα ειδικά σε περιπτώσεις μεγάλων συνόλων δεδομένων και πολύπλοκων συναρτήσεων πυρήνα.

4.3.1.3 Δένδρα απόφασης (Decision Trees)

Το δένδρο απόφασης είναι ένας μη παραμετρικός αλγόριθμος εποπτευόμενης μηχανικής μάθησης που χρησιμοποιείται σε προβλήματα ταξινόμησης. Είναι ένα δενδροειδές μοντέλο κατηγοριοποίησης ιεραρχικής δομής που αποτελείται από εσωτερικούς κόμβους, διακλαδώσεις και κόμβους φύλλα.

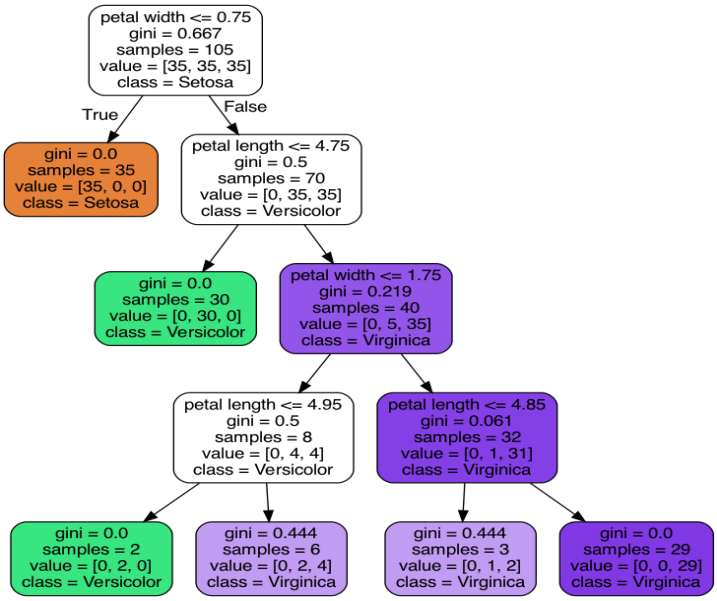
Αναλυτικότερα, το δένδρο απόφασης αποτελείται από έναν αρχικό κόμβο ρίζα (root node) ο οποίος καλείται και κόμβος γονιός (parent node) ενώ όλοι οι υπόλοιποι κόμβοι καλούνται κόμβοι παιδιά (child nodes). Ο κόμβος ρίζα δεν έχει εισερχόμενες διακλαδώσεις και αναπαριστά ολόκληρο το σύνολο δεδομένων το οποίο παρακάτω διαιρείται σε δύο ή περισσότερα ομογενή υποσύνολα. Οι εσωτερικοί κόμβοι χαρακτηρίζονται και ως κόμβοι απόφασης καθώς σε κάθε έναν από αυτούς λαμβάνεται και μία απόφαση με βάση το χαρακτηριστικό και την τιμή που αντιπροσωπεύει ο εκάστοτε κόμβος και κατόπιν σύμφωνα με την αντίστοιχη διακλάδωση οδηγούμαστε στον επόμενο κόμβο και ούτω καθεξής. Οι κόμβοι απόφασης διαθέτουν πολλαπλές διακλαδώσεις-ακμές. Τέλος, οι κόμβοι φύλλα (leaf nodes) ή τερματικοί κόμβοι (terminal nodes) αποδίδουν τα αποτελέσματα του αλγορίθμου δηλαδή των αποφάσεων που πάρθηκαν στους εσωτερικούς κόμβους και δε διαθέτουν περεταίρω διακλαδώσεις.

Η εκπαίδευση δένδρου απόφασης επιστρατεύει μια στρατηγική διαίρεσης διεξάγοντας έρευνα για την εξακρίβωση του βέλτιστου σημείου διαχωρισμού μέσα στο δέντρο. Η διαδικασία διαχωρισμού (splitting) γίνεται με επαναλαμβανόμενο εκ των άνω προς τα κάτω τρόπο μέχρις ότου όλες ή οι περισσότερες από τις εγγραφές του συνόλου δεδομένων να ταξινομηθούν σε συγκεκριμένες κλάσεις. Το πόσο ικανοποιητική θα είναι η κατηγοριοποίηση εξαρτάται από την πολυπλοκότητα του δένδρου. Τα μικρά δένδρα είναι πιο πιθανό να καταλήξουν σε τερματικούς κόμβους με τιμές χαρακτηριστικών ενώ όσο πιο σύνθετο γίνεται το δένδρο τόσο δυσκολότερο γίνεται να διατηρηθεί ξεκάθαρη η κατηγοριοποίηση. Το γεγονός αυτό χαρακτηρίζεται και ως κατακερματισμός δεδομένων (data fragmentation) και συχνά μπορεί να οδηγήσει σε υπερπροσαρμογή των δεδομένων.

Επειδή τα πολύ μεγάλα και σύνθετα δένδρα έχουν αυξημένο κίνδυνο εμφάνισης υπερπροσαρμογής και αντίστοιχα ένα απλούστερο δένδρο είναι πιθανό να χάσει σημαντική πληροφορία, χρησιμοποιείται μια τεχνική που εξυπηρετεί στη μείωση του μεγέθους των δένδρων χωρίς να μειώνει την ακρίβεια των αποτελεσμάτων. Η τεχνική αυτή ονομάζεται pruning («κλάδεμα»).

Το δένδρο απόφασης είναι από τους πιο εύκολους αλγορίθμους όσον αφορά στην κατανόησή του αφού ακολουθεί τον τρόπο σκέψης του ανθρώπινου εγκεφάλου και είναι εξαιρετικά χρήσιμο στην επίλυση προβλημάτων που σχετίζονται με λήψη απόφασης. Επιπλέον, απαιτεί μικρότερη διαδικασία προεπεξεργασίας των δεδομένων σε σχέση με άλλους αλγορίθμους αφού μπορεί αυτόματα να διαχειριστεί το πρόβλημα ελλειπουσών τιμών επιλέγοντας την τιμή με τη μεγαλύτερη συχνότητα ή προβλέποντας κάποια τιμή βάσει άλλων κριτηρίων και μπορεί να διαχειριστεί τόσο αριθμητικά όσο και κατηγορικά δεδομένα.

Στα αρνητικά χαρακτηριστικά του συγκαταλέγονται η πολυπλοκότητά του εξαιτίας των πολλών στρωμάτων από κόμβους απόφασης καθώς και η τάση να παρουσιάζει υπερπροσαρμογή στα δεδομένα. Όσο περισσότερες είναι οι προκαθορισμένες κλάσεις τόσο αυξάνεται και η πολυπλοκότητα του δένδρου.



Σχήμα 4.6: Διαγραμματική απεικόνιση Δένδρου Απόφασης
(Πηγή: https://vitalflux.com/visualize-decision-tree-python-sklearn-library/)

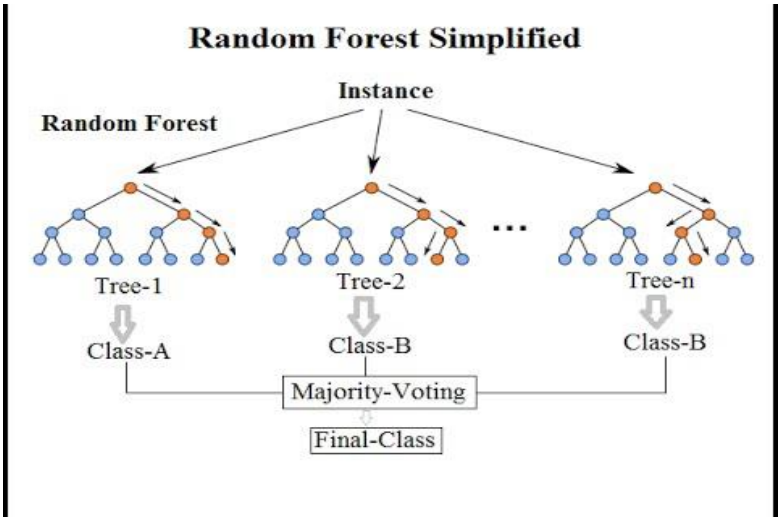
4.3.1.4 Τυχαία Δάση (Random Forests)

Ο ταξινομητής τυχαίου δάσους είναι ένας εύχρηστος και ευέλικτος αλγόριθμος μηχανικής μάθησης με επίβλεψη που συνδυάζει τα αποτελέσματα πολλαπλών δένδρων απόφασης για να οδηγηθεί σε ένα τελικό αποτέλεσμα. Δημιουργήθηκε για να επιλύσει το πρόβλημα που αντιμετώπιζαν τα δένδρα απόφασης ως προς την μεροληψία και την υπερπροσαρμογή.

Στο τυχαίο δάσος κάθε δένδρο απόφασης εκπαιδεύεται σε ένα τυχαίο υποσύνολο δεδομένων εισόδου που λήφθηκε από το αρχικό σύνολο δεδομένων με επανατοποθέτηση (δειγματοληψία bootstrap) και σε κάθε κόμβο του δένδρου μόνο ένα τυχαίο υποσύνολο των χαρακτηριστικών λαμβάνεται υπόψιν χρησιμοποιώντας τη μέθοδο τυχαίου υποχώρου χαρακτηριστικών (random subspace method). Η ύπαρξη της τυχαιότητας τόσο στα χαρακτηριστικά όσο και στο σύνολο δεδομένων εκπαίδευσης συμβάλλει στη μείωση της υπερπροσαρμογής. Όσον αφορά την ακρίβεια των αποτελεσμάτων, αυτή αυξάνεται αν τα μεμονωμένα δένδρα του δάσους είναι μεταξύ τους ασυσχέτιστα.

Τα τυχαία δάση έχουν βασικές υπερπαραμέτρους που χρειάζεται να καθοριστούν πριν την εκπαίδευση του αλγορίθμου όπως είναι ο αριθμός των δένδρων απόφασης που θα έχει το τυχαίο δάσος (number of trees), το μέγεθος του κόμβου που υποδηλώνει των ελάχιστο αριθμό παρατηρήσεων στον τερματικό κόμβο (node size) και τον αριθμό των τερματικών κόμβων (max terminal node) μεταξύ άλλων. Για την πραγματοποίηση πρόβλεψης ο αλγόριθμος συνδυάζει τις προβλέψεις όλων των επιμέρους δένδρων ώστε να οδηγηθεί σε ένα τελικό αποτέλεσμα.

Ο αλγόριθμος τυχαίου δάσους βρίσκει εφαρμογή σε αρκετά επιστημονικά πεδία όπως η ιατρική για την εξακρίβωση συμπτωμάτων ασθενειών, τα χρηματοοικονομικά για την πρόβλεψη του ρίσκου δανεισμού και η τοπογραφία για την αναγνώριση εδαφών με παρόμοια χαρακτηριστικά. Διαθέτει, τελικώς, αρκετά πλεονεκτήματα όπως η μείωση του ρίσκου για υπερπροσαρμογή, η ευελιξία και η ευκολία στο να προσδιορίσει τη σημαντικότητα των χαρακτηριστικών και τέλος η ικανότητα να χειριστεί μεγάλα σύνολα δεδομένων σε πολλές διαστάσεις.



Σχήμα 4.7: Διαγραμματική απεικόνιση Τυχαίου Δάσους
(Πηγή: https://www.nvidia.com/en-us/glossary/random-forest/)

4.3.1.5 Extreme Gradient Boosting (XGBoost)

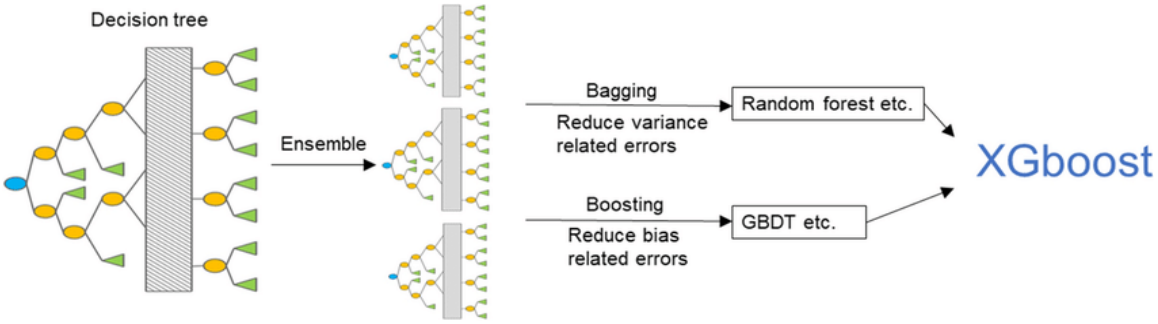
Ο αλγόριθμος eXtreme Gradient Boosting ή XGBoost είναι ένας αλγόριθμος μηχανικής μάθησης με επίβλεψη που ανήκει στην κατηγορία των αλγορίθμων συλλογικής μάθησης (ensemble learning) και ειδικότερα στο γενικό πλαίσιο των αλγορίθμων Gradient Boosting. Αποτελεί μία εκ των τεσσάρων βασικών boosting αλγορίθμων που συνδυάζει πολλαπλούς αδύναμους ταξινομητές, κυρίως δένδρα απόφασης, για τη δημιουργία ενός ισχυρότερου ταξινομητή. Είναι γνωστός για την υπολογιστική του δύναμη, την ανάλυση σημαντικότητας των χαρακτηριστικών που εισέρχονται στο μοντέλο και την ικανότητα διαχείρισης ελλειπουσών τιμών αφού μαθαίνει αυτόματα να κάνει προβλέψεις για την κατάλληλη τιμή όταν συναντά κενές εγγραφές.

Ως μοντέλο συλλογικής μάθησης λειτουργεί προσθέτοντας επαναληπτικά αδύναμους μαθητές (weak learners) και συγκεκριμένα δένδρα απόφασης τα οποία εκπαιδεύονται διαδοχικά και στοχεύουν στην μείωση του σφάλματος των προηγούμενων δένδρων. Έτσι το κάθε επόμενο στη σειρά δένδρο θα μάθει από το προηγούμενο. Οι αδύναμοι μαθητές έχουν υψηλή μεροληψία και μικρή προβλεπτική ικανότητα, όμως καθένας τους συνεισφέρει μερική σημαντική πληροφορία για την τελική πρόβλεψη. Αυτό βοηθάει την τεχνική ενίσχυσης να δημιουργήσει ένα ισχυρό μοντέλο συνδυάζοντας αυτούς του μαθητές.

Εν αντιθέσει με τις τεχνικές σακούλας (bagging) που χρησιμοποιούν τα τυχαία δάση στα οποία τα δένδρα απόφασης φτάνουν το μεγαλύτερο δυνατό μέγεθος, η μέθοδος boosting χρησιμοποιεί δέντρα μικρότερου μεγέθους τα οποία είναι εξαιρετικά ερμηνεύσιμα. Οι παράμετροι του XGBoost όπως ο αριθμός των δένδρων ή των επαναλήψεων, το επίπεδο στο οποίο ο gradient boosting μαθαίνει και το μέγεθος κάθε δένδρου επιλέγονται με χρήση τεχνικών επικύρωσης όπως η διασταυρούμενη επικύρωση K-πτυχών.

Στα πλεονεκτήματα του XGBoost είναι η υψηλή ακρίβεια των αποτελεσμάτων του, η δυνατότητα επέκτασης χάρη στην οποία διαχειρίζεται μεγάλα σύνολα δεδομένων με εκατομμύρια εγγραφές, η αποτελεσματικότητά του όσον αφορά τους υπολογισμούς και την ταχεία εκπαίδευση μοντέλων και η ευελιξία του. Επιπλέον, με χρήση τεχνικών κανονικοποίησης μειώνει το πρόβλημα υπερπροσαρμογής των δεδομένων και βοηθάει στη γενίκευση του σε νέα δεδομένα. Τέλος, δίνει σαν αποτέλεσμα σκορ για τη σημαντικότητα των

χαρακτηριστικών, βοηθώντας του χρήστες να καταλάβουν ποιες μεταβλητές είναι χρησιμότερες για προβλέψεις.



Σχήμα 4.8: Διαγραμματική απεικόνιση διαδικασίας XGBoost
(Πηγή: https://www.researchgate.net/figure/Schematic-representation-of-the-XGBoost-model-XGBoost-extreme-gradient-boosting-GBDT_fig6_349960301)

4.3.1.6 Κ Κοντινότεροι Γείτονες (K Nearest Neighbours, KNN)

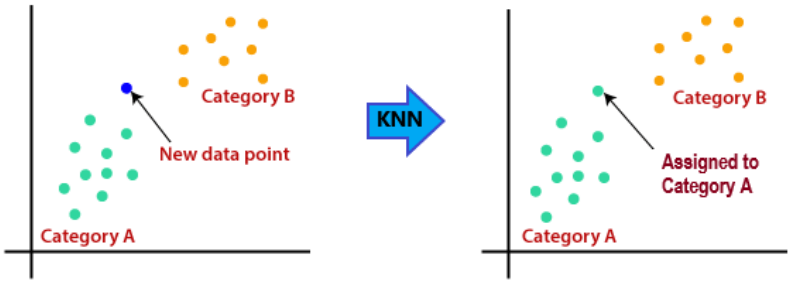
Ο αλγόριθμος Κ κοντινότερων γειτόνων είναι ένας από τους πιο δημοφιλείς και απλούς αλγορίθμους ταξινόμησης εποπτευόμενης μάθησης για κατηγοριοποίηση ή πρόβλεψη της ομάδας ενός συγκεκριμένου σημείου δεδομένου (data point). Είναι μη παραμετρικός πράγμα που σημαίνει ότι δε χρειάζονται υποθέσεις για την κατανομή που ακολουθούν τα δεδομένα ή για τη σχέση μεταξύ της εξαρτημένης και των ανεξάρτητων μεταβλητών. Βασίζεται στην ιδέα ότι παρόμοια σημεία δεδομένων τείνουν να έχουν κοινές ετικέτες ή τιμές.

Κατά τη διάρκεια της φάσης εκπαίδευσης ο KNN αποθηκεύει ολόκληρο το σύνολο δεδομένων εκπαίδευσης σαν αναφορά. Στη συνέχεια, για την πρόβλεψη υπολογίζει την απόσταση μεταξύ του νέου δεδομένου εισόδου και των δεδομένων εκπαίδευσης χρησιμοποιώντας ένα προκαθορισμένο μέτρο απόστασης. Τα συνήθη μέτρα απόστασης είναι η Ευκλείδεια απόσταση, η απόσταση Manhattan και η απόσταση Minkowski. Κατόπιν, εξακριβώνονται οι Κ πλησιέστεροι γείτονες στο δεδομένο που εισέρχεται στον αλγόριθμο βάσει των αποστάσεων τους. Για τη διαδικασία κατηγοριοποίησης ο αλγόριθμος δίνει την πιο κοινή ετικέτα (label) μεταξύ των κοντινότερων γειτόνων ως την προβλεπόμενη για το νέο δεδομένο.

Η επίδοση του αλγορίθμου καθορίζεται τόσο από την επιλογή του Κ από τον επόπτη όσο και από το εκάστοτε μέτρο απόφασης. Η τιμή του Κ υποδηλώνει των αριθμό των γειτόνων που θα ελεγχθούν για να καθορίσουν την κλάση του νέου σημείου δεδομένου. Ο καθορισμός του Κ επηρεάζει την ισορροπία του αλγορίθμου αφού διαφορετικές τιμές του μπορεί να οδηγήσουν σε προβλήματα είτε υπερπροσαρμογής (overfitting) είτε υποπροσαρμογής (underfitting) των νέων δεδομένων στο μοντέλο. Επιπροσθέτως, μικρές τιμές του Κ μπορεί να αποδώσουν υψηλή διακύμανση και μικρή μεροληψία ενώ αντίστοιχα μεγάλες τιμές του Κ αποδίδουν μεγάλη μεροληψία και μικρότερη διασπορά. Η επιλογή το Κ βασίζεται κυρίως στο είδος του εκάστοτε συνόλου δεδομένων. Τα σύνολα δεδομένων με περισσότερο θόρυβο (noise) και ακραίες τιμές (outliers) είναι πιθανό να δώσουν ακριβέστερα αποτελέσματα για μεγάλες τιμές του Κ. Γενικά, προτείνεται περιττός αριθμός για το Κ προς αποφυγή ισοπαλιών στην ταξινόμηση και χρήση μεθόδων διασταυρούμενης επικύρωσης για την εύρεση της βέλτιστης τιμής του Κ σε κάθε περίπτωση.

Στα θετικά του αλγορίθμου KNN είναι η απλότητα και η ακρίβεια των αποτελεσμάτων, η προσαρμοστικότητα που επιδείχνει σε νέα δεδομένα και η χρήση μόνο δύο υπερπαραμέτρων,

του K και του μέτρου απόφασης. Παρ' όλ' αυτά έχει και αρκετούς περιορισμούς όπως η χρήση περισσότερης μνήμης για αποθήκευση δεδομένων με αποτέλεσμα να καταναλώνει περισσότερο χρόνο και χρήματα. Ακόμη, είναι θύμα της επονομαζόμενης κατάρας της διαστατικότητας που σημαίνει ότι έχει χαμηλή απόδοση όταν χειρίζεται σύνολα δεδομένων με πολλές διαστάσεις. Τέλος, έχει την τάση να υπερπροσαρμόζεται στα δεδομένα και να δίνει παραπλανητικά αποτελέσματα.



Σχήμα 4.9: Διαγραμματική απεικόνιση του αλγορίθμου KNN
(Πηγή: https://ai.plainenglish.io/k-nearest-neighbors-knn-769bd39514c6)

Μέτρα απόστασης των K κοντινότερων γειτόνων

Όπως προαναφέρθηκε ο αλγόριθμος K κοντινότερου γείτονα έχει δύο υπερπαραμέτρους. Μία εξ αυτών είναι το μέτρο απόστασης και χρησιμεύει στον προσδιορισμό των K κοντινότερων σημείων των δεδομένων ώστε να δοθεί μια ετικέτα κλάσης στο νέο σημείο που εισέρχεται στο μοντέλο. Στην ουσία οι αποστάσεις αυτές αποτελούν έναν κανόνα απόφασης για την επιλογή των K γειτόνων.

Τα πιο δημοφιλή μέτρα απόστασης αναφέρονται παρακάτω:

- Ευκλείδεια απόσταση (Euclidean distance): Η ευκλείδεια απόσταση αποτελεί την πιο απλή κα γνωστή απόσταση ανάμεσα σε συνεχή δεδομένα. Μετράει την απόσταση με μία ευθεία γραμμή ανάμεσα στο υπό μελέτη δεδομένο και στα ήδη υπάρχοντα. Εξαρτάται κυρίως από την κλίμακα μέτρησης και επομένως αλλάζοντας την κλίμακα λαμβάνουμε τελείως διαφορετικές τιμές. Ακόμη, η συνεισφορά μεταβλητών με μεγάλες απόλυτες τιμές είναι συνήθως πολύ μεγαλύτερη από τη συνεισφορά μεταβλητών με μικρές απόλυτες τιμές με αποτέλεσμα η απόσταση μεταξύ των παρατηρήσεων να καθορίζεται σχεδόν αποκλειστικά από τις μεγάλες τιμές.
  Ο μαθηματικός της τύπος είναι ο εξής:

$$d(x_i, x_j) = \sqrt{\sum_{r=1}^{p} (x_{ir} - x_{jr})^2}$$

Όπου, x_i = (x_{i1}, x_{i2}, …, x_{ip}) το διάνυσμα των παρατηρήσεων για τα p χαρακτηριστικά που αντιστοιχεί στην i εγγραφή (i = 1, 2, …, n) και x_j = (x_{j1}, x_{j2}, …, x_{jp}) το διάνυσμα των παρατηρήσεων που αντιστοιχεί στην j εγγραφή.

- Απόσταση Manhattan ή City-block Distance ή Taxicab Distance: Η απόσταση Manhattan μοιάζει αρκετά με την ευκλείδεια απόσταση με τη διαφορά ότι μετρά την απόλυτη τιμή μεταξύ δύο σημείων δεδομένων αντί της χρήσης τετραγωνικών ριζών. Δίνει περίπου ίδια αποτελέσματα με την ευκλείδεια απόσταση εκτός από τη περίπτωση ύπαρξης έκτροπων παρατηρήσεων (outliers). Στην περίπτωση αυτή επειδή τους δίνει μικρότερο βάρος αφού δεν υψώνει τις διαφορές στο τετράγωνο μπορεί να οδηγήσει σε πιο ανθεκτικά αποτελέσματα. Χαρακτηρίζεται επίσης και ως απόσταση οικοδομικού τετραγώνου (city-block distance) ή απόσταση ταξί (taxicab distance) αφού απεικονίζεται με τη μορφή πλέγματος αποτυπώνοντας τον τρόπο με τον οποίο μπορεί κανείς να καθοδηγηθεί από μία διεύθυνση σε μία άλλη μέσω δρόμων πόλης.
  Ο μαθηματικός της τύπος είναι ο εξής:

$$d(x_i, x_j) = (\sum_{r=1}^{p} |x_{ir} - y_{jr}|)$$

  Όπου, x_i = (x_{i1}, x_{i2}, …, x_{ip}) το διάνυσμα των παρατηρήσεων για τα p χαρακτηριστικά που αντιστοιχεί στην i εγγραφή (i = 1, 2, …, n) και x_j = (x_{j1}, x_{j2}, …, x_{jp}) το διάνυσμα των παρατηρήσεων που αντιστοιχεί στην j εγγραφή.

- Απόσταση Minkowski: Η απόσταση αυτή αποτελεί γενίκευση της ευκλείδειας απόστασης και της απόστασης Manhattan αφού η πρώτη προκύπτει για φ=2 και η δεύτερη για φ=1.
  Ο μαθηματικός της τύπος είναι ο εξής:

$$d(x_i, x_j) = \left(\sum_{r=1}^{p} |x_{ir} - x_{jr}|^{\varphi}\right)^{1/\varphi}$$

  Όπου, x_i = (x_{i1}, x_{i2}, …, x_{ip}) το διάνυσμα των παρατηρήσεων για τα p χαρακτηριστικά που αντιστοιχεί στην i εγγραφή (i = 1, 2, …, n) και x_j = (x_{j1}, x_{j2}, …, x_{jp}) το διάνυσμα των παρατηρήσεων που αντιστοιχεί στην j εγγραφή

- Απόσταση Hamming: Η μέθοδος αυτή χρησιμοποιείται κυρίως για Boolean διανύσματα ή διανύσματα συμβολοσειρών, εντοπίζοντας τα σημεία δεδομένων όπου τα διανύσματα δεν ταιριάζουν. Για αυτό χαρακτηρίζεται και ως μετρική επικάλυψη.
  Ο μαθηματικός της τύπος είναι ο εξής:

$$D_H = \left(\sum_{r=1}^{p} |x_{ir} - x_{jr}| \right)$$

Όπου, xi = (xi1, xi2, …, xip) το διάνυσμα των παρατηρήσεων για τα p χαρακτηριστικά που αντιστοιχεί στην i εγγραφή (i = 1, 2, …, n) και xj = (xj1, xj2, …, xjp) το διάνυσμα των παρατηρήσεων που αντιστοιχεί στην j εγγραφή

4.3.1.7 Γραμμική Διαχωριστική Ανάλυση (Linear Discriminant Analysis, LDA)

Η γραμμική διαχωριστική ανάλυση (Linear Discriminant Analysis or Normal Discriminant Analysis or Discriminant Function Analysis) αποτελεί γενίκευση της γραμμικής διαχωριστικότητας του Fisher και είναι μέθοδος εποπτευόμενης μηχανικής μάθησης που χρησιμοποιείται για την επίλυση προβλημάτων ταξινόμησης πολλαπλών κλάσεων. Για την ακρίβεια βρίσκει γραμμικούς συνδυασμούς παραγόντων οι οποίοι διαχωρίζουν τα δεδομένα σε δύο ή περισσότερες κατηγορίες. Ο τελικός συνδυασμός μπορεί να χρησιμοποιηθεί σαν γραμμικός ταξινομητής ή συνηθέστερα για μείωση διαστατικότητας (dimensionality reduction). Η LDA είναι παραμετρική μέθοδος και επομένως κάνει υποθέσεις σχετικά με την κατανομή των δεδομένων.

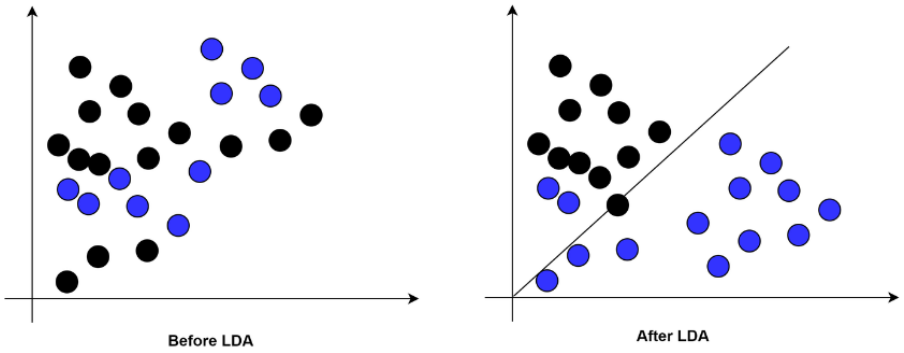
Η ταξινομική ανάλυση διακρίνεται σε δύο κατηγορίες. Η πρώτη αφορά στην LDA δύο κλάσεων όπου υπάρχουν μόνο δύο κλάσεις για να ταξινομηθούν τα δεδομένα. Τότε η LDA μειώνει τις διαστάσεις των δεδομένων σε μία και ταξινομεί τα δεδομένα βάσει του προκαθορισμένου κριτηρίου απόφασης. Στην περίπτωση που υπάρχουν περισσότερες από δύο κατηγορίες για την ταξινόμηση των δεδομένων η LDA μειώνει τις διαστάσεις των δεδομένων ώστε να είναι κατά μία λιγότερες από τις προκαθορισμένες κλάσεις.

Για την επίτευξη του παραπάνω χρειάζεται να βρεθεί ένα σύνολο από γραμμικούς διαχωριστές που θα μεγιστοποιούν το λόγο της διασποράς μεταξύ των κλάσεων προς τη διασπορά εντός των κλάσεων. Ο αλγόριθμος LDA υπολογίζει πρώτα το διάνυσμα που περιλαμβάνει τη μέση τιμή κάθε χαρακτηριστικού για κάθε κλάση και τον πίνακα συνδυακύμανσης που μετρά τη διασπορά μεταξύ και εντός των κλάσεων. Κατόπιν, υπολογίζεται ο πίνακας διασποράς που είναι στην ουσία το άθροισμα των πινάκων συνδυακύμανσης κάθε κλάσης και ο οποίος μετρά τη διακύμανση μεταξύ των κλάσεων με σκοπό την εύρεση των ιδιοδιανυσμάτων. Τα ιδιοδιανύσματα αντιπροσωπεύουν τις κατευθύνσεις στις οποίες τα δεδομένα πρέπει να προβληθούν για την επίτευξη του μέγιστου διαχωρισμού κλάσεων. Ακολούθως, οι ιδιοτιμές αντιπροσωπεύουν το σύνολο της διακύμανσης που εξηγείται από τα αντίστοιχα ιδιοδιανύσματα. Στη μέθοδο LDA οι ιδιοτιμές υποδηλώνουν τη σημαντικότητα των αντίστοιχων διαχωριστικών διανυσμάτων. Μεγαλύτερες ιδιοτιμές υποδεικνύουν μεγαλύτερο διαχωρισμό των κλάσεων στα αντίστοιχα ιδιοδιανύσματα. Έπειτα τα μετασχηματισμένα δεδομένα χρησιμοποιούνται για την εκπαίδευση ταξινομητή που θα προβλέπει την κλάση των νέων δεδομένων.

Η μέθοδος LDA έχει αρκετά πλεονεκτήματα όπως η απλότητα και αποτελεσματικότητα στους υπολογισμούς. Επιπλέον, μπορεί να διαχειριστεί δεδομένα πολλαπλών διαστάσεων και είναι αποτελεσματική όταν ο αριθμός των χαρακτηριστικών είναι μεγαλύτερος από τον αριθμό των εγγραφών του συνόλου δεδομένων εκπαίδευσης. Γι’ αυτό μπορεί να χρησιμοποιηθεί σε εφαρμογές όπως η ανάλυση κειμένου και η αναγνώριση εικόνας. Τέλος, διαχειρίζεται το πρόβλημα της πολυσυγγραμμικότητας (multicollinearity) της παρουσίας δηλαδή υψηλής συσχέτισης μεταξύ των διαφορετικών χαρακτηριστικών αφού μετατρέπει το

σύνολο δεδομένων σε χώρο χαμηλότερων διαστάσεων ενώ παράλληλα διατηρεί ακέραια την πληροφορία.

Ωστόσο, η γραμμική διαχωριστική ανάλυση παρουσιάζει και κάποια αρνητικά χαρακτηριστικά όπως η δυσκολία χειρισμού των κατανομών με κοινό μέσο. Πιο συγκεκριμένα, όταν οι κατανομές των κλάσεων παρουσιάζουν κοινό μέσο υπάρχει δυσκολία στη δημιουργία αξόνων που θα διαχωρίζουν γραμμικά τις δύο κλάσεις. Δηλαδή κλάσεις με επικάλυψη (overlapping) δε διαχωρίζονται αποτελεσματικά. Παραδείγματος χάριν, αν δύο είδη λουλουδιών έχουν πολύ παρόμοιο μήκος και πλάτος πετάλου τότε ο LDA αλγόριθμος θα αντιμετωπίζει πρόβλημα στον διαχωρισμό των ειδών αυτών αν βασιστεί μόνο στα παραπάνω δύο χαρακτηριστικά. Σε τέτοιες περιπτώσεις προτιμώνται εναλλακτικά μέθοδοι μη γραμμικής διαχωριστικής ανάλυσης. Επίσης, χρησιμοποιείται αυστηρά σε επισημασμένα δεδομένα, ενώ στην περίπτωση μη επισημασμένων δεδομένων χρησιμοποιείται η μέθοδος ανάλυσης κύριων συνιστωσών (principal component analysis, PCA). Τέλος, παρουσιάζει ευαισθησία σε ακραίες τιμές και διαθέτει περιορισμό ίσων πινάκων συνδυακύμανσης σε όλες τις κλάσεις.



Σχήμα 4.10: Διαγραμματική απεικόνιση πριν και μετά την εφαρμογή της μεθόδου LDA
(Πηγή: https://medium.com/@gajendra.k.s/linear-discriminant-analysis-lda-8b8d0c163e08)

Διαχωριστική Ανάλυση: Μαθηματική Φόρμουλα

Για τη μαθηματική εφαρμογή της διαχωριστικής ανάλυσης χρειάζεται αρχικά ο υπολογισμός της διακύμανσης μεταξύ των κλάσεων, της απόστασης δηλαδή μεταξύ των μέσων τιμών των διαφόρων κλάσεων και ο υπολογισμός της διακύμανσης εντός κλάσης που είναι η απόσταση μεταξύ της μέσης τιμής και του δείγματος κάθε κλάσης. Οι σχετικοί τύποι δίνονται παρακάτω ως εξής:

- **Διακύμανση μεταξύ των κλάσεων (Between-class variance)**

$$S_B = \sum_{i=1}^{c} N_i(\bar{X}_i - \bar{X})(\bar{X}_i - \bar{X})^T$$

- **Διακύμανση εντός κλάσης (Within-class variance)**

$$S_W = \sum_{i=1}^{c}(N_i - 1)S_i = \sum_{i=1}^{c}\sum_{j=1}^{N_i}(\bar{X}_{i,j} - \bar{X})(\bar{X}_{i,j} - \bar{X})^T$$

Τέλος, για την κατασκευή του χώρου χαμηλότερης διάστασης που μεγιστοποιεί τη διακύμανση μεταξύ των κλάσεων και ελαχιστοποιεί τη διακύμανση εντός των κλάσεων αρκεί να υπολογιστεί ο παρακάτω λόγος:

$$J(P) = \frac{P^T S_B P}{P^T S_W P},$$

όπου P η προβολή του χώρου χαμηλότερης διάστασης γνωστή και ως κριτήριο του Fisher.

4.3.2 Μέθοδοι Παλινδρόμησης (Regression Methods)

Η ανάλυση παλινδρόμησης αποτελεί τεχνική εποπτευόμενης μηχανικής μάθησης η οποία βοηθάει στην εύρεση συσχέτισης μεταξύ μεταβλητών με σκοπό την πρόβλεψη τιμών συνεχών μεταβλητών βασιζόμενη σε μία ή περισσότερες προβλεπτικές μεταβλητές. Είναι μέθοδος που μοντελοποιεί τη σχέση μεταξύ μιας εξαρτημένης (dependent) μεταβλητής και μιας ή περισσότερων ανεξάρτητων (independent) μεταβλητών. Η εξαρτημένη μεταβλητή καλείται, επίσης, μεταβλητή απόκρισης (response variable) ή μεταβλητή στόχος (target variable). Βοηθά στην κατανόηση του τρόπου με τον οποίο μεταβάλλεται η μεταβλητή απόκρισης σε σχέση με μια εκ των ανεξάρτητων μεταβλητών όταν οι υπόλοιπες προβλεπτικές μεταβλητές παραμένουν σταθερές.

Στόχος της ανάλυσης παλινδρόμησης είναι η εκτίμηση των παραμέτρων του μοντέλου εκείνου που θα προβλέψει καλύτερα την τιμής της μεταβλητής απόκρισης για ένα δοθέν σύνολο δεδομένων. Η μεταβλητή απόκρισης είναι συνεχής και μπορεί να αντιστοιχεί σε πραγματικά μεγέθη όπως η θερμοκρασία, η ηλικία, ο μισθός. Η παλινδρόμηση χρησιμοποιείται κυρίως για πρόβλεψη, πρόγνωση καιρού, μοντελοποίηση χρονοσειρών και καθορισμό σχέσης αιτίου-αποτελέσματος μεταξύ μεταβλητών.

Στην παρούσα εργασία θα αναλυθούν οι εξής μέθοδοι παλινδρόμησης:

- Γραμμική Παλινδρόμηση (Linear Regression)
- Παλινδρόμηση Κορυφογραμμής (Ridge Regression)
- Παλινδρόμηση LASSO (LASSO Regression)
- Παλινδρόμηση Elastic Net (Elastic Net Regression)
- Παλινδρόμηση με Δένδρα Απόφασης (Decision Trees Regression)
- Παλινδρόμηση με Τυχαία Δάση (Random Forest Regression)
- Παλινδρόμηση Gradient Boosting (Gradient Boosting Regression)

4.3.2.1 Γραμμική Παλινδρόμηση (Linear Regression)

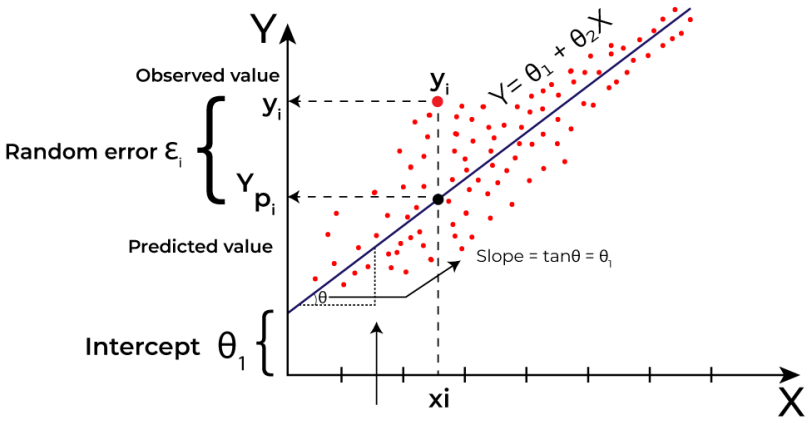
Η ανάλυση γραμμικής παλινδρόμησης (Linear Regression) είναι μία από τις εύκολες και δημοφιλής τεχνικές μηχανικής μάθησης. Αποτελεί στατιστική μέθοδο που χρησιμοποιείται για προβλεπτική ανάλυση συνεχών μεταβλητών όπως μισθός, ηλικία, τιμή προϊόντος. Ο αλγόριθμος γραμμικής παλινδρόμησης δείχνει μια γραμμική σχέση μεταξύ της μεταβλητής απόκρισης (Y) και μιας ή περισσότερων ανεξάρτητων μεταβλητών (Xi) που σημαίνει ότι δείχνει πως η τιμή τής εξαρτημένης μεταβλητής μεταβάλλεται βάσει της τιμής τής ανεξάρτητης μεταβλητής. Όταν στο μοντέλο υπάρχει μία μόνο ανεξάρτητη μεταβλητή τότε η

γραμμική παλινδρόμηση καλείται απλή ενώ όταν υπάρχουν περισσότερες από μία ανεξάρτητες μεταβλητές η παλινδρόμηση καλείται πολλαπλή.

Η σχέση μεταξύ της εξαρτημένης και των ανεξάρτητων μεταβλητών δίνεται από την παρακάτω γραμμική εξίσωση η οποία καλείται και ευθεία παλινδρόμησης ως εξής:

$$Y = \beta_0 + \beta_i X_i + \varepsilon_i$$

όπου Y είναι η μεταβλητή απόκρισης, x_1, x_2, …, x_n είναι οι ανεξάρτητες μεταβλητές, β0 είναι ο σταθερός όρος και β_1, β_2, …, β_n είναι οι συντελεστές γραμμικής παλινδρόμησης που αντιπροσωπεύουν τη μεταβολή που θα επιφέρει στην εξαρτημένη μεταβλητή Y μία μοναδιαία μεταβολή σε καθεμία από τις ανεξάρτητες μεταβλητές. Κάθε φορά μεταβάλλεται μόνο μία ανεξάρτητη μεταβλητή διατηρώντας τις υπόλοιπες σταθερές. Τα ε_1, ε_2, …, ε_n είναι τα τυχαία σφάλματα.



Σχήμα 4.11: Διαγραμματική απεικόνιση της ευθείας παλινδρόμησης
(Πηγή: https://www.geeksforgeeks.org/ml-linear-regression/)

Στόχος του μοντέλου είναι η εκτίμηση των συντελεστών που ελαχιστοποιούν το άθροισμα των τετραγωνικών διαφορών μεταξύ των προβλεπόμενων τιμών και των πραγματικών τιμών της μεταβλητής απόκρισης. Η διαδικασία αυτή ονομάζεται μέθοδος ελαχίστων τετραγώνων (method of least squares). Η μαθηματική φόρμουλα της μεθόδου είναι η εξής:

$$\sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y_i - \beta_0 \pm \beta_i x_i)^2$$

Συνεπώς, οι τιμές των β_0 και βi που ελαχιστοποιούν την παραπάνω εξίσωση καλούνται εκτιμήτριες ελαχίστων τετραγώνων και δίνονται από τις παρακάτω σχέσεις:

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_\iota \bar{x}$$

και

$$\hat{\beta}_i = \frac{n\sum_{i=1}^{n} x_i y_i - (\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i)}{n\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2}$$

όπου $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ και $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$.

Η γραμμική παλινδρόμηση είναι μέθοδος που χρησιμοποιείται σε πάρα πολλούς τομείς, από τη βιολογία και την περιβαλλοντολογική επιστήμη μέχρι τις κοινωνικές επιστήμες και την επιχειρηματική δραστηριότητα, και αυτό οφείλεται στα πλεονεκτήματα που διαθέτει. Η απλότητα και ταχύτητα με την οποία γίνονται οι υπολογισμοί αλλά και η ερμηνευσιμότητα των αποτελεσμάτων είναι μερικά από αυτά. Παρ' όλες όμως τις ευκολίες που προσφέρει δεν παύει να διαθέτει και περιορισμούς όπως η υπόθεση για την ύπαρξη γραμμικής σχέσης μεταξύ της μεταβλητής απόκρισης και των ερμηνευτικών μεταβλητών, ευαισθησία σε ακραίες τιμές και η υπόθεση ότι οι μεταβλητές που εισέρχονται στο μοντέλο είναι μεταξύ τους ανεξάρτητες ώστε να μην παρουσιάζεται πρόβλημα πολυσυγγραμμικότητας. Τέλος, λόγω της απλότητάς της αλλά και των παραπάνω περιορισμών είναι συχνά δύσκολο να εφαρμοστεί σε προβλήματα με πραγματικά δεδομένα.

4.3.2.2 Παλινδρόμηση Κορυφογραμμής (Ridge Regression)

Η παλινδρόμηση κορυφογραμμής (Ridge Regression or L2 Regularization) είναι στατιστική τεχνική κανονικοποίησης (regularization) που χρησιμοποιείται όταν το σύνολο δεδομένων αποτελείται από μεταβλητές που συσχετίζονται σε μεγάλο βαθμό μεταξύ τους ή όταν υπάρχει πρόβλημα υπερπροσαρμογής του μοντέλου στα δεδομένα.

Πρόκειται για μία μέθοδο που βασίζεται στην ελαχιστοποίηση του ποινικοποιημένου αθροίσματος τετραγώνων (penalized sum of squares). Πιο συγκεκριμένα, στην παλινδρόμηση ridge οι συντελεστές των ερμηνευτικών μεταβλητών του μοντέλου εκτιμώνται ελαχιστοποιώντας μία συνάρτηση κόστους που περιλαμβάνει έναν όρο ποινής (penalty term). Ο όρος αυτός βοηθάει στη διόρθωση των υψηλών τιμών των συντελεστών του μοντέλου οι οποίες αποτελούν συνήθως ένδειξη υπερπροσαρμογής. Η συνάρτηση κόστους δίνεται από τον κάτωθι τύπο:

$$RSS_{L2} = \sum_{i=1}^{n}\left(Y_i - \hat{Y}_L\right)^2 + \lambda\sum_{j=1}^{p}\beta_j^2$$

Η επίδραση του όρου ποινής στο μοντέλο ελέγχεται από την παραπάνω παράμετρο λ η οποία καθορίζει και το μέγεθος της κανονικοποίησης που θα εφαρμοστεί στο μοντέλο γι' αυτό και κάθε φορά το βέλτιστο μοντέλο επιλέγεται δοκιμάζοντας ένα εύρος τιμών για το λ.

Η παραπάνω διαδικασία που μειώνει την τιμή των συντελεστών προκειμένου να διαχειριστεί την ύπαρξη πολυσυγγραμμικότητας, η οποία προκαλεί πρόβλημα στην γραμμική παλινδρόμηση, αποτελεί και ένα από τα βασικότερα πλεονεκτήματα της παλινδρόμησης ridge. Η ύπαρξη πολυσυγγραμμικότητας υποδηλώνει ότι δύο ή περισσότερες μεταβλητές έχουν πολύ άμεση γραμμική σχέση μεταξύ τους και αυτό προκαλεί δυσκολία στην ερμηνευσιμότητα του μοντέλου. Ένα ακόμη χαρακτηριστικό της παλινδρόμησης κορυφογραμμής είναι η ανταλλαγή της διακύμανσης με τη μεροληψία αφού σε κάποιες περιπτώσεις είναι θεμιτή η λήψη μεροληπτικών αποτελεσμάτων προκειμένου να μειωθεί η διακύμανση.

Παρά τα πλεονεκτήματα που προσφέρει η συγκεκριμένη μέθοδος διαθέτει και κάποιους περιορισμούς όπως η υπόθεση ότι όλες οι ανεξάρτητες μεταβλητές του συνόλου δεδομένων είναι σημαντικές για το μοντέλο. Ωστόσο, αυτό δεν ισχύει σε όλες τις εφαρμογές καθώς ορισμένες από τις μεταβλητές πολύ συχνά δεν προσφέρουν καμία πληροφορία στο μοντέλο και η συμβολή τους στην ερμηνεία των αποτελεσμάτων είναι αμελητέα. Στην προκειμένη περίπτωση άλλες μέθοδοι, όπως η παλινδρόμηση LASSO που θα αναλυθεί παρακάτω, ίσως είναι καταλληλότερες. Ακόμη, χρειάζεται μεγάλη προσοχή για την επιλογή της κατάλληλης τιμής της υπερπαραμέτρου η οποία επηρεάζει σημαντικά τα αποτελέσματα του μοντέλου. Τέλος, στην παλινδρόμηση ridge παρουσιάζεται δυσκολία ερμηνείας των αποτελεσμάτων.

4.3.2.3 Παλινδρόμηση LASSO (LASSO Regression)

Η παλινδρόμηση LASSO (Least Absolute Shrinkage and Selection Operator) είναι μία δημοφιλής τεχνική που χρησιμοποιείται στη στατιστική μοντελοποίηση και στη μηχανική μάθηση για να εκτιμήσει τις σχέσεις μεταξύ των μεταβλητών και να κάνει προβλέψεις. Χρησιμοποιείται κυρίως για τη διαχείριση συνόλων δεδομένων πολλαπλών διαστάσεων αφού πραγματοποιεί αυτόματα επιλογή των χαρακτηριστικών του μοντέλου κατά την εφαρμογή της. Αυτό γίνεται προσθέτοντας στο άθροισμα τετραγώνων έναν όρο ποινής ο οποίος πολλαπλασιάζεται με την υπερπαράμετρο κανονικοποίησης λ που ελέγχει το μέγεθος της κανονικοποίησης.

Η παλινδρόμηση LASSO καλείται και L1 regularization καθώς εκτελεί κανονικοποίηση L1, προσθέτει δηλαδή έναν όρο ποινής ο οποίος είναι ίσος με το άθροισμα των απόλυτων τιμών των συντελεστών του μοντέλου. Αυτή η διαδικασία είναι πιθανό να οδηγήσει σε απλούστερα μοντέλα με λιγότερες ανεξάρτητες μεταβλητές γεγονός που βοηθάει στην αποφυγή προβλήματος πολυσυγγραμμικότητας και υπερπροσαρμογής. Αυτό γίνεται γιατί ορισμένοι από τους συντελεστές του μοντέλου μπορεί να μηδενιστούν. Με αυτόν τον τρόπο μπορεί να βελτιωθεί και η ερμηνευσιμότητα του μοντέλου εν αντιθέσει με άλλες μεθόδους όπως η παλινδρόμηση Ridge που αναλύθηκε παραπάνω.

Στην παλινδρόμηση LASSO η συνάρτηση που πρέπει να ελαχιστοποιηθεί είναι η εξής:

$$RSS_{L2} = \sum_{i=1}^{n} \left(Y_i - \hat{Y}_L\right)^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$

Όπως ήδη αναφέρθηκε και για την παλινδρόμηση κορυφογραμμής, πολύ σημαντικό ρόλο παίζει η σωστή επιλογή της τιμής της υπερπαραμέτρου λ αφού επηρεάζει το μέγεθος της κανονικοποίησης και κατ' επέκταση τον αριθμό των ανεξάρτητων μεταβλητών που τελικά θα έχει το μοντέλο. Αναλυτικότερα, όταν το λ=0 τότε δεν εξαλείφεται καμία παράμετρος και το τελικό μοντέλο παλινδρόμησης είναι ισοδύναμο με αυτό της γραμμικής παλινδρόμησης. Όταν το $\lambda \rightarrow \infty$ τότε όλο και περισσότεροι συντελεστές μηδενίζονται με αποτέλεσμα να απλοποιείται το μοντέλο όλο και περισσότερο. Ακόμη, υπογραμμίζεται πως όταν αυξάνεται η τιμή του λ αυξάνεται και η μεροληψία του μοντέλου ενώ όταν μειώνεται η τιμή του λ αυξάνεται η διασπορά του μοντέλου.

Στα πλεονεκτήματα της παλινδρόμησης LASSO είναι η επιλογή χαρακτηριστικών μέσω της διαδικασίας συρρίκνωσης των συντελεστών, διαδικασία που βοηθάει στην αποφυγή πολύπλοκων και δυσνόητων μοντέλων με πολλές μεταβλητές οι οποίες δεν προσφέρουν καμία πληροφορία. Επιπλέον, η παλινδρόμηση LASSO μπορεί να διαχειριστεί σύνολα δεδομένων με πολλαπλές διαστάσεις καθώς και να αποφύγει το πρόβλημα της

υπερπροσαρμογής των δεδομένων στο μοντέλο. Ωστόσο, διαθέτει και κάποια μειονεκτήματα όπως υψηλή μεροληψία στην επιλογή των ανεξάρτητων μεταβλητών.

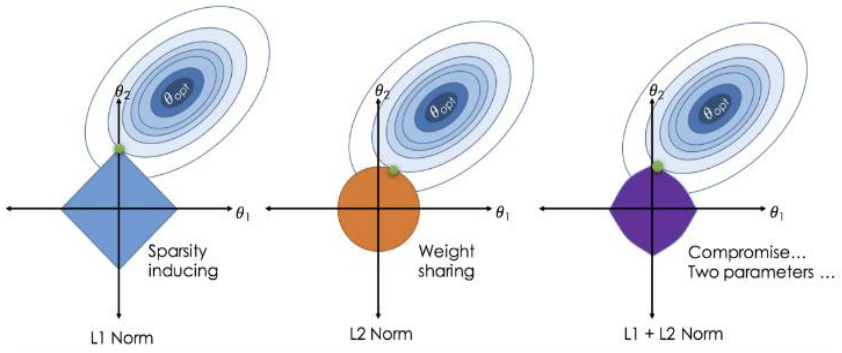
4.3.2.4 Παλινδρόμηση Elastic Net (Elastic Net Regression)

Η παλινδρόμηση Elastic Net αποτελεί συνδυασμό των δύο πιο δημοφιλών εκδοχών κανονικοποίησης της γραμμικής παλινδρόμησης, της Ridge και της LASSO οι οποίες και αναλύθηκαν παραπάνω. Ενώ η ridge χρησιμοποιεί την L2 ποινικοποίηση και η LASSO την L1, η Elastic Net δε διαλέγει μεταξύ των δύο μοντέλων καθώς χρησιμοποιεί ταυτόχρονα και τους δύο όρους ποινής. Πιο συγκεκριμένα, η συνάρτηση που πρέπει να ελαχιστοποιηθεί ως προς β είναι η εξής:

$$RSS_{L1,L2} = \sum_{i=1}^{n}\left(Y_i - \hat{Y}_L\right)^2 + \lambda\left\{a\sum_{j=1}^{p}|\beta_j| + (a-1)\sum_{j=1}^{p}\beta_j^2\right\}$$

Όπως φαίνεται από τον παραπάνω τύπο, οι παλινδρομήσεις Ridge και LASSO είναι ειδικές περιπτώσεις της παλινδρόμησης Elastic Net καθώς προκύπτουν για α=0 και α=1 αντίστοιχα. Για τιμές του α μεταξύ του διαστήματος 0 και 1, η Elastic Net μοιράζεται ιδιότητες και των μεθόδων μιας και επιλέγει μεταβλητές όπως η LASSO και συρρικνώνει τους συντελεστές των συσχετισμένων μεταβλητών όπως η Ridge.

Είναι από τις πιο σύγχρονες μεθόδους καθώς εισήχθη το 2005 από τους Zou και Hastie με αρκετά πλεονεκτήματα όπως η δυνατότητα επιλογής χαρακτηριστικών που οδηγεί σε μοντέλα με λιγότερες ανεξάρτητες μεταβλητές τα οποία είναι ευκολότερο να ερμηνευθούν και λιγότερο πιθανό να παρουσιάσουν πρόβλημα υπερπροσαρμογής. Επιπροσθέτως, αποτελεί μια πιο ισχυρή και ανθεκτική μέθοδο αφού συνδυάζει την ανθεκτικότητα και των δύο άλλων μεθόδων και μπορεί να διαχειριστεί σχετισμένες μεταβλητές και μεταβλητές διαφορετικής κλίμακας. Τέλος, φαίνεται πως αποδίδει καλύτερα από τις προαναφερθείσες μεθόδους, ειδικά σε σύνολα δεδομένων με μεγάλο αριθμό μεταβλητών. Εφαρμόζεται σε αρκετά επιστημονικά πεδία όπως στη βιοπληροφορική για την ταυτοποίηση γονιδίων που σχετίζονται με συγκεκριμένες ασθένειες, στην επεξεργασία εικόνων για την απομάκρυνση θορύβου από τις εικόνες και την αναπαράσταση αλλοιωμένων εικόνων και στα χρηματοοικονομικά για τη κατασκευή μοντέλων για πρόβλεψη τιμών μετοχών και άλλων προϊόντων.



Σχήμα 4.12: Διαγραμματική απεικόνιση των περιοχών περιορισμού των LASSO, Ridge και Elastic Net (από αριστερά προς τα δεξιά)
(Πηγή: https://freeiplmk.best/product_details/24642171.html)

4.3.2.5 Παλινδρόμηση με Δένδρα Απόφασης (Decision Trees Regression)

Η παλινδρόμηση με δένδρα απόφασης (Decision Trees Regression) είναι μια τεχνική μηχανική μάθησης που κατασκευάζει ένα μοντέλο με δομή δένδρου για την πρόβλεψη συνεχών μεταβλητών εν αντιθέσει με τα προβλήματα ταξινόμησης στα οποία τα αποτελέσματα αφορούν κατηγορικές μεταβλητές.

Αναλυτικότερα, στον πυρήνα της παλινδρόμησης με δένδρα απόφασης υπάρχει η δομή δένδρου. Αυτό πρακτικά σημαίνει ότι ο αλγόριθμος ξεκινά κατασκευάζοντας έναν αρχικό κόμβο ρίζα που αναπαριστά ολόκληρο το σύνολο δεδομένων. Στη συνέχεια, διαχωρίζει τα δεδομένα εισόδου σε υποσύνολα βάσει κατάλληλων κανόνων απόφασης ή κατωφλίων, με κάθε εσωτερικό κόμβο να περιέχει συγκεκριμένο εύρος τιμών βάσει του εκάστοτε κανόνα διαχωρισμού. Η διαδικασία αυτή επαναλαμβάνεται αναδρομικά έως ότου ικανοποιηθεί ένα κριτήριο τερματισμού. Το κριτήριο αυτό ορίζεται συνήθως από μεγέθη όπως το μέγιστο όριο βάθους του δένδρου, εξασφαλίζοντας ότι το δένδρο δε θα γίνει πολύ περίπλοκο. Εναλλακτικά, μπορεί να χρησιμοποιηθεί ο ελάχιστος αριθμός σημείων δεδομένων σε καθένα από τους τελικούς κόμβους φύλλα για την αποφυγή υπερπροσαρμογής. Η τελική πρόβλεψη για μια νέα τιμή εισόδου πραγματοποιείται με τη διάσχιση του δένδρου από τον αρχικό κόμβο ρίζα μέχρι και τον κατάλληλο τερματικό κόμβο φύλλο που προκύπτει από τις συνθήκες των εσωτερικών κόμβων.

Χαρακτηριστικό πλεονέκτημα της παλινδρόμησης με δένδρα απόφασης είναι η ερμηνευσιμότητα των αποτελεσμάτων καθώς η διαδικασία που ακολουθείται μπορεί να αναπαρασταθεί πολύ εύκολα από κατάλληλα διαγράμματα που διευκολύνουν στην κατανόηση των αποτελεσμάτων. Τέλος, αξίζει να σημειωθεί ότι μπορεί να διαχειριστεί γραμμικές και μη γραμμικές σχέσεις μεταξύ της μεταβλητής στόχου και των ερμηνευτικών μεταβλητών.

4.3.2.6 Παλινδρόμηση με Τυχαίο Δάσος (Random Forest Regression)

Η παλινδρόμηση τυχαίου δάσους (Random Forest Regression) είναι μία πολύπλευρη τεχνική μηχανικής μάθησης που χρησιμοποιείται για την πρόβλεψη συνεχών αποτελεσμάτων. Συνδυάζει τις προβλέψεις πολλαπλών δένδρων απόφασης με στόχο τον περιορισμό του προβλήματος της υπερπροσαρμογής και την βελτίωση της ακρίβειας του μοντέλου. Για την ακρίβεια, η παλινδρόμηση τυχαίου δάσους είναι συλλογική μέθοδος μάθησης που συνδυάζει τις προβλέψεις πολλαπλών δένδρων απόφασης για την παραγωγή μιας ισχυρότερης πρόβλεψης.

Για την κατασκευή ενός μοντέλου παλινδρόμησης τυχαίου δάσους επιλέγεται ένα τυχαίο σύνολο δεδομένων εκπαίδευσης για τη δημιουργία ενός δένδρου απόφασης. Το δένδρο απόφασης διαθέτει εσωτερικούς κόμβους με κανόνες απόφασης και τερματικό κόμβο όπως η μορφολογία που αναλύθηκε παραπάνω. Κατ’ αυτόν τον τρόπο δημιουργούνται πολλαπλά δένδρα απόφασης χρησιμοποιώντας κάθε φορά διαφορετικό τυχαίο υποσύνολο δεδομένων εκπαίδευσης. Για την τελική πρόβλεψη χρησιμοποιούνται τα αποτελέσματα όλων των δένδρων και το αποτέλεσμα που δίνει ο αλγόριθμος παλινδρόμησης τυχαίου δάσους είναι η μέση των προβλεπόμενων τιμών των δέντρων.

Η παλινδρόμηση τυχαίου δάσους διαθέτει κοινά πλεονεκτήματα με την παλινδρόμηση με δένδρα απόφασης όπως η διαχείριση τόσο γραμμικών όσο και μη γραμμικών σχέσεων μεταξύ της μεταβλητής απόκρισης και των υπόλοιπων μεταβλητών του συνόλου δεδομένων. Επιπλέον, μπορεί και αυτή να απεικονισθεί διαγραμματικά γεγονός που συμβάλλει στην καλύτερη κατανόηση και ερμηνεία των αποτελεσμάτων. Τέλος, η συγκεκριμένη μέθοδος έχει

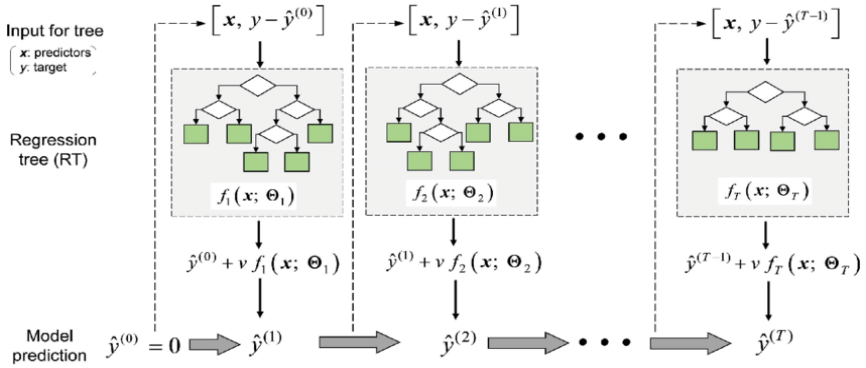
ανθεκτικότητα στην παρουσία ακραίων τιμών και μπορεί να χρησιμοποιηθεί για σύνολα δεδομένων πολλαπλών διαστάσεων.

4.3.2.7 Παλινδρόμηση με Gradient Boosting

Η παλινδρόμηση με Gradient Boosting είναι μία από τις συλλογικές μεθόδους μάθησης καθώς χρησιμοποιεί πολλαπλούς αδύναμους μαθητές, ένα συνδυασμό δηλαδή αδύναμων μοντέλων, συνήθως δένδρων απόφασης, σε αλληλουχία ώστε να κατασκευαστεί ένα ισχυρότερο μοντέλο.

Για την κατασκευή ενός μοντέλου παλινδρόμησης με Gradient Boosting εκπαιδεύεται συνήθως ένα αδύναμο μοντέλο, παραδείγματος χάριν ένα δένδρο απόφασης, για να κάνει προβλέψεις. Κατόπιν υπολογίζεται η διαφορά μεταξύ των προβλεπόμενων και των πραγματικών τιμών και χρησιμοποιείται για την εκπαίδευση ενός νέου δένδρου. Στην ουσία κάθε αδύναμο μοντέλο στοχεύει στην βελτίωση της ακρίβειας του προηγούμενου ακόμη πιο αδύναμου μοντέλου. Έτσι ο αλγόριθμος έχει την ικανότητα να δημιουργήσει ένα ισχυρό προβλεπτικό μοντέλο εκπαιδεύοντας τα αδύναμα μοντέλα ακολουθιακά. Η τελική πρόβλεψη για μία νέα τιμή δεδομένων προέρχεται από τις προβλέψεις όλων των αδύναμων μοντέλων.

Η παλινδρόμηση Gradient Boosting χρησιμοποιείται κυρίων λόγω της ικανότητάς της να χειρίζεται μη γραμμικές σχέσεις μεταξύ των ερμηνευτικών μεταβλητών και της μεταβλητής απόκρισης. Τέλος, παρουσιάζει μικρότερη τάση για υπερπροσαρμογή των δεδομένων στο μοντέλο.



Σχήμα 4.13: Διαγραμματική απεικόνιση παλινδρόμησης με Gradient Boosting
(Πηγή: https://www.researchgate.net/figure/Schematic-diagram-of-the-gradient-boosted-regression-tree_fig2_342270212)

4.3.3 Τεχνικές Μη Εποπτευόμενης Μάθησης

Τα μοντέλα μη εποπτευόμενης μάθησης χρησιμοποιούνται σε τρεις βασικές διαδικασίες οι οποίες είναι η ομαδοποίηση, οι κανόνες συσχέτισης και η μείωση της διαστατικότητας. Όσον αφορά την ομαδοποίηση (clustering) αποτελεί τεχνική εξόρυξης δεδομένων η οποία ομαδοποιεί μη επισημασμένα δεδομένα βασιζόμενη στις ομοιότητες ή τις διαφορές που αυτά εμφανίζουν. Έτσι οι αλγόριθμοι ομαδοποίησης χρησιμοποιούνται για να επεξεργαστούν τα αδρά, μη ταξινομημένα δεδομένα και να τα ταξινομήσουν σε ομάδες βάσει μοτίβων και δομών που προκύπτουν από τις πληροφορίες που προσφέρουν. Υπάρχουν διάφοροι τύποι αλγορίθμων συσταδοποίησης όπως η ιεραρχική συσταδοποίηση (hierarchical clustering), η συσταδοποίηση βάσει πυκνότητας (density-based clustering), η συσταδοποίηση βάσει κεντροειδούς (centroid-based clustering), η συσταδοποίηση βάσει κατανομής (distribution-

based clustering) και ο αλγόριθμος k-means. Η επιλογή κάθε αλγορίθμου εξαρτάται κάθε φορά από το πρόβλημα και το εκάστοτε σύνολο δεδομένων.

Η δεύτερη βασική διαδικασία αφορά στην δημιουργία κανόνων συσχέτισης. Πιο συγκεκριμένα, οι κανόνες συσχέτισης αποτελούν μία μέθοδο που βασίζεται σε συγκεκριμένους κανόνες ώστε να βρει συσχετίσεις μεταξύ των μεταβλητών ενός δοθέντος συνόλου δεδομένων στοχεύοντας στην εξόρυξη χρήσιμης πληροφορίας. Αυτές οι μέθοδοι χρησιμοποιούνται συνήθως στην ανάλυση καλαθιού αγοράς (market basket analysis) βοηθώντας τις εταιρείες στην αποτελεσματικότερη προώθηση και πώληση προϊόντων μέσω της κατανόησης των συνηθειών των καταναλωτών. Ο ευρέως χρησιμοποιούμενος αλγόριθμος για τους κανόνες συσχέτισης είναι ο Apriori. Χρησιμοποιείται σε δεδομένα συναλλαγών για την εξακρίβωση συνόλων αντικειμένων ή προϊόντων.

Τέλος, σημαντικό πρόβλημα στη μη εποπτευόμενη μάθηση αποτελεί η μείωση της διαστατικότητας (dimensionality reduction). Η μείωση της διαστατικότητας είναι τεχνική που χρησιμοποιείται όταν δίνεται σύνολο δεδομένων πολλαπλών διαστάσεων ή όταν ο αριθμός των χαρακτηριστικών του συνόλου δεδομένων είναι αρκετά μεγάλος. Εφαρμόζεται συνήθως στο στάδιο προπαρασκευής των δεδομένων. Υπάρχουν διάφορες τεχνικές μείωσης της διαστατικότητας όπως η ανάλυση κύριων συνιστωσών (PCA) και η ανάλυση σε ιδιάζουσες τιμές (SVD). Η επιλογή της εκάστοτε τεχνικής εξαρτάται κάθε φορά από το πρόβλημα και το δοθέν σύνολο δεδομένων.

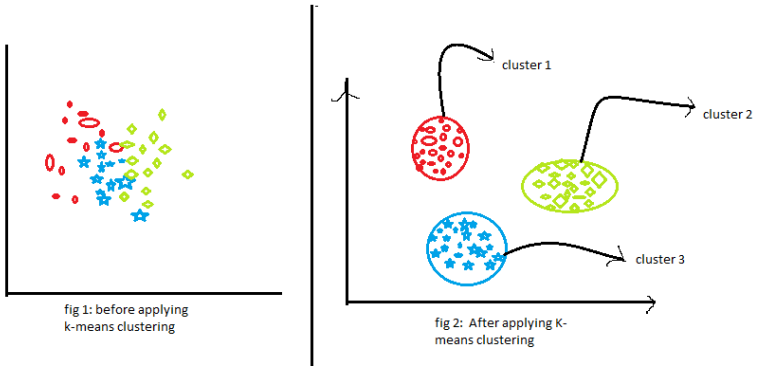
4.3.3.1 Αλγόριθμος Κ μέσων (K-means Algorithm)

Ο αλγόριθμος συσταδοποίησης Κ μέσων (K-means) είναι αλγόριθμος μη εποπτευόμενης μηχανικής μάθησης που χρησιμοποιείται για την επίλυση προβλημάτων συσταδοποίησης. Στοχεύει στην τμηματοποίηση και ομαδοποίηση των δεδομένων. Αναλυτικότερα, το αντικείμενό του είναι να διαχωρίσει ένα πεπερασμένο σύνολο δεδομένων σε k διακριτές μη επικαλυπτόμενες συστάδες (clusters), με κάθε παρατήρηση του συνόλου δεδομένων να εκχωρείται στη συστάδα της οποίας ο μέσος (ή κεντροειδές) είναι κοντινότερα, και έτσι να λειτουργεί ως αντιπρόσωπος αυτής της συστάδας. Αφού καθοριστεί η παράμετρος k ο αλγόριθμος επιλέγει τον αντίστοιχο αριθμό κεντροειδών τα οποία αντιπροσωπεύουν και τα κέντρα των συστάδων. Κατόπιν, εκχωρούνται τα σημεία δεδομένων στο πλησιέστερο σε αυτά κεντροειδές με επαναληπτική διαδικασία. Ως κριτήριο απόστασης του εκάστοτε σημείου από το πιθανό κεντροειδές επιλέγεται η ευκλείδεια απόσταση. Έπειτα καθορίζεται εκ νέου το κεντροειδές κάθε συστάδας βάσει της μέσης τιμής των εκχωρημένων δεδομένων. Η διαδικασία αυτή επαναλαμβάνεται έως ότου ο αλγόριθμος να συγκλίνει.

Στα πλεονεκτήματα του αλγορίθμου Κ μέσων περιλαμβάνονται η απλότητα και η ευκολία ως προς την εφαρμογή του. Επιπλέον, μπορεί να διαχειριστεί ταχύτατα μεγάλα σύνολα δεδομένων με πολλαπλές διαστάσεις και έχει τη δυνατότητα κλιμάκωσης σε περίπτωση ακόμη μεγαλύτερου συνόλου δεδομένων. Τέλος, διαθέτει ευελιξία καθώς προσαρμόζεται στις διάφορες εφαρμογές χρησιμοποιώντας κάθε φορά διαφορετικό μέτρο απόστασης. Παρ' όλα αυτά διαθέτει και κάποια μειονεκτήματα όπως η ευαισθησία στην επιλογή των κατάλληλων αρχικών κεντροειδών τα οποία παίζουν καθοριστικό ρόλο στα τελικά αποτελέσματα. Ακόμη, απαιτείται προκαθορισμός του Κ από τον ερευνητή, διαδικασία αρκετά απαιτητική για κάποιες περιπτώσεις. Τέλος, παρουσιάζει ευαισθησία στην ύπαρξη ακραίων τιμών γεγονός που μπορεί να έχει σημαντικό αντίκτυπο στη δημιουργία των τελικών συστάδων.

Μολονότι ο αλγόριθμος k-means διαθέτει και κάποια αρνητικά χαρακτηριστικά αποτελεί έναν από τους πιο δημοφιλής αλγορίθμους μη εποπτευόμενης μάθησης που μπορεί να

χρησιμοποιηθεί για ακαδημαϊκούς σκοπούς, για διαγνωστικά συστήματα, μηχανές αναζήτησης και ασύρματους αισθητήρες δικτύων.



Σχήμα 4.14: Διαγραμματική απεικόνιση πριν και μετά την εφαρμογή του k-means
(Πηγή: https://www.analyticsvidhya.com/blog/2020/10/a-simple-explanation-of-k-means-clustering/)

Διαδικασία αλγορίθμου K-means

Έστω ότι δίνεται συσταδοποίηση C = {C_1, C_2, …, C_k}. Χρειαζόμαστε κάποια συνάρτηση βαθμολόγησης η οποία θα αξιολογεί την ποιότητα (quality) ή καταλληλότητα (goodness) της συσταδοποίησης. Αυτή η συνάρτηση βαθμολόγησης βασίζεται στο άθροισμα των τετραγώνων των σφαλμάτων (sum of squared errors, SSE) και ορίζεται ως

$$SSE(C) = \sum_{i=1}^{k} \sum_{x_j \in C_i} \|x_j - \mu_i\|^2$$

Όπου μ_i είναι ο μέσος όλων των σημείων της συστάδας i για i=1, …, k που ονομάζεται επίσης κέντρο βάρους (κεντροειδές) και δίνεται από τον παρακάτω τύπο ως εξής

$$\mu_i = \frac{1}{n_i} \sum_{x_j \in C_i} x_j$$

Όπου $n_i = |C_i|$ είναι το πλήθος των σημείων που ανήκουν στην συστάδα C_i.

Στόχος είναι η εύρεση της συσταδοποίησης εκείνης που ελαχιστοποιεί τη βαθμολογία SSE:

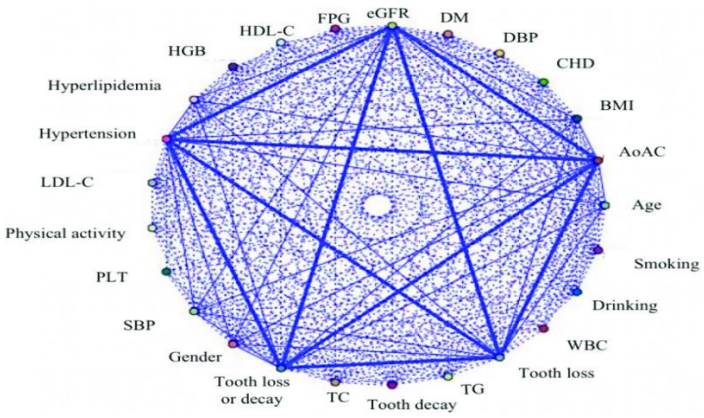
$$C^* = arg \min_C \{SSE(C)\}$$

Ο K-means χρησιμοποιεί μια πλεονεκτική (greedy), επαναληπτική τεχνική για να βρει μια συσταδοποίηση που ελαχιστοποιεί την αντικειμενική συνάρτηση SSE. Κατά συνέπεια, μπορεί να συγκλίνει σε τοπικά βέλτιστα (local optima) αντί σε μία καθολικά βέλτιστη συσταδοποίηση.

4.3.3.2 Αλγόριθμος Apriori (Apriori Algorithm)

Ο αλγόριθμος Apriori προτάθηκε από τους Agrawal και Srikant το 1994 και σχεδιάστηκε για να εφαρμόζεται σε βάσεις δεδομένων που περιέχουν στοιχεία συναλλαγών όπως συλλογές από αντικείμενα που αγόρασαν οι καταναλωτές ή λεπτομέρειες παρουσίας σε μία ιστοσελίδα. Στην απλοϊκή μέθοδο απαριθμούνται όλα τα πιθανά στοιχειοσύνολα σε μία προσπάθεια να εξακριβωθεί ποια στοιχειοσύνολα είναι συχνά. Ως συχνά στοιχειοσύνολα χαρακτηρίζονται αυτά των οποίων η υποστήριξη είναι μεγαλύτερη από ένα προκαθορισμένο κατώφλι ή μία ελάχιστη τιμή υποστήριξης ορισμένη από τον ερευνητή. Ο αλγόριθμος στηρίζεται στη έρευνα κατά επίπεδα (level-wise) ή στην πρώτα κατά πλάτος έρευνα με χρήση δένδρου Hash για τον αποτελεσματικό υπολογισμό των συσχετίσεων του στοιχειοσυνόλου. Είναι, στην ουσία, μία επαναληπτική μέθοδος για την εύρεση συχνών στοιχειοσυνόλων από μεγάλα σύνολα δεδομένων.

Χρησιμοποιείται κυρίως στην ανάλυση καλαθιού αγοράς και βοηθά στη εύρεση των προϊόντων που συνήθως αγοράζονται μαζί όπως για παράδειγμα καφές-ζάχαρη-τσάι ή ψωμί-βούτυρο-μαρμελάδα. Μπορεί, επίσης, να χρησιμοποιηθεί στον τομέα της υγείας για την διαπίστωση αντιδράσεων των φαρμάκων στους ασθενείς. Στα πλεονεκτήματά του περιλαμβάνονται η ευκολία στη κατανόηση της δομής και λειτουργίας του καθώς και ευκολία χρήσης του σε μεγάλα σύνολα δεδομένων. Ωστόσο, διαθέτει σημαντικούς περιορισμούς όπως ότι είναι πιο αργός στη λειτουργία συγκριτικά με άλλους αντίστοιχους αλγορίθμους. Ακόμη, η συνολική απόδοσή του μπορεί να περιοριστεί μετά τη διάσχιση του συνόλου δεδομένων πολλαπλές φορές. Τέλος, η πολυπλοκότητα του χώρου του αλγορίθμου είναι $O(2^D)$ δηλαδή πολύ υψηλή αφού το D υποδηλώνει το οριζόντιο πλάτος του εκάστοτε συνόλου δεδομένων.



Σχήμα 4.15: Διαγραμματική απεικόνιση κανόνων συσχέτισης βάσει του αλγορίθμου Apriori
(Πηγή: https://www.researchgate.net/figure/Visualization-of-the-association-rules-based-on-the-Apriori-algorithm-The-thickness-of_fig1_366273637)

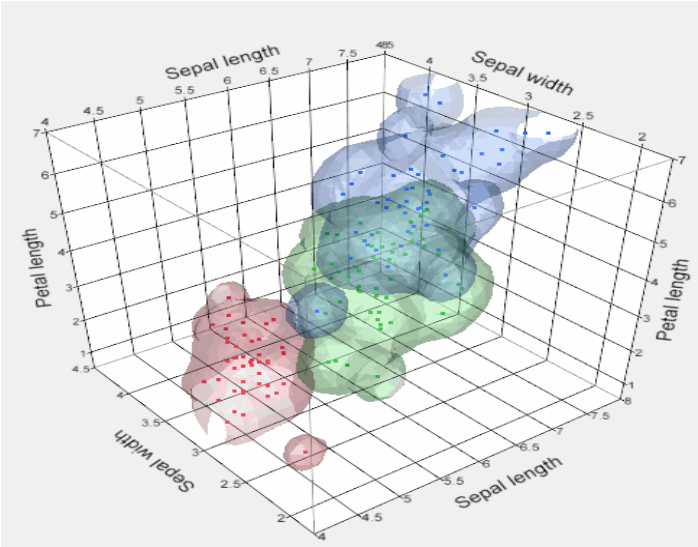
4.3.3.3 Ανάλυση Κύριων Συνιστωσών (Principal Component Analysis)

Η ανάλυση κύριων συνιστωσών (PCA) είναι μία εξαιρετικά δημοφιλής και αποτελεσματική μέθοδος για απεικόνιση και εξερεύνηση συνόλων δεδομένων πολλαπλών διαστάσεων ή συνόλων δεδομένων με πολλές μεταβλητές αφού μπορεί εύκολα να αναγνωρίσει τάσεις, μοτίβα και ακραίες τιμές. Χρησιμοποιείται ευρέως κυρίως στο στάδιο της προπαρασκευής των δεδομένων. Εξάγει τις μεταβλητές εκείνες που δίνουν την περισσότερη πληροφορία. Η διαδικασία αυτή ελαττώνει την πολυπλοκότητα του μοντέλου μιας και η προσθήκη νέων μεταβλητών συχνά τείνει να επηρεάζει αρνητικά το μοντέλο. Το

πρόβλημα αυτό χαρακτηρίζεται και ως «κατάρα της διαστατικότητας» (curse of dimensionality). Επίσης, προβάλλοντας ένα σύνολο δεδομένων πολλαπλών διαστάσεων σε μικρότερο χώρο χαρακτηριστικών, η PCA ελαχιστοποιεί προβλήματα όπως η πολυσυγγραμμικότητα και η υπερπροσαρμογή.

Η PCA συγκεντρώνει την πληροφορία, που περιέχεται σε μεγάλα σύνολα δεδομένων, σε μικρότερα σύνολα από ασυσχέτιστες μεταβλητές γνωστές ως κύριες συνιστώσες. Αυτές οι κύριες συνιστώσες είναι γραμμικοί συνδυασμοί των αρχικών μεταβλητών που επεξηγούν το μεγαλύτερο μέρος της μεταβλητότητας συγκριτικά με άλλους πιθανούς γραμμικούς συνδυασμούς. Οι συνιστώσες αυτές αποδίδουν όση περισσότερη πληροφορία είναι δυνατόν από το αρχικό σύνολο δεδομένων. Η πρώτη κύρια συνιστώσα (PC1) είναι η κατεύθυνση στο χώρο στον οποίο τα σημεία δεδομένων έχουν την υψηλότερη διακύμανση. Είναι δηλαδή η γραμμή που αναπαριστά καλύτερα το σχήμα των προβαλλόμενων δεδομένων. Όσο μεγαλύτερη είναι η μεταβλητότητα που αποτυπώνεται στην πρώτη συνιστώσα τόσο μεγαλύτερη είναι και η πληροφορία που διατηρείται από τα αρχικά δεδομένα. Καμία από τις επόμενες συνιστώσες δε μπορεί να έχει υψηλότερη μεταβλητότητα. Η δεύτερη κύρια συνιστώσα (PC2) υπολογίζεται όπως ακριβώς και η πρώτη, έχει τη δεύτερη μεγαλύτερη διακύμανση στο σύνολο δεδομένων και πρέπει να είναι ασυσχέτιστη με την πρώτη. Για να είναι ασυσχέτιστες οι κύριες συνιστώσες, κάθε μία από αυτές είναι ορθογώνια ως προς τις άλλες. Για την εύρεσή τους πρέπει να υπολογιστεί ο πίνακας συνδιακύμανσης των αρχικών μεταβλητών αφού τα ιδιοδιανύσματα (eigenvectors) του πίνακα συνδιακύμανσης είναι οι κύριες συνιστώσες και οι αντίστοιχες ιδιοτιμές τους (eigenvalues) αντιπροσωπεύουν το ποσοστό μεταβλητότητας που εξηγείται από αυτές.

Η PCA χρησιμοποιείται κυρίως για την εξαγωγή των πιο χρήσιμων μεταβλητών, με τη μεγαλύτερη πληροφορία, από σύνολα δεδομένων με μεγάλο αριθμό μεταβλητών. Συνήθεις περιπτώσεις που χρησιμοποιείται η PCA είναι στη διαδικασία συμπίεσης εικόνας (image compression) καθώς επιτρέπει τη δημιουργία συμπιεσμένων αναπαραστάσεων της εικόνας που διευκολύνει την αποθήκευσή τους. Ακόμη, εφαρμόζεται στην απεικόνιση δεδομένων αφού μειώνει τις διαστάσεις του χώρου καθιστώντας εφικτή τη διαδικασία αναπαράστασης. Τέλος, αφαιρεί από το σύνολο δεδομένων το θόρυβο και τις περιττές πληροφορίες. Ωστόσο, λόγω της υπόθεσης ότι τα δεδομένα σχετίζονται γραμμικά είναι πιθανό να μη λειτουργεί σε μη γραμμικά δεδομένα.



Σχήμα 4.16: Διαγραμματική 3D απεικόνιση της μεθόδου PCA
(Πηγή: https://www.oreilly.com/library/view/jmp-11-essential/9781612906850/EG_05_3d_Scatterplot.xhtml)

4.4 Τεχνικές Αξιολόγησης Μοντέλων (Model Evaluation Techniques)

Η αξιολόγηση των μοντέλων είναι η διαδικασία χρήσης διαφορετικών μέτρων αξιολόγησης για την κατανόηση της απόδοσης ενός μοντέλου μηχανικής μάθησης καθώς και της εύρεσης των δυνατοτήτων και αδυναμιών του. Είναι σημαντική για την εκτίμηση της αποτελεσματικότητας ενός μοντέλου κατά τη διάρκεια των αρχικών φάσεων της έρευνας και παίζει επίσης ρόλο στην παρακολούθηση των μοντέλων.

Αναλυτικότερα χρησιμοποιούνται ποσοτικά μέτρα για την αξιολόγηση της επίδοσης και αποτελεσματικότητας στατιστικών μοντέλων ή μοντέλων μηχανικής μάθησης. Τα μέτρα αυτά δίνουν πληροφορίες για το πώς αποδίδουν τα μοντέλα σε αδρά και ανεπεξέργαστα δεδομένα συγκριτικά με άλλα μοντέλα ή αλγορίθμους. Με τη χρήση αυτών, οι εταιρείες μπορούν να βελτιώσουν την ακρίβεια των προβλέψεων και να πάρουν σημαντικές αποφάσεις βάσει δεδομένων.

Οι τεχνικές αξιολόγησης των μοντέλων ποικίλουν ανάλογα με το είδος των μοντέλων που χρησιμοποιείται κάθε φορά και ανάλογα το στάδιο της έρευνας κατά το οποίο εφαρμόζεται.

Γενικότερα, έχουμε τις παρακάτω τεχνικές:

- **Holdout validation:** Είναι τεχνική κατά την οποία το αρχικό σύνολο δεδομένων διαχωρίζεται με τυχαίο τρόπο σε σύνολο δεδομένων εκπαίδευσης και σύνολο δεδομένων ελέγχου. Οι συνήθεις αναλογίες διαχωρισμού είναι 70% για το σύνολο εκπαίδευσης και 30% για το σύνολο ελέγχου ή 80% για την εκπαίδευση και 20 για τον έλεγχο. Βάσει της τεχνικής αυτή το μοντέλο εκπαιδεύεται με τα δεδομένα εκπαίδευσης και αξιολογείται η απόδοση του στα δεδομένα ελέγχου.
- **Διασταυρούμενη επικύρωση (Cross-validation):** Η διασταυρούμενη επικύρωση βασίζεται στην τεχνική της δειγματοληψίας με επανάληψη. Πιο συγκεκριμένα, διαχωρίζει τα δεδομένα σε k ισομεγέθη υποσύνολα (πτυχές-folds). Ένα από τα υποσύνολα αποθηκεύεται ως υποσύνολο ελέγχου και τα υπόλοιπα ως υποσύνολα εκπαίδευσης. Η διαδικασία επαναλαμβάνεται πολλές φορές χρησιμοποιώντας κάθε φορά διαφορετικό υποσύνολο για τον έλεγχο. Κατόπιν, υπολογίζεται ο μέσος όρος από τα βήματα επικύρωσης για να παραχθεί ένα πιο ισχυρό και αποδοτικό μοντέλο.
- **Μετρικά (Metrics):** Τα μετρικά βοηθούν στην αξιολόγηση της απόδοσης των μοντέλων αλλά και στη σύγκριση των αποτελεσμάτων παρόμοιων μοντέλων παρέχοντας ποσοτικά μέτρα όπως η ακρίβεια (accuracy/precision), η ανάκληση (recall) ή κάλυψη (coverage), το F μέτρο (F1-score) και η καμπύλη ROC.
- **Καμπύλη μάθησης ή εκπαίδευσης (learning or training curve):** Αυτή η τεχνική αναπαριστά γραφικά τη βέλτιστη τιμή μιας συνάρτησης ζημίας (loss function), ενός μοντέλου για το σύνολο δεδομένων εκπαίδευσης κατ' αντιπαράσταση με τη συνάρτηση ζημίας για τα δεδομένα ελέγχου. Αναφέρεται, επίσης και ως καμπύλη σφάλματος (error curve), καμπύλη εμπειρίας (experience curve) ή καμπύλη βελτίωσης (improvement curve). Βοηθά στον εντοπισμό τυχόν υπερπροσαρμογής ή υποπροσαρμογής του μοντέλου στα δεδομένα.
- **Διερεύνηση πλέγματος (grid search):** Η διερεύνηση πλέγματος είναι τεχνική συντονισμού των υπερπαραμέτρων που περιέχει κάποιος αλγόριθμος. Πραγματοποιεί πολλούς υπολογισμούς στις υπερπαραμέτρους που εμπεριέχει ο κάθε αλγόριθμος προκειμένου να καταλήξει στο ιδανικό σύνολο των τιμών των υπερπαραμέτρων που βοηθούν στην επίτευξη καλύτερων αποτελεσμάτων.

4.4.1 Τεχνικές Αξιολόγησης Μοντέλων Ταξινόμησης

Η μήτρα σύγχυσης (confusion matrix) είναι ένας πίνακας που συνοψίζει την απόδοση ενός μοντέλου μηχανικής μάθησης για ένα σύνολο δεδομένων εκπαίδευσης. Στην ουσία, είναι ένα μέσο παρουσίασης των συμβάντων που ταξινομήθηκαν με ακρίβεια ή ανακρίβεια βάσει των προβλέψεων του μοντέλου. Χρησιμοποιείται κυρίως για την αξιολόγηση της απόδοσης μοντέλων κατηγοριοποίησης. Στην περίπτωση δυαδικής κατηγοριοποίησης βοηθά στον υπολογισμό των ποσοστών αληθώς θετικών, αληθώς αρνητικών, ψευδώς θετικών και ψευδώς αρνητικών ταξινομήσεων. Μπορεί να χρησιμοποιηθεί και σε προβλήματα ταξινομητών με περισσότερες των δύο κλάσεων, ωστόσο εδώ θα αναλυθεί η περίπτωση μόνο της δυαδικής κατηγοριοποίησης.

Αναλυτικότερα τα προαναφερθέντα ποσοστά είναι τα εξής:

- **Αληθώς θετικά (TP):** Το πλήθος των περιπτώσεων που ο κατηγοριοποιητής προβλέπει σωστά ως θετικά.
- **Αληθώς αρνητικά (TN):** Το πλήθος των περιπτώσεων που ο κατηγοριοποιητής προβλέπει σωστά ως αρνητικά.
- **Ψευδώς θετικά (FP):** Το πλήθος των περιπτώσεων που ο κατηγοριοποιητής προβλέπει ως θετικά και στην πραγματικότητα ανήκουν στην αρνητική κατηγορία.
- **Ψευδώς αρνητικά (FN):** Το πλήθος των περιπτώσεων που ο κατηγοριοποιητής προβλέπει ως αρνητικά και στην πραγματικότητα ανήκουν στην θετική κατηγορία.

Η μορφής της μήτρας σύγχυσης δίνεται παρακάτω ως εξής:

	Πρόβλεψη	
Πραγματικότητα	**Θετικό**	**Αρνητικό**
Θετικό	Ορθά θετικό	Εσφαλμένα αρνητικό
Αρνητικό	Εσφαλμένα θετικό	Ορθά αρνητικό

Πίνακας 4.1: Μήτρα σύγχυσης (Confusion matrix)

Με βάση τις τιμές της μήτρας σύγχυσης υπολογίζονται οι εξής μετρικές:

- **Ορθότητα (Accuracy):** Είναι το ποσοστό των σωστών προβλέψεων του μοντέλου και υπολογίζεται ως εξής:

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

- **Ακρίβεια εξειδικευμένη για κάθε κατηγορία (Precision):** Η ακρίβεια για τη θετική και την αρνητική κατηγορία δίνεται από τις σχέσεις:

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

$$prec_N = \frac{TN}{TN + FN}$$

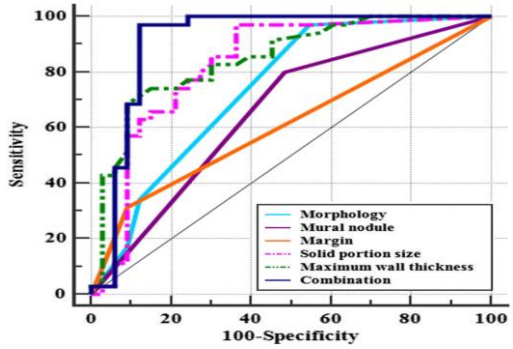
- **Ανάκληση (Recall):** Μετρά την αναλογία των αληθώς θετικών περιπτώσεων μεταξύ όλων των πραγματικών θετικών προβλέψεων:

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

- **F-μέτρο (F1-score):** Είναι ο αρμονικός μέσος της ακρίβειας και της ανάκλησης και χρησιμοποιείται για την αξιολόγηση της συνολικής απόδοσης του μοντέλου ταξινόμησης. Δίνεται από τη σχέση:

$$F1score = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

- **Καμπύλη ROC (ROC Curve):** Η καμπύλη ROC (Receiver Operating Characteristic, χαρακτηριστική καμπύλη λειτουργίας δέκτη) είναι μια δημοφιλής μέθοδος για την αξιολόγηση της απόδοσης των κατηγοριοποιητών για δυαδικό μοντέλο ταξινόμησης. Απαιτεί από τον κατηγοριοποιητή να δώσει ως έξοδο μια τιμή-βαθμολογία για τη θετική κατηγορία, για κάθε σημείο του συνόλου ελέγχου. Οι βαθμολογίες αυτές μπορούν να χρησιμοποιηθούν στη συνέχεια για την ταξινόμηση των σημείων κατά φθίνουσα σειρά. Ωστόσο, στην περίπτωση των δυαδικών κατηγοριοποιητών επιλέγεται κάποιο θετικό κατώφλι βαθμολογίας ρ. Τα σημεία με βαθμολογία μεγαλύτερη από ρ κατηγοριοποιούνται ως θετικά ενώ τα υπόλοιπα ως αρνητικά. Επειδή ο καθορισμός ενός τέτοιου κατωφλίου τείνει να είναι αυθαίρετος η ανάλυση ROC σχεδιάζει τη γραφική παράσταση της απόδοσης του κατηγοριοποιητή για όλες τις πιθανές τιμές της παραμέτρου του κατωφλίου ρ. Πιο συγκεκριμένα, για κάθε τιμή του ρ, σχεδιάζει τη γραφική παράσταση του ψευδώς θετικού ποσοστού (άξονας x) και του αληθώς θετικού ποσοστού (άξονας y). Αυτή η γραφική παράσταση ονομάζεται καμπύλη ROC. Το εμβαδόν της περιοχής κάτω από την καμπύλη, που ονομάζεται AUC (Area Under Curve) χρησιμοποιείται ως μέτρο απόδοσης για τον ταξινομητή. Αφού το συνολικό εμβαδόν του διαγράμματος είναι ίσο με 1, το AUC ανήκει στο διάστημα [0,1] και επομένως, όσο ψηλότερη είναι η τιμή του τόσο καλύτερος είναι ο ταξινομητής.



Σχήμα 4.17: Καμπύλη ROC
(Πηγή: https://www.researchgate.net/figure/Receiver-operating-characteristic-ROC-curve-analysis-was-conducted-to-discriminate-BOTs_fig4_339868711)

4.4.2 Τεχνικές Αξιολόγησης Μοντέλων Παλινδρόμησης

Όπως συμβαίνει για τα μοντέλα ταξινόμησης, έτσι και για τα μοντέλα παλινδρόμησης μπορούν να χρησιμοποιηθούν πολλές διαφορετικές μετρικές για την αξιολόγηση της απόδοσής τους. Αναλυτικότερα:

- **Μέτρο R^2 (R^2 Score):** Είναι στατιστική μετρική που χρησιμοποιείται συνήθως για την αξιολόγηση της καλής προσαρμογής ενός μοντέλου παλινδρόμησης στα δεδομένα. Αναφέρεται συχνά και ως συντελεστής προσδιορισμού (coefficient of determination). Μετρά το ποσοστό της διακύμανσης της εξαρτημένης μεταβλητής που εξηγείται από το μοντέλο. Το εύρος τιμών του είναι μεταξύ 0 και 1 με τιμές κοντά στη μονάδα να υποδηλώνουν καλή προσαρμογή. Δίνεται από τον τύπο:

$$R^2 = 1 - \frac{SSR}{SSTO}$$

  Όπου:
  SSR: το άθροισμα τετραγώνων της παλινδρόμησης (regression sum of squares)
  SSTO: το συνολικό άθροισμα τετραγώνων (total sum of squares)

- **Προσαρμοσμένο μέτρο R^2 (Adjusted R^2):** Ο συντελεστής προσδιορισμού μειονεκτεί στην περίπτωση που προστίθενται νέες μεταβλητές καθώς είτε αυξάνεται είτε παραμένει σταθερός αλλά ποτέ δε μειώνεται καθώς υποθέτει πως με την προσθήκη νέων μεταβλητών αυξάνεται το ποσοστό διακύμανσης της εξαρτημένης μεταβλητής που εξηγείται από το μοντέλο. Το πρόβλημα έγκειται στο γεγονός ότι μπορεί με την προσθήκη νέων μεταβλητών, οι οποίες προσφέρουν ελάχιστη πληροφορία, να αυξηθεί το R2, κάτι μη ορθό. Για τον έλεγχο του παραπάνω προβλήματος χρησιμοποιείται πολλές φορές ο προσαρμοσμένος συντελεστής προσδιορισμού που δίνεται από τον παρακάτω τύπο:

$$R^2_{adj} = 1 - \left[\left(\frac{n-1}{n-k-1}\right) \times (1 - R^2)\right]$$

  Όπου:
  n: ο αριθμός των παρατηρήσεων
  k: ο αριθμός των ανεξάρτητων μεταβλητών

- **Μέσο απόλυτο σφάλμα (Mean Absolute Error, MAE):** Μετρά την απόλυτη διαφορά μεταξύ των πραγματικών και προβλεπόμενων τιμών. Έχει εύρος τιμών από 0 έως ∞. Γενικά, όσο μικρότερη είναι η τιμή τόσο καλύτερο είναι το μοντέλο μιας και MAE = 0 υποδηλώνει τέλεια πρόβλεψη. Δίνεται από τον παρακάτω τύπο:

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |y_i - \hat{y}_i|$$

- **Μέσο τετραγωνικό σφάλμα (Mean Squared Error, MSE):** Υπολογίζει τη μέση τετραγωνική διαφορά μεταξύ των πραγματικών και προβλεπόμενων τιμών. Όσο

μικρότερη η τιμή του τόσο μικρότερο είναι και το σφάλμα του μοντέλου, επομένως, προτιμώνται τιμές που τείνουν στο 0. Δίνεται από τον παρακάτω τύπο:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

- **Ρίζα μέσου τετραγωνικού σφάλματος (Root MSE, RMSE):** Είναι η τετραγωνική ρίζα το MSE. Χρησιμοποιείται αντί του MSE καθώς είναι περισσότερο κατανοητό ως προς την ερμηνεία του.

$$RMSE = \sqrt{MSE}$$

- **Ρίζα μέσου τετραγωνικού λογαριθμικού σφάλματος (Root Mean Squared Logarithmic Error, RMSLE):** Υπολογίζεται εφαρμόζοντας λογαρίθμους στις πραγματικές και προβλεπόμενες τιμές και υπολογίζοντας μετά τη διαφορά τους. Το RMSLE είναι πιο ισχυρό στην ύπαρξη ακραίων τιμών αφού τα μικρά και μεγάλα σφάλματα επεξεργάζονται ομοιόμορφα.

$$RMSLE = \sqrt{(\log(y_i + 1) - log(\hat{y}_i + 1))^2}$$

4.4.3 Διασταυρούμενη Επικύρωση (Cross Validation, CV)

Η διασταυρούμενη επικύρωση είναι τεχνική που χρησιμοποιείται στη μηχανική μάθηση για την αξιολόγηση της απόδοσης ενός μοντέλου σε νέα, αθέατα δεδομένα. Διαχωρίζει τα διαθέσιμα δεδομένα σε πολλαπλά υποσύνολα (πτυχές) χρησιμοποιώντας κάθε φορά ένα από αυτά ως υποσύνολο επικύρωσης και τα υπόλοιπα υποσύνολα χρησιμοποιούνται για την εκμάθηση του μοντέλου. Η διαδικασία επαναλαμβάνεται πολλές φορές, χρησιμοποιώντας κάθε φορά διαφορετικό υποσύνολο ως υποσύνολο επικύρωσης. Τελικώς, υπολογίζεται ο μέσος όρος των αποτελεσμάτων των βημάτων επικύρωσης για την παραγωγή μιας πιο ισχυρής εκτίμησης της απόδοσης του μοντέλου.

Ο βασικός στόχος της χρήσης διασταυρούμενης μάθησης είναι η αποφυγή της υπερπροσαρμογής που προκύπτει όταν ένα μοντέλο είναι πολύ καλά εκπαιδευμένο αλλά παρουσιάζει πολύ κακή απόδοση σε νέα, αθέατα δεδομένα. Υπάρχουν διάφοροι τύποι διασταυρούμενης επικύρωσης η επιλογή των οποίων γίνεται βάσει του όγκου και της φύσης των δεδομένων καθώς και των ειδικών απαιτήσεων που εμφανίζει ο εκάστοτε αλγόριθμος μηχανικής μάθησης.

Αναλυτικότερα:

- **Leave-one-out cross validation (LOOCV):** Στη μέθοδο αυτή, χρησιμοποιείται για την εκπαίδευση του μοντέλου ολόκληρο το σύνολο δεδομένων αφήνοντας κάθε φορά εκτός μία μόνο παρατήρηση. Αυτό σημαίνει ότι κάθε φορά το μοντέλο εκπαιδεύεται για n-1 παρατηρήσεις και αξιολογείται βάσει της μίας που μένει εκτός, επαναλαμβάνοντας τη διαδικασία για καθεμία παρατήρηση ξεχωριστά. Το πλεονέκτημα αυτής της μεθόδου είναι ότι χρησιμοποιώντας όλες τις παρατηρήσεις μειώνεται η μεροληψία. Όμως, το βασικότερο μειονέκτημα είναι ότι μπορεί να οδηγήσει σε πολύ υψηλή διακύμανση αφού ελέγχεται κάθε φορά ένα μόνο σημείο

δεδομένου το οποίο μπορεί να είναι ακραία τιμή. Ακόμη, οι υπολογισμοί καταναλώνουν αρκετό χρόνο αφού γίνονται για κάθε παρατήρηση ξεχωριστά.

- **Στρωματοποιημένη διασταυρούμενη επικύρωση (Stratified CV):** Είναι τεχνική που χρησιμοποιείται στη μηχανική μάθηση για την εξασφάλιση της διατήρησης της ίδιας κατανομής κλάσεων σε κάθε πτυχή της διαδικασίας διασταυρούμενης μάθησης όπως αυτή παρουσιάζεται και στο αρχικό σύνολο δεδομένων. Είναι χρήσιμη στη διαχείριση ανομοιόμορφων συνόλων δεδομένων στα οποία μερικές κλάσεις είναι πιθανό να μην εκπροσωπούνται επαρκώς. Ακόμη, είναι σημαντική σε περιπτώσεις που η διατήρηση της κατανομής των κλάσεων είναι καίρια για τη γενίκευση του μοντέλου σε νέα δεδομένα.

- **Διασταυρούμενη επικύρωση Κ-πτυχών (K-fold CV):** Στη διασταυρούμενη επικύρωση Κ πτυχών το αρχικό σύνολο δεδομένων διαιρείται σε k υποσύνολα (πτυχές). Το εκάστοτε μοντέλο εκπαιδεύεται σε k-1 υποσύνολα και αξιολογείται η απόδοσή του σε αυτό που περισσεύει. Η διαδικασία αυτή επαναλαμβάνεται k φορές κρατώντας εκτός κάθε φορά ένα διαφορετικό υποσύνολο ως σύνολο επικύρωσης. Ο συνήθης χρησιμοποιούμενος αριθμός πτυχών είναι 10.

4.5 Προεπεξεργασία Δεδομένων (Data Pre-processing)

Η προεπεξεργασία των δεδομένων είναι η διαδικασία προετοιμασίας των αδρών δεδομένων (raw data) κάνοντάς τα κατάλληλα για τα μοντέλα μηχανικής μάθησης. Αποτελεί το πρώτο και πιο κρίσιμο στάδιο σε ένα ερευνητικό σχέδιο μηχανικής μάθησης. Αξίζει να αναφερθεί ότι στις περισσότερες περιπτώσεις τα δεδομένα που πρόκειται να χρησιμοποιηθούν δεν είναι κατάλληλα. Χρειάζεται πρώτα να εφαρμοστούν σε αυτά διαδικασίες όπως καθαρισμός (data cleaning), αφαίρεση ελλειπουσών τιμών (missing values) ή αντικατάστασή τους, αφαίρεση θορύβου (outliers/noise). Επιπλέον, είναι βασική η επιλογή μόνο των μεταβλητών που δίνουν την περισσότερη πληροφορία σε περιπτώσεις μεγάλων συνόλων δεδομένων ώστε να αποφευχθεί το πρόβλημα της διαστατικότητας. Ακόμη, απαιτείται κλιμάκωση των δεδομένων (data scaling) ώστε τα δεδομένα να βρίσκονται στη ίδια κλίμακα και να επιτρέπεται η από κοινού επεξεργασία τους.

Με τις παραπάνω διαδικασίες βελτιώνεται η ποιότητα των δεδομένων και αυξάνεται η ευκολία διαχείρισής τους. Απόρροια των παραπάνω είναι η κατασκευή αξιόπιστων μοντέλων με μικρότερη πιθανότητα εμφάνισης σφαλμάτων και αποκλίσεων.

4.5.1 Διαχείριση Ελλειπουσών Τιμών (Handling Missing Values)

Υπάρχουν διάφορες αιτίες για την απουσία τιμών από το σύνολο δεδομένων και είναι σημαντικό να γνωστοποιούνται καθώς επηρεάζεται και ο τρόπος διαχείρισής τους. Πιο συγκεκριμένα, παρατηρούνται ελλείπουσες τιμές επειδή τα δεδομένα έχουν συλλεχθεί πριν από αρκετά μεγάλο χρονικό διάστημα και η πληροφορία που προσφέρουν δε είναι σημαντική γι’ αυτό και αφαιρέθηκαν. Ακόμη, μπορεί να μην έχουν καταγραφεί παρατηρήσεις λόγω ανθρώπινου λάθους ή γιατί ο ερωτώμενος αρνήθηκε να απαντήσει. Τέλος, δεδομένα μπορεί να μην παρέχονται επιτηδευμένα λόγω προστασίας προσωπικών δεδομένων. Ο τρόπος χειρισμού τους πρέπει να εξετάζεται προσεκτικά καθώς μπορεί να επηρεαστούν σημαντικά τα τελικά αποτελέσματα.

4.5.1.1 Τύποι ελλειπουσών τιμών (Types of missing data)

- **Εντελώς τυχαία ελλείπουσες τιμές (Missing Completely At Random, MCAR):** Τέτοιες ελλείπουσες τιμές προκύπτουν όταν η πιθανότητα ύπαρξης ελλειπουσών τιμών είναι ομοιόμορφη σε όλο το σύνολο δεδομένων. Δεν παρατηρείται κάποιο μοτίβο στην έλλειψη συγκεκριμένων τιμών.
- **Τυχαία ελλείπουσες τιμές (Missing At Random, MAR):** Τέτοιου τύπου ελλείπουσες τιμές προκύπτουν όταν η πιθανότητα ύπαρξης ελλειπουσών τιμών εξαρτάται μόνο από τα παρατηρηθέντα δεδομένα, όχι από τις ίδιες τις ελλείπουσες τιμές. Παραδείγματος χάριν, σε μία μελέτη, τιμές της μεταβλητής «Ηλικία» μπορεί να λείπουν για όσους δε συμπλήρωσαν το φύλο τους. Εδώ οι απώλειες από τη μεταβλητή «Ηλικία» εξαρτάται από τη μεταβλητή «Φύλο» όμως οι ελλείπουσες τιμές της μεταβλητής «Ηλικία» κατανέμονται τυχαία στο σύνολο αυτών που δε συμπλήρωσαν το φύλο τους.
- **Επιτηδευμένα ελλείπουσες τιμές (Missing Not At Random, MNAR):** Τέτοιες ελλείπουσες τιμές προκύπτουν όταν δεν έχουν εσκεμμένα παρατηρηθεί τιμές για κάποιες μεταβλητές. Αυτός ο τύπος ελλειπουσών τιμών παρουσιάζει μοτίβο και δε μπορεί να επεξηγηθεί από τις παρατηρούμενες μεταβλητές.

4.5.1.2 Τρόποι διαχείρισης ελλειπουσών τιμών (Handling missing values)

- **Αφαίρεση (Deletion):** Με αυτήν την τεχνική αφαιρούνται από το σύνολο δεδομένων γραμμές ή στήλες που περιέχουν ελλείπουσες τιμές. Ωστόσο, μπορεί να φανεί προβληματική αν αφαιρεθεί σημαντικό μέρος του αρχικού συνόλου δεδομένων καθώς μπορεί να επηρεάζει την ακρίβεια και ορθότητα των προβλέψεων.
- **Υπολογισμός (Imputation):** Με αυτήν την τεχνική αντικαθίστανται οι ελλείπουσες τιμές με εκτιμήσεις. Μερικοί από τους κοινούς τρόπους αντικατάστασης των ελλειπουσών τιμών είναι:
    - Χρήση μέσου όρου (Mean)
    - Χρήση διαμέσου (Median)
    - Χρήση πιο συχνής τιμής (Mode)
    - Χρήση αλγορίθμου Κ κοντινότερων γειτόνων (KNN)
    - Υπολογισμός με χρήση μοντέλου που βασίζεται σε άλλα χαρακτηριστικά

4.5.2 Κωδικοποίηση κατηγορικών μεταβλητών (Label Encoding)

Η κωδικοποίηση ετικέτας (label encoding) είναι μία συνήθης τεχνική για την μετατροπή τω κατηγορικών μεταβλητών σε αριθμητικές τιμές. Είναι διαδικασία απαραίτητη αφού οι περισσότεροι αλγόριθμοι μηχανικής μάθησης λαμβάνουν μόνο αριθμητικές μεταβλητές. Στην κωδικοποίηση ετικέτας κάθε μοναδική κατηγορία μιας κατηγορικής μεταβλητής αντιστοιχίζεται με έναν μοναδικό ακέραιο βάσει αριθμητικής ή αλφαβητικής διάταξης. Παραδείγματος χάριν, για την κατηγορική μεταβλητή φύλο (male, female, non-binary) η κατηγορίες μπορούν να κωδικοποιηθούν ως εξής: male: 0, female: 1, non-binary: 2.

Γενικά, αποτελεί πολύ σημαντικό βήμα της προεπεξεργασίας των δεδομένων αφού χωρίς αυτό οι περισσότεροι αλγόριθμοι δε λειτουργούν σωστά και τα αποτελέσματα που δίνουν δεν είναι έγκυρα και ακριβή.

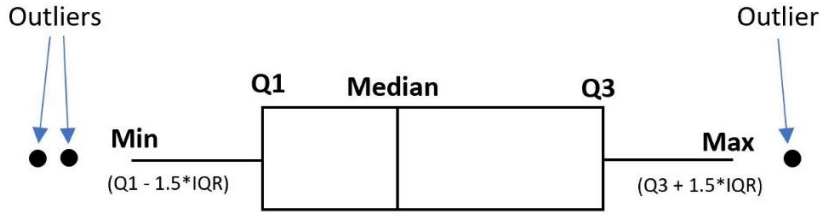
4.5.3 Ακραίες Τιμές (Outliers)

Στη μηχανική μάθηση μια ακραία τιμή είναι μία μεμονωμένη παρατήρηση η οποία διαχωρίζεται σημαντικά από τα υπόλοιπα δεδομένα. Μπορεί να είναι είτε πολύ μεγαλύτερη είτε πολύ μικρότερη από τα άλλα σημεία δεδομένων και η παρουσία της μπορεί να έχει εξαιρετικά μεγάλο αντίκτυπο στα αποτελέσματα των αλγορίθμων. Προκύπτουν συνήθως λόγω ανθρώπινου λάθους ή λανθασμένου υπολογισμού, όμως μπορεί να αντιπροσωπεύουν και μία πολύ σπάνια ή ασυνήθιστη περίπτωση.

Υπάρχουν δύο βασικοί τύποι ακραίων τιμών. Οι καθολικές ακραίες τιμές (global outliers) οι οποίες είναι απομονωμένα σημεία δεδομένων σε μεγάλη απόσταση από το κύριο σώμα δεδομένων. Είναι συνήθως εύκολο να εντοπιστούν και να αφαιρεθούν. Όσον αφορά στο δεύτερο τύπο ακραίων τιμών, αυτές χαρακτηρίζονται ως σχετικές ακραίες τιμές (contextual outliers) και μπορεί για ένα συγκεκριμένο περιεχόμενο να χαρακτηρίζονται ως ακραίες τιμές αλλά πιθανόν σε ένα διαφορετικό περιεχόμενο είναι παρατηρήσεις του βασικού σώματος δεδομένων.

4.5.3.1 Μέθοδοι εντοπισμού ακραίων τιμών (Outlier Detection Methods)

- **Στατιστικές Μέθοδοι:** Αυτές οι τεχνικές βασίζονται σε στατικές μετρικές όπως το Z-score και το ενδοτεταρτημοριακό εύρος (interquartile range, IQR).
- **Μέθοδοι βάσει απόστασης (Distance-Based Methods):** Συνήθως χαρακτηρίζουν τα outliers ως τους μακρινότερους γείτονες. Τέτοιες μέθοδοι είναι ο K-nearest neighbours και ο τοπικός παράγοντας ακραίων τιμών (Local Outlier Factor, LOF).
- **Μέθοδοι βάσει συσταδοποίησης (Clustering-Based Methods):** Τέτοιες μέθοδοι είναι ο αλγόριθμος συσταδοποίησης βασισμένος στην πυκνότητα (Density-Based Spatial Clustering and Applications with Noise, DBSCAN) και οι ιεραρχικοί αλγόριθμοι.
- **Άλλες μέθοδο:** Άλλες μέθοδοι εντοπισμού των ακραίων τιμών είναι ο αλγόριθμος δάσους απομόνωσης (isolation forest) και οι μηχανές διανυσμάτων υποστήριξης μίας κλάσης (One-class Support Vector Machines, OCSVM).



Σχήμα 4.18: Εντοπισμός ακραίων τιμών μέσω IQR
(Πηγή: https://www.statology.org/how-to-read-box-plot-with-outliers/)

4.5.3.2 Τεχνικές διαχείρισης ακραίων τιμών (Handling Outliers Techniques)

- **Αφαίρεση (Removal):** Η διαδικασία αυτή μπορεί να επιτευχθεί είτε ορίζοντας κάποιο κατώφλι, είτε με χρήση των μεθόδων συσταδοποίησης.
- **Μετασχηματισμός (Transformation):** Ο μετασχηματισμός των δεδομένων μπορεί να γίνει είτε με χρήση μεθόδων κλιμάκωσης (scaling), είτε αντικαθιστώντας τα με την πιο κοντινή σε αυτά τιμή που δε θεωρείται όμως ακραία, είτε μετασχηματίζοντας τα δεδομένα με χρήση λογαρίθμου.
- **Εκτίμηση ισχύς (Robust Estimation):** Η τεχνική αυτή εμπεριέχει τη χρήση αλγορίθμων που δείχνουν ανθεκτικότητα στην παρουσία ακραίων τιμών όπως οι robust regression, M-estimators και DBSCAN.

4.5.4 Διαχωρισμός συνόλου δεδομένων (Data Split)

Ο διαχωρισμός του συνόλου των δεδομένων αποτελεί ένα από τα βασικά βήματα του σταδίου της προεπεξεργασίας των δεδομένων κατά την οποία το αρχικό σύνολο δεδομένων χωρίζεται σε σύνολο δεδομένων εκπαίδευσης (training dataset) και σύνολο δεδομένων ελέγχου (test dataset).

Σκοπός της διαδικασίας αυτής είναι η αποφυγή του προβλήματος υπερπροσαρμογής (overfitting) του μοντέλου στα δεδομένα, καθώς η απόδοση του εκάστοτε μοντέλου αξιολογείται κάθε φορά από διαφορετικό σύνολο δεδομένων από αυτό που χρησιμοποιήθηκε για την εκμάθηση του αλγορίθμου. Αξίζει να σημειωθεί ότι ο διαχωρισμός πραγματοποιείται με τυχαίο τρόπο προς αποφυγή τυχόν μεροληψίας σε κάποιο συγκεκριμένο υποσύνολο. Πάντα το μεγαλύτερο μέρος των δεδομένων χρησιμοποιείται για την εκπαίδευση του αλγορίθμου με τις συνήθεις αναλογίες να είναι 70% / 30%, 75% / 25% και 80% / 20%.

4.5.5 Τεχνικές επαναδειγματοληψίας (Resampling Techniques)

Οι τεχνικές επαναδειγματοληψίας είναι στατιστικές τεχνικές που χρησιμοποιούνται για την παραγωγή νέων σημείων δεδομένων σε ένα σύνολο δεδομένων επιλέγοντας τυχαία σημεία δεδομένων από το ήδη υπάρχον σύνολο δεδομένων. Οι τεχνικές αυτές βοηθάνε στη δημιουργία ενός συνθετικού συνόλου δεδομένων για τη εκπαίδευση των αλγορίθμων μηχανικής μάθησης. Με τον τρόπον αυτό αντιμετωπίζεται το πρόβλημα ανομοιόμορφων κλάσεων που παρουσιάζεται συχνά στα σύνολα δεδομένων. Χαρακτηρίζεται ως ανισόρροπη κατανομή κλάσεων (imbalanced class distribution) και πρακτικά σημαίνει ότι οι παρατηρήσεις μιας κλάσης είναι πολύ περισσότερες ή πολύ λιγότερες από τις υπόλοιπες κλάσεις.

Οι πιο κοινές τεχνικές επαναδειγματοληψίας είναι η διασταυρούμενη επικύρωση και η μέθοδος bootstrap, η οποία και θα αναλυθεί. (Η διασταυρούμενη μάθηση σχολιάστηκε στην παράγραφο 4.4.3). Το bootstrap είναι μία τεχνική επαναδειγματοληψίας κατά την οποία χρησιμοποιούνται δεδομένα από ένα δείγμα με σκοπό την παραγωγή μιας κατανομής δειγματοληψίας, παίρνοντας επαναλαμβανόμενα τυχαία δείγματα από το αρχικό δείγμα, με επανατοποθέτηση. Χρησιμοποιείται κυρίως για τον υπολογισμό της εκτίμησης της τυπικής απόκλισης των εκτιμητών και κατ' επέκταση για τον υπολογισμό διαστημάτων εμπιστοσύνης. Στα θετικά τη μεθόδου bootstrap είναι η ευκολία εκτέλεσης της διαδικασίας καθώς βασίζεται σε καμία υπόθεση για την κατανομή του πληθυσμού.

Μια ακόμη δημοφιλής τεχνική επαναδειγματοληψίας είναι η Τεχνική Συνθετικής Υπερδειγματοληψίας Μειονότητας (Synthetic Minority Over-sampling Technique, SMOTE),

κατά την οποία παράγονται νέα συνθετικά δείγματα της μειονοτικής κλάσης. Επικεντρώνεται στον χώρο των χαρακτηριστικών για να παράξει τα νέα παραδείγματα με χρήση παρεμβολής μεταξύ των ήδη υπαρχόντων παραδειγμάτων. Συμβάλλει έτσι στην εξισορρόπηση του πληθυσμού των διαφορετικών κλάσεων ώστε να αποφευχθεί τυχόν μεροληψία του αλγορίθμου ως προς την πλειοψηφική κλάση γεγονός που θα επηρέαζε τα αποτελέσματα πρόβλεψης.

4.5.6 Κλιμάκωση και κανονικοποίηση δεδομένων (Data scaling and normalization)

Η κλιμάκωση των δεδομένων είναι μία τεχνική που χρησιμοποιείται στα προπαρασκευαστικά στάδια επεξεργασίας των δεδομένων στη μηχανική μάθηση και αποσκοπεί στον χειρισμό της μεγάλης διασποράς που μπορεί να παρουσιάζουν οι τιμές των μεταβλητών. Αν δεν πραγματοποιηθεί κλιμάκωση των δεδομένων τότε οι αλγόριθμοι τείνουν να δίνουν μεγαλύτερο βάρος στις υψηλές τιμές και να θεωρούν τις μικρότερες τιμές λιγότερο σημαντικές ανεξαρτήτως της μονάδας μέτρησης που χρησιμοποιείται κάθε φορά. Η διαδικασία της κλιμάκωσης εγγυάται ότι όλα τα χαρακτηριστικά είναι συγκρίσιμα μεταξύ τους και διαθέτουν συγκρίσιμα εύρη τιμών. Κατά αυτόν τον τρόπο βελτιώνεται η απόδοση των αλγορίθμων. Ακόμη, αποφεύγεται η αριθμητική αστάθεια και εξασφαλίζεται ότι όλα τα χαρακτηριστικά λαμβάνονται ισάξια υπόψιν κατά τη διάρκεια της διαδικασίας εκμάθησης.

Αφού διαφοροποιείται η αρχική κλίμακα των χαρακτηριστικών ώστε να γίνουν συγκρίσιμα εμπίπτουν τελικώς όλα σε ένα συγκεκριμένο εύρος τιμών μεταξύ 0 και 1 ή μεταξύ -1 και 1.

Οι συνήθεις τεχνικές κλιμάκωσης και κανονικοποίησης παρουσιάζονται παρακάτω:

- **Απόλυτη μέγιστη κλιμάκωση (Absolute Maximum Scaling):** Αυτή η μέθοδος κλιμακώνει τα δεδομένα έτσι ώστε να έχουν εύρος τιμών μεταξύ -1 και 1. Ωστόσο, δε χρησιμοποιείται συχνά γιατί είναι αρκετά ευαίσθητη στην παρουσία ακραίων τιμών. Δίνεται από τον τύπο:

$$x_{scaled} = \frac{x_i - \max(|x|)}{\max(|x|)}$$

- **Κλιμάκωση min-max (Min-Max Scaling):** Η μέθοδος αυτή είναι επίσης ευαίσθητη στις ακραίες τιμές αφού χρησιμοποιούνται η μέγιστη και η ελάχιστη παρατήρηση. Κλιμακώνει τα δεδομένα έτσι ώστε να έχουν εύρος τιμών μεταξύ 0 και 1. Δίνεται από τον τύπο:

$$x_{scaled} = \frac{x_i - \min(x)}{\max(x) - \min(x)}$$

- **Δεκαδική κλιμάκωση (Decimal Scaling):** Αυτή η μέθοδος κλιμακώνει τα δεδομένα χρησιμοποιώντας μετατόπιση του δεκαδικού σημείου των τιμών των δεδομένων. Τα δεδομένα θα έχουν εύρος τιμών μεταξύ -1 και 1.

- **Κανονικοποίηση Z-score (Z-score normalization/standardization):** Αυτή η τεχνική κανονικοποίησης μετασχηματίζει τα δεδομένα έτσι ώστε να έχουν μέση τιμή ίση με 0 και τυπική απόκλιση ίση με 1. Με την κανονικοποίηση των δεδομένων πραγματοποιείται ευκολότερα και με περισσότερη ακρίβεια η σύγκριση χαρακτηριστικών με διαφορετικά εύρη τιμών. Δίνεται από τον τύπο:

$$x_{norm} = \frac{x - mean(x)}{std(x)}$$

4.6 Νευρωνικά Δίκτυα (Neural Networks)

Τα νευρωνικά δίκτυα αποτελούν ομάδα αλγορίθμων της μηχανικής μάθησης. Είναι υπολογιστικά μοντέλα που λαμβάνουν αποφάσεις μιμούμενα τις περίπλοκες λειτουργίες του ανθρώπινου εγκεφάλου. Πιο συγκεκριμένα, μιμούνται τον τρόπο λειτουργίας των βιολογικών νευρώνων για την εξακρίβωση φαινομένων, τη στάθμιση επιλογών και την εξαγωγή συμπερασμάτων. Ιστορικά, τα νευρωνικά δίκτυα ξεκίνησαν σε πρώιμη μορφή τις δεκαετίες του ’40 και του ’50 με την εμφάνιση του πρώτου μαθηματικού μοντέλου τεχνητών νευρώνων από τους McCulloch και Pitts, όμως οι υπολογιστικοί περιορισμοί δυσκόλεψαν την εξέλιξή τους. Τη δεκαετία του ’80 έγινε εφικτή η εκμάθηση δικτύων με πολλαπλές συστάδες (Multi-layer network) λόγω της εφεύρεσης της μεθόδου οπισθοδιάδοσης (backpropagation method) από τους Rumelhart, Hinton και Williams. Έπειτα από μια δεκαετία, που χαρακτηρίστηκε ως χειμώνας για την εξέλιξη των νευρωνικών δικτύων, παρατηρείται τη δεκαετία του 2000 επάνοδος όσον αφορά τη διαχείριση μεγάλων συνόλων δεδομένων, τις καινοτόμες δομές και την ενισχυμένη ικανότητα επεξεργασίας και έτσι η βαθιά μάθηση (Deep learning) επιδεικνύει εξαιρετική αποτελεσματικότητα σε πληθώρα τομέων. Από το 2010 μέχρι σήμερα δύο είναι οι αρχιτεκτονικές βαθιάς μάθησης που κυριαρχούν στη μηχανική μάθηση, τα συνελικτικά νευρωνικά δίκτυα (Convolutional Neural Networks - CNN) και τα επαναλαμβανόμενα νευρωνικά δίκτυα (Recurrent Neural Networks - RNN) με την ισχύ τους να επιδεικνύεται χάρη σε καινοτομίες στον τομέα των ηλεκτρονικών παιχνιδιών, της αναγνώρισης εικόνας και της επεξεργασίας φυσικής γλώσσας.

Όσον αφορά τη δομή τους, τα νευρωνικά δίκτυα αποτελούνται από στρώματα (layers) κόμβων συνδεδεμένων μεταξύ τους, γνωστά και ως νευρώνες (neurons). Συγκεκριμένα, υπάρχει το στρώμα εισόδου (input layer) στο οποίο κάθε χαρακτηριστικό του συνόλου δεδομένων αντιπροσωπεύεται από έναν κόμβο ο οποίος λαμβάνει τα εισερχόμενα δεδομένα. Επιπλέον, καίριο ρόλο στη δομή των νευρωνικών δικτύων έχουν τα βάρη (weights) και οι συνδέσεις (connections). Το βάρος κάθε συνδεδεμένου νευρώνα υποδηλώνει πόσο ισχυρή είναι η σύνδεση και αυτό διαφοροποιείται κατά τη διάρκεια της διαδικασίας εκμάθησης. Στη συνέχεια, τα κρυμμένα στρώματα (hidden layers) επεξεργάζονται τα δεδομένα εισόδου πολλαπλασιάζοντάς τα με τα αντίστοιχα βάρη, προσθέτοντάς τα και δίνοντάς τα σε μία προκαθορισμένη συνάρτηση ενεργοποίησης. Με τη διαδικασία αυτή παρουσιάζεται μη γραμμικότητα στα δεδομένα η οποία βοηθάει το δίκτυο να αναγνωρίσει πολύπλοκα μοντέλα. Το τελικό αποτέλεσμα παράγεται επαναλαμβάνοντας την παραπάνω διαδικασία μέχρι το στρώμα εξόδου (Output layer).

Όπως είναι αντιληπτό τα νευρωνικά δίκτυα χρησιμοποιούνται σε ευρύ φάσμα προβλημάτων όπως η αυτόματη οδήγηση, η μηχανική, η ρομποτική, η αεροδιαστημική βιομηχανία και οι τηλεπικοινωνίες.

4.6.1 Νευρωνικά Δίκτυα Πρόσθιας Τροφοδότησης (Feedforward Neural Networks, FNN)

Τα νευρωνικά δίκτυα πρόσθιας τροφοδότησης (Feedforward NN), γνωστά και ως πολυστρωματικά (ή πολυεπίπεδα) νευρωνικά δίκτυα (multilayer perceptron - MLP) είναι η πιο απλή μορφή νευρωνικών δικτύων αποτελούμενα από στρώματα (ή επίπεδα - layers) συνδεδεμένων κόμβων. Στα FNN οι κόμβοι δε σχηματίζουν βρόχους και επομένως όλη η πληροφορία ρέει μόνο προς μία κατεύθυνση. Αυτό σημαίνει ότι δεν υπάρχουν συνδέσεις που να επιτρέπουν στην πληροφορία που αφήνει το στρώμα εξόδου του νευρωνικού δικτύου να σταλεί πίσω στο δίκτυο. Αυτού του τύπου τα νευρωνικά δίκτυα αποτελούνται από ένα στρώμα εισόδου, ένα στρώμα εξόδου και έναν ή περισσότερους κρυφούς νευρώνες. Η έξοδος κάθε νευρώνα σε ένα στρώμα καθορίζεται από το σταθμισμένο άθροισμα των εισόδων του προηγούμενου στρώματος το οποίο εισέρχεται πρώτα σε μία συνάρτηση ενεργοποίησης.

Για να εξαχθεί το επιθυμητό αποτέλεσμα θα πρέπει οι νευρώνες να είναι κατάλληλα συνδεδεμένοι μεταξύ τους με συνδέσμους καθένας από τους οποίους θα πρέπει να έχει κατάλληλο βάρος (ισχύ σύναψης). Οι νευρώνες θα πρέπει επίσης να έχουν κατάλληλο κατώφλι ενεργοποίησης (bias). Τα βάρη και τα κατώφλια ενεργοποίησης καθορίζονται κατά τη διάρκεια της εκμάθησης του νευρωνικού δικτύου με χρήση κατάλληλου αλγορίθμου.

Η μαθηματική φόρμουλα του FNN θεωρώντας ότι αυτό διαθέτει ένα μόνο κρυφό στρώμα δίνεται παρακάτω ως εξής:

$$y = f(W_2 * f(W_1 x + b_1) + b_2)$$

Όπου $x \in R^n$ είναι το διάνυσμα εισόδου, με n τον αριθμό των χαρακτηριστικών του συνόλου δεδομένων. Το y είναι το διάνυσμα εξόδου, W_1 και W_2 είναι οι πίνακες των βαρών και b_1, b_2 είναι τα διανύσματα των κατωφλίων ενεργοποίησης.

Όσον αφορά το κρυμμένο στρώμα η είσοδός του δίνεται από την εξίσωση:

$$z_1 = W_1 x + b_1$$

Εφαρμόζοντας την κατάλληλη συνάρτηση ενεργοποίησης στην είσοδο του κρυμμένου νευρώνα παίρνουμε το παρακάτω αποτέλεσμα:

$$a_1 = f(z_1)$$

Επομένως, η είσοδος του στρώματος εξόδου θα δίνεται από την εξίσωση:

$$z_2 = W_2 a_1 + b_2$$

Και εφαρμόζοντας εκ νέου συνάρτηση ενεργοποίησης η οποία μπορεί να είναι ίδια ή διαφορετική από την προηγούμενη λαμβάνουμε το τελικό y:

$$y = g(z_2) = f(W_2 * f(W_1 x + b_1) + b_2)$$

Οι συναρτήσεις ενεργοποίησης είναι εξαιρετικά σημαντικές καθώς μέσω αυτών καθορίζεται αν ένας νευρώνας θε πρέπει να ενεργοποιηθεί ή όχι υπολογίζοντας τα σταθμισμένα αθροίσματα και προσθέτοντας κάθε φορά το κατώφλι ενεργοποίησης. Ο σκοπός της χρήσης συνάρτησης ενεργοποίησης είναι να εισάγει τη μη γραμμικότητα στο αποτέλεσμα του νευρώνα. Συνεπώς, η χρήση της συνάρτησης ενεργοποίησης είναι καθοριστική για το

νευρωνικό δίκτυο αφού χωρίς αυτήν το νευρωνικό δίκτυο είναι απλά ένα μοντέλο γραμμικής παλινδρόμησης. Η συνάρτηση ενεργοποίησης πραγματοποιεί ένα μη γραμμικό μετασχηματισμό στην είσοδο του νευρώνα καθιστώντας τον ικανό να μάθει και να εκτελέσει πιο περίπλοκες εργασίες.

Οι συχνότερα χρησιμοποιούμενες συναρτήσεις ενεργοποίησης παρατίθενται παρακάτω:

- **Σιγμοειδής συνάρτηση (Sigmoid function):** Η σιγμοειδής συνάρτηση ενεργοποίησης είναι μια συνεχής διαφοροποιήσιμη συνάρτηση η οποία λαμβάνει πραγματικές τιμές μετασχηματίζοντάς τες ώστε να ανήκουν σε διάστημα μεταξύ 0 και 1. Τα αποτελέσματά της μπορούν εύκολα να ερμηνευθούν ως ποσοστά και έτσι χρησιμοποιείται σε προβλήματα δυαδικής κατηγοριοποίησης. Δίνεται από τον κάτωθι τύπο:

$$f(x) = \frac{1}{1 + e^{-x}}$$

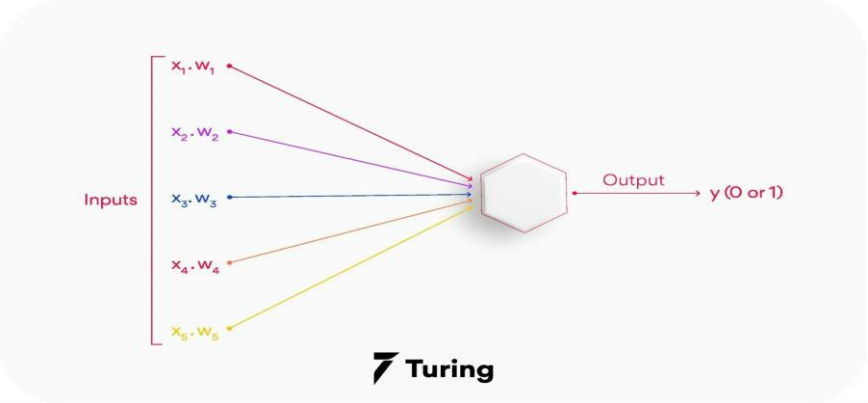
- **Συνάρτηση υπερβολικής εφαπτομένης (Hyperbolic tangent):** Η συνάρτηση υπερβολικής εφαπτομένης έχει πεδίο τιμών μεταξύ του -1 και του 1 και έτσι μπορεί να χειριστεί αποτελεσματικότερα αρνητικές τιμές εν αντιθέσει με την σιγμοειδή συνάρτηση που έχει πεδίο τιμών μεταξύ του 0 και του 1. Ακριβώς επειδή τα αποτελέσματα της συνάρτησης εφαπτομένης βρίσκονται μεταξύ των τιμών -1 και 1, παρουσιάζει ισχυρότερες κλίσεις από τη σιγμοειδή συνάρτηση πράγμα που συχνά οδηγεί σε ταχύτερη εκμάθηση και σύγκλιση του αλγορίθμου. Η συνάρτηση υπερβολικής εφαπτομένης συχνά χρησιμοποιείται ως συνάρτηση ενεργοποίησης στα κρυμμένα στρώματα του νευρωνικού δικτύου καθώς μπορεί να οδηγήσει σε πιο αποτελεσματική εκμάθηση όταν τα δεδομένα εισόδου είναι κανονικοποιημένα λόγω της συμμετρικότητας που παρουσιάζει ως προς το 0. Δίνεται από τον κάτωθι τύπο:

$$f(x) = \tanh(x) = \frac{e^{x} - e^{-x}}{e^{x} + e^{-x}}$$

- **Συνάρτηση διορθωμένης γραμμικής μονάδας (Rectified Linear Unit - ReLU):** Η συγκεκριμένη συνάρτηση ενεργοποίησης φράζει τα δεδομένα εισόδου στο 0 επιστρέφοντας 0 για κάθε αρνητική τιμή, διαφορετικά αν η τιμή εισόδου είναι θετική την επιστρέφει αυτούσια. Επομένως, για τις θετικές τιμές η συνάρτηση ReLU λειτουργεί σαν μια απλή γραμμική συνάρτηση διατηρώντας την κλίση της σταθερή και ίση με τη μονάδα. Η επιστροφή μηδενικής τιμής για κάθε αρνητική τιμή εισόδου οδηγεί σε σποραδική ενεργοποίηση των νευρώνων δηλαδή κάθε φορά ενεργοποιείται μόνο ένα υποσύνολο νευρώνων και έτσι επιτυγχάνονται πιο αποτελεσματικοί υπολογισμοί. Η συνάρτηση ReLU είναι υπολογιστικά οικονομική και αυτό επιτρέπει στα νευρωνικά δίκτυα να κλιμακώνουν σε πολλά στρώματα, χωρίς ιδιαίτερη αύξηση στο όριο πολυπλοκότητας των υπολογισμών, σε σύγκριση με άλλες πιο περίπλοκες συναρτήσεις ενεργοποίησης όπως η σιγμοειδής και η υπερβολική εφαπτομένη. Δίνεται από τον κάτωθι τύπο:

$$f(x) = \max(0, x)$$

- **Συνάρτηση ενεργοποίησης Softmax:** Η συνάρτηση ενεργοποίησης Softmax, γνωστή και ως κανονικοποιημένη εκθετική συνάρτηση, χρησιμοποιείται κυρίως σε προβλήματα ταξινόμησης πολλαπλών κλάσεων. Τα αποτελέσματα της συνάρτησης αυτής ακολουθούν κατανομή πιθανότητας που αθροίζει στη μονάδα. Κάθε στοιχείο του αποτελέσματος αναπαριστά την πιθανότητα το στοιχείο εισόδου να ανήκει σε συγκεκριμένη κλάση. Η softmax παρουσιάζει ευαισθησία σε ακραίες τιμές και μπορεί να οδηγήσει σε υπερπροσαρμογή του μοντέλου.



Σχήμα 4.19: Διαγραμματική απεικόνιση τη δομής ενός FNN
(Πηγή: https://www.turing.com/kb/mathematical-formulation-of-feed-forward-neural-network)

4.6.2 Συνελικτικά Νευρωνικά Δίκτυα (Convolutional Neural Networks, CNN)

Το συνελικτικό νευρωνικό δίκτυο (Convolutional Neural Network, CNN - CovNet) είναι τύπος βαθέος νευρωνικού δικτύου που χρησιμοποιείται κυρίως στην υπολογιστική (ή μηχανική ή τεχνητή) όραση (computer vision), ένα επιστημονικό πεδίο της τεχνητής νοημοσύνης το οποίο επιχειρεί να αναπαράγει αλγοριθμικά την αίσθηση της όρασης, συνήθως σε ηλεκτρονικό υπολογιστή ή ρομπότ. Στην ουσία επιτρέπει στον υπολογιστή να κατανοήσει και να ερμηνεύσει δεδομένα απεικόνισης και εικόνες.

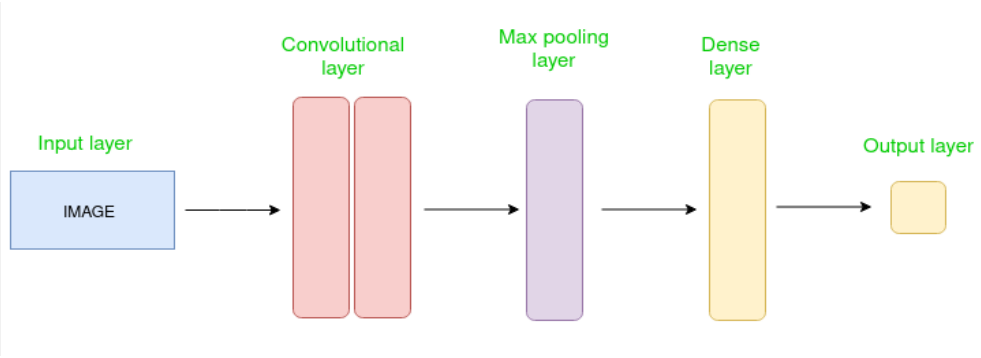
Τα συνελικτικά νευρωνικά δίκτυα έχουν πιο σύνθετη δομή από τα νευρωνικά δίκτυα πρόσθιας τροφοδότησης καθώς αποτελούνται από πολλαπλά στρώματα καθένα από τα οποία εκτελεί και διαφορετική διεργασία στα δεδομένα εισόδου. Πιο συγκεκριμένα, όσον αφορά τη δομή τα CNNs αποτελούνται από τρεις διαφορετικούς τύπους στρωμάτων. Αρχικά, υπάρχουν τα στρώματα συνελικτικού τύπου (convolutional layers) κατά τα οποία εφαρμόζονται διάφορα φίλτρα και καθένα από αυτά χρησιμοποιείται για την αναγνώριση ειδικών μοτίβων σε μία εικόνα όπως η καμπυλότητα των ψηφίων, οι κορυφές, το συνολικό σχήμα του ψηφίου μεταξύ άλλων. Κατόπιν, υπάρχει το στρώμα συγκέντρωσης (pooling layer) που στοχεύει στη συγκέντρωση των πιο σημαντικών χαρακτηριστικών εφαρμόζοντας συναρτήσεις συνάθροισης που μειώνουν τις διαστάσεις του πίνακα χαρακτηριστικών. Έτσι μειώνεται και η μνήμη που χρησιμοποιείται όσο εκπαιδεύεται το νευρωνικό δίκτυο. Χρησιμεύει, επίσης, και στη μείωση του προβλήματος υπερπροσαρμογής των δεδομένων. Οι συνήθεις χρησιμοποιούμενες συναρτήσεις συνάθροισης είναι οι εξής:

- max pooling: μέγιστη τιμή του χάρτη χαρακτηριστικών

- sum pooling: αντιστοιχεί στο άθροισμα όλων των τιμών του χάρτη χαρακτηριστικών
- average pooling: είναι ο μέσος όρος των τιμών του χάρτη χαρακτηριστικών

Τέλος, τα πλήρως συνδεδεμένα στρώματα (fully connected layers) αποτελούν το τελευταίο επίπεδο της δομής του συνελικτικού νευρωνικού δικτύου και η είσοδος τους αντιστοιχεί σε ένα συνεπτυγμένο πίνακα μίας διάστασης που παράχθηκε από το τελευταίο στρώμα συγκέντρωσης. Σε αυτά τα στρώματα εφαρμόζεται συνήθως η συνάρτηση ενεργοποίησης ReLU για την επίτευξη της μη γραμμικότητας. Στο τέλος το αποτέλεσμα των πλήρως συνδεδεμένων στρωμάτων εισέρχεται σε μια λογιστική συνάρτηση για εφαρμογές ταξινόμησης όπως είναι η συνάρτηση ενεργοποίησης softmax και η σιγμοειδής συνάρτηση οι οποίες μετατρέπουν το αποτέλεσμα για κάθε κλάση στο σκορ πιθανότητας της κάθε κλάσης.

Γενικά, τα συνελικτικά νευρωνικά δίκτυα μπορούν να χρησιμοποιηθούν με κατάλληλη δομή και συναρτήσεις ενεργοποίησης σε πληθώρα εφαρμογών όπως αναγνώριση προσώπου, αυτόνομη οδήγηση και ανάλυση ιατρικών εικόνων.



Σχήμα 4.20: Διαγραμματική απεικόνιση της δομής ενός CNN
(Πηγή: https://www.geeksforgeeks.org/introduction-convolution-neural-network/)

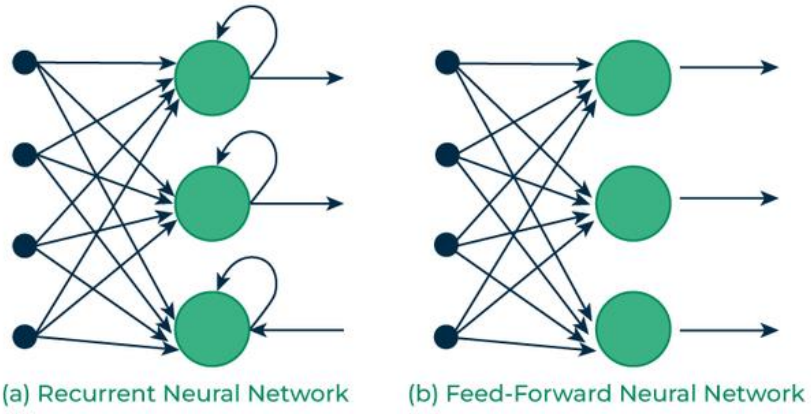
4.6.3 Επαναλαμβανόμενα Νευρωνικά Δίκτυα (Recurrent Neural Networks, RNN)

Τα επαναλαμβανόμενα νευρωνικά δίκτυα (Recurrent Neural Networks, RNN) είναι τύπος νευρωνικού δικτύου που λειτουργεί καλύτερα σε διαδοχικά δεδομένα όπως δεδομένα χρονοσειρών και δεδομένα κειμένου. Στο συγκεκριμένο τύπο νευρωνικού δικτύου το αποτέλεσμα του προηγούμενου βήματος παρέχεται ως είσοδος στο τρέχον βήμα. Αντίθετα, στα απλούστερα νευρωνικά δίκτυα τα δεδομένα εισόδου και εξόδου ήταν ανεξάρτητα το ένα από το άλλο. Ωστόσο, σε περιπτώσεις, παραδείγματος χάριν, πρόβλεψης της επόμενης λέξης σε μία πρόταση, απαιτούνται οι προηγούμενες λέξεις και επομένως υπάρχει ανάγκη απομνημόνευσης των λέξεων αυτών.

Γι’ αυτό κατασκευάστηκαν τα επαναλαμβανόμενα νευρωνικά δίκτυα που έλυσαν το άνωθεν πρόβλημα με τη βοήθεια ενός κρυμμένου στρώματος. Το κυριότερο και σημαντικότερο χαρακτηριστικό των RNNs είναι η κρυμμένη κατάσταση (Hidden State), η οποία απομνημονεύει πληροφορίες για τη σειρά. Αυτή η κατάσταση ονομάζεται και κατάσταση μνήμης (Memory State) μιας και απομνημονεύει την προηγούμενη είσοδο του δικτύου. Για κάθε είσοδο χρησιμοποιείται η ίδια παράμετρος αφού πραγματοποιείται η ίδια

εργασία σε όλες τις εισόδους ή τα κρυμμένα στρώματα ώστε να παραχθεί το αποτέλεσμα. Αυτό μειώνει την περιπλοκότητα της χρήσης διαφορετικών παραμέτρων εν αντιθέσει με άλλους τύπους νευρωνικών δικτύων.

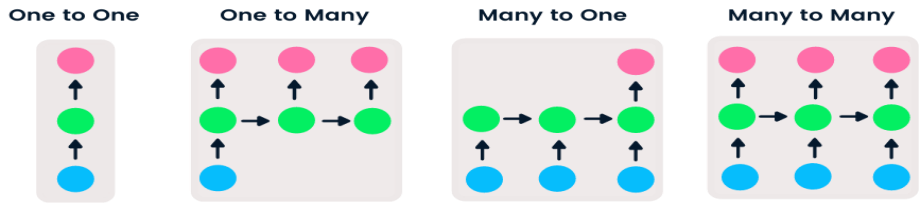
Όσον αφορά τη δομή τους, το στρώμα εισόδου επεξεργάζεται τα δεδομένα και στη συνέχεια τα προωθεί στα κρυμμένα στρώματα όπου καθένα από αυτά διαθέτει συναρτήσεις ενεργοποίησης, βάρη και κατώφλια ενεργοποίησης. Αυτές οι παράμετροι παραμένουν σταθερές σε όλα τα κρυμμένα στρώματα έτσι ώστε αντί να δημιουργηθούν πολλαπλά κρυμμένα στρώματα να δημιουργηθεί ένα επαναλαμβανόμενο. Τα RNNs δε χρησιμοποιούν την κλασσική μέθοδο οπισθοδιάδοσης αλλά χρησιμοποιούν αλγορίθμους οπισθοδιάδοσης μέσω χρόνου (Backpropagation Through Time, BPTT) για τον καθορισμό της κλίσης (gradient).



Σχήμα 4.21: Διαγραμματική απεικόνιση της διαφοράς της δομής των RNN και FNN
(Πηγή: https://www.geeksforgeeks.org/introduction-to-recurrent-neural-network/)

Υπάρχουν τέσσερις διαφορετικοί τύποι επαναλαμβανόμενων νευρωνικών δικτύων μιας και τα RNNs έχουν την ευελιξία να διαθέτουν διαφορετικά μήκη δεδομένων εισόδου και εξόδου ως εξής:

- **Ένα-προς-ένα (one-to-one):** Αυτό είναι απλό νευρωνικό δίκτυο. Χρησιμοποιείται συνήθως για προβλήματα μηχανικής μάθησης που έχουν μονή είσοδο και έξοδο.
- **Ένα-προς-πολλά (one-to-many):** Τέτοια RNN έχουν μονή είσοδο και πολλαπλές εξόδους και χρησιμοποιούνται συνήθως για την παραγωγή λεζάντας σε εικόνες.
- **Πολλά-προς-ένα (many-to-one):** Αυτά λαμβάνουν μία διαδοχή από πολλαπλές εισόδους και προβλέπουν ένα μοναδικό αποτέλεσμα. Είναι περισσότερο δημοφιλή σε ταξινόμηση δεδομένων που αφορούν σε γνώμη κοινού όπου η είσοδος είναι σε μορφή κειμένου και η έξοδος είναι μία κατηγορία.
- **Πολλά-προς-πολλά (many-to-many):** Αυτά λαμβάνουν πολλαπλές εισόδους και δίνουν πολλαπλά αποτελέσματα. Η πιο συχνή χρήση τους αφορά σε μετάφραση μηχανής.



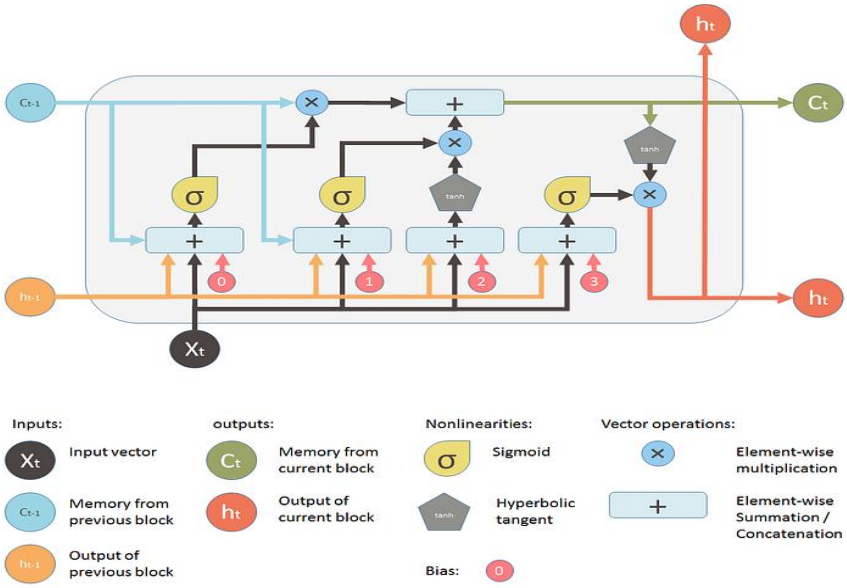
Σχήμα 4.22: Τύποι επαναλαμβανόμενων νευρωνικών δικτύων
(Πηγή: https://www.datacamp.com/tutorial/tutorial-for-recurrent-neural-network)

4.6.4 Δίκτυα Μακράς Βραχυπρόθεσμης Μνήμης (Long Short-Term Memory Networks, LSTM)

Τα δίκτυα μακράς βραχυπρόθεσμης μνήμης (Long Short-Term Memory Networks, LSTM) είναι μια βελτιωμένη εκδοχή των επαναλαμβανόμενων νευρωνικών δικτύων από τους Hochreiter και Schmidhuber. Τα παραδοσιακά RNNs διαθέτουν μονή κρυμμένη κατάσταση γεγονός που δημιουργεί δυσκολία στο δίκτυο να εκπαιδευτεί σε μακροπρόθεσμες εξαρτήσεις. Τα μοντέλα LSTM επιλύουν αυτό το πρόβλημα χρησιμοποιώντας ένα κελί μνήμης (memory cell), ένα δοχείο δηλαδή που μπορεί να διατηρήσει τις πληροφορίες για κάποιο χρονικό διάστημα. Ωστόσο, οι αρχιτεκτονικές των LSTM παρουσιάζουν τη δυνατότητα εκμάθησης μακροπρόθεσμων εξαρτήσεων και επομένως λειτουργούν αποτελεσματικά σε εργασίες όπως η μετάφραση γλώσσας, η αναγνώριση ομιλίας και η πρόβλεψη χρονοσειρών.

Όσον αφορά στη δομή τους, τα δίκτυα μακράς βραχυπρόθεσμης μνήμης περιέχουν το κελί μνήμης το οποίο ελέγχεται από τρεις πύλες, την πύλη εισόδου (input gate), την πύλη λήθης (forget gate) και την πύλη εξόδου (output gate) από το κελί μνήμης. Αναλυτικότερα, η πύλη εισόδου ελέγχει αν οι πληροφορίες που προστίθενται στο κελί μνήμης είναι χρήσιμες με τη βοήθεια μιας σιγμοειδούς συνάρτησης. Οι πληροφορίες που δεν προσφέρουν στο δίκτυο αφαιρούνται από το κελί μνήμης μέσω της πύλης λήθης. Τέλος, η πύλη εξόδου αναλαμβάνει να εξάγει χρήσιμη πληροφορία από το τρέχον κελί ώστε να παρουσιαστεί σαν αποτέλεσμα. Η παραπάνω διαδικασία επιτρέπει στα LSTM να κρατούν ή όχι πληροφορίες οι οποίες επιτρέπουν την εκμάθηση μακροπρόθεσμων εξαρτήσεων. Τέλος, το LSTM διαθέτει μια κρυμμένη κατάσταση η οποία δρα σαν βραχυπρόθεσμη μνήμη για το δίκτυο. Αυτή η κρυμμένη κατάσταση ενημερώνεται βάσει της εισόδου, της προηγούμενης κρυμμένης κατάστασης και της τρέχουσας κατάστασης του κελιού μνήμης.

Εν κατακλείδι, τα δίκτυα LSTM αποτελούν εξαιρετικά χρήσιμα εργαλεία και βρίσκουν εφαρμογή σε πολλές εργασίες όπως η μοντελοποίηση φυσικής γλώσσας, η αναγνώριση ομιλίας, η ανάλυση βίντεο και η ανίχνευση ανωμαλιών και ακραίων τιμών.



Σχήμα 4.23: Διαγραμματική απεικόνιση δομής LSTM
Πηγή: https://blog.mlreview.com/understanding-lstm-and-its-diagrams-37e2f46f1714

ΚΕΦΑΛΑΙΟ 5ο

Εφαρμογές

5.1 Στόχος Ανάλυσης

Η ανάλυση δεδομένων της Formula 1 στοχεύει στην πρόβλεψη της θέσης τερματισμού των οδηγών σε κάθε αγώνα. Πηγή εύρεσης των δεδομένων είναι η ιστοσελίδα Kaggle.com (https://www.kaggle.com/datasets/rohanrao/formula-1-world-championship-1950-2020/data?select=circuits.csv), η οποία περιέχει πολλά διαφορετικά σύνολα δεδομένων (datasets) σε μορφή αρχείων CSV, που συλλέγονται από την ιστοσελίδα http://ergast.com/mrd/, αφορούν στους αγώνες της F1 από το 1950 μέχρι το 2024 και ανανεώνονται συνεχώς.

Για την πρόβλεψη των θέσεων τερματισμού των οδηγών χρησιμοποιήθηκαν αλγόριθμοι μηχανικής μάθησης και νευρωνικά δίκτυα. Συγκεκριμένα έγινε χρήση συλλογικών μεθόδων βασισμένων σε δένδρα απόφασης, μεθόδων παλινδρόμησης και τριών διαφορετικών νευρωνικών δικτύων. Όλα τα στάδια της ανάλυσης πραγματοποιήθηκαν με χρήση της γλώσσας προγραμματισμού python και το περιβάλλον ανοιχτού κώδικα Jupyter Notebook 7.0.8 στο οποίο αποκτήθηκε πρόσβαση από το περιβάλλον του Anaconda Navigator.

5.2 Παρουσίαση Δεδομένων

Τα αρχικά σύνολα δεδομένων που λήφθηκαν από την ιστοσελίδα Kaggle.com είναι τα εξής:

- circuits.csv (χαρακτηριστικά των circuits)
- constructor_results.csv (κατάσταση αποτελέσματος κατασκευαστών)
- constructor_standings.csv (κατάταξη κατασκευαστών)
- constructors.csv (πληροφορίες κατασκευαστών)
- driver_standings.csv (κατάταξη οδηγών)
- drivers.csv (πληροφορίες οδηγών)
- lap_times.csv (χρόνοι γύρων αγώνα)
- pit_stops.csv (πληροφορίες για pit stops)
- qualifying.csv (πληροφορίες κατατακτήριων δοκιμών)
- races.csv (πληροφορίες τοποθεσίας και χρόνου αγώνα)
- results.csv (αποτελέσματα αγώνων)
- seasons.csv (χρονιές)
- sprint_results.csv (αποτελέσματα αγώνα sprint)
- status.csv (κατάσταση μονοθέσιου κατά τον τερματισμό)

Το τελικό σύνολο δεδομένων που χρησιμοποιήθηκε για την πρόβλεψη της θέσης τερματισμού κάθε οδηγού διαμορφώθηκε από τη ένωση των συνόλων constructors.csv, drivers.csv, qualifying.csv, races.csv, results.csv και status.csv. Χρησιμοποιήθηκε, επίσης, το σύνολο δεδομένων constructor_standings.csv για την κατασκευή διαγράμματος. Πριν την

συγχώνευση των συνόλων δεδομένων πραγματοποιήθηκε έλεγχος για την ύπαρξη διπλότυπων εγγραφών (duplicates) και κατόπιν έγινε αλλαγή των ονομάτων μερικών χαρακτηριστικών στα σύνολα δεδομένων, λόγω του ότι υπάρχουν σε διαφορετικά σύνολα δεδομένων στήλες με την ίδια ονομασία, που όμως περιέχουν διαφορετική πληροφορία. Οι στήλες που μετονομάστηκαν σε κάθε dataset παρουσιάζονται στον παρακάτω πίνακα:

Όνομα Συνόλου Δεδομένων	**Αρχική Ονομασία Στήλης**	**Τελική Ονομασία Στήλης**
drivers.csv	number	driver_number
	surname	driver_name
	nationality	driver_nationality
constructors.csv	name	constructors_name
	nationality	constructor_nationality
qualifying.csv	number	car_number
	position	qualifying_position
races.csv	date	GP_date
	time	GP_time
	name	circuit_name
results.csv	number	car_number
	grid	starting_position
	position	final_position
	rank	fastestLapRank
	points	drivers_points

Πίνακας 5.1: Μετονομασία στηλών από τα αρχικά σύνολα δεδομένων

Η συγχώνευση των έξι συνόλων δεδομένων έγινε με χρήση της εντολής merge και των κοινών τους στηλών, resultId, raceId, driverId, constructorId, circuitId, statuId, qualifyId οι οποίες έπαιξαν το ρόλο των πρωτευόντων κλειδιών.

Στο προκύπτον σύνολο δεδομένων περιέχονται μεταξύ άλλων, εκτός των στηλών τύπου Id, οι κάτωθι μεταβλητές οι οποίες είναι σημαντικές για την ανάλυσή μας:

Όνομα	**Περιγραφή**
starting_position	Θέση οδηγού στη σχάρα εκκίνησης
final_position	Θέση τερματισμού οδηγού
positionText	Θέση τερματισμού οδηγού αποτυπωμένη με αριθμούς ή κείμενο
positionOrder	Τελική κατάταξη οδηγού μετά τον αγώνα
drivers_points	Κερδισμένοι πόντοι του οδηγού βάσει της θέσης τερματισμού για κάθε αγώνα
laps	Αριθμός γύρων που ολοκλήρωσε ο οδηγός σε κάθε αγώνα
time	Χρόνος πρώτου οδηγού και χρονική διαφορά από τους υπόλοιπους οδηγούς
milliseconds	Χρονική διάρκεια ολοκλήρωσης αγώνα για κάθε οδηγό δοσμένη σε milliseconds
fastestLapTime	Χρόνος ταχύτερου γύρου

fastestLapSpeed	Ταχύτητα ταχύτερου γύρου
driver_name	Όνομα οδηγού (μόνο επίθετο)
dob	Χρονολογία γεννήσεως οδηγούς
driver_nationality	Εθνικότητα οδηγού
qualifying_position	Θέση οδηγού στις κατατακτήριες δοκιμές
q1	Χρόνοι γύρου οδηγών στον πρώτο γύρω κατατακτήριων δοκιμών
q2	Χρόνοι γύρου οδηγών στον δεύτερου γύρω κατατακτήριων δοκιμών
q3	Χρόνοι γύρου οδηγών στον τρίτο γύρω κατατακτήριων δοκιμών
year	Έτος αγώνα
round	Αριθμός αγώνα βάσεις calendar
circuit_name	Όνομα πίστας
constructors_name	Όνομα κατασκευαστή
GP_date	Ημερομηνία διεξαγωγής GP
status	Κατάσταση οδηγού/μονοθέσιου πριν, κατά τη διάρκεια ή στο τέλος του αγώνα
constructor_nationality	Εθνικότητα κατασκευαστή

Πίνακας 5.2: Παρουσίαση μεταβλητών τελικού συνόλου δεδομένων

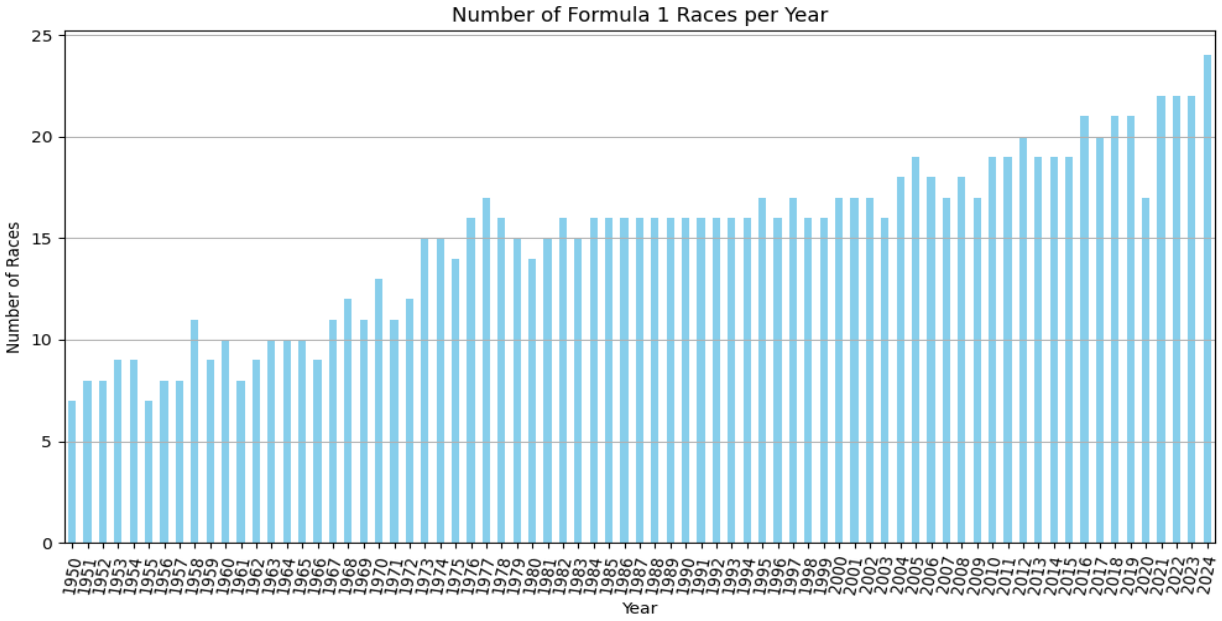
Ωστόσο, πολλές από τις παραπάνω μεταβλητές δε μπορούν να χρησιμοποιηθούν αυτούσιες διότι περιέχουν πληροφορία η οποία προέρχεται από το μέλλον καθώς πολλά από τα χαρακτηριστικά του πίνακα 5.2 δεν είναι γνωστά πριν την έναρξη του αγώνα. Αναλυτικότερα, πριν την έναρξη κάθε GP μπορούμε να γνωρίζουμε μόνο πληροφορίες όπως τα ονόματα των οδηγών, των κατασκευαστών και των circuit, την ημερομηνία διεξαγωγής του αγώνα, την εθνικότητα των κατασκευαστών και των οδηγών, τον αριθμό του αγώνα στο καλεντάρι και το έτος. Γνωρίζουμε, επίσης, τους χρόνους των κατατακτήριων δοκιμών και τις θέσεις των οδηγών στη σχάρα εκκίνησης.

Όμως, αν δεν τερματίσει ο αγώνας δε είναι γνωστοί οι γύροι που ολοκλήρωσε ο κάθε οδηγός, αν πέτυχε ή όχι ταχύτερο γύρο, και αν ναι, ποια είναι τα χαρακτηριστικά αυτού, το χρόνο διάρκειας του αγώνα, την κατάσταση του οδηγού ή του μονοθέσιου, τους κερδισμένους πόντους και προφανώς την τελική κατάταξη των οδηγών. Επομένως, για τις μεταβλητές final_position, positionText, positionOrder, drivers_points, laps, time, milliseconds, fastestLapTime, fastestLapSpeed και status γνωρίζουμε για κάθε αγώνα μόνο τις πληροφορίες από τους προηγούμενους αγώνες και τα προηγούμενα έτη. Έτσι, για να προβλεφθεί η θέση τερματισμού σε κάθε αγώνα θα πρέπει να δημιουργηθούν νέες μεταβλητές που παρέχουν όσο το δυνατόν περισσότερη πληροφορία, η οποία συλλέγεται κάθε φορά από τους προηγούμενους αγώνες.

Η παραπάνω διαδικασία αποτελεί το τρίτο και τελευταίο στάδιο που θα διαμορφώσει το τελικό σύνολο δεδομένων το οποίο και θα χρησιμοποιηθεί για την εφαρμογή των αλγορίθμων μηχανικής μάθησης. Προηγούνται η διερευνητική ανάλυση των δεδομένων και η προπαρασκευή τους για την αντιμετώπιση ελλειπουσών ή και ακραίων τιμών και άλλων προβλημάτων.

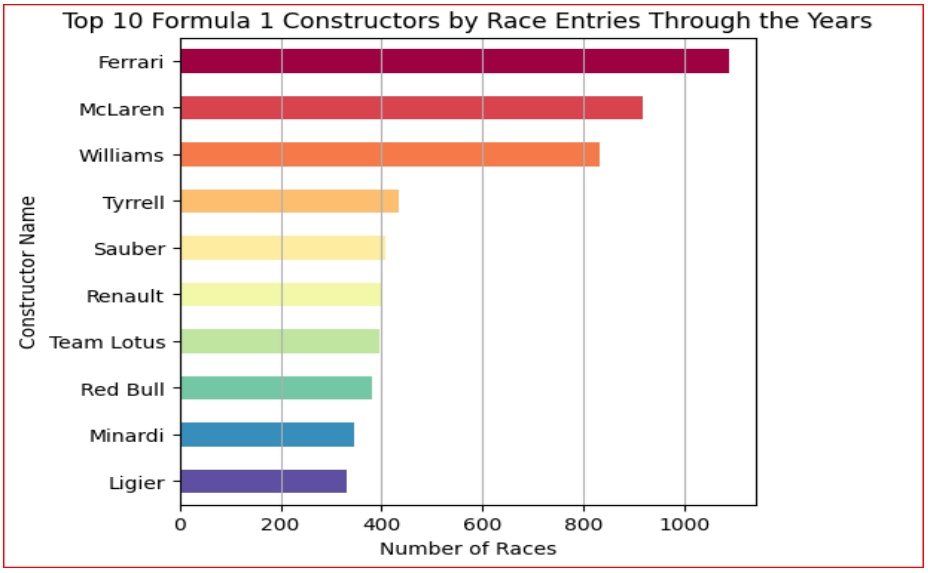
5.3 Διερευνητική Ανάλυση

Κατά τη διερευνητική ανάλυση κατασκευάστηκαν διαγράμματα που παρέχουν πληροφορίες σχετικά με τα ιστορικά στοιχεία της F1. Όπως, λοιπόν, φαίνεται στο παρακάτω ραβδόγραμμα, που περιλαμβάνει τον αριθμό των αγώνων για κάθε έτος από το 1950 μέχρι και το 2024, με την πάροδο των χρόνων ο αριθμός των GPs έχει αυξηθεί σημαντικά. Συγκεκριμένα, μέχρι και το 1966 ο αριθμός των αγώνων δεν ξεπερνά τους δέκα με εξαίρεση το έτος 1958. Ωστόσο, από το 2016 και έπειτα ο αριθμός των GPs ξεπερνά τα είκοσι ανά έτος με εξαίρεση το έτος 2020 κατά το οποίο πολλοί αγώνες ακυρώθηκαν λόγω της πανδημίας COVID-19.



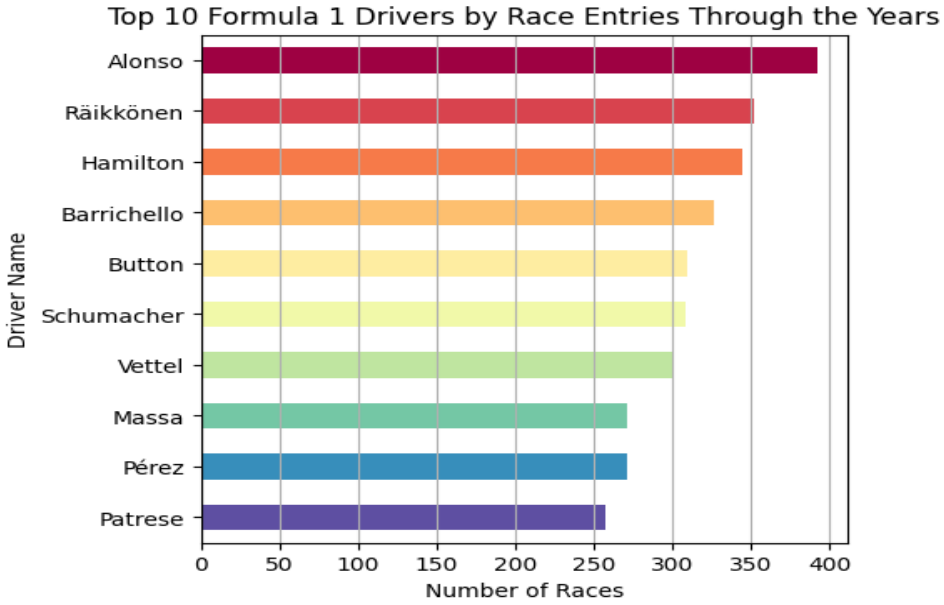
Σχήμα 5.1: Αριθμός αγώνων ανά έτος (1950-2024)

Αξίζει, επιπλέον, να σημειωθεί ότι λίγοι είναι οι κατασκευαστές που κατορθώνουν να συμμετέχουν σε αρκετές σεζόν της F1 εξαιτίας του εξαιρετικά υψηλού προϋπολογισμού που απαιτείται για την κατασκευή ανταγωνιστικού μονοθέσιου (βλ. Κεφ. 2). Η Ferrari είναι η μοναδική ομάδα που συμμετέχει σε κάθε πρωτάθλημα από το 1950 μέχρι και σήμερα έχοντας λάβει συμμετοχή σε πάνω από 1000 GPs όπως φαίνεται στο κάτωθι ραβδόγραμμα. Η δεύτερη μακροβιότερη ομάδα στην F1 είναι η McLaren με πάνω από 900 συμμετοχές και τρίτη είναι η Williams έχοντας πάνω από 800 συμμετοχές στους αγώνες.



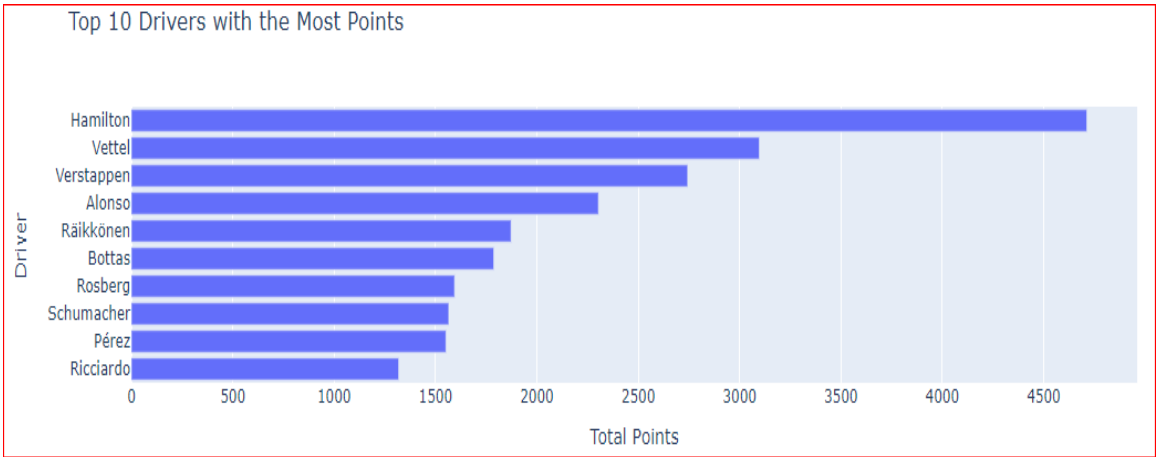
Σχήμα 5.2: Τοπ 10 κατασκευαστές με τις περισσότερες συμμετοχές σε αγώνες ανά τα χρόνια

Όσον αφορά τους οδηγούς, ο Fernando Alonso είναι ο οδηγός με τις περισσότερες συμμετοχές σε αγώνες στην ιστορία του αθλήματος, έχοντας συμμετάσχει σε πάνω από 350 με δεύτερο τον Kimi Räikkönen με οριακά πάνω από 350 αγώνες και τρίτο τον Lewis Hamilton ο οποίος πλησιάζει τις 350 συμμετοχές.



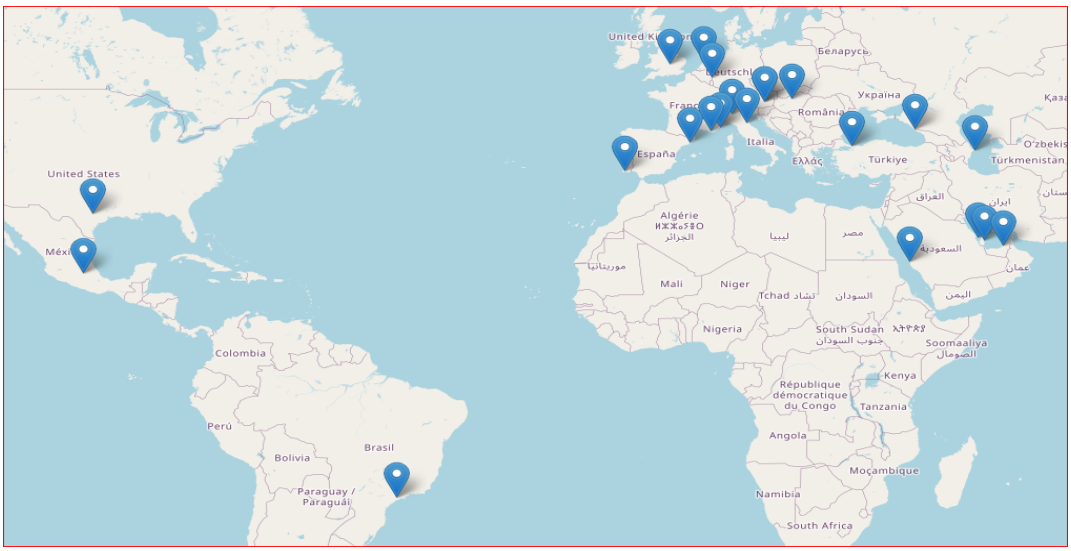
Σχήμα 5.3: Τοπ 10 οδηγοί με τις περισσότερες συμμετοχές σε αγώνες ανά τα χρόνια

Ο οδηγός με τους περισσότερους πόντους στην ιστορία της F1, σύμφωνα με το σχήμα 5.4 φαίνεται να είναι ο Lewis Hamilton έχοντας συγκεντρώσει μέχρι στιγμής πάνω από 4700 πόντους. Δεύτερος στην κατάταξη των οδηγών με τους περισσότερους πόντους βρίσκεται ο Sebastian Vettel με πάνω από 3000 πόντους και τρίτος ο Max Verstappen συγκεντρώνοντας μέχρι στιγμής πάνω από 2700 πόντους.



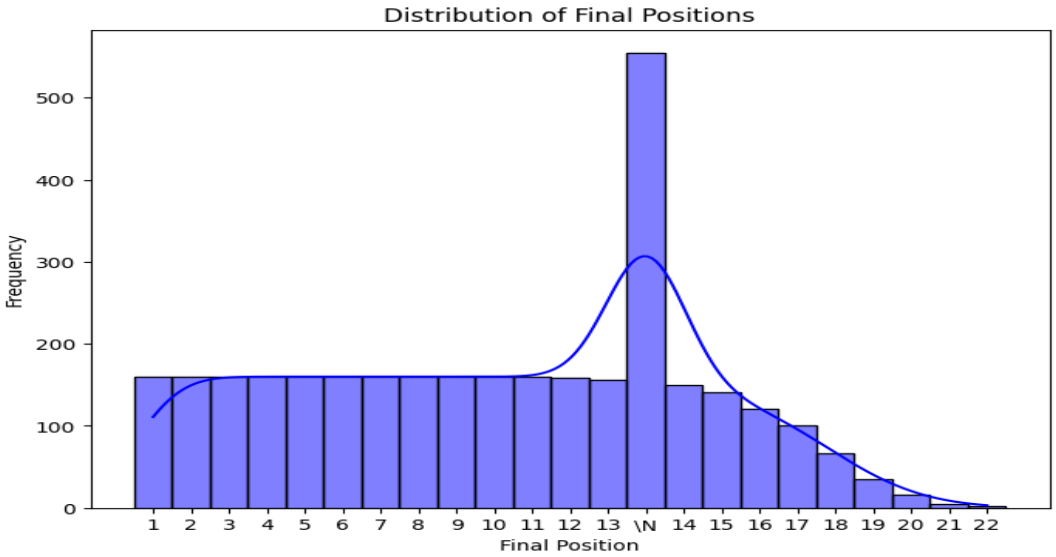
Σχήμα 5.4: Τοπ 10 οδηγοί με τους περισσότερους πόντους

Παρακάτω παρατίθεται, επίσης, χάρτης στον οποίο είναι επισημασμένα με πινέζες όλα τα GPs για το 2021.



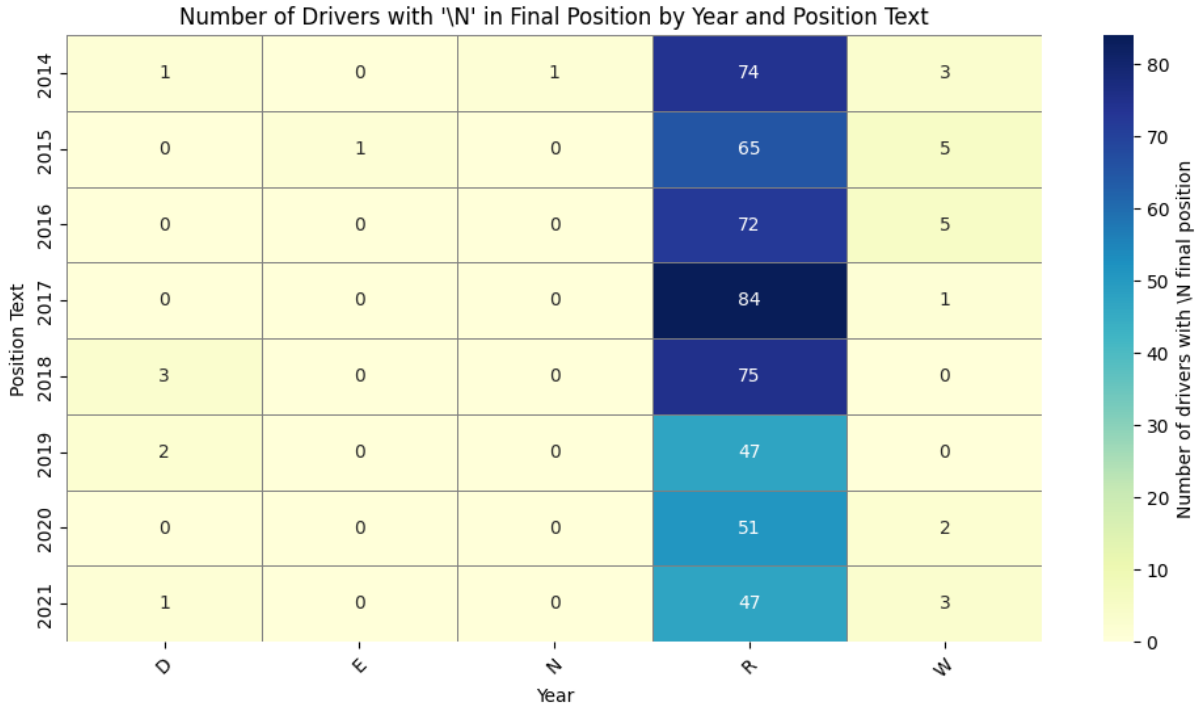
Σχήμα 5.5: Χάρτης με τις τοποθεσίες των GPs για το 2021

Μετά τη συγχώνευση των συνόλων δεδομένων επιλέγονται, ως περίοδος ανάλυσης, τα έτη από το 2014 μέχρι το 2021 μιας και αποτελούν την πιο σύγχρονη περίοδο χωρίς αξιοσημείωτες διαφοροποιήσεις των κανονισμών τόσο ως προς την κατασκευή των μονοθέσιων όσο και ως προς τη διεξαγωγή των αγώνων. Στο σχήμα 5.6 παρατίθεται η κατανομή των τελικών θέσεων βάσει της μεταβλητής final_position.



Σχήμα 5.6: Κατανομή των τελικών θέσεων

Από το διάγραμμα φαίνεται πως υπάρχουν πολλές κενές εγγραφές (\N) οι οποίες και πρέπει είτε να αφαιρεθούν είτε να αντικατασταθούν με κάποια μέθοδο ώστε να μη δημιουργήσουν πρόβλημα στα αποτελέσματα των αλγορίθμων. Ο τρόπος διαχείρισής τους αναλύεται στην επόμενη ενότητα (5.4).

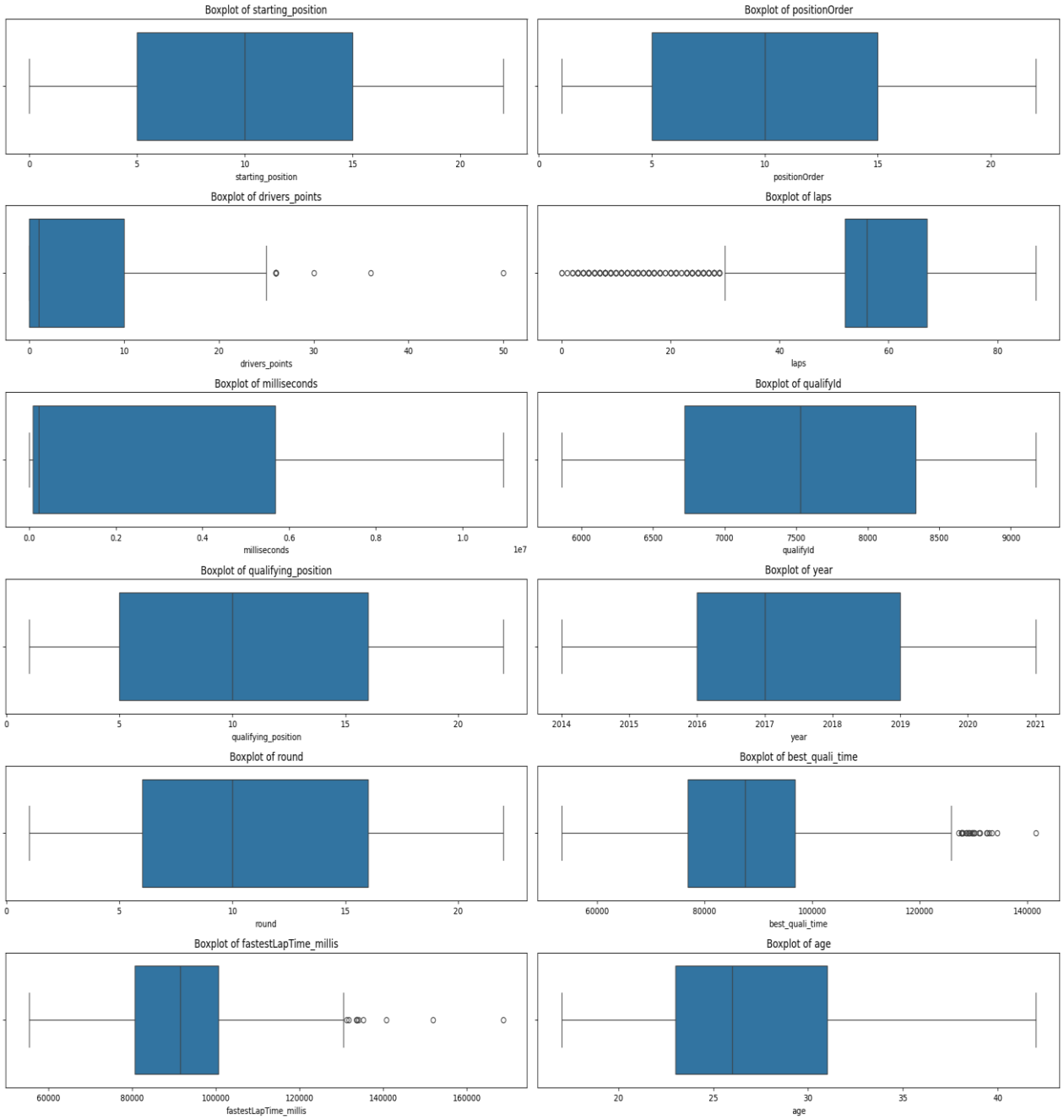


Σχήμα 5.7: Αριθμός οδηγών με κενή εγγραφή στη θέση τερματισμού ανά χρονιά και θέση σε μορφή κειμένου. (Η μεταβλητή positionText περιέχει θέσεις τερματισμού με αριθμούς για οδηγούς που τερμάτισαν επιτυχώς και θέσεις με γράμματα τα οποία υποδηλώνουν το λόγο που δεν τερμάτισε ο εκάστοτε οδηγός.)

Ακόμη από το σχήμα 5.7 παρατηρούμε ότι οι κενές εγγραφές στις τελικές θέσεις αφορούν κατά κύριο λόγο σε μονοθέσια που δεν έχουν τερματίσει καθώς έχουν αποσυρθεί πριν τη

λήξη του αγώνα είτε λόγω σύγκρουσης με άλλο μονοθέσιο ή με τα προστατευτικά κιγκλιδώματα, είτε άλλων τεχνικών και μηχανικών προβλημάτων.

Κατασκευάστηκαν, επίσης, για τις αριθμητικές μεταβλητές θηκογράμματα (boxplots) για τον εντοπισμό πιθανών ακραίων τιμών τα οποία και παρατίθενται στο κάτωθι σχήμα.



Σχήμα 5.8: Θηκογράμματα αριθμητικών μεταβλητών

Από το διάγραμμα φαίνεται πως για τις μεταβλητές starting_position, positionOrder, milliseconds, qualifying_position, year, round και age δεν παρατηρούνται ακραίες τιμές. Στη μεταβλητή drivers_points παρατηρήθηκαν ελάχιστες ακραίες τιμές. Αυτές είναι πολύ πιθανό

να προκύπτουν από λάθος καθώς δε γίνεται να έχει δοθεί βαθμολογία σε οδηγό μεγαλύτερη των 25 βαθμών. Ωστόσο, στις υπόλοιπες μεταβλητές δεν πραγματοποιήθηκε διαδικασία αφαίρεσης των ακραίων τιμών. Αυτό συνέβη επειδή όσον αφορά στις μεταβλητές που σχετίζονται με το χρόνο (best_quali_time, fastestLapTime_millis) δε μπορούμε να γνωρίζουμε από που προέρχονται. Εκτός από λάθος στην καταγραφή, μπορεί οι ακραίες διαφορές στο χρόνο να οφείλονται σε εξαιρετικά καλή απόδοση κάποιου μονοθέσιου με αποτέλεσμα να επιτύχει πολύ μικρότερο χρόνο γύρου σε σχέση με τα υπόλοιπα. Ακόμη οι καιρικές συνθήκες παίζουν σημαντικό ρόλο στο χρόνο γύρου καθώς σε περιπτώσεις έντονης βροχόπτωσης τα μονοθέσια κινούνται με πολύ μικρότερη ταχύτητα και επομένως ο χρόνος γύρου αυξάνεται σημαντικά. Όλα αυτά μπορεί να επηρεάσουν την έκβαση του αγώνα, γι' αυτό και προτιμήθηκε να μην αφαιρεθούν από το σύνολο δεδομένων λαμβάνοντας υπόψιν ότι μπορεί να επηρεαστούν τα αποτελέσματα των μοντέλων.

5.4 Προπαρασκευή Δεδομένων

Κατά τη διαδικασία προεπεξεργασίας των δεδομένων αφαιρέθηκαν, αρχικά, στήλες που δε προσφέρουν πληροφορία στην ανάλυση μας. Πιο συγκεκριμένα, οι στήλες που περιέχουν τον αριθμό του οδηγού και του μονοθέσιου δεν περιλαμβάνονται στο τελικό dataset αφού ο αριθμός που χαρακτηρίζει έναν οδηγό και κατ' επέκταση το μονοθέσιο που οδηγεί είναι μοναδικός για τον καθένα μόνο κατά τη διάρκεια της ίδιας σεζόν. Επιτρέπεται, δηλαδή, δύο διαφορετικοί οδηγοί να χρησιμοποιούν το ίδιο νούμερο αν δεν συμμετέχουν στην ίδια σεζόν της F1. Ακόμη, αφαιρούνται οι στήλες που περιέχουν είτε το όνομα των οδηγών είτε τη συντομογραφία του επωνύμου (π.χ. Ham για τον Hamilton στη στήλη code) και παραμένει στο τελικό σύνολο δεδομένων μόνο η στήλη που περιέχει τα επίθετα των οδηγών. Τέλος, αφαιρούνται στήλες που αφορούν τις ελεύθερες δοκιμές, τους αγώνες sprint (ισχύουν από το 2022 και έπειτα) καθώς και τις στήλες που περιέχουν τις ώρες διεξαγωγής των αγώνων, των κατατακτήριων δοκιμών και των αγώνων sprint αφού είναι συγκεκριμένες και σταθερές κάθε σεζόν και δεν επηρεάζουν τις θέσεις τερματισμού των οδηγών.

Σχετικά με τις ελλείπουσες τιμές, στο σύνολο δεδομένων με τις μεταβλητές του πίνακα 5.2 φαίνεται να υπάρχουν 16 ελλείπουσες τιμές που αφορούν τις κατατακτήριες δοκιμές και βρίσκονται στις στήλες qualifyId, qualifying_position, q1, q2 και q3. Οι γραμμές που περιέχουν τις κενές εγγραφές για τις παραπάνω στήλες αφαιρέθηκαν από το σύνολο δεδομένων.

Υπάρχουν, ωστόσο, πολλές στήλες που δεν φαίνεται να έχουν κενές εγγραφές χαρακτηρισμένες ως NaN αλλά έχουν \N, δηλαδή Null. Αναλυτικότερα, όσον αφορά τις στήλες q1, q2 και q3, η στήλη q1 είχε 38 γραμμές με null εγγραφές οι οποίες και διαγράφηκαν με αποτέλεσμα η στήλη q2 να περιέχει 835 null εγγραφές και η στήλη q3 να περιέχει 1660 null εγγραφές. Οι στήλες αυτές περιέχουν τους χρόνους γύρου που κάνουν οι οδηγοί στις κατατακτήριες δοκιμές. Είναι γνωστό ότι στο πρώτο μέρος των κατατακτήριων δοκιμών (q1) συμμετέχουν και οι είκοσι οδηγοί και βγαίνουν εκτός (ζώνη αποκλεισμού - elimination zone) οι πέντε τελευταίοι για τους οποίους καταγράφηκαν οι πιο αργοί χρόνοι γύρου. Αντίστοιχα, στη δεύτερη φάση των κατατακτήριων δοκιμών (q2) συμμετέχουν οι δεκαπέντε οδηγοί που πέρασαν από το q1 και πάλι βγαίνουν εκτός οι πέντε τελευταίοι ώστε να περάσουν στο q3 μόνο οι πρώτοι δέκα οδηγοί. Επομένως, οι τιμές \N στα q2, q3 αφορούν τους οδηγούς που δεν πέρασαν στην αντίστοιχη φάση. Για να αντιμετωπιστεί το πρόβλημα των κενών εγγραφών κατασκευάστηκε μια νέα στήλη, η best_quali_time, η οποία περιέχει τους καλύτερους χρόνους που σημείωσαν οι οδηγοί στις κατατακτήριες δοκιμές και τελικώς οι στήλες q1, q2, q3 διαγράφηκαν από το dataset.

Ακόμη, κατόπιν ελέγχου, εγγραφές \N παρατηρούνται και στις στήλες final_position, milliseconds, fastestLapTime και fastestLapSpeed. Η στήλη milliseconds περιέχει το χρόνο που έκανε ο εκάστοτε οδηγός για να ολοκληρώσει τον αγώνα. Όπως είναι λογικό, ο οδηγός που τερματίζει πρώτος κάνει λιγότερο χρόνο από το δεύτερο κ.ο.κ.. Υπάρχει περίπτωση, όμως, ο οδηγός που βρίσκεται στην πρώτη θέση να «ρίξει γύρο ή γύρους» σε κάποιο μονοθέσιο, να το προσπεράσει δηλαδή για έναν η περισσότερους γύρους. Αυτό προφανώς συμβαίνει στα μονοθέσια που βρίσκονται στις τελευταίες θέσεις κατά τη διάρκεια του αγώνα. Στην προκειμένη περίπτωση, λοιπόν, δεν καταγράφεται η χρονική διάρκεια του αγώνα για τα συγκεκριμένα μονοθέσια αλλά μετριέται ο αριθμός των γύρων για τους οποίους ο πρώτος οδηγός έχει προσπεράσει τους τελευταίους. Γι' αυτό οι εγγραφές αυτές μένουν ως \N στη στήλη milliseconds αλλά καταγράφονται στη στήλη status μετρώντας πόσους γύρους πίσω έχει μείνει ο εκάστοτε οδηγός από τον πρώτο (π.χ. +1 Lap, +2 Laps κ.ο.κ.). Για αυτό, για να αντικατασταθούν τα \N της στήλης milliseconds υπολογίστηκε μια εκτίμηση του χρόνου που πραγματοποίησαν τα μονοθέσια που έχουν φάει γύρο ως εξής: ο χρόνος γύρου που έκανε ο εκάστοτε οδηγός στις κατατακτήριες δοκιμές ο οποίος και λαμβάνεται από την στήλη best_quali_time πολλαπλασιάζεται κάθε φορά με τον αριθμό των γύρων που τον έχει προσπεράσει ο πρώτος οδηγός και κατόπιν προστίθεται σε αυτό το αποτέλεσμα ο χρόνος διάρκειας αγώνα του πρώτου οδηγού. Τέλος, αξίζει να σημειωθεί ότι η μεταβλητή time περιέχει την ίδια πληροφορία με τη μεταβλητή milliseconds αλλά σε διαφορετική μορφή και για αυτό αφαιρέθηκε από το σύνολο δεδομένων.

Όσον αφορά τη στήλη final_position, αυτή περιέχει τις θέσεις τερματισμού των οδηγών στο τελικό dataset. Στα αρχικά σύνολα δεδομένων, όμως, και συγκεκριμένα στο results.csv υπάρχουν δύο στήλες, η position που περιέχει τις θέσεις τερματισμού των οδηγών καθώς και εγγραφές \N και η positionText που διαθέτει ακριβώς τις ίδιες εγγραφές με την position αλλά στις αντίστοιχες τιμές \N έχει γράμματα που αντιπροσωπεύουν το λόγο για τον οποίο στον εκάστοτε οδηγό δεν έχει καταχωρηθεί θέση τερματισμού (βλ. σχήμα 5.7). Συγκεκριμένα, διαθέτει τα γράμματα 'N', 'R', 'D', 'W' και 'E'. Το 'N' σημαίνει Not Classified και αφορά στους οδηγούς που τερμάτισαν αλλά δεν έχουν ολοκληρώσει ένα προκαθορισμένο ποσοστό του αγώνα και έτσι δεν επιβραβεύονται με θέση. Το 'R' σημαίνει Retired και αφορά στους οδηγούς των οποίων το μονοθέσιο αποσύρθηκε πριν τη λήξη του αγώνα λόγω ατυχήματος ή τεχνικών/μηχανικών προβλημάτων. Το 'D' σημαίνει Disqualified και αφορά στους οδηγούς που αποκλείονται από τον αγώνα είτε στο τέλος είτε κατά τη διάρκεια αυτού λόγω παραβίασης κανονισμών. Αντίστοιχη σημασία έχει και το 'E' που σημαίνει Excluded. Τέλος, το 'W' σημαίνει Withdrawn και αφορά στους οδηγούς που δεν εκκίνησαν στον αγώνα λόγω μηχανικών προβλημάτων που συνήθως προκύπτουν από συγκρούσεις στις κατατακτήριες δοκιμές. Αφού, λοιπόν, για τους παραπάνω λόγους δεν έχει καταχωρηθεί θέση τερματισμού οι τιμές \N στη στήλη position (τελική ονομασία: final_position) αντικαταστάθηκαν με 0 και η στήλη positionText δε χρησιμοποιήθηκε στο τελικό dataset. Τέλος, οι εγγραφές \N στις στήλες fastestLapTime και fastestLapSpeed διαγράφηκαν αφού αφορούν κυρίως οδηγούς που δεν τερμάτισαν και τα χαρακτηριστικά του ταχύτερου γύρου τους τελικώς δεν καταγράφηκαν. Δεν αντικαταστάθηκαν από την τιμή 0 αφού έτσι θα σήμαινε ότι συμπεριλήφθηκε ο ταχύτερος γύρος αλλά είχε μηδενική διάρκεια και ταχύτητα.

Η στήλη status περιέχει εγγραφές που περιγράφουν με λεπτομέρεια είτε διάφορα προβλήματα, μηχανικά και τεχνικά, που εμφανίστηκαν στα μονοθέσια κατά τη διάρκεια του αγώνα είτε αν τα μονοθέσια τερμάτισαν, φορτώθηκαν γύρους, αποσύρθηκαν ή αποκλείστηκαν είτε αν ενεπλάκησαν σε σύγκρουση. Ακόμη αναφέρει και περιπτώσεις ασθένειας οδηγών. Επειδή οι διαφορετικές εγγραφές που περιέχει είναι 58, δημιουργήθηκε

νέα στήλη με το όνομα new_status η οποία περιέχει πιο διευρυμένες κατηγορίες των ανωτέρω. Συγκεκριμένα, δημιουργήθηκαν οι εξής 8 κατηγορίες:

- **finished** για τα μονοθέσια που τερμάτισαν ή φορτώθηκαν γύρους
- **engine_related_problems** για μονοθέσια που παρουσίασαν μηχανικά προβλήματα
- **technical_problems** για μονοθέσια που παρουσίασαν τεχνικά προβλήματα
- **crash_damage** για μονοθέσια που ενεπλάκησαν σε σύγκρουση
- **retired** για μονοθέσια που αποσύρθηκαν
- **not_ranked** για τα μονοθέσια που αποκλείστηκαν ή δεν εκκίνησαν καθόλου
- **illness** για ασθένεια οδηγών
- **other**

Ακόμη, υπάρχουν κατασκευαστές που κατά τη διάρκεια της περιόδου που μελετάμε άλλαξαν την επωνυμία τους. Σε αυτήν την περίπτωση όλες οι παλιές επωνυμίες άλλαξαν και στο dataset χρησιμοποιείται μόνο η πιο πρόσφατη. Κάτι αντίστοιχο συνέβη και με τα ονόματα των circuit αφού κάποιες χρονιές τα circuit καταγράφηκαν με το όνομα της επετείου που εορταζόταν εκείνη τη στιγμή και όχι με την κλασσική τους ονομασία. Οι αλλαγές που έγιναν παρουσιάζονται αναλυτικά στους κάτωθι πίνακες:

Προηγούμενη Επωνυμία Κατασκευαστή	**Τρέχουσα Επωνυμία Κατασκευαστή**
Toro Rosso	Alpha Tauri
Force India	Aston Martin
Racing Point	
Sauber	Alpha Romeo
Marussia	Manor Marussia
Lotus F1	Alpine F1 Team
Renault	

Πίνακας 5.3: Αλλαγή επωνυμίας κατασκευαστών

Αρχική Ονομασία Circuit	**Τελική Ονομασία Circuit**
European Grand Prix	Azerbaijan Grand Prix
Eifel Grand Prix	German Grand Prix
70th Anniversary Grand Prix	British Grand Prix

Πίνακας 5.4: Αλλαγή ονομασίας circuit

Οι χρόνοι που περιέχουν οι μεταβλητές best_quali_time και fastestLapTime μετατράπηκαν από λεπτά σε χιλιοστά του δευτερολέπτου. Δημιουργήθηκε μια ακόμη μεταβλητή, η age η οποία καταγράφει την ηλικία κάθε οδηγού σε κάθε αγώνα αφού συνηθίζεται μεγαλύτεροι ηλικιακά οδηγοί που διαθέτουν περισσότερη εμπειρία να έχουν καλύτερες επιδόσεις. Τέλος, από το σύνολο δεδομένων διαγράφηκαν οι στήλες των πρωτευόντων κλειδιών μιας και δεν προσφέρουν καμία πληροφορία στην ανάλυση.

Αφού ολοκληρώθηκε το στάδιο προπαρασκευής των δεδομένων για τις ήδη υπάρχουσες μεταβλητές προχωράμε στην κατασκευή νέων μεταβλητών ώστε να αντιμετωπιστεί το πρόβλημα που δημιουργείται από τις μεταβλητές που μεταφέρουν πληροφορία από το μέλλον. Η διαδικασία κατασκευής των μεταβλητών αυτών παρατίθεται αναλυτικά στην επόμενη ενότητα

5.5 Αντιμετώπιση προβλήματος target leakage

Όταν τα μοντέλα μηχανικής μάθησης περιέχουν δεδομένα τα οποία δε θα είναι διαθέσιμα την στιγμή που γίνεται η πρόβλεψη τότε παρουσιάζεται το λεγόμενο target leakage. Στην ουσία, το σύνολο δεδομένων ελέγχου περιέχει πληροφορία για τη μεταβλητή στόχο η οποία προέρχεται από το μέλλον με αποτέλεσμα οι προβλέψεις να είναι εξαιρετικά καλές, αν όχι τέλειες γεγονός μη πραγματικό. Στο παρόν case study δε μπορούμε πριν από τον αγώνα να γνωρίζουμε τους γύρους που πραγματοποίησε κάθε οδηγός, αν πέτυχε ταχύτερο γύρο και αν ναι, τον χρόνο και τη διάρκεια αυτού. Ακόμη, αν δε λήξει ο αγώνας δεν είναι γνωστό αν κάποιο μονοθέσιο παρουσιάσει τεχνικό ή μηχανικό πρόβλημα ή αν θα συγκρουστεί. Τέλος, είναι προφανές ότι αν δεν πέσει η καρό σημαία δεν είναι γνωστές οι θέσεις που τερμάτισαν οι οδηγοί και κατ' επέκταση η βαθμολογία τους. Γνωρίζουμε μόνο τι συνέβη στους προηγούμενους αγώνες αλλά και στις κατατακτήριες δοκιμές. Για το λόγο αυτό, κατασκευάζονται νέες μεταβλητές οι οποίες θα δίνουν όσο το δυνατόν περισσότερη πληροφορία για τους προηγούμενος αγώνες.

Οι νέες μεταβλητές που κατασκευάστηκαν είναι οι εξής:

- driverTop3
- driverTop5
- driverTop10
- driverBottom3
- driverBottom5
- driverBottom10
- driverPointsHist
- millisecondsHist
- positionOrderHist

Αναλυτικότερα, η μεταβλητή driverTop3 περιέχει πληροφορία για τον αν ο εκάστοτε οδηγός τερμάτισε στις πρώτες τρείς θέσεις στον προηγούμενο αγώνα. Αν ναι, τότε καταγράφεται η θέση στην οποία τερμάτισε, διαφορετικά καταγράφεται η τιμή μηδέν. Στον πρώτο αγώνα κάθε έτους καταγράφεται μηδενική θέση αφού δεν υπάρχει προηγούμενος αγώνας. Αντίστοιχα, η μεταβλητή driverTop5 δίνει πληροφορία για το αν ο εκάστοτε οδηγός τερμάτισε στις πέντε πρώτες θέσεις στον προηγούμενο αγώνα. Αν ναι, καταγράφεται η θέση που τερμάτισε αλλιώς η μεταβλητή παίρνει τιμή μηδέν. Πάλι στον πρώτο αγώνα κάθε έτους καταγράφεται μηδέν για όλους τους οδηγούς αφού δεν υπάρχει προηγούμενος αγώνας. Το ίδιο συμβαίνει και για τη μεταβλητή driverTop10 για τις δέκα πρώτες θέσεις.

Με το ίδιο σκεπτικό κατασκευάστηκε και η μεταβλητή driverBottom3. Συγκεκριμένα, αν ο οδηγός τερμάτισε στις τρεις τελευταίες θέσεις στον προηγούμενο αγώνα τότε καταγράφεται η θέση τερματισμού του ενώ αν δεν βρίσκεται στις τρεις τελευταίες θέσεις καταγράφεται η τιμή μηδέν. Πάλι, για τον πρώτο αγώνα κάθε έτους η μεταβλητή έχει μηδενική τιμή για όλους τους οδηγούς. Το ίδιο συμβαίνει και για τις μεταβλητές driverBottom5 και driverBottom10 για τις πέντε τελευταίες και τις δέκα τελευταίες θέσεις αντίστοιχα.

Όσον αφορά τη μεταβλητή driverPointsHist, αυτή καταγράφει τους πόντους που συγκέντρωσε ο οδηγός κάθε φορά στον προηγούμενο αγώνα. Για τον πρώτο αγώνα κάθε έτους η μεταβλητή κατέγραψε, αρχικά, ελλείπουσες τιμές. Οι κενές αυτές εγγραφές αντικαταστάθηκαν στη συνέχεια με το μέσο όρο των βαθμών που συγκέντρωσε ο εκάστοτε οδηγός στη διάρκεια κάθε σεζόν. Ωστόσο, μερικές από τις ελλείπουσες τιμές αφορούν σε

οδηγούς που συμμετείχαν μόνο για έναν αγώνα και δεν υπήρχε καταγραφή πόντων σε κάποιον προηγούμενο. Οι εγγραφές αυτές αντικαταστάθηκαν με 0.

Η μεταβλητή millisecondsHist καταγράφει για κάθε έτος το χρόνο που πραγματοποίησε ο εκάστοτε οδηγός στον προηγούμενο αγώνα. Και εδώ για τον πρώτο αγώνα κάθε έτους καταχωρούνται ελλείπουσες τιμές οι οποίες, στη συνέχεια, αντικαθίστανται με τον μέσο όρο του χρόνου που πραγματοποιεί ο οδηγός στους υπόλοιπους αγώνες κάθε έτους. Πάλι υπάρχουν κενές εγγραφές για τους οδηγούς που αγωνίστηκαν μία μόνο φορά οι οποίες και πάλι αντικαταστάθηκαν από μηδέν.

Τέλος, η μεταβλητή positionOrderHist περιέχει για κάθε αγώνα τη θέση που τερμάτισε ο εκάστοτε οδηγός στον προηγούμενο αγώνα. Και εδώ οι θέσεις των οδηγών για τον πρώτο αγώνα κάθε έτους καταγράφηκαν αρχικά ως ελλείπουσες τιμές και στη συνέχεια αντικαταστάθηκαν με το μέσο όρο των θέσεων που τερμάτισε ο οδηγός στους επόμενους αγώνες. Για τους οδηγούς που αγωνίστηκαν μόνο μία φορά οι ελλείπουσες τιμές διαγράφηκαν.

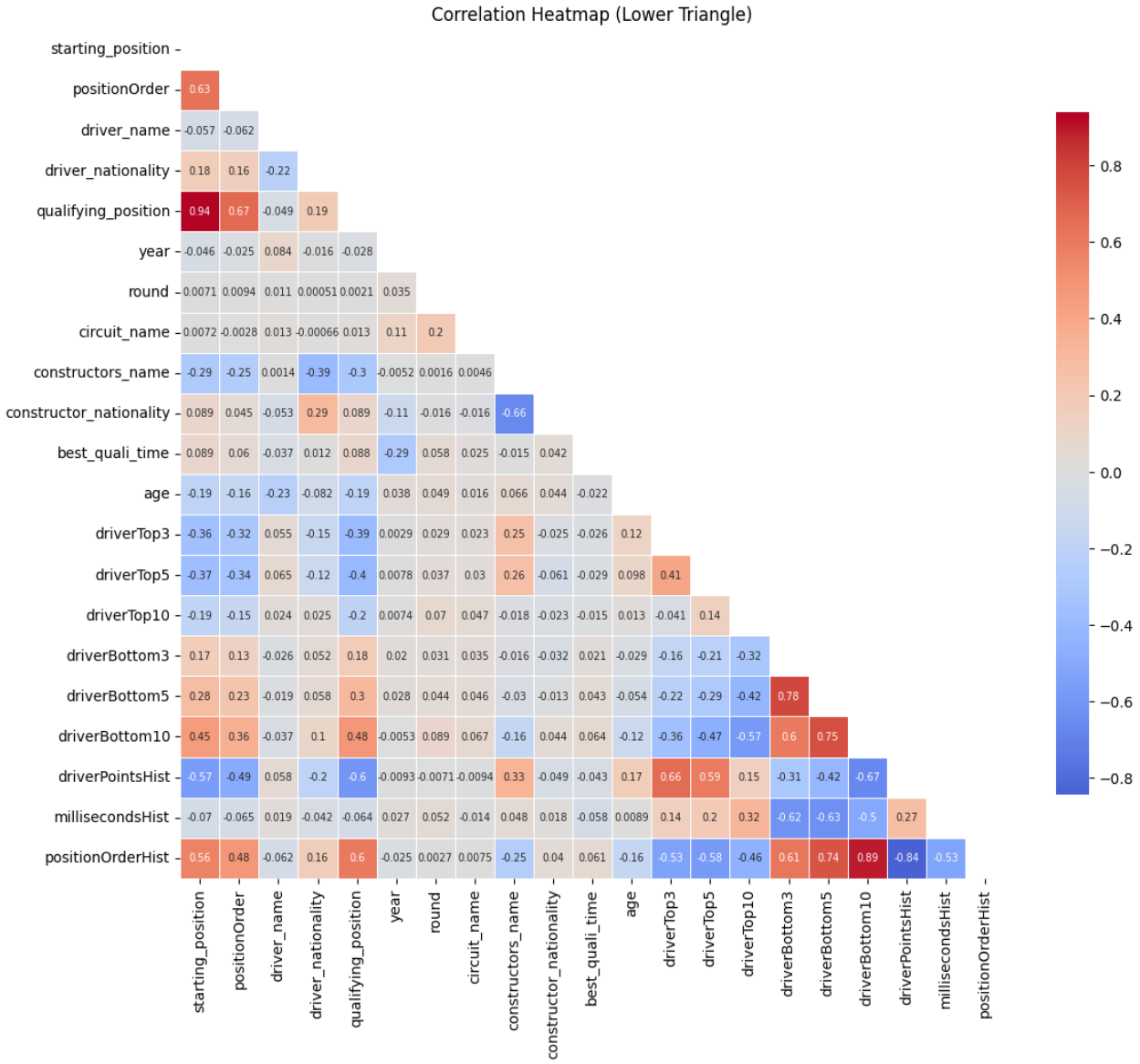
Έπειτα από την κατασκευή των νέων μεταβλητών το τελικό σύνολο δεδομένων διαμορφώνεται ως εξής:

Όνομα	**Περιγραφή**
starting_position	Θέση οδηγού στη σχάρα εκκίνησης
positionOrder	Τελική κατάταξη οδηγού μετά τον αγώνα
driver_name	Όνομα οδηγού (μόνο επίθετο)
driver_nationality	Εθνικότητα οδηγού
year	Έτος αγώνα
round	Αριθμός αγώνα βάσει calendar
circuit_name	Όνομα πίστας
constructors_name	Όνομα κατασκευαστή
constructor_nationality	Εθνικότητα κατασκευαστή
best_quali_time	Χρόνος καλύτερου γύρου στις κατατακτήριες δοκιμές
age	Ηλικία οδηγού σε κάθε αγώνα
driverTop3	Τερμάτισε ή όχι ο οδηγός στις 3 πρώτες θέσεις
driverTop5	Τερμάτισε ή όχι ο οδηγός στις 5 πρώτες θέσεις
driverTop10	Τερμάτισε ή όχι ο οδηγός στις 10 πρώτες θέσεις
driverBottom3	Τερμάτισε ή όχι ο οδηγός στις 3 τελευταίες θέσεις
driverBottom5	Τερμάτισε ή όχι ο οδηγός στις 5 τελευταίες θέσεις
driverBottom10	Τερμάτισε ή όχι ο οδηγός στις 10 τελευταίες θέσεις
driverPointsHist	Πόντοι οδηγού από προηγούμενο αγώνα
millisecondsHist	Χρόνος οδηγού από προηγούμενο αγώνα
positionOrderHist	Τελική θέση οδηγού από προηγούμενο αγώνα

Πίνακας 5.5: Τελικό σύνολο δεδομένων

Το τελικό dataset αποτελείται από 3,033 γραμμές και 21 στήλες. Η κωδικοποίηση των κατηγορικών μεταβλητών έγινε με την τεχνική κωδικοποίησης ετικέτας (label encoding) και κλιμάκωση των δεδομένων (scaling) έγινε μόνο στις μεταβλητές που σχετίζονται με χρόνο (millisecondsHist, best_quali_time).

Τέλος, δίνεται παρακάτω ο χάρτης θερμότητας (heatmap) με τις συσχετίσεις των μεταβλητών μεταξύ τους.



Σχήμα 5.9: Heatmap των μεταβλητών

Φαίνεται να υπάρχουν τέσσερις υψηλές θετικές συσχετίσεις. Αρχικά η μεταβλητή qualifying_position παρουσιάζει εξαιρετικά υψηλή συσχέτιση (0,94) με την μεταβλητή starting_position, κάτι λογικό και αναμενόμενο αφού οι τελικές θέσεις των οδηγών στις κατατακτήριες δοκιμές είναι ίδιες με τις αρχικές θέσεις των οδηγών στη γραμμή εκκίνησης για τον αγώνα πλην ορισμένων ελάχιστων περιπτώσεων κατά τις οποίες η θέση του οδηγού διαφέρει από τις κατατακτήριες δοκιμές στη σχάρα εκκίνησης λόγω επιβολής κάποιας ποινής. Επίσης, υψηλή θετική συσχέτιση (0,89) παρουσιάζουν και οι μεταβλητές driverBottom10 και

positionOrderHist. Ακόμη, υψηλές συσχετίσεις παρουσιάζουν και οι μεταβλητές driverBottom3 και driverBottom5 και οι driverBottom5 με την driverBottom10.

Παρακάτω παρουσιάζεται και ο πίνακας συσχετίσεων όλων των μεταβλητών με τη μεταβλητή στόχο final_position.

Όνομα μεταβλητής	**Συσχέτιση**
qualifying_position	0.668
starting_position	0.635
positionOrderHist	0.483
driverBottom10	0.362
driverBottom5	0.231
driver_nationality	0.156
driverBottom3	0.135
best_quali_time	0.06
constructor_nationality	0.045
round	0.009
circuit_name	-0.003
year	-0.025
driver_name	-0.062
millisecondsHist	-0.065
driverTop10	-0.153
age	-0.157
constructors_name	-0.253
driverTop3	-0.321
driverTop5	-0.341
driverPointsHist	-0.495

Πίνακας 5.6: Συσχετίσεις μεταβλητών με τη μεταβλητή στόχος

Όπως φαίνεται στον πίνακα 5.6 οι μεταβλητές qualifying_position, starting_position, positionOrderHist και driverPointsHist παρουσιάζουν σχετικά ισχυρές συσχετίσεις (θετικές και αρνητική) με τη μεταβλητή στόχο γεγονός που υποδηλώνει ύπαρξη γραμμικότητας. Ωστόσο, οι περισσότερες μεταβλητές δεν παρουσιάζουν ισχυρές συσχετίσεις με την final_position. Γι' αυτό πέραν των μοντέλων παλινδρόμησης ridge, lasso και elastic net τα οποία πιθανόν δε θα είναι ιδιαίτερα αποδοτικά, χρησιμοποιούνται επίσης τμηματικά πολυώνυμα (piecewise polynomial splines), συλλογικές μέθοδοι βασισμένες σε δένδρα απόφασης (random forest, XGBoost, AdaBoost, Decision Tree) και νευρωνικά δίκτυα.

5.6 Εφαρμογή αλγορίθμων παλινδρόμησης και νευρωνικών δικτύων

Για την ανάλυση θα χρησιμοποιηθούν αλγόριθμοι παλινδρόμησης, συλλογικές μέθοδοι βασισμένες σε δένδρα απόφασης και νευρωνικά δίκτυα. Ως εξαρτημένη μεταβλητή ορίζεται η positionOrder. Ως σύνολο δεδομένων εκπαίδευσης χρησιμοποιούνται τα δεδομένα για τα έτη 2014 έως 2020 και ως σύνολο δεδομένων ελέγχου χρησιμοποιούνται τα δεδομένα για το έτος 2021. Επίσης, για την βελτιστοποίηση των παραμέτρων χρησιμοποιείται η τεχνική αναζήτησης πλέγματος (GridSearch). Ακόμη, χρησιμοποιείται η τεχνική διασταυρούμενης επικύρωσης 10 πτυχών (10-fold Cross Validation). Χρησιμοποιήθηκε seed 123

Οι αλγόριθμοι παλινδρόμησης που χρησιμοποιήθηκαν είναι οι εξής:

- Random Forest Regression
- Extreme Gradient Boosting Regression
- Ridge Regression
- Lasso Regression
- Elastic Net Regression
- AdaBoost Regression
- Decision Tree Regression
- Piecewise Polynomial Splines

Τα νευρωνικά δίκτυα που χρησιμοποιήθηκαν είναι τα εξής:

- Feedforward Neural Network
- Convolutional Neural Network
- Long Short-Term Memory

Τα μέτρα που χρησιμοποιήθηκαν για την αξιολόγηση των μοντέλων παλινδρόμησης και τω νευρωνικών δικτύων είναι τα εξής:

- Mean Squared Error, MSE
- Root Mean Squared Error, RMSE
- Mean Absolute Error, MAE
- Median Absolute Error, MedAE
- R^2

5.6.1 Πρόβλεψη θέσης τερματισμού των οδηγών

Τα αποτελέσματα των μέτρων για κάθε μοντέλο παρουσιάζονται στον παρακάτω πίνακα

Model	MSE	RMSE	MAE	MedAE	R^2
Random Forest	16.073	4.009	3.074	2.532	0.491
XGB	15.542	3.942	3.022	2.422	0.508
Ridge	15.813	3.977	3.0395	2.443	0.499
Lasso	15.831	3.979	3.057	2.414	0.499
Elastic Net	15.82	3.977	3.049	2.404	0.499
Splines	15.52	3.939	2.981	2.342	0.509
AdaBoost	17.076	4.132	3.226	2.651	0.459
Decision Tree	19.920	4.463	3.415	2.806	0.369
FNN	16.769	4.095	3.19	2.563	0.469
CNN	16.284	4.035	3.069	2.47	0.485
LSTM	15.838	3.98	3.07	2.52	0.499

Πίνακας 5.7: Μέτρα μοντέλων (Τα παραπάνω αποτελέσματα είναι στρογγυλοποιημένα στα τρία δεκαδικά ψηφία)

Παρατηρώντας τον πίνακα 5.7 διαπιστώνουμε ότι το μοντέλο Piecewise Polynomial Splines (Splines) παρουσιάζει την καλύτερη εικόνα σε όλα τα μέτρα καθώς έχει το χαμηλότερο μέσο τετραγωνικό σφάλμα (MSE: 15.52), το χαμηλότερο RMSE (3.939), το χαμηλότερο MAE (2.981) και το χαμηλότερο MedAE (2.342). Επίσης, έχει το υψηλότερο R^2 score (0.509). Αμέσως μετά, με μικρές διαφορές, έρχεται το μοντέλο XGBoost με MSE 15.542 και R^2 score 0.508. Στη συνέχεια ακολουθούν τα μοντέλα Ridge, Lasso, Elastic Net και LSTM παρουσιάζουν R^2 score 0.499.

Ικανοποιητικά αποτελέσματα δείχνουν τα μοντέλα τυχαίου δάσους, CNN και FNN με συντελεστές προσδιορισμού R^2 score 0.491, 0.485 και 0.469 αντιστοίχως. Τέλος, τα μοντέλα δένδρου απόφασης και AdaBoost δείχνουν να μην προσαρμόζονται καλά στα δεδομένα καθώς παρουσιάζουν τα υψηλότερα μέσα τετραγωνικά σφάλματα, 19.920 και 17.076 αντίστοιχα ενώ έχουν τα χαμηλότερα R^2 score, 0.369 και 0.459 αντίστοιχα.

Στους παρακάτω πίνακες παρουσιάζεται η πραγματική θέση τερματισμού του κάθε οδηγού καθώς και η θέση που προέβλεψε το εκάστοτε μοντέλο, ενδεικτικά για τον 1ο, τον 7ο, τον 16ο και τον 22ο και τελευταίο αγώνα του έτους 2021.

Driver	Actual Position	RF	XGB	Ridge	Lasso	El. Net	DT	AdaBoost	FFNN	CNN	Spline	LSTM
Hamilton	1	3.1	3.1	3.9	4.1	4.1	1.6	3.7	4.2	3.9	3.1	4.1
Verstappen	2	4.1	4.1	3.5	3.6	3.6	1.6	3.9	3.6	3.0	3.2	4.3
Bottas	3	4.8	4.6	5.7	5.8	5.8	5.3	4.3	5.7	5.6	4.3	5.7
Norris	4	8.4	8.8	8.0	8.0	8.0	7.3	8.6	8.1	7.7	9.9	8.6
Perez	5	9.6	11.4	9.2	9.3	9.3	7.1	9.8	7.8	6.7	12.2	9.0
Leclerc	6	6.0	5.4	6.4	6.5	6.5	3.9	6.4	6.7	6.4	7.4	7.3
Ricciardo	7	8.0	7.6	8.1	8.0	8.1	5.7	7.3	8.3	7.7	7.5	8.7
Sainz	8	8.4	8.6	8.2	8.3	8.3	7.0	9.7	8.6	8.2	9.8	9.0

Tsunoda	**9**	12.2	12.4	13.4	13.2	13.3	11.5	12.3	12.6	12.3	14.1	12.5
Stroll	**10**	10.2	10.4	11.6	11.3	11.4	8.6	10.0	11.3	10.5	10.3	11.4
Räikkönen	**11**	13.1	13.6	13.4	13.3	13.4	11.9	14.4	13.3	12.9	13.1	13.6
Giovinazzi	**12**	12.5	12.4	13.1	12.9	13.0	13.2	11.8	12.5	12.7	14.8	13.0
Ocon	**13**	12.2	12.8	14.5	14.3	14.4	11.5	12.2	14.0	13.3	12.5	12.9
Russel	**14**	14.2	13.7	14.2	14.1	14.1	15.6	14.6	12.3	12.7	13.3	14.5
Vettel	**15**	13.4	13.6	16.0	15.8	15.9	11.5	12.6	15.2	14.9	13.0	14.8
Schumacher	**16**	15.4	16.1	18.2	17.7	17.9	16.1	14.8	15.5	16.3	14.7	16.9
Gasly	**17**	9.5	9.7	8.2	8.1	8.1	11.1	8.5	8.6	8.1	7.0	9.0
Latifi	**18**	15.0	14.3	15.9	15.6	15.7	15.6	14.6	13.6	13.8	13.8	15.9
Alonso	**19**	10.5	11.6	10.9	10.6	10.7	10.7	11.3	11.0	10.6	9.2	10.8

Πίνακας 5.8: Σύγκριση πραγματικών θέσεων τερματισμού σε σχέση με αυτές που προέβλεψαν τα μοντέλα για τον 1ο αγώνα του 2021

Driver	Actual Position	RF	XGB	Ridge	Lasso	El. Net	DT	AdaBoost	FFNN	CNN	Spline	LSTM
Verstappen	**1**	4.3	4.2	5.0	5.2	5.1	4.6	3.9	5.2	5.0	4.1	4.4
Hamilton	**2**	4.7	3.5	5.8	5.6	5.7	4.6	3.8	7.1	7.1	4.1	4.5
Perez	**3**	6.0	5.5	4.8	4.6	4.7	2.6	5.5	4.1	4.3	5.6	4.7
Bottas	**4**	5.1	4.2	6.6	6.5	6.6	2.6	4.2	7.4	7.2	4.3	6.2
Norris	**5**	8.7	8.8	7.6	7.8	7.8	10.5	10.1	7.9	6.9	7.3	8.2
Ricciardo	**6**	10.1	10.5	10.4	10.3	10.4	10.9	10.0	10.5	10.8	11.4	11.4
Gasly	**7**	8.0	9.1	7.4	7.1	7.1	11.1	8.3	7.6	7.3	8.4	8.2
Alonso	**8**	10.2	10.8	10.0	9.7	9.8	12.6	11.2	10.5	10.4	9.6	10.4
Vettel	**9**	9.4	10.2	9.9	9.6	9.7	11.5	10.1	9.8	9.3	10.6	9.1
Stroll	**11**	7.5	7.1	6.5	6.6	6.5	8.7	7.3	7.8	6.8	6.3	6.3
Sainz	**12**	14.7	13.8	11.4	11.9	11.7	15.6	14.6	11.2	12.0	12.6	12.0
Russel	**14**	10.2	11.1	11.5	11.6	11.5	8.7	10.3	11.4	11.5	12.2	12.0
Tsunoda	**15**	13.0	13.1	11.4	11.8	11.6	14.4	12.6	12.2	12.0	12.4	11.9
Ocon	**16**	7.6	8.0	7.3	7.4	7.4	7.3	8.5	7.5	7.3	8.5	7.9
Giovinazzi	**17**	13.8	13.7	13.2	13.5	13.3	11.9	14.0	14.2	14.0	11.8	12.9
Leclerc	**18**	15.6	14.4	13.9	13.7	13.8	15.6	14.6	13.8	13.8	13.9	14.6
Räikkönen	**19**	14.4	14.2	13.8	13.5	13.7	11.9	14.6	13.7	13.8	14.7	13.4
Latifi	**20**	15.5	15.7	15.8	15.5	15.6	16.1	14.7	15.2	15.8	15.7	14.4

Πίνακας 5.9: Σύγκριση πραγματικών θέσεων τερματισμού σε σχέση με αυτές που προέβλεψαν τα μοντέλα για τον 7ο αγώνα του 2021

Driver	Actual Position	RF	XGB	Ridge	Lasso	El. Net	DT	AdaBoost	FFNN	CNN	Spline	LSTM
Bottas	**1**	3.2	3.6	4.5	4.9	4.8	1.5	3.9	4.7	4.5	3.9	4.2
Verstappen	**2**	4.7	4.9	6.1	6.3	6.2	5.3	4.2	6.1	6.1	5.1	4.7
Perez	**3**	8.1	7.6	7.7	7.9	7.8	8.1	8.4	8.1	8.1	6.9	6.8
Leclerc	**4**	5.8	5.3	7.0	6.9	7.0	3.9	4.3	7.9	7.9	6.8	6.3
Hamilton	**5**	5.0	4.3	3.3	3.5	3.5	4.6	6.5	4.0	4.9	2.9	4.1
Gasly	**6**	7.7	7.8	8.4	8.4	8.4	3.9	7.5	9.0	8.6	8.1	8.2
Norris	**7**	8.8	9.4	7.8	8.1	8.0	14.4	10.4	8.5	8.1	9.0	7.3
Sainz	**8**	9.2	9.5	11.0	11.1	11.0	13.2	11.9	11.0	11.1	9.4	10.2
Stroll	**9**	9.7	9.5	10.0	10.0	10.0	8.6	9.9	10.8	9.7	9.4	10.0
Ocon	**10**	12.7	11.9	12.4	12.3	12.4	11.5	11.2	12.5	11.8	12.4	12.4
Giovinazzi	**11**	14.6	14.6	14.7	14.6	14.7	16.1	14.7	15.1	15.1	15.1	14.3
Räikkönen	**12**	13.1	13.8	14.4	14.7	14.5	13.2	12.2	14.5	14.6	13.0	14.4
Ricciardo	**13**	10.2	10.9	11.8	12.2	12.1	13.2	12.0	12.1	12.5	12.4	12.9
Tsunoda	**14**	10.1	10.4	11.4	11.1	11.2	6.9	9.9	11.6	10.9	12.2	10.7
Russel	**15**	11.9	12.2	10.7	11.2	11.0	15.6	12.5	10.9	11.4	10.8	10.4
Alonso	**16**	8.0	8.0	8.2	8.1	8.1	5.8	7.3	8.7	8.7	6.7	8.6
Latifi	**17**	15.1	14.4	13.4	13.8	13.6	15.6	14.7	12.6	13.4	12.9	13.6
Vettel	**18**	9.5	10.2	11.1	11.0	11.1	8.6	10.0	11.7	10.7	10.2	11.0
Schumacher	**19**	14.1	13.8	13.0	13.1	13.1	13.9	14.6	12.4	13.3	14.1	13.3
Mazepin	**20**	15.7	15.9	16.6	16.2	16.4	16.1	14.8	15.9	16.9	16.3	15.2

Πίνακας 5.10: Σύγκριση πραγματικών θέσεων τερματισμού σε σχέση με αυτές που προέβλεψαν τα μοντέλα για τον 16ο αγώνα του 2021

Driver	Actual Position	RF	XGB	Ridge	Lasso	El. Net	DT	AdaBoost	FFNN	CNN	Spline	LSTM
Verstappen	**1**	3.3	4.0	3.8	3.5	3.6	4.4	3.9	3.7	3.0	3.6	4.3
Hamilton	**2**	4.1	3.4	3.8	4.0	4.0	4.6	3.9	7.0	5.3	4.6	4.4
Sainz	**3**	8.0	7.1	6.7	6.8	6.8	5.8	7.1	8.1	7.2	6.3	7.1
Tsunoda	**4**	10.4	10.1	10.4	10.2	10.3	12.2	9.7	10.7	10.7	11.6	10.7
Gasly	**5**	11.4	12.0	11.5	11.4	11.5	7.8	11.7	11.9	11.8	11.2	11.4
Bottas	**6**	6.2	6.0	6.9	6.7	6.7	4.6	7.0	6.7	6.5	6.0	6.5
Norris	**7**	5.9	4.6	6.0	5.9	6.0	8.2	4.2	7.4	6.9	4.2	6.7
Alonso	**8**	11.5	11.9	12.3	12.0	12.2	13.2	11.4	12.9	12.9	11.0	13.0
Ocon	**9**	9.7	9.9	9.3	9.2	9.3	12.2	9.8	9.6	9.0	10.5	9.8
Leclerc	**10**	8.9	8.4	7.9	8.0	8.0	8.1	8.9	8.7	8.4	8.3	8.6
Vettel	**11**	13.9	13.5	13.4	13.7	13.6	13.9	14.6	13.7	13.0	12.3	13.2
Ricciardo	**12**	9.6	9.6	9.2	9.4	9.4	10.9	9.8	9.4	9.0	10.3	10.5
Stroll	**13**	13.7	12.8	12.6	12.4	12.5	13.9	14.5	13.3	13.0	11.6	12.7
Schumacher	**14**	15.1	15.3	15.8	15.9	15.9	16.1	14.8	15.0	16.0	16.3	16.1
Perez	**15**	5.9	5.8	6.8	7.0	7.0	8.2	6.3	7.2	7.2	6.4	6.4
Latifi	**16**	14.9	14.0	13.5	13.5	13.5	15.6	14.6	13.5	13.3	13.2	14.3
Giovinazzi	**17**	13.3	13.2	12.2	12.4	12.3	10.5	12.3	13.2	13.2	12.8	12.2
Russel	**18**	15.1	14.5	13.6	14.0	13.8	15.6	14.7	13.0	13.8	13.5	14.2
Räikkönen	**19**	14.7	14.9	15.0	14.9	15.0	16.1	14.7	16.1	16.1	13.4	14.8

Πίνακας 5.11: Σύγκριση πραγματικών θέσεων τερματισμού σε σχέση με αυτές που προέβλεψαν τα μοντέλα για τον 22ο αγώνα του 2021

Επιπροσθέτως, στον κάτωθι πίνακα παρουσιάζονται τα ποσοστά ικανοποιητικής πρόβλεψης κάθε μοντέλου με περιθώριο λάθους +-1, +-2, +-3.

Model	Within +-1	Within +-2	Within +-3
Random Forest	18.83%	39.19%	61.32%
XGB	19.08%	40.46%	62.09%
Ridge	18.83%	40.97%	61.07%
Lasso	18.58%	39.69%	60.31%
Elastic Net	19.08%	39.95%	60.81%
Splines	19.85%	42.24%	63.10%
AdaBoost	16.03%	38.93%	56.49%
Decision Tree	21.12%	39.95%	52.16%
FNN	17.81%	36.64%	58.78%
CNN	18.83%	41.98%	58.52%
LSTM	20.61%	38.42%	58.78%

Πίνακας 5.12: Ποσοστό αναμενόμενων σωστών προβλέψεων με εύρος λάθους στην πρόβλεψη +-1, +-2, +-3.

Παρατηρούμε ότι με περιθώριο λάθους +-1 το μοντέλου δένδρου απόφασης δίνει το μεγαλύτερο ποσοστό ικανοποιητικών προβλέψεων, περίπου 21.12%. Ωστόσο, για εύρος λάθους +-2 και +- 3 το μοντέλο των τμηματικών πολυωνύμων δίνει τα μεγαλύτερα ποσοστά ικανοποιητικών προβλέψεων, 42.24% και 63.10% αντίστοιχα.

ΚΕΦΑΛΑΙΟ 6ο

Συμπεράσματα

Στην παρούσα εργασία παρουσιάστηκε περιληπτικά η ιστορία της F1 κάνοντας αναφορά σε σημαντικά γεγονότα και περιόδους του αθλήματος. Ακόμη, έγινε όσο το δυνατόν πιο λεπτομερής περιγραφή των κανονισμών που τη διέπουν και αφορούν στη διεξαγωγή των πρωταθλημάτων, στα χαρακτηριστικά των μονοθέσιων, τις προδιαγραφές των circuits, τις ακαδημίες νέων οδηγών και τους προϋπολογισμούς των ομάδων. Σχετικά με το τελευταίο έγινε εκτενής περιγραφή των οικονομικών στοιχείων της F1. Συγκεκριμένα αναλύθηκαν τα σταθερά και μεταβλητά κόστη που πρέπει να καλύψουν οι ομάδες για κατασκευή και ανακατασκευή του μονοθέσιου, κατασκευή και εξέλιξη κινητήρα, μισθούς προσωπικού και οδηγών και άλλα κόστη που απαιτούνται για την επίτευξη του καλύτερου δυνατού αποτελέσματος. Επιπλέον, αναφέρθηκαν και άλλες οικονομικές δραστηριότητες που σχετίζονται με τη Formula 1 και αφορούν τα κόστη διεξαγωγής GP, κατασκευής και συντήρησης των circuit καθώς και τις εταιρείες που χορηγούν τις ομάδες και τη διοργάνωση στο σύνολό της. Τέλος, έγινε αναφορά σε μερικές από τις δημοφιλέστερες καινοτομίες τις F1 που βρίσκουν εφαρμογή στα αυτοκίνητα πόλης.

Στη συνέχεια, έγινε περιγραφή του θεωρητικού μέρους των αλγορίθμων μηχανικής μάθησης, των τρόπων αξιολόγησής τους καθώς και των πιο βασικών τεχνικών προπαρασκευής των δεδομένων. Με λιγότερες λεπτομέρειες έγινε αναφορά και στους βασικούς τύπους νευρωνικών δικτύων.

Έπειτα έγινε εφαρμογή των μεθόδων μηχανικής μάθησης σε δεδομένα της F1 που αφορούν τους οδηγούς, τους κατασκευαστές, τα αποτελέσματα κατατακτήριων δοκιμών και αγώνων. Ουσιαστικά, χρησιμοποιήθηκαν σύνολα από δεδομένα τα οποία παρουσιάζονται, μεμονωμένα για κάθε αγώνα, στο κοινό, είτε αυτό βρίσκεται στην εξέδρα και παρακολουθεί ζωντανά των αγώνα, είτε τα λαμβάνει στην οθόνη του.

Στην ανάλυσή μας χρησιμοποιήθηκαν αρκετοί αλγόριθμοι παλινδρόμησης, συλλογικές μέθοδοι βασισμένες σε δένδρα απόφασης και τρία διαφορετικά νευρωνικά δίκτυα με στόχο την πρόβλεψη της θέσης τερματισμού κάθε οδηγού στον αγώνα. Για να εφαρμοστούν, όμως, τα παραπάνω χρειάστηκε να δημιουργηθούν νέες μεταβλητές που αποτρέπουν το πρόβλημα χρήσης μελλοντικής πληροφορίας που θα επηρεάσει τις προβλέψεις των αλγορίθμων και θα δώσει λανθασμένα αποτελέσματα. Καλύτερη εφαρμογή στα δεδομένα φαίνεται να έχει το μοντέλο τμηματικών πολυωνύμων (piecewise polynomial splines) ενώ δεύτερο με ελάχιστες διαφορές έρχεται το μοντέλο extreme gradient boosting. Αντίθετα, τα μοντέλα decision tree και AdaBoost φάνηκαν να έχουν τη χειρότερη εφαρμογή στα δεδομένα.

Αξίζει να σημειωθεί, πως αν και η χρήση splines και XGBoost δείχνει να έχει καλύτερα αποτελέσματα, οι τιμές των R2 scores παραμένουν σχετικά χαμηλές και τα μέσα τετραγωνικά σφάλματα αρκετά υψηλά. Αυτό συμβαίνει γιατί η πρόβλεψη των θέσεων τερματισμού των οδηγών είναι εξαιρετικά περίπλοκη διαδικασία και ο παράγοντας τύχη παίζει καίριο ρόλο στο τελικό αποτέλεσμα. Παρ' όλα αυτά οι ομάδες της F1 διαθέτουν, πέραν των δεδομένων που χρησιμοποιήθηκαν στην παρούσα εργασία, και άλλα απόρρητα στο κοινό δεδομένα που συλλέγονται από ειδικούς αισθητήρες πάνω στο μονοθέσιο και αφορούν στην τηλεμετρία.

Είναι εύκολα αντιληπτό πως η αξιοποίηση κάθε είδους δεδομένων με σωστό και αποτελεσματικό τρόπο σε συνδυασμό με τις τεχνικές μηχανικής μάθησης μπορεί να συμβάλει στην δημιουργία ανταγωνιστικών στρατηγικών και στη λήψη ταχύτατων αποφάσεων τόσο πριν όσο και κατά τη διάρκεια του αγώνα.

Βιβλιογραφία

Ξένη

- *Mourao, P. (2017). The economics of motorsports: The case of Formula One, Springer Nature, London, United Kingdom.*
- *Tomar, A., Malik, H., Kumar, P., Iqbal, A. (Marc 2021). Proceedings of 3rd Conference on Machine Learning, Advances in Computing, Renewable Energy and Communication, Springer Nature, Singapore.*
- *Cheng, K. (2023). A Need For Speed: Enhancing F1 Race Cars with a Novel Computational Fluid Dynamics and Machine Learning Method, Crescent School, Toronto, Canada.*
- *Tejada, L. G. (2023). Applying Machine Learning to Forecast Formula 1 Race Outcomes, School of Science, Aalto University, Espoo, Finland.*
- *Boettinger, M., Klotz, D. (2023). Mastering Nordschleife -- A comprehensive race simulation for AI strategy decision-making in motorsports, Cornell University, New York, USA.*
- *Evans, B. D., Jordaan, H. W., Engelbrecht, H. A. (2023). Safe reinforcement learning for high-speed autonomous racing, Stellenbosch University, Stellenbosch, South Africa.*
- *Zhang, C., Huang, Z., Wang, S., Hong, Y. (2023). Decision-making for Overtaking in Specific Unmanned Driving Scenarios based on Deep Reinforcement Learning, Published in: 2023 IEEE 3rd International Conference on Information Technology, Big Data and Artificial Intelligence (ICIBA).*
- *Ju, S., van Vliet, P., Arenz, O., Peters, J. (2023). Digital Twin of a Driver-in-the-Loop Race Car Simulation With Contextual Reinforcement Learning, Published in: IEEE Robotics and Automation Letters.*
- *Patil, A., Jain, N., Agrahari, R., Hossari, M., Orlandi, F., Dev, S. (2023). A Data-Driven Analysis of Formula 1 Car Races Outcome, Published in: Artificial Intelligence and Cognitive Science 30th Irish Conference, AICS 2022, Munster, Ireland, December 8-9, 2022, Revised Selected Papers.*
- *von Schleinitz1, J., Schwarzhuber, T., W¨orle1, L., Graf, M., Eichberger, A., Trutschnig, W., Schr¨oder, A., (2022). Race Driver Evaluation at a Driving Simulator using a physical Model and a Machine Learning Approach, Cornell University, New York, USA.*
- *Den Hartog, I.H.D. (2022). Data to drive-Personalized visualization in Formula One racing, Eindhoven University of Technology, Eindhoven, Netherlands.*
- *Amsury, F., Ruhyana, N., Mardiana, T. (2022). Comparison of Classification Algorithms for Analysis Sentiment of Formula E Implementation in Indonesia, Universitas Nusa Mandiri, East Jakarta, Indonesia.*
- *Villegas-Ch, W., Garcia-Ortiz, J., Jaramillo-Alcazar, A. (2023). An Approach Based on Recurrent Neural Networks and Interactive Visualization to Improve Explainability in AI Systems, The Universidad de Las Americas, Quito, Ecuador.*

- *Greasley, A., Panchal, G., Samvedi, A. (2022). The Use of Simulation with Machine Learning and Optimization for a Digital Twin-A Case on Formula 1 DSS, Published in: 2022 Winter Simulation Conference (WSC), Singapore.*
- *Heilmeier, A., Thomaser, A., Graf, M., Betz, J. (2020). Virtual Strategy Engineer: Using Artificial Neural Networks for Making Race Strategy Decisions in Circuit Motorsport, Technical University of Munich, Garching, Germany.*
- *Sicoie, H. (2022). Machine Learning Framework for Formula 1Race Winner and Championship Standings Predictor, Tilburg University, Tilburg, Netherlands.*
- *Mohammed, J. Z., Wagner, M. Jr. (2020). Data Mining and Analysis: Fundamental Concepts and Algorithms, TJ International Ltd., Padstow, Cornwall, United Kingdom.*
- *Hastie, T., Tibshirani, R., Friedman, J. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Springer, Stanford, California.*
- *Maini, V., Sabri, S. (2017). Machine Learning for Humans*
- *Sutton, R. S., Barto, A. G. (2014, 2015). Reinforcement Learning: An Introduction, A Bradford Book, The MIT Press, Cambridge, Massachusetts, London, England.*
- *Bishop, C. M. (1995). Neural Networks for Pattern Recognition, Department of Computer Science and Applied Mathematics, Aston University, Birmingham, UK. p. 116-119*
- *McKinney, W. (2022). Python for Data Analysis: Data Wrangling with pandas, NumPy & Jupyter, O'Reilly, 3rd Edition.*

Διαδίκτυο

- https://en.wikipedia.org/wiki/History_of_Formula_One *(Τελευταία πρόσβαση: 08/2024)*
- https://www.fia.com/events/formula-2-championship/season-2023/fia-formula-2 *(Τελευταία πρόσβαση: 08/2024)*
- https://www.fia.com/events/fia-formula-3-championship/season-2023/fia-formula-3 *(Τελευταία πρόσβαση: 08/2024)*
- https://www.fia.com/regulation/category/110 *(Τελευταία πρόσβαση: 08/2024)*
- https://us.motorsport.com/f1/news/fia-track-grades-requirements-f1-potential/6508331/ *(Τελευταία πρόσβαση: 08/2024)*
- https://racingnews365.com/how-much-each-circuit-on-the-calendar-pays-to-formula-1 *(Τελευταία πρόσβαση: 08/2024)*
- https://rtrsports.com/en/blog/differences-in-the-cost-of-f1-sponsorship/ *(Τελευταία πρόσβαση: 08/2024)*
- https://www.formula1.com/en/latest/article.all-10-formula-1-teams-to-have-f1-academy-drivers-and-liveries-for-the-2024.hxfPWROnTFy5cPaDzuPku.html *(Τελευταία πρόσβαση: 08/2024)*
- https://www.sportingnews.com/au/formula-1/news/how-much-f1-drivers-paid-salaries-teams-2023/kr2jqdm7hofmbpyohlckmuy5 *(Τελευταία πρόσβαση: 08/2024)*
- https://racingnews365.com/f1-driver-salaries-2023 *(Τελευταία πρόσβαση: 08/2024)*
- https://formulapedia.com/f1-pit-crew-salary/ *(Τελευταία πρόσβαση: 08/2024)*

- https://en.number13.de/f1-23-which-team-drives-which-engine/ *(Τελευταία πρόσβαση: 08/2024)*
- https://formulapedia.com/f1-engine-manufacturers-suppliers/ *(Τελευταία πρόσβαση: 08/2024)*
- https://www.forbes.com/sites/csylt/2019/11/10/revealed-the-14-billion-cost-of-developing-f1-engines/?sh=387daa152755 *(Τελευταία πρόσβαση: 08/2024)*
- https://www.forbes.com/sites/mikeozanian/2023/07/19/formula-1s-most-valuable-teams-2023/?sh=586daeeb2adb *(Τελευταία πρόσβαση: 08/2024)*
- https://www.formula1.com/en/latest/article.fia-confirm-all-10-f1-teams-operated-below-cost-cap-in-2022.1MSZ7OgE8XF6GTMBLH4LxS.html *(Τελευταία πρόσβαση: 08/2024)*
- https://theathletic.com/4834040/2023/09/05/f1-cost-cap-2022/ *(Τελευταία πρόσβαση: 08/2024)*
- https://www.racv.com.au/royalauto/transport/cars/f1-innovations-in-road-cars.html *(Τελευταία πρόσβαση: 08/2024)*
- https://www.makeuseof.com/f1-tech-in-road-car/ *(Τελευταία πρόσβαση: 08/2024)*
- https://us.motorsport.com/f1/news/the-car-that-changed-formula-1-history/4779401/ *(Τελευταία πρόσβαση: 08/2024)*
- https://www.bmw.com/en/performance/carbon-fiber-in-a-car.html *(Τελευταία πρόσβαση: 08/2024)*
- https://www.24h-lemans.com/en/news/the-history-of-the-paddle-shift-2922 *(Τελευταία πρόσβαση: 08/2024)*
- https://www.forbes.com/sites/csylt/2017/03/13/the-1-billion-cost-of-hosting-an-f1-race/?sh=7e1ec7904f79 *(Τελευταία πρόσβαση: 08/2024)*
- https://www.forbes.com/sites/csylt/2019/05/19/revealed-sponsors-fuel-formula-one-with-30-billion/?sh=7dd2a6e62416 *(Τελευταία πρόσβαση: 08/2024)*
- https://racingnews365.com/f1-hands-out-billion-dollar-prize-money-thats-how-much-red-bull-and-other-teams-get *(Τελευταία πρόσβαση: 08/2024)*