ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ

ΣΧΟΛΗ ΤΕΧΝΟΛΟΓΙΩΝ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΕΠΙΚΟΙΝΩΝΙΩΝ

ΤΜΗΜΑ ΨΗΦΙΑΚΩΝ ΣΥΣΤΗΜΑΤΩΝ

ΠΜΣ ΠΛΗΡΟΦΟΡΙΑΚΑ ΣΥΣΤΗΜΑΤΑ ΚΑΙ ΥΠΗΡΕΣΙΕΣ

ΚΑΤΕΥΘΥΝΣΗ: ΜΕΓΑΛΑ ΔΕΔΟΜΕΝΑ ΚΑΙ ΑΝΑΛΥΤΙΚΗ

Γερασόπουλος Αλέξανδρος

Πτυχιακή Εργασία

Θέμα : «Εφαρμογή Τεχνικών Εξόρυξης Γνώσης σε Οικονομικά Δεδομένα – Πλεονεκτήματα και Μειονεκτήματα σε μια Τράπεζα και στις Πιστωτικές με Χρήση Πακέτων R/Matlab/Python»

<u>Επιβλέπων Καθηγητής</u> : Μιχαήλ Φιλιππάκης

ΠΕΙΡΑΙΑΣ, 06-2023

Περιεχόμενα

Κατάλογος Εικόνων

Κατάλογος Πινάκων

Περίληψη

Τα τελευταία χρόνια λόγω της ραγδαίας ανάπτυξης της τεχνολογίας σε επίπεδο επικοινωνιών αλλά επίσης σε επίπεδο αποθήκευσης και επεξεργαστικής ισχύος οδήγησε την παραγωγή δεδομένων σε επίπεδα που κανένας δεν μπορούσε να φανταστεί. Πλέον οι επιχειρήσεις παγκοσμίως θα πρέπει να επενδύσουν στο τομέα της ανάλυσης των δεδομένων προκειμένου να παραμείνουν ανταγωνιστικές. Οι περισσότεροι κλάδοι της σύγχρονης οικονομίας έχουν επηρεαστεί από την έκρηξη αυτής της πληροφορίας καθώς η αξιοποίησή της προσφέρει κέρδη η προλαμβάνει δυσάρεστες καταστάσεις.

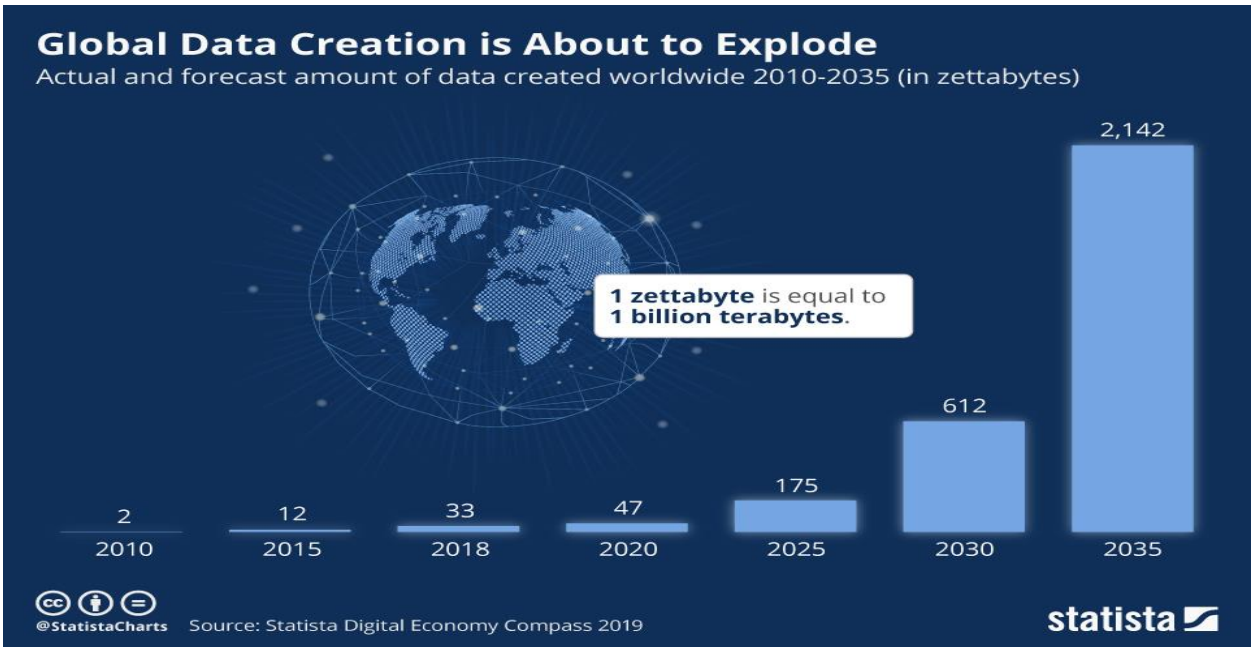
Σαφώς το τραπεζικό σύστημα δεν θα μπορούσε να μείνει εκτός αυτής της παγκόσμιας πορείας που θέλει κάθε οργανισμό να αξιοποιεί στο έπακρο τα διαθέσιμα δεδομένα προς όφελός του. Τα σύγχρονα τραπεζικά ιδρύματα διαθέτουν ένα «ωκεανό» δεδομένων που μέσω των τεχνικών εξόρυξης γνώσης χρησιμεύουν αρχικά για την επέκταση των οργανισμών αλλά επίσης και για την προστασία τους. Μία απειλή για κάθε τραπεζικό ίδρυμα είναι αυτή της απάτης και πιο συγκεκριμένα της απάτης μέσω πιστωτικών καρτών. Οι τράπεζες οφείλουν να προστατέψουν τους πελάτες τους από τέτοια περιστατικά αλλά ο μεγάλος όγκος συναλλαγών στέκεται τροχοπέδη στην επιτυχημένη ανίχνευση των περιστατικών απάτης που δεν είναι πολλά σε σχέση με αυτά των νόμιμων.

Στόχος της παρούσας διπλωματικής εργασίας είναι η ανάδειξη του προβλήματος της ανίχνευσης της απάτης των συναλλαγών μέσω πιστωτικών καρτών καθώς επίσης και η διερευνητική διαδικασία μέσω της οποίας δημιουργούνται μοντέλα μηχανικής μάθησης με σκοπό την αποτελεσματική ταξινόμηση μιας συναλλαγής ως νόμιμη ή παράνομη.

1. Μεγάλα Δεδομένα (Big Data)

Η αλματώδης τεχνολογική εξέλιξη των τελευταίων χρόνων έχει επηρεάσει, αν όχι σε όλους, σε πολλούς τομείς την ζωή των ανθρώπων. Έχει επηρεάσει τόσο πολύ όπου πλέον δημιουργούνται νέες νόρμες είτε στην επικοινωνία των ανθρώπων μεταξύ τους είτε, στην επιχειρηματική τους δραστηριότητα. Η κύρια τεχνολογία που ώθησε την ανθρωπότητα σε αυτή την αλλαγή είναι η εξέλιξη και κατ' επέκταση η αύξηση της ταχύτητας του διαδικτύου. Πλέον οι άνθρωποι χρησιμοποιώντας τις νέες τεχνολογίες της εποχής όπως τα έξυπνα κινητά τηλέφωνα σε συνδυασμό με τις υψηλές ταχύτητες των δικτύων, είτε οικιστικών είτε κυψελωδών τέταρτης και πέμπτης γενιάς, μπορούν να διαχειρίζονται από παντού την επιχείρησή τους, τις τραπεζικές τους κινήσεις καθώς επίσης μπορούν έχουν και διαρκή ενημέρωση νέων τάσεων της αγοράς.

Έχοντας πλέον πρόσβαση σχεδόν από παντού, οι άνθρωποι παράγουν συνεχώς δεδομένα είτε αυτά είναι φωτογραφίες που ανεβαίνουν σε κάποιο μέσω κοινωνικής δικτύωσης είτε είναι γεωγραφικά δεδομένα, είτε είναι η εκδήλωση ενδιαφέροντος για κάποιο προϊόν, είτε είναι κινήσεις τραπεζικών συναλλαγών. Η αύξηση των παραγόμενων δεδομένων είναι τόσο ραγδαία όπου όπως αναφέρεται χαρακτηριστικά στο [1] αν έπρεπε να αποθηκεύσουμε σε Blue-ray δίσκους των 50 GB έκαστος, τα δεδομένα που παράχθηκαν το 2018, τα οποία είναι 33 zettabytes, θα έπρεπε να επενδυθούν 660 δισεκατομμύρια δολάρια. Αναφορικά παρουσιάζεται η εικόνα 1 όπου φαίνεται η πρόβλεψη της δημιουργίας των παγκόσμιων δεδομένων μέχρι το 2035.



Εικόνα 1 Συγκριτικό διάγραμμα δημιουργίας δεδομένων [1]

Γίνεται αντιληπτή η ανάγκη λοιπόν για υποδομές όπου θα μπορούν αρχικά να αποθηκεύσουν τα δεδομένα αυτά αλλά μεγαλύτερη πρόκληση είναι η επεξεργασία τους. Πλέον η αποθήκευση των δεδομένων μέσω της ανάπτυξης μεγάλων datacenters έχει γίνει πιο φθηνή από ποτέ καθώς επίσης η ενοικίαση πόρων για την εξεργασία τους. Πλέον ζούμε στην εποχή όπου μία επιχείρηση δεν σπαταλάει πόρους στην αγορά υπολογιστικού εξοπλισμού αντιθέτως νοικιάζουν τον εξοπλισμό αναλόγως με τις

ανάγκες τους που μπορεί να διαφέρουν από μέρα σε μέρα. Τα πλεονεκτήματα είναι αμέτρητα καθώς εκτός από την χρονο-μίσθωση του εξοπλισμού παρέχεται και η ασφάλεια ότι το site που μπορεί να έχουν σηκώσει δεν έχει καθόλου διακοπές λειτουργίας (downtime) σε συνδυασμό του ότι δεν θα χρειάζεται η κάθε επιχείρηση να έχει κάποιο τμήμα στο μισθολόγιό της που να διαχειρίζεται αυτή την υπηρεσία.

Μια άλλη πολύ ενδιαφέρουσα στατιστική δημοσιεύθηκε από το FORBES [2] όπου αναφέρονται χαρακτηριστικά τα οικονομικά οφέλη από την ανάλυση των παγκόσμιων δεδομένων. Αναφορικά προβλέπεται αύξηση κατά 10.48% στα έσοδα της παγκόσμιας αγοράς των μεγάλων δεδομένων για λογισμικό και υπηρεσίες από τα 42 δις δολάρια που ήταν το 2018 στα 103 δις το 2026. Σε προέκταση αυτού έχουν δημιουργηθεί κλάδοι όπου χρησιμοποιώντας το τεράστιο όγκο δεδομένων αποφέρουν αποτελέσματα στην επιχειρησιακή λειτουργία και απόδοση σε βαθμό που 10 χρόνια πριν ούτε θα το φανταζόμασταν. Για παράδειγμα με τη χρήση αλγορίθμων μηχανικής μάθησης μπορούμε πλέον διαδικασίες που ήταν καθαρά χειρωνακτικές να της αυτοματοποιήσουμε αλλά παράλληλα η αποδοτικότητα του αλγορίθμου να αυξάνεται καθώς αυξάνονται τα δεδομένα.

1.1 Χαρακτηριστικά Μεγάλων Δεδομένων

Τα μεγάλα δεδομένα είναι μία «οντότητα» τόσο μεγάλη και πολύπλοκη που τα παραδοσιακά συστήματα αδυνατούν να τα διαχειριστούν είτε αναφερόμαστε σε αποθήκευση, ανάκτηση δεδομένων πόσο μάλλον για την επεξεργασία τους. Για το λόγο αυτό έχουν αναπτυχθεί συστήματα αποθήκευσης και επεξεργασίας τα οποία είναι κατανεμημένα όπως η MongoDB , Hadoop και SPARK τα οποία είναι από τα πιο διαδεδομένα συστήματα διαχείρισης μεγάλων δεδομένων. Τι είναι αυτό που ξεχωρίζει τα μεγάλα δεδομένα από τα «απλά» δεδομένα; Σίγουρα έχει να κάνει με το μέγεθος των δεδομένων σαν πλήθος αλλά δεν είναι μόνο αυτό. Η πιο ακριβής προσέγγιση για να χαρακτηρίσει κάποιος τα μεγάλα δεδομένα περιγράφεται εν συντομία από τα 5 V's. Τα πέντε V εκπροσωπούν όπως φαίνεται και στην εικόνα 2 [3], την «οντότητα» μεγάλα δεδομένα [4].

Εικόνα 2 5V's of Big Data [3]

Μέγεθος-Όγκος (Volume):

Όπως αναφέρει και το όνομα του χαρακτηριστικού έχει να κάνει με το καθαρό μέγεθος, αν μιλάμε για Bytes, αριθμό, αν μιλάμε για εγγραφές σε μία βάση και γενικά, για όγκο αν μιλάμε συνδυαστικά και για αριθμό εγγραφών σε πολλαπλές βάσεις αλλά και για μέγεθος αυτών των αποθηκευμένων εγγραφών. Με το ρυθμό που παράγονται τα δεδομένα αποτελεί πραγματική πρόκληση η αποθήκευσή τους, πόσο μάλλον η επεξεργασία τους.

Ταχύτητα (Velocity):

Με τον όρο ταχύτητα εννοείται η συνεχής αυξανόμενη ταχύτητα δημιουργίας των δεδομένων η οποία επιβάλλει με τη σειρά της υψηλές ταχύτητες επεξεργασίας και αποθήκευσης. Επιπλέον η σαν ταχύτητα

ορίζεται εκτός από τον ρυθμό παραγωγής δεδομένων αλλά επίσης και η ταχύτητα με την οποία τα δεδομένα μεταφέρονται.

Ποικιλομορφία (Variety):

Η ποικιλομορφία των δεδομένων έχει να κάνει με την μορφή, την δομή και το είδος των δεδομένων. Είναι εξαιρετικά απίθανό τα μεγάλα δεδομένα να είναι σε δομημένη μορφή, καθιστώντας την διαδικασία εισαγωγής τους σε σχεσιακές βάσεις δεδομένων μία πρόκληση. Είναι ένα από τους λόγους αυτό που η NoSQL βάσεις έχουν γίνει τόσο δημοφιλείς καθώς πλέον έχουμε αφήσει την εποχή που τα δεδομένα προς επεξεργασία ήταν δομημένα. Όπως πολύ χαρακτηριστικά αναφέρεται και στο [4], το 90% των παραγόμενων δεδομένων είναι μη-δομημένα.

Εγκυρότητα (Veracity):

Όπως προαναφέρθηκε η συντριπτική πλειοψηφία των δεδομένων προς αποθήκευση και εξεργασία είναι σε μη-δομημένη μορφή διατρέχοντας μεγάλο κίνδυνο για ανακρίβειες και λανθασμένα στοιχεία. Συνεπώς η ανάγκη για καθαρισμό των δεδομένων πριν την αποθήκευση είναι ζωτικής σημασίας καθώς ανεξαρτήτως μεγέθους δεδομένων και ταχύτητας παραγωγής, τα δεδομένα θα πρέπει να είναι έγκυρα προκειμένου να λάβουμε σωστές αποφάσεις.

Αξία (Value):

Τελευταίο αλλά για πολλούς το πιο σημαντικό των πέντε V’s είναι η αξία των δεδομένων. Όπως σε όλο τον επιχειρηματικό κόσμο δεν υπάρχει λόγος δημιουργίας ενός προϊόντος, όσο καλό και αν είναι αυτό, αν δεν υπάρχει κάποιο προβλεπόμενο κέρδος από την διαδικασία αυτή. Η διαδικασία της αποθήκευσης, της επεξεργασίας και της ανάλυσης των μεγάλων δεδομένων είναι αρκετά «ακριβή», είτε μιλάμε για εργατοώρες, είτε μιλάμε για εξοπλισμό. Επομένως αν δεν υπάρχει λόγος μία επιχείρηση να επενδύσει σε αυτούς τους τομείς (εξειδικευμένο προσωπικό, εξοπλισμό κλπ.) αν δεν υπάρχει η αντίστοιχη απόδοση, η αλλιώς return on investment.

1.2 Τομείς Εφαρμογής Μεγάλων Δεδομένων

Τα τελευταία χρόνια γίνεται αντιληπτή η ανάγκη για ανάλυση των παραγόμενων δεδομένων από τις επιχειρήσεις καθώς μέσω αυτής, αυξάνεται και ενισχύεται το επιχειρηματικό τους πλάνο. Το πιο φωτεινό παράδειγμα αυτής της ανάλυσης είναι η μηχανή αναζήτηση της google όπου αυτή εκτός από το ότι γίνεται συνολικά πιο αποδοτική σύμφωνα με τις αναζητήσεις που γίνονται παγκοσμίως από τους χρήστες αλλά επιπλέον παρέχει προϊόν αναζήτησης προσωποποιημένο για τον εκάστοτε χρήστη σύμφωνα με τις προηγούμενες αναζητήσεις του αλλά από την κίνησή του στο διαδίκτυο. Βλέπουμε λοιπόν ότι εκτός από την ανάγκη πλέον για καθολικά αναπτυγμένα συστήματα οι επιχειρήσεις θα πρέπει να προσανατολιστούν σε μία πιο πελατοκεντρική προσέγγιση. Με αυτό το τρόπο οι επιχειρήσεις όχι μόνο μπορούν σε κάθε χρήστη να προωθούν τα προϊόντα που κατά πάσα πιθανότητα θα αγόραζε, αυξάνοντας τα έσοδά τους, αλλά επίσης μπορούν πιο αποτελεσματικά από ποτέ να κάνουν προβλέψεις για το μέλλον τόσο για τις τάσεις της αγοράς επενδύοντας αναλόγως, όσο επίσης για το αν η επιχείρηση είναι βιώσιμη σε βάθος χρόνου.

Στη συνέχεια παρουσιάζονται κλάδοι όπου η εφαρμογή τακτικών ανάλυσης δεδομένων αύξησε, σημαντικά την αποδοτικότητά τους.

Τηλεπικοινωνίες:

Κυριότερος από τους κλάδους που έχει επηρεάσει σημαντικά η ανάλυση των δεδομένων είναι ο κλάδος των τηλεπικοινωνιακών παρόχων. Οι τηλεπικοινωνιακοί πάροχοι συλλέγουν το δευτερόλεπτο ασύλληπτα μεγάλους όγκους δεδομένων των χρηστών τους. Ενδεικτικά κάποια δεδομένα που συλλέγονται σύμφωνα με [5] είναι τα εξής:

- Δημογραφικά στοιχεία (Χώρα, Πόλη, Φύλλο, Ηλικία, Αξιολόγηση Πελάτη κλπ)
- Αναφορές κατανάλωσης (Μηνιαίο πάγιο, Τύπος Πακέτου, Συνολικές ώρες κλήσεων, Μέγεθος κίνησης δεδομένων, Τύπος υπηρεσίας κλπ)
- Αναφορές προτίμησης (Ώρα χρησιμοποίησης υπηρεσιών, Προτίμηση εφαρμογής, Μάρκα Τερματικού κλπ)

Βλέπουμε λοιπόν ότι θα ήταν αφέλεια από την πλευρά των παρόχων να μην χρησιμοποιήσουν τα συγκεκριμένα δεδομένα έτσι ώστε να πραγματοποιήσουν τιμολογιακές πολιτικές προσαρμοσμένες σε κάθε έναν από τους χρήστες τους. Με αυτό το τρόπο μπορούμε από την πλευρά του καταναλωτή να απολαμβάνουμε και να πληρώνουμε τις υπηρεσίες που μας ενδιαφέρουν κυρίως, αδιάκοπα χωρίς να ξεφεύγουμε από το όριο του παγίου. Από την πλευρά του παρόχου, βλέπουμε ότι τα πλεονεκτήματα είναι πολυεπίπεδα, καθώς με την ανάλυση των προτιμήσεων των χρηστών και σε συνδυασμό με τα δημογραφικά στοιχεία, υπάρχει η δυνατότητα της κατάλληλης παραμετροποίησης των υποδομών τους έτσι ώστε να μην παρατηρούνται πχ προβλήματα υπερφόρτωσης των δικτύων. Επιπλέον υπάρχει πληθώρα στοχευμένων προσαρμογών στα καταναλωτικά προϊόντα αναλόγως τις ανάγκες του κάθε χρήστη που έχουν σαν αποτέλεσμα την μεγιστοποίηση του κέρδους των εταιριών αυτών.

Υγεία:

Βέβαια δεν μπορούσε να λείπει ο τομέας της υγείας από αυτή την πελατοκεντρική προσέγγιση ή καλύτερα, προσαρμοσμένη εφαρμογή της ιατρικής στον εκάστοτε ασθενή. Η λήψη της απόφασης όσον αφορά την θεραπεία που θα δώσει ένας γιατρός σε ένα ασθενή είναι μία χρονοβόρα διαδικασία που δυστυχώς λόγω των πολλαπλών μεταβλητών που τη συνθέτουν υπάρχει πιθανότητα να είναι λανθασμένη. Η ανάλυση των δεδομένων μας προσφέρει εργαλεία έτσι ώστε να μπορέσουμε να κατηγοριοποιήσουμε επιτυχώς τον κάθε ασθενή σύμφωνα με το ιστορικό του, αλλά και να προβλέψουμε με κάποια πιθανότητα, παρενέργειες από μία θεραπεία ή την εύρεση της καλύτερης βάσει των καταγραφών από τα παγκόσμια ιστορικά της εκάστοτε ασθένειας. Ένα επιπλέον πολύ σημαντικό στοιχείο που προκύπτει από την ανάλυση των δεδομένων υγείας είναι ότι μπορούμε να διαχειριστούμε καλύτερα καταστάσεις έκτακτης ανάγκης όπως, παραδείγματος χάρει, η παγκόσμια πανδημία του Covid-19. Σε καθημερινή βάση βλέπουμε τον χάρτη, οποίος ήταν αποτέλεσμα της ανάλυσης δεδομένων των κρουσμάτων καθώς και δημογραφικών στοιχείων του κάθε ατόμου, με τις πιο «επικίνδυνες» περιοχές λόγω των κρουσμάτων και εφαρμόζοντας τα αντίστοιχα μέτρα πρόληψης στις συγκεκριμένες περιοχές μπορούμε να περιορίσουμε την εξάπλωση.

Μεταφορές:

Ένας αρκετά ανερχόμενος κλάδος εφαρμογής της επιστήμης των δεδομένων είναι ανάπτυξη «Έξυπνων συστημάτων μετακίνησης». Αυτό που παρακίνησε την ανάπτυξη των έξυπνων συστημάτων μετακίνησης ήταν πρωταρχικά η ελαχιστοποίηση των ατυχημάτων και η ενίσχυση της ασφάλειας στους δρόμους καθώς σύμφωνα με το [6] κάθε χρόνο πραγματοποιούνται περίπου 8 εκατομμύρια τροχαία ατυχήματα όπου έχουν σαν αποτέλεσμα τον τραυματισμό 7 εκατομμυρίων ανθρώπων και τον θάνατο άλλων 1.3

εκατομμυρίων. Είναι φυσικό επόμενο πλέον οι αυτοκινητοβιομηχανίες να ενσωματώνουν αισθητήρες περιμετρικά των οχημάτων αλλά και εσωτερικά με σκοπό τη συνεχή συλλογή δεδομένων την ανάλυσή τους, σε πραγματικό χρόνο, και τέλος της λήψη μιας απόφασης που μπορεί να είναι κρίσιμη όπως ένα σύστημα αυτόματου φρεναρίσματος, η αποτροπή αλλαγής λωρίδας. Επιπλέον σύστημα που αυτή τη στιγμή είναι επιχειρησιακό εμπλέκει μία κάμερα που εστιάζει συνεχώς στα μάτια του οδηγού όπου όταν αντιληφθεί ότι η συχνότητα που ανοιγοκλείνει τα μάτια του ο οδηγός αυξάνεται και σε συνδυασμό με δεδομένα από άλλους αισθητήρες, όπως αυτός της αλλαγής κατεύθυνσης, ανεβάζει την ένταση του ραδιοφώνου στο μέγιστο με σκοπό να αφυπνίσει τον οδηγό. Σημαντικό θέμα επίσης είναι η εξοικονόμηση χρόνου και ενέργειας, καθώς έρευνες έχουν δείξει ότι 90 εκατομμύρια ώρες περίπου, δαπανώνται λόγο κυκλοφοριακών προβλημάτων. Τέλος οι μετρήσεις των εκπομπών διοξειδίου του άνθρακα για το 2021 προβλέπεται να αυξηθούν κατά 4.8%, σε σχέση με αυτές του 2020 όπου υπήρχε σημαντική πτώση λόγω της παγκόσμιας πανδημίας του Covid-19, φτάνοντας τους 33 GT Co2 (1 Gt = 10^9 tones) [7]. Είναι απαραίτητη η εφαρμογή μέτρων και τεχνικών εκσυγχρονίζοντας τους τρόπους μετακίνησης παγκοσμίως για την ελαχιστοποίηση των εκπομπών διοξειδίου του άνθρακα. Γίνεται αντιληπτή λοιπόν η ανάγκη για εύρεση διαδρομών με λιγότερη κίνηση επιλέγοντας πιο ασφαλείς δρόμους μειώνοντας σημαντικά το χρόνο όπου το κάθε αυτοκίνητο ρυπαίνει το περιβάλλον.

Ναυτιλία.

Όντας η ναυτιλία η κινητήριος δύναμη του παγκοσμίου εμπορίου, με πάνω από τα τέσσερα πέμπτα των εμπορευμάτων να μεταφέρονται μέσω της θάλασσας, δεν θα μπορούσε να μην επηρεαστεί από τις μεγάλες αλλαγές της τεχνολογίας που συζητήθηκαν και νωρίτερα. Βρισκόμαστε στο μέσω μιας εποχής όπου όλο και περισσότερες εταιρείες επενδύουν στην τεχνολογία των σκαφών τους με στόχο την πιο άμεση αναφορά προβλημάτων αλλά κυριότερα την ταχύτερη λήψη μιας απόφασης. Η αυτοματοποίηση των συστημάτων διεύθυνσης, η αυτοματοποίηση της κατασκευής των σκαφών, η τηλεμετρία με ηλεκτρονικούς αισθητήρες αντί των προγενέστερων μηχανικών (θερμοκρασίας, πίεσης κλπ.) καθώς επίσης και η έρευνα για την αυτοματοποίηση ακόμα και των πιο απλών διαδικασιών μέσα σε ένα πλοίο, παράγουν χρήσιμη πληροφορία η οποία, μπορεί να αποτρέψει την οικονομική ζημία μιας «κακής» απόφασης. Τέλος ο συνδυασμός όλων αυτών των δεδομένων είναι απαραίτητος προκειμένου να φτάσουμε στο σημείο όπου θα μπορούμε να έχουμε μία αποδοτική οικονομικά, περιβαλλοντολογικά ευαισθητοποιημένη, μη-επανδρωμένη ναυτιλία [8].

Αθλητισμός:

Ο αθλητισμός παράγει καθημερινά πολύ μεγάλο πλήθος δεδομένων τα οποία είναι σχετικά με τους παίχτες, τις επιδόσεις της ομάδας αλλά και το κοινό. Καθώς οι απαιτήσεις για πιο έγκυρα στατιστικά αυξάνονται έτσι λοιπόν η ανάλυση των μεγάλων αυτών όγκων δεδομένων είναι η μόνη λύση. Από την παρακολούθηση πιθανών ταλέντων, την επιλογή της σωστής τακτικής έναντι ενός συγκεκριμένου συστήματος σε ένα αγώνα ποδοσφαίρου, τις προωθητικές ενέργειες της ομάδας προς τους φιλάθλους, η εποπτεία της αγωνιστικής κατάστασης των παικτών από το προπονητικό επιτελείο έως και την ανάλυση και θέσπιση υλοποιήσιμων στόχων (πρωταθλημάτων, κυπέλων κλπ.) για τη χρονιά αναλύοντας τα δεδομένα των «αντίπαλων» ομάδων, είναι όλα προϊόν ανάλυσης μεγάλων δεδομένων. Είναι μία εξαιρετικά μεγάλη αγορά αν αναλογιστεί κανείς σε παγκόσμιο επίπεδο, εκτιμάται μία αύξηση από τα 388.3 δισεκατομμύρια δολάρια, που ήταν το 2020, στα 599 δισεκατομμύρια δολάρια το 2025 σύμφωνα με [9].

Οικονομία:

Τέλος μεγαλύτερη αξιοποίηση των μεγάλων δεδομένων φαίνεται να πραγματοποιείται από το κλάδο της οικονομίας και κατ' επέκταση το χρηματοπιστωτικό και τραπεζικό σύστημα. Σε καθημερινή βάση τράπεζες συλλέγουν και αξιολογούν πολλές εκατοντάδες Giga byte δεδομένων, όπως δημογραφικά στοιχεία πελατών, κινήσεις πελατών, είτε αυτές είναι εμβάσματα είτε πληρωμές με χρεωστικές και πιστωτικές κάρτες, γεωγραφικά στοιχεία, ύψος συναλλαγών κλπ. Ο όγκος της πληροφορίας είναι πολύ μεγάλος και χρησιμοποιείται για την διασφάλιση της καλής ποιότητας των υπηρεσιών, την αποφυγή κακόβουλων ενεργειών, προώθηση προϊόντων στοχευμένα σε πελάτες που η τράπεζα είτε θέλει να διατηρήσει και είναι πιθανό να φύγουν (churners), είτε να επιβραβεύσει τους «καλούς» πελάτες. Ο στόχος όπως προαναφέρθηκε είναι ο ίδιος, η πελατοκεντρική προσέγγιση των προϊόντων, η εύρεση τρόπων αναγνώρισης οικονομικής απάτης αλλά και επίτευξη ενός αυτοματοποιημένου τρόπου ικανοποίησης των απαιτούμενων διεργασιών χωρίς να αυξάνεται η πολυπλοκότητα από την πλευρά του χρήστη-πελάτη.

2. Τα Μεγάλα Δεδομένα στο Τραπεζικό Τομέα

Όπως προαναφέρθηκε η ανάπτυξη της τεχνολογίας της πληροφορικής η οποία έχει ενσωματωθεί σε πολλαπλά πεδία της σύγχρονης αγοράς έφερε μεγάλες αλλαγές και στον τρόπο λειτουργίας στο τομέα της τραπεζικής. Με την ψηφιοποίηση των τραπεζικών συναλλαγών, τα δεδομένα που συλλέγονται παγκοσμίως για την διατήρηση της καλής ποιότητας των προϊόντων τους αλλά και την βελτιστοποίηση της σχέσης με τους πελάτες τους, καθώς και η αξιοποίηση νέων τεχνολογιών όπως τα υπολογιστικά νέφη (cloud-computing), blockchain και η τεχνητή-νοημοσύνη (AI), επέφεραν μία πρωτοφανή ευκαιρία για την ανάπτυξη του τραπεζικού συστήματος παγκοσμίως. Χωρίς καμία αμφιβολία η κατοχή των δεδομένων, είναι ένας εν δυνάμει πολύ αποδοτικός πόρος, αλλά λόγω της ποικιλομορφίας και του όγκου των δεδομένων, η σωστή αξιοποίησή τους αποτελεί μία πρόκληση και σίγουρα η τραπεζική βιομηχανία θα πρέπει να κάνει την έρευνά της για την αποδοτικότερη αξιοποίησή τους. Είναι κατανοητό λοιπόν ότι με την αξιοποίηση των δεδομένων, ο τραπεζικός τομέας μπορεί να επωφεληθεί σε παραπάνω από ένα σημεία. Συνεπώς τη δεδομένη στιγμή ο τραπεζικός τομέας θα πρέπει να επικεντρωθεί στην λήψη όλο και περισσότερων δεδομένων από πολλαπλές πηγές με σκοπό την αξιολόγησή τους σε πραγματικό χρόνο (real-time) και την πιο επικυρωμένη, ως προς το χρόνο, λήψη απόφασης. Όπως αναφέρεται χαρακτηριστικά και στο [10], θα πρέπει τραπεζική βιομηχανία να επεκτείνει τα κανάλια εισροής των δεδομένων, καθώς τα δεδομένα συναλλαγών που κυρίως κατέχει είναι ιστορικά δεδομένα και δεν είναι δεδομένα πραγματικού χρόνου. Συνεπώς θα πρέπει να στραφεί σε πηγές όπως κυβερνητικά δεδομένα, IoT εφαρμογές αλλά κυρίως να διαμορφώσει «συμμαχίες» με άλλους κλάδους με σκοπό την διαμοίραση δεδομένων που εκείνοι διαθέτουν.

Έχοντας σαν γνώμονα την πελατοκεντρική προσέγγιση, είτε του marketing είτε και του ίδιου του προϊόντος, θα πρέπει ο τραπεζικός κλάδος να χαρτογραφεί την κίνηση των καταναλωτών, παραδείγματος χάρει σε κοινωνικά δίκτυα (social media), για την ανακάλυψη καταναλωτικών αναγκών, που έχει σαν σκοπό την άμεση ανταπόκριση της τράπεζας για την κάλυψη αυτών. Ένα άλλο παράδειγμα εφαρμογής έχει να κάνει με την ανίχνευση απάτης, και για το πως τα δεδομένα πραγματικού χρόνου από τρίτους φορείς μπορούν να ενισχύσουν ή και να επισπεύσουν την έναρξη των αντίμετρων που έχουν οριστεί, με σκοπό την διασφάλιση των πελατών τους. Αν παραδείγματος χάρει ένα τερματικό ηλεκτρονικής πώλησης (POS) ζητήσει άδεια για την πληρωμή ενός προϊόντος ενώ το κινητό του χρήστη,

μέσω των δεδομένων που συλλέγουν οι εταιρείες κινητής τηλεφωνίας, γνωρίζουμε ότι βρίσκεται σε διαφορετική περιοχή από αυτή που γίνεται η χρέωση, τότε η τράπεζα μπορεί να μπλοκάρει αυτή τη συναλλαγή αποτρέποντας την χρέωση του πελάτη της από κάποιο άτομο που πιθανόν του έκλεψε την χρεωστική κάρτα. Μία τέτοια συνεργασία μεταξύ παρόχων κινητής τηλεφωνίας και τραπεζών αυξάνει κατά πολύ τα επίπεδα ασφάλειας των συναλλαγών της δεύτερης, η οποία μπορεί με την σειρά της να το διαφημίσει και να κερδίσει μεγάλο μέρος της αγοράς καθώς η ασφάλεια των συναλλαγών είναι ένα κύριο ζήτημα. Βέβαια η αύξηση των καναλιών εισροής των δεδομένων απαιτεί την ανάπτυξη βελτιωμένων τεχνικών εξόρυξης γνώσης καθώς, τα σημερινά συστήματα αναλύουν κυρίως δομημένα δεδομένα (structured data) και όχι μερικώς (semi-structured data) ή τελείως αδόμητα (unstructured data). Είναι λοιπόν τεχνική η πρόκληση της ανάλυσης και η εξαγωγή της πληροφορίας από αδόμητα δεδομένα, που είναι ένα βίντεο ή ανάλυση κειμένου, για την λήψη feedback μέσω των κοινωνικών δικτύων φέρ' ειπείν.

Είναι γνωστό ότι παγκοσμίως ο τραπεζικός τομέας, αναγνωρίζει την σημασία της πληροφορίας που απορρέει από τα δεδομένα των συναλλαγών των πελατών τους και για το λόγο αυτό διατηρεί μία από τις πιο «πλούσιες» πηγές δεδομένων αναφορικά με πληροφορίες πελατών (δημογραφικά στοιχεία, αρχεία συναλλαγών, μοτίβα χρησιμοποίησης πιστωτικών καρτών κλπ.) [11].

Στη συνέχεια παρουσιάζονται μερικές από τις πιο διαδιδόμενες εφαρμογές των μεγάλων δεδομένων στα σύγχρονα τραπεζικά συστήματα, στις οποίες, η εξόρυξη γνώσης από τα δεδομένα είναι απαραίτητη.

2.1 Ανίχνευση Απάτης (Fraud Detection)

Αρχικά θα πρέπει να δοθεί ένας ορισμός για το τι είναι ακριβώς είναι η απάτη και πιο συγκεκριμένα η οικονομική απάτη. Απάτη ορίζεται η εν γνώσει παράσταση ψευδών γεγονότων σαν αληθινών με σκοπό την παράνομη οικονομική ωφέλεια [12]. Υπάρχουν πολλαπλοί τρόποι άσκησης απάτης στα πλαίσια των τραπεζικών συναλλαγών με εξαιρετικά δυσμενή αποτελέσματα καθώς η οικονομική ζημία είναι τεράστια. Αναφορικά το 2020 σύμφωνα με τη Javelin Strategy & research [13] το κόστος ανήλθε στο ύψος των 56 δισεκατομμυρίων δολαρίων. Μερικές από τις πιο διαδεδομένες μορφές απάτης στον τραπεζικό τομέα είναι οι εξής:

- Απάτες σχετικές με χρεωστικές ή πιστωτικές κάρτες. Εδώ συμπεριλαμβάνονται όλες οι κινήσεις που επικυρώνονται από μία κάρτα, είτε αυτή είναι απομακρυσμένη χρέωση είτε είναι από κοντά που είναι προϊόν κλοπής ή αντιγραφής της κάρτας.
- Απάτη επιταγών.
- Απάτη μη εξουσιοδοτημένης απομακρυσμένης πρόσβασης η οποία μπορεί να επιτευχθεί είτε επιτυγχάνοντας πρόσβαση στο λογαριασμό μέσω του διαδικτύου, είτε μέσω τηλεφώνου ή τέλος είτε μέσω των εφαρμογών κινητών τηλεφώνων.
- Τέλος η εξουσιοδοτημένη εντολή πληρωμής σε κάποιο λογαριασμό που ανήκει στον θύτη καθώς το θύμα έχει εξαπατηθεί.

Σύμφωνα με [14] η οποία είναι μία αναφορά που περιλαμβάνει μία επισκόπηση των στατιστικών των περιπτώσεων απάτης στο Ηνωμένο Βασίλειο, παρατηρείται μία μείωση της τάξης του 5% σε σχέση με το 2019. Όμως κοιτάζοντας λίγο πιο προσεκτικά θα δούμε παρόλο που υπάρχει μία γενική μείωση κάποιες μορφές απάτης έχουν αρχίσει και αυξάνονται και σε αριθμό περιστατικών αλλά και σε αθροιστική ζημία. Για παράδειγμα καθώς λόγω της ευκολίας που μας παρέχει η αξιοποίηση της τεχνολογίας μπορούμε να φανταστούμε ότι οι απάτες μέσω του διαδικτύου θα αυξηθούν ενώ παραδείγματος χάρει οι απάτες επιταγών θα μειωθούν καθώς δεν χρησιμοποιούν και πολλοί επιταγές σήμερα. Οι εικόνες 3 και 4 δείχνουν ακριβώς το παραπάνω.

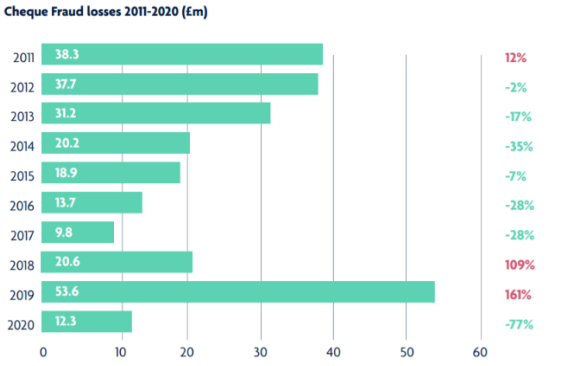
Remote Banking Fraud losses 2013-2020

Remote banking values	2013	2014	2015	2016	2017	2018	2019	2020	19/20 % Change
Internet banking	£58.8m	£81.4m	£133.5m	£101.8m	£121.2m	£123.0m	£111.8m	£159.7m	43%
Telephone banking	£13.1m	£16.8m	£32.3m	£29.6m	£28.4m	£22.0m	£23.6m	£16.1m	-32%
Mobile banking	N/A	N/A	£2.8m	£5.7m	£6.5m	£7.9m	£15.2m	£21.6m	41%
TOTAL	**£71.7m**	**£71.9m**	**£98.2m**	**£137.0m**	**£156.1m**	**£152.9m**	**£150.7m**	**£197.3m**	**31%**

Annual case volumes Remote Banking fraud 2013-2020

Remote banking cases	2013	2014	2015	2016	2017	2018	2019	2020	19/20 % Change
Internet banking	13,799	16,041	19,691	20,088	21,745	20,904	25,849	55,995	117%
Telephone banking	5,596	5,578	11,380	10,495	9,577	7,937	11,199	7,490	-33%
Mobile banking	N/A	N/A	2,235	2,809	3,424	2,956	6,872	10,155	48%
Total	**19,395**	**21,819**	**33,306**	**33,392**	**34,746**	**31,797**	**43,920**	**73,640**	**68%**

Εικόνα 3 Διαχρονική Επισκόπηση Περιστατικών Απάτης internet Banking [14]



Εικόνα 4 Διαχρονική Επισκόπηση Μεταβολής της Απάτης Επιταγών [14]

Βλέπουμε ότι η απάτη των επιταγών είτε μέσω των αντίμετρων που έχουν εφαρμόσει οι τράπεζες του Ηνωμένου Βασιλείου, είτε επειδή δεν χρησιμοποιούνται πλέον ή ακόμη, επειδή υπάρχουν πιο «εύκολοι» τρόποι να πραγματοποιηθεί η απάτη, βλέπουμε μία πτώση το 2020 σε σχέση με το 2019 της τάξης του 77%. Από την άλλη πλευρά στην εικόνα 4 του internet banking βλέπουμε ότι έχουμε αύξηση της συνολικής ζημίας από το 2019 στο 2020 κατά 31% που ανέρχεται στο ύψος των 197,3 εκατομμυρίων λιρών αλλά ακόμη πιο ανησυχητικό είναι ότι υπάρχει αύξηση κατά 68% των συνολικών περιστατικών.

2.2 Προσωποποιημένη Προώθηση Προϊόντων (Personalized Marketing)

Η προώθηση των προϊόντων είναι μία από τις πρώτες, αλλά και πιο επικερδείς εφαρμογές των τεχνικών της εξόρυξης γνώσης. Με μία πιο γενική οπτική χωρίς να αναφερόμαστε αποκλειστικά στο τραπεζικό τομέα, μέσω της εξόρυξης γνώσης μπορεί να γίνει αναγνώριση και κατηγοριοποίηση των πελατών ενός οργανισμού με σκοπό την επιλογή εκείνων που είναι πιο πιθανό να ανταποκριθούν σε μία προωθητική ενέργεια. Με αυτό το τρόπο το τμήμα προώθησης, μπορεί να περιορίσει κατά πολύ το κόστος, που μπορεί παραδείγματος χάρει να είναι η σπατάλη εργατοωρών για την εύρεση των πιο «αποδοτικών» ενεργειών χωρίς τη χρήση των τεχνικών της εξόρυξης γνώσης, ή ακόμα αν αναλυθεί η συνολικό κόστος της προωθητικής ενέργειας ανά πελάτη μπορεί να διαστασιοποιηθεί το κόστος και να επιτευχθεί μεγαλύτερη απόδοση της επένδυσης (Return on investment - ROI) και κατ' επέκταση μεγαλύτερα κέρδη αν επικεντρωθεί η στρατηγική στους πελάτες που είναι πιο πιθανόν να ανταποκριθούν. Αυτά τα δεδομένα προέρχονται από βάσεις δεδομένων που συντηρούν είτε ίδιοι οι οργανισμοί είτε από δεδομένα που είναι ελεύθερα όπως π.χ. η εκδήλωση ενδιαφέροντος καταναλωτών σε ορισμένα προϊόντα από μέσα κοινωνικής δικτύωσης, από κινήσεις πιστωτικών καρτών, ύψος μισθού κλπ. Βάσει των προαναφερόμενων χαρακτηριστικών και όχι μόνο, οι οργανισμοί προσπαθούν να μοντελοποιήσουν τη δραστηριότητα των πελατών με απώτερο σκοπό την πρόβλεψη των ατόμων που θα αγοράσουν ένα προϊόν. Αυτή η κίνηση των οργανισμών είναι αναγκαία καθώς η σύγχρονη αγορά συνεχώς αλλάζει και η εξόρυξη γνώσης είναι ένα «όπλο» των επιχειρήσεων να προβλέπουν της αλλαγές αυτές αντί του απλά να αντιδρούν αφού αυτές έχουν πραγματοποιηθεί.

Όπως χαρακτηριστικά αναφέρεται στο[15] υπάρχουν πολλές εφαρμογές της εξόρυξης γνώσης στη προώθηση προιόντων εκ των οποίων οι πιο διαδεδομένες είναι οι εξής:

- Απόκτηση νέων πελατών (Customer Acquisition): Μέσω των τεχνικών της εξόρυξης γνώσης γνωστοποιούνται τα χαρακτηριστικά εκείνα που εντοπίζονται σε πελάτες που ανταποκρίθηκαν θετικά σε μία προωθητική ενέργεια ή έχουν αγοράσει ένα προιόν. Στη συνέχεια τα

χαρακτηριστικά αυτά συγκρίνονται με εκείνα ατόμων που δεν είναι πελάτες με σκοπό την προσέγγιση νέων αγοραστών.

- Διατήρηση πελατών (Customer Retention): Βάσει των δεδομένων ενός οργανισμού είναι δυνατή η πρόβλεψη για το αν ένα άτομο θα συνεχίσει να είναι πελάτης ή όχι. Σε περίπτωση που τα δεδομένα δείχνουν ότι το συγκεκριμένο θα τερματίσει αυτή τη πελατειακή σχέση τότε ο οργανισμός μπορεί να το αποτρέψει αυτό κάνοντας προωθητικές ενέργειες με προσφορές και διάφορα άλλα κίνητρα με σκοπό να διατηρήσει το άτομο αυτό στο πελατολόγιο του.
- Εγκατάλειψη πελάτη (Customer Abandonment): Σε αντίθεση με τις προηγούμενες δύο εφαρμογές που έχουν σαν σκοπό την εύρεση περισσότερων πελατών ή την διατήρηση των υπαρχόντων, η εξόρυξη γνώσης μπορεί να ανιχνεύσει αν κάποιος πελάτης έχει αρνητικό αντίκτυπο στο συνολικό αποτέλεσμα της επιχείρησης.
- Ανάλυση καλαθιού αγοράς (Market Basket Analysis): Σκοπός της εφαρμογής αυτής είναι η εύρεση μοτίβων μεταξύ αγορασμένων προϊόντων των συνολικών πελατών ενός οργανισμού με στόχο την ανίχνευση συσχετίσεων στις αγοραστικές συνήθειες και ανάγκες. Έτσι λοιπόν όταν ένας νέος πελάτης αγοράσει ένα προϊόν που η πλειοψηφία των πελατών το συμπλήρωσε με κάτι άλλο τότε αυτοματοποιημένα το σύστημα προωθεί το συμπληρωματικό προϊόν στο νέο αυτό πελάτη. Για παράδειγμα αν η πλειοψηφία των πελατών ενός καταστήματος ηλεκτρονικών ειδών έχει αγοράσει ένα ηλεκτρονικό υπολογιστή και ένα πληκτρολόγιο και ένας νέος πελάτης έχει αγοράσει ένα υπολογιστή και δεν έχει επιλέξει πληκτρολόγιο, τότε βάσει της ανάλυσης αυτής θα του προωθηθεί και το αντίστοιχο πληκτρολόγιο.

Στο τραπεζικό τομέα πιο συγκεκριμένα όπου διατηρούνται εκτενής βάσεις δεδομένων με δημογραφικά στοιχεία πελατών, συναλλαγές κλπ. όπως προαναφέρθηκε έτσι και εδώ, υπάρχει η δυνατότητα προσέγγισης των πελατών με τη χρήση της προσωποποιημένης προώθησης προϊόντων. Με την αντίστοιχη ανάλυση οι τράπεζες έχουν την δυνατότητα να δημιουργούν πελατειακά προφίλ όπου επιχειρείται η προώθηση προϊόντων ή υπηρεσιών που με μεγάλη πιθανότητα να ενδιαφέρουν τους αντίστοιχους πελάτες.

2.3 Διαχείριση Ρίσκου (Risk Management)

Υψηλή προτεραιότητα δίνουν οι τραπεζικοί οργανισμοί στην διαχείριση του ρίσκου μέσω της μελέτης των μεγάλων δεδομένων, καθώς μέσω αυτής μπορούν να αξιολογήσουν αν κάποιος, βάσει χαρακτηριστικών, μπορεί να αποπληρώσει ένα πιθανό δάνειο. Στην τελική το κύριο προϊόν που προσφέρει ένας τραπεζικός οργανισμός είναι ο δανεισμός. Μέσω της εξόρυξης γνώσης είναι δυνατό να αξιολογηθεί το επίπεδο του ρίσκου που λαμβάνει η τράπεζα δανείζοντας σε κάποιον ο οποίος μπορεί να μην είναι ικανός για την αποπληρωμή του δανείου του. Σε αυτές τις περιπτώσεις η τράπεζα είναι σε θέση να αξιολογήσει τους πιθανούς πελάτες της βάσει των ιστορικών στοιχείων που κατέχει από «καλούς» αλλά και «κακούς» δανειολήπτες, με σκοπό την κατηγοριοποίηση των νέων σε ομάδες προβλέποντας την συμπεριφορά τους. Εκτός από το δανεισμό αυτό κάθε αυτό, η διαχείριση ρίσκου είναι επιτακτική να χρησιμοποιείται και σε μικρότερης κλίμακας ποσά που είναι διαθέσιμα μέσω των πιστωτικών καρτών. Είναι δηλαδή υπολογίσιμο το μηνιαίο ποσό που μπορεί να διαθέσει η τράπεζα στο πελάτη της και αυτό αναλόγως την μεταβολή των χαρακτηριστικών που τον περιγράφουν (πελάτη) να αυξηθεί αλλά και να μειωθεί, αν παραδείγματος χάρει ο μισθολογικός παράγοντας μεταβληθεί αντίστοιχα. Συνήθως αυτή η αξιολόγηση γίνεται με τη χρήση ενός μέτρου που ονομάζεται πιστωτική βαθμολογία (credit scoring). Όταν ένας πιθανός πελάτης ζητάει δάνειο από μία τράπεζα είναι απαραίτητο να προσκομίσει τα απαραίτητα έγγραφα που πιστοποιούν τα κριτήρια που ορίζει η πολιτική της τράπεζας. Έχοντας συλλέξει όλα τα στοιχεία η τράπεζα πρέπει να πάρει μία απόφαση για το αν θα δανειοδοτήσει το συγκεκριμένο

άτομο ή όχι. Βάσει των στοιχείων αλλά και των κριτηρίων αυτών αποδίδεται η πιστωτική βαθμολογία του ατόμου και αν αυτή συμφωνεί με τη πολιτική της τράπεζας τότε μπορεί να προχωρήσει τη διαδικασία. Ο πιστωτικός έλεγχος αποσκοπεί στην επιτάχυνσης της συνολικής διαδικασίας αλλά κυρίως στον υπολογισμό της συνέπειας της αποπληρωμής του δανείου με το πιο αποδοτικό από πλευράς τράπεζας επιτοκίου [16].

2.4 Αξία του Χρόνου Ζωής του Πελάτη (Customer Lifetime Value Prediction - CLV)

Η αξία του χρόνου ζωής των πελατών είναι ένας από τους σημαντικότερους δείκτες απόδοσης (Key Performance Indicator - KPI) καθώς διαστασιοποιεί το όφελος ενός οργανισμού, σε όλη τη διάρκεια της συνεργασίας τους, από ένα πελάτη ή αντίστοιχα από μία ομάδα πελατών με κοινά χαρακτηριστικά. Η σημασία του συγκεκριμένου δείκτη μπορεί να κατανοηθεί πολύ καλύτερα αν αναλογιστεί κανείς ότι σύμφωνα με [17, 18] είναι πέντε με εφτά φορές ποιο ακριβό να αποκτήσεις νέους πελάτες από το να προσπαθήσεις να διατηρήσεις τους ήδη υπάρχοντες. Επιπλέον στη σημερινή ανταγωνιστική αγορά, όπως έχουν δείξει έρευνες [19], τα αποτελέσματα είναι πιο καταστροφικά αν ένας οργανισμός, π.χ. μία τράπεζα, δεν επενδύει καθόλου στην προσέλκυση νέων πελατών και επικεντρώνεται στη διατήρηση των πελατών του. Μία τέτοια ενέργεια χαρακτηρίζει το οργανισμό ως στάσιμο και σε ένα συνεχώς μεταβαλλόμενο περιβάλλον αν δεν προσαρμοστεί το πιο πιθανό είναι να μην επιβιώσει.

Η χρησιμότητα του δείκτη CLV γίνεται κατανοητή εδώ καθώς μέσω αυτού, ο οργανισμός μπορεί να επενδύσει στην απόκτηση νέων πελατών στοχεύοντας εκείνο το σύνολο που θα έχει μεγαλύτερο όφελος. Έτσι μπορεί να υπάρξει μία ισορροπία μεταξύ των δύο αντικρουόμενων στρατηγικών, που είναι η απόκτηση νέων πελατών και η διατήρηση των ήδη υπαρχόντων. Βλέπουμε λοιπόν ότι είναι αναγκαία η κατηγοριοποίηση σύμφωνα με την προβλεπόμενη αξία του πελάτη, εφαρμόζοντας τις αντίστοιχες προωθητικές ενέργειες σε αυτούς που επιθυμούμε να αποκτήσουμε ή αντίθετα να μην επιχειρήσουμε να διατηρήσουμε κάποιο πελάτη που είναι «χαμηλής» αξίας. Η μεγιστοποίηση του κέρδους και η ελαχιστοποίηση του κόστους είναι ο απώτερος στόχος, και μέσω της εξόρυξης γνώσης είναι δυνατή η λήψη των σωστών αποφάσεων προς αυτή τη κατεύθυνση.

2.5 Τμηματοποίηση Πελατών (Customer Segmentation)

Με τη αποτελεσματική τμηματοποίηση των πελατών των τραπεζών είναι δυνατές οι περισσότερες από τις εφαρμογές που προαναφέρθηκαν όπως το προσωποποιημένο marketing ή, η εκτίμηση της αξίας χρόνου ζωής του πελάτη ή ακόμη μπορεί να αποτελέσει και μία ένδειξη για κάποια μελλοντική πράξη απάτης. Είναι λοιπόν κατανοητό ότι η τμηματοποίηση, αξιοποιώντας τα δεδομένα που συλλέγουν οι τράπεζες είτε εσωτερικά (καταγραφή συναλλαγών), είτε με τις συνεργασίες με άλλους φορείς, δημόσιους και ιδιωτικούς είναι απαραίτητη, προκειμένου να είναι ανταγωνιστική στην «εχθρική» αυτή αγορά. Τα πιο διαδεδομένα χαρακτηριστικά τμηματοποίησης είναι γεωγραφικά, δημογραφικά και οικονομικά. Παρόλα αυτά με τη χρήση των ψηφιακών αποτυπωμάτων π.χ. που αφήνουν καθημερινά οι χρήστες του Internet banking, είναι δυνατή η κατηγοριοποίηση με βάση τα συμπεριφορικά δεδομένα των καταναλωτών, ή αλλιώς ψυχογραφική τμηματοποίηση (psychographic segmentation) [20], κατηγοριοποιώντας ακόμη περισσότερο τους καταναλωτές που μπορεί με τα παραδοσιακά χαρακτηριστικά να βρίσκονται στην ίδια ομάδα. Με την ανάλυση των συμπεριφορικών δεδομένων σε συνδυασμό με τα παραδοσιακά χαρακτηριστικά τμηματοποίησης, ο τραπεζικός τομέας βρίσκεται στη προνομιούχα θέση να «γνωρίζει» καλύτερα από τον καθένα τους πελάτες του. Χρησιμοποιώντας αυτή τη γνώση ο τραπεζικός τομέας δημιουργεί προσωποποιημένα προϊόντα, στοχεύοντας αποτελεσματικότερα τις διαφημιστικές εκστρατείες και κάνοντας προσφορές στις ομάδες των καταναλωτών που έχουν χαρακτηριστεί ως αξιόπιστοι ή υψηλής «αξίας». Πιο συγκεκριμένα η αξιοποίηση της τμηματοποίησης των πελατών επηρεάζει θετικά:

- Την απόκτηση νέων πελατών
- Την αύξηση των πωλήσεων των προϊόντων
- Την αύξηση των πιο προσοδοφόρων πελατών, σε συνδυασμό με την πρόβλεψη αξίας χρόνου ζωής των πελατών της τράπεζας, βρίσκοντας πιθανούς πελάτες που παρουσιάζουν κοινά χαρακτηριστικά με τους ήδη υπάρχοντες,
- Τη μείωση των churners, δηλαδή των πελατών που αφήνουν τον οργανισμό για κάποιο ανταγωνιστή. Με την τμηματοποίηση η ικανοποίηση των πελατών αυξάνεται που οδηγεί και στην αύξηση της πιθανότητας παραμονής στον οργανισμό
- Τις προωθητικές ενέργειες των τραπεζών

2.6 Συστήματα Συστάσεων (Recommendation Systems)

Τα συστήματα συστάσεων έχουν μονοπωλήσει το ενδιαφέρον καθώς με την εφαρμογή αλγορίθμων εξόρυξης γνώσης σε συνδυασμό με αλγορίθμους μηχανικής μάθησης μπορούν να παρέχουν αξιόπιστες συστάσεις, οδηγώντας σε μεγαλύτερη κερδοφορία τους οργανισμούς που τα αξιοποιούν. Έτσι μην αποτελώντας εξαίρεση και ο τραπεζικός τομέας μπορεί να επωφεληθεί σημαντικά από αυτά. Όπως φαίνεται και στο [21] από το ιστορικό των συναλλαγών των πιστωτικών καρτών των χρηστών σε συνδυασμό με τα γεωγραφικά δεδομένα που μοιράζονται οι χρήστες μέσω των κινητών τους τηλεφώνων, μπορούν να προτείνουν ευκαιρίες για αγορά προϊόντων κοντά στο χρήστη. Όπως αναφέρεται χαρακτηριστικά και στη σχετική έρευνα, η πλειοψηφία των χρηστών βρήκε την εφαρμογή αυτού του συστήματος ικανοποιητική.

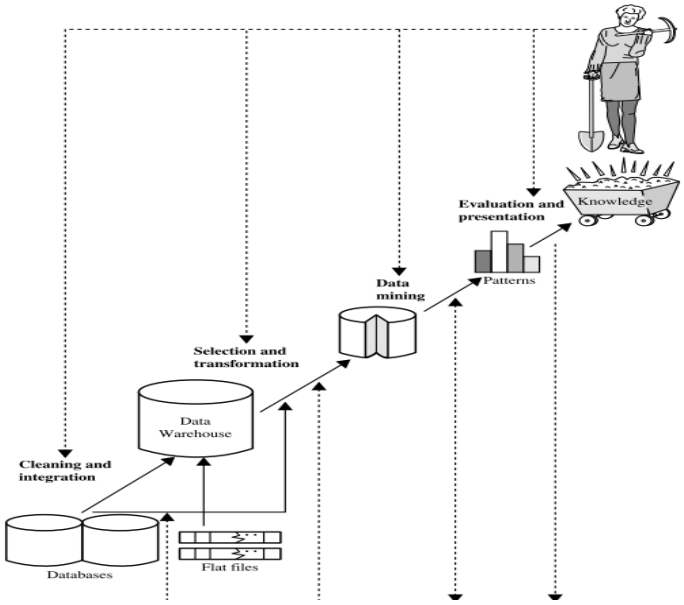
Εκτός από την προσωποποιημένη σύσταση για επίσκεψη σε κάποιο κατάστημα, μέσω του internet banking είναι δυνατό, ένας πελάτης σύμφωνα με τις καταναλωτικές του συνήθειες, να του προταθούν πακέτα δανεισμού, που θα του επιτρέψουν από τη μία να ικανοποιήσει τις ανάγκες του μετριάζοντας ταυτόχρονα το κίνδυνο από πλευρά τράπεζας, καθώς η πρόταση θα γίνει αν και μόνο αν πελάτης πληροί τις προϋποθέσεις. Μπορεί να εφαρμοστεί σαν ένα μέσω πληροφόρησης για τα διαθέσιμα σε αυτόν, τραπεζικά προϊόντα, δίχως να είναι απαραίτητη η σχετική και χρονοβόρα έρευνα από τη πλευρά του καταναλωτή, αλλά επίσης χωρίς τη δέσμευση πόρων, ανθρώπινων και όχι μόνο, από τη πλευρά της τράπεζας.

Ένας άλλο τομέας που μπορεί να αξιοποιηθεί από τον τραπεζικό τομέα είναι η προώθηση συστάσεων σε πελάτες που δείχνουν ενδιαφέρον σε στεγαστικό δανεισμό. Απόκτηση ακίνητης περιουσίας είτε πρόκειται για την κάλυψη της ανάγκης της στέγασης είτε για επενδυτικούς σκοπούς, είναι μία ακριβή συναλλαγή, πράγμα που κάνει τους αγοραστές να δαπανούν πολύ μεγάλο μέρος του χρόνου τους στην εύρεση της καλύτερης επιλογής. Η επιλογή αυτή αποτελείται από πολλαπλές μεταβλητές που σχετίζονται με την τιμή του ακινήτου, την τοποθεσία του, την εκτίμηση της αξίας του σε βάθος χρόνου και πολλές άλλες. Χαρακτηριστικές έρευνες είναι αυτές των Yuan [22] και Daly [23], όπου στην πρώτη γίνεται λόγος για την εξοικονόμηση χρόνου στην εύρεση του κατάλληλου ακινήτου, ενώ η δεύτερη πραγματεύεται ένα τρόπο υπολογισμού του χρόνου που θα δαπανάται στη μεταφορά, από και προς την οικία, την εργασία και άλλα σημεία ενδιαφέροντος, με σκοπό την σύσταση εκείνων των ακινήτων που θα έχουν το μικρότερο μέσο χρόνο μετακίνησης.

3. Εξόρυξη Δεδομένων (Data Mining)

Εξόρυξη δεδομένων (data mining) ή αλλιώς όπως αναφέρεται στη βιβλιογραφία ανακάλυψη της γνώσης (knowledge discovery from data- KDD), ορίζεται η πληροφορία - γνώση - που μπορούμε να εξάγουμε από ένα σύνολο δεδομένων η οποία έχει αξία για εμάς καθώς είτε μπορούμε να τη χρησιμοποιήσουμε για μελλοντικές αποφάσεις είτε για την εξήγηση κάποιων φαινομένων. Η εξόρυξη γνώσης είναι ένα πεδίο που συγκεντρώνει εξαιρετικά μεγάλο ενδιαφέρον από ερευνητές, αλλά και από επιχειρήσεις παγκοσμίως, καθώς η δυνατότητα εξαγωγής έγκυρων συμπερασμάτων από τα δεδομένα δυσκολεύει καθώς τα δεδομένα αυξάνονται σε αριθμό ή διαφέρουν σε μορφή. Όπως αναφέρεται χαρακτηριστικά στον βιβλίο " Data mining: concepts and techniques" [24], χαρακτηρίζει την εξόρυξη γνώσης ως μία «φυσική» εξέλιξη της πληροφορικής ύστερα από την αλματώδη ανάπτυξη που γνώρισε στον τομέα της αποθήκευσης και επεξεργασίας των δεδομένων και κατ' επέκταση γίνεται η επισήμανση της ανάγκης αποδοτικών αναλυτικών τεχνικών από τεράστιους όγκους δεδομένων. Επιπλέον μπορούμε να δούμε την εξόρυξη γνώσης ως μία διαδικασία κατά την οποία τα δεδομένα υπόκεινται σε μία σειρά από ενέργειες προκειμένου να οδηγηθούμε στο επιθυμητό αποτέλεσμα που είναι αποκόμιση της γνώσης που κρύβουν τα δεδομένα.

Έχοντας εκμεταλλευτεί πλήρως εργαλεία της σύγχρονης επιστήμης της πληροφορικής όπως εργαλεία ETL, web scrapping, API's κλπ. τα δεδομένα που είναι διαθέσιμα δεν έχουν την ίδια μορφή και περιέχουν πληροφορία που δεν την χρειαζόμαστε σύμφωνα με τους στόχους της επιθυμητής ανάλυσης. Συνεπώς θα πρέπει τα δεδομένα να υποβληθούν σε μία προ-επεξεργασία προκειμένου να γίνουν αξιοποιήσιμα από τα σύγχρονα εργαλεία. Όπως φαίνονται και σχηματικά την εικόνα 5 το πρώτο στάδιο είναι ο καθαρισμός των δεδομένων με την παράληψη αυτών που προσθέτουν θόρυβο ή χαρακτηρίζονται ως ασυνεπή δεδομένα. Σε συνέχεια του καθαρισμού των δεδομένων ακολουθεί ο συνδυασμός των πολλαπλών πηγών των δεδομένων. Έχοντας κρατήσει όλα τα δυνατά δεδομένα που είναι διαθέσιμα, θα πρέπει να γίνει επιλογή αυτών των χαρακτηριστικών που είναι «χρήσιμα» και θα μας οδηγήσουν αποδοτικότερα προς τον στόχο της ανάλυσης.



Εικόνα 5 Σύνολο Διαδικασιών που Συνθέτουν τη Εξόρυξη Γνώσης Σύμφωνα με [24]

Το στάδιο της προ-επεξεργασίας τερματίζεται με τον μετασχηματισμό όλων των δεδομένων σε μορφές που βολεύουν την ανάλυσή τους σύμφωνα με τα διαθέσιμα εργαλεία. Για παράδειγμα αν το πρόγραμμα με το οποίο θα πραγματοποιηθεί η ανάλυση δέχεται σαν είσοδο comma separated value (.CSV) αρχεία θα πρέπει να γίνουν οι κατάλληλες ενέργειες έτσι ώστε να εκπληρωθεί η απαίτηση αυτή. Με το τέλος των προηγούμενων βημάτων σειρά έχει η εξόρυξη δεδομένων, όπου σε αυτό το στάδιο εντοπίζονται πρότυπα που συνθέτουν τα δεδομένα χρησιμοποιώντας μία οι περισσότερες τεχνικές αναγνώρισης προτύπων που έχουν αναπτυχθεί διαχρονικά με την αξιοποίηση κυρίως των επιστημών της στατιστικής και της πληροφορικής. Τέλος σειρά έχουν η αξιολόγηση των προτύπων που βρέθηκαν και η οπτικοποίηση των αποτελεσμάτων για τον διαμοιρασμό της πληροφορίας στα αρμόδια άτομα, ή αρχές, που στο τέλος θα κληθούν να αποφασίσουν με βάσει τα δεδομένα.

Η εξόρυξη της γνώσης παραμένει το πιο χρήσιμο βήμα όλης της προαναφερόμενης διαδικασίας καθώς είναι ο πυρήνας όλων των βημάτων, καθώς όλη η προ-επεξεργασία έχει σκοπό την σωστή αξιοποίηση των δεδομένων έτσι ώστε να είναι ορατά τα πρότυπα των χαρακτηριστικών και να βγουν ορθά αποτελέσματα κατά την αξιολόγηση. Η εξόρυξη χρησιμοποιείται κυρίως για την εύρεση γνώσης σε σχέση με τους 6 παρακάτω στόχους [25]:

- Ομαδοποίηση και τμηματοποίηση (clustering) των δεδομένων σε μικρότερα σύνολα σύμφωνα με κάποιο μέτρο ομοιότητας.
- Ταξινόμηση (classification) των δεδομένων ύστερα από σχετική εξέταση των δεδομένων κατηγοριοποιώντας τα αντικείμενα σε μία προκαθορισμένη κατηγορία. Η ταξινόμηση των δεδομένων σε αυτές τις προκαθορισμένες κατηγορίες είναι ιδιαιτέρως δημοφιλή τεχνική στο σημερινό επιχειρηματικό κόσμο καθώς χρησιμοποιούνται κατά κόρον για την εύρεση των πελατών που δαπανούν περισσότερα για παράδειγμα, ή ακόμη για τον εντοπισμό μίας κακόβουλης συναλλαγής.
- Η εκτίμηση (estimation) κάποιων μεγεθών που δεν βρίσκουν αναγκαία απάντηση με ένα απλό ναι ή όχι, αντί αυτού χρησιμοποιείται κάποια κλίμακα (scoring), με σκοπό την πιο έγκυρη κατηγοριοποίηση των δεδομένων. Ένα πολύ σύνηθες παράδειγμα είναι το credit scoring με το οποίο μία τράπεζα μπορεί να προβλέψει την αποπληρωμή ενός δανείου από ένα πιθανό πελάτη της.
- Πρόβλεψη (prediction) συμπεριφοράς αντικειμένων βάσει ιστορικών δεδομένων.
- Ομαδοποίηση «συγγενικών» χαρακτηριστικών η οποία με την σειρά τους συνθέτουν και «συγγενικά» αντικείμενα.
- Τέλος η περιγραφή των ευρημάτων με σκοπό την εξήγηση αυτών.

Οι προαναφερόμενοι στόχοι επιτυγχάνονται με μία σειρά από αναλύσεις μερικές από τις οποίες είναι ανάλυση συσχετίσεων (association analysis), ανάλυση παλινδρόμησης (regression analysis), ανάλυση ακραίων τιμών (outlier analysis) η οποία είναι εξαιρετικά διαδεδομένη για την εύρεση ανωμαλιών σε συναλλαγές η οποία είναι γνωστή και ως ανίχνευση απάτης με ανάλυση συστάδων (fraud detection cluster analysis) κ.α.

Στην συνέχεια της ενότητας αυτής, πραγματοποιείται μία λεπτομερής περιγραφή των κύριων τεχνικών και αλγορίθμων που χρησιμοποιούνται σήμερα με σκοπό την εξόρυξη της γνώσης από μεγάλα δεδομένα καθώς και τις ιδιότητες των χαρακτηριστικών ανά αλγόριθμο, δίνοντας έμφαση σε αυτές που χρησιμοποιούνται στο πειραματικό στάδιο αυτής της διπλωματικής εργασίας.

3.1 Τεχνικές Προ-επεξεργασίας Δεδομένων

Όπως προαναφέρθηκε στην αρχή του κεφαλαίου αυτού, η προ επεξεργασία των δεδομένων είναι μία διαδικασία απαραίτητη καθώς τα πραγματικά δεδομένα συνοδεύονται από μία πληθώρα προβλημάτων. Σε αυτή την ενότητα γίνεται αναφορά στις πιο σύνηθες περιπτώσεις όπου το «πείραγμα» των δεδομένων είναι απαραίτητο, καθώς επίσης και τους πιθανούς τρόπους αντιμετώπισης για την κάθε περίπτωση.

Με μία πιο προσεκτική ματιά στα δεδομένα ένας αναλυτής θα πρέπει να είναι σε θέση να αναγνωρίζει προβλήματα όπως οι διπλοεγγραφές, τιμές που παραβιάζουν λογικούς κανόνες, χρήση συντομογραφιών και φυσικά το πιο διαδεδομένο πρόβλημα είναι αυτό των χαμένων τιμών (missing values). Χαμένες τιμές προκύπτουν παραδείγματος χάριν όταν δημιουργείται η καρτέλα ενός καταναλωτή και δεν συμπληρώνεται πλήρως ή μπορεί ακόμη κάποιος χρήστης του συστήματος να θεωρήσει ότι κάποια πληροφορία δεν είναι σημαντική και να την διαγράψει. Επιπλέον σημαντικό πρόβλημα που εντοπίζεται στα δεδομένα είναι ο θόρυβος. Σαν θόρυβος χαρακτηρίζονται οι λανθασμένες τιμές που υπάρχουν στα δεδομένα, οι ακραίες τιμές (outliers) καθώς δεν προσφέρουν κάποια χρήσιμη πληροφορία στην ανάλυση επειδή περιγράφουν σπάνιες, μεμονωμένες περιπτώσεις που δεν μας ενδιαφέρουν. Ο χειρισμός των outliers πρέπει να γίνεται με προσοχή καθώς ο συνυπολογισμός τους στην ανάλυση μπορεί να οδηγήσει τους αλγορίθμους στην εξαγωγή μη έγκυρων συμπερασμάτων. Έγινε λόγος για προσεκτικό χειρισμό των outliers και όχι ο αποκλεισμός τους, καθώς σε συγκεκριμένες περιπτώσεις είναι απαραίτητη η παρατήρησή τους όπως στη ανίχνευση απάτης. Λόγω του ότι οι απατηλές δοσοληψίες μέσα σε μία ημέρα είναι πολύ λίγες σε σχέση με τις κανονικές μπορούν να θεωρηθούν ως ακραίες (outliers). Έτσι λοιπόν με την παρατήρηση αυτών των ακραίων τιμών (outliers) μπορούν τα σύγχρονα συστήματα να εντοπίζουν τις απάτες σε πολύ σύντομο χρονικό διάστημα.

Η διαδικασία αυτή της προπεξεργασίας των δεδομένων χαρακτηρίζεται ως καθαρισμός των δεδομένων που είναι απαραίτητος πριν την ανάλυσή τους καθώς μπορεί να επηρεάσει αρνητικά τα αποτελέσματα των αλγορίθμων. Εκτός της περίπτωσης των ελλιπών δεδομένων σημαντική είναι και η περίπτωση κατά την οποία ένας αλγόριθμος συμπεριφέρεται καλύτερα παραδείγματος χάρει με τιμές που κυμαίνονται από το 0 έως το 1, όπου η κανονικοποίηση των αριθμητικών τιμών των μεταβλητών είναι απαραίτητη. Επιπρόσθετα της κανονικοποίησης άλλοι αλγόριθμοι δέχονται σαν είσοδο ονομαστικές τιμές και αν οι μεταβλητές είναι αριθμητικές θα πρέπει να γίνει η διακριτοποίησή τους προκειμένου να αξιοποιηθούν αποδοτικά οι συγκεκριμένοι αλγόριθμοι. Όπως για παράδειγμα η περίπτωση όπου μέθοδοι που διαχειρίζονται συνεχείς τιμές, όπως αυτή των κ-κοντινότερων γειτόνων, επηρεάζονται πολύ από τις υψηλές τιμές μεταξύ των μεταβλητών με αποτέλεσμα τα συμπεράσματα που αποδίδονται να μην είναι αξιόπιστα καθώς σχηματίζονται από τις τιμές μίας ή το πολύ δύο μεταβλητών με τις υψηλότερες τιμές. Τέλος σαν αποτέλεσμα της ακατάσχετης παραγωγής δεδομένων από το κάθε ένα μας στην εποχή αυτή καθίσταται πολλές φορές αδύνατη η ανάλυση αυτών των μεγάλων όγκων δεδομένων, καθώς ο χρόνος ή το κόστος επεξεργασίας τους είναι ασύμφορο. Με γνώμονα λοιπόν την ανάγκη της ανάλυσης και εξαγωγής συμπερασμάτων, ανεξαρτήτως του όγκου των δεδομένων, χρησιμοποιούνται ευρέως τεχνικές μείωσης δεδομένων όπου επιλέγονται τα πιο σημαντικά χαρακτηριστικά (feature selection) μεταξύ των μεταβλητών που συνθέτουν μία εγγραφή. Στη συνέχεια της ενότητας αυτής παρουσιάζονται οι τεχνικές με τις οποίες επιτυγχάνονται οι λύσεις των προβληματικών, ή καλύτερα, των θορυβώδη δεδομένων κατά το στάδιο της προπεξεργασίας.

3.1.1 Ελλιπείς Τιμές (Missing Values)

Λόγω των πιθανών καταστάσεων που αναφέρθηκαν νωρίτερα είτε λόγω διαγραφής, είτε μη πλήρους καταχώρησης στοιχείων είτε πρόσθεσης πεδίου που δεν έχουν παλαιότερα δεδομένα είναι πολύ συχνό

φαινόμενο να υπάρχει σημαντικός αριθμός ελλιπών στοιχείων. Πολλοί αλγόριθμοι εξόρυξης γνώσης όπως τα δένδρα αποφάσεων τύπου C4.5 αντιμετωπίζουν το πρόβλημα των χαμένων τιμών του συνόλου των δεδομένων. Όμως σε μία πειραματική ανάλυση για το ποιος αλγόριθμος είναι πιο αποδοτικός θα πρέπει, τουλάχιστον στο στάδιο της προεπεξεργασίας, οι χαμένες τιμές να αντιμετωπίζονται με τον ίδιο τρόπο για την έγκυρη εξαγωγή των αποτελεσμάτων. Αυτό συμβαίνει καθώς, η κάθε μέθοδος εξόρυξης γνώσης, που αντιμετωπίζει «εσωτερικά» το πρόβλημα των χαμένων τιμών, παράγει διαφορετικά σύνολα δεδομένων, με αποτέλεσμα τη αδύνατη εξαγωγή αντικειμενικών συγκριτικών συμπερασμάτων μεταξύ των υπό διερεύνηση αλγορίθμων. Έτσι λοιπόν προτιμάται η επίλυση του προβλήματος των χαμένων τιμών πριν την εφαρμογή των αλγορίθμων εξόρυξης γνώσης. Παρακάτω παρουσιάζονται μερικές από τις πιο διαδεδομένες τεχνικές επίλυσης του προβλήματος αυτής της ενότητας:

- Διαγραφή ολόκληρης της γραμμής. Σύμφωνα με το [26] δεν προτείνεται καθώς αναλόγως το πλήθος των χαμένων τιμών μπορεί να διαγραφεί σημαντικό πλήθος δεδομένων με τον κίνδυνο να βγουν λανθασμένα συμπεράσματα.
- Η πιο σωστή προσέγγιση αλλά ταυτόχρονα και η πιο δύσκολη ή και ασύμφορη (λόγω χρόνου) είναι η συμπλήρωση των κενών τιμών με τις πραγματικές. Γίνεται αντιληπτό ότι αυτή η προσέγγιση για την συμπλήρωση, παραδείγματος χάρει όλων των χαμένων ηλικιών των πελατών μιας τράπεζας, θα πρέπει η τράπεζα να δαπανήσει τεράστιο αριθμό εργατοωρών.
- Αντικατάσταση της χαμένης τιμής με το μέσο όρο, αν είναι αριθμητική η μεταβλητή, ή με τη πιο σύνηθες τιμή, αν είναι ονομαστική. Η αντικατάσταση των χαμένων τιμών μπορεί να γίνει λίγο πιο στοχευμένα αξιοποιώντας πιθανές κλάσεις από άλλες μεταβλητές. Με αυτό το τρόπο γίνεται πιο ορθολογική η αντικατάσταση των χαμένων τιμών καθώς ο μέσος όρος, παραδείγματος χάρει, του ύψους του εισοδήματος σε τραπεζικά δεδομένα δεν υπολογίζεται από όλους τους πελάτες αλλά από αυτούς που προσεγγίζουν άλλα χαρακτηριστικά όπως η ηλικία, το φύλλο, η βαθμίδα εκπαίδευσης κλπ.
- Μία άξια αναφοράς προσέγγιση, που μπορεί να χρησιμοποιηθεί κυρίως σε ονομαστικές μεταβλητές, είναι η χρησιμοποίηση όλων των πιθανών ονομαστικών τιμών που εντοπίζονται στη στήλη που εντοπίζεται η χαμένη τιμή. Όπως και προηγουμένως μπορεί να γίνει πιο ορθολογική αντικατάσταση της χαμένης τιμής χρησιμοποιώντας μόνο εκείνες τις ονομαστικές τιμές που εντοπίζονται στη κλάση.
- Τέλος ως πιο αποδοτική τακτική που προτείνεται στο [26] είναι η πρόβλεψη της τιμής της μεταβλητής. Με αυτό το τρόπο ανεξαρτήτως είδους μεταβλητής, ονομαστική ή αριθμητική, μπορεί να γίνει πρόβλεψη βάσει των υπολοίπων χαρακτηριστικών.

3.1.2 Θορυβώδη Δεδομένα

Τα θορυβώδη δεδομένα όπως προαναφέρθηκε, αναφέρονται σε δεδομένα που οι μεταβλητές τους περιέχουν λανθασμένες τιμές και τιμές οι οποίες είναι σημαντικά διαφορετικές από το μέσο όρο, γνωστές και ως τιμές εξαιρέσεις (outliers). Τα δεδομένα με αυτά τα χαρακτηριστικά δημιουργούν σημαντικά προβλήματα στους αλγορίθμους εξόρυξης γνώσης καθώς, επηρεάζουν σε μεγάλο βαθμό το συμπεράσματα. Θα πρέπει να αναφερθεί ότι η αντιμετώπιση του προβλήματος του θορύβου δεν είναι πανάκεια και διαφέρει αναλόγως με το τελικό στόχο. Για παράδειγμα στη δεύτερη περίπτωση αφού γίνει ο εντοπισμός των ακραίων τιμών είτε διαγράφονται από το σύνολο δεδομένων είτε τροποποιούνται οι ακραίες τιμές κατάλληλα. Βέβαια όλη η προσέγγισή μας αλλάζει αν η πρόβλεψη των ακραίων τιμών είναι αυτή που μας ενδιαφέρει, όπως παραδείγματος χάρει στην πρόβλεψη απατηλών χρηματικών κινήσεων μέσω πιστωτικών καρτών. Βέβαια αυτή η περίπτωση θα διερευνηθεί εκτενέστερα αργότερα καθώς είναι ένα από τα κύρια ζητούμενα της παρούσας εργασίας.

Τα θορυβώδη δεδομένα μπορούν να ομαλοποιηθούν αντικαθιστώντας τις τιμές των ταξινομημένων κατά αύξουσα σειρά μεταβλητών και χωρίζοντάς τα σε ίσα διαστήματα πλάτους ή συχνότητας. Αυτή η τεχνική ονομάζεται κατακερματισμός σε διαστήματα και αντικατάσταση τιμών. Τα διαστήματα συχνότητας είναι ίσου πλήθους εγγραφών και τα διαστήματα πλάτους έχουν το ίδιο εύρος τιμών και αφού υπολογιστούν οι νέες τιμές αντικαθιστούν όλες τις τιμές του εκάστοτε διαστήματος. Οι νέες τιμές είναι είτε ο μέσος όρος του κάθε διαστήματος, είτε γίνεται αντικατάσταση των οριακών τιμών με τη μικρότερη τιμή ή τη μεγαλύτερη τιμή του διαστήματος που εντοπίζονται.

Περιληπτικά υπάρχει η ο στατιστικός εντοπισμός των εξαιρέσεων όπου υπολογίζεται η μέση τιμή και η τυπική απόκλιση των εγγραφών μιας μεταβλητής. Στη συνέχεια αν η τιμή μιας εγγραφής είναι μεγαλύτερη από το άθροισμα της μέσης τιμής με το γινόμενο της τυπικής απόκλισης και μιας σταθεράς k, που ορίζει ο χρήστης τότε, η τιμή αυτή χαρακτηρίζεται ως εξαίρεση. Ως εξαίρεση χαρακτηρίζεται επίσης η τιμή αν είναι μικρότερη από τη διαφορά της μέσης τιμής και του γινομένου της τυπικής απόκλισης και της k. Ο υπολογισμός της τιμής του k, που επί του πρακτέου είναι ο πολλαπλασιαστής που ορίζει ο χρήστης, για το τι θεωρείται ως ακραία τιμή, είναι μία εμπειρική διαδικασία που προϋποθέτει πολύ καλή γνώση των δεδομένων και του στόχου της ανάλυσης.

Εντοπισμός ακραίων τιμών γίνεται επίσης μέσω της ανάλυσης συστάδων (cluster analysis) όπου τα δεδομένα ομαδοποιούνται σύμφωνα με τα επικρατέστερα χαρακτηριστικά. Κάποιες εγγραφές κατά τη διαδικασία της συσταδοποίησης μπορεί να μην μπορούν να ομαδοποιηθούν σε κάποια από τις ομάδες που συνέταξαν τα υπόλοιπα δεδομένα. Οι εγγραφές που δεν ομαδοποιήθηκαν χαρακτηρίζονται ως εξαιρέσεις.

3.1.3 Μετασχηματισμός Δεδομένων (Κανονικοποίηση)

Η διαδικασία μετασχηματισμού των δεδομένων σε τιμές που βολεύουν την εκάστοτε ανάλυση ονομάζεται κανονικοποίηση και αναλόγως τη μέθοδο που επιλέγουμε κατά την ανάλυση των δεδομένων, χρησιμοποιούμε τον αντίστοιχο κατάλληλο μετασχηματισμό. Όπως αναφέρεται χαρακτηριστικά στο [26] ορισμένες τεχνικές εξόρυξης γνώσης όπως τα νευρωνικά δίκτυα και η ανάλυση συστάδων απαιτούν μία προ επεξεργασία και μετασχηματισμό των δεδομένων πριν την εφαρμογή τους. Πιο συγκεκριμένα τα νευρωσικά δίκτυα συμπεριφέρονται καλύτερα όταν οι τιμές των χαρακτηριστικών που μελετώνται είναι μεταξύ των τιμών μηδέν και ένα. Συνεπώς θα πρέπει να γίνει η κανονικοποίηση των τιμών των μεταβλητών στο εύρος (0,1). Τέλος κατά την ανάλυση συστάδων οι τιμές των διαφορετικών χαρακτηριστικών δεν θα πρέπει να αποκλίνουν σημαντικά καθώς δίνεται μεγαλύτερη βαρύτητα στις μεγάλες τιμές με αποτέλεσμα να την εσφαλμένη ανάλυση και λανθασμένα αποτελέσματα. Σε αυτή τη περίπτωση θα πρέπει να υλοποιηθεί ένας μετασχηματισμός με τον οποίο θα διατηρούνται οι διαφορές εντός τις ίδιας μεταβλητής αλλά οι απόλυτες τιμές μεταξύ των χαρακτηριστικών θα πρέπει να έχουν συγκρίσιμη διαφορά. Η κανονικοποίηση των δεδομένων μπορεί να επιτευχθεί με πολλαπλούς τρόπους και στη συνέχεια παρουσιάζονται οι πιο διαδεδομένοι.

Πρώτη σε σειρά είναι η κανονικοποίηση ελαχίστου-μεγίστου. Με τη συγκεκριμένη τεχνική μπορεί να γίνει αντικατάσταση των τιμών ενός συνόλου δεδομένων σύμφωνα με ένα νέο εύρος που ορίζουμε εμείς. Η αντιστοίχιση των τιμών γίνεται με γραμμικό μετασχηματισμό όπως φαίνεται και στον τύπο παρακάτω:

$$x' = \frac{x - min_A}{max_A - min_A}(new\_max_A - new\_min_A) + new\_min_A$$

Έτσι λοιπόν προσαρμόζεται το σύνολο τιμών σε ένα νέο εύρος με την μέθοδο κανονικοποίησης ελαχίστου-μεγίστου, εντός ενός επιθυμητού εύρους τιμών αλλά επίσης με αυτό το τρόπο διατηρείται η αναλογία των τιμών σε σχέση με το αρχικό σύνολο δεδομένων.

Η δεύτερη τεχνική αξιοποιεί περιγραφικά χαρακτηριστικά, πιο συγκεκριμένα τη μέση τιμή και στη τυπική απόκλιση, του συνόλου δεδομένων και βάσει αυτών προσαρμόζει τις τιμές των χαρακτηριστικών. Μπορεί να θεωρηθεί πιο αποδοτική από την κανονικοποίηση ελαχίστου-μεγίστου, καθώς με την ύπαρξη ακραίων τιμών, ο μετασχηματισμός θα τοποθετούσε σε ένα μικρό τμήμα του δηλωθέν εύρους τις τιμές των «κανονικών» παρατηρήσεων και το υπόλοιπο εύρος θα χρησιμοποιούνταν για τις ακραίες τιμές.

$$x' = \frac{x - M_A}{\sigma_A}$$

Τελευταία μέθοδος κανονικοποίησης είναι ο υποδεκαπλασιασμός των τιμών του συνόλου δεδομένων. Σκοπός της δεκαδικής κλιμάκωσης είναι ο σχηματισμός μιας δύναμης του 10 που να καθιστά το μέγιστο να είναι μικρότερο του 1.

$$x' = \frac{x}{10^{\kappa}}$$

3.1.4 Μείωση Διαστάσεων – Επιλογή Χαρακτηριστικών (Feature Selection)

Ίσως από τις πιο σημαντικές διαδικασίες στο πλαίσιο των διεργασιών της προεπεξεργασίας των δεδομένων είναι η μείωση των διαστάσεων έχοντας πάντα σαν αρχή, την διατήρηση της πληροφορίας. Η ανάγκη για την μείωση των διαστάσεων κατά τη επεξεργασία των δεδομένων είναι συνεχώς αυξανόμενη στη εποχή των μεγάλων δεδομένων. Το πλήθος των διαθέσιμων δεδομένων είναι τεράστιο, έχοντας σαν αποτέλεσμα την ακριβή σε υπολογιστικούς πόρους αλλά επίσης και τη κοστοβόρα από πλευράς χρόνου επεξεργασίας τους. Για αυτό το λόγο έχουν προταθεί τρόποι μείωσης διαστάσεων ενός συνόλου δεδομένου που θα αναλυθούν μεταγενέστερα σε αυτή την ενότητα. Προβλήματα όπως στο ότι σε ένα σύνολο δεδομένων μπορεί να υπάρχει επικαλυπτόμενη πληροφορία, δηλαδή η συμπεριφορά μιας στήλης να είναι ίδια ή ανάλογη με μία άλλη, μπορεί να επιλυθούν μέσω των τεχνικών μείωσης διαστάσεων. Σε αυτή τη περίπτωση αναλόγως και την επιθυμητή ανάλυση συνήθως παραλείπεται μία από τις στήλες αυτές. Μία άλλη πολύ συνηθισμένη κατάσταση είναι όταν ένα χαρακτηριστικό είναι «άσχετο» σύμφωνα με την ανάλυση πράγμα που μας οδηγεί στην παράληψή του έτσι ώστε να μειώσουμε την πολυπλοκότητα. Η μείωση της πολυπλοκότητας της επεξεργασίας, εκπαίδευσης αλλά και δοκιμής των μοντέλων θα πρέπει να είναι από τα πρώτα προβλήματα προς επίλυση καθώς αλγόριθμοι όπως οι «Κ Κοντινότεροι Γείτονες» είναι δύστροποι όταν εφαρμόζονται σε δεδομένα με πολλές διαστάσεις. Τέλος οι στατιστικές μέθοδοι ανάλυσης κυρίως υποθέτουν ότι χρησιμοποιούνται μόνο σημαντικές μεταβλητές και ότι είναι ασυσχέτιστες μεταξύ τους.

Η επιλογή των χαρακτηριστικών μπορεί να γίνει «χειροκίνητα» από κάποιο «ειδικό» που γνωρίζει πολύ καλά τα δεδομένα και το στόχο της ανάλυσης και μπορεί να διακρίνει ποιες μεταβλητές είναι οι πιο σημαντικές. Το κύριο αρνητικό χαρακτηριστικό αυτής της προσέγγισης είναι ότι δεν υπάρχει πάντα αυτός ο «ειδικός» πράγμα που θα καθιστούσε αδύνατη τη επιλογή των σημαντικών χαρακτηριστικών. Επιπλέον ανεξαρτήτως της γνώσης του «ειδικού» στο πεδίο και τα δεδομένα η επιλογή των χαρακτηριστικών εμπεριέχει μία υποκειμενική προσέγγιση για το πιο είναι πιο σημαντικό σε σχέση με κάποιο άλλο χαρακτηριστικό. Είναι εμφανές λοιπόν η ανάγκη για αυτοματοποιημένους αντικειμενικούς τρόπους εύρεσης των σημαντικών χαρακτηριστικών και αναλόγως κάποιων τιμών κατωφλίων, για το περιθώριο

λάθους παραδείγματος χάρει, να γίνεται η επιλογή του καταλληλότερου υποσυνόλου χαρακτηριστικών του αρχικού συνόλου δεδομένων.

Υπάρχουν πολλοί τρόποι μείωσης διαστάσεων και θα αναλυθούν σε αυτή την ενότητα οι πιο γνωστοί. Στο [26] γίνεται μία κατηγοριοποίηση των τεχνικών αυτών σε σχέση με το αν εφαρμόζονται από τον αλγόριθμο ανάλυσης ή αν γίνεται η μείωση πριν την εισαγωγή των δεδομένων στον αλγόριθμο. Πιο συγκεκριμένα όταν η μείωση των διαστάσεων πραγματοποιείται πριν την εφαρμογή του εκάστοτε αλγορίθμου είναι μέθοδοι τύπου φίλτρου (filter). Ενώ όταν ο αλγόριθμος της ανάλυσης κάνει επιλογή των χαρακτηριστικών είναι τύπου περιτυλίγματος (wrapper). Κάθε ένας από τους τύπους αυτούς έχει τα αρνητικά και τα θετικά σημεία τους, όπως για παράδειγμα οι τύπου φίλτρου (filter) μέθοδοι είναι σημαντικά πιο γρήγοροι στο χρόνο και μπορούν να συνδυαστούν με πολλαπλούς αλγορίθμους. Αντιθέτως οι μέθοδοι τύπου περιτυλίγματος (wrapper) είναι πιο «αργοί» αλλά είναι πιο εύστοχοι ως προς την επιλογή των χαρακτηριστικών καθώς ο ίδιος αλγόριθμος της ανάλυσης επιλέγει ποια χαρακτηριστικά τον «βολεύουν» καλύτερα.

3.1.4.1 Πρόσθια Επιλογή και Οπίσθια Εξάλειψη

Η πρόσθια επιλογή χαρακτηριστικών όπως και η οπίσθια εξάλειψη χαρακτηριστικών είναι ευρετικές μέθοδοι που έχουν σαν σκοπό την επιλογή των πιο σημαντικών χαρακτηριστικών αποφεύγοντας με αυτό το τρόπο, την εξαντλητική δοκιμή όλων των πιθανών συνδυασμών του συνόλου δεδομένων. Έτσι λοιπόν στην πρώτη περίπτωση αφού αναγνωριστεί η πιο σημαντική μεταβλητή αφαιρείται από το αρχικό σύνολο δεδομένων και τοποθετείται στο σύνολο δεδομένων με τις πιο σημαντικές μεταβλητές. Στη συνέχεια διενεργείται εκ νέου η ανάλυση για την εύρεση των πιο σημαντικών μεταβλητών στο αρχικό σύνολο που είναι κατά μία διάσταση μικρότερο, και η επικρατούσα μεταβλητή αφαιρείται από το αρχικό και το τοποθετείται με τη σειρά της στο νέο σύνολο. Η επαναληπτική αυτή διαδικασία πραγματοποιείται μέχρις ότου να ικανοποιηθεί μια συνθήκη εξόδου. Αντίστοιχα στην οπίσθια εξάλειψη αφαιρείται ανά επανάληψη η πιο «ασήμαντη» διάσταση του συνόλου δεδομένων και ομοίως, όπως στην πρώτη περίπτωση, η διαδικασία τερματίζεται με την ικανοποίηση της συνθήκης εξόδου.

3.1.4.2 Επιλογή Χαρακτηριστικών Βάσει Συσχέτισης (Correlation Based Feature Extraction)

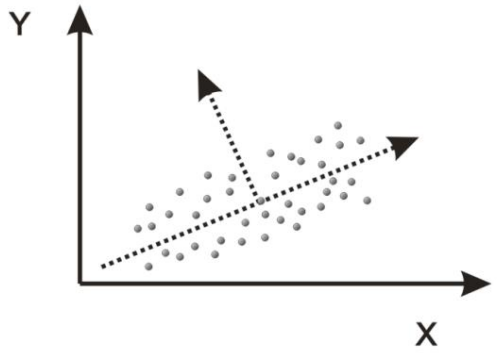
Στην συνέχεια έχουμε την επιλογή χαρακτηριστικών σύμφωνα με τη συσχέτιση (correlation based feature selection - CFS) που έχουν με τη μεταβλητή στόχου. Όπως αναφέρεται χαρακτηριστικά στα [26, 27] είναι μία μέθοδος τύπου filter δηλαδή εφαρμόζεται πριν τον αλγόριθμο ανάλυσης, και κυρίως συναντάτε σε προβλήματα κατηγοριοποίησης (classification). Σκοπός είναι η εύρεση των χαρακτηριστικών εκείνων που έχουν ισχυρή συσχέτιση με τη μεταβλητή στόχου αλλά χαμηλή συσχέτιση μεταξύ τους αποδίδοντας έτσι ένα σύνολο ανεξάρτητων μεταβλητών. Στο [27] αναφέρεται ότι κατά το πειραματικό στάδιο χρησιμοποιήθηκαν φυσικά αλλά και τεχνητά δεδομένα σε τρεις αλγορίθμους μηχανικής μάθησης. Πιο συγκεκριμένα στα τεχνητά δεδομένα με μεγάλη ταχύτητα αναγνώριζε τα σημαντικά χαρακτηριστικά και στα φυσικά δεδομένα παρατηρήθηκε ότι περιοριζόταν στα μισά και πλέον χαρακτηριστικά. Στις περισσότερες περιπτώσεις τα αποτελέσματα της ακρίβειας της κατηγοριοποίησης του μειωμένου συνόλου δεδομένων ήταν καλύτερα από αυτά που χρησιμοποιούνταν το πλήρες σύνολο δεδομένων. Στη χειρότερη, δε περίπτωση, τα αποτελέσματα της ακρίβειας ήταν ίσα μειώνοντας πάντα το χρόνο επεξεργασίας του εκάστοτε αλγορίθμου μηχανικής μάθησης. Τέλος συγκρίνοντας τις μεθόδους επιλογής χαρακτηριστικών σύμφωνα με τη συσχέτιση με τεχνικές μείωσης διαστάσεων περιτυλίγματος, σε πολλές περιπτώσεις τα αποτελέσματα ήταν συγκρίσιμα με τη μόνη διαφορά ότι γενικά ήταν σημαντικά πιο γρήγορη η πρώτη, πράγμα που είναι επιθυμητό ειδικά όταν εξετάζουμε μεγάλα σύνολα δεδομένων.

3.1.4.3 Ανάλυση Κύριων Συνιστωσών – ΑΚΣ (Principal Components Analysis – PCA)

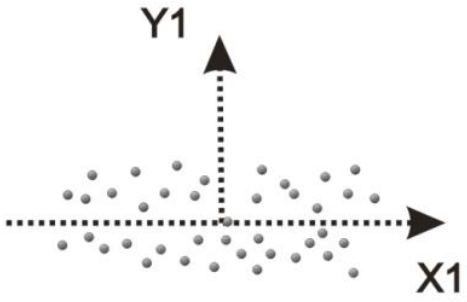
Η ανάλυση κύριων συνιστωσών αποτελεί μία γραμμική μέθοδο συμπίεσης δεδομένων η οποία προκύπτει από τον επαναπροσδιορισμό των συντεταγμένων ενός συνόλου δεδομένων σε ένα νέο σύστημα συντεταγμένων. Οι νέες αυτές συντεταγμένες είναι το αποτέλεσμα ενός γραμμικού συνδυασμού των αρχικών μεταβλητών. Πιο συγκεκριμένα αν τα δεδομένα έχουν N μεταβλητές με τιμές $(x_1 \dots x_n)$ τότε η κύρια συνιστώσα *i* έχει τη μορφή:

$$PC_i = a_{i1}x_1 + a_{i2}x_2 + \cdots + a_{in}x_n$$

Οι συντελεστές a_{ik} καθορίζουν το βαθμό με στον οποίο κάθε μεταβλητή επηρεάζει τη βασική συνιστώσα. Η ΑΚΣ αποτελεί μέθοδο μείωσης των διαστάσεων των δεδομένων, με προβολή τους σε ένα χώρο λιγότερων, αλλά διαφορετικών διαστάσεων, συνεπώς, δεν είναι μία μέθοδος επιλογής σημαντικών χαρακτηριστικών με την έννοια της επιλογής κάποιων μεταβλητών από το αρχικό σύνολο δεδομένων. Η εξάλειψη συνιστωσών που δεν συγκεντρώνουν σημαντική πληροφορία είναι έργο μεταγενέστερης επεξεργασίας και όχι της ανάλυσης των κυρίων συνιστωσών.



Εικόνα 6 Ανάλυση Κύριων Συνιστωσών 1η φάση
Πηγή: [26]



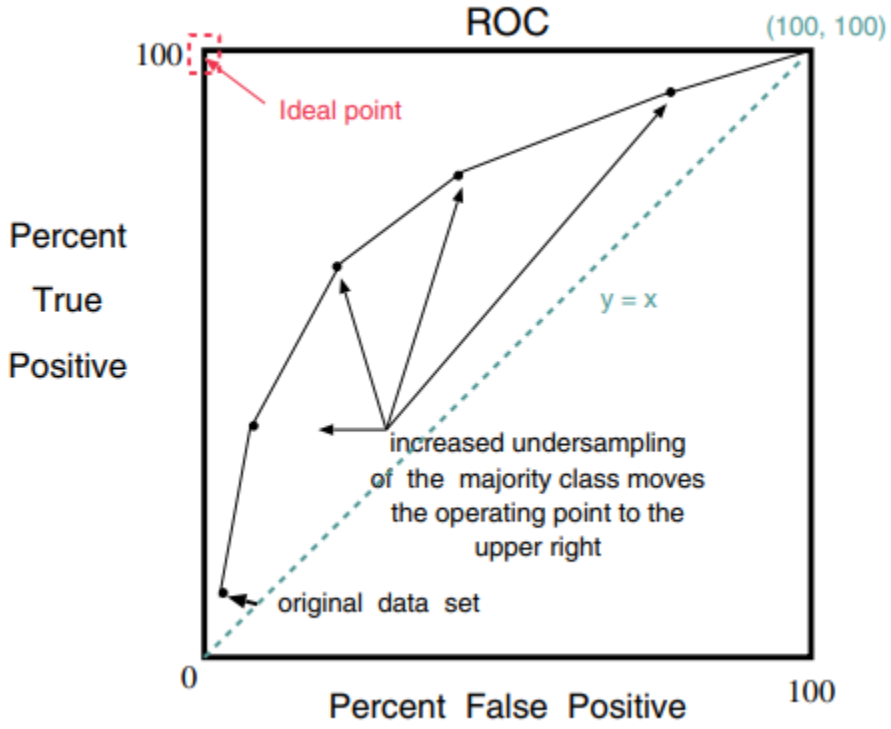
Εικόνα 7 Ανάλυση Κύριων Συνιστωσών 2η φάση
Πηγή: [26]

Στις εικόνες 6 και 7 γίνεται μία προσπάθεια απεικόνισης της ΑΚΣ. Στη πρώτη φάση στην εικόνα 6 διαγράφονται με διακεκομμένες κάθετες γραμμές οι 2 κύριες συνιστώσες των δεδομένων και στη δεύτερη φάση απεικονίζονται τα ίδια δεδομένα όχι πλέον στο σύστημα αξόνων X, Y αλλά στο σύστημα των 2 κύριων συνιστωσών που βρέθηκε κατά τη φάση 1. Παρατηρείται ότι οι προβολές των σημείων κατά μήκος του άξονα X1 περιέχουν σημαντικά περισσότερη πληροφορία σε σχέση με τις προβολές των ίδιων σημείων στον άξονα Y1. Έτσι λοιπόν αν εξαλειφθεί ο άξονας Y1 θα διατηρήσουμε μεγάλο ποσοστό της πληροφορίας σχετικά με τη διασπορά των δεδομένων.

3.1.5 Δειγματοληψία Ισορροπίας (Balance Sampling)

Όπως πολύ χαρακτηριστικά αναφέρεται στο βιβλίο [28] το πρόβλημα των ανομοιόμορφων κατανομών ως προς το πεδίο της μεταβλητής κατηγοριοποίησης έγινε αισθητό κυρίως σε εφαρμογές μηχανικής μάθησης σε προβλήματα του πραγματικού κόσμου. Μερικά σημαντικά πεδία όπου συναντάει κανείς το πρόβλημα των ανομοιόμορφων κατανομών σύμφωνα με το [29] είναι, η ανίχνευση απάτης και εισβολής, διαχείριση ρίσκου, κατηγοριοποίηση κειμένου, ιατρική διάγνωση και επίβλεψη καθώς επίσης και πολλοί άλλοι τομείς. Το πρόβλημα εντοπίζεται καθώς οι αλγόριθμοι μηχανικής μάθησης αδυνατούν να κατηγοριοποιήσουν σωστά μία παρατήρηση, που ανήκει σε μία κλάση μειοψηφίας, λόγω του πολύ μεγάλου αριθμού παρατηρήσεων μιας δεύτερης κλάσης. Πληθώρα εργασιών έχουν δημοσιευθεί με

σκοπό την εύρεση του καταλληλότερου χειρισμού των ανομοιόμορφων κατανομών και ποιος (χειρισμός) είναι πιο αποδοτικός ή «συνεργάζεται» καλύτερα με τους διαθέσιμους αλγόριθμους μηχανικής μάθησης και πάντα αναλόγως το πρόβλημα. Δεν άργησε να γίνει κατανοητό ότι η μετρική της ακρίβειας των εκτιμητών δεν ήταν επαρκείς και έτσι προτιμήθηκε ευρέως η χρήση των καμπυλών ROC. Η ανάλυση των καμπυλών ROC δίνει τη δυνατότητα ελέγχου των συνολικών αποτελεσμάτων των κατηγοριοποιητών παραθέτοντας της σωστά θετικές κατηγοριοποιήσεις έναντι των λανθασμένων θετικών.



Εικόνα 8 Κίνηση Καμπύλης ROC πηγή: [28]

Στο σχήμα φαίνεται ότι με τη μείωση των παρατηρήσεων της κλάσης πλειοψηφίας το ποσοστό των λανθασμένων θετικών κατηγοριοποιήσεων μειώνεται και με ταυτόχρονα αυξάνεται το ποσοστό των ορθά θετικών κατηγοριοποιήσεων. Με τη βοήθεια των καμπυλών ROC προσπαθούμε να προσεγγίσουμε το ιδανικό σημείο (ideal point) στην πάνω αριστερή γωνία του σχήματος. Γίνεται αντιληπτό λοιπόν, ότι με τη μείωση των παρατηρήσεων της πλειοψηφικής κλάσης επετεύχθη η δημιουργία ενός πιο «έξυπνου» μοντέλου ικανού κατηγοριοποίησης μιας παρατήρησης εξαίρεσης στην σωστή κλάση δίχως να επηρεάζεται σημαντικά από τον δυσανάλογο αριθμό των παρατηρήσεων δεύτερων ή τρίτων κλάσεων. Ο υπολογισμός των ποσοστών που συνθέτουν την καμπύλη ROC γίνεται με τη βοήθεια του παρακάτω πίνακα 1 ο οποίος ορίζει τι είναι το ορθά θετικό και λανθασμένη θετική κατηγοριοποίηση.

		Προβλέψεις	
		Αρνητική (Negative)	Θετική (Positive)
Πραγματικότητα	Αρνητική (Negative)	Ορθά Αρνητική Πρόβλεψη (True Negative: TN)	Λανθασμένα θετική Πρόβλεψη (False Positive: FP)
	Θετική (Positive)	Λανθασμένα Αρνητική Πρόβλεψη (False Negative: FN)	Ορθά θετική Πρόβλεψη (True Positive: TP)

Πίνακας 1 Confusion Matrix

Σαν στήλες έχουν ορισθεί οι προβλέψεις του κατηγοριοποιητή και σαν γραμμές οι πραγματικές τιμές των παρατηρήσεων. Συνεπώς για να πραγματοποιηθεί μία εκτίμηση του μοντέλου θα πρέπει να είναι γνωστές οι πραγματικές κατηγοριοποιήσεις των παρατηρήσεων που εξετάζονται έτσι ώστε να μπορεί να αξιολογηθεί βάσει του πίνακα 1. Αν λοιπόν πραγματοποιηθεί μία πρόβλεψη για μία παρατήρηση και κατηγοριοποιηθεί σαν Θετική αλλά στην πραγματικότητα είναι αρνητική τότε εντοπίζεται στο επάνω δεξιά κελί του πίνακα 1 και χαρακτηρίζεται ως Λανθασμένα θετική πρόβλεψη (False Positive: FP). Αν σε ένα δεύτερο παράδειγμα ο εκτιμητής τοποθετήσει μία παρατήρηση ως αρνητική και στην πραγματικότητα η παρατήρηση είναι αρνητική τότε βρισκόμαστε στο επάνω αριστερά κελί του πίνακα και η κατηγοριοποίηση χαρακτηρίζεται ως ορθά αρνητική πρόβλεψη (True Negative: TN). Τα υπόλοιπα κελιά ερμηνεύονται με τον ίδιο τρόπο αντίστοιχα για τις ορθά θετικές προβλέψεις (True Positives: TP) και λανθασμένα αρνητικές προβλέψεις (False Negatives: FN).

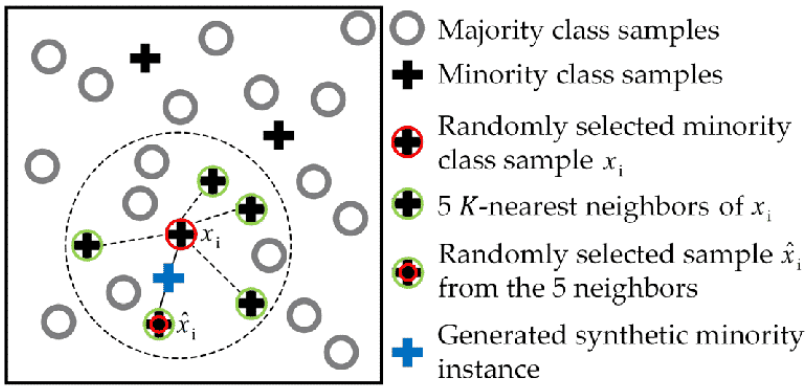
Έχοντας λοιπόν σαν στόχο την αύξηση της αποδοτικότητας των αλγορίθμων κατηγοριοποίησης έχουν προταθεί πολλαπλοί τρόποι εξισορροπημένης δειγματοληψίας και χωρίζονται κυρίως σε κύριες κατηγορίες όπως σημειώνονται στο [30]:

1) Εξισορροπημένη Δειγματοληψία σε Επίπεδο Δεδομένων:
    a. Τυχαιοποιημένη υπερ-δειγματοληψία της κλάσης μειοψηφίας, Αύξηση δηλαδή του αριθμού των παρατηρήσεων της κλάσης με τις λιγότερες παρατηρήσεις με τυχαίο τρόπο
    b. Τυχαία υπο-δειγματοληψία της κλάσης πλειοψηφίας. Μείωση του αριθμού παρατηρήσεων της κλάσης πλειοψηφίας.
    c. Καθοδηγούμενη υπερ-δειγματοληψία της κλάσης μειοψηφίας, όπου δεν παράγονται μεν καινούριες παρατηρήσεις αλλά η επιλογή των παρατηρήσεων που αντικαθίστανται δεν είναι τυχαία
    d. Καθοδηγούμενη υπο-δειγματοληψία της κλάσης πλειοψηφίας μέσω της οποίας με μη-τυχαίο τρόπο «αφαιρούνται» από το σύνολο δεδομένων παρατηρήσεις ενδιαφέροντος
    e. Καθοδηγούμενη υπερ-δειγματοληψία της κλάσης μειοψηφίας, όπου γίνεται παραγωγή νέων παρατηρήσεων
    f. Τέλος ο συνδυασμός των προηγούμενων
2) Εξισορροπημένη Δειγματοληψία σε Αλγοριθμικό Επίπεδο
    a. Ρύθμιση του κόστους (βάρους) συγκεκριμένων κλάσεων με στόχο της εξισορρόπηση των αποτελεσμάτων λόγο της ανομοιομορφίας των κατανομών
    b. Ρύθμιση της πιθανολογικής εκτίμησης στα φύλα των δέντρων (όταν χρησιμοποιούνται δέντρα αποφάσεων)
    c. Ρύθμιση των τιμών κατωφλίων απόφασης
    d. Εκπαίδευση με στόχο την αναγνώριση παρατηρήσεων της μειοψηφικής κλάσης και μόνο αντί της προσέγγισης της διάκρισης μεταξύ δύο η περισσοτέρων (κλάσεων)

Στην συνέχεια σημειώνονται τα κυριότερα λειτουργικά χαρακτηριστικά των πιο διαδεδομένων από τις μεθόδους που προαναφέρθηκαν παραθέτοντας τα αρνητικά και θετικά σημεία τους.

Για την επίτευξη μιας ομοιόμορφης κατανομής μεταξύ των κλάσεων της μεταβλητής στόχου, στο επίπεδο του συνόλου δεδομένων, θα πρέπει είτε να μειωθεί με κάποιο τρόπο ο αριθμός των παρατηρήσεων της πλειοψηφικής κλάσης είτε να αυξηθεί ο αριθμός των παρατηρήσεων της μειοψηφικής κλάσης. Με την τυχαία μείωση των παρατηρήσεων της κλάσης πλειοψηφίας ερχόμαστε αντιμέτωποι με τον κίνδυνο παράληψης «χρήσιμων» παρατηρήσεων έχοντας σαν αποτέλεσμα την δημιουργία ενός σχετικά «αδύναμου» προβλεπτικού μοντέλου. Απάντηση σε αυτό το πρόβλημα έρχεται να δώσει η τυχαιοποιημένη υπερ-δειγματοληψία της κλάσης μειοψηφίας. Αυτή η μέθοδος δημιουργεί με τυχαίο τρόπο αντίγραφα παρατηρήσεων συνεπώς δεν υπάρχει πιθανότητα απώλειας πληροφορίας, αλλά σύμφωνα με το [30] αυξάνει τη πιθανότητα εμφάνισης του φαινόμενου της υπερ-εκπαίδευσης (over-fitting). Εύκολα κανείς καταλαβαίνει γιατί συμβαίνει αυτό καθώς οι κανόνες που σχηματίζει το μοντέλο μπορεί να είναι μεν ακριβής αλλά έχουν σαν αναφορά «μία» παρατήρηση (και τα αντίγραφά της). Εκτός του κινδύνου της υπέρ-εκπαίδευσης (over-fitting), παράλληλα με την αύξηση των παρατηρήσεων, αυξάνεται και το κόστος των υπολογιστικών πόρων που αποτελεί από μόνο του μία πρόκληση ειδικά όταν γίνεται λόγος για επεξεργασία ανομοιόμορφων σχετικά μεγάλων συνόλων δεδομένων.

Μία μερική απάντηση στο πρόβλημα της υπερ-εκπαίδευσης έρχεται να δώσει η μέθοδος SMOTE (*Synthetic Minority Over-Sampling Technique*). Η SMOTE όπως χαρακτηριστικά αναφέρεται στο [31] ήταν έμπνευση από την προσέγγιση που ακολουθήθηκε από τους Ha & Bunke το 1997 [32] όπου για την αναγνώριση χειρόγραφων χαρακτήρων, ήταν απαραίτητο να εμπλουτιστεί το σύνολο εκπαίδευσης με επιπλέον παρατηρήσεις οι οποίες ήταν προϊόν επεξεργασίας (περιστροφή και σκίαση εικόνων π.χ.) των πραγματικών δεδομένων. Με ελαφρός διαφορετικό τρόπο η SMOTE δημιουργεί για κάθε «πραγματική» παρατήρηση της κλάσης μειοψηφίας [33], ένα σύνολο από παρατηρήσεις το οποίο τοποθετείται μεταξύ ενός ή περισσοτέρων κοντινότερων γειτόνων της αρχικής «πραγματικής» παρατήρησης. Στη συνέχεια γίνεται τυχαία επιλογή αυτών παρατηρήσεων σύμφωνα με το επιθυμητό πλήθος. Έχοντας ένα τόσο «δυνατό» εργαλείο όπως η SMOTE υπάρχει η δυνατότητα να εξαλείψει κανείς το πρόβλημα των ανομοιόμορφων κατανομών χωρίς την παράληψη χρήσιμης πληροφορίας και χωρίς τον κίνδυνο της υπερ-εκπαίδευσης.



Εικόνα 9 SMOTE [33]

Υπάρχουν παραδείγματα όπου η χρήση υπερ-δειγματοληψίας δημιουργεί προβλήματα σε ένα αλγόριθμο μηχανικής μάθησης, όπως φαίνεται και στο [34], όπου η απόδοση του αλγορίθμου (C4.5) δεν βελτιώθηκε σχεδόν καθόλου. Επιπροσθέτως αναφέρεται ότι το «κλάδεμα» των φύλλων του δέντρου απόφασης είναι εμφανώς μικρότερο έχοντας σαν αποτέλεσμα την μειωμένη γενίκευση του μοντέλου σε σύγκριση με την υπο-δειγματοληψία της πλειοψηφικής κλάσης.

3.2 Ανασκόπηση Αλγορίθμων Εξόρυξης Γνώσης

Η εξόρυξη γνώσης χωρίζεται σε δύο κύριες κατηγορίες ανάλυσης. Η πρώτη είναι η περιγραφική ανάλυση με σκοπό την ανάδειξη των ομαδοποιήσεων των εγγραφών αναλογικά με τις τιμές που τους έχουν ανατεθεί και η προεπισκόπηση διάφορων άλλων ιδιοτήτων των δεδομένων, που βοηθάνε στην καλύτερη κατανόησή τους. Η δεύτερη ανάλυση είναι η προγνωστική και στόχος της είναι, όπως αναφέρεται στην ονομασία της, η πρόγνωση μελλοντικών τιμών βάσει κάποιου μοντέλου που έχει κατασκευαστεί. Οι μέθοδοι της εξόρυξης γνώσης χωρίζονται σε δύο κύριες κατηγορίες. Ή πρώτη κατηγορία είναι αυτή που την αποτελούν οι αλγόριθμοι εποπτευόμενης μάθησης (supervised learning) και η δεύτερη κατηγορία είναι αυτή της μη-εποπτευόμενης μάθησης (unsupervised learning).

3.2.1 Εποπτευόμενη Μάθηση (Supervised Learning)

Οι μέθοδοι που ανήκουν στην εποπτευόμενη μάθηση είναι αυτοί που για την εξαγωγή συμπερασμάτων χρειάζονται ένα ιστορικό σύνολο δεδομένων στο οποίο ορίζεται μία μεταβλητή «στόχος». Σκοπός λοιπόν των μεθόδων αυτών είναι η ανάπτυξη μοντέλων που να συνδέουν τις απλές μεταβλητές με τη μεταβλητή στόχο. Με αυτό το τρόπο ένα καλά εκπαιδευμένο μοντέλο μπορεί να προβλέψει την τιμή της μεταβλητής στόχου, μιας νέας εγγραφής, εφόσον είναι διαθέσιμες οι υπόλοιπες ανεξάρτητες μεταβλητές. Τις συγκεκριμένες μεθόδους τις συναντάει κανείς όταν το αντικείμενο της ανάλυσης είναι απόδοση μιας διακριτής τιμής σε μία μεταβλητή στόχο, δηλαδή σε προβλήματα ταξινόμησης/κατηγοριοποίησης ενώ όταν ο στόχος της ανάλυσης είναι η προσέγγιση μιας συνεχής αριθμητικής τιμής, τα προβλήματα ονομάζονται παλινδρόμησης. Γίνεται κατανοητό λοιπόν ότι μέσω των μεθόδων εποπτευόμενης μάθησης δημιουργείται ένας μηχανισμός λήψης απόφασης (πρόβλεψης), για την μεταβλητή στόχο, αναλόγως των τιμών των ανεξάρτητων μεταβλητών. Σημαντικό μειονέκτημα των αλγορίθμων αυτής της κατηγορίας είναι ότι η εκπαίδευσή τους απαιτεί μεγάλο σύνολο δεδομένων συνεπώς είναι χρονοβόροι. Τέλος θα πρέπει να αναφερθεί ότι καθοριστικό ρόλο παίζει η ποιότητα του συνόλου δεδομένων εκπαίδευσης .

3.2.1.1 Αλγόριθμοι Κατηγοριοποίησης (Classification)

Οι αλγόριθμοι κατηγοριοποίησης έρχονται να απαντήσουν στα προβλήματα ταξινόμησης που θεωρούνται από τα πιο διαδεδομένα προβλήματα στο κόσμο της ανάλυσης. Σε αυτά τα προβλήματα κατηγοριοποίησης είναι γνωστό το γεγονός ότι τα δεδομένα χωρίζονται σε κατηγορίες/κλάσεις (διακριτών τιμών και όχι συνεχών) και αυτή η πληροφορία εντοπίζεται σε ένα τουλάχιστον χαρακτηριστικό το οποίο αναγνωρίζεται σαν μεταβλητή «στόχος». Αλγόριθμοι κατηγοριοποίησης εντοπίζονται σε πληθώρα εφαρμογών του σύγχρονου κόσμου με κυριότερες αυτές της οικονομίας, της προώθησης προϊόντων, της ιατρικής των τηλεπικοινωνιών κ.ο.κ.

Στη συνέχεια θα παρατεθούν μερικοί από τους πλέον εδραιωμένους αλγόριθμους κατηγοριοποίησης και θα παρουσιαστεί το θεωρητικό υπόβαθρο του κάθε ένα ξεχωριστά. Ιδιαίτερη έμφαση θα δοθούν σε αυτούς που θα χρησιμοποιηθούν αργότερα και στο πειραματικό στάδιο της συγκεκριμένης εργασίας.

3.2.1.1.1 Λογιστική Παλινδρόμηση (Logistic Regression)

Η λογιστική παλινδρόμηση πρόκειται για ένα στατιστικό μοντέλο το οποίο καλείται να ταξινομήσει μία εξαρτημένη μεταβλητή (συνήθως δυαδική) στηριζόμενο στις σχέσεις που την συνδέουν με μία η περισσότερες ανεξάρτητες. Δηλαδή ο ταξινομητής βασισμένος σε κάποια χαρακτηριστικά μίας παρατήρησης την χαρακτηρίζει παραδείγματος χάρει ως αληθής ή ψευδείς ή 0 ή 1 αναλόγως το σύνολο δεδομένων. Εκτός από τη περίπτωση της δυαδικής εξαρτημένη μεταβλητής υπάρχουν και περιπτώσεις όπου η λογιστική παλινδρόμηση χρησιμοποιείται για την ταξινόμηση παρατηρήσεων σε πάνω από 2

κατηγορίες. Η δεύτερη περίπτωση εξαρτημένης μεταβλητής είναι να της αποδίδονται τιμές οι οποίες όπως χαρακτηριστικά αναφέρεται στο [35] είναι περισσότερες από 2 σε πλήθος και μεταξύ αυτών ισχύει η έννοια της ανισότητας. Η μεταβλητή αυτή ονομάζεται τακτική. Η τρίτη περίπτωση χαρακτηρίζει την εξαρτημένη μεταβλητή ως ονοματική ή πολυωνυμική, έχοντας σαν χαρακτηριστικό την απόδοση τιμών όπου το πλήθος τους, ομοίως με την τακτική, είναι μεγαλύτερο των δύο αλλά σε αντίθεση με την προηγούμενη δεν υπάρχει κάποια σχέση μεταξύ των κατηγοριών αυτών.

Μερικά παραδείγματα χρησιμοποίησης της λογιστικής παλινδρόμησης είναι τα εξής:

- Στην ιατρική π.χ. έχοντας σαν είσοδο ένα σύνολο χαρακτηριστικών του ασθενή μπορεί να εκτιμηθεί η πιθανότητα εμφάνισης μίας ασθένειας που δεν έχει ακόμα εκδηλωθεί.
    - Σύνολο: {Ηλικία, Φύλλο, Καπνιστής, Ιστορικό, Αποτελέσματα εξετάσεων κλπ.}
    - Κατηγορίες: {Διαβήτης: 30%, ΧΑΠ: 20% κλπ.}
- Στην προώθηση προϊόντων και πρόβλεψη της πρόθεσης αγοράς ενός προϊόντος
- Στην πρόβλεψη της πιθανότητας αν ένας δανειολήπτης θα είναι συνεπής στις υποχρεώσεις του απέναντι στη τράπεζα που του παρείχε το δάνειο

Γίνεται αντιληπτή η πληθώρα των εφαρμογών που η χρήση της λογιστικής παλινδρόμησης παρέχει πολύτιμες προβλέψεις για την έκβαση καταστάσεων πριν ακόμα αυτές πραγματοποιηθούν.

Η δυαδική λογιστική παλινδρόμηση εκφράζεται από την σχέση:

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

- z: Μεταβλητή εισόδου
- $f(z)$: Πιθανότητα αποτελέσματος

Πιο συγκεκριμένα η μεταβλητή *z* εκφράζει τη δράση μιας ομάδας ανεξάρτητων μεταβλητών και το *f(z)* είναι η πιθανότητα ενός αποτελέσματος λόγω της δράσης αυτής. Το *z* προσδιορίζει τη δράση όλων των ανεξάρτητων μεταβλητών και αυτό επιτυγχάνεται με την εξής σχέση.

$$z = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_\kappa X_\kappa$$

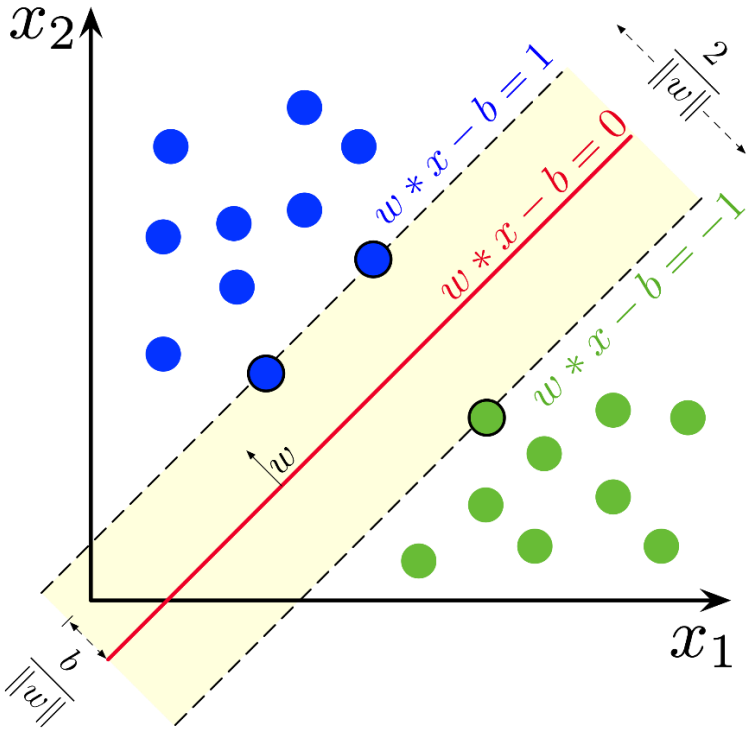
- β_0: Ύψος κλίσης της γραμμής παλινδρόμησης και ισούται με το *z* όταν οι τιμές όλων των ανεξάρτητων μεταβλητών είναι μηδέν
- β_i: Συντελεστές παλινδρόμησης εκφράζοντας τον βαθμό συνεισφοράς της συγκεκριμένης μεταβλητής

Θετική τιμή των συντελεστών παλινδρόμησης σημαίνει ότι η συγκεκριμένη ανεξάρτητη μεταβλητή αυξάνει τη πιθανότητα επιτυχημένης πρόβλεψης του συγκεκριμένου γεγονότος. Αντίθετα η αρνητική τιμή ενός συντελεστή υποδηλώνει ότι η τιμή της συγκεκριμένης ανεξάρτητης μεταβλητής μειώνει τη πιθανότητα εμφάνισης του γεγονότος.

3.2.1.1.2 Μηχανές Διανυσμάτων Υποστήριξης (Support Vector Machine)

Οι μηχανές διανυσμάτων υποστήριξης είναι γραμμικά μοντέλα που χρησιμοποιούνται για την επίλυση προβλημάτων κατηγοριοποίησης και παλινδρόμησης. Η κύρια ιδέα του αλγορίθμου είναι εύρεση μια "γραμμής" η οποία θα είναι σε θέση να διαχωρίσει ικανοποιητικά τις κλάσεις. Η τοποθέτηση της γραμμής αυτής γίνεται αρχικά με την εύρεση δύο σημείων (διαφορετικών κλάσεων) που είναι τα κοντινότερα προς την άλλη κλάση αλλά ταυτόχρονα μεγιστοποιούν την απόσταση από τη γραμμή. Αυτά τα σημεία

ονομάζονται διανύσματα υποστήριξης. Σε αυτό το σημείο πρέπει να ειπωθεί ότι αν είναι επιτακτική ανάγκη το μοντέλο κάνει υπέρθεση του συνόλου δεδομένων σε διαστάσεις μεγαλύτερες από αυτό με στόχο την εύρεση της γραμμής (επιπέδου για το τρισδιάστατο χώρο κλπ.) αυτής που θα διαχωρίζει αποδοτικότερα τις δύο κλάσεις. Έτσι λοιπόν όταν κληθεί να ταξινομήσει μια νέα παρατήρηση αρκεί να γίνει σύγκριση του σημείου εκείνου με τη γραμμή (επίπεδο κλπ.) και αναλόγως από ποια πλευρά της γραμμής βρίσκεται το νέο σημείο τους το κατηγοριοποιεί στην αντίστοιχη κλάση. Στη συνέχεια παρουσιάζεται εικόνα που διευκολύνει τη κατανόηση της δημιουργίας της γραμμής διαχωρισμού των δύο κλάσεων στο δισδιάστατο χώρο.



Εικόνα 10 Απεικόνιση Διανυσμάτων Υποστήριξης Σε Δύο Διαστάσεις [36]

Τα σημεία που εντοπίζονται επάνω στις διακεκομμένες γραμμές ονομάζονται διανύσματα υποστήριξης και επιλέγονται αυτά καθώς μεγιστοποιείται η απόσταση της ευθείας $w * x - b = 0$ από τις ευθείες $w * x - b = 1$ και $w * x - b = -1$. Το μοντέλο αγνοεί όλα τα υπόλοιπα σημεία από τη στιγμή που δημιουργεί την ευθεία διαχωρισμού των κλάσεων καθώς όποια παρατήρηση βρίσκεται αριστερά ή δεξιά της ευθείας κατηγοριοποιείται αναλόγως.

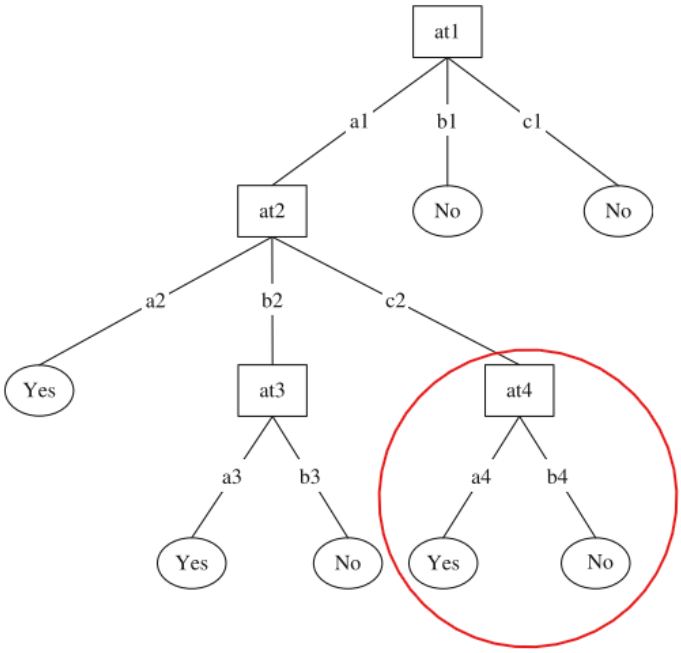
Αξίζει να σημειωθεί ότι η παραπάνω διαδικασία μπορεί να υλοποιηθεί και για σύνολα δεδομένων που δεν μπορούν οι κλάσεις να διαχωριστούν γραμμικά, με την χρήση της μεθόδου του κόλπου πυρήνα (kernel trick) [37]. Η μέθοδος αυτή αντιστοιχίζει τα δεδομένα σε ένα χώρο μεγαλύτερης διάστασης με σκοπό την ευκολότερη εύρεση ενός υπέρ-επιπέδου ικανού να διαχωρίσει αποτελεσματικά τις κλάσεις των δεδομένων. Τρεις είναι οι επικρατέστερες μέθοδοι πυρήνων (Kernel) που χρησιμοποιούνται σύμφωνα με το [37] για την εύρεση του κατάλληλου υπέρ-επιπέδου για δεδομένα που δεν διαχωρίζονται γραμμικά:

- Πολυωνυμικός πυρήνας (Polynomial Kernel)
- RBF – πυρήνας (Radial Basis Function Kernel)
- Σιγμοειδής Πυρήνας (Sigmoid Kernel)

3.2.1.1.3 Δέντρα Αποφάσεων (Decision Trees)

Πρόκειται για ένα αλγόριθμο εποπτευόμενης μηχανικής μάθησης με απώτερο σκοπό τη λήψη σωστής απόφασης (κατηγοριοποίηση) για μία νέα παρατήρηση που αποτελείται από κάποια χαρακτηριστικά. Ονομάζονται δέντρα αποφάσεων λόγω της δενδροειδής τους μορφής καθώς δημιουργούν κόμβους απόφασης και εκτείνονται σε πολλαπλά βάθη αν κριθεί απαραίτητο. Κύρια χαρακτηριστικά των δέντρων απόφασης είναι:

- Η ρίζα: Ως ρίζα χαρακτηρίζεται ο αρχικός κόμβος, όπως θα αναλυθεί αργότερα αποτελείται από το χαρακτηριστικό με το μεγαλύτερο βάρος, χωρίζοντας το σύνολο εκπαίδευσης σε δύο η περισσότερα υποσύνολα
- Εσωτερικοί κόμβοι: Ονομάζονται οι κόμβοι που δεν είναι ούτε στην αρχή του δέντρου αποφάσεων αλλά ούτε και στο τέλος (φύλλα).
- Φύλλα: Ονομάζονται οι κόμβοι που δεν έχουν «απογόνους» και είναι οι τελικοί κόμβοι ενός δέντρου απόφασης. Συνεπώς αναλόγως σε ποιο φύλλο θα καταλήξει η διαδοχική λήψη αποφάσεων, ανάλογη των κόμβων, θα αποδοθεί και η κατάλληλη κατηγοριοποίηση στη παρατήρηση.



Εικόνα 11 Απεικόνιση Δέντρου Απόφασης Πηγή:[38]

Όλοι οι κόμβοι εκτός των τελικών (φύλλα) εκφράζονται από συνθήκες βάση των οποίων γίνεται η διάσπαση των δεδομένων σε κάθε επίπεδο. Άρα μέσω μιας επαναληπτικής διαδικασίας διαμερισμού των δεδομένων με το καταλληλότερο, κάθε στιγμή, χαρακτηριστικό πραγματοποιείται η φάση της ανάπτυξης του δέντρου και διαρκεί είτε μέχρι να ικανοποιηθεί κάποιο κριτήριο τερματισμού, είτε να φτάσει σε τέτοιο επίπεδο όπου δεν είναι δυνατή η περαιτέρω διάσπαση του συνόλου δεδομένων. Μετά τη φάση της ανάπτυξης του δέντρου απόφασης, σειρά έχει η φάση του «κλαδέματος» (pruning). Σκοπός του είναι γενίκευση του μοντέλου αποφεύγοντας το φαινόμενο της υπέρ προσαρμογής στα δεδομένα εκπαίδευσης, καθιστώντας το πιο ικανό να κατηγοριοποιήσει σωστά, άγνωστες για αυτό μέχρι στιγμής παρατηρήσεις. Το «κλάδεμα» γίνεται κατά κύριο λόγο μετά την ολοκλήρωση της ανάπτυξης του δέντρου αφαιρώντας ένα τμήμα του. Αν π.χ. αφαιρούσαμε το «κλαδί» που είναι εντός του κόκκινου κύκλου στην εικόνα 11.

Υπάρχουν και χρησιμοποιούνται πληθώρα κριτηρίων διαχωρισμού των δεδομένων στους κόμβους και τα πιο διαδεδομένα είναι τα εξής:

- Κέρδος Πληροφορίας (Information Gain)
- Λόγος Κέρδους Πληροφορίας (Gain Ratio)
- Gini Index
- Απόσταση χ^2 (Chi-square)
- Ακρίβεια (Accuracy)

Η πιο διαδεδομένη τεχνική διαμέρισης του συνόλου δεδομένων στα δέντρα αποφάσεων είναι αυτή του κέρδους πληροφορίας και βασίζεται στο υπολογισμό της εντροπίας κάθε πιθανής τιμής κατωφλίου έχοντας σαν σκοπό την πιο «καθαρή» διαμέριση των κλάσεων χρησιμοποιώντας ένα χαρακτηριστικό. Ο τύπος του κέρδους πληροφορίας και της εντροπίας είναι οι εξής[38]:

$$Information\ Gain\ (IG) = Entropy(S) - \sum \frac{S_v}{S} Entropy(S_v)$$

$$Entropy = \sum -p_i \log\left(p_i\right)$$

- $Entropy(S)$: Εντροπία του κόμβου «πατέρα» ο οποίος αν είναι η ρίζα θα έχει τη μέγιστη τιμή που μπορεί να έχει μία εντροπία δηλαδή το ένα (1)
- *v*: Ο κόμβος «παιδί» για την οποία μετράμε την εντροπία
- $\frac{S_v}{S}$: Βάρος της εντροπίας του κόμβου *v* που εκφράζεται από το λόγο του πλήθους των διαμοιρασμένων δεδομένων προς το σύνολο των δεδομένων του κόμβου προγόνου *S.*
- p_i: Η πιθανότητα εμφάνισης της κλάσης *i* στο κόμβο

Η εντροπία υπολογίζεται για όλες τις πιθανές κλάσεις σε κάθε κόμβο του ίδιου επιπέδου και σε συνδυασμό με την εντροπία του κόμβου προγόνου είναι δυνατή η εύρεση του κέρδους πληροφορίας. Αν η τιμή της εντροπίας για μία τιμή διαχωρισμού είναι μηδέν σε ένα κόμβο απόγονο, τότε εντοπίζονται παρατηρήσεις μόνο μίας κλάσης στο κόμβο αυτό και η διαδικασία τερματίζεται εκεί για τον κόμβο αυτό καθώς δεν μπορεί να διαχωριστεί περαιτέρω. Η διαδικασία θα συνεχιστεί για τους υπόλοιπους κόμβους που δεν έχουν μηδενική εντροπία. Η τιμή του κατωφλίου που μεγιστοποιεί το κέρδος πληροφορίας διατηρείται για τον υπό εξέταση κόμβο σαν συνθήκη διαχωρισμού του συνόλου δεδομένων και η διαδικασία επαναλαμβάνεται για τα υπόλοιπα χαρακτηριστικά.

Εξαιρετικά σύνηθες τεχνικές διαχωρισμού των συνόλου δεδομένων είναι η Gini Index που ορίζεται σαν κριτήριο ανομοιομορφίας των συνόλων, μετρώντας τις αποκλίσεις μεταξύ των κατανομών πιθανότητας των τιμών του χαρακτηριστικού και ο λόγος του κέρδους πληροφορίας ο οποίος κανονικοποιεί το κέρδος πληροφορίας. Λόγω του ότι δεν μπορεί να προσδιοριστεί όταν η εντροπία είναι μηδενική για ένα κόμβο πρώτα υπολογίζονται όλα τα κέρδη πληροφορίας για όλα τα χαρακτηριστικά και επιλέγεται αυτό με το καλύτερο λόγο. Έρευνες έχουν δείξει ότι τείνει να ξεπερνάει το απλό κέρδος πληροφορίας όσον αφορά την ακρίβεια αλλά και τη πολυπλοκότητα σύμφωνα με το[39].

3.2.1.1.4 Gaussian Naive Bayes Ταξινομητής

Οι ταξινομητές που είναι βασισμένοι στο θεώρημα του Bayes πρόκειται για απλούς πιθανοτικούς κατηγοριοποιητές και στηρίζονται στην υπόθεση της πλήρους ανεξαρτησίας των χαρακτηριστικών. Κύρια εφαρμογή του συγκεκριμένου κατηγοριοποιητή εντοπίζεται στον χαρακτηρισμό εγγράφων κειμένων ως αθλητικό ή πολιτικό π.χ. αξιολογώντας τις συχνότητες εμφάνισης των λέξεων. Άλλη πολύ δημοφιλή σύγχρονη εφαρμογή των συγκεκριμένων κατηγιορποιητών γίνεται από τους παρόχους ηλεκτρονικής αλληλογραφίας για το χαρακτηρισμό των μηνυμάτων ως ανεπιθύμητα, διαφημιστικά κλπ. Μία δεύτερη υπόθεση είναι ότι οι τιμές των χαρακτηριστικών των παρατηρήσεων ακολουθούν την κανονική κατανομή και είναι υπολογίσιμη η μέση τιμή και η διασπορά τους (Εξού και το όνομα της, Gaussian Naive Bayes).

Πιο συγκεκριμένα ο Gaussian Naive Bayes χρησιμοποιεί το θεώρημα της δεσμευμένης πιθανότητας του Bayes της οποίας ο τύπος σύμφωνα με το [40]:

$$P\left(\frac{C}{X}\right) = P(C) * \frac{P\left(\frac{X}{C}\right)}{P(X)}$$

- *P(C)* : η πιθανότητας εμφάνισης της κατηγορίας C στο σύνολο δειγμάτων εκπαίδευσης
- $P\left(\frac{X}{C}\right)$: η πιθανότητα εμφάνισης των χαρακτηριστικών X δεδομένου ότι τα χαρακτηριστικά ανήκουν στην κατηγορία C
- P(X): η πιθανότητα εμφάνισης των χαρακτηριστικών X ανεξαρτήτου κατηγορίας

Η μέθοδος αυτή συγκεντρώνει πολλά θετικά χαρακτηριστικά με κύριο το γεγονός ότι μπορεί να αποδώσει αποτελέσματα για διαφορετικούς τύπους δεδομένων συμπεριλαμβανομένων των κειμένων και των εικόνων[41, 42]. Επιπλέον πλεονέκτημα του Naive Bayes είναι ότι απαιτεί μόνο ένα μικρό αριθμό δεδομένων εκπαίδευσης για την εκτίμηση νέων παρατηρήσεων το οποίο μπορεί να μεταφραστεί και σαν "γρήγορος" αλγόριθμος.

Κύριο μειονέκτημα της μεθόδου αυτής είναι η υπόθεση της ανεξαρτησίας των χαρακτηριστικών που κατά πάσα πιθανότητα δεν είναι ακριβείς ειδικότερα όταν γίνεται λόγος για προβλήματα πραγματικού κόσμου. Τέλος επηρεάζεται σε σημαντικό βαθμό από την ανισορροπία των κλάσεων.

Παραλλαγές του Gaussian Naive Bayes είναι εκείνες που χρησιμοποιούν ξανά το θεώρημα της δεσμευμένης πιθανότητας αλλά χωρίς την υπόθεση εκείνη που "θέλει" τις τιμές των χαρακτηριστικών να ακολουθούν την κανονική κατανομή. Αντίστοιχα υπάρχουν ταξινομητές κατά Bayes για σύνολα δεδομένων που ακολουθούν την κατανομές poisson, πολυωνυμικές και Bernoulli [43].

3.2.1.1.5 Πολύ-Επίπεδοι Αισθητήρες (Multilayer Perceptron)

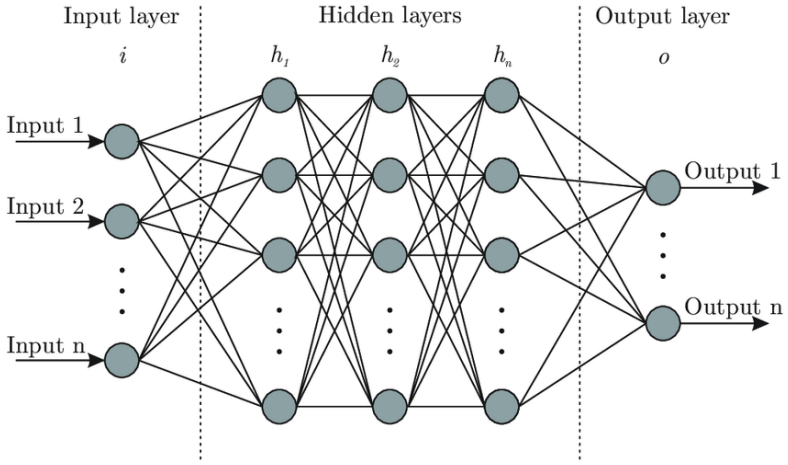
Ο αλγόριθμος των πολυ-επίπεδων αισθητήρων ανήκει σε μία διαφορετική κατηγορία αλγορίθμων από τους προηγούμενους που είδαμε στην ενότητα αυτή. Για την ακρίβεια οι προηγούμενοι αλγόριθμοι εντάσσονται κάτω από την ομπρέλα της μηχανικής μάθησης ενώ αντίθετα ο αλγόριθμος των πολυ-επίπεδων αισθητήρων είναι χαρακτηριστικό παράδειγμα αλγορίθμου νευρωνικών δικτύων. Τα τεχνικά νευρωνικά δίκτυα εμπνεύστηκαν από τη λειτουργία του εγκεφάλου ο οποίος αποτελείται από δισεκατομμύρια νευρώνες οι οποίοι είναι συνδεδεμένοι μεταξύ τους με τις νευρωνικές συνάψεις. Έτσι λοιπόν και τα τεχνητά νευρωνικά δίκτυα αποτελούνται από απλούς υπολογστικούς κόμβους (νευρώνες) οι οποίοι συνδέονται μεταξύ τους (συνάψεις). Κάθε διασύνδεση αποδίδει μία τιμή βάρους στο αντίστοιχο χαρακτηριστικό το οποίο προσαρμόζεται κατά το στάδιο της εκπαίδευσης του μοντέλου. Στόχος των βαρών είναι να οδηγηθεί η αρχική είσοδος προς την τιμή της επιθυμητής εξόδου. Όπως πολύ

χαρακτηριστικά αναφέρεται και στο [44] η λειτουργία των νευρωνικών δικτύων μπορεί να χαρακτηρισθεί και ως μία διεργασία χαρτογράφησης (mapping) των αναγκαίων βαρών προκειμένου από τα δεδομένα εισόδου να φτάσουμε στα επιθυμητά αποτελέσματα.

Οι νευρώνες κατηγοριοποιούνται σε επίπεδα (layers) τα οποία με τη σειρά τους χωρίζονται σε επίπεδο εισόδου (input layer), κρυφό επίπεδο (hidden layer) και επίπεδο εξόδου (output layer).

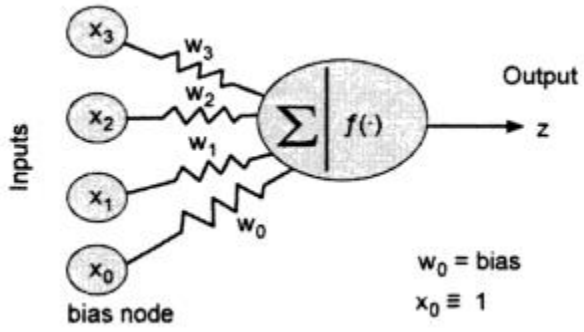
- Επίπεδο εισόδου: Αποτελείται από κόμβους όπου γίνεται η είσοδος των δεδομένων
- Κρυφό επίπεδο: Ως κρυφό επίπεδο αναγνωρίζεται όποιο επίπεδο κόμβων βρίσκεται μεταξύ του επιπέδου εισόδου και εξόδου. Τέλος μπορεί να είναι παραπάνω από ένα τα κρυφά επίπεδα. Ο αριθμός τους προσδιορίζεται κυρίως από τις ανάγκες του μοντέλου κατηγοριοποίησης αλλά επίσης και από την διαθέσιμη επεξεργαστική ισχύ καθώς αναφερόμαστε σε δαπανηρές ως προς το χρόνο διαδικασίες.
- Επίπεδο εξόδου: Είναι το τελευταίο επίπεδο του νευρωνικού δικτύου από το οποίο προκύπτουν και τα αποτελέσματα

Τα νευρωνικά δίκτυα χωρίζονται σε δύο κύριες κατηγορίες αναλόγως την ροή της πληροφορίας. Η πρώτη κατηγορία που ανήκει και ο αλγόριθμος των πολυ-επίπεδων αισθητήρων, είναι αυτή της πρόσθιας τροφοδότησης (Feed Forward), όπου η πληροφορία κατευθύνεται συνεχώς από το ένα επίπεδο στο άλλο με μοναδική κατεύθυνση από το επίπεδο εισόδου σε αυτό της εξόδου. Δεν υπάρχει, δηλαδή, κάποια ανατροφοδότηση από την έξοδο του. Η δεύτερη κατηγορία είναι εκείνη που χαρακτηρίζεται από την ανατροφοδότηση (Feed Back) της εξόδου ενός τουλάχιστον κόμβου σε είσοδο ενός προγενέστερου κόμβου. Στην παρακάτω εικόνα γίνεται αντιληπτή η δομή ενός πολυεπίπεδου τεχνητού νευρωνικού δικτύου μονόδρομης τροφοδότησης.



Εικόνα 12 Δομή Τεχνητού Νευρωνικού Δικτύου Πολλαπλών Επιπέδων πηγή: [45]

Το επίπεδο εισόδου είναι εκείνο που δέχεται τα αρχικά δεδομένα και δεν γίνεται σε αυτό κάποια επεξεργασία. Με τη σειρά τους τα κρυφά επίπεδα λαμβάνουν την πληροφορία από το επίπεδο εισόδου και μετά την εφαρμογή της κατάλληλης επεξεργασίας των δεδομένων προωθούνται στο επίπεδο εξόδου. Βλέποντας τον νευρώνα της εικόνας 13 γίνεται αντιληπτή η διαδικασία μέσω της οποία ένας νευρώνας κρυφού επιπέδου επεξεργάζεται την πληροφορία αποδίδοντας τη έξοδο (output) *z*.



Εικόνα 13 Απεικόνιση Διεργασιών Νευρώνα πηγή: [44]

Σε κάθε ένα νευρώνα λοιπόν υπολογίζεται η έξοδος z βάσει των εισόδων σε συνδυασμό με τα αντίστοιχα βάρη και αυτή η διεργασία πραγματοποιείται σε όλα τα κρυφά επίπεδα. Το z της εικόνας 13 υπολογίζεται με την βοήθεια του παρακάτω τύπου:

$$z = f(\sum_{i=0}^{3} w_i x_i)$$

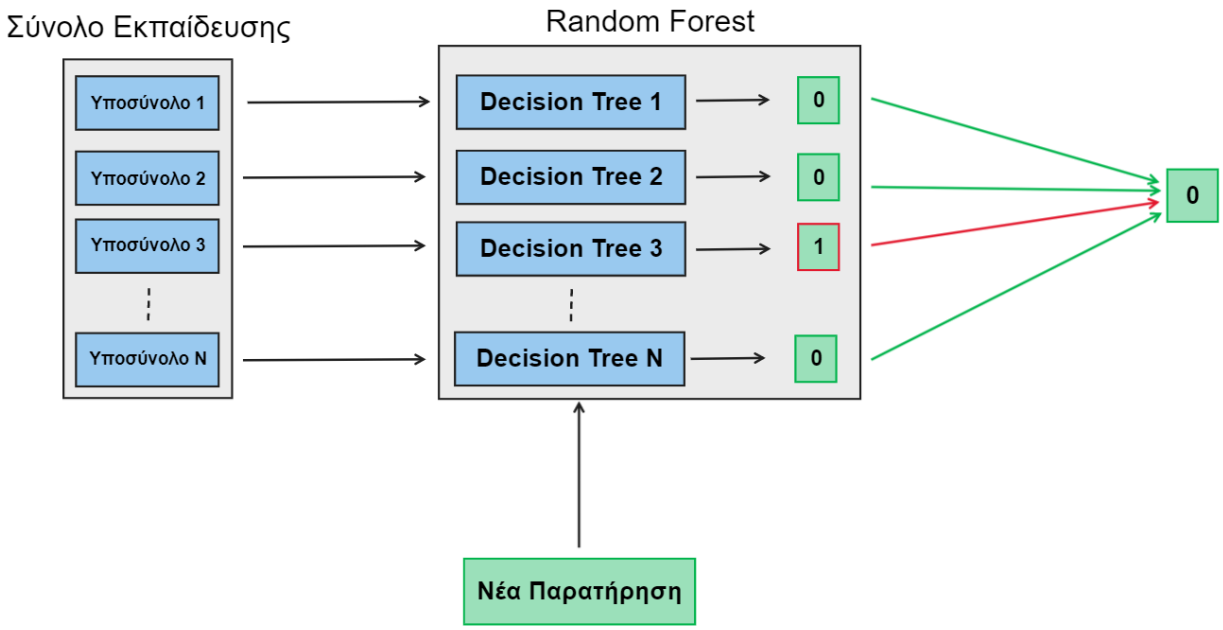
Πρώτα εντός της παρένθεσης υπολογίζεται ένα σταθμισμένο άθροισμα των δεδομένων εισόδου και των αντίστοιχων βαρών τους και στη συνέχεια κάνοντας χρήση μίας από τις συναρτήσεις μεταφοράς f υπολογίζεται η z. Η πιο συνήθης συνάρτηση μεταφοράς ειδικότερα όταν ερευνάται πρόβλημα κατηγοριοποίησης δυαδικών μεταβλητών είναι αυτή της σιγμοειδούς (λογιστική) ο τύπος της οποίας παρουσιάζεται στη συνέχεια:

$$z = \frac{1}{1 + e^{-(\sum_i w_i x_i + w_0)}}$$

3.2.1.1.6 Τυχαία Δάση (Random Forest)

Ο αλγόριθμος των τυχαίων δασών είναι ένας αλγόριθμος μηχανικής μάθησης που η δημοφιλία του συνεχώς αυξάνεται και έχει φτάσει στο σημείο να ανταγωνίζεται τα δέντρα αποφάσεων (βάσει των οποίων δημιουργείται). Χρησιμοποιείται σε προβλήματα κατηγοριοποίησης και παλινδρόμησης. Δεδομένου του προβλήματος της συγκεκριμένης εργασίας που δεν είναι άλλο από την κατηγοριοποίηση νέων παρατηρήσεων θα αναλυθεί σε βάθος η λειτουργία αυτού του αλγορίθμου που στοχεύει στην ορθή ταξινόμηση των νέων παρατηρήσεων. Επιπλέον θα αναφερθούν πλεονεκτήματα και μειονεκτήματα της συγκεκριμένης μεθόδου καθώς επίσης θα γίνει και μία σύγκριση με τα απλά δέντρα αποφάσεων που είδαμε σε προηγούμενη ενότητα.

Αρχικά τα δάση αποτελούνται από πολλά δέντρα τα οποία δημιουργούνται από τυχαία διασπασμένα υποσύνολα του συνόλου εκπαίδευσης. Έτσι δημιουργούνται πολλαπλά δέντρα απόφασης τα οποία διαφέρουν σημαντικά ως προς την δομή τους αφού εκπαιδεύονται με “διαφορετικά”, φαινομενικά, δεδομένα. Με την είσοδο των χαρακτηριστικών μιας νέας παρατήρησης προς ταξινόμηση, ο ταξινομητής προβάλει τα χαρακτηριστικά αυτά σε κάθε ένα από τα δέντρα που αποτελούν το “δάσος” και το αποτέλεσμα αυτού είναι η πλειοψηφική κλάση, ή διαφορετικά το αποτέλεσμα μιας συνάρτησης που λαμβάνει υπόψη πιθανά βάρη που έχουν ανατεθεί.



Εικόνα 14 Αναπαράσταση Διαδικασίας Αλγορίθμου Τυχαίων Δασών (Random Forest)

Συνεπώς ο αλγόριθμος των τυχαίων δασών αποτελεί μία περίπτωση εκμάθησης συνόλου ensemble learning. Δηλαδή ανήκει στην ομάδα των ταξινομητών οι οποίοι αποτελούνται από περισσότερους από ένα αλγορίθμους κατηγοριοποίησης με στόχο την καλύτερη απόδοση του μοντέλου, από ότι θα είχε ο κάθε ένας από τους ταξινομητές που το συνθέτουν ξεχωριστά.

Το κλειδί για τον επιτυχημένο συνδυασμό πολλαπλών μοντέλων είναι η χαμηλή συσχέτιση μεταξύ των μοντέλων. Έτσι λοιπόν τα δέντρα μεταξύ τους αλληλοπροστατεύονται από λάθος εκτιμήσεις, αρκεί να μην κάνουν πολλά "υπό"-μοντέλα το ίδιο λάθος, πράγμα απίθανο για χαμηλής συσχέτισης ταξινομητές. Ο αλγόριθμος έχει κάποιους "μηχανισμούς" οι οποίοι προστατεύουν το τελικό μοντέλο δημιουργώντας ασυσχέτιστα επιμέρους δέντρα αποφάσεων.

Η πρώτη μέθοδος ονομάζεται Bagging (Bootstrap Aggregation). Με αυτή τη μέθοδο τα υποσύνολα εκπαίδευσης που δημιουργούνται προστίθενται τυχαίες παρατηρήσεις (πολλές φορές επαναλαμβανόμενες) με σκοπό την δημιουργία ενός συνόλου ίσου μήκους με το αρχικό. Επειδή τα δέντρα αποφάσεων είναι πολύ "ευαίσθητα" και επηρεάζεται σημαντικά η δομή τους από τα δεδομένα εκπαίδευσης ο random forest δημιουργεί μοντέλα χαμηλής συσχέτισης. Αυτό βέβαια αν δεν το παραμετροποιήσουμε αναλόγως, δηλαδή να δημιουργούμε δέντρα αποφάσεων με μικρότερα σύνολα δεδομένων σε μήκος.

Η δεύτερη μέθοδος χαρακτηριζεται από την τυχαία επιλογή των χαρακτηριστικών διάσπασης. Στα απλά δέντρα απόφασης ένας κόμβος διαχωρισμού αποτελείται από το χαρακτηριστικό εκείνο που χωρίζει καλύτερα το σύνολο δεδομένων στους αντίστοιχους απογονικούς κομβους. Για την δημιουργία ακόμα πιο ασυσχέτιστων μοντέλων ο random forest επιλέγει τυχαία χαρακτηριστικά και τα αποδίδει στους αντίστοιχους κόμβους διαχωρισμού.

Πλεονεκτήματα:

- Ισχυρός αλγόριθμος που δεν απαιτεί μεγάλη παραμετροποίηση

- Ανθεκτικός στο φαινόμενο της υπερ προσαρμογής
- Μπορεί να χρησιμοποιηθεί για μεγάλα σύνολα δεδομένων
- Πραγματοποιείται feature selection για την μείωση της διάστασης των χαρακτηριστικών

Μειονεκτήματα:

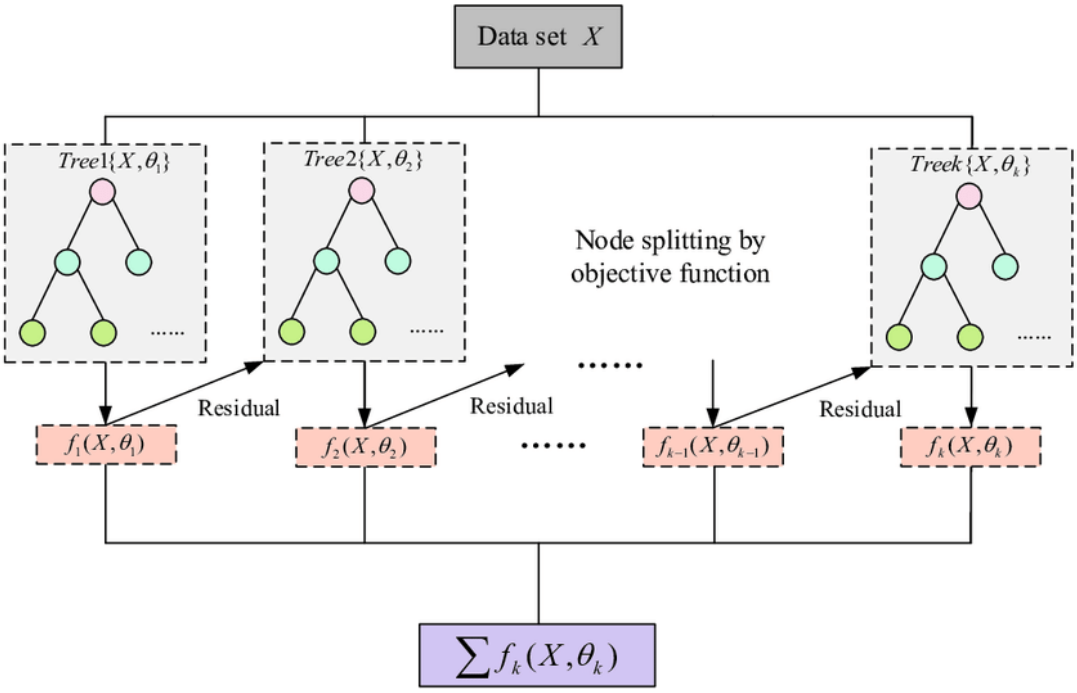
- Σημαντικά αυξημένος χρόνος εκπαίδευσης
- Αυξημένες απαιτήσεις σε μνήμη (αποθήκευση πολλαπλών δέντρων απόφασης)
- Δυσκολία ερμηνείας αποτελεσμάτων των επιμέρους δέντρων

3.2.1.1.7 XGBOOST

Πρόκειται για έναν αλγόριθμο εξαιρετικά ισχυρό ο οποίος κερδίζει έδαφος όλο και περισσότερο το τελευταίο καιρό. Χρησιμοποιείται για προβλήματα κατηγοριοποίησης αλλά και παλινδρόμησης. Ομοίως με τον αλγόριθμων των τυχαίων δασών που παρουσιάστηκε προηγουμένως, είναι ένας αλγόριθμος εκμάθησης συνόλου όπου εκμεταλλευόμενος πολλαπλά απλούστερα μοντέλα μηχανικής μάθησης (συνηθέστερα δέντρα απόφασης) κατασκευάζει ένα ταξινομητή κάνοντας ανάθεση κατάλληλων βαρών στο εκάστοτε «υπό» κατηγοριοποιητή. Πληθώρα εφαρμογών έχουν υλοποιηθεί με την βοήθεια του XGBOOST μερικές εκ των οποίων είναι οι εξής σύμφωνα με το [46].

- Κατηγοριοποίηση κειμένου στο διαδίκτυο
- Εκτίμηση συμπεριφοράς πελατών
- Ανίχνευση κίνησης
- Κατηγοριοποίηση malware
- Συστήματα συστάσεων (Netflix prize)
- Εκτίμηση πωλήσεων καταστημάτων

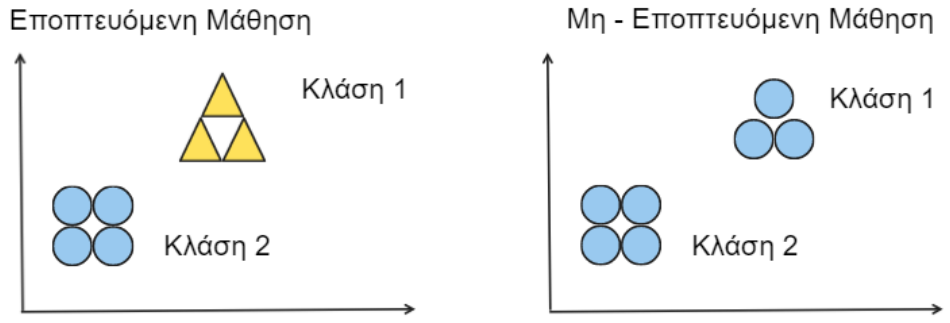
Ο XGBOOST αποτελεί μία επέκταση του αλγορίθμου βελτιστοποίησης κλίσης (Gradient Boosting) ενσωματώνοντας λειτουργίες που ελαχιστοποιούν την πιθανότητα υπερ-προσαρμογής του μοντέλου αλλά επίσης βελτιστοποιεί την επιλογή των σημαντικότερων χαρακτηριστικών. Ένα από τα κύρια πλεονεκτήματα του XGBOOST είναι η επεκτασιμότητά του ανεξαρτήτως σεναρίου εφαρμογής καθώς επίσης το γεγονός ότι είναι σημαντικά ταχύτερος αλγοριθμικά από όμοιες λύσεις. Σε συνδυασμό με την παράλληλη επεξεργασία ο XGBOOST είναι ένα «υπέρ όπλο» στα χέρια των επιστημόνων των δεδομένων παρέχοντάς τους την δυνατότητα να επεξεργάζονται εκατοντάδες χιλιάδες παραδείγματα σε ένα προσωπικό υπολογιστή, πόσο μάλλον δε σε μία συστοιχία υπολογιστικών πόρων. Τέλος είναι ανθεκτικός σε περιπτώσεις όπου τα σύνολα δεδομένων εμπεριέχουν ελλιπείς τιμές, πράγμα καθόλου απίθανό ειδικά όταν γίνεται λόγος για δεδομένα πραγματικού κόσμου [47].



Εικόνα 15 Απεικόνιση Μεθόδου XGBOOST πηγή: [47]

3.2.2 Μη Εποπτευόμενη Μάθηση (Unsupervised Learning)

Οι αλγόριθμοι που ανήκουν στη κατηγορία της μη-εποπτευόμενης μάθησης σε αντίθεση με αυτούς της εποπτευόμενης το αποτέλεσμα τους στηρίζεται στις ιδιότητες των μεταβλητών. Δηλαδή οι αλγόριθμοι έχοντας σαν εισόδους τις τιμές των μεταβλητών κατασκευάζουν συστάδες, ομάδες και συσχετίσεις όμοιων στοιχείων. Βασική διαφορά από την κατηγορία αλγορίθμων εποπτευόμενης μάθησης είναι ότι δεν είναι γνωστή η κλάση (μεταβλητή στόχος) του εκάστοτε στοιχείου που συνθέτει το σύνολο δεδομένων εκπαίδευσης.



Εικόνα 16 Απεικόνιση Διαφορών Εποπτευόμενης και Μη-Εποπτευόμενης Μηχανικής Μάθησης

Όπως φαίνεται πολύ χαρακτηριστικά στην εικόνα 16 στη περίπτωση των αλγορίθμων εποπτευόμενης μηχανικής μάθησης είναι γνωστό σε ποια κατηγορία εντάσσονται τα δεδομένα. Αντιθέτως οι αλγόριθμοι μη εποπτευόμενης μηχανικής μάθησης τα δεδομένα σε πρώτη «ανάγνωση» είναι ίδια χωρίς να υπάρχει

κάποια «ταμπέλα» που να τα ξεχωρίζει. Στην συνέχεια οι αλγόριθμοι αυτής της κατηγορίας «επιχειρούν» να «ανακαλύψουν» ποια περίπτωση ανήκει σε ποια κατηγορία. Αλγόριθμοι τέτοιου τύπου χρησιμοποιούνται συνήθως για την εύρεση μοτίβων (patterns) στα δεδομένα με ελάχιστη αν όχι καθόλου την ανθρώπινη επίδραση επί των αλγορίθμων. Κάποιες από τις πιο κύριες εφαρμογές των αλγορίθμων μη εποπτευόμενης μάθησης εντοπίζονται:

- Στην «εξερεύνηση» των δεδομένων
- Στη τμηματοποίηση πελατών
- Στην ανάπτυξη συστημάτων συστάσεων
- Στην ανάπτυξη πιο προσποιημένων προωθητικών ενεργειών
- Στην επεξεργασία δεδομένων και οπτικοποίησή τους
- Στην εύρεση ανωμαλιών σε μεγάλα σύνολα δεδομένων

3.2.2.1 Συσταδοποίηση (Clustering)

Η συσταδποίηση είναι η πιο διαδεδομένη υποκατηγορία των αλγορίθμων μη εποπτευόμενης μάθησης της οποίας ο στόχος είναι η ομαδοποίηση των παρατηρήσεων βάσει ομοιοτήτων, η διαφορών που παρατηρούνται ανάμεσά τους. Υπάρχουν πολλαπλοί αλγόριθμοι οι οποίοι χρησιμοποιούνται ευρέως με στόχο την ομαδοποίηση των δεδομένων. Στην συνέχεια θα παρουσιαστούν οι πιο γνωστοί/βασικοί αλγόριθμοι της κατηγορίας αυτής, αφού πρώτα παρουσιαστούν οι μετρικές, βάσει των οποίων γίνεται η εύρεση των ομοιοτήτων μεταξύ των παρατηρήσεων.

3.2.2.1.1 Μετρικές Απόστασης - Συσχέτισης

Ανεξαρτήτως του αλγορίθμου συσταδοποίησης που θα χρησιμοποιηθεί θα πρέπει να επιλεγεί η καταλληλότερη μετρική απόστασης - συσχέτισης προκειμένου να υπάρχουν τα επιθυμητά αποτελέσματα. Στην ουσία η απόσταση δείχνει το βαθμό διαχωρισμού μεταξύ των στοιχείων των δεδομένων «παρατηρώντας» τα χαρακτηριστικά της κάθε παρατήρησης. Σύμφωνα με [48], όπου γίνεται μία συγκεντρωτική έρευνα για το ποια μετρική απόστασης είναι η «βέλτιστη», δεν υπάρχει κανόνας όσον αφορά την επιλογή της καταλληλότερης μεθόδου. Σημαντικό ρόλο στην επιλογή της μεθόδου υπολογισμού της απόστασης-συσχέτισης είναι ο τύπος των δεδομένων αλλά επίσης και το πρόβλημα συσταδοποίησης αυτό κάθε αυτό.

Ονομασία	Μαθηματικός Τύπος	Μεταβλητές-Επεξηγήσεις
Ευκλείδεια Απόσταση	$D_e(x_i, x_j) = \left(\sum_{l=1}^{d} \lvert x_{il} - x_{jl} \rvert^2\right)^{1/2}$	Απόσταση μεταξύ (x_i, x_j) παρατηρήσεων όπου τα (x_{il}, x_{jl}) συμβολίζουν την l διάσταση της παρατήρησης
Manhattan Απόσταση	$D_{Mn}(x_i, x_j) = \sum_{l=1}^{d} \lvert x_{il} - x_{jl} \rvert$	Άθροισμα των απολύτων τιμών των διαφορών των διαστάσεων των παρατηρήσεων (x_i, x_j). Μειωμένο επεξεργαστικό κόστος από την Ευκλείδεια
Μέτρο Συνημίτονου	$D_{cos}(x_i, x_j) = 1 - \left(\frac{x_i^T x_j}{\lvert x_i \rvert \lvert x_j \rvert}\right)$	Αποτελεί κυρίως ένα μέτρο συνοχής των ομάδων
Συσχέτιση κατά Pearson	$D_{Corr}(x_i, x_j) = 1 - S_{CR}(x_i, x_j)$	Η συσχέτιση κατά Pearson χρησιμοποιείται κυρίως για την μέτρηση της γραμμικής

	$S_{CR}(x_i, x_j) = \frac{\sum_{k=1}^{d}(m_{ik})(m_{jk})}{\sqrt{\sum_{k=1}^{d}(m_{ik})^2 \ \sum_{k=1}^{d}(m_{jk})^2}}$ $m_{ik} = x_{ik} - \overline{x_i}$ $m_{jk} = x_{jk} - \overline{x_j}$ $\overline{x_i} = \frac{1}{d}\sum_{k=1}^{d}(x_{ik})$ $\overline{x_j} = \frac{1}{d}\sum_{k=1}^{d}(x_{jk})$	εξάρτησης μεταξύ δύο σημειακών δεδομένων.

Πίνακας 2 Συγκεντρωτικός Πίνακας Μετρικών Απόστασης και Συναρτήσεων Συσχέτισης

3.2.2.1.2 Αλγόριθμος K-Μέσων (K-means)

Είναι πιθανότατα ο πιο γνωστός αλγόριθμος ομαδοποίησης παρατηρήσεων. Παρότι έχει ευρέως κατηγοριοποιηθεί σαν αλγόριθμος μη-εποπτευόμενης μάθησης το γεγονός ότι χρειάζεται σαν είσοδο τον αριθμό των κλάσεων που θα διερευνήσει δεν τον κάνει και τελείως «μη-εποπτευόμενο». Υπάρχουν παραλλαγές του αλγορίθμου που κυρίως διαφοροποιούνται ως προς τη μετρική που χρησιμοποιείται κάθε φορά, αλλά μεγαλύτερο ενδιαφέρον έχουν οι διαφοροποιήσεις που επιχειρούν να αποδεσμεύσουν τον αλγόριθμο από την είσοδο του αριθμού των κλάσεων. Χαρακτηριστική περίπτωση είναι του [49] όπου εντοπίζεται ο αριθμός των κλάσεων που διαχωρίζει σε όσο το δυνατό καλύτερο επίπεδο.

Η διαδικασία κατά την οποία ομαδοποιεί τα δεδομένα ο απλός αλγόριθμος κ-μέσων (με την είσοδο δηλαδή των επιθυμητών κλάσεων από το χρήστη) αποτελείται από τα εξής βήματα:

1. Δημιουργία τυχαίων κεντρικών σημείων ίσο με τον αριθμό των κλάσεων που δόθηκε στην είσοδο του αλγορίθμου
2. Υπολογισμός των αποστάσεων (συνηθέστερα Ευκλείδεια απόσταση) από τα τυχαία κεντρικά σημεία και ανάθεση των παρατηρήσεων στις ομάδες που είναι πιο κοντινές σε αυτές.
3. Μετά την ολοκλήρωση της ανάθεσης όλων των παρατηρήσεων στις ομάδες, υπολογίζονται εκ νέου τα κεντρικά σημεία σύμφωνα με το μέσο όρο των παρατηρήσεων της αντίστοιχης ομάδας
4. Η διαδικασία επαναλαμβάνει τα προηγούμενα βήματα έως ότου:
    a. Τα κεντρικά σημεία δεν μετακινηθούν περαιτέρω μεταξύ των επαναλήψεων
    b. Ενεργοποίηση συνθήκης τερματισμού όπως είναι ο μέγιστος αριθμός επαναλήψεων της διαδικασίας

Ο συγκεκριμένος αλγόριθμος έχει συγκεντρώσει μεγάλο ενδιαφέρον καθώς είναι απλός ως προς την υλοποίησή του και «φθηνός» από άποψη κόστους καθώς η πολυπλοκότητά του είναι γραμμική ως προς το χρόνο διεκπεραίωσης. Το σημείο το οποίο ο αλγόριθμος κ-μέσων παρουσιάζει σημαντική αδυναμία είναι εκείνο κατά το οποίο γίνεται τυχαία απόδοση των κεντρικών σημείων για την ομαδοποίηση των δεδομένων εισόδου. Τέλος πολύ σημαντική πληροφορία βάσει του [39] που θα πρέπει να σημειωθεί είναι η ευαισθησία που εντοπίζεται να χαρακτηρίζει τον αλγόριθμο αυτό στα θορυβώδη δεδομένα εισόδου και σε σύνολα δεδομένων που περιέχουν ακραίες τιμές.

3.2.2.1.3 Ιεραρχικές Μέθοδοι Συσταδοποίησης (Hierarchical)

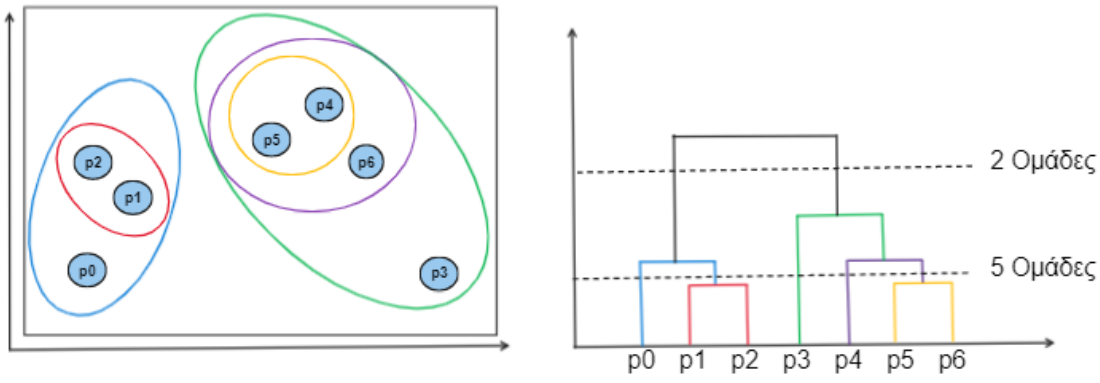
Οι ιεραρχικές μέθοδοι συσταδοποίησης είναι αποτέλεσμα αναδρομικών τμηματοποιήσεων του συνόλου δεδομένων, έχοντας σαν αποτέλεσμα μια δενδροειδή μορφή όπου σε κάθε επίπεδο ορίζεται ο αριθμός των συστάδων. Συνεπώς οι ιεραρχικοί αλγόριθμοι δεν χρειάζονται καμία είσοδο από τη πλευρά των επιστημόνων των δεδομένων (π.χ. αριθμό clusters) παρά μόνο τα ίδια τα δεδομένα. Χωρίζονται σε δύο κύριες κατηγορίες που διαφοροποιούνται κυρίως από το σημείο εκκίνησης του εκάστοτε αλγόριθμου και την κατεύθυνση δημιουργίας του δέντρου των πιθανών συστάδων [39].

Οι αλγόριθμοι της πρώτης κατηγορίας (Agglomerative Hierarchical Clustering) «θεωρούν» κάθε παρατήρηση ως μία ομάδα και έχουν σημείο εκκίνησης τα φύλλα του δέντρου (επίπεδο 0). Στην συνέχεια ομαδοποιούνται οι γειτονικές ομάδες διαμορφώνοντας με αυτό το τρόπο το επόμενο επίπεδο του δέντρου. Η διαδικασία αυτή συνεχίζεται μέχρι να μην είναι δυνατές άλλες ομαδοποιήσεις, δηλαδή αναφερόμαστε σε μία ομάδα που εμπεριέχει όλα τα δεδομένα, ή να φτάσουμε στο επιθυμητό για εμάς αριθμό ομάδων. Όταν η διαδικασία τερματιστεί καθώς δεν είναι δυνατές επιπλέον ομαδοποιήσεις τότε βρισκόμαστε στο μέγιστο δυνατό επίπεδο του δέντρου, δηλαδή στην ρίζα του δέντρου.

Η δεύτερη προσέγγιση (Divisive Hierarchical Clustering) ξεκινάει από την ρίζα του δέντρου όπου θεωρείται ότι όλα τα δεδομένα ανήκουν σε μία ομάδα και αναδρομικά διαιρούνται τα δεδομένα σε μικρότερες υπό-ομάδες. Με τη διαίρεση της προγονικής ομάδας δεδομένων αυξάνεται κατά ένα το επίπεδο του δέντρου. Η διαδικασία αυτή συνεχίζεται μέχρι να μην είναι εφικτή επιπλέον διάσπαση των ομάδων, έχουμε δηλαδή φτάσει στο σημείο όπου κάθε παρατήρηση αποτελεί και μία συστάδα. Η διαδικασία αυτή μπορεί να τερματιστεί νωρίτερα αν φτάσουμε στο επιθυμητό σχήμα και αριθμό συστάδων.

Η διαίρεση ή, η ομαδοποίηση των ομάδων κατώτερων επιπέδων γίνονται βάσει των μετρικών απόστασης-συσχετίσεων που είδαμε προηγουμένως ικανοποιώντας ταυτόχρονα κάποιο επιλεγμένο κριτήριο αξιολόγησης ή βελτιστοποίησης.

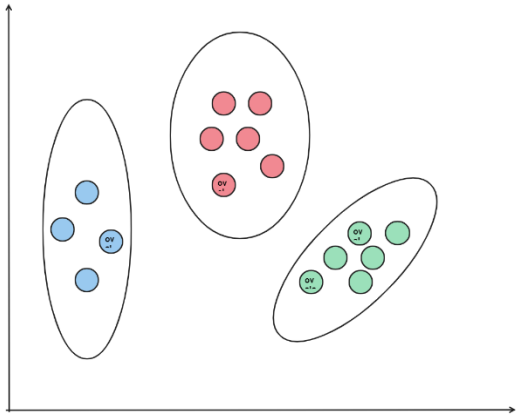
το Κυριότερο μειονέκτημα της μεθόδου αυτής είναι η αυξημένη πολυπλοκότητα καθώς ο χρόνος διεκπεραίωσης της διαδικασίας είναι κατ' ελάχιστον της τάξης $O(n^2)$, όπου n είναι ο αριθμός των παρατηρήσεων, δεν είναι δηλαδή γραμμικός. Συνεπώς η προσπάθεια τμηματοποίησης μεγάλου όγκου δεδομένων με τη χρήση ιεραρχικών αλγορίθμων αυξάνει σημαντικά το κόστος.



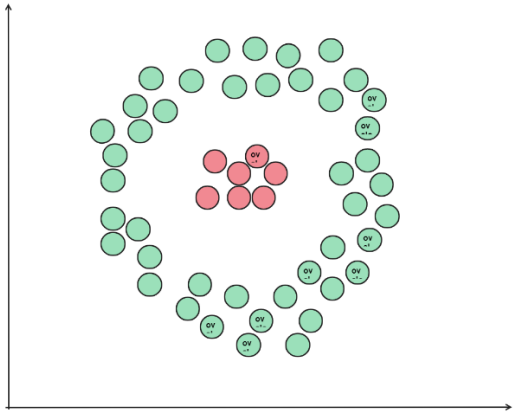
Εικόνα 17 Διαγραμματική Απεικόνιση Ιεραρχικών Αλγορίθμων Συσταδοποίησης

3.2.2.1.4 Συσταδοποίηση Βάσει Πυκνότητας (DB-SCAN)

Οι δύο προηγούμενοι αλγόριθμοι είναι εξαιρετικά δημοφιλείς καθώς είναι απλοί στην υλοποίηση, εύκολα κατανοητοί και παραμετροποίησιμοι στις ανάγκες της κάθε περίπτωσης αλλά επίσης είναι και πολύ αποτελεσματικοί. Ο αλγόριθμος κ-μέσων και ο ιεραρχικός αλγόριθμος συσταδοποίησης αν και είναι οι αλγόριθμοι που συγκεντρώνουν όλα τα θετικά στοιχεία που επιθυμεί ένας αναλυτής δεδομένων υπάρχει μία σημαντική αδυναμία. Η αποτελεσματικότητά τους αυτή στηρίζεται στη δομή των δεδομένων και στο γεγονός ότι τα σημεία των κλάσεων δεν μπερδεύονται μεταξύ άλλων. Στις παρακάτω εικόνες γίνεται μία σχηματική απεικόνιση του προβλήματος που αντιμετωπίζουν οι προηγούμενοι αλγόριθμοι.



Εικόνα 18 Σχηματική Απεικόνιση Ιδανικών Συνθηκών Συσταδοποιησης για K-means και Ιεραρχικού Αλγόριθμου



Εικόνα 19 Σχηματική Απεικόνιση Μη-Ιδανικών Συνθηκών Συσταδοποίησης για K-means και Ιεραρχικού Αλγόριθμου

Για διατάξεις σημείων όπως αυτή της εικόνας 19 οι προαναφερθέντες αλγόριθμοι θα αδυνατούσαν να διαχωρίσουν αποτελεσματικά τις δύο κλάσεις. Για αυτές τις περιπτώσεις και όχι μόνο προτιμάται η χρήση αλγορίθμων που η ομαδοποίηση των σημείων και κατ' επέκταση των κλάσεων, γίνεται με βάση την πυκνότητα εμφάνισης των σημείων.

Όπως χαρακτηριστικά αναφέρεται στο άρθρο [50] η ιδέα «κλειδί» της συσταδοποίησης βάσει πυκνότητας είναι ότι εντός μίας ορισμένης ακτίνας πρέπει να περιέχει τουλάχιστον ένα αριθμό από παρατηρήσεις. Δηλαδή από άποψης εισόδου από πλευράς χρήστη ο αλγόριθμος περιμένει μία τιμή κατωφλίου που ορίζει από ποιον αριθμό παρατηρήσεων και πάνω μπορεί να θεωρηθεί μία ομάδα κλάση, και ποια η ελάχιστη απόσταση που πρέπει να έχουν δύο σημεία μεταξύ τους. Η παρακάτω μαθηματική σχέση περιγράφει πως σχηματίζεται η γειτονιά ενός σημείου p, με τη χρησιμοποίηση των τιμών κατωφλίων που εισάγει ο χρήστης με τις παρατηρήσεις των συνόλου δεδομένων.

$$N_{Eps} = \{q \in D \mid dist(p,q) < Eps\}$$

- Eps: Η ακτίνα που έχει ορισθεί από το χρήστη
- D: Σύνολο παρατηρήσεων

Αν η γειτονιά N_{Eps}, του σημείου p περιέχει τουλάχιστον τον αριθμό των σημείων που έχει δοθεί σαν είσοδος κατά την εκκίνηση του αλγορίθμου τότε το σημείο αυτό χαρακτηρίζεται ως πυρήνας

$$N_{Eps}(P) > MinPts$$

Η διαδικασία επαναλαμβάνεται για κάθε παρατήρηση του συνόλου δεδομένων η οποία επαναληπτική διαδικασία μπορεί να οδηγήσει σε ενσωμάτωση μικρότερων γειτονιών. Ο αλγόριθμος τερματίζεται όταν δεν μπορεί να σχηματιστεί καινούρια γειτονιά από τις παρατηρήσεις.

Η ομαδοποίηση όπως όλες οι μέθοδοι συγκεντρώνουν θετικά και αρνητικά στοιχεία από τα οποία τα πιο σημαντικά παρουσιάζονται στην συνέχεια.

Πλεονεκτήματα:

- Δεν απαιτείται ο εκ των προτέρων προσδιορισμός του αριθμού των κλάσεων
- Η ομαδοποίηση είναι ανεξάρτητη της διάταξης των παρατηρήσεων
- Αντιμετωπίζει αποτελεσματικά το φαινόμενο του θορύβου στα δεδομένα
- Δεν επηρεάζεται από ακραίες τιμές (outliers)

Μειονεκτήματα:

- Απαιτεί τη γνώση του πεδίου μελέτης για τον καθορισμό των αποστάσεων και τον αριθμό των γειτονικών παρατηρήσεων
- Δεν μπορεί να γίνει σωστή ομαδοποίηση των δεδομένων που εμφανίζουν ισχυρές διαφορές πυκνότητας καθώς δεν μπορεί εύκολα να βρεθούν οι τιμές κατωφλιών που απαιτεί για είσοδο η μέθοδος

3.2.3 Μέτρα αξιολόγησης

Τα μέτρα αξιολόγησης χρησιμοποιούνται αφού εκπαιδευτεί ένα μοντέλο και βάσει αυτών είμαστε σε θέση να συγκρίνουμε διαφορετικούς ταξινομητές. Τα βασικότερα μέτρα τα οποία χρησιμοποιούνται για την αξιολόγηση της αποτελεσματικότητας ενός αλγορίθμου είναι η συνολική ακρίβεια ή αλλιώς accuracy, η ακρίβεια precision, η ανάκληση recall και το f1-score.

Η μετρική αξιολόγησης της συνολικής ακρίβειας είναι η πιο απλή στη κατανόηση μετρική καθώς εκφράζεται από τις σωστά ταξινομημένες παρατηρήσεις προς το σύνολο των παρατηρήσεων. Υψηλή συνολική ακρίβεια δεν είναι σε όλες τις περιπτώσεις ικανή για την αξιολόγηση ενός αλγορίθμου. Για παράδειγμα όταν το μοντέλο εκπαιδεύεται με σύνολα δεδομένων τα οποία είναι ανομοιόμορφα ως προς τη πυκνότητα των κλάσεων της μεταβλητής στόχου. Σε αυτή τη περίπτωση μπορεί το μοντέλο να έχει 99% συνολική ακρίβεια αλλά να μην είναι σε θέση να αναγνωρίσει ούτε μία περίπτωση της κλάσης μειοψηφίας. Συνεχίζοντας παρουσιάζεται ο μαθηματικός τύπος της συνολικής ακρίβειας.

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

a. TP: Ορθά θετική Εκτίμηση (True Positive)
b. TN: Ορθά Αρνητική Εκτίμηση (True Negative)
c. FP: Λανθασμένα Θετική Εκτίμηση (False Positive)
d. FN: Λανθασμένα Αρνητική Εκτίμηση (False Negative)

Συνέχεια έχει η μετρική της ακρίβειας (Precision) η οποία ερμηνεύεται ως ο λόγος των ορθά θετικών εκτιμήσεων προς το συνολικό αριθμό των θετικών εκτιμήσεων. Είναι χρήσιμη αυτή η μετρική όταν μας ενδιαφέρει περισσότερο ο αριθμός των λανθασμένων θετικών εκτιμήσεων έναντι των λανθασμένων αρνητικών. Περιπτώσεις που μπορεί αυτή η μετρική να βρίσκει εφαρμογή είναι στα συστήματα συστάσεων όπου μία λανθασμένη εκτίμηση μπορεί να προκαλέσει δυσαρέσκεια στους πελάτες. Μπορεί να υπάρχει δηλαδή σημαντικό κόστος από τη λανθασμένη θετική εκτίμηση. Αντίθετα δεν υπάρχει το ίδιο κόστος αν εκτιμηθεί μια παρατήρηση ως αρνητική ενώ ήταν θετική, καθώς το σύστημα δεν θα προβεί σε κάποια ενέργεια σε αυτή τη περίπτωση.

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

Η ανάκληση (recall) εξηγεί πόσες πραγματικά θετικές παρατηρήσεις κατάφερε το μοντέλο να ταξινομήσει ορθά. Είναι δηλαδή ο λόγος των ορθά θετικών εκτιμήσεων προς τον αριθμό των πραγματικά θετικών παρατηρήσεων. Σε αντίθεση με τη precision, που είδαμε νωρίτερα, η ανάκληση (recall) είναι χρήσιμη όταν μας ενδιαφέρει η λανθασμένη αρνητική εκτίμηση περισσότερο από την ορθά αρνητική. Αυτή η μετρική έχει εφαρμογή παραδείγματος χάρει στην ιατρική καθώς μία λανθασμένη θετική διάγνωση δεν αποτελεί σημαντικό πρόβλημα αλλά μία περίπτωση η οποία είναι πραγματικά θετική δεν πρέπει να περνάει απαρατήρητη. Ο μαθηματικός τύπος της recall είναι ο εξής:

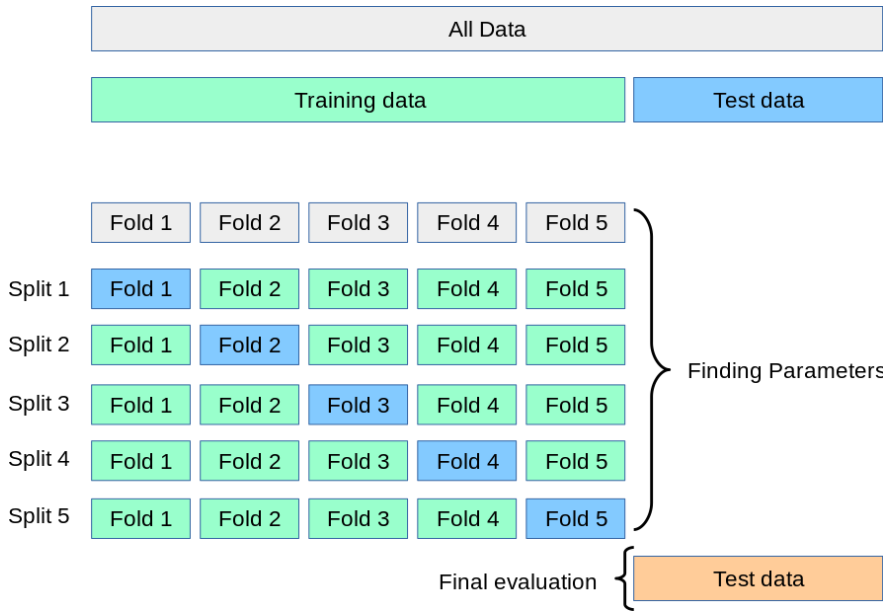
$$Recal = \frac{TP}{TP + FN}$$

Τέλος η F1-score είναι ένας συνδυασμός των δύο τελευταίων μετρικών αξιολογήσεων και χρησιμοποιείται όταν οι λανθασμένες αρνητικές και θετικές εκτιμήσεις έχουν το ίδιο κόστος. Είναι αρκετά ανθεκτική όταν προστίθενται επιπλέον παρατηρήσεις και δεν αλλοιώνεται το αποτέλεσμα. Ο λόγος της F1-score είναι:

$$F1 - Score = \frac{2}{\frac{1}{Recall} + \frac{1}{Precision}}$$

3.2.4 Διασταυρούμενη Επικύρωση Κ-Φορών (K-Fold Cross Validation)

Η μέθοδος K-fold cross Validation ή στα ελληνικά «διασταυρούμενη επικύρωση k φορών» είναι ένας στατιστικός τρόπος επικύρωσης των αποτελεσμάτων των αλγορίθμων μηχανικής μάθησης. Αρχικά η μέθοδος χωρίζει το σύνολο δεδομένων σε Κ μέρη όπως της εικόνας 20 όπου το αρχικό σύνολο δεδομένων διαχωρίζεται σε σύνολα εκπαίδευσης και δοκιμής. Στη συνέχεια το σύνολο εκπαίδευσης χωρίζεται σε 5 ίσα μέρη και ο αλγόριθμος εκπαιδεύεται διαδοχικά σε 5 διαφορετικά φαινομενικά σύνολα καθώς ένα από τα 5 τμήματα που διαχωρίστηκαν χρησιμοποιείται ως σύνολο δοκιμής. Εφόσον ολοκληρωθεί η επαναληπτική διαδικασία το μοντέλο αξιολογείται ως προς την ακρίβεια των εκτιμήσεων σε σύνολο δοκιμής που δεν έχει «δει» προηγουμένως κατά τη διαδικασία της εκπαίδευσης.



Εικόνα 20 Απεικόνιση Ενδεικτικής Διάσπασης Δεδομένων Κατά την Εκτέλεση της Διαδικασίας Κ(5)-Fold Cross Validation
Πηγή: [51]

Σύμφωνα με το [52] η συνηθέστερη επιλογή διαχωρισμού του συνόλου εκπαίδευσης είναι η διάσπαση σε 10 τμήματα (10-Fold Cross Validation). Υπάρχουν πολλές επιπρόσθετες λειτουργίες του αλγορίθμου επικύρωσης και μερικές από αυτές είναι:

- Hold out Validation: Για την αποφυγή της υπέρ-εκπαίδευσης η τελική αξιολόγηση γίνεται σε σύνολο δεδομένων που δεν έχει εκπαιδευτεί σε προγενέστερο επίπεδο.
- Επαναλαμβανόμενη διασταυρούμενη επικύρωση: Με την μέθοδο αυτή έχουμε καλύτερη άποψη για την ακρίβεια του εκτιμητή καθώς η διαδικασία που περιεγράφηκε παραπάνω επαναλαμβάνεται παράγοντας σημαντικά περισσότερα αποτελέσματα.
- Διασταυρούμενη επικύρωση με Παράλειψη ενός (Leave-One-Out) διασταυρούμενη επικύρωση: η δοκιμή σε κάθε επανάληψη γίνεται με μία παρατήρηση και το αποτέλεσμα του αλγορίθμου ως προς την ακρίβεια πηγάζει από τις Κ εκτιμήσεις.
- Στρατηγική Διασταυρούμενη επικύρωση κ-φορών: Χρησιμοποιείται κυρίως για την διατήρηση των ιδιαιτεροτήτων του συνόλου δεδομένων για την παραγωγή αποτελεσμάτων που η τυχαία επιλογή ενδέχεται να μην συμπεριλάβει.

4. Σύνολα Δεδομένων (Datasets)

4.1 Πειραματικό Περιβάλλον

Η παρακάτω πειραματική διαδικασία πραγματοποιήθηκε με την χρήση της γλώσσας python η οποία είναι μία γλώσσα προγραμματισμού ευρείας χρήσης όπου το εύκολο συντακτικό της την έχει ορίσει ως την πιο διαδεδομένη γλώσσα για την ανάλυση δεδομένων και όχι μόνο. Όλο και περισσότεροι επαγγελματίες την χρησιμοποιούν καθώς είναι αποτελεσματική ως προς το χρόνο εκμάθησης, έχοντας σαν αποτέλεσμα την δημιουργία τεράστιου πλήθους βιβλιοθηκών και εργαλείων. Πέρα του ποικιλόμορφου «οπλοστασίου» της python αυτό που την έχει αναδείξει είναι η ενεργή κοινότητα που με ζήλο θα επιχειρήσει να διορθώσει όποιο πρόβλημα παρουσιαστεί. Επανερχόμενοι στο πειραματικό περιβάλλον κατ' επέκταση της python για την εκπόνηση της παρούσας εργασίας χρησιμοποιήθηκαν οι βιβλιοθήκες pandas και numpy, για διαχείριση και επεξεργασία των δεδομένων, η seaborn και matplotlib για την οπτικοποίηση των αποτελεσμάτων μέσω γραφημάτων και μία πληθώρα μεθόδων της sklearn που παρέχει μοντέλα μηχανικής μάθησης, τα οποία με τις ελάχιστες γραμμές κώδικα είναι διαθέσιμα και μπορεί ο οποιοσδήποτε να τα χρησιμοποιήσει.

4.2 Περιγραφή Συνόλου Δεδομένων

Το σύνολο δεδομένων που χρησιμοποιήθηκε σε αυτή τη πειραματική διερεύνηση συντάχθηκε και αναλύθηκε στα πλαίσια έρευνας που διεξήχθη από το πανεπιστήμιο Libre de Bruxelles (ULB) πάνω στην εξόρυξη μεγάλων δεδομένων και την ανίχνευση απάτης.

Το συγκεκριμένο σύνολο δεδομένων αποτελείται από αρχεία συναλλαγών πιστωτικών καρτών που πραγματοποιήθηκαν το Σεπτέμβριο του 2013 από χρήστες εντός της Ευρωπαϊκής Ένωσης. Ειδικότερα οι συναλλαγές που διατηρούνται στο σύνολο είναι διαστήματος δύο (2) ημερών και το συνολικό πλήθος τους είναι 284807. Οι συναλλαγές που έχουν χαρακτηριστεί ως απατηλές αποτελούν μόλις το 0,172% του συνόλου δεδομένων, αφού ανέρχονται στις 492 συνολικά. Συνεπώς γίνεται λόγος για ένα εξαιρετικά ανομοιόμορφο σύνολο όπου η προσπάθεια κατηγοριοποίησης των παρατηρήσεων χωρίς την εξισορρόπηση των κατανομών των κλάσεων, της μεταβλητής στόχου, να είναι μάταιη.

Λόγω της ευαισθησίας τέτοιου είδους δεδομένων αλλά και της νομοθεσίας περί διαφύλαξης των προσωπικών δεδομένων, που στη συγκεκριμένη περίπτωση χαρακτηρίζονται ως απόρρητα, τα χαρακτηριστικά των παρατηρήσεων έχουν μετασχηματιστεί με τη βοήθεια της ανάλυσης σε κύριες συνιστώσες (PCA) και δεν είναι γνωστό τίποτα άλλο πέρα από αυτό. Τα μόνα χαρακτηριστικά που δεν έχουν «αλλοιωθεί» από την PCA είναι αυτά του χρόνου και του ύψους του ποσού της εκάστοτε συναλλαγής. Οι τιμές του χαρακτηριστικού «Time» είναι σε δευτερόλεπτα, τα οποία σηματοδοτούν το πότε έγινε η συγκεκριμένη συναλλαγή σε σχέση με την πρώτη παρατήρηση του συνόλου δεδομένων ενώ στο χαρακτηριστικό «Amount» είναι διαθέσιμη η πληροφορία σχετικά με το ύψος της συναλλαγής που πραγματοποιήθηκε. Τέλος υπάρχει το χαρακτηριστικό «Class» που παίρνει τιμές μηδέν ή ένα (0 ή 1) και χρησιμοποιείται για το χαρακτηρισμό της συναλλαγής σαν απατηλή (εάν έχει αποδοθεί η τιμή 1) ή σαν νόμιμη (εάν έχει την τιμή 0).

Όπως γίνεται αντιληπτό με μία πρώτη ανάγνωση των δεδομένων σε συνεργασία με τον συγκεντρωτικό πίνακα 3 οι παρατηρήσεις χαρακτηρίζονται από τριάντα (30) μεταβλητές και μιας επιπλέον μεταβλητής στόχου (Class). Όλα τα χαρακτηριστικά φαίνεται να είναι πλήρως συμπληρωμένα δεν υπάρχουν δηλαδή διαλείπουσες τιμές και όλες οι τιμές των χαρακτηριστικών είναι αριθμητικές.

Ονομασία Χαρακτηριστικού	Ελλιπείς Τιμές	Σύνολο Εγγραφών	Τύπος Δεδομένων
Time	0	284807	Float64
V1	0	284807	Float64
V2	0	284807	Float64
V3	0	284807	Float64
V4	0	284807	Float64
V5	0	284807	Float64
V6	0	284807	Float64
V7	0	284807	Float64
V8	0	284807	Float64
V9	0	284807	Float64
V10	0	284807	Float64
V11	0	284807	Float64
V12	0	284807	Float64
V13	0	284807	Float64
V14	0	284807	Float64
V15	0	284807	Float64
V16	0	284807	Float64
V17	0	284807	Float64
V18	0	284807	Float64
V19	0	284807	Float64
V20	0	284807	Float64
V21	0	284807	Float64
V22	0	284807	Float64
V23	0	284807	Float64
V24	0	284807	Float64
V25	0	284807	Float64
V26	0	284807	Float64
V27	0	284807	Float64
V28	0	284807	Float64
Amount	0	284807	Float64
Class	0	284807	Int64

Πίνακας 3 Επεξηγηματικός Πίνακας Χαρακτηριστικών Συνόλου Δεδομένων

4.3 Προ-επεξεργασία Δεδομένων

Όπως προαναφέρθηκε δεν υπάρχουν ελλιπείς τιμές στο σύνολο των δεδομένων συνεπώς δεν θα χρειαστεί να προβούμε σε κάποια διορθωτική ενέργεια. Δυστυχώς παρατηρούνται όμως διπλότυπες παρατηρήσεις όπου θα πρέπει να αφαιρεθούν καθώς παίζουν το ρόλο του θορύβου στα δεδομένα μας. Επίσης είναι ατυχές το γεγονός αυτό καθώς υπάρχουν και 19 παρατηρήσεις από την κλάση των απατηλών συναλλαγών πράγμα που σημαίνει ότι η κλάση της μειοψηφίας μειώνεται περεταίρω. Πιο συγκεκριμένα από τις 284807 συναλλαγές παρατηρούνται 1081 διπλότυπες παρατηρήσεις οι οποίες αφαιρούνται από το σύνολο έχοντας σαν αποτέλεσμα την αλλαγή του πλήθους των συναλλαγών σε αυτό των 283726. Αξίζει να σημειωθεί ότι η επίπτωση μετά την εκκαθάριση των διπλών παρατηρήσεων είναι σημαντική καθώς η κλάση μειοψηφίας μειώθηκε κατά 3.861% ενώ αντίθετα η πλειοψηφική κλάση μειώθηκε κατά 0.3728%. Έτσι λοιπόν η διαφορά του πλήθους μεταξύ των κλάσεων αυξήθηκε.

	Αρχικό Σύνολο Δεδομένων		Διπλότυπες Παρατηρήσεις		Νέο Σύνολο Δεδομένων	
	Αριθμός Παρατηρήσεων	Ποσοστό (%)	Αριθμός Παρατηρήσεων	Ποσοστό (%)	Αριθμός Παρατηρήσεων	Ποσοστό (%)
Νόμιμη	284315	99.827 %	1062	0.3728 %	283253	99.8332 %
Απάτη	492	0.172 %	19	3.861 %	473	0.1667 %

Πίνακας 4 Επισκόπηση Πλήθους Διπλότυπων Παρατηρήσεων

Παρά το γεγονός ότι οι απατηλές συναλλαγές είναι κάτω του 0.2%, το συνολικό ύψος της ζημίας ανέρχεται σε κάτι περισσότερο από τα 58.000€ έχοντας σαν μέση τιμή αυτή των 123€ ανά απατηλή συναλλαγή. Όπως φαίνεται και στον παρακάτω συγκριτικό πίνακα ύψους συναλλαγών, το 0.233% του συνολικού ποσού των συναλλαγών του δείγματος είναι απάτες. Ενώ μπορεί να μην φαίνεται πολύ μεγάλο ποσοστό θα πρέπει να θυμηθούμε ότι το δείγμα αυτό περιέχει δεδομένα δύο μόλις ημερών, που σημαίνει ότι περίπου 30.000€ «χάνονται» κάθε μέρα λόγω κακόβουλων ενεργειών και αυτά μόνο μέσω πιστωτικών καρτών. Είναι πολύ σημαντικό λοιπόν να βρεθούν οι μηχανισμοί αυτοί που θα ανιχνεύουν αποτελεσματικά, κατά προτίμηση σε πραγματικό χρόνο, της κακόβουλες ενέργειες.

	sum	**std**	**mean**
	Amount	**Amount**	**Amount**
Class			
0	25043410	250.379	88.41357
1	58591.39	260.211	123.8719
All	25102002	250.399	88.47269

Πίνακας 5 Συγκριτικός Πίνακας Ύψους Συναλλαγών Ανά Κατηγορία

4.4 Διερεύνηση Δεδομένων

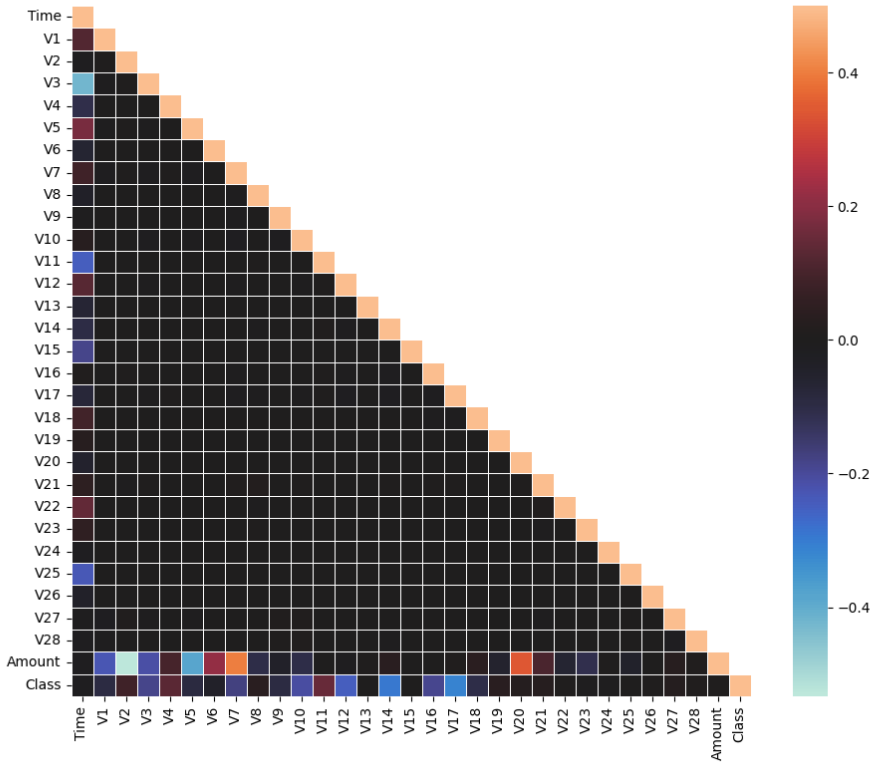
Για την καλύτερη κατανόηση των δεδομένων και την σχέση τους μεταξύ τους θα ακολουθήσει πίνακας (πίνακας 6) περιγραφικής στατιστικής ομαδοποιώντας και συγκρίνοντας τα πεδία των τιμών χαρακτηριστικών του δείγματος. Εκ πρώτης όψεως φαίνεται ότι οι μέση τιμή των μεταβλητών, εκτός του χρόνου και του ύψους του ποσού των συναλλαγών, κυμαίνεται από -0.004134 έως 0.00591. Αντίστοιχα η μέση τιμή του ποσού είναι 88.4726 που σημαίνει ότι θα πρέπει να κανονικοποιηθεί προκειμένου να μην επηρεάσει τα αποτελέσματα των ταξινομητών η μεγάλη αυτή διαφορά μεταξύ των χαρακτηριστικών όπως προαναφέρθηκε και προηγουμένως στην αντίστοιχη ενότητα.

	count	mean	std	min	25%	50%	75%	max
Time	283726	94811.08	47481.05	0	54204.75	84692.5	139298	172792
V1	283726	0.005917	1.948026	-56.4075	-0.91595	0.020384	1.316068	2.45493
V2	283726	-0.00413	1.646703	-72.7157	-0.60032	0.063949	0.800283	22.05773
V3	283726	0.001613	1.508682	-48.3256	-0.88968	0.179963	1.02696	9.382558
V4	283726	-0.00297	1.414184	-5.68317	-0.85013	-0.02225	0.739647	16.87534
V5	283726	0.001828	1.377008	-113.743	-0.68983	-0.05347	0.612218	34.80167
V6	283726	-0.00114	1.331931	-26.1605	-0.76903	-0.27517	0.396792	73.30163
V7	283726	0.001801	1.227664	-43.5572	-0.55251	0.040859	0.570474	120.5895
V8	283726	-0.00085	1.179054	-73.2167	-0.20883	0.021898	0.325704	20.00721
V9	283726	-0.0016	1.095492	-13.4341	-0.64422	-0.0526	0.595977	15.59499
V10	283726	-0.00144	1.076407	-24.5883	-0.53558	-0.09324	0.453619	23.74514
V11	283726	0.000202	1.01872	-4.79747	-0.76165	-0.03231	0.739579	12.01891
V12	283726	-0.00071	0.994674	-18.6837	-0.4062	0.139072	0.616976	7.848392
V13	283726	0.000603	0.99543	-5.79188	-0.64786	-0.01293	0.663178	7.126883
V14	283726	0.000252	0.952215	-19.2143	-0.42573	0.050209	0.492336	10.52677
V15	283726	0.001043	0.914894	-4.49894	-0.58145	0.049299	0.650104	8.877742
V16	283726	0.001162	0.873696	-14.1299	-0.46686	0.067119	0.523512	17.31511
V17	283726	0.00017	0.842507	-25.1628	-0.48393	-0.06587	0.398972	9.253526
V18	283726	0.001515	0.837378	-9.49875	-0.49801	-0.00214	0.501956	5.041069
V19	283726	-0.00026	0.813379	-7.21353	-0.45629	0.003367	0.458508	5.591971
V20	283726	0.000187	0.769984	-54.4977	-0.21147	-0.06235	0.133207	39.4209
V21	283726	-0.00037	0.723909	-34.8304	-0.2283	-0.02944	0.186194	27.20284
V22	283726	-1.5E-05	0.72455	-10.9331	-0.5427	0.006675	0.528245	10.50309
V23	283726	0.000198	0.623702	-44.8077	-0.1617	-0.01116	0.147748	22.52841
V24	283726	0.000214	0.605627	-2.83663	-0.35445	0.041016	0.439738	4.584549
V25	283726	-0.00023	0.52122	-10.2954	-0.31749	0.016278	0.350667	7.519589
V26	283726	0.000149	0.482053	-2.60455	-0.32676	-0.05217	0.240261	3.517346
V27	283726	0.001763	0.395744	-22.5657	-0.07064	0.001479	0.091208	31.6122
V28	283726	0.000547	0.328027	-15.4301	-0.05282	0.011288	0.078276	33.84781
Amount	283726	88.47269	250.3994	0	5.6	22	77.51	25691.16
Class	283726	0.001667	0.040796	0	0	0	0	1

Πίνακας 6 Πίνακας Περιγραφής Πεδίων Τιμών των Χαρακτηριστικών

Παρά το γεγονός ότι στα δεδομένα έχει εφαρμοσθεί ανάλυση σε κύριες συνιστώσες (PCA), που σαν αποτέλεσμα έχει τη μείωση διαστάσεων του αρχικού συνόλου, θα πρέπει να γίνει εκ νέου διερεύνηση των συσχετίσεων μεταξύ των μεταβλητών με στόχο, το αν κάποιες από αυτές είναι υψηλά συσχετιζόμενες να διατηρήσουμε ένα υποσύνολο αυτών. Με αυτή την ενέργεια θα διατηρηθεί η πληροφορία αλλά παράλληλα απλοποιείται σημαντικά το σύνολο δεδομένων κάνοντας τις εργασίες των αλγορίθμων μηχανικής μάθησης πιο αποδοτικές ως προς το χρόνο διεκπεραίωσης χωρίς να έχουμε σημαντικές αποκλίσεις στην ακρίβεια των κατηγοριοποιήσεων των μοντέλων.

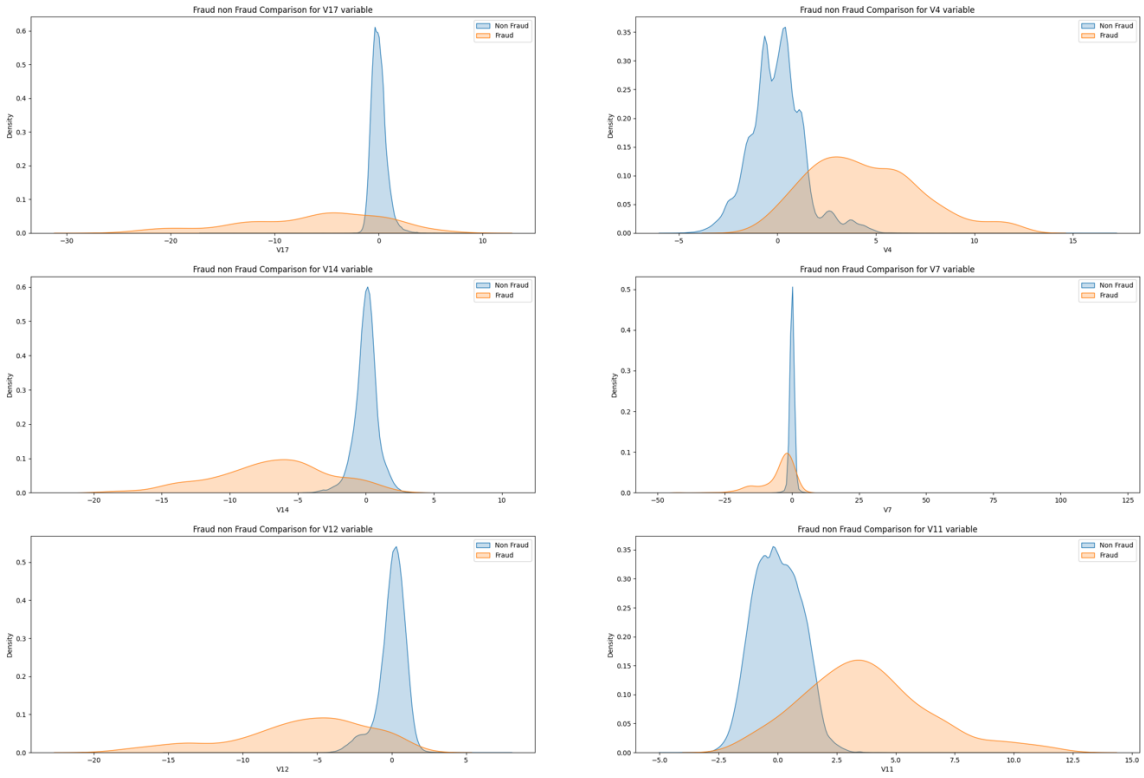
Έτσι λοιπόν στην παρακάτω εικόνα (εικόνα 21) παρουσιάζεται ένα διάγραμμα συσχετίσεων μεταξύ των μεταβλητών του συνόλου δεδομένων.



Εικόνα 21 Διαγραμματική Απεικόνιση Συσχετίσεων Μεταβλητών

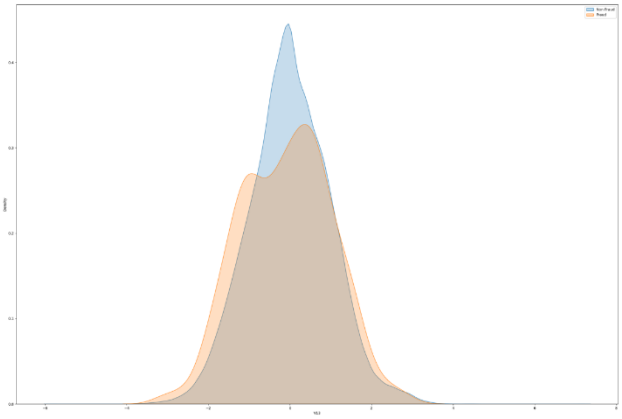
Σκοπός σε αυτό το στάδιο έχοντας παρουσιάσει τις συσχετίσεις των μεταβλητών είναι η εξαγωγή συμπερασμάτων που συνδέουν τις ανεξάρτητες μεταβλητές με τη μεταβλητή στόχο που στην δικιά μας περίπτωση είναι η μεταβλητή Class. Οι τρεις μεταβλητές με την μεγαλύτερη συσχέτιση είναι οι V2, V4, V11 με τιμές 0.084624, 0.129326, 0.149067 αντίστοιχα. Βέβαια καλύτερη εικόνα μας δίνουν οι μεταβλητές με τη μικρότερη συσχέτιση που είναι οι V17, V14 και V12 με τις αντίστοιχες τιμές -0.313498, -0.293375 και -0.250711. Αντλούμε σημαντική πληροφορία από τις τελευταίες μεταβλητές καθώς σε απόλυτη τιμή απέχουν περισσότερο από το μηδέν έτσι αυτές οι τρεις μεταβλητές είναι πολύ σημαντικές για την ταξινόμηση μιας συναλλαγής ως νόμιμης ή απάτης.

Ενδιαφέρον έχει να δούμε τις κατανομές των τιμών των μεταβλητών με τη μικρότερη αλλά και με τη μεγαλύτερη συσχέτιση συγκρίνοντάς τες μεταξύ τους αλλά να γίνει και μία οπτικοποίηση των διαφορών και των ομοιοτήτων των δύο κλάσεων ανά μεταβλητή.



Εικόνα 22 Συγκριτική Απεικόνιση Υψηλά & Χαμηλά Συ-σχετιζόμενων Μεταβλητών

Είναι απόλυτα σαφές από την εικόνα 22 ότι το μοντέλο ταξινόμησης που θα επιλεγεί στο τέλος της εργασίας αυτής θα «δώσει ιδιαίτερη» σημασία σε μεταβλητές σαν αυτές της εικόνας, καθώς παρατηρείται σημαντική διαφορά των κατανομών των νόμιμων και των απατηλών συναλλαγών. Αντίθετα μεταβλητές σαν και την V13 της εικόνας 23 δεν παρέχει σημαντική διαφοροποίηση μεταξύ των κατανομών των δύο κλάσεων, συνεπώς είναι μηδαμινή η συνεισφορά της μεταβλητής αυτής για την ανίχνευση μιας απατηλής συναλλαγής. Θεωρητικά θα μπορούσαμε να αφαιρέσουμε εξ' ολοκλήρου μεταβλητές με παρόμοιες κατανομές χωρίς να έχουμε σημαντικές απώλειες στην αποδοτικότητα του μοντέλου μας. Η μη χρησιμοποίηση των μεταβλητών αυτών βοηθάει επίσης και στην εξάλειψη του φαινομένου της υπέρ-προσαρμογής του μοντέλου στο σύνολο εκπαίδευσης καθώς γενικεύεται σημαντικά.



Εικόνα 23 Κατανομή Απατηλών/Νόμιμων Συναλλαγών της Μεταβλητής V13

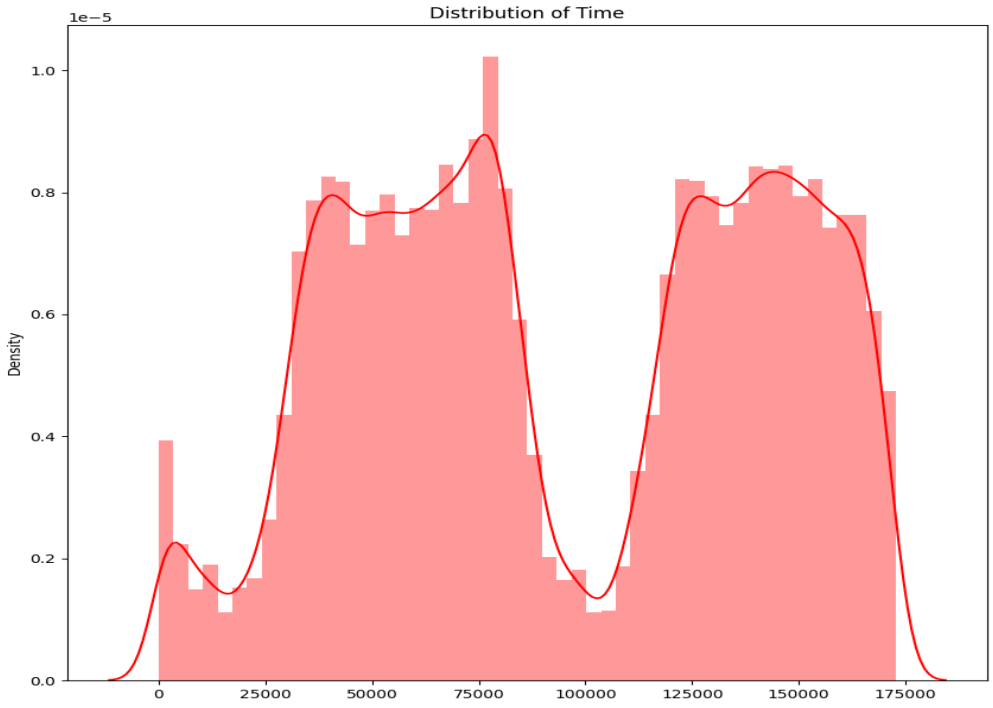
Θα ήταν ιδιαίτερα ενδιαφέρον να γνωρίζαμε την περισσότερα για τις μεταβλητές αυτές αλλά λόγω του μετασχηματισμού που έχει πραγματοποιηθεί είναι αδύνατον. Μπορούμε μόνο να κάνουμε εκτιμήσεις με σκοπό την διαστασιοποίηση των ευρημάτων αυτών των εικόνων 22 και 23. Για παράδειγμα η μεταβλητή V13 θα μπορούσε να είναι ο τρόπος με τον οποίο έγινε η συναλλαγή και μπορεί να παίρνει τιμές μηδέν ή ένα αν ήταν παραδείγματος χάρει ανέπαφη συναλλαγή (χωρίς την εισαγωγή PIN στο τερματικό χρέωσης). Δηλαδή είναι ένα πεδίο που είναι εξαιρετικά δημοφιλές για την πραγματοποίηση νόμιμων συναλλαγών αλλά και απατηλών καθώς δεν χρειάζεται τίποτα άλλο για την πραγματοποίηση αυτών πέρα από την φυσική κάρτα. Από την άλλη η μεταβλητή V17 που διαχωρίζει ικανοποιητικά τις συναλλαγές σε νόμιμες και απατηλές θα μπορούσε να είναι η ηλικία του κατόχου της κάρτας που έγινε η χρέωση. Είναι πιθανότερο άτομα μεγαλύτερης ηλικίας να πέφτουν θύματα απάτης μέσω κλοπής της κάρτας τους καθώς δεν γίνεται αντιληπτό σε σύντομο χρονικό διάστημα η απώλειά της. Έτσι δεν γίνονται άμεσα οι απαραίτητες ενέργειες φραγής των συναλλαγών μέσω της κάρτας, με αποτέλεσμα να γίνονται χρεώσεις χωρίς να το γνωρίζουν.

Στη συνέχεια προχωράει η διερεύνηση στις μεταβλητές που δεν αλλοιώθηκαν από τον PCA μετασχηματισμό που δεν είναι άλλες από αυτή του χρόνου και αυτή του ύψους του ποσού των συναλλαγών. Στην εικόνα 24 με πράσινο χρώμα απεικονίζεται συγκέντρωση των παρατηρήσεων για συγκεκριμένο ύψος ποσού συναλλαγής. Βλέπουμε ότι η συντριπτική πλειοψηφία εντοπίζεται μεταξύ των 0€ και των 2,500€. Το γεγονός ότι είναι τόσο πολύ μετατοπιζόμενο το ιστόγραμμα προς τον αριστερό κάθετο άξονα του γραφήματος μας δίνει μία εικόνα ότι υπάρχουν ακραίες τιμές στην μεταβλητή αυτή. Αυτή τη διαίσθηση έρχεται να μας επιβεβαιώσει η προτελευταία γραμμή του πίνακα 6 που φαίνεται ότι η μέγιστη τιμή των ποσών των συναλλαγών είναι 25.691€ ενώ μέση είναι 88,5€ περίπου και η τιμή της τυπικής απόκλισης είναι 250.39. Σε επόμενο στάδιο της διερεύνησης και προ-επεξεργασίας των δεδομένων πραγματοποιείται μελέτη εύρεσης των ακραίων τιμών της μεταβλητής «Amount».

	count	mean	std	min	25%	50%	75%	max
Amount	283726	88.47269	250.3994	0	5.6	22	77.51	25691.16

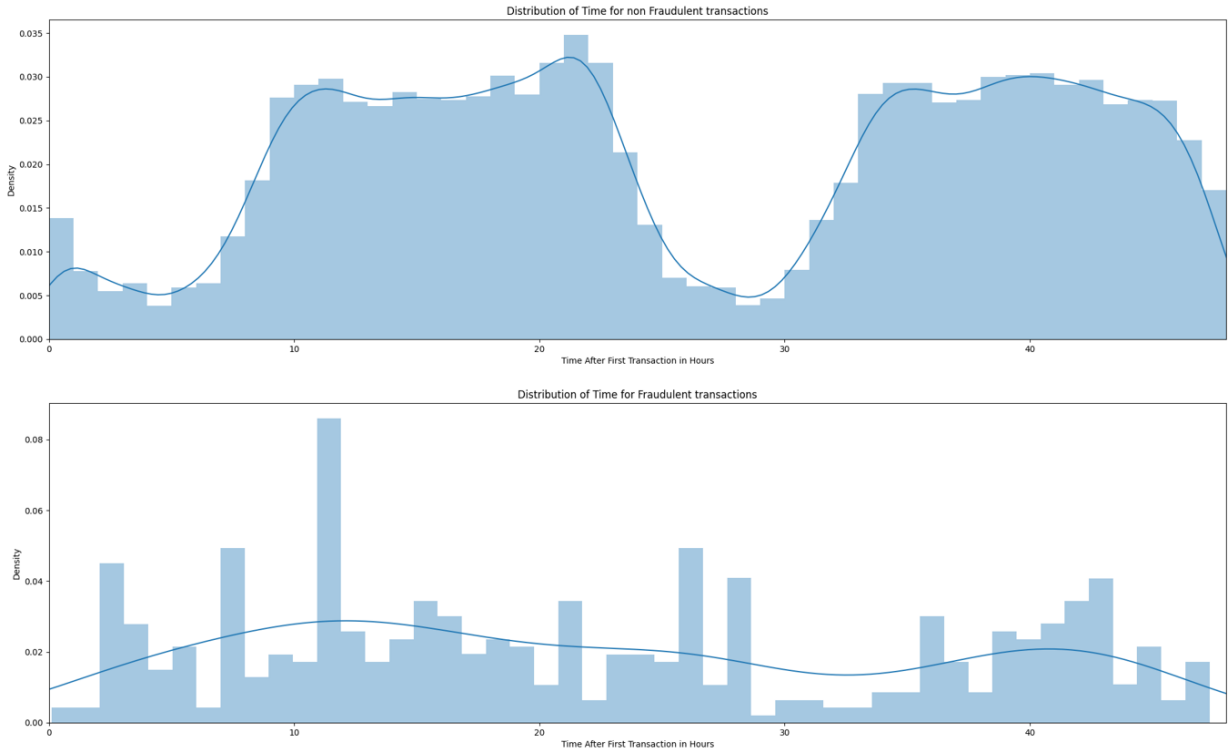
Εικόνα 24 Κατανομή Ύψους Ποσού Συναλλαγών

Συνεχίζοντας με την εικόνα 25 με κόκκινο χρώμα απεικονίζεται η κατανομή των συναλλαγών ανά χρονική στιγμή. Υπενθυμίζοντας ότι τα διαθέσιμα δεδομένα είναι δεδομένα συναλλαγών που γίνανε σε διάστημα δύο ημερών βλέπουμε μία περιοδικότητα στην πυκνότητα εμφάνισής τους. Σε αυτό το σημείο μπορούμε να κάνουμε μία υπόθεση για τις ώρες που γίνονται οι συναλλαγές. Υποθέτουμε λοιπόν ότι η αυξημένη πυκνότητα εμφάνισης των παρατηρήσεων συμβαίνει πιθανότερα κατά τη διάρκεια της ημέρας. Αυτή η υπόθεση είναι σημαντική προκειμένου να καταλάβουμε για το αν υπάρχει κάποια σχέση μεταξύ της ώρας και της μεταβλητής στόχου (Class).



Εικόνα 25 Κατανομή Χρόνου Συναλλαγών

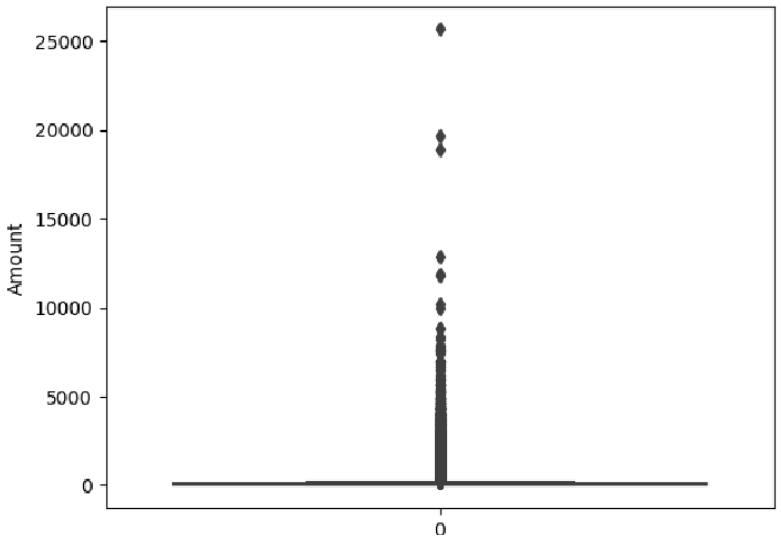
Στην εικόνα 26 παρουσιάζεται ένα συγκριτικό γράφημα των κατανομών των απατηλών και νόμιμων συναλλαγών σε σχέση με το χρόνο διεξαγωγής τους. Επιπλέον πραγματοποιήθηκε μετασχηματισμός της στήλης χρόνου από δευτερόλεπτα σε ώρες που σε συνδυασμό με την υπόθεση που έγινε προηγουμένως σαν 0 θεωρείται η ώρα 12.00 πμ. Είναι εμφανές ότι από τις 12.00 πμ μέχρι και τις 8.00 πμ περίπου, η πυκνότητα των πραγματοποιημένων συναλλαγών είναι μικρότερη από αυτή που παρατηρείται από τις 10.00 πμ έως τις 11.00 μμ. Στην συνέχεια υπάρχει πάλι αισθητή μείωση και αύξηση στα ίδια επίπεδα με αυτά της πρώτης μέρας. Συγκρίνοντας τώρα τα δύο γραφήματα φαίνεται η περιοδικότητα που παρουσιάζουν οι νόμιμες συναλλαγές δεν εντοπίζεται στις απάτες. Επιπλέον εκτός από μία περίπτωση μεγάλης συγκέντρωσης απατηλών συναλλαγών στις 12.00 μμ της πρώτης ημέρας, βλέπουμε ότι τα υπόλοιπα «peaks» της πυκνότητας τοποθετούνται σε βραδινές ώρες.



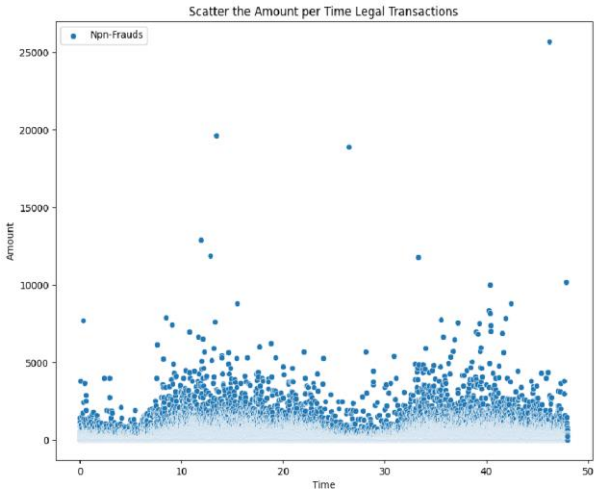
Εικόνα 26 Σύγκριση Νόμιμων και Απατηλών Κατανομών σε Σχέση με το Χρόνο

4.5 Ανάλυση Ακραίων Τιμών (Outlier Analysis)

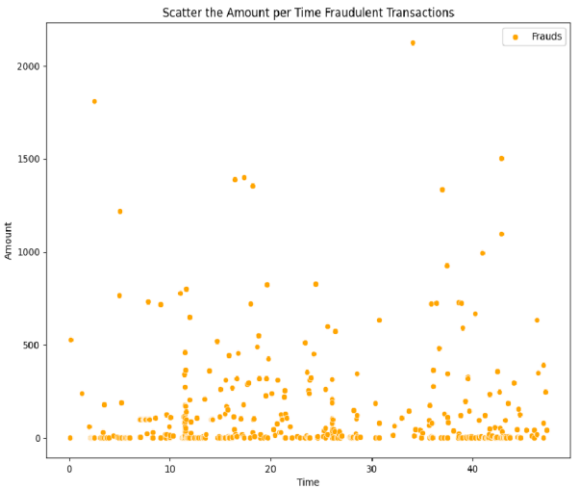
Όπως αναφέρθηκε προηγουμένως η μεταβλητή που εκφράζει το ύψος των ποσών των συναλλαγών του συνόλου δεδομένων παρουσιάζει μέση τιμή 88,5€ περίπου και μέγιστη τιμή τα 25.000€. Θα πρέπει να βρεθούν βάσει κάποιας συνάρτησης ή μέσω της οπτικοποίησης των τιμών οι ακραίες τιμές προκειμένου οι συγκεκριμένες παρατηρήσεις να μην τις λάβουμε υπόψη μας κατά το στάδιο της εκπαίδευσης των αλγορίθμων μηχανικής μάθησης. Ένα σχήμα που συνήθως μας βοηθάει να εντοπίζουμε εύκολα και πολύ ευδιάκριτα τις ακραίες τιμές είναι αυτό που ονομάζεται «boxplot όπου παραθέτοντας σε μία «ευθεία» όλες οι τιμές μίας μεταβλητής βλέπουμε αν είναι συνεχείς οι παρατηρήσεις ή διακεκομμένες όπως φαίνεται και στην εικόνα 27 από τις 8.000€ περίπου και μετά.



Εικόνα 27 Απεικόνιση Ακραίων Τιμών

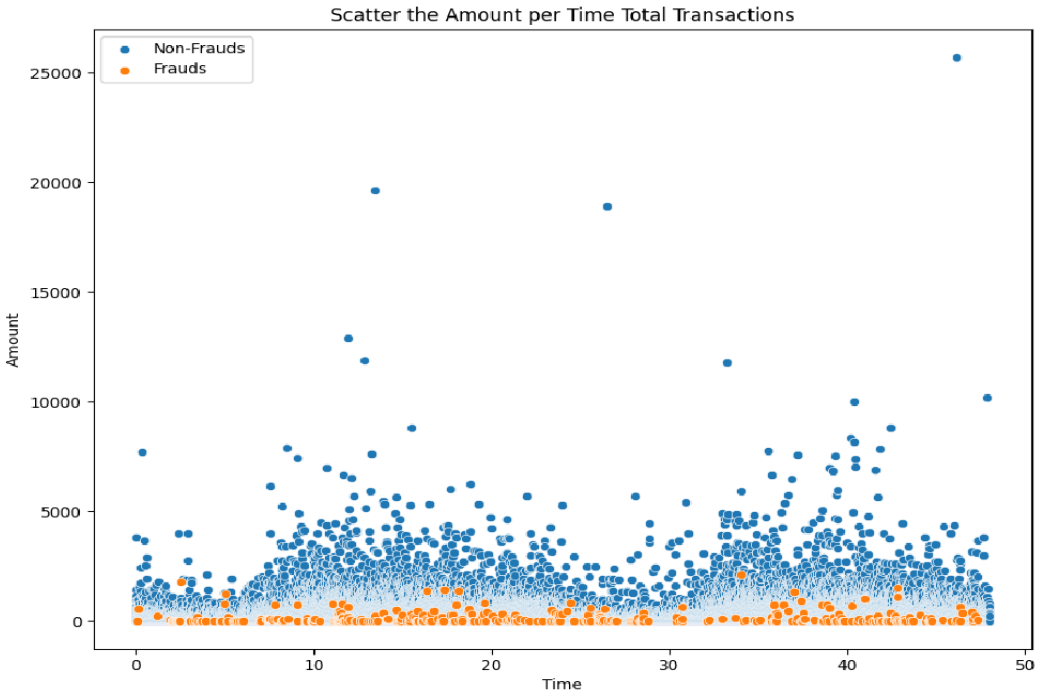


Εικόνα 28 Σημειακή Απεικόνιση Μεταβλητής Amount Νόμιμων Συναλλαγών



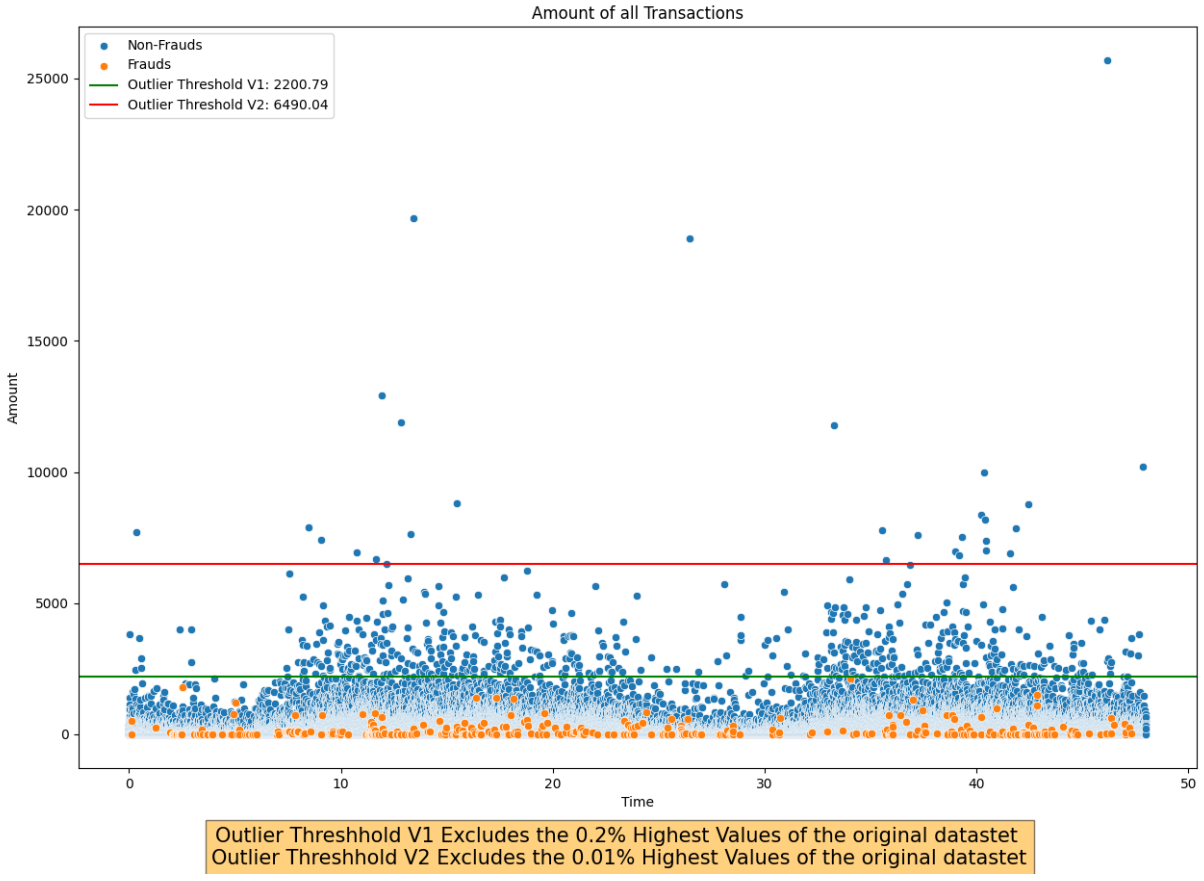
Εικόνα 29 Σημειακή Απεικόνιση Μεταβλητής Amount Απατηλών Συναλλαγών

Στα σχήματα των εικόνων 28 και 29 απεικονίζονται τα ποσά των νόμιμων και απατηλών συναλλαγών αντίστοιχα για κάθε ώρα παρατήρησης. Το ύψος των ποσών φαίνεται να διατηρεί την περιοδικότητα που παρατηρήθηκε και στην πυκνότητα εμφάνισης των νόμιμων συναλλαγών ενώ αντίθετα στις απατηλές δεν φαίνεται να ακολουθούν κάποιο μοτίβο. Συνεπώς συμπεραίνουμε ότι τις βραδινές ώρες τα ποσά που δαπανούνται είναι μικρότερα από αυτά που δαπανούνται κατά τη διάρκεια της ημέρας. Ένα επίσης σημαντικό εύρημα είναι ότι τα ποσά των απατηλών ενεργειών είναι σημαντικά μικρότερα από αυτά των νόμιμων συναλλαγών. Γίνεται οπτικά διακριτό όταν τα δύο αυτά σχήματα συνδυάζονται όπως γίνεται στην εικόνα 30.



Εικόνα 30 Συνδυαστική Απεικόνιση των 24, 25

Με την βοήθεια του προηγούμενου συνδυαστικού γραφήματος γίνονται αντιληπτές τιμές που αποσπώνται από την «κανονικότητα» που εντοπίζεται μέχρι τα 2.500€ περίπου. Στόχος είναι η εύρεση εκείνης της εξίσωσης που θα αφαιρεί τις παρατηρήσεις με οριακές τιμές χωρίς να αφαιρέσουμε κάποια παρατήρηση που είναι απάτη. Μας ενδιαφέρει να διατηρήσουμε όλες τις απατηλές παρατηρήσεις καθώς το σύνολό τους είναι ήδη πολύ μικρό και δεν θέλουμε να το περιορίσουμε περεταίρω. Αναλύοντας τη κατανομή των τιμών της μεταβλητής Amount βρίσκουμε ότι το 99,8% των παρατηρήσεων είναι κάτω του ποσού των 2.200,80€ και το 99,99% των παρατηρήσεων είναι κάτω του ποσού των 6.940,04€. Αν ορίσουμε σαν ακραίες τιμές το 0,01% των παρατηρήσεων θα αφαιρέσουμε μόλις 29 οριακές παρατηρήσεις ενώ αντίστοιχα αν ορίσουμε το 0,2% των παρατηρήσεων ως οριακές τιμές θα αφαιρέσουμε 568 νόμιμες συναλλαγές. Για την επιλογή του κατάλληλου ποσοστού ακραίων τιμών μας βοηθάει το επόμενο γράφημα (εικόνα 31).



Εικόνα 31 Απεικόνιση Τιμών Κατωφλίου Οριακών Τιμών

Επιλέγεται η κόκκινη γραμμή που αντιστοιχεί στο 0,01% παρατηρήσεων καθώς από εκείνο το σημείο και μετά εμφανίζονται οι παρατηρήσεις πολύ πιο αραιές. Δεν επιλέγεται η πράσινη γραμμή καθώς παρά το ότι σέβεται τη «συνθήκη» περί μη απόρριψης απατηλής συναλλαγής αφαιρεί σημαντικό μέρος του συνόλου που δεν αποτελείται από οριακές τιμές βάση σχήματος. Η αφαίρεση των όλων των παρατηρήσεων από το ποσό των 2.200€ και άνω θα δημιουργούσε μοντέλα ταξινόμησης που θα ήταν επιρρεπή σε λάθη καθώς συναλλαγές αυτών των ποσών είναι πιθανό να γίνουν. Τέλος το 0,2% των παρατηρήσεων είναι μεγάλο ποσοστό για να θεωρηθεί ως οριακές τιμές όταν το ποσοστό της κλάσης μειοψηφίας είναι μόλις 0,0167%.

Μετά την απόρριψη του 0,01% του συνόλου δεδομένων ως οριακές τιμές έχουμε πλέον ένα σύνολο δεδομένων μήκους 283.224 νόμιμων παρατηρήσεων και 473 παρατηρήσεων απάτης. Επιπλέον η νέα μέγιστη τιμή της ανεξάρτητης μεταβλητής «Amount» ανέρχεται στα 6554,74€. Σε αυτό το σημείο πριν ξεκινήσουμε την εφαρμογή των μεθόδων της μηχανικής μάθησης πραγματοποιήθηκε κανονικοποίηση των τιμών της μεταβλητής διαμορφώνοντας το νέο μέγιστο το 28,07277 και νέο ελάχιστο το -0,38578.

4.6 Μηχανική Μάθηση

Η παρούσα πειραματική διαδικασία ξεκινά με την εξισορρόπηση των κατανομών των κλάσεων με τυχαία υπό – δειγματοληψία και συνεχίζει με την εύρεση του καταλληλότερου συνδυασμού υπέρ-παραμέτρων του εκάστοτε αλγορίθμου όπου γίνεται σύγκρισή τους ως προς τη μετρική της ακρίβειας της κατηγοριοποίησης. Τέλος πραγματοποιείται διερεύνηση κατά την οποία γίνεται εξισορρόπηση των κλάσεων με τη μέθοδο SMOTE (υπέρ - δειγματοληψία) και γίνεται σύγκριση των αποτελεσμάτων της Στρατηγικής Διασταυρούμενης επικύρωσης κ-φορών μεθόδου με αυτά των περιπτώσεων υπό – δειγματοληψίας.

Έχοντας λοιπόν ένα σύνολο με υψηλή ανομοιομορφία των κλάσεων προς κατηγοριοποίηση, θα πρέπει να βρεθεί μία τεχνική υπό-δειγματοληψίας, της κλάσης πλειοψηφίας, ή υπέρ-δειγματοληψίας, της κλάσης μειοψηφίας, για την εξισορρόπηση των κατανομών των δύο κλάσεων σε αποδεκτά επίπεδα. Λόγω του περιορισμού των υπολογιστικών πόρων στη παρούσα διερεύνηση χρησιμοποιείται η τυχαία υπό-δειγματοληψία για το σύνολο 7 αλγορίθμων ενώ για την διερεύνηση της υπέρ – δειγματοληψίας θα χρησιμοποιηθούν μόνο 6 αλγόριθμοι (όχι ο Support Vector Machine). Γίνεται λόγος για τους υπολογιστικούς πόρους καθώς αν διαμορφώσουμε με την τεχνική SMOTE το υπό εξέταση σύνολο δεδομένων και εξισορροπήσουμε τις δύο κλάσεις, θα δημιουργηθεί ένα σύνολο των 600 χιλιάδων παρατηρήσεων κάνοντας την διερεύνηση αυτού του αλγορίθμου μηχανικής μάθησης αδύνατη για το περιβάλλον του collab, που υλοποιήθηκε η συγκεκριμένη εργασία, καθώς υπάρχει περιορισμός στο χρόνο εκτέλεσης των διεργασιών.

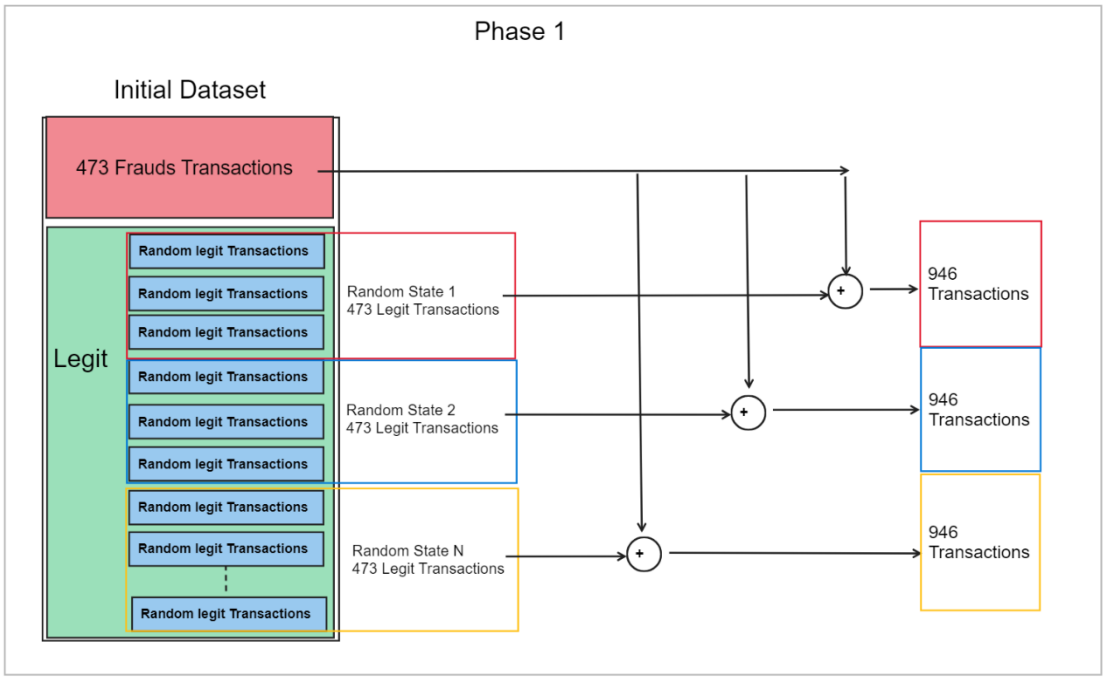
4.6.1 Υπό – Δειγματοληψία Περιγραφή Διαδικασίας

Έχοντας πλέον "καθαρίσει" τα δεδομένα διαγράφοντας τις διπλές και οριακές τιμές, βλέπουμε ότι η αναλογία μεταξύ των δύο κλάσεων δεν έχει διαφοροποιηθεί προς όφελος μας. Οι απατηλές συναλλαγές παραμένουν η κλάση μειοψηφίας συγκεντρώνοντας το ποσοστό του 0.1667% του συνολικού συνόλου δεδομένων. Αν προχωρήσουμε σε μία ανάπτυξη ταξινομητών με αυτές τις αναλογίες των κλάσεων οι αλγόριθμοι θα αποδίδουν τις νέες παρατηρήσεις στην κλάση της πλειοψηφίας καθώς είναι η συντριπτικά πιο πιθανή επιλογή.

Στόχος λοιπόν είναι η δημιουργία ενός συνόλου δεδομένων με όλες τις παρατηρήσεις απάτης, που σημειώνονται στο αρχικό dataset και ένα τυχαίο υποσύνολο νόμιμων συναλλαγών ίσο σε αριθμό με αυτό των συναλλαγών απάτης. Έτσι λοιπόν στη συνέχεια παρουσιάζεται ο τρόπος με τον οποίο γίνεται η τυχαία υπό-δειγματοληψία της κλάσης πλειοψηφίας με απώτερο σκοπό την δημιουργία ενός ή περισσοτέρων, όπως θα δούμε παρακάτω, συνόλων δεδομένων με ίσο αριθμό παρατηρήσεων απάτης και νόμιμων συναλλαγών.

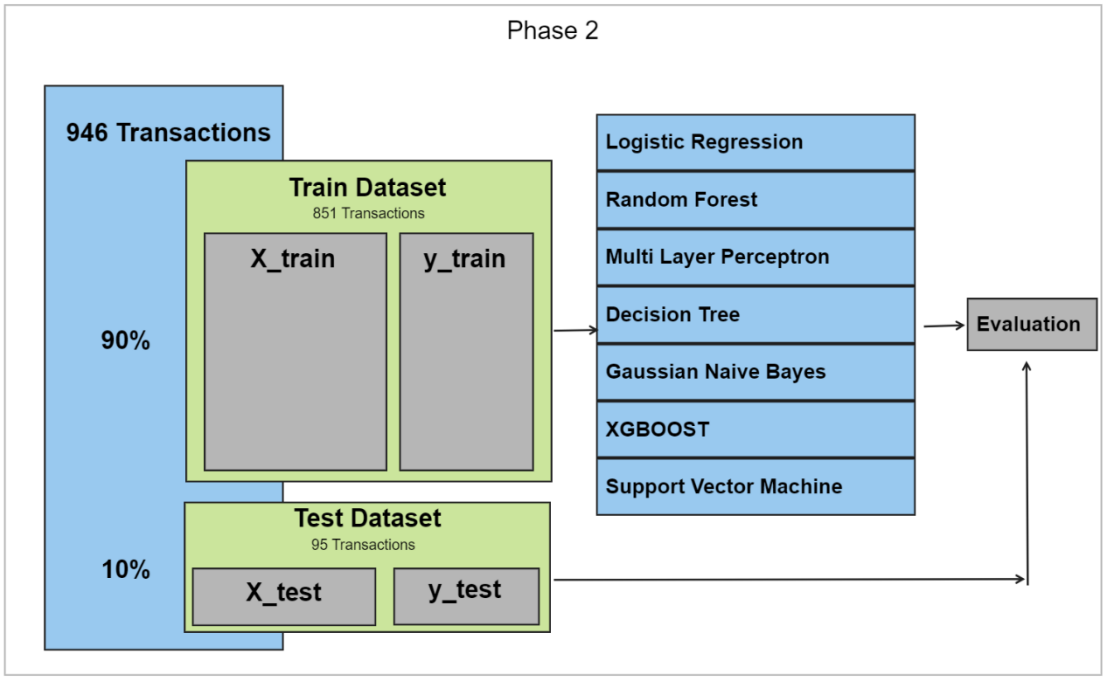
Μετά την δημιουργία ενός συνόλου δεδομένων με ίσο αριθμό παρατηρήσεων μεταξύ των δύο κλάσεων πραγματοποιείται μία σύντομη περιήγηση στους υπό εξέταση αλγορίθμους δημιουργώντας μοντέλα για κάθε ένα από τους αλγόριθμους που αναλύθηκαν στο αντίστοιχο κεφάλαιο της παρούσας εργασίας. Σε αυτή τη πρώτη φάση της πειραματικής διαδικασίας δεν πραγματοποιείται καμία παραμετροποίηση αλγορίθμων και γίνεται μία σύγκριση της αποδοτικότητας τους, χρησιμοποιώντας τις προκαθορισμένες

τιμές των υπέρ-παραμέτρων που αποδίδει η βιβλιοθήκη scikit-learn στις μεθόδους της. Επίσης κατά την επισκόπηση των αποτελεσμάτων των αλγορίθμων θα χρησιμοποιηθεί μόνο ένα τυχαίο υποσύνολο του δείγματος που θα μοιράζει στη μέση τις παρατηρήσεις σε απάτες και νόμιμες. Δηλαδή θα δημιουργηθεί δείγμα συνολικού μήκους 946 παρατηρήσεων εκ των οποίων το 50% θα είναι απατηλές συναλλαγές και 50% νόμιμες συναλλαγές.



Εικόνα 32 Σχηματική Απεικόνιση Τρόπου Υπό - δειγματοληψίας των Δεδομένων

Στην συνέχεια μέσω της μεθόδου train_test_split της scikit-learn, μπορούμε να δημιουργήσουμε σύνολα εκπαίδευσης δεδομένων αλλά και σύνολα δοκιμής. Η μέθοδος δέχεται σαν είσοδο το σύνολο των δεδομένων ορίζοντας ποια χαρακτηριστικά είναι αυτά των ανεξάρτητων μεταβλητών και ποιο το χαρακτηριστικό στόχος. Η μέθοδος στην έξοδό της δίνει τέσσερις πίνακες όπως φαίνεται και στην εικόνα 33 X_train, y_train, X_test, y_test. Έχει οριστεί για όλη τη διάρκεια της διεξαγωγής της πειραματικής διαδικασίας ότι από τις 946 συναλλαγές το 90% θα παρέχεται στους εκτιμητές για να εκπαιδευτούν και μόλις 10% για δοκιμές. Οι πίνακες «X» περιέχουν τη πληροφορία των ανεξάρτητων μεταβλητών ενώ οι πίνακες «y» περιέχουν τις τιμές της μεταβλητής στόχου (class). Συνεπώς τα μοντέλα εκπαιδεύονται έχοντας σαν είσοδο τους πίνακες X_train και y_train όπου βρίσκονται οι κατάλληλοι μετασχηματισμοί που οδηγούν στο επιθυμητό αποτέλεσμα δεδομένης κάποιας εισόδου.



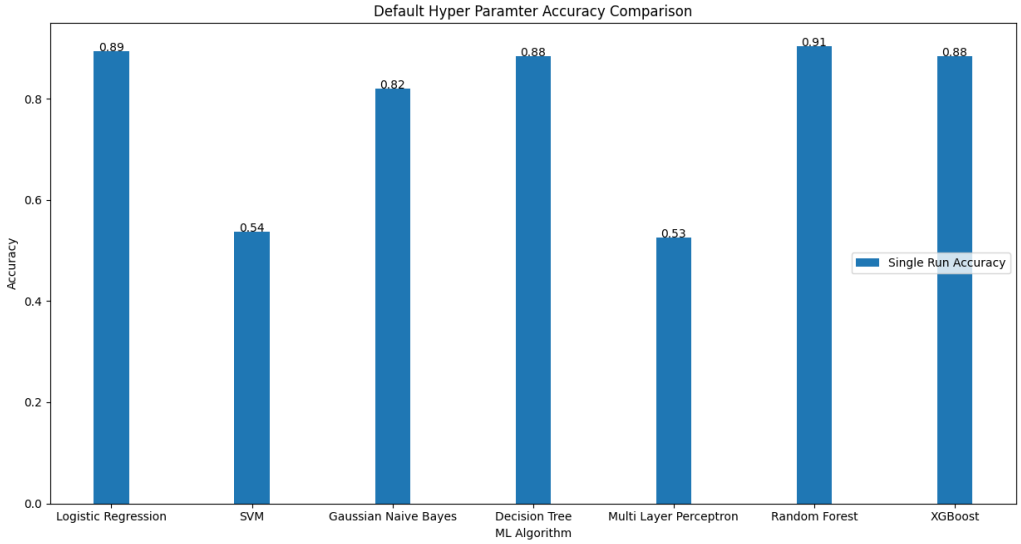
Εικόνα 33 Διαχωρισμός Τυχαίου Υποσυνόλου σε Σύνολα Εκπαίδευσης και Σύνολα Δοκιμής

Αφού ολοκληρωθεί η εκπαίδευση των αλγορίθμων με τα δεδομένα εκπαίδευσης σειρά έχει η διαδικασία της αξιολόγησης (evaluation) όπου χρησιμοποιούνται δεδομένα που ο αλγόριθμος δεν έχει «ξαναδεί». Γίνεται η εκτίμηση της τιμής στόχου κάθε παρατήρησης και έπειτα συγκρίνεται με τα πραγματικά δεδομένα της μεταβλητής που είναι τα y_test. Αφού ολοκληρωθούν οι εκτιμήσεις για όλο σύνολο δοκιμής μετράτε το ποσοστό των εύστοχων εκτιμήσεων. Τα αποτελέσματα των αλγορίθμων χωρίς να επέμβουμε στην παραμετροποίηση τους και εκπαιδεύοντάς τους μόνο με ένα τυχαία επιλεγμένο υπό-σύνολο δεδομένων, παρουσιάζονται στον παρακάτω πίνακα (πίνακας 7).

Εκτιμητής (Classifier)	Υπέρ – Παράμετροι (Hyper Parameters)	Ακρίβεια (Accuracy) %
Logistic Regression	Προκαθορισμένο (Default)	89.00%
Support Vector Machine	Προκαθορισμένο (Default)	54.00%
Decision Tree	Προκαθορισμένο (Default)	88.00%
Gaussian Naïve Bayes	Προκαθορισμένο (Default)	82.00%
Multi-Layer Perceptron	Προκαθορισμένο (Default)	53.00%
Random Forest	Προκαθορισμένο (Default)	91.00%
XGBoost	Προκαθορισμένο (Default)	88.00%

Πίνακας 7 Συγκριτικός Πίνακας Ακρίβειας Αλγορίθμων Προκαθορισμένων Υπέρ – Παραμέτρων Για μία Περίπτωση Τυχαίου Δείγματος

Φαίνεται ότι καλύτερα αποδίδουν οι αλγόριθμοι των τυχαίων δασών, της λογιστικής Παλινδρόμησης και XGBoost. Για τους Random Forest και XGBoost δεν αποτελεί έκπληξη αυτή ακρίβεια καθώς ανήκουν στη κατηγορία των αλγορίθμων εκμάθησης συνόλου όπου, συνδυάζοντας πάνω από ένα μοντέλο ταξινόμησης, πετυχαίνουν επιδόσεις που οι μεμονωμένοι ταξινομητές δεν μπορούν να επιτύχουν. Αντίθετα τις χειρότερες επιδόσεις πετυχαίνει ο αλγόριθμος διανυσμάτων υποστήριξης με 54.00% και το νευρωνικό δίκτυο πολύ-επίπεδων αισθητήρων με 53%. Προσφέρεται και το συγκριτικό γράφημα (γράφημα) όπου παρουσιάζονται οι επιδόσεις των αλγορίθμων του πίνακα 7 ως προς την ακρίβεια των εκτιμήσεων τους.



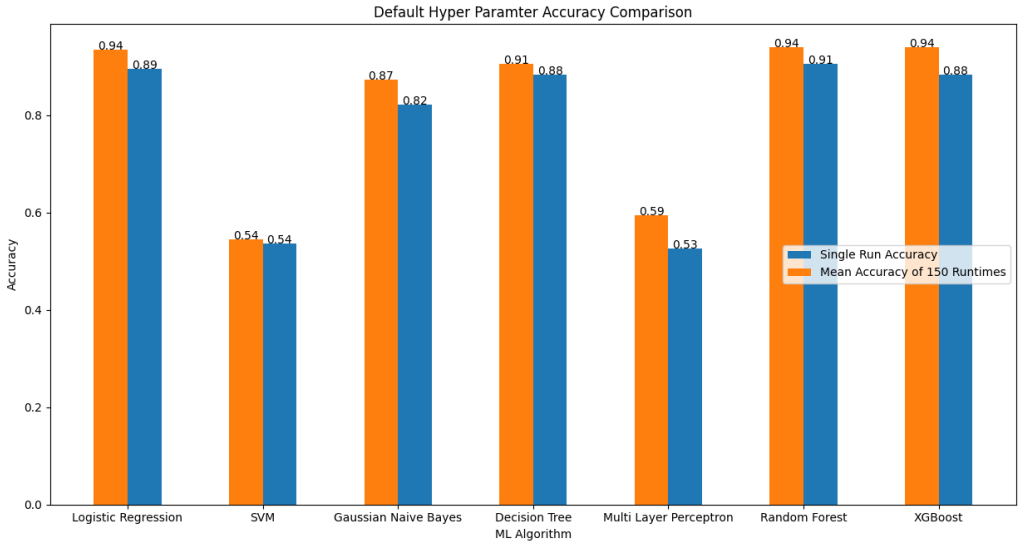
Εικόνα 34 Απεικόνιση Ποσοστού Ακριβείας Με Default Παραμέτρους

Τα αποτελέσματα αυτά όπως αναφέρθηκε και νωρίτερα είναι προϊόν ενός μόνο τυχαίου δείγματος κατά την υπό – δειγματοληψία. Έτσι λοιπόν θα χρειαστεί η διαδικασία που αναφέρθηκε προηγουμένως να επαναληφθεί αλλά για διαφορετικά τυχαία σύνολα όπως φαίνεται και στην εικόνα 32. Για την ακρίβεια χρησιμοποιούνται 150 διαφορετικές περιπτώσεις σύνθεσης του δείγματος προς εξέταση, αλλάζοντας συνεχώς τα δεδομένα που εκπαιδεύονται και κατηγοριοποιούν οι αλγόριθμοι. Αφού λοιπόν υπολογιστεί η ακρίβεια του εκάστοτε αλγορίθμου για τη κάθε περίπτωση δείγματος υπολογίζεται ο μέσος όρος της.

Εκτιμητής (Classifier)	Υπέρ – Παράμετροι (Hyper Parameters)	Μέση Ακρίβεια (Mean Accuracy) %
Logistic Regression	Προκαθορισμένο (Default)	94.00%
Support Vector Machine	Προκαθορισμένο (Default)	54.00%
Decision Tree	Προκαθορισμένο (Default)	91.00%
Gaussian Naïve Bayes	Προκαθορισμένο (Default)	87.00%
Multi-Layer Perceptron	Προκαθορισμένο (Default)	59.00%
Random Forest	Προκαθορισμένο (Default)	94.00%
XGBoost	Προκαθορισμένο (Default)	94.00%

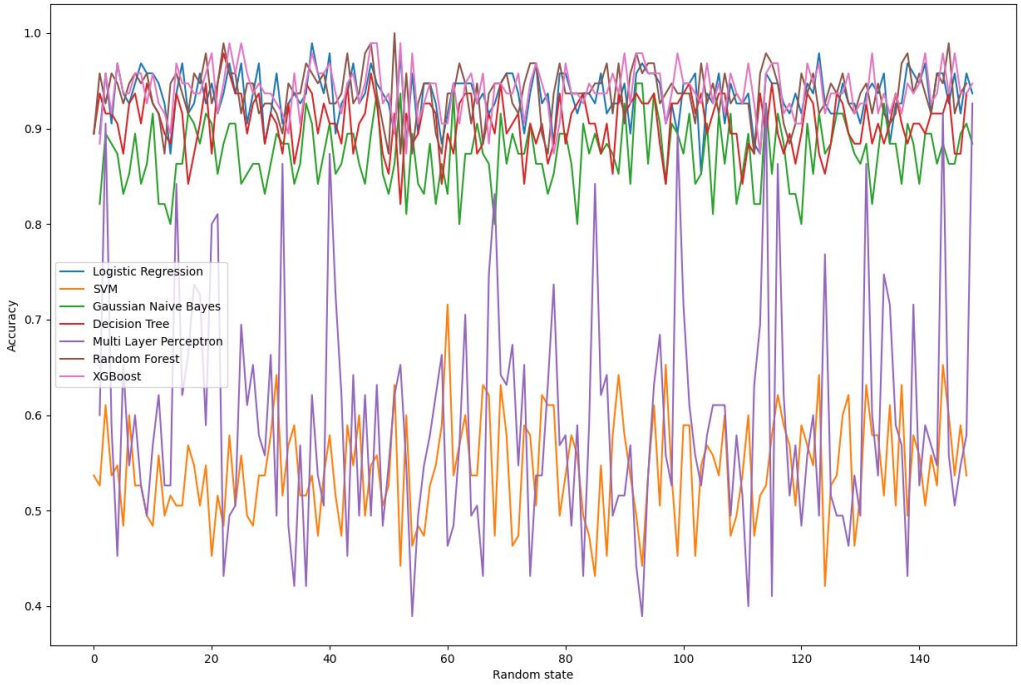
Πίνακας 8 Συγκριτικός Πίνακας Ακρίβειας Αλγορίθμων Προκαθορισμένων Υπέρ – Παραμέτρων Για 150 Διαφορετικές Περιπτώσεις Τυχαίου Δείγματος

Συγκρίνοντας τα αποτελέσματα των πινάκων 7 και 8 βλέπουμε ότι υπάρχει διαφορά όσον αφορά τα αποτελέσματα των αλγορίθμων. Κάποιοι βελτιώθηκαν και κάποιοι παρέμειναν ίδιοι. Θεωρούμε ότι ο πίνακας 8 είναι πιο ακριβής ως προς τις εκτιμήσεις του καθώς αναγράφει τα αποτελέσματα των αλγορίθμων από μεγαλύτερο, αν το δούμε αθροιστικά, υποσύνολο του αρχικού. Έτσι λοιπόν η σύγκριση των αποτελεσμάτων μετά τον υπολογισμό των υπέρ – παραμέτρων θα γίνει με τα αποτελέσματα του πίνακα 8. Παρακάτω παρουσιάζεται μία σχηματική απεικόνιση της σύγκρισης των αποτελεσμάτων του πίνακα 8 με τις τιμές του πίνακα 7. Έχοντας αποκτήσει πλέον μία ευρύτερη άποψη σχετικά με την ακρίβεια των εκτιμητών φαίνεται ότι ο αλγόριθμος τυχαίων δασών και XGBoost παραμένουν οι καταλληλότεροι για το συγκεκριμένο πρόβλημα ταξινόμησης ενώ τρίτος καταλληλότερος αλγόριθμος πλέον είναι αυτός της Λογιστικής Παλινδρόμησης αντί των Δέντρων Αποφάσεων.

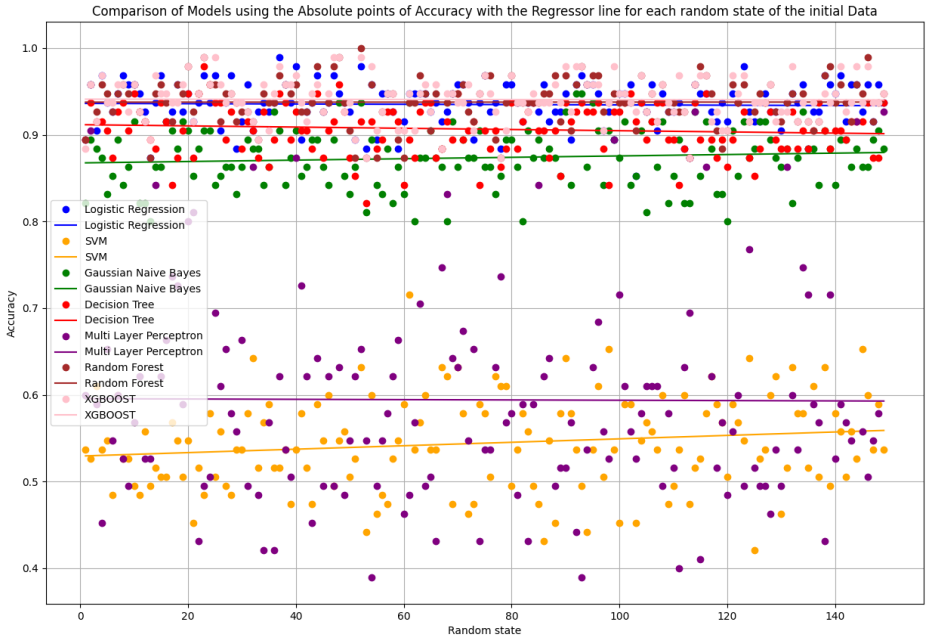


Εικόνα 35 Απεικόνιση Σύγκρισης Τιμών των Πινάκων 7 και 8

Στην Εικόνα 36 παρουσιάζεται μία απεικόνιση της διακύμανσης της ακρίβειας των μοντέλων των διαφορετικών που προκαλείται από τα διαφορετικά σύνολα δεδομένων. Βλέπουμε ότι η τιμές τις ακρίβειας των αλγορίθμων της λογιστικής παλινδρόμησης, του Gaussian Naive Bayes, του αλγόριθμου τυχαίων δασών και του XGBoost παρουσιάζουν αποκλίσεις αλλά όχι της τάξης των διανυσμάτων υποστήριξης και του πολύ-επίπεδων αισθητήρων. Επιπλέον δίνεται και το σχήμα όπου αναπαρίστανται σημειακά οι τιμές της ακρίβειας και υπολογίζεται η ευθεία παλινδρόμησης για κάθε ένα από τους αλγορίθμους στην εικόνα 37.



Εικόνα 36 Απεικόνιση Ακρίβειας για Κάθε Περίπτωση του Τυχαίου Συνόλου



Εικόνα 37 Σημειακή Απεικόνιση των Τιμών Ακρίβειας του Κάθε Τυχαίου Συνόλου

Βλέπουμε ότι τέσσερεις από τους 7 κατηγοριοποιητές που εξετάζουμε πετυχαίνουν πολύ καλές επιδόσεις όσον αφορά την συνολική ακρίβεια της εκτίμησης. Όμως σε πολλές περιπτώσεις μπορεί το ποσοστό ακρίβειας να είναι πολύ μεγάλο και παράλληλα το μοντέλο μας να είναι πλήρως αποτυχημένο. Μία τέτοια περίπτωση είναι όταν οι κατανομές των κλάσεων της μεταβλητής στόχου είναι εξαιρετικά ανομοιόμορφες. Αυτό συμβαίνει καθώς ο εκτιμητής αποδίδει όλες τις προβλέψεις στην κλάση της πλειοψηφίας. Για παράδειγμα αν επιχειρήσουμε να εκτιμήσουμε 100 συναλλαγές από τις οποίες μόλις οι 5 είναι απάτες και οι υπόλοιπες είναι νόμιμες και ο κατηγοριοποιητής τις αποδώσει όλες σαν νόμιμες, η ακρίβειά του, υπολογίζεται να είναι της τάξης του 95%. Όμως θα έχει αποτύχει πλήρως να εντοπίσει τις απάτες που είναι και το ζητούμενο. Για αυτό το λόγω έχουν ορισθεί επιπλέον μετρικές οι οποίες επικεντρώνονται στις επιδόσεις ανά κλάση. Μας ενδιαφέρει λοιπόν η αναλογία της υψηλής ορθώς θετικής/αρνητικής εκτίμησης και κατ' επέκταση της χαμηλής λανθασμένης θετικής/αρνητικής εκτίμησης.

Κάνοντας μία βαθύτερη εισαγωγή στην επιχειρησιακή λογική και των αποφάσεων που είναι καθοδηγούμενες από τα αποτελέσματα των αλγορίθμων μηχανικής μάθησης, θα πρέπει να υπάρχει ιδανικά μία συνάρτηση κόστους. Μέσω της συνάρτησης κόστους θα μπορούσαμε να πάρουμε σωστές αποφάσεις προς όφελος τις επιχείρησης/οργανισμού καθώς θα είναι οι λιγότερο ζημιογόνες.

Για παράδειγμα, σε πρώτη φάση θα θεωρήσουμε το positive να είναι μία συναλλαγή απάτης. Μπορεί να δημιουργηθεί λοιπόν, ένα μοντέλο το οποίο μπορεί να πετυχαίνει υψηλό ποσοστό συνολικής ακρίβειας αλλά παράλληλα να έχει και ένα σημαντικό ποσοστό λανθασμένων θετικών. Σαν αποτέλεσμα της εκτίμησης ότι μία συναλλαγή είναι απάτη το σύστημα θα πρέπει να την μπλοκάρει και να μην την ολοκληρώσει. Αν όντως ήταν απάτη τότε καλώς το αυτόματο σύστημα έκανε και μπλόκαρε την συναλλαγή. Αν όμως το μοντέλο εκτιμήσει ότι μία συναλλαγή είναι απάτη ενώ στην πραγματικότητα δεν είναι (λόγω υψηλού false positive είναι πολύ πιθανό) και μπλοκάρει την συναλλαγή, τότε το αποτέλεσμα θα είναι ένας πολύ δυσαρεστημένος πελάτης. Αυτό αποτελεί σημαντικό κόστος για την τράπεζα και θα πρέπει να διαστασιοποιηθεί αποτελεσματικά και να μεταφραστεί σε μία συνάρτηση κόστους. Αν η συνάρτηση παραδείγματος χάρει έχει σαν αποτέλεσμα ότι το κόστος είναι μικρότερο με το να αφήσει την απατηλή συναλλαγή να πραγματοποιηθεί από το να έχουμε δυσαρεστημένους πελάτες τότε δεν θα μπλοκάρει τη συναλλαγή το σύστημα. Αντίθετα αν εκτιμηθεί ότι το κόστος είναι μεγαλύτερο αν πραγματοποιηθεί η συναλλαγή τότε θα την μπλοκάρει.

Σύμφωνα με τα παραπάνω πρώτη μετρική που θα πρέπει να συλλογιστούμε είναι η F1 – Score καθώς είναι η αυτή που θα μας δώσει μία αναλογία των λανθασμένων θετικών που είναι τόσο σημαντική για την περίπτωση της ανίχνευσης απάτης.

Για την απόδοση των μετρικών ανά κλάση κατηγοριοποίησης χρησιμοποιείται ο πίνακας σύγχυσης (confusion matrix) όπως είδαμε και σε προηγούμενη ενότητα. Στην συνέχεια παρουσιάζεται ο συγκριτικός πίνακας (πίνακας 9) των επιδόσεων ανά κλάση για τους αλγορίθμους χωρίς παραμετροποίηση των υπέρ – παραμέτρων του καθώς και οι πίνακες σύγχυσης και καμπύλες ROC ανά ταξινομητή.

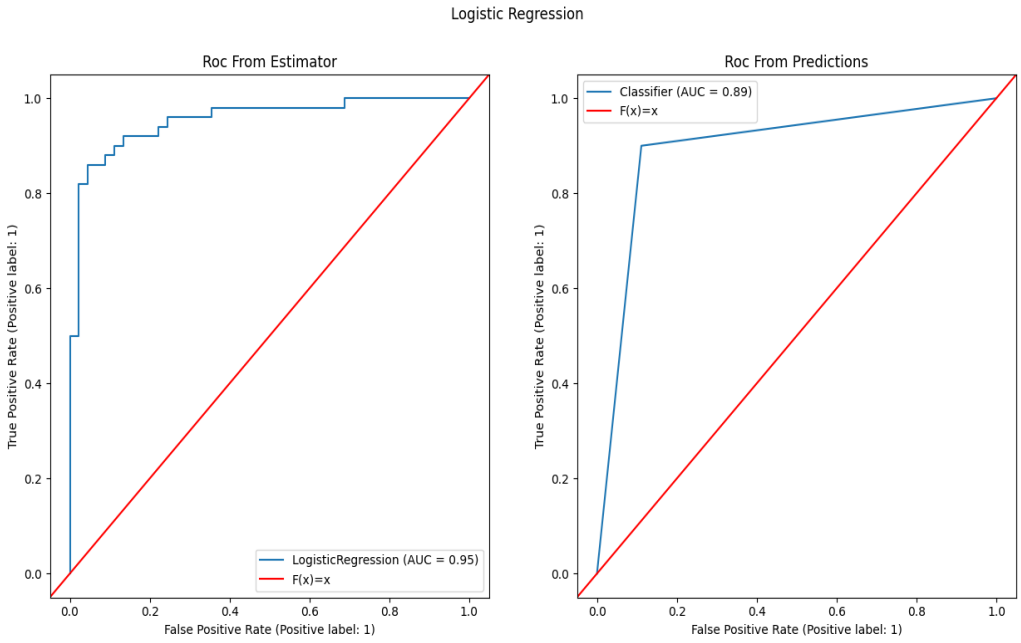
Εκτιμητής (Classifier)	Υπέρ – Παράμετροι (Hyper Parameters)	Κλάση	Precision	Recall	F1 - Score	Support
Logistic Regression	Προκαθορισμένο (Default)	0	0.89	0.89	0.89	45
		1	0.9	0.9	0.9	50
Support Vector Machine	Προκαθορισμένο (Default)	0	0.51	0.78	0.61	45
		1	0.62	0.32	0.42	50
Decision Tree	Προκαθορισμένο (Default)	0	0.86	0.93	0.89	45
		1	0.93	0.86	0.9	50
Gaussian Naïve Bayes	Προκαθορισμένο (Default)	0	0.73	0.98	0.84	45
		1	0.97	0.68	0.8	50
Multi-Layer Perceptron	Προκαθορισμένο (Default)	0	0.00	0.00	0.00	45
		1	0.53	1.00	0.69	50
Random Forest	Προκαθορισμένο (Default)	0	0.86	0.96	0.91	45
		1	0.96	0.86	0.91	50
XGBoost	Προκαθορισμένο (Default)	0	0.84	0.93	0.88	45
		1	0.93	0.84	0.88	50

Πίνακας 9 Συγκριτικός Πίνακας Μετρικών Κλάσεων Κατηγοριοποίησης (Προκαθορισμένοι Παράμετροι)

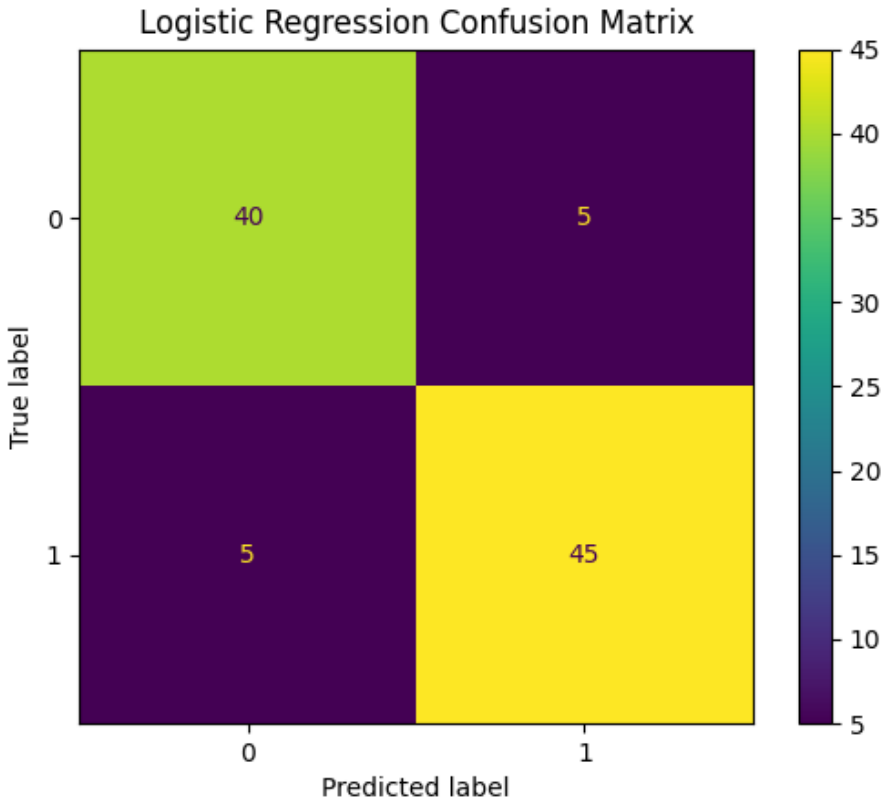
Ξεκινώντας την αξιολόγηση των γραφημάτων των εικόνων 38 έως και 51 θα πρέπει να οριστεί το περιβάλλον και ο επιθυμητός στόχος της διερεύνησης. Μετά την τυχαία υποδειγματοληψία της κλάσης μειοψηφίας και το διαμοιρασμό του δείγματος όπου το 90% θα είναι το σύνολο εκπαίδευσης και 10% το σύνολο δοκιμής έχουμε τις πρώτες κατηγοριοποιήσεις των εκτιμητών. Ειδικότερα οι 7 αλγόριθμοι επιχειρούν να κατηγοριοποιήσουν, μετά την εκπαίδευσή τους φυσικά, 95 παρατηρήσεις. Στόχος μας είναι η υψηλή ακρίβεια συνολικά αλλά και το υψηλό f1-score που δηλώνει χαμηλό λόγο λανθασμένων θετικών εκτιμήσεων.

Με την βοήθεια του πίνακα 9 και των παρακάτω εικόνων βλέπουμε ότι για την κλάση 1 (που σημαίνει ότι η συναλλαγή είναι απάτη) ο εκτιμητής που έχει χαμηλότερο λόγο λανθασμένων θετικών είναι ο αλγόριθμος τυχαίων δασών με 0,91. Στο πίνακα σύγχυσης της εικόνας 49 παρατηρούμε ότι μόλις 2 παρατηρήσεις έχουν εκτιμηθεί ως απάτες και δεν ήταν. Επίσης είναι καλύτερος από όλους του υπόλοιπους αλγορίθμους στην εκτίμηση και των νόμιμων συναλλαγών καθώς έχει κατηγοριοποίησει λάθος μόλις 7 παρατηρήσεις.

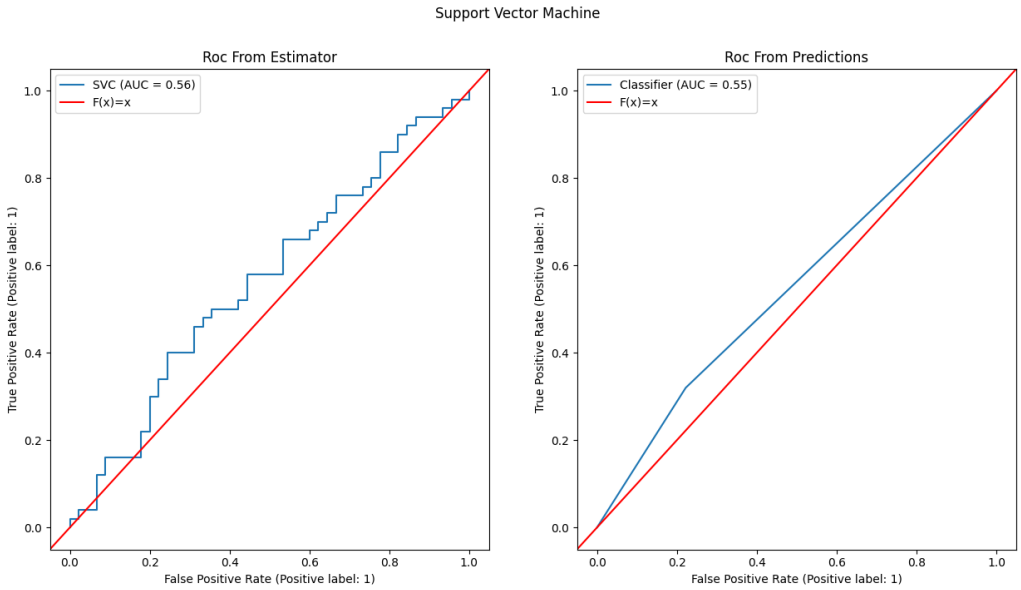
Χειρότερος από τους 7 αλγορίθμους είναι αυτός των διανυσμάτων μηχανών υποστήριξης (support vector machine) όπου το f1-score για την κλάση 1 είναι μόλις 0,42. Από την εικόνα 40 φαίνεται ότι οι καμπύλες ROC είναι δεν απέχουν πολύ από την ευθεία y=x, δηλαδή την τυχαία εκτίμηση της παρατήρησεις. Αυτό απεικονίζεται και στον πίνακα σύγχυσης της εικόνας 41 όπου έχει 10 λανθασμένα θετικές παρατηρήσεις αλλά ακόμα πιο ανακριβής είναι για την κλάση 0 όπου έχει κατηγοριοποιήσει λανθασμένα 34 παρατηρήσεις. Συνεπώς καταλαβαίνουμε ότι από τις 50 παρατηρήσεις απάτης στο σύνολο δοκιμής ο αλγόριθμος κατηγοριοποίησε σωστά τις 16 και του «ξέφυγαν» οι 34. Η κατάσταση αυτή εκτιμάται να διαφοροποιηθεί μετά την ρύθμιση των υπέρ – παραμέτρων του αλγορίθμου αλλά οι απώλειες είναι πολύ μεγάλες προκειμένου να βρεθεί ως ο ιδανικός αλγόριθμος για την επίλυση αυτού του προβλήματος.



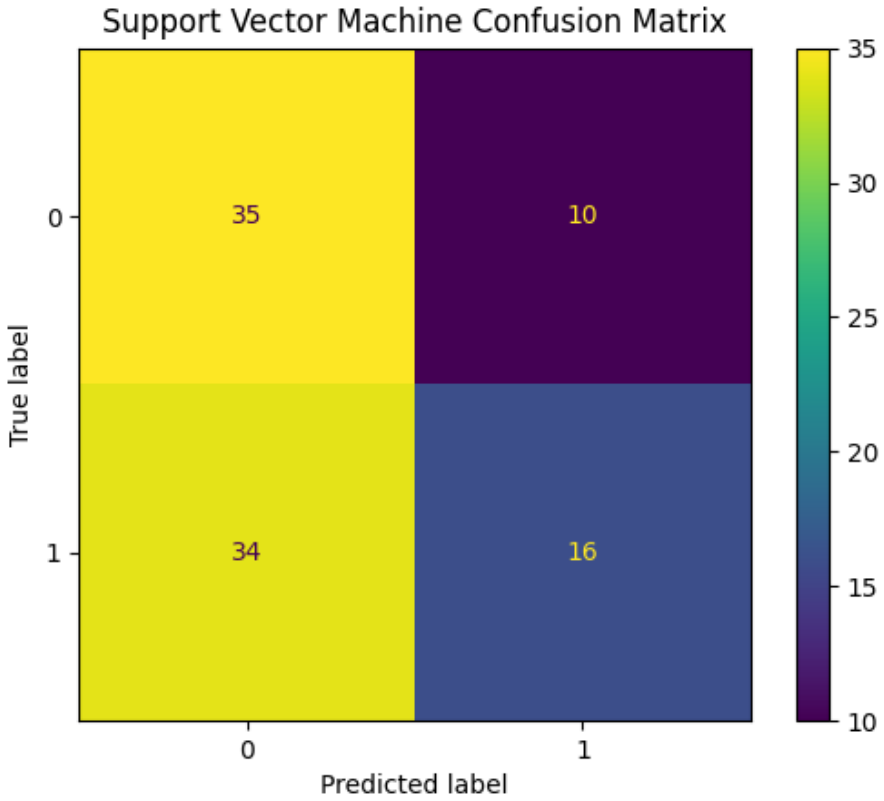
Εικόνα 38 Απεικόνιση Καμπύλης ROC Αριστερά του Εκτιμητή και Δεξιά των Εκτιμήσεων για τον Αλγόριθμο της Λογιστικής Παλινδρόμησης (Logistic Regression) (Default Parameters)



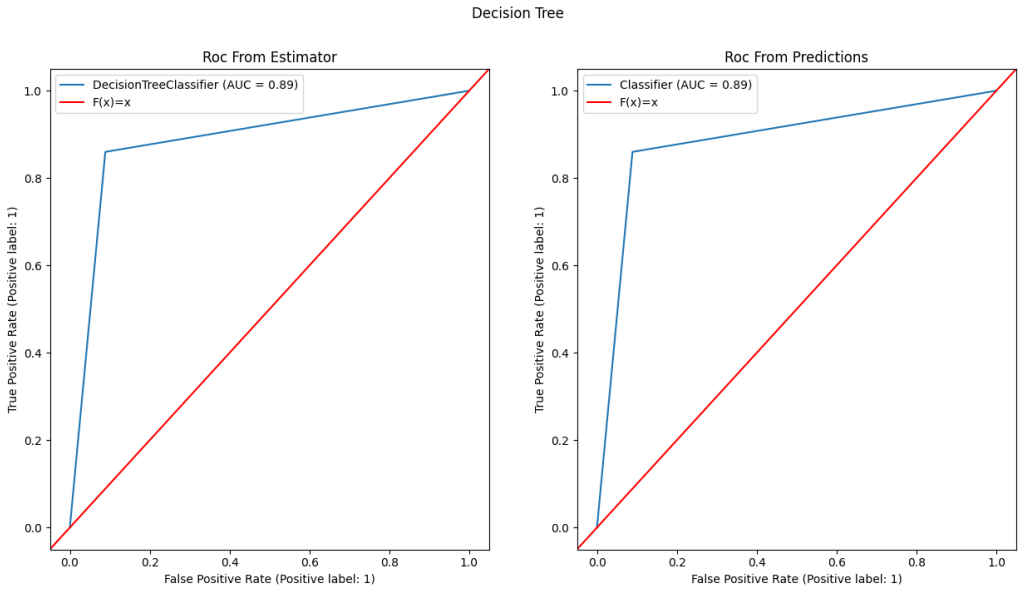
Εικόνα 39 Πίνακας Σύγχυσης Αποτελεσμάτων Λογιστικής Παλινδρόμησης (Logistic Regression) (Default Parameters)



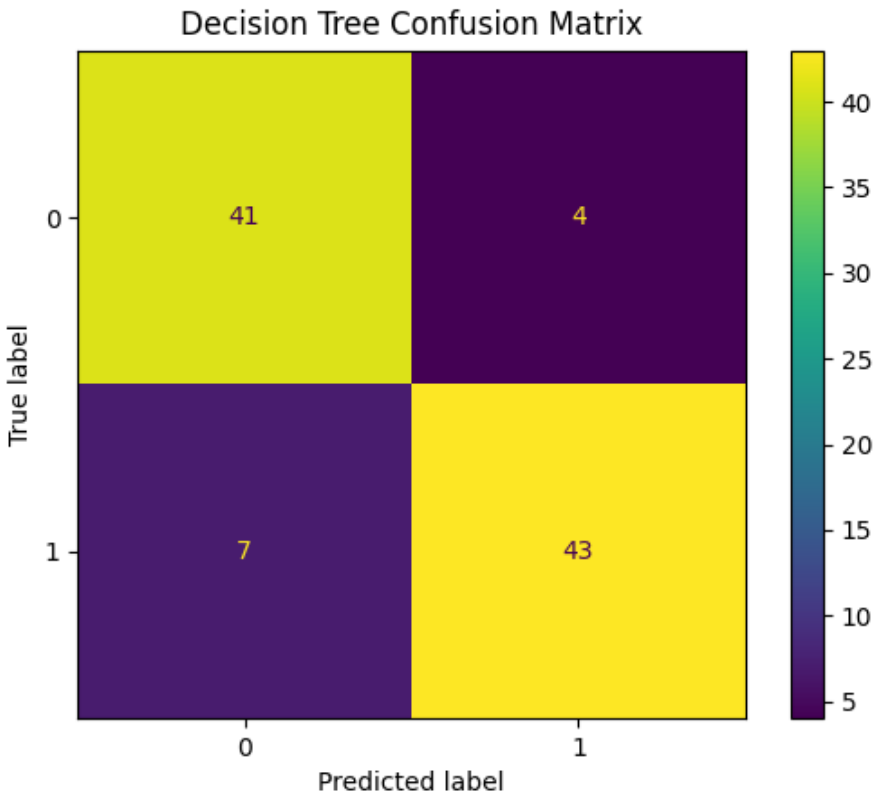
Εικόνα 40 Απεικόνιση Καμπύλης ROC Αριστερά του Εκτιμητή και Δεξιά των Εκτιμήσεων για τον Αλγόριθμο Διανυσμάτων Μηχανών Υποστήριξης (Support Vector Machine)(Default Parameters)



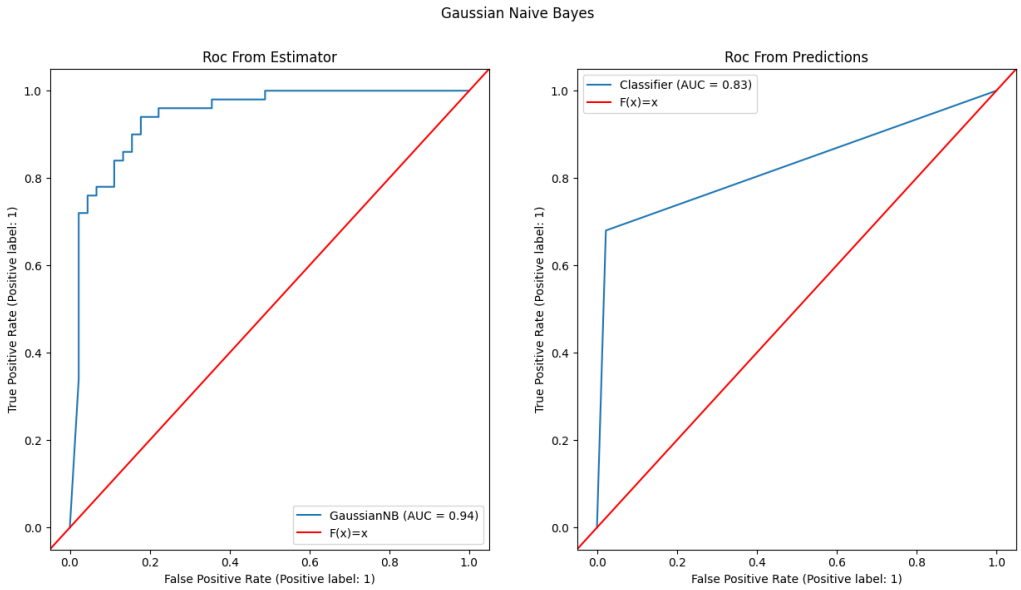
Εικόνα 41 Πίνακας Σύγχυσης Αποτελεσμάτων για τον Αλγόριθμο Διανυσμάτων Μηχανών Υποστήριξης (Support Vector Machine) (Default Parameters)



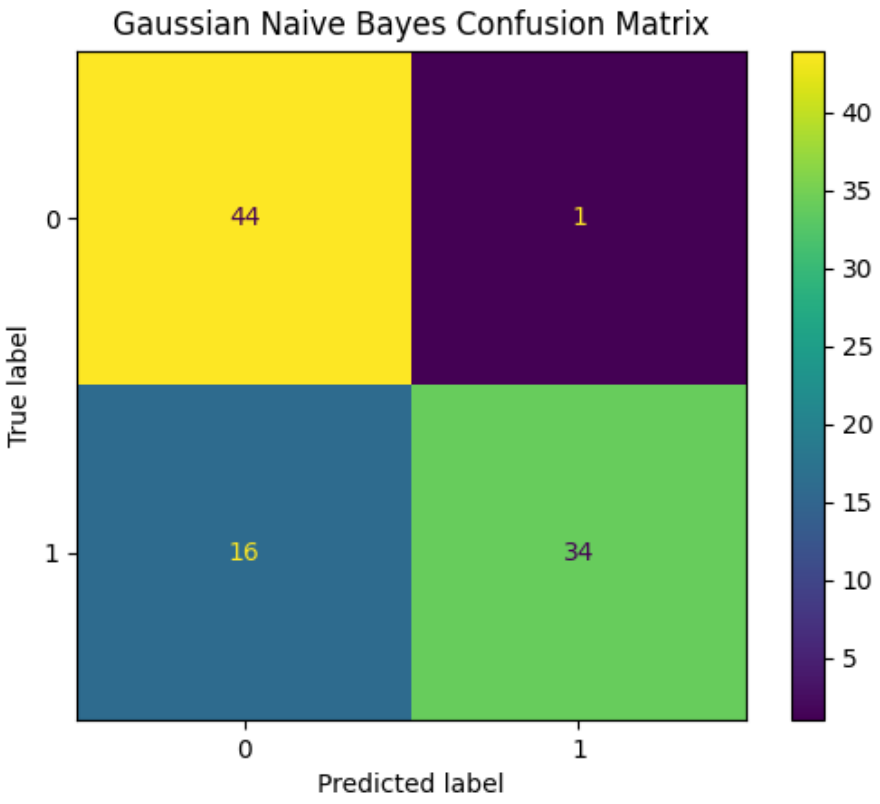
Εικόνα 42 Απεικόνιση Καμπύλης ROC Αριστερά του Εκτιμητή και Δεξιά των Εκτιμήσεων για τον Αλγόριθμο Δέντρων Αποφάσεων (Decision Trees) (Default Parameters)



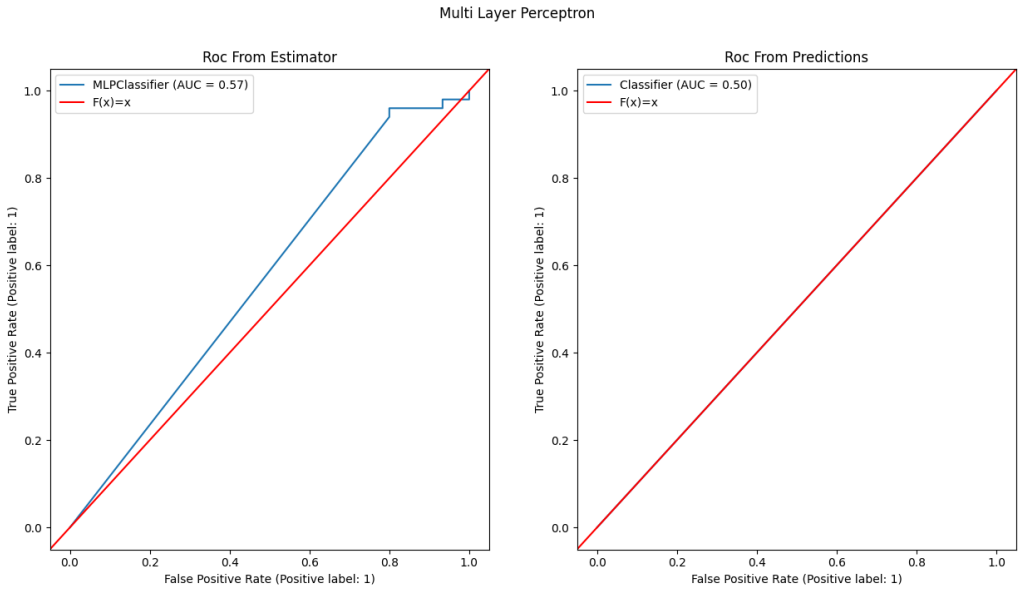
Εικόνα 43 Πίνακας Σύγχυσης Αποτελεσμάτων για τον Αλγόριθμο Δέντρων Αποφάσεων (Decision Trees) (Default Parameters)



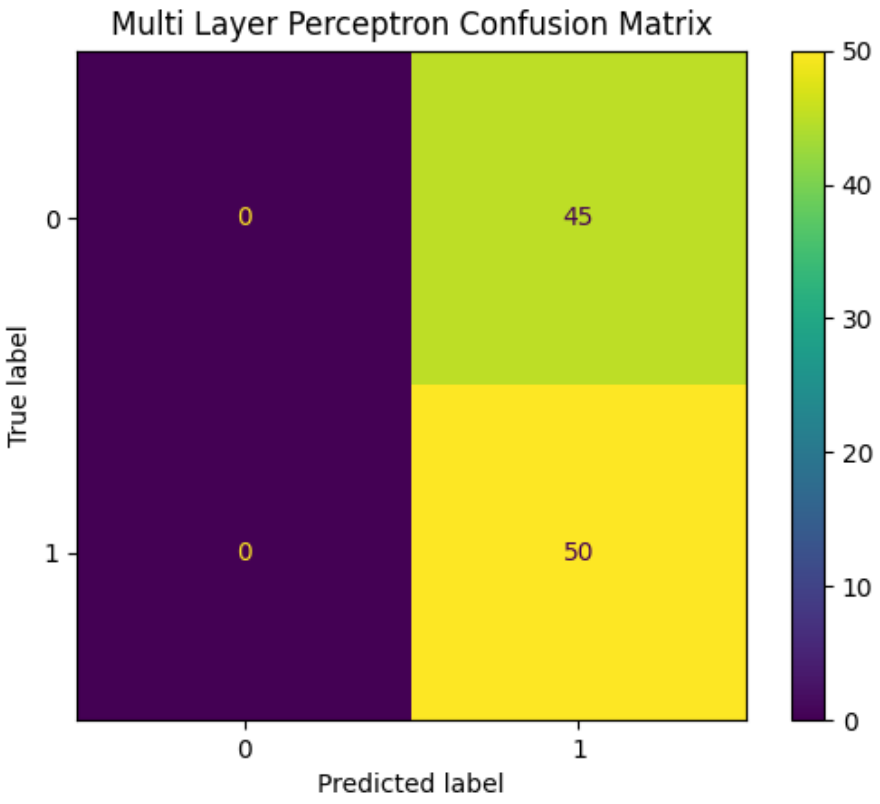
Εικόνα 44 Απεικόνιση Καμπύλης ROC Αριστερά του Εκτιμητή και Δεξιά των Εκτιμήσεων για τον Αλγόριθμο Gaussian Naive Bayes (Default Parameters)



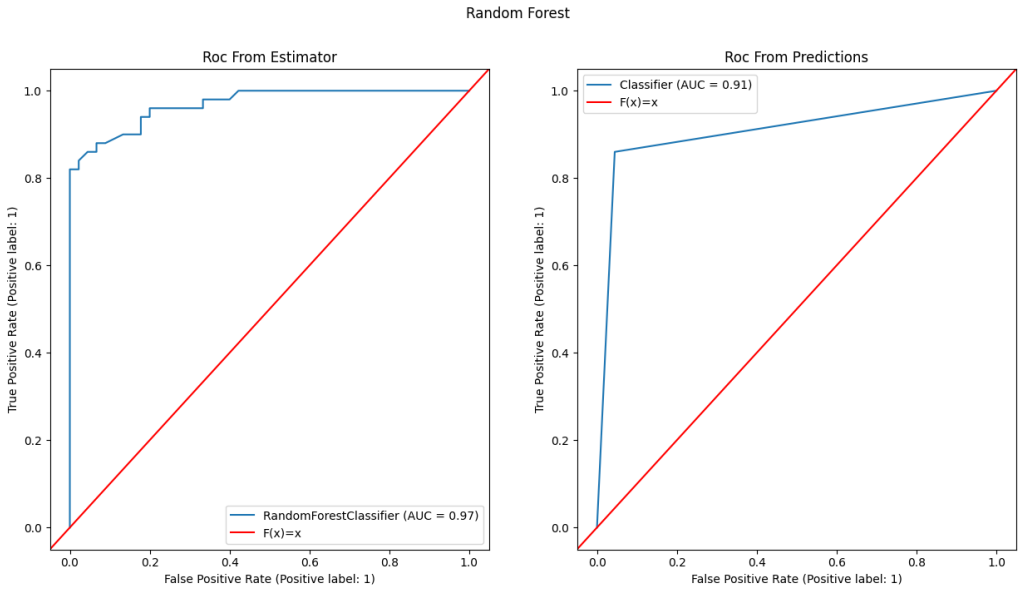
Εικόνα 45 Πίνακας Σύγχυσης Αποτελεσμάτων Gaussian Naive Bayes (Default Parameters)



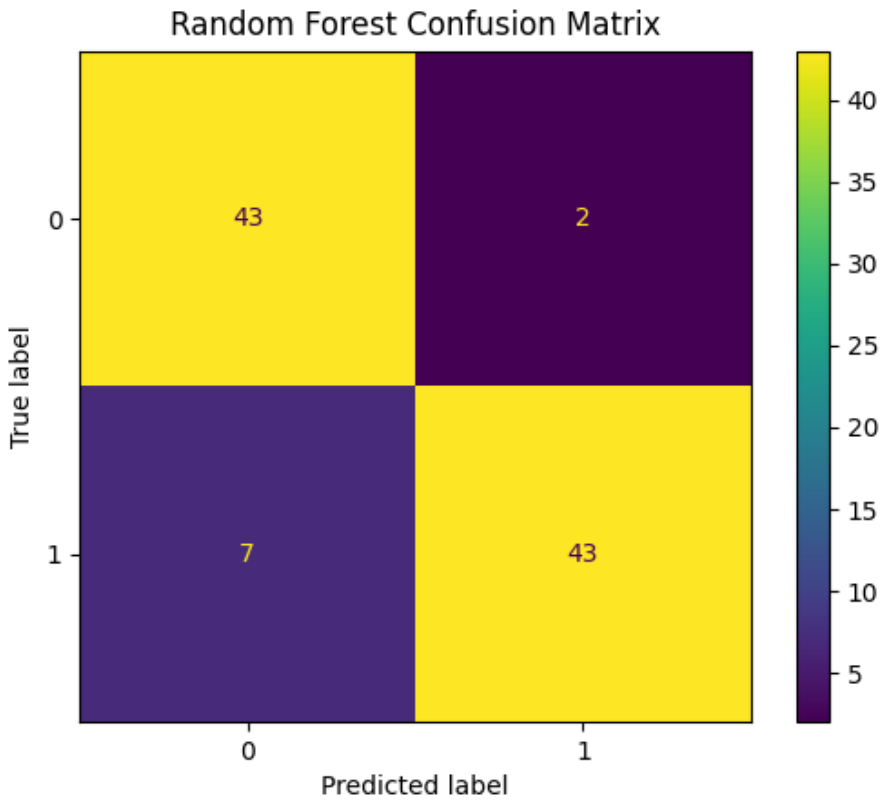
Εικόνα 46 Απεικόνιση Καμπύλης ROC Αριστερά του Εκτιμητή και Δεξιά των Εκτιμήσεων για τον Αλγόριθμο Πολύ-Επίπεδων Αισθητήρων (Multi-Layer Perceptron) (Default Parameters)



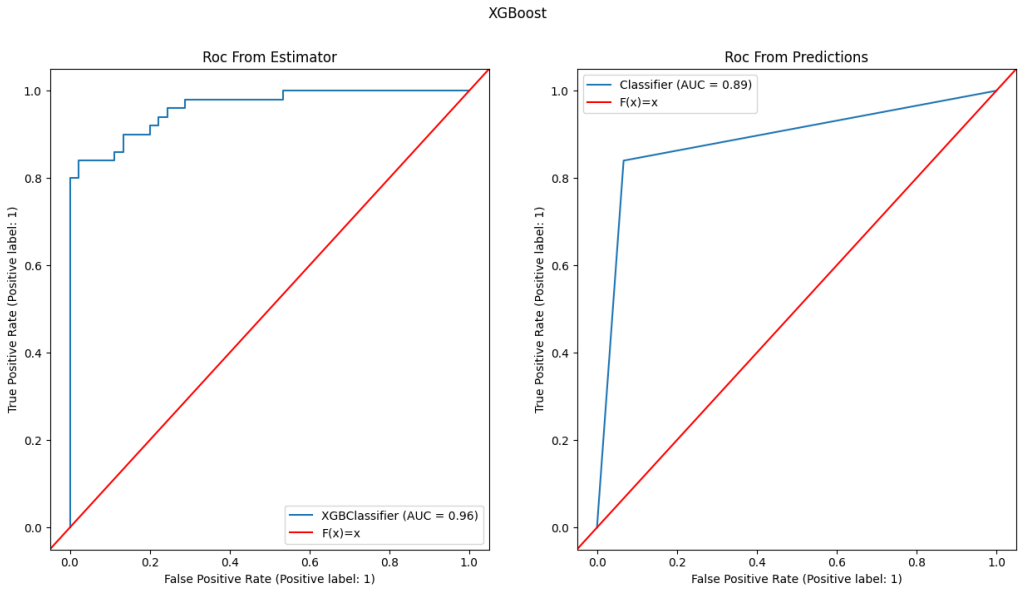
Εικόνα 47 Πίνακας Σύγχυσης Αποτελεσμάτων για τον Αλγόριθμο Πολύ-Επίπεδων Αισθητήρων (Multi-Layer Perceptron) (Default Parameters)



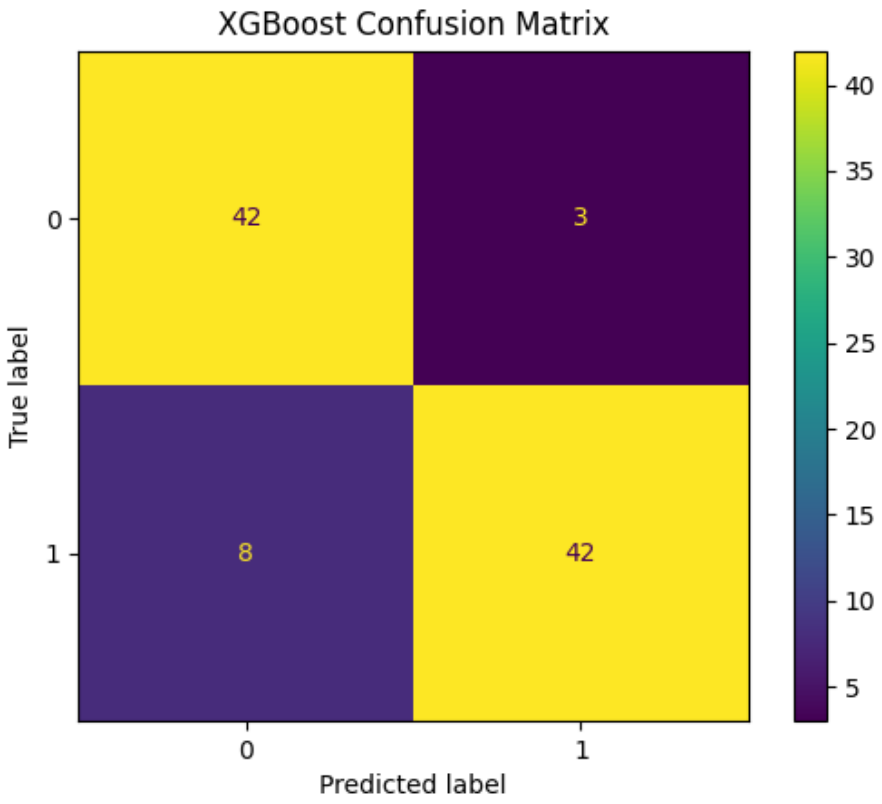
Εικόνα 48 Απεικόνιση Καμπύλης ROC Αριστερά του Εκτιμητή και Δεξιά των Εκτιμήσεων για τον Αλγόριθμο Τυχαίων Δασών (Random Forest) (Default Parameters)



Εικόνα 49 Πίνακας Σύγχυσης Αποτελεσμάτων για τον Αλγόριθμο Τυχαίων Δασών (Random Forest) (Default Parameters)



Εικόνα 50 Απεικόνιση Καμπύλης ROC Αριστερά του Εκτιμητή και Δεξιά των Εκτιμήσεων για τον Αλγόριθμο XGBoost (Default Parameters)

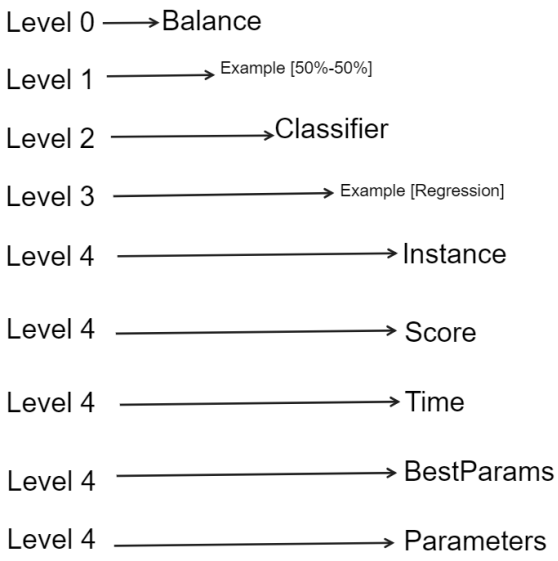


Εικόνα 51 Πίνακας Σύγχυσης Αποτελεσμάτων XGBoost (Default Parameters)

4.6.2 Ρύθμιση Υπέρ - Παραμέτρων

Για την ρύθμιση των παραμέτρων κάθε αλγορίθμου χρησιμοποιείται η μέθοδος GridSearch η οποία δεν κάνει τίποτα περισσότερο από μία εξοντωτική αναζήτηση όλων των πιθανών συνδυασμών των τιμών των παραμέτρων που περνάμε εμείς στον εκάστοτε αλγόριθμο. Τέλος η μέθοδος της αναζήτησης διατηρεί την παραμετροποίηση εκείνη που μεγιστοποιεί την τιμή κάποιας επιλεγμένης από εμάς μετρικής. Έτσι λοιπόν δημιουργήθηκε μία δομή τύπου dictionary όπου διατηρεί όλη την πληροφορία ανά αλγόριθμο για το ποιες τιμές των παραμέτρων μελετήθηκαν ποιες είναι οι καλύτερες τον χρόνο εκτέλεσης της διερεύνησης κλπ.

Το dictionary ακολουθεί την παρακάτω δομή της εικόνας 52.



Εικόνα 52 Δομή Dictionary

Στην περίπτωσή μας υπάρχουν πολλαπλά επίπεδα 3 όπου παίρνουν τις τιμές των 7 υπό εξέταση αλγορίθμων. Για κάθε ένα από αυτούς αρχικοποιείται κάτω από το κλειδί instance ο εκάστοτε αλγόριθμος για μεγαλύτερη ευκολία κατά την κλήση του. Συνεχίζοντας το κλειδί score αντιστοιχίζεται στην τιμή της συνολικής ακρίβειας (accuracy) που πέτυχε ο αλγόριθμος κατά την εκπαίδευσή του. Στο κλειδί time αντιστοιχεί ο χρόνος που δαπανήθηκε για την ολοκλήρωση της αναζήτησης των καλύτερων παραμέτρων. Στο κλειδί BestParams αποθηκεύονται οι καλύτερες παράμετροι και τέλος στο κλειδί Parameters αποθηκεύονται οι τιμές των παραμέτρων που εξετάσθηκαν.

Μία τέτοια τεχνική ενδείκνυται για τέτοιου είδους πειραματικές διαδικασίες καθώς υπάρχει η δυνατότητα αποθήκευσης του συγκεκριμένης δομής τύπου dictionary. Έτσι λοιπόν διατηρούνται οι καλύτεροι παράμετροι χωρίς να είναι απαραίτητη η εκ νέου αναζήτηση τους η οποία, είναι εξαιρετικά χρονοβόρα. Στην συνέχεια παρουσιάζεται το dictionary και οι παράμετροι όπως υπολογίστηκαν από την GridSearch.

Εύρος Παραμέτρων	Καλύτεροι Παράμετροι

"Regression": {
"instance": LogisticRegression(),
"Score": [],
"Time": [],
"BestParams": None,
"Parameters": {
"max_iter": [0.01, 0.1, 1, 10, 100, 10000],
"penalty": ["l2", "l1"],
"solver": ["liblinear", "lbfgs", "saga"],
"C": [0.001, 0.01, 0.1, 1, 10, 100, 1000]
}
}
	{
	'C': 10,
	'class_weight': None,
	'dual': False,
	'fit_intercept': True,
	'intercept_scaling': 1,
	'l1_ratio': None,
	'max_iter': 10,
	'multi_class': 'auto',
	'n_jobs': None,
	'penalty': 'l1',
	'random_state': None,
	'solver': 'liblinear',
	'tol': 0.0001,
	'verbose': 0,
	'warm_start': False
	}

"SupportvectorMachine": {
"instance": SVC(),
"Score": [],
"Time": [],
"BestParams": None,
"Parameters": {
"kernel": ["linear", "poly", "rbf",
"sigmoid"],
"C": [1, 10, 100, 1000],
"max_iter": [0.01, 0.1, 1, 10, 100, 1000,
10000, 100000],
"gamma": ["scale", "auto"]
}
}
	{
	'C': 1,
	'break_ties': False,
	'cache_size': 200,
	'class_weight': None,
	'coef0': 0.0,
	'decision_function_shape': 'ovr',
	'degree': 3,
	'gamma': 'scale',
	'kernel': 'linear',
	'max_iter': 100000,
	'probability': False,
	'random_state': None,
	'shrinking': True,
	'tol': 0.001,
	'verbose': False
	}

"GaussianNaiveBayes": {
"instance": GaussianNB(),
"Score": [],
"Time": [],
"BestParams": None,
"Parameters": {
'var_smoothing': np.logspace(0,-9, num=100)
}
}
	{
	'priors': None,
	'var_smoothing': 1e-09
	}

"Decisiontree": {
"instance": DecisionTreeClassifier(),
"Score": [],
"Time": [],
"BestParams": None,
"Parameters": {
"criterion": ["gini", "entropy", "log_loss"],
"splitter": ["best", "random"],
"max_depth": list(range(1, 30)),
"min_samples_split": [2, 3, 4, 5, 6, 7, 8, 9,
10],
"min_samples_leaf": [1, 2, 3],
"max_features": [None, "auto", "sqrt", "log2"]
}
}
	{
	'ccp_alpha': 0.0,
	'class_weight': None,
	'criterion': 'log_loss',
	'max_depth': 7,
	'max_features': None,
	'max_leaf_nodes': None,
	'min_impurity_decrease': 0.0,
	'min_samples_leaf': 1,
	'min_samples_split': 6,
	'min_weight_fraction_leaf': 0.0,
	'random_state': None,
	'splitter': 'random'
	}

"MultiLayerPerceptron": {
"instance": MLPClassifier(),
"Score": [],
"Time": [],
"BestParams": None,
"Parameters": {
"hidden_layer_sizes": [(50,10,), (50,20,),
(50,50,), (50,)],
	{
	'activation': 'tanh',
	'alpha': 0.001,
	'batch_size': 'auto',
	'beta_1': 0.9,
	'beta_2': 0.999,
	'early_stopping': False,
	'epsilon': 1e-08,
	'hidden_layer_sizes': (50, 50),

```
        "activation": ["identity", "logistic", "tanh",
"relu"],
        "solver": ["lbfgs", "sgd", "adam"],
        "alpha": [0.0001, 0.00001, 0.001],
        "learning_rate": ["constant", "invscalling",
"adaptive"],
        "learning_rate_init": [0.001, 0.01, 0.0001],
        "max_iter": [200, 220, 250]
    }
}
```

```
    'learning_rate': 'adaptive',
    'learning_rate_init': 0.0001,
    'max_fun': 15000,
    'max_iter': 250,
    'momentum': 0.9,
    'n_iter_no_change': 10,
    'nesterovs_momentum': True,
    'power_t': 0.5,
    'random_state': None,
    'shuffle': True,
    'solver': 'lbfgs',
    'tol': 0.0001,
    'validation_fraction': 0.1,
    'verbose': False,
    'warm_start': False
}
```

```
"RandomForest": {
    "instance": RandomForestClassifier(),
    "Score": [],
    "Time": [],
    "BestParams": None,
    "Parameters": {
        "n_estimators": [10, 50, 100, 150, 200],
        "criterion": ["gini", "entropy", "log_loss"],
        "max_features": ["auto", "sqrt", "log2",
None],
        "max_depth": [5, 10, 20, 30, 40]
    }
}
```

```
{
    'bootstrap': True,
    'ccp_alpha': 0.0,
    'class_weight': None,
    'criterion': 'entropy',
    'max_depth': 5,
    'max_features': None,
    'max_leaf_nodes': None,
    'max_samples': None,
    'min_impurity_decrease': 0.0,
    'min_samples_leaf': 1,
    'min_samples_split': 2,
    'min_weight_fraction_leaf': 0.0,
    'n_estimators': 100,
    'n_jobs': None,
    'oob_score': False,
    'random_state': None,
    'verbose': 0,
    'warm_start': False
}
```

```
"XGBoost": {
    "instance":
xgboost.XGBClassifier(objective='binary:logistic'),
    "Score": [],
    "Time": [],
    "BestParams": None,
    "Parameters": {
        "booster": ["gbtree", "gblinear"],
        "eta": [0.1, 0.3, 0.5, 0.8],
        "max_depth": [2, 4, 6, 8, 10],
        "subsample": [0.25, 0.5, 0.75, 1],
        "lambda": [0, 1],
        "alpha": [0, 1],
        "grow_policy": ["depthwise", "lossguide"],
        "num_parallel_tree": [7]
    }
}
```

```
{
    'alpha': 0,
    'base_score': None,
    'booster': 'gblinear',
    'callbacks': None,
    'colsample_bylevel': None,
    'colsample_bynode': None,
    'colsample_bytree': None,
    'early_stopping_rounds': None,
    'enable_categorical': False,
    'eta': 0.8,
    'eval_metric': None,
    'feature_types': None,
    'gamma': None,
    'gpu_id': None,
    'grow_policy': 'depthwise',
    'importance_type': None,
    'interaction_constraints': None,
    'lambda': 0,
    'learning_rate': None,
    'max_bin': None,
    'max_cat_threshold': None,
    'max_cat_to_onehot': None,
    'max_delta_step': None,
    'max_depth': 2,
    'max_leaves': None,
    'min_child_weight': None,
    'missing': nan,
    'monotone_constraints': None,
    'n_estimators': 100,
    'n_jobs': None,
    'num_parallel_tree': 1,
```

	'objective': 'binary:logistic', 'predictor': None, 'random_state': None, 'reg_alpha': None, 'reg_lambda': None, 'sampling_method': None, 'scale_pos_weight': None, 'subsample': 0.25, 'tree_method': None, 'use_label_encoder': None, 'validate_parameters': None, 'verbosity': None }

Πίνακας 10 Συγκριτικός Πίνακας Υπέρ Παραμέτρων Αλγορίθμων

Για την εύρεση των καλύτερων παραμέτρων για άλλη μια φορά χρησιμοποιήθηκε η τυχαία υπό – δειγματοληψία. Έτσι λοιπόν με μόλις ένα σύνολο εκπαίδευσης και άλλο ένα δοκιμής καταλήξαμε στα αποτελέσματα των παραμέτρων του πίνακα 10. Δυστυχώς δεν ήταν δυνατό να πραγματοποιηθεί μία πιο εξαντλητική διερεύνηση χρησιμοποιώντας όλα τα πιθανά υποσύνολα (διατηρώντας πάντα το 50 -50 μεταξύ των νόμιμων και των απατηλών συναλλαγών), αλλά το γεγονός ότι είναι τυχαίο δείγμα μας δίνει μία καλή κατεύθυνση για το που πρέπει να κινηθούν οι παράμετροι και για τις υπόλοιπες συναλλαγές.

Εκτιμητής (Classifier)	Υπέρ – Παράμετροι (Hyper Parameters)	Κλάση	Precision		Recall		F1 - Score		Support
Logistic Regression	Προκαθορισμένο (Default)	0	-0.02	0.87	0.02	0.91	0	0.89	45
		1	0.02	0.92	-0.02	0.88	0	0.9	50
Support Vector Machine	Προκαθορισμένο (Default)	0	0.1	0.61	-0.34	0.44	-0.1	0.51	45
		1	-0.02	0.6	0.42	0.74	0.24	0.66	50
Decision Tree	Προκαθορισμένο (Default)	0	-0.03	0.83	-0.04	0.89	-0.03	0.86	45
		1	-0.04	0.89	-0.02	0.84	-0.03	0.87	50
Gaussian Naïve Bayes	Προκαθορισμένο (Default)	0	0	0.73	0	0.98	0	0.84	45
		1	0	0.97	0	0.68	0	0.8	50
Multi-Layer Perceptron	Προκαθορισμένο (Default)	0	0.83	0.83	0.89	0.89	0.86	0.86	45
		1	0.36	0.89	-0.16	0.84	0.18	0.87	50
Random Forest	Προκαθορισμένο (Default)	0	-0.03	0.83	0.02	0.98	-0.01	0.9	45
		1	0.02	0.98	-0.04	0.82	-0.02	0.89	50
XGBoost	Προκαθορισμένο (Default)	0	-0.01	0.83	0.03	0.96	0.01	0.89	45
		1	0.02	0.95	-0.02	0.82	0	0.88	50

Πίνακας 11 Συγκριτικός Πίνακας Μετρικών Κλάσεων Κατηγοριοποίησης (Βέλτιστοι Παράμετροι)

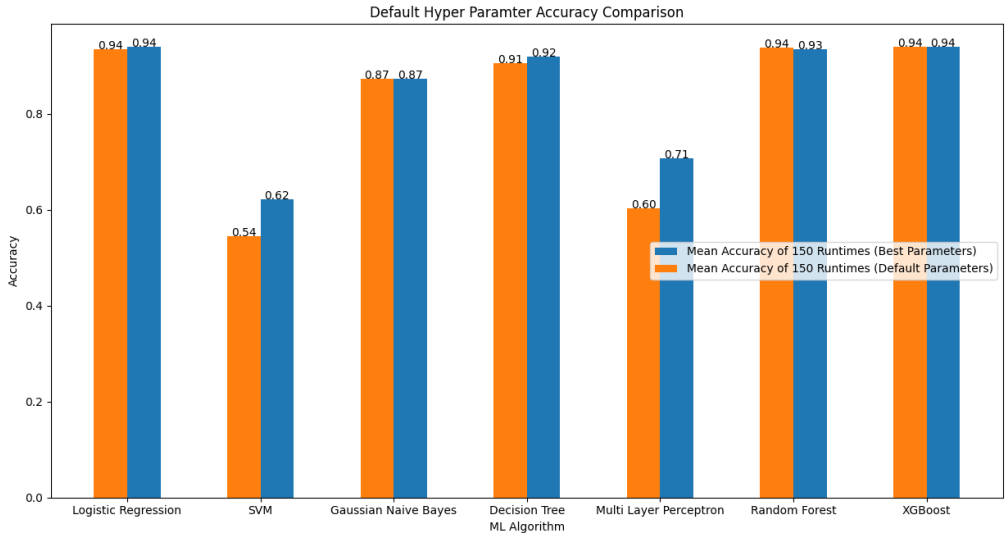
Στον πίνακα 11 βλέπουμε τις μεταβολές των μετρικών μετά την παραμετροποίηση των αλγορίθμων μηχανικής μάθησης. Μετά την παραμετροποίηση των αλγορίθμων εμφανώς βελτιωμένος εντοπίζεται ο αλγόριθμος των πολύ-επίπεδων αισθητήρων (MLP) και ο αλγόριθμος των διανυσμάτων μηχανών υποστήριξης (Support Vector Machine) όπου το F1-score σημείωσε σημαντική άνοδο. Τη μεγαλύτερη πτώση της τάξης του 3% φαίνεται ότι τη σημείωσε ο Decision Tree.

Ομοίως με τη διαδικασία που ακολουθήθηκε και με τις προκαθορισμένες (default) παραμέτρους πραγματοποιείται μία επαναληπτική διαδικασία εκπαίδευσης των αλγορίθμων με 150 τυχαία δείγματα. Αυτά τα δείγματα έχουν συνεχώς σταθερό μήκος 946 παρατηρήσεων όπου οι 473 είναι συναλλαγές απάτης και οι υπόλοιπες είναι νόμιμες. Στην ουσία πραγματοποιείται τυχαία δειγματοληψία για το υποσύνολο των νόμιμων παρατηρήσεων. Μετά το πέρας της επαναληπτικής διαδικασίας είμαστε σε θέση να συγκρίνουμε την συνεισφορά των παραμέτρων στις μέσες τιμές της συνολικής ακρίβειας (accuracy).

Εκτιμητής (Classifier)	Υπέρ – Παράμετροι (Hyper Parameters)	Μέση Ακρίβεια (Mean Accuracy) %
Logistic Regression	Βέλτιστοι (Best)	94.00 %
Support Vector Machine	Βέλτιστοι (Best)	62.00 %
Gaussian Naïve Bayes	Βέλτιστοι (Best)	87.00 %
Decision Tree	Βέλτιστοι (Best)	92.00 %
Multi-Layer Perceptron	Βέλτιστοι (Best)	71.00 %
Random Forest	Βέλτιστοι (Best)	93.00 %
XGBoost	Βέλτιστοι (Best)	94.00 %

Πίνακας 12 Συγκριτικός Πίνακας Ακρίβειας Αλγορίθμων Βέλτιστων Υπέρ – Παραμέτρων Για 150 Διαφορετικές Περιπτώσεις Τυχαίου Δείγματος

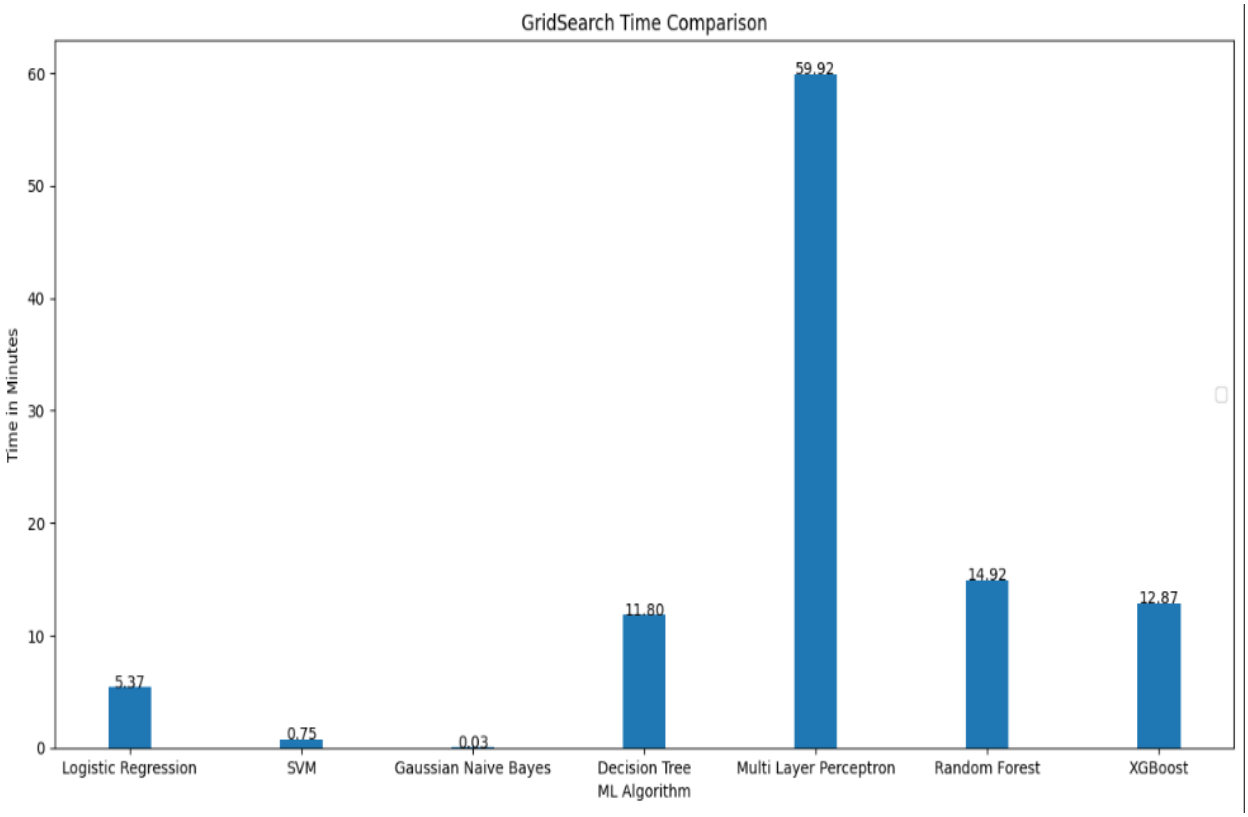
Αυτή η ρύθμιση των παραμέτρων των αλγορίθμων φαίνεται να λειτούργησε προς όφελος των αλγορίθμων Support Vector Machine, Decision Tree και Multi-Layer Perceptron. Ειδικότερα συγκριτικά με τον πίνακα 8 όπου παρουσιάστηκε η απόδοση του κάθε αλγορίθμου με τις προκαθορισμένες παραμέτρους έχουμε μία αύξηση στη μέση ακρίβεια (accuracy) της τάξης του 8% για τον SVM, 1% για το Decision Tree και 12% για τον MLP. Η σχηματική απεικόνιση των τιμών των πινάκων 8 και 12 δίνει σαφέστατη εικόνα για τις μεταβολές και το πόσο ωφέλησε τους αλγορίθμους αυτή η διαδικασία της εύρεσης των βέλτιστων παραμέτρων.



Εικόνα 53 Απεικόνιση Σύγκρισης Μέσης Ακρίβειας για 150 Διαφορετικές Περιπτώσεις Τυχαίου Δείγματος

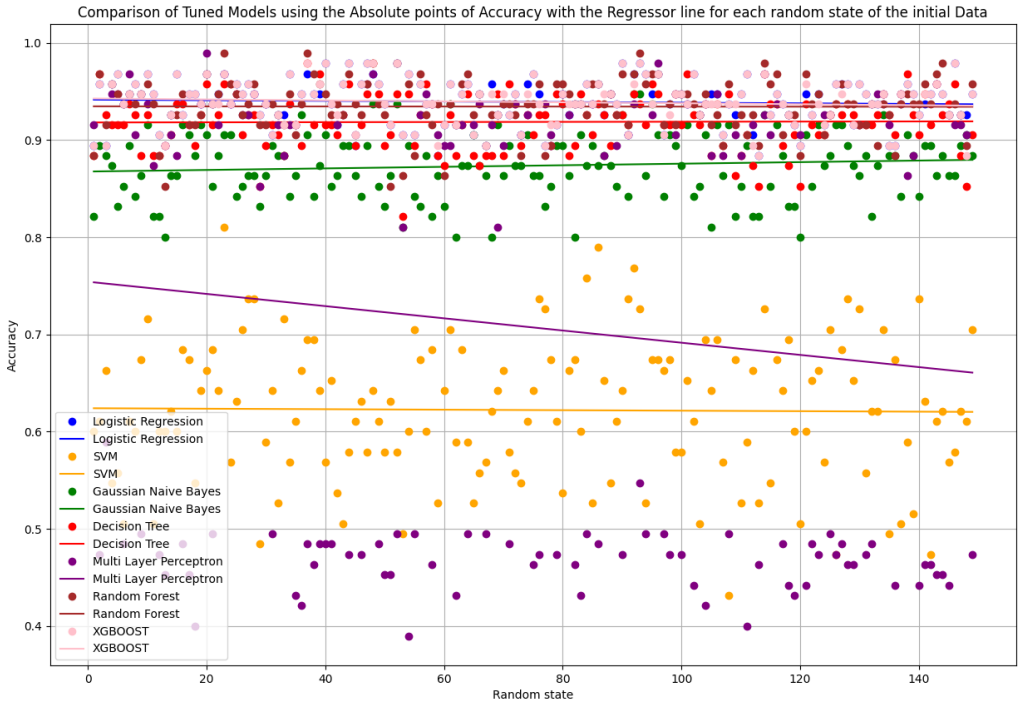
Δεδομένων των υπολογιστικών πόρων και του προβλήματος που καλούμαστε να απαντήσουμε η διαφορά μετά την ρύθμιση των παραμέτρων δεν είναι σημαντική. Επιπλέον πραγματοποιώντας τη ρύθμιση των παραμέτρων με μόλις 946 παρατηρήσεις από τις 300,000 μπορεί να μην αντιπροσωπεύει το πληθυσμό, καθώς είναι πολύ πιθανό το τυχαίο δείγμα που θα συνταχθεί να μην μεταφέρει όλες τις ιδιαιτερότητες του συνόλου. Καθώς το ζητούμενό μας είναι να βρούμε εκείνο το μοντέλο που με μεγαλύτερη πιθανότητα θα κατηγοριοποιήσει σωστά μία παρατήρηση είναι κατανοητό ότι οι MLP, SVM και ίσως και ο Gaussian Naïve Bayes έχουν σημαντικό μειονέκτημα έναντι των υπολοίπων εκτιμητών. Συνεπώς τα μοντέλα που επωφελήθηκαν περισσότερο από τη ρύθμιση των παραμέτρων παραμένουν σημαντικά πιο αδύναμοι.

Αυτό που ίσως αξίζει να σημειωθεί, είναι το πόσο «εύπλαστος» είναι ο αλγόριθμος των πολύ-επίπεδων αισθητήρων (MLP) καθώς υπάρχει πληθώρα παραμέτρων που μπορεί να δεχτεί και να βελτιώσει άμεσα την αποδοτικότητα του. Βέβαια αυτό το θετικό του, συνοδεύεται από ένα μεγάλο μειονέκτημα. Ο χρόνος εκπαίδευσης αλλά και αναζήτησης της βέλτιστης παραμετροποίησης είναι σημαντικά μεγαλύτερος από κάθε άλλο αλγόριθμο που εξετάστηκε. Ενδεικτικά παρουσιάζεται το επόμενο σχήμα όπου αναπαριστάτε ο χρόνος που δαπανήθηκε στην αναζήτηση των βέλτιστων παραμέτρων. Ο MLP χρειάστηκε περίπου μία ώρα για την εύρεση των παραμέτρων όταν ο δεύτερος πιο δαπανηρός σε χρόνο αλγόριθμος χρειάστηκε μόλις δεκαπέντε λεπτά.



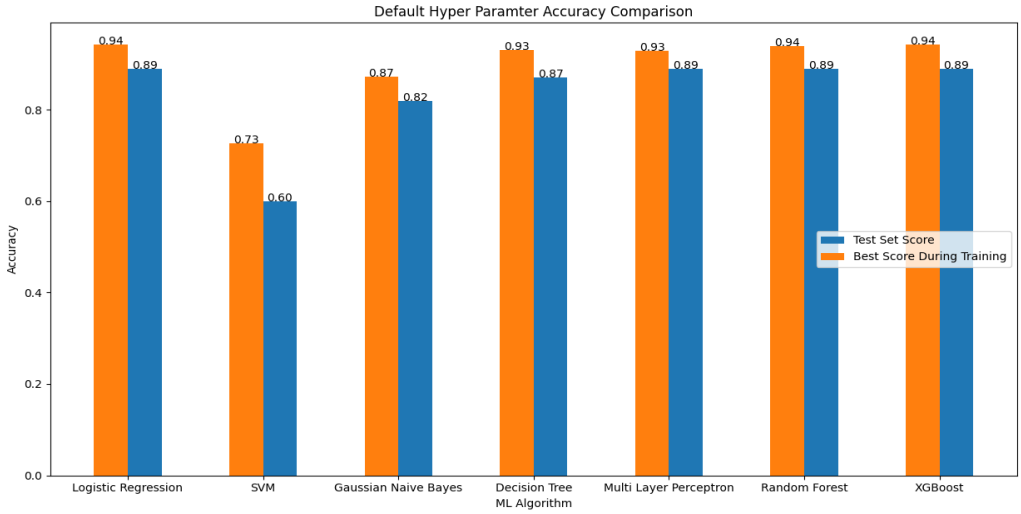
Εικόνα 54 Απεικόνιση Σύγκρισης Χρόνου Αναζήτησης Βέλτιστων Παραμέτρων

Παρά το γεγονός ότι βελτιώθηκε η αποτελεσματικότητα κάποιων αλγορίθμων το επόμενο διάγραμμα απεικόνισης των σημείων της ακρίβειας για κάθε μία από τις 150 επαναλήψεις έρχεται να αποδείξει το γεγονός ότι το σύνολο που χρησιμοποιήθηκε για την εύρεση των παραμέτρων δεν αντιπροσωπεύει το σύνολο των δεδομένων. Στους αλγορίθμους των πολύ-επίπεδων αισθητήρων και διανυσμάτων μηχανών υποστήριξης είναι πιο ορατό από την γραμμή παλινδρόμησης που έχει φθίνουσα τάση καθώς αυξάνονται οι επαναλήψεις. Αυτό δηλώνει ότι η αποτελεσματικότητα του αλγορίθμου αυξήθηκε σημαντικά με την ρύθμιση των υπέρ – παραμέτρων. Όμως καθώς τα σύνολα διαφοροποιούνται όλο και περισσότερο η αποτελεσματικότητά του φθίνει. Στη συγκεκριμένη περίπτωση μπορεί να γίνει λόγος για overfitting του αλγορίθμου στο υπό διερεύνηση (το πρώτο υποσύνολο) υποσύνολο εκπαίδευσης. Καταλήγουμε στο συμπέρασμα ότι η διερεύνηση για την εύρεση των βέλτιστων παραμέτρων δεν μπορεί να γίνει παίρνοντας ένα τόσο μικρό σύνολο καθώς λείπει σημαντικό μέρος της πληροφορίας προκειμένου να ολοκληρωθεί με επιτυχία η συγκεκριμένη διαδικασία.



Εικόνα 55 Σημειακή Απεικόνιση των Τιμών Ακρίβειας του Κάθε Τυχαίου Συνόλου (Βέλτιστοι Παράμετροι)

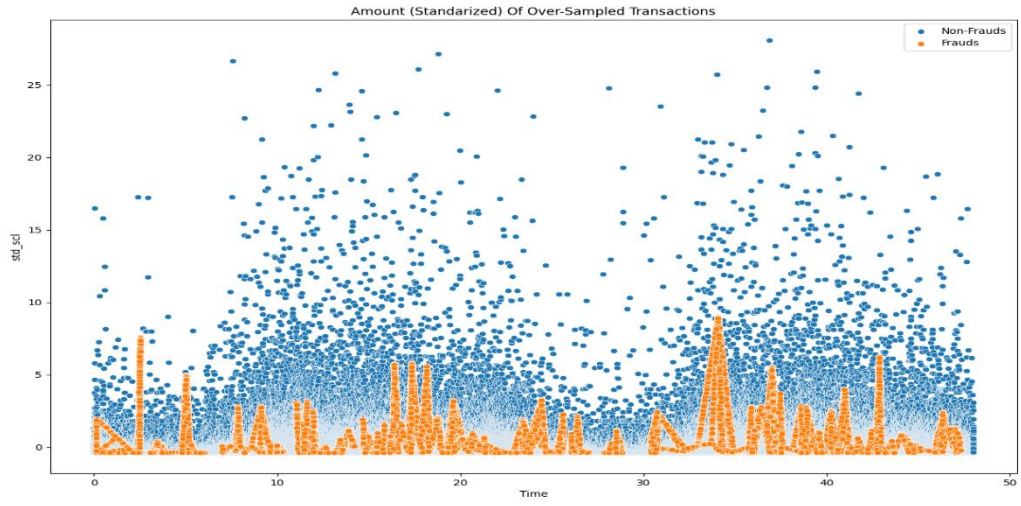
Τέλος ένα ακόμα διάγραμμα το οποίο ενισχύει την υπόθεση της υπέρ - προσαρμογής είναι αυτό της εικόνας 56. Στο συγκεκριμένο διάγραμμα απεικονίζεται η σύγκριση των καλύτερων εκτιμήσεων κατά το στάδιο της εκπαίδευσης (οι οποίες όρισαν και τις παραμέτρους του εκάστοτε αλγορίθμου) με την ακρίβεια που πέτυχε ο κάθε ένας στο σύνολο δοκιμής.



Εικόνα 56 Απεικόνιση Σύγκρισης Εκτιμήσεων Κατά το Στάδιο της Εκπαίδευσης και Δοκιμής

4.6.3 Υπέρ – δειγματοληψία (SMOTE)

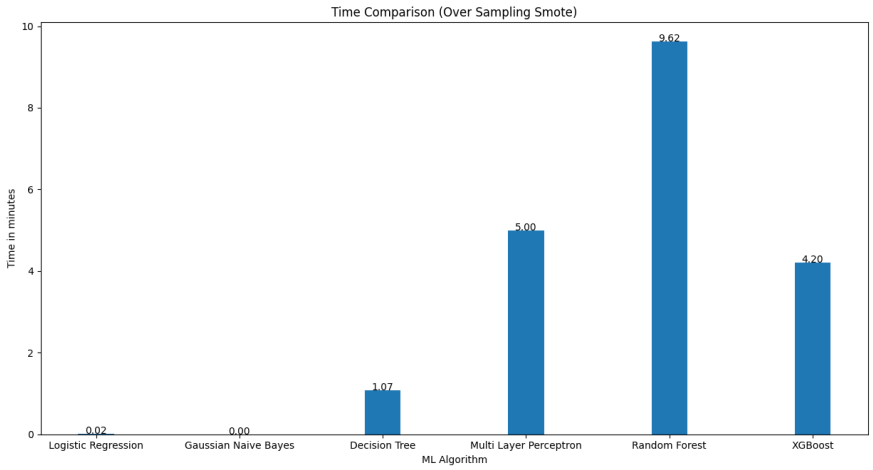
Σε αυτή την ενότητα θα πραγματοποιηθεί υπέρ – δειγματοληψία της κλάσης μειοψηφίας χρησιμοποιώντας τη μέθοδο SMOTE αποσκοπώντας στην εκμετάλλευση όλης της διαθέσιμης πληροφορίας που παρέχουν οι παρατηρήσεις του συνόλου δεδομένων. Έτσι λοιπόν από το αρχικό σύνολο δεδομένων των 283.224 νόμιμων συναλλαγών και 473 συναλλαγών απάτης δημιουργείται ένα σύνολο το οποίο περιέχει ίσο αριθμό παρατηρήσεων των δύο κλάσεων συνολικού αριθμού 566.448. Λόγω της δραματικής αύξησης των παρατηρήσεων (οριακά διπλασιασμό τους) θα δούμε σημαντικά αυξημένους χρόνους εκπαίδευσης των αλγορίθμων. Η διαφορά γίνεται αντιληπτή από το γράφημα της εικόνας 57 όπου παρατηρείται το πόσο διαφοροποιήθηκε η κατανομή της κλάσης μειοψηφίας συγκρίνοντάς τη με αυτή της εικόνας 30.



Εικόνα 57 Απεικόνιση Ύψους Κανονικοποιημένου Amount ως Προ το Χρόνο Μετά την SMOTE

Στη συνέχεια πραγματοποιείται διαχωρισμός σε σύνολο εκπαίδευσης και σύνολο δοκιμής αναλογίας 90-10 κατ' αντιστοιχία με την προηγούμενη διερευνητική διαδικασία. Όπως ήταν λογικό οι χρόνοι εκπαίδευσης αλλά και δοκιμής αυξήθηκαν σημαντικά για όλους τους αλγορίθμους. Επίσης όσον αφορά τον αλγόριθμο διανυσμάτων υποστήριξης, ήταν αδύνατη η εκπαίδευσή του μέσω του περιβάλλοντος του collab καθώς ο χρόνος εκπαίδευσής του ξεπερνούσε τις δυόμιση ώρες και το σύστημα διέκοπτε συνεχώς της διαδικασία. Για αυτό το λόγο δεν θα μπορέσουμε να συμπεριλάβουμε στην διαδικασία σύγκρισης το συγκεκριμένο αλγόριθμο λόγω του περιορισμού των διαθέσιμων υπολογιστικών πόρων.

Στην εικόνα 58 παρουσιάζονται οι χρόνοι εκπαίδευσης των 6 αλγορίθμων όπως αυτοί υπολογίστηκαν για το επαυξημένο σύνολο δεδομένων μετά την υπέρ – δειγματοληψία της κλάσης μειοψηφίας.



Εικόνα 58 Απεικόνιση Χρόνων Εκπαίδευσης Των Αλγορίθμων (SMOTE)

Εκτιμητής (Classifier)	Παράμετροι	Κλάση	Train Time	Accuracy	Precision	Recall	F1-Score	Support
Logistic Regression	(Default)	0	0:02	0.957	0.95	0.97	0.96	28327
		1			0.97	0.95	0.96	28318
Support Vector Machine	(Default)	0						
		1						
Decision Tree	(Default)	0	1:05	0.998	1	1	1	28327
		1			1	1	1	28318
Gaussian Naïve Bayes	(Default)	0	0:01	0.871	0.8	0.99	0.89	28327
		1			0.99	0.75	0.85	28318
	(Default)	0	5:00	0.975	0.96	0.99	0.98	28327

Multi-Layer Perceptron		1			0.96	0.96	0.98	28318
Random Forest	(Default)	0	9:37	0.999	1	1	1	28327
		1			1	1	1	28318
XGBoost	(Default)	0	4:12	0.998	1	1	1	28327
		1			1	1	1	28318

Πίνακας 13 Πίνακας Αποτελεσμάτων Μετρικών Αξιολογήσεων (SMOTE)

Παράλληλα σημειώνεται σημαντική αύξηση των επιδόσεων των μετρικών αξιολόγησης όλων σχεδόν των αλγορίθμων. Όπως φαίνεται και στους παρακάτω πίνακες η γενική συνολική ακρίβεια (accuracy) δεν πέφτει κάτω από το 87% του Gaussian Naïve bayes. Οι αποδοτικότεροι αλγόριθμοι είναι κατά σειρά προτεραιότητας είναι αυτοί των τυχαίων δασών, XGBoost, δέντρων αποφάσεων, πολύ-επίπεδων αισθητήρων και τέλος της λογιστικής παλινδρόμησης.

Οι διαφορές ως προς τη συνολικής ακρίβεια (accuracy) των τριών πρώτων αλγορίθμων είναι πολύ μικρές και όσον αφορά το στρογγυλοποιημένο f1-score δεν παρατηρείται καμία. Ρίχνοντας μία πιο προσεκτική ματιά στους πίνακες σύγχυσης των εκτιμητών του συγκεντρωτικού πίνακα 14 βλέπουμε ότι οι αλγόριθμοι τυχαίων δασών και XGBoost δεν τους «ξέφυγε» καμία συναλλαγή απάτης και το μόνο λάθος του τυχαίων δαών ήταν το γεγονός ότι κατηγοριοποίησε 6 συναλλαγές ως απάτες ενώ δεν ήταν (6 λανθασμένα θετικές εκτιμήσεις). Ο επόμενος αλγόριθμος ως προς τη συνολική ακρίβεια (accuracy) XGBoost είχε μόλις 9 λανθασμένα θετικές εκτιμήσεις ενώ το δέντρο απόφασης με μία επίσης πολύ καλή επίδοση είχε 21 λανθασμένα αρνητικές και 59 λανθασμένα θετικές εκτιμήσεις. 0 αριθμός αυτός των 80 λανθασμένων παρατηρήσεων είναι πολύ μικρός (σχεδόν μηδενικός) σε σχέση με το αρχικό σύνολο δεδομένων.

Regression

Confusion Matrix			
		Predicted	
		1	0
Actual	1	26814	1504
	0	904	27423

Multi-Layer Perceptron

Confusion Matrix			
		Predicted	
		1	0
Actual	1	27123	1195
	0	187	28140

Gaussian Naïve Bayes

Confusion Matrix			
		Predicted	
		1	0
Actual	1	21261	7057
	0	222	28105

Random Forest

Confusion Matrix			
		Predicted	
		1	0
Actual	1	28318	0
	0	6	28321

Decision Tree

Confusion Matrix			
		Predicted	
		1	0
Actual	1	28297	21
	0	59	28268

XGBoost

Confusion Matrix			
		Predicted	
		1	0
Actual	1	28318	0
	0	9	28318

Πίνακας 14 Συγκριτικός Πίνακας Confusion Matrix

4.6.4 Εφαρμογή Διασταυρούμενης Επικύρωσης Κ-φορών (K-Fold Cross Validation)

Τελευταία μέθοδο αξιολόγησης αυτής της πειραματικής διαδικασίας είναι αυτή της διασταυρούμενης επικύρωσης κ-φορών μέσω της οποίας το σύνολο εκπαίδευσης χωρίζεται σε Κ τμήματα εκ των οποίων τα Κ-1 τμήματα χρησιμοποιούνται ως σύνολο εκπαίδευσης και το σύνολο που δεν χρησιμοποιήθηκε στην εκπαίδευση του αλγορίθμου το εκμεταλλευόμαστε ως σύνολο δοκιμής. Με αυτό το τρόπο αποφεύγουμε το φαινόμενο της υπέρ – προσαρμογής πράγμα που είδαμε να μας προβληματίζει στα πλαίσια αυτής της εργασίας κατά της διαδικασία τις υπό – δειγματοληψίας.

Για την εφαρμογή της συγκεκριμένης μεθόδου επιστρατεύεται η στρατηγική διασταυρούμενη επικύρωση κ-φορών η οποία εξασφαλίζει σε κάθε τμήμα της διάσπασης του συνόλου εκπαίδευσης την κατανομή των κλάσεων της μεταβλητής στόχου. Έτσι λοιπόν κάθε τμήμα που δημιουργείται είναι της τάξης του 50% νόμιμων συναλλαγών και 50% συναλλαγών απάτης. Το αποτέλεσμα της μεθόδου αυτής είναι η μέση τιμή των μετρικών που δηλώνουμε. Στη συγκεκριμένη πειραματική διαδικασία χρησιμοποιούνται επιστρέφονται οι εξής μετρικές αξιολόγησης:

- Συνολική Ακρίβεια (Accuracy)
- Ακρίβεια (Precision)
- Ανάκληση (Recall)
- F1 – Score

Έτσι θα είναι δυνατή μία υποτυπώδη σύγκριση μεταξύ των δύο μεθόδων εκπαίδευσης και αξιολόγησης των υπό εξέταση αλγορίθμων αλλά επίσης και με τη τυχαία υπό - δειγματοληψία που μελετήθηκε σε προγενέστερο στάδιο της παρούσας εργασίας.

algorithm	fittime	test_accuracy	test_precision	test_recall	test_f1-score
Regression	4.628417873	0.965454937	0.969031382	0.962700902	0.965590408
GaussianNaiveBayes	0.222557449	0.87067305	0.988665244	0.74995056	0.852871222
Decisiontree	53.51722822	0.806347568	0.789061344	0.999237352	0.863979064
MultiLayerPerceptron	218.9949366	0.962794499	0.952073762	0.977523094	0.963870399
RandomForest	446.5542849	0.845337835	0.825421172	0.999996469	0.88841254
XGBoost	247.4545406	0.833748126	0.799681642	0.999985877	0.875576175

Πίνακας 15 Πίνακας Επιδόσεων Μοντέλων (5-Fold Cross Validation)

Οι αναγραφόμενες τιμές του πίνακα 15 είναι αποτέλεσμα της μεθόδου στρατηγικής διασταυρούμενης επικύρωσης 5-φορών, όπου αναπαρίστανται οι μέσες τιμές από τις 5 υπολογιζόμενες εκδόσεις των μέτρων αξιολόγησης της μεθόδου.

5. Σύνοψη και Συμπεράσματα

Στην παρούσα διπλωματική εργασία πραγματοποιήθηκε μία επισκόπηση των βασικότερων μεθόδων που χρησιμοποιούνται στα σύγχρονα προβλήματα κατηγοριοποίησης. Σκοπός ήταν η ανάδειξη της σημασίας των μεγάλων δεδομένων και των εφαρμογών τους στα σημερινά τραπεζικά συστήματα και όχι μόνο. Μία από τις σημαντικότερες εφαρμογές είναι αυτή της ανίχνευσης απάτης καθώς πέρα από το οικονομική ζημία των τραπεζών και των πελατών τους είναι θέμα γοήτρου καθώς οι τράπεζες θα πρέπει να εμπνέουν αξιοπιστία και ασφάλεια με σκοπό την προσέλκυση περισσότερων πελατών αλλά και διατήρηση των ήδη υπαρχόντων. Οι μορφές απάτης που έρχεται αντιμέτωπο το τραπεζικό σύστημα είναι πολλαπλές και πολύ δύσκολα ανιχνεύσιμες. Μία από τις πλέον διαδεδομένες μορφές απάτης στην εποχή του «πλαστικού» χρήματος είναι η απάτη πιστωτικών καρτών. Η δημοφιλία της οφείλεται στο γεγονός ότι είναι σχετικά απλή χωρίς να χρειάζεται κάποια εξειδικευμένη γνώση ή να απαιτεί μεγάλο ρίσκο από την πλευρά του παραβάτη. Αρκεί το να έχει στην κατοχή του μία πιστωτική κάρτα που μπορεί να κάνει συναλλαγές χωρίς την επικύρωση με κάποιο PIN από το κάτοχό της.

Από πλευράς της τράπεζας υπάρχει μεγάλη ανάγκη για ένα σύστημα το οποίο μπορεί να εποπτεύει τις συναλλαγές και να μπορεί να κατηγοριοποιήσει αποδοτικά αν μία συναλλαγή είναι απάτη ή νόμιμη. Η συγκεκριμένη άσκηση εμπεριέχει αρκετές δυσκολίες καθώς οι συναλλαγές των πελατών των τραπεζών θεωρούνται απόρρητες με αποτέλεσμα να μην υπάρχουν πολλά διαθέσιμα σύνολα δεδομένων. Ο περιορισμένος αριθμός των συνόλων δυσχεραίνει το έργο των τραπεζικών συστημάτων στον αγώνα που δίνουν για την αποδοτική ανίχνευση των απατηλών συναλλαγών. Ένα πρόβλημα που εντοπίζεται επίσης είναι η μεγάλη ανομοιομορφία των δεδομένων καθώς ο αριθμός των νόμιμων συναλλαγών είναι συντριπτικά μεγαλύτερος από αυτό των συναλλαγών απάτης.

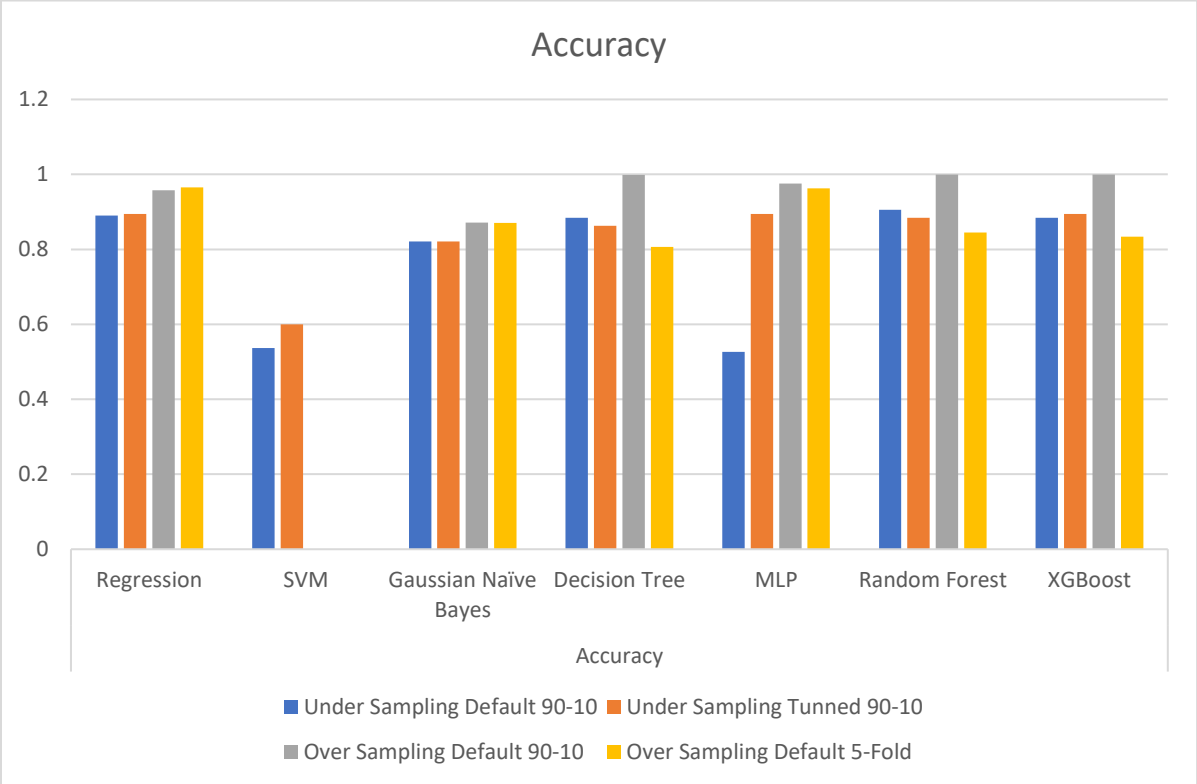
Στη παρούσα πειραματική διαδικασία χρησιμοποιήθηκε σύνολο δεδομένων πραγματικών συναλλαγών πιστωτικών καρτών από Ευρωπαίους πολίτες που πραγματοποιήθηκαν σε διάστημα 2 ημερών στο 2013. Λόγω του εμπιστευτικού χαρακτήρα των συναλλαγών στο σύνολο δεδομένων εφαρμόσθηκε PCA που είχε σαν αποτέλεσμα να μην γνωρίζουμε τις ακριβείς τιμές των χαρακτηριστικών. Αφού έγινε μία περιήγηση στα δεδομένα αναδεικνύοντας συσχετίσεις μεταξύ χαρακτηριστικών, απεικόνιση των πιο σημαντικών πεδίων και εξάλειψη των διπλοεγγραφών, πραγματοποιήθηκε ανάλυση εύρεσης και εξάλειψης ακραίων τιμών στο πεδίο του ποσού (Amount). Τέλος η προ - επεξεργασία του συνόλου ολοκληρώνεται με τη κανονικοποίηση των τιμών της Amount καθώς θα δημιουργούσε προβλήματα κατά το στάδιο της κατηγοριοποίσηης καθώς μερικοί από τους αλγόριθμους παρουσιάζουν ευαισθησία όσον αφορά τις μεγάλες διαφορές των τιμών μεταξύ των χαρακτηριστικών του συνόλου.

Για την διαδικασία της εύρεσης του καταλληλότερου εκτιμητή χρησιμοποιήθηκαν 7 αλγόριθμοι μηχανικής μάθησης, Logistic Regression, Support Vector Machine, Gaussian Naïve Bayes, Decision Tree, Multi-Layer Perceptron, Random Forest και XGBoost. Για την επίλυση του προβλήματος της ανομοιομορφίας των κλάσεων χρησιμοποιήθηκαν 2 τεχνικές. Η πρώτη είναι η τυχαία υπό – δειγματοληψία (random under sampling) της κλάσης πλειοψηφίας και η δεύτερη είναι υπέρ δειγματοληψία (oversampling) της κλάσης μειοψηφίας με τη μέθοδο SMOTE. Για την περίπτωση του under sampling έγινε μία προσπάθεια εύρεσης των καλύτερων παραμέτρων των αλγορίθμων και πραγματοποιήθηκε μία σύγκριση με τις περιπτώσεις των ίδιων εκτιμητών αλλά με τις προκαθορισμένες (default) παραμέτρους. Τα αποτελέσματα έδειξαν ότι κατά την προσπάθεια της ρύθμισης των παραμέτρων, με σημαντικά μικρότερο δείγμα για την διασφάλιση της ομοιομορφίας των κλάσεων της

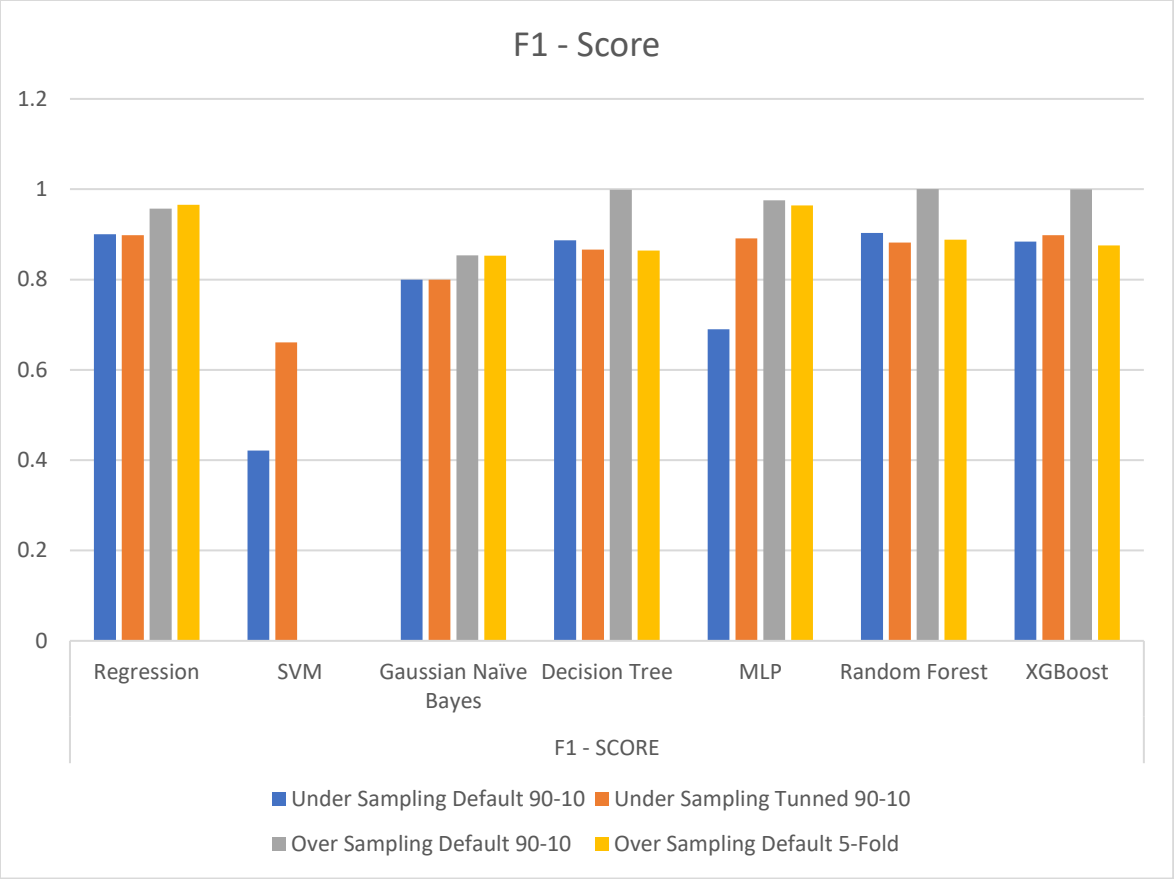
μεταβλητής στόχου, παρουσιάστηκε το φαινόμενο της υπέρ – προσαρμογής των εκτιμητών στο μικρό αυτό δείγμα οδηγώντας στην μη ουσιαστική βελτίωση των αποτελεσμάτων των ταξινομητών.

Στην συνέχεια έγινε χρήση της SMOTE αποσκοπώντας στην αύξηση των παρατηρήσεων απάτης επιτυγχάνοντας με αυτό το τρόπο την εξισορρόπηση των κλάσεων της μεταβλητής στόχου. Η διερεύνηση περιλαμβάνει πλέον 6 αλγορίθμους καθώς η δραματική αύξηση των παρατηρήσεων κατέστησε αδύνατη την εκπαίδευση του Support Vector Machine. Για την εκπαίδευση και αξιολόγηση του συνόλου δεδομένων μετά τη SMOTE χρησιμοποιήθηκαν δύο τεχνικές. Η πρώτη είναι ο κλασικός διαχωρισμός του αρχικού συνόλου δεδομένων σε σύνολα εκπαίδευσης και σύνολα δοκιμής και η δεύτερη προσέγγιση ήταν αυτή της Stratified K-Fold Cross Validation διαχωρίζοντας τα δεδομένα σε 5 υπό -σύνολα (5-Fold). Από τις δύο μεθόδους η 5-Fold φαίνεται να αποδυνάμωσε τους αλγορίθμους εκτός από αυτό της Λογιστικής Παλινδρόμησης και του Multi-Layer Perceptron.

Παρατηρώντας το συγκεντρωτικό πίνακα αποτελεσμάτων της πειραματικής διαδικασίας (πίνακας 16) γίνεται αντιληπτό ότι οι περισσότεροι αλγόριθμοι συμπεριφέρονται καλύτερα όταν επιλέγεται υπέρ – δειγματοληψία της κλάσης μειοψηφίας έναντι της υπό – δειγματοληψίας. Καλύτερες επιδόσεις πετυχαίνουν οι Random Forest, XGBoost και Decision Tree με ποσοστό 99,9% στην ακρίβεια αλλά και στο F1 – Score που ενδιαφέρει ίσως περισσότερο από την γενική ακρίβεια καθώς είναι μετρική αξιολόγησης εκτίμησης σε επίπεδο κλάσης. Συνεχίζοντας τις χειρότερες επιδόσεις τις σημείωσε ο Support Vector Machine καθώς σε όσες περιπτώσεις ήταν δυνατή η εκπαίδευσή του σε ξεπέρασε το 60% accuracy.



Εικόνα 59 Συγκριτική Απεικόνιση Αποτελεσμάτων Ακρίβειας (Accuracy) Πειραματικής Διαδικασίας



Εικόνα 60 Συγκριτική Απεικόνιση Αποτελεσμάτων F1 Πειραματικής Διαδικασίας

Accuracy

	Parameters	Validation	Regression	SVM	Gaussian Naïve Bayes	Decision Tree	MLP	Random Forest	XGBoost
Under Sampling	Default	90-10	0.89	0.537	0.8211	0.8842	0.526	0.9053	0.8842
	Tunned	90-10	0.8947	0.6	0.8211	0.8632	0.895	0.8842	0.8947
Over Sampling	Default	90-10	0.9575		0.8714	0.9985	0.976	0.9998	0.9998
	Default	5-Fold	0.9654		0.8706	0.8063	0.963	0.8453	0.8337

PRECISION

	Parameters	Validation	Regression	SVM	Gaussian Naïve Bayes	Decision Tree	MLP	Random Forest	XGBoost
Under Sampling	Default	90-10	0.9	0.615	0.9714	0.9149	0.526	0.9767	0.9333
	Tunned	90-10	0.9167	0.597	0.9714	0.8936	0.976	0.9535	0.9167
Over Sampling	Default	90-10	0.9674		0.9897	0.9979	0.993	0.9998	0.9997
	Default	5-Fold	0.969		0.9886	0.789	0.952	0.8254	0.7996

RECALL

	Parameters	Validation	Regression	SVM	Gaussian Naïve Bayes	Decision Tree	MLP	Random Forest	XGBoost
Under Sampling	Default	90-10	0.9	0.32	0.68	0.86	1	0.84	0.84
	Tunned	90-10	0.88	0.74	0.68	0.84	0.82	0.82	0.88
Over Sampling	Default	90-10	0.9469		0.7508	0.9993	0.958	1	1
	Default	5-Fold	0.9627		0.7499	0.9992	0.978	0.9999	0.9999

F1 - SCORE

	Parameters	Validation	Regression	SVM	Gaussian Naïve Bayes	Decision Tree	MLP	Random Forest	XGBoost
Under Sampling	Default	90-10	0.9	0.421	0.8	0.8866	0.69	0.9032	0.8842
	Tunned	90-10	0.898	0.661	0.8	0.866	0.891	0.8817	0.898
Over Sampling	Default	90-10	0.957		0.8538	0.9986	0.975	0.9999	0.9998
	Default	5-Fold	0.9655		0.8528	0.8639	0.964	0.8884	0.8757

Πίνακας 16 Πίνακας Αποτελεσμάτων Πειραματικής Διαδικασίας

6. Μελλοντικές Επεκτάσεις

Η παρούσα εργασία έδειξε εφαρμόζοντας σύγχρονες μεθόδους κατηγοριοποίησης αξιοποιώντας μόνο τα χαρακτηριστικά της εκάστοτε παρατήρησης ότι μπορούμε ικανοποιητικά να αξιολογήσουμε αν μία συναλλαγή είναι απατηλή ή νόμιμη. Παρά την υψηλή αποδοτικότητα υπάρχει πολύ χώρος για βελτίωση και επέκταση της συγκεκριμένης εργασίας.

Αρχικά θα μπορούσε να μελετηθεί η επίπτωση που θα επέφερε στα αποτελέσματα των μετρικών αξιολόγησης, η παράλειψη συγκεκριμένων χαρακτηριστικών του συνόλου δεδομένων (ιδανικά αυτά που έχουν χαμηλή συσχέτιση με τη μεταβλητή στόχο) από την διαδικασία εκπαίδευσης του εκάστοτε αλγορίθμου. Η συγκεκριμένη διερεύνηση συγκεντρώνει σημαντικά πλεονεκτήματα καθώς αν η αποτελεσματικότητα δεν επηρεαστεί σε επίπεδα απαγορευτικά, το μειωμένο πλήθος χαρακτηριστικών θα έκανε την διαδικασία εκπαίδευσης σημαντικά ταχύτερη. Ίσως με αυτό το τρόπο να είμαστε σε θέση να αξιολογήσουμε και τον αλγόριθμο μηχανών διανυσμάτων υποστήριξης μετά την υπέρ – δειγματοληψία της κλάσης μειοψηφίας πράγμα που ήταν αδύνατο να γίνει στο σύνολο των χαρακτηριστικών όπως διαπιστώθηκε στα πλαίσια της παρούσας εργασίας.

Συνεχίζοντας, ιδιαίτερο ενδιαφέρον θα είχε η συμπεριφορική διερεύνηση των συναλλαγών των καταναλωτών έχοντας σαν στόχο, την σύγκριση μιας συναλλαγής προς εκτίμηση, με αυτές του ίδιου ατόμου που έχουν ήδη πραγματοποιηθεί στο παρελθόν. Αυτό θα αποτελούσε ένα πολύ ισχυρό εργαλείο στη μάχη της ανίχνευσης της απάτης καθώς θα ήταν δυνατή η ανίχνευση οποιασδήποτε ανωμαλίας όσον αφορά την κίνηση του συγκεκριμένου πελάτη ξεφεύγοντας από τα ευρύτερα μοτίβα των συναλλαγών απάτης. Βλέπουμε λοιπόν ότι με αυτό το τρόπο ακόμη και ο μηχανισμός ανίχνευσης της απάτης θα μπορούσε να αποκτήσει μία πελατο – κεντρική διάσταση συμβαδίζοντας με το ρεύμα της εποχής. Η συγκεκριμένη διερεύνηση απαιτεί την σύνδεση πολλαπλών συναλλαγών με μία οντότητα και ίσως θα πρέπει να διαθέτουμε μεγαλύτερης έκτασης σύνολα δεδομένων, καθώς εντός δύο ημερών (όπως ήταν τα σύνολα της παρούσας εργασίας) πόσες συναλλαγές μπορεί να έχει κάνει ένα άτομο προκειμένου να βγει και ένα συμπέρασμα;

Τέλος ο συνδυασμός όλων των προηγουμένων σίγουρα είναι μία ιδέα άξια μελέτης. Ένας συνδυασμός δηλαδή, που στοχεύει στην βελτιστοποίηση της εκπαίδευσης των αλγορίθμων ταξινόμησης ως προς το χρόνο αλλά επίσης στην ενίσχυση των ορθών εκτιμήσεων, συνδυάζοντας τα αποτελέσματα της συμπεριφορικής διερεύνησης με αυτά των εκτιμητών. Το προτεινόμενο αυτό σύστημα ανίχνευσης απάτης αν και αυξημένης πολυπλοκότητας συγκεντρώνει πιθανά πλεονεκτήματα που αξίζουν περεταίρω διερεύνησης.

ΒΙΒΛΙΟΓΡΑΦΙΑ

1. Martin, A., *Global Data Creation is About to Explode.* 2019.
2. Columbus, L., *10 Charts That Will Change Your Perspective Of Big Data's Growth.* 2018.
3. Gutta, S., *5V's of Big Data*. 2020.
4. Ishwarappa and J. Anuradha, *A Brief Introduction on Big Data 5Vs Characteristics and Hadoop Technology.* Procedia Computer Science, 2015. **48**: p. 319-324.
5. Jia, Y., et al. *Telecom Big Data based Precise User Classification Scheme*. in *2019 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computing, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI)*. 2019. IEEE.
6. Darwish, T.S.J. and K.A. Bakar, *Fog Based Intelligent Transportation Big Data Analytics in The Internet of Vehicles Environment: Motivations, Architecture, Challenges, and Critical Issues.* IEEE Access, 2018. **6**: p. 15679-15701.
7. IEA (2021), G.E.R., IEA, Paris
8. Κουτσουλιά, Κ., *Αυτόνομα πλοία: η νέα πρόκληση στη ναυτιλία.* 2019.
9. *Global Sports Market Opportunities and Strategies Report 2021: Sports Market Forecast to Reach $599.9 billion by 2025 as COVID-19 Lockdowns Ease*. 22-7-2021; Available from: https://finance.yahoo.com/news/global-sports-market-opportunities-strategies-080800261.html?guccounter=1&guce_referrer=aHR0cHM6Ly93d3cuZ29vZ2xlLmNvbS8&guce_referrer_sig=AQAAAFFYQ923yckuNJH3SSydhIjxcBIW0emJVI0tvlltGbRogzmF2JhMEgfXrj22Mxpi72ReWzT0YcUwhk0m23z1WSdG09QO4xuAbk55603IEAR-eI-a_ghIRWghHH772hPzuqwsYPOUep-q35TfXW1WIiU88_eWhn60wpLF35MtR3Jq.
10. Sun, S., et al. *Application and research of Big data analysis in commercial Banks*. in *2020 International Conference on Big Data and Social Sciences (ICBDSS)*. 2020.
11. Moin, K.I. and D.Q.B. Ahmed, *Use of data mining in banking.* International Journal of Engineering Research and Applications, 2012. **2**(2): p. 738-742.
12. *Άρθρο 386 - Ποινικός Κώδικας (Νόμος 4619/2019) - Απάτη*. Available from: https://www.lawspot.gr/nomikes-plirofories/nomothesia/n-4619-2019/arthro-386-poinikos-kodikas-nomos-4619-2019-apati.
13. Research, J.S. *Total Identity Fraud Losses Soar to $56 Billion in 2020*. Available from: https://www.businesswire.com/news/home/20210323005370/en/Total-Identity-Fraud-Losses-Soar-to-56-Billion-in-2020.
14. *FRAUD - THE FACTS 2021 THE DEFINITIVE OVERVIEW OF PAYMENT INDUSTRY FRAUD*. 2021, UK FINANCE.
15. Hormozi, A.M. and S. Giles, *Data mining: A competitive weapon for banking and retail industries.* Information systems management, 2004. **21**(2): p. 62-71.
16. Kala, K., *A Customized Approach for Risk Evaluation and Prediction based on Data Mining Technique.* INTERNATIONAL JOURNAL OF ENGINEERING RESEARCH & TECHNOLOGY (IJERT) RACMS – 2015, 2018.
17. *Customer Acquisition vs. Customer Retention: What Data Says?* 2021; Available from: https://www.markinblog.com/customer-loyalty-retention-statistics/?gclid=Cj0KCQiAgP6PBhDmARIsAPWMq6k33R7m7zxWk9L8bsDWglo7OWx2W_hmFj2g7r4KIj84JQ8Mj7YUpu4aAj9PEALw_wcB
18. *Is Acquiring New Customers More Expensive Than Keeping Them?* 2021; Available from: https://www.europeanbusinessreview.com/is-acquiring-new-customers-more-expensive-than-keeping-them/.

19. Reinartz, W., J.S. Thomas, and V. Kumar, *Balancing acquisition and retention resources to maximize customer profitability.* Journal of marketing, 2005. **69**(1): p. 63-79.
20. Boote, A.S., *Interactions in psychographics segmentation: Implications for advertising.* Journal of Advertising, 1984. **13**(2): p. 43-48.
21. Gallego, D. and G. Huecas. *An empirical case of a context-aware mobile recommender system in a banking environment*. in *2012 third FTRA international conference on mobile, ubiquitous, and intelligent computing*. 2012. IEEE.
22. Yuan, X., et al., *Toward a user-oriented recommendation system for real estate websites.* Information Systems, 2013. **38**(2): p. 231-243.
23. Daly, E.M., et al. *Multi-criteria journey aware housing recommender system*. in *Proceedings of the 8th ACM Conference on Recommender systems*. 2014.
24. Han, J., Pei, Jian, Kamber, Micheline, *Data mining: concepts and techniques*. 2011: Elsevier.
25. Loshin, D., *Chapter 21 - Quick Reference Guide*, in *Business Intelligence (Second Edition)*, D. Loshin, Editor. 2013, Morgan Kaufmann. p. 333-353.
26. Κύρκος, Ε.Γ., *Επιχειρηματική Ευφυΐα και Εξόρυξη Δεδομένων*. 2015.
27. Hall, M.A., *Correlation-based feature selection for machine learning.* 1999.
28. Chawla, N.V., *Data Mining for Imbalanced Datasets: An Overview*, in *Data Mining and Knowledge Discovery Handbook*, O. Maimon and L. Rokach, Editors. 2010, Springer US: Boston, MA. p. 875-886.
29. Chawla, N.V., N. Japkowicz, and A. Kotcz, *Editorial: special issue on learning from imbalanced data sets.* SIGKDD Explor. Newsl., 2004. **6**(1): p. 1–6.
30. Kotsiantis, S., D. Kanellopoulos, and P. Pintelas, *Handling imbalanced datasets: A review.* GESTS international transactions on computer science and engineering, 2006. **30**(1): p. 25-36.
31. Chawla, N.V., et al., *SMOTE: synthetic minority over-sampling technique.* Journal of artificial intelligence research, 2002. **16**: p. 321-357.
32. Bunke, H., et al. *Recovery of temporal information of cursively handwritten words for on-line recognition*. in *Proceedings of the Fourth International Conference on Document Analysis and Recognition*. 1997.
33. Ma, H., et al., *Integrating Growth and Environmental Parameters to Discriminate Powdery Mildew and Aphid of Winter Wheat Using Bi-Temporal Landsat-8 Imagery.* Remote Sensing, 2019. **11**: p. 846.
34. Drummond, C. and R.C. Holte. *C4. 5, class imbalance, and cost sensitivity: why under-sampling beats over-sampling*. in *Workshop on learning from imbalanced datasets II*. 2003.
35. Πετρίδης, Δ., *ΛΟΓΙΣΤΙΚΗ ΠΑΛΙΝΔΡΟΜΗΣΗ.* 2016.
36. By Larhmam - Own work, C.B.-S., S. Graph_1, Editor.
37. Yu, H. and S. Kim, *SVM Tutorial-Classification, Regression and Ranking.* Handbook of Natural computing, 2012. **1**: p. 479-506.
38. Kotsiantis, S.B., *Decision trees: a recent overview.* Artificial Intelligence Review, 2013. **39**(4): p. 261-283.
39. Rokach, L. and O. Maimon, *Decision trees.* Data mining and knowledge discovery handbook, 2005: p. 165-192.
40. Koch, K.-R. and K.-R. Koch, *Bayes' theorem.* Bayesian Inference with Geodetic Applications, 1990: p. 4-8.
41. Sun, S., et al., *Active Learning With Gaussian Process Classifier for Hyperspectral Image Classification.* IEEE Transactions on Geoscience and Remote Sensing, 2015. **53**(4): p. 1746-1760.
42. Xu, S., *Bayesian Naïve Bayes classifiers to text classification.* Journal of Information Science, 2018. **44**(1): p. 48-59.

43. Di Nunzio, G.M., *A new decision to take for cost-sensitive Naïve Bayes classifiers.* Information Processing & Management, 2014. **50**(5): p. 653-674.
44. Priddy, K.L. and P.E. Keller, *Artificial neural networks: an introduction*. Vol. 68. 2005: SPIE press.
45. Bre, F., J. Gimenez, and V. Fachinotti, *Prediction of wind pressure coefficients on building surfaces using Artificial Neural Networks.* Energy and Buildings, 2017. **158**.
46. Chen, T. and C. Guestrin, *XGBoost: A Scalable Tree Boosting System*, in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2016, Association for Computing Machinery: San Francisco, California, USA. p. 785–794.
47. Guo, R., et al., *Degradation state recognition of piston pump based on ICEEMDAN and XGBoost.* Applied Sciences, 2020. **10**: p. 6593.
48. Kumar, V., J.K. Chhabra, and D. Kumar, *Performance evaluation of distance metrics in the clustering algorithms.* INFOCOMP Journal of Computer Science, 2014. **13**(1): p. 38-52.
49. Sinaga, K.P. and M.S. Yang, *Unsupervised K-Means Clustering Algorithm.* IEEE Access, 2020. **8**: p. 80716-80727.
50. Khan, K., et al. *DBSCAN: Past, present and future*. in *The Fifth International Conference on the Applications of Digital Information and Web Technologies (ICADIWT 2014)*. 2014.
51. *K-Fold Cross-validation Image*. scikit-learn.org.
52. Refaeilzadeh, P., L. Tang, and H. Liu, *Cross-Validation*, in *Encyclopedia of Database Systems*, L. Liu and M.T. ÖZsu, Editors. 2009, Springer US: Boston, MA. p. 532-538.