Εθνικο Μετσοβιο Πολυτεχνειο
Σχολη Ηλεκτρολογων Μηχανικων και Μηχανικων Υπολογιστων
Τομεας Ηλεκτρικης Ισχυος

Εντοπισμός ρευματοκλοπών με μηχανική μάθηση

Διπλωματικη Εργασια

του

ΜΗΤΣΕΛΟΥ ΑΘΑΝΑΣΙΟΥ

Επιβλέπων: Χατζηαργυρίου Νικόλαος
Καθηγητής Ε.Μ.Π.

Εργαστηριο Συστηματων Ηλεκτρικης Ενεργειας
Αθήνα, Οκτώβριος 2017

Εθνικό Μετσόβιο Πολυτεχνείο
Σχολή Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών
Τομέας Ηλεκτρικής Ισχύος
Εργαστήριο Συστημάτων Ηλεκτρικής Ενέργειας

Εντοπισμός ρευματοκλοπών με μηχανική μάθηση

Διπλωματικη Εργασια

του

ΜΗΤΣΕΛΟΥ ΑΘΑΝΑΣΙΟΥ

Επιβλέπων: Χατζηαργυρίου Νικόλαος
Καθηγητής Ε.Μ.Π.

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 666 Οκτωβρίου 2017.

(Υπογραφή)	*(Υπογραφή)*	*(Υπογραφή)*
.....	.....	.....
Χατζηαργυρίου Νικόλαος	Παπαθανασίου Σταύρος	Γεωργιλάκης Παύλος
Καθηγητής Ε.Μ.Π.	Αν. Καθηγητής Ε.Μ.Π.	Επ. Καθηγητής Ε.Μ.Π.

Αθήνα, Οκτώβριος 2017

(Υπογραφή)

.....

ΜΗΤΣΕΛΟΣ ΑΘΑΝΑΣΙΟΣ

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

Εθνικό Μετσόβιο Πολυτεχνείο
Σχολή Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών
Τομέας Ηλεκτρικής Ισχύος
Εργαστήριο Συστημάτων Ηλεκτρικής Ενέργειας

Ευχαριστίες

Θα ήθελα να ευχαριστήσω τον επιβλέποντα καθηγητή κ. Νικόλαο Χατζηαργυρίου για την ευκαιρία που μου έδωσε να εκπονήσω τη παρούσα διπλωματική.

Επίσης, θα ήθελα να ευχαριστήσω τους καθηγητές κ. Σταύρο Παπαθανασίου και κ. Παύλο Γεωργιλάκη για την τιμή που μου έκαναν να συμμετάσχουν στην επιτροπή εξέτασης της διπλωματικής.

Ευχαριστώ ιδιαίτερα τον υποψήφιο διδάκτορα Γιώργη Μεσσήνη για την καθοδήγηση, στήριξη και καθοριστική βοήθεια που μου παρείχε με συνέπεια και σοβαρότητα.

Τέλος, θα ήθελα να ευχαριστήσω την οικογένειά μου και τους φίλους μου που παρέχουν πάντοτε ένα χέρι βοήθειας σε ό,τι χρειαστώ και ειδικότερα τον πατέρα μου για τη γλωσσική επιμέλεια αυτής της διπλωματικής.

Αφιερώνω αυτή τη διπλωματική στον Αλέξανδρο, στην Αμαλία και στη Μαρία.

Περίληψη

Οι εταιρίες παροχής ηλεκτρισμού αντιμετωπίζουν το ολοένα και αυξανόμενο πρόβλημα της διείσδυσης μη τεχνικών απωλειών στις καταναλώσεις των πελατών τους. Το γεγονός αυτό πλήττει σημαντικά τις εταιρίες, μειώνοντας το εισόδημά τους και θέτει σε κίνδυνο τους ανειδίκευτους καταναλωτές που επεμβαίνουν στις υποδομές του παρόχου. Η προσέγγιση αυτού του προβλήματος έγινε με προσομοίωση ρευματοκλοπών σε ετήσιες χρονοσειρές καταναλωτών και δοκιμάστηκαν πληθώρα αλγορίθμων επιβλεπόμενης, μη επιβλεπόμενης και ημι-επιβλεπόμενης μηχανικής μάθησης για την ανίχνευση των καταναλωτών με διείσδυση μη τεχνικών απωλειών. Τα αποτελέσματα αναδεικνύουν τις δυνατότητες των συστημάτων μη επιβλεπόμενης και ημι-επιβλεπόμενης μάθησης σε σχέση με τη δεδομένη επιτυχία των αλγορίθμων επιβλεπόμενης μάθησης. Τα συστήματα που δημιουργήθηκαν έχουν ικανοποιητική απόδοση που δεν αποκλίνει σημαντικά από τους αλγορίθμους αναφοράς της επιβλεπόμενης μάθησης. Καθίσταται λοιπόν σαφές πως η ανίχνευση μη τεχνικών απωλειών είναι εφικτή με συστήματα μηχανικής μάθησης.

Λέξεις Κλειδιά

Μη τεχνικές απώλειες, Ρευματοκλοπές, Χρονοσειρές, Μηχανική μάθηση, Επιβλεπόμενοι αλγόριθμοι, Μη επιβλεπόμενοι αλγόριθμοι, Ημι-επιβλεπόμενοι αλγόριθμοι.

Abstract

Power companies face the problem of increasing intrusion of non-technical losses on consumptions of their clients. That fact hurts significantly power companies by reducing their economical growth and sets on danger unskilled consumers who intervene with the power infastracture. This problem was approached by simulating frauds on yearly timeseries and by testing many different algorithms of supervised, unsupervised and semi-supervised machine learning in order to detect consumers with non-technical loss intrusion. The results show the potencial of the unsupervised and semi-supervised learning in relation with the given success of supervised algorithms. The created systems have satisfactory performance which does not diverge significantly from the reference algorithms of supervised learning. Concluding the detection of non-technical losses is achievable with machine learning systems.

Keywords

Non-technical losses, power fraud, Timeseries, Machine learning, Supervised algorithms, Unsupervised algorithms, Semi-supervised algorithms.

Περιεχόμενα

Κατάλογος Σχημάτων

Κατάλογος Πινάκων

Κεφάλαιο 1

Εισαγωγή

Είναι ευρέως διαδεδομένο πως η καθημερινότητα πολλών ανθρώπων συνδέεται άρρηκτα με τη χρήση ηλεκτρικών συσκευών, αλλά και με την ανάγκη ύπαρξης βιομηχανικών εγκαταστάσεων για την εκπλήρωση των καταναλωτικών τους επιθυμιών. Αυτό δημιουργεί μια αυξανόμενη ζήτηση στον τομέα της παραγωγής, της μεταφοράς και διανομής ηλεκτρικής ενέργειας, που με τη σειρά της οδηγεί στον συνεχή εκσυγχρονισμό των εγκαταστάσεων. Παράλληλα, διανύοντας την εποχή της Ψηφιακής Επανάστασης, παρατηρείται η μετάβαση από τις αναλογικές τεχνολογίες στις ψηφιακές, γεγονός που δεν θα μπορούσε να αφήσει ανεπηρέαστο τον τομέα της ηλεκτρικής ενέργειας. Η μετάβαση αυτή στον τομέα που μελετάται σε αυτή τη διπλωματική εργασία σηματοδοτείται από την χρήση έξυπνων μετρητών, οι οποίοι έχουν τη δυνατότητα να παρέχουν μεγάλο όγκο δεδομένων για τα επίπεδα της κατανάλωσης κάθε πελάτη.

Ανοίγεται, λοιπόν, ένας νέος ορίζοντας εποπτείας και αναλυτικής μελέτης των χρονοσειρών που παράγονται από κάθε καταναλωτή. Η ταυτόχρονη και συνεχής αύξηση των ρευματοκλοπών στις περισσότερες περιοχές του κόσμου καθιστά επιτακτική ανάγκη την εύρεση μεθόδων εντοπισμού τους . Σύμφωνα με τα επίσημα στοιχεία του Διαχειριστή Δικτύου (ΔΕΔΔΗΕ), το 2016 εντοπίστηκαν 10.616 κρούσματα ρευματοκλοπών, μέγεθος που είναι το υψηλότερο όλων των εποχών, έναντι 400 το 2006 [34]. Άμεσο επακόλουθο της επίλυσης αυτού του προβλήματος είναι η ομαλή λειτουργία των παρόχων ενέργειας και η βελτίωση της ποιότητας των υπηρεσιών που παρέχουν. Στη συνέχεια θα αναπτυχθεί το βαθύτερο αίτιο της παρούσας διατριβής και μια επισκόπηση του περιεχομένου της [28].

1.1 Κίνητρο και υπόβαθρο διπλωματικής

Το πρόβλημα της παράνομης αφαίρεσης ηλεκτρικής ενέργειας ενδιαφέρει τους διαχειριστές δικτύων. Οι χρήστες συχνά παραβιάζουν τους νόμους, προσπαθώντας να αλλοιώσουν τα συστήματα μέτρησης. Σε κάποιες χώρες μόνο κάποιο κομμάτι της παραγωγής χρεώνεται. Παραδείγματος χάριν στην Ινδία το 55% της παραγωγής ηλεκτρικής ενέργειας χρεώνεται και το υπόλοιπο καταναλώνεται χωρίς να περάσει από μετρητικές συσκευές [26]. Παράλληλα, μόνο ένα μέρος της πληρωμής καταλήγει στον πάροχο, λόγω απλήρωτων λογαριασμών και περιφερειακών χρεώσεων. Παρ΄ όλα αυτά, η παράνομη χρήση ενέργειας λαμβάνει χώρα και

σε ευρωπαϊκές χώρες. Το σημαντικότερο κίνητρο για το λανσάρισμα των αυτοματοποιημένων υποδομών ανάγνωσης μετρητών (Automated Meter Reading) από τον ιταλικό πάροχο ενέργειας (ENEL) ήταν η προσπάθεια ελαχιστοποίησης των μη τεχνικών απωλειών στο δίκτυο διανομής του. Η μείωση των ρευματοκλοπών βοήθησε στην αιτιολόγηση μεγάλων επενδύσεων σε AMR και επί του παρόντος η Ιταλία πρωταγωνιστεί στην διείσδυση AMR [7].
Οι εταιρίες παραγωγής, μεταφοράς και διανομής αναλαμβάνουν την ευθύνη της κάλυψης των ενεργειακών αναγκών των πελατών. Μερικοί μπορεί να υποστηρίζουν ότι αυτές οι εταιρίες παρέχουν κακή εξυπηρέτηση, υπερχρεώνουν, κερδίζουν ανεξέλεγκτα αρκετά χρήματα και ως εκ τούτου, ένα ποσοστό κλοπής δεν θα καταστρέψει την εταιρία ή δεν θα επηρεάσει δραστικά τις λειτουργίες και την κερδοφορία της. Άλλοι, παρατηρώντας την ίδια κατάσταση, θα υποστήριζαν ότι η κλοπή είναι έγκλημα και δεν θα έπρεπε να επιτρέπεται. Η Διεθνής Εταιρία Προστασίας Εσόδων των Παρόχων (International Utilities Revenue Protection Association) έχει καθιερωθεί για να προάγει τον εντοπισμό και την πρόληψη της κλοπής ρεύματος κυρίως για την οικονομική ασφάλεια των εταιριών παροχής ενέργειας.

Οι συνέπειες της κλοπής είναι εξαιρετικά σημαντικές και μπορούν να επηρεάσουν άμεσα τη βιωσιμότητα των υπηρεσιών που παρέχονται. Οι συνδυασμένες απώλειες (συμπεριλαμβάνοντας και τους απλήρωτους λογαριασμούς) σε μερικά συστήματα έχουν σοβαρές επιπτώσεις, με συνέπεια οι εγκαταστάσεις να λειτουργούν σε καθεστώς μεγάλων απωλειών. Όταν οι εταιρίες παραγωγής, μεταφοράς και διανομής λειτουργούν σε καθεστώς αναποτελεσματικότητας και διαφθοράς, η παροχή αξιόπιστων υπηρεσιών επιτυγχάνεται με μεγάλη δυσκολία. Ακόμη και σε αποτελεσματικά συστήματα ισχύος, όπως η Tenaga της Μαλαισίας, η κλοπή ρεύματος ανέρχεται στα $132 εκατομμύρια ετησίως [18]. Αντίστοιχα, στην Ελλάδα η συνολική εγχεόμενη ενέργεια στο Δίκτυο Διανομής ανήλθε το 2016 σε 47.655.372 MWh, ενώ το σύνολο των ρευματοκλοπών εκτιμάται σε 1.525.292 MWh. Στην πραγματικότητα, όμως, το μέγεθος των ρευματοκλοπών είναι αρκετά μεγαλύτερο, επιβαρύνει δε κατά κύριο λόγο τη Δημόσια Επιχείρηση Ηλεκτρισμού (ΔΕΗ). Ωστόσο, παίρνοντας ως δεδομένη την ποσότητα που αναγνωρίζει η Ρυθμιστική Αρχή Ενέργειας (ΡΑΕ), τα έσοδα που διαφεύγουν κάθε χρόνο λόγω των ρευματοκλοπών με βάση τις μοναδιαίες τιμές του 2016 έχουν ως εξής [32]:

Εταιρίες	εκατ. €
ΔΕΗ	120-125
Υπηρεσίες Κοινής Ωφέλειας (ΥΚΩ)	21
ΕΤΜΕΑΡ	32
ΑΔΜΗΕ	7,3
ΔΕΔΔΗΕ	26,5
Σύνολο	206,8-211,8

Πίνακας 1.1: Διαφεύγοντα έσοδα ελληνικών εταιριών λόγω ρευματοκλοπών

1.1.1 Ορίζοντας τις ρευματοκλοπές

Σύμφωνα με το εγχειρίδιο ρευματοκλοπών της PAE, ρευματοκλοπή ορίζεται εν γένει η αυθαίρετη και με δόλο επέμβαση σε εξοπλισμό ή εγκαταστάσεις του Δικτύου, με σκοπό την κατανάλωση ηλεκτρικής ενέργειας, χωρίς αυτή να καταγράφεται ή να αντιστοιχίζεται με Εκπρόσωπο Φορτίου, και χωρίς να τιμολογείται [35]. Υπάρχουν τέσσερις επικρατούσες μέθοδοι «κλοπής» σε όλα τα συστήματα ενέργειας. Η έκταση της κλοπής εξαρτάται από πλήθος παραγόντων (πολιτιστικούς, πολιτικούς, οργανωτικούς κ.ά.).

Επέμβαση στο μετρητή

Επέμβαση στο μετρητή θεωρείται όταν ο καταναλωτής σκοπίμως παρεμβαίνει στη μετριτική διάταξη με σκοπό τη χαμηλότερη χρέωση. Μια συνήθης πρακτική είναι να παραβιάζει το μετρητή, ώστε να καταγράφει χαμηλότερα ποσά ενέργειας από τα πραγματικά. Αυτό εν γένει είναι μια επικίνδυνη διαδικασία για έναν ερασιτέχνη, και σε πολλές περιπτώσεις έχουν καταγραφεί περιστατικά ηλεκτροπληξίας. Στην Ελλάδα πρόκειται για τη συνηθέστερη περίπτωση ρευματοκλοπής [35].

Απευθείας Σύνδεση

Η κλοπή ενέργειας μπορεί να επιτευχθεί τραβώντας μια γραμμή από το δίκτυο διανομής μέχρι το επιθυμητό σημείο και παρακάμπτοντας τον μετρητή. Ένας καθιερωμένος τρόπος κλοπής ενέργειας στην Ελλάδα είναι η απευθείας σύνδεση με αγκίστρωση στους αγωγούς του εναέριου δικτύου, απουσία μετρητικής διάταξης ή παροχής ή νομίμως υφιστάμενου κτίσματος [35].

Ακανόνιστες χρεώσεις

Οι ακανόνιστες χρεώσεις μπορούν να προέλθουν από πολλές πηγές. Κάποιοι οργανισμοί παροχής ενέργειας μπορεί να μην είναι αρκετά αποτελεσματικοί στη μέτρηση της ενέργειας που έχει καταναλωθεί και ακούσια μπορεί να δώσουν υψηλότερη ή χαμηλότερη μέτρηση από την πραγματική. Αυτές οι ακανόνιστες χρεώσεις μπορεί να ισοζυγιστούν με την πάροδο του χρόνου. Παρ΄ όλα αυτά, είναι πολύ εύκολο σε μερικά συστήματα να έρθει σε επαφή εργαζόμενος με καταναλωτή, για να ορίσουν πολύ χαμηλότερους λογαριασμούς από τους πραγματικούς. Εργαζόμενοι μπορεί να δωροδοκηθούν για να καταγράψουν μικρότερο νούμερο από αυτό που εμφανίζεται στον μετρητή. Ο καταναλωτής πληρώνει μικρότερο λογαριασμό και ο εργαζόμενος που καταγράφει τις μετρήσεις αποκτά ανεπίσημο μισθό.

Απλήρωτοι λογαριασμοί

Κάποια άτομα και κάποιοι οργανισμοί δεν πληρώνουν αυτά που οφείλουν για ηλεκτρική ενέργεια. Οικιακοί ή επιχειρηματικοί καταναλωτές μπορεί να έχουν φύγει από την πόλη ή την εγκατάσταση λόγω χρεοκοπίας. Στη Νότιο Αμερική, υπάρχει «καθεστώς μη πληρωμής» [16].

Στην Αρμενία, τα επίπεδα μη πληρωμής είναι της τάξης του 80-90% για τον οικιακό τομέα. Οι απώλειες των μετασχηματιστών και της διανομής είναι άνω του 40% [29].

Σε όλες τις χώρες, καθώς η τιμή της ηλεκτρικής ενέργειας αυξάνεται, κάποιοι άνθρωποι αδυνατούν να πληρώσουν με συνέπεια τους λογαριασμούς τους. Αυτό τους ενθαρρύνει να αναζητήσουν τρόπους να μειώσουν τους λογαριασμούς, όπως η επέμβαση τους μετρητές.

1.2 Δομή Διπλωματικής

Στην παρούσα διπλωματική γίνεται μια διεξοδική αναζήτηση μεθόδων ανίχνευσης απάτης με μια πληθώρα διαφορετικών αλγορίθμων από τη σκοπιά της μηχανικής μάθησης. Δεδομένου του εύρους των δυνατοτήτων της μηχανικής μάθησης, έγινε προσπάθεια για αντιμετώπιση του προβλήματος από διαφορετικές οπτικές γωνίες, ώστε να επιτευχθεί η βέλτιστη αντιστάθμιση μεταξύ απόδοσης και πρακτικότητας. Η εξισορρόπηση αυτών των παραγόντων είναι κύριο μέλημα κάθε μηχανικού. Ειδικότερα, συνοψίζοντας κάθε κεφάλαιο εξάγεται η παρακάτω δομή:

Κεφάλαιο 1

Γνωστοποιείται το βασικό κίνητρο αυτής της διπλωματικής και δίνεται ένας σαφής ορισμός του προς αντιμετώπιση προβλήματος.

Κεφάλαιο 2

Γίνεται μια εισαγωγή στα εργαλεία που χρησιμοποιούνται για τη λήψη των αρχικών χρονοσειρών, την επεξεργασία τους και ταξινόμηση των καταναλωτών, αλλά και για τις συνιστώσες που λαμβάνονται υπόψιν για τα τελικά αποτελέσματα.

Κεφάλαιο 3

Αναπτύσσεται η μορφή και η φύση των δεδομένων, αλλά και η μεθοδολογία προεπεξεργασίας τους. Παράλληλα, διευκρινίζεται ο τρόπος προσομοίωσης και μοντελοποίησης της ρευματοκλοπής.

Κεφάλαιο 4

Δημιουργείται ένας άξονας αναφοράς για τα αποτελέσματα με τη χρήση αλγορίθμων επιβλεπόμενης μάθησης που φημίζονται για την μεγάλη ευστοχία τους, αλλά και τη δυσκολία εφαρμογής τους σε πραγματικά προβλήματα.

Κεφάλαιο 5

Εξετάζονται λεπτομερώς τα συστατικά των αλγορίθμων μη επιβλεπόμενης μάθησης, ενώ παράλληλα διεξάγονται δοκιμές για την εξερεύνηση των διαφορετικών μεθόδων επίλυσης του προβληματος.

Κεφάλαιο 6

Επεξηγούνται οι δυσκολίες που αντιμετωπίστηκαν από διαφορετικά πρίσματα. Αναλυτικότερα, γνωστοποιούνται τα τεχνικά εμπόδια που αντιμετωπίστηκαν, αλλά και οι που θα υποστούν οι καταναλωτές.

Κεφάλαιο 7

Παρουσιάζονται σφαιρικά τα αποτελέσματα με γνώμονες τη φύση κάθε αλγορίθμου και την ευστοχία στην ταξινόμηση των καταναλωτών και διατυπώνονται προτάσεις για μελλοντική επέκταση.

Κεφάλαιο 2

Θεωρητικό υπόβαθρο

Η ηλεκτρική ενέργεια είναι ζωτικής σημασίας για την καθημερινότητά μας αλλά και ο ακρογωνιαίος λίθος της βιομηχανίας. Για αυτό τον λόγο, η έννοια των μελλοντικών δικτύων (έξυπνα δίκτυα) στοχεύει στην αύξηση της αξιοπιστίας, της ποιότητας και της ασφάλειας της μελλοντικής παροχής ενέργειας. Για να συμβεί αυτό, απαιτούνται περαιτέρω πληροφορίες για τη λειτουργία και την κατάσταση των δικτύων διανομής. Μια από τις σημαντικότερες προκλήσεις στα μελλοντικά δίκτυα διανομής είναι η αυξανόμενη διείσδυση διεσπαρμένης παραγωγής (Distributed Generation) και η μετάβαση από την έννοια της παραδοσιακής παραγωγής ενέργειας με κυρίαρχους μεγάλους σταθμούς παραγωγής και ροές ενέργειας μονής κατεύθυνσης σε κατανεμημένα μοντέλα. Οι πληροφορίες λειτουργίας θα είναι καίριας σημασίας για τη λειτουργικότητα των μελλοντικών δικτύων διανομής και για τους διαχειριστές του δικτύου (Distribution Network Operators). Μια από τις πηγές πληροφορίας θα είναι οι προηγμένες υποδομές μέτρησης. Εκτός των άλλων, οι έξυπνοι μετρητές πρέπει να διευρύνουν τους γνωστικούς ορίζοντες των καταναλωτών για την ηλεκτρική ενέργεια. Η έννοια αυτή θα παράξει ακόμη περισσότερη πληροφορία στους διαχειριστές δικτύου. Τα δεδομένα των έξυπνων μετρητών δίνουν τη δυνατότητα στον διαχειριστή του δικτύου να αναλύσει ροές ενέργειας και να εντοπίσει πιθανή κλοπή ρεύματος [20].

2.1 Έξυπνοι μετρητές

Η διανομή είναι ένας τομέας στον οποίο η εξέλιξη είναι σταδιακή, τουλάχιστον όσον αφορά τα στοιχεία του δικτύου. Παρ΄ όλα αυτά, ο κλάδος των τηλεπικοινωνιών και της εξαγωγής και επεξεργασίας δεδομένων εξελίσσεται ραγδαία τα τελευταία χρόνια. Οι απομακρυσμένες μετρήσεις και η συνεχής καταγραφή και παρακολούθηση της κατανάλωσης θεωρούνται ως προηγμένη υποδομή μέτρησης (Advanced Metering Infrastructure). Η δραστική μείωση στις τιμές των μετρητών και στον εξοπλισμό τηλεπικοινωνιών καθιστά την απόκτησή τους οικονομικά βιώσιμη, ξεκινώντας με μεγάλους καταναλωτές και σταδιακά εγκαθιστώντας τους στους μέσους και μικρούς. Η αποτελεσματικότητα των εργαλείων στην αναγνώριση και αποθάρρυνση της κλοπής και άλλων τρόπων παράκαμψης μετρητών είναι τεράστια, όπως φαίνεται να

συμβαίνει σε αναπτυσσόμενες χώρες (συμπεριλαμβανομένου της Δομινικανικής Δημοκρατίας, της Ονδούρας και της Βραζιλίας) [2].

Η ευρεία εφαρμογή AMI μπορεί να συμβάλλει σημαντικά στη συνεχή ανάπτυξη και την αποτελεσματική λειτουργία των ενεργειακών δομών. Τα AMI παρέχουν ισχυρά εργαλεία, ικανά να μειώσουν τις συνολικές απώλειες και να αυξήσουν τα έσοδα των εταιριών.

2.1.1 Θετικά αντίκτυπα εφαρμογής AMI

Η εφαρμογή των AMI θα έχει τα ακόλουθα θετικά αποτελέσματα:

1. Αίσθηση παρακολούθησης στους χρήστες. Οι καταναλωτές αντιλαμβάνονται πως ο πάροχος ενέργειας μπορεί να παρακολουθεί την κατανάλωση. Αυτό επιτρέπει στην εταιρία τη γρήγορη ανίχνευση οποιασδήποτε ανωμαλίας στην κατανάλωση, λόγω αλλοίωσης του μετρητή ή παράκαμψής του και της δίνει τη δυνατότητα να κάνει διορθωτικές κινήσεις. Το αποτέλεσμα είναι η πειθάρχηση των καταναλωτών.

2. Ενίσχυση της εταιρικής διακυβέρνησης και καταπολέμησης της διαφθοράς. Τα παραδείγματα κλοπής μεγάλων καταναλωτών συνήθως συμπεριλαμβάνουν συνεννόηση μεταξύ αυτών και των ελεγκτών των μετρητών. Η διαφθορά είναι επίσης πιθανό να παρατηρηθεί και στις ενέργειες που συσχετίζονται με την αποσύνδεση του μετρητή, λόγω απλήρωτων λογαριασμών. Η εγκατάσταση των AMI καθιστά τις πληροφορίες των έξυπνων μετρητών διαθέσιμες στους καταναλωτές και τους διαχειριστές, επιβάλλοντας διαφάνεια.

3. Υλοποίηση προπληρωμένων καταναλώσεων. Η προ-πλήρωση των λογαριασμών είναι γενικώς πολύ θετικό για τους καταναλωτές μικρού εισοδήματος. Τα AMI δίνουν τη δυνατότητα αντιγραφής του επιχειρηματικού μοντέλου των εταιριών κινητής τηλεφωνίας και στον τομέα της ενέργειας.

4. Ελαχιστοποίηση απωλειών σε δυσπρόσιτες και απομακρυσμένες περιοχές. Τα AMI διαδραματίζουν καθοριστικό ρόλο στην προσέγγιση της διανομής μέσης τάσης (Medium-Voltage Distribution), που χρησιμοποιείται για την κατασκευή και λειτουργία ηλεκτρικών δικτύων, για την παροχή ενέργειας σε περιοχές που η πρόσβαση της εταιρίας είναι περιορισμένη για λόγους ασφαλείας. Στα MVD δίκτυα κάθε σύνδεση καταναλωτή ξεκινάει απευθείας από το μετασχηματιστή μέσης σε χαμηλή τάση, με το δίκτυο χαμηλής τάσης να εκλείπει.

5. Διαχείριση από την πλευρά της ζήτησης για μεγιστοποίηση της αποτελεσματικότητας στην παροχή και κατανάλωση ενέργειας. Τα AMI μέσα σε έξυπνο δίκτυο επιτρέπουν την βελτιστοποίηση της κατανάλωσης ενέργειας, ενημερώνοντας τους χρήστες έγκαιρα για τις τιμές, την αρχή και το τέλος των περιόδων αιχμής της κατανάλωσης, το άθροισμα της κατανάλωσης, συναγερμούς κτλ [2].

2.2 Μηχανική μάθηση

Υπάρχουν διαφορετικοί τρόποι που ένας αλγόριθμος μπορεί να μοντελοποιήσει ένα πρόβλημα βασισμένος στα δεδομένα εισόδου. Είναι δημοφιλές στα βιβλία μηχανικής μάθησης και τεχνητής νοημοσύνης να εξετάζεται ο τρόπος μάθησης που ένας αλγόριθμος μπορεί να υιοθετήσει. Υπάρχουν μόνο μερικοί βασικοί τρόποι εκμάθησης ή μοντέλα εκμάθησης που ένας αλγόριθμος μπορεί να χρησιμοποιήσει. Θα αναφερθεί κάθε μοντέλο εκμάθησης με λίγα παραδείγματα από αλγορίθμους και τύπους προβλημάτων που ταιριάζει στο καθένα. Αυτή η ταξινόμηση ή ο τρόπος οργάνωσης των αλγορίθμων είναι χρήσιμος, καθώς αναγκάζει τον χρήστη να σκεφτεί τον ρόλο των δεδομένων εισόδου και το μοντέλο επεξεργασίας και να επιλέξει τον κατάλληλο αλγόριθμο για το πρόβλημα, με στόχο τα βέλτιστα αποτελέσματα. Παρακάτω αναλύονται οι τρεις διαφορετικές κατηγορίες αλγορίθμων μηχανικής μάθησης με βάση τον τρόπο εκμάθησης.

2.2.1 Επιβλεπόμενη μάθηση

Τα δεδομένα εισόδου καλούνται δεδομένα εκπαίδευσης και είναι γνωστά τα αποτελέσματά τους (κλάσεις). Τέτοια προβλήματα ορίζονται, όταν ένα παράδειγμα ταξινομείται σε αρνητική κλάση ή θετική κλάση ή αναζητείται αριθμητικό αποτέλεσμα σε μια ορισμένη χρονική περίοδο (παλινδρόμηση), ενώ έχει προηγηθεί εκπαίδευση μοντέλου με ζευγάρια δεδομένων αποτελεσμάτων. Ένα μοντέλο χτίζεται στη φάση της εκπαίδευσης κατά την οποία απαιτείται να κάνει προβλέψεις και να τις διορθώσει όταν είναι λάθος. Η διαδικασία της εκπαίδευσης συνεχίζει μέχρι το μοντέλο να επιτύχει το επίπεδο ευστοχίας στα δεδομένα εκπαίδευσης. Τέτοια προβλήματα είναι τα προβλήματα ταξινόμησης και παλινδρόμησης. Κάποιοι από τους δημοφιλείς αλγορίθμους είναι η λογιστική παλινδρόμηση και τα νευρωνικά δίκτυα.

2.2.2 Μη επιβλεπόμενη μάθηση

Τα δεδομένα εισόδου σε αυτούς τους αλγορίθμους δεν έχουν έχουν γνωστά αποτελέσματα. Ένα μοντέλο προετοιμάζεται, εξάγοντας χαρακτηριστικά από τα δεδομένα εισόδου. Εν συνεχεία, εφαρμόζονται γενικοί κανόνες που βασίζονται στα υπάρχοντα χαρακτηριστικά. Αυτό συνήθως συμβαίνει μέσω κάποιας μαθηματικής διαδικασίας που μειώνει συστηματικά την επαναληψιμότητα του αλγορίθμου ή με οργάνωση των δεδομένων βάσει ομοιότητας. Τέτοιου είδους προβλήματα είναι η συσταδοποίηση, η μείωση διάστασης και η εκπαίδευση μέσω κανόνων συσχέτισης. Αντιπροσωπευτικοί αλγόριθμοι είναι το K-Means και το Principal Component Analysis (PCA).

2.2.3 Ημι-επιβλεπόμενη μάθηση

Τα δεδομένα εισόδου είναι μια μείξη γνωστών και άγνωστων δυαδικών χαρακτηριστικών. Υπάρχει μια επιθυμητή πρόβλεψη τους προβλήματος, αλλά το μοντέλο πρέπει να μάθει τη δομή για να οργανώσει τα δεδομένα αλλά και να κάνει τις τελικές προβλέψεις. Τέτοια προβλήματα είναι η ταξινόμηση και η παλινδρόμηση. Οι αλγόριθμοι που χρησιμοποιούνται είναι

επέκταση άλλων ευέλικτων μεθόδων που κάνουν υποθέσεις για το μοντέλο χωρίς τα δυαδικά χαρακτηριστικά [4].

2.3 Μετρικές μηχανικής μάθησης

Για να γίνει αξιολόγηση της ταξινόμησης, χρειάζεται να ληφθούν υπόψη κάποια κριτήρια και μετρικές. Ο ρυθμός ευστοχίας ή η μέση τιμή του λάθους αδυνατούν να μας περιγράψουν σαφώς τον ταξινομητή, οπότε εισάγεται η έννοια του confusion matrix. Σύμφωνα με τον πίνακα μετράμε τις εξής τιμές:

		Πρόβλεψη		
		p	**n**	**Συνολικά**
Πραγματική Τιμή	**p′**	True Positive	False Negative	P′
	n′	False Positive	True ενΝεγατιε	N′
	Συνολικά	P	N	

Σχήμα 2.1: Confusion Matrix

TP=πλήθος των σωστών προβλέψεων στο θετικό αποτέλεσμα
TN=πλήθος των σωστών προβλέψεων στο αρνητικό αποτέλεσμα
FN=πλήθος των λανθασμένων προβλέψεων στο θετικό αποτέλεσμα (αρνητική πρόβλεψη)
FP=πλήθος των λανθασμένων προβλέψεων στο αρνητικό αποτέλεσμα (θετική πρόβλεψη)

Με τις παραπάνω τιμές είναι δυνατό να δομηθούν τα κριτήρια ευστοχίας του συστήματος. Οι τέσσερις βασικοί άξονες της μέτρησης είναι το ποσοστό αναγνώρισης DR (Detection Rate), το ποσοστό λάθος συναγερμού FPR(False Positive Rate), το ποσοστό της ευστοχίας (Accuracy) και το F1 score, που είναι ένας συνδυασμός μετρικών, για να αποκτηθεί μια γενικότερη εικόνα της ακρίβειας του συστήματος.

Με τα πλήθη προβλέψεων να έχουν οριστεί έχουμε τις θεμελιώδεις μετρικές:

$$DR = \frac{TP}{TP+FN},\ FPR = \frac{FP}{FP+TN}$$
$$Accuracy = \frac{TP+TN}{TP+FP+FN+TN},\ F1 = 2\frac{precision \cdot recall}{precision+recall}$$
δεδομένου $Precision = \frac{TP}{TP+FP}$, $Recall = DR = \frac{TP}{TP+FN}$.

Ακόμη θα χρησιμοποιηθεί το ποσοστό αναγνώρισης του Bayes και η αντίστοιχή του άρνηση για να μας δώσουν μια πιθανοτική σκοπιά για τη διάκριση απάτης από τη φυσιολογική κατανάλωση. Η $P(I)$ είναι η πιθανότητα να υπάρχει απάτη στα δεδομένα και αυτό σε πραγματικές συνθήκες δεν είναι εύκολο να υπολογιστεί με ακρίβεια. Το ενδεχόμενο A αντιστοιχεί στον συναγερμό που ενεργοποιείται στην αναγνώριση απάτης. Μπορεί στα συγκεκριμένα δεδομένα να οριστεί ως η πιθανότητα μια τυπική μέρα να βρεθεί απάτη στις μετρήσεις.

Τεχνικά οι δύο πιθανότητες ορίζονται ως εξής:

- $P(I|A)-$ ένας συναγερμός πραγματικά ενδεικνύει απάτη
- $P(\neg I|\neg A)-$ η μη ενεργοποίση του συναγερμού υποδηλώνει απουσία απάτης

Αυτό που έχει σημασία είναι και οι δύο πιθανότητες να παραμείνουν όσο το δυνατόν μεγαλύτερες [3].

Μπορούμε να αντιστοιχίσουμε τα βασικά κριτήρια με τις πιθανότητες στο ποσοστό αναγνώρισης του Bayes.

$$P(I|A) = \frac{P(I)P(A|I)}{P(I)P(A|I)+P(\neg I)\cdot P(A|\neg I)},\ P(\neg I|\neg A) = \frac{P(\neg I)\cdot P(\neg A|\neg I)}{P(\neg I)\cdot P(\neg A|\neg I)+P(I)\cdot P(\neg A|I)}$$
$$\text{για } P(A|I) = DR,\ P(A|\neg I) = FPR,\ P(\neg A|I) = 1 - P(A|I),\ P(\neg A|\neg I) = 1 - P(A|\neg I)$$
$$\text{έχουμε } BDR = \frac{P(I)DR}{P(I)\cdot DR+P(\neg I)\cdot FPR}$$

Κεφάλαιο 3

Περιγραφή και οργάνωση δεδομένων

Απαραίτητη φάση της διαδικασίας εξόρυξης δεδομένων είναι η συλλογή και η προετοιμασία των δεδομένων. Η φάση κατανόησης των δεδομένων περιλαμβάνει τη συλλογή και εξερεύνησή τους. Ρίχνοντας μια πιο προσεκτική ματιά στα δεδομένα, καθίσταται εφικτός ο καθορισμός του πόσο καλά μπορούμε να αντιμετωπίσουμε το πρόβλημα. Η προσεκτική προετοιμασία δεδομένων μπορεί να βελτιώσει δραστικά τις πληροφορίες που μπορούν να εξαχθούν από την εξόρυξη δεδομένων [19].

3.1 Περιγραφή δεδομένων

Τα δεδομένα υπό εξερεύνηση αποτελούνται από καταναλώσεις έξυπνων μετρητών για σχεδόν 5.000 οικιακά νοικοκυριά και 600 επιχειρήσεις. Πιο συγκεκριμένα, προέρχονται από την Commision for Energy Regulation (CER), η οποία αποτελεί την ανεξάρτητη αρχή για την ενέργεια και το νερό της Ιρλανδίας [9]. Οι ενδιαφερόμενοι πελάτες παρείχαν εθελοντικά τα δεδομένα των καταναλώσεων και ερωτηματολόγια για τις καταναλωτικές τους συνήθειες και τις υποδομές τους, πράγμα που έδωσε τη δυνατότητα να αναλυθούν διεξοδικά τα δεδομένα. Τα αντιπροσωπευτικά αυτά δείγματα συλλέχθηκαν ανώνυμα σε χρονικό παράθυρο σχεδόν 2 ετών, από το 2009-2011 και με συχνότητα λήψης 30 λεπτά. Οι πληροφορίες των έξυπνων μετρητών είναι αποθηκευμένες σε έξι διαφορετικά αρχεία κειμένου (.txt), που καθένα έχει 24 εκατομμύρια καταχωρήσεις οι οποίες αντιστοιχούν σε διάφορες μετρήσεις ενέργειας. Ο Πίνακας 3.1 αντιπροσωπεύει ένα μικρό δείγμα των αρχείων κειμένου, το οποίο αποτελείται από 3 στήλες. Η πρώτη στήλη αναπαριστά το ID του έξυπνου μετρητή, που είναι ξεχωριστό για κάθε πελάτη. Η δεύτερη στήλη δείχνει κωδικοποιημένα την ημερομηνία και την ώρα της συγκεκριμένης μέτρησης, ενώ η τρίτη στήλη αποτελεί την αντίστοιχη μέτρηση ενέργειας που καταναλώθηκε σε κιλοβατώρες (kWh).

ID Μετρητή	Κωδικοποιημένη ημερομηνία/ώρα	Κατανάλωση ενέργειας kWh
1392	19503	0.140
1392	19504	0.138
...	...	...
1187	22028	1.367
1187	22029	1.425
1392	19940	0.234

Πίνακας 3.1: Στιγμιότυπα αρχείου δεδομένων

3.1.1 Επισκόπηση χρονοσειρών

Έχοντας διευκρινίσει, λοιπόν, την προέλευση και τη δομή των δεδομένων, αξίζει να γίνει μια αναλυτική επισκόπησή τους. Επειδή, καθίσταται αδιανόητη η μελέτη 4.500 ετήσιων καταναλώσεων, επιλέγονται ομάδες που να αντιπροσωπεύουν τον πληθυσμό. Για την ομαδοποίηση των δεδομένων επιλέγεται ο αλγόριθμος K-Means, που αναλύεται στο Κεφάλαιο 5.2.1. Δημιουργήθηκαν 6 συστάδες (ομάδες) που να εκφράζουν είτε τη μορφή της καμπύλης είτε το ύψος της ημερήσιας κατανάλωσης. Με αυτό τον τρόπο ομαδοποιούνται τα δεδομένα και διευκολύνεται η διαδικασία παρατήρησης των χαρακτηριστικών 6 διαφορετικών ομάδων βάσει 2 διαφορετικών κριτηρίων. Επιλέχθηκαν 6 συστάδες, καθώς έτσι επιτυγχάνεται ομοιομορφία στο πλήθος των μελών. Έτσι, κάθε συστάδα έχει ικανοποιητικό μέγεθος πληθυσμού, το οποίο την αντιπροσωπεύει.

Με δεδομένου την ύπαρξη 4.500 καταναλωτών με ημερήσιες μετρήσεις για ένα έτος, δημιουργήθηκε πίνακας $m \times n$ με m παρατηρήσεις και n χαρακτηριστικά. Σαν είσοδος λοιπόν του αλγορίθμου K-Means χρησιμοποιήθηκαν 4.500 καταναλωτές με 365 χαρακτηριστικά που συσταδοποιήθηκαν με και χωρίς κανονικοποίηση στα ετήσια διανύσματα $\{x_1, ..., x_n\}$. Η συσταδοποίηση με κανονικοποίηση σε εύρος [-1,1] δίνει τη δυνατότητα παρατήρησης της καμπύλης της χρονοσειράς ανεξαρτήτως του επιπέδου κατανάλωσης. Παράλληλα, η κανονικοποίηση επιτυγχάνει ομαλοποίηση της καμπύλης, μειώνοντας τις έντονες διακυμάνσεις της ενέργειας. Από την άλλη πλευρά, χωρίς κανονικοποίηση η συσταδοποίηση επηρεάζεται σημαντικά από το ύψος καταναλώσεων. Στον Πίνακα 3.2 φαίνονται τα αποτελέσματα με τα μέλη κάθε συστάδας.

Παρατηρείται, λοιπόν πως στον Πίνακα 3.2α΄ η συσταδοποίηση βάσει των μορφών των χρονοσειρών έχει 3 πολυμελείς συστάδες, που συνοψίζουν τους 3.074 από τους 4.500 καταναλωτές, που επιλέχθηκαν για τη δοκιμή, δημιουργώντας σχετικά ομοιόμορφες συστάδες. Παράλληλα στον Πίνακα 3.2β΄ η συσταδοποίηση βάσει του ύψους της κατανάλωσης έχει 3 πολυμελείς συστάδες που συνοψίζουν τους 4.226 από τους 4.500 καταναλωτές που επιλέχθηκαν. Πρόκειται για απλούς οικιακούς πελάτες, κρίνοντας από την μέση ημερήσια κατανάλωση κάθε συστάδας. Δεν μπορεί να παραλειφθεί σε αυτό το σημείο το γεγονός πως υπάρχουν 2 ολιγομελείς ομάδες που απαριθμούν αθροιστικά 97 μέλη και έχουν πολλαπλάσιες ημερήσιες καταναλώσεις από τους υπόλοιπους.

Συστάδα	Μέλη
1	532
2	892
3	567
4	944
5	327
6	1238

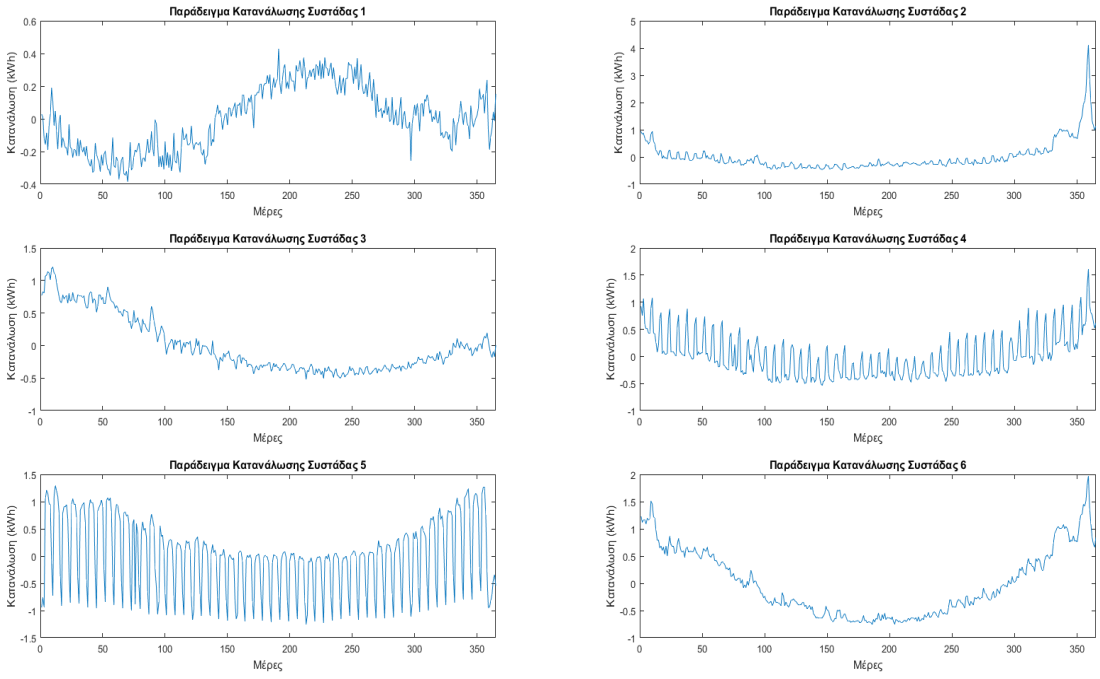
(α′) Συσταδοποίηση με κανονικοποίηση στα ετήσια διανύσματα

Συστάδα	Μέλη	Μέση ημ. κατανάλωση(kWh)
1	1398	11.6
2	1767	23.74
3	13	572.20
4	177	98.54
5	84	243.42
6	1061	40.52

(β′) Συσταδοποίηση στα ετήσια διανύσματα

Πίνακας 3.2: Ομαδοποιήσεις με 2 κριτήρια

Για περαιτέρω εξερεύνηση των κριτηρίων ομαδοποίησης και των συστάδων δημιουργήθηκαν δύο σχήματα που αποτελούνται από παραδείγματα μελών κάθε συστάδας. Αναλυτικότερα, στο Σχήμα 3.2 και στο Σχήμα 3.1 φαίνονται οι καταναλώσεις των κέντρων κάθε συστάδας. Έτσι, δίνεται η δυνατότητα να αναλυθεί η μορφή 6 διαφορετικών ομάδων, αλλά και να παρατηρηθεί ο διαχωρισμός των καταναλωτών και των χρονοσειρών τους με γνώμονα την ημερήσια κατανάλωσή τους κατά τη διάρκεια ενός έτους.



Σχήμα 3.1: Παραδείγματα χρονοσειρών συσταδοποίησης βάσει της μορφής των χρονοσειρών

Όπως φαίνεται στο Σχήμα 3.1, υπάρχουν κάποιες αξιοσημείωτες ομοιότητες και διαφορές μεταξύ των μορφών των καμπυλών.

- Η συστάδα 1 δείχνει μια ύφεση στο τέλος του χειμώνα στην κατανάλωση, η οποία επιστρέφει σε υψηλότερα επίπεδα μέσα στην άνοιξη. Στη συνέχεια παρατηρείται ξανά πτώση της κατανάλωσης στα μέσα του φθινοπώρου.

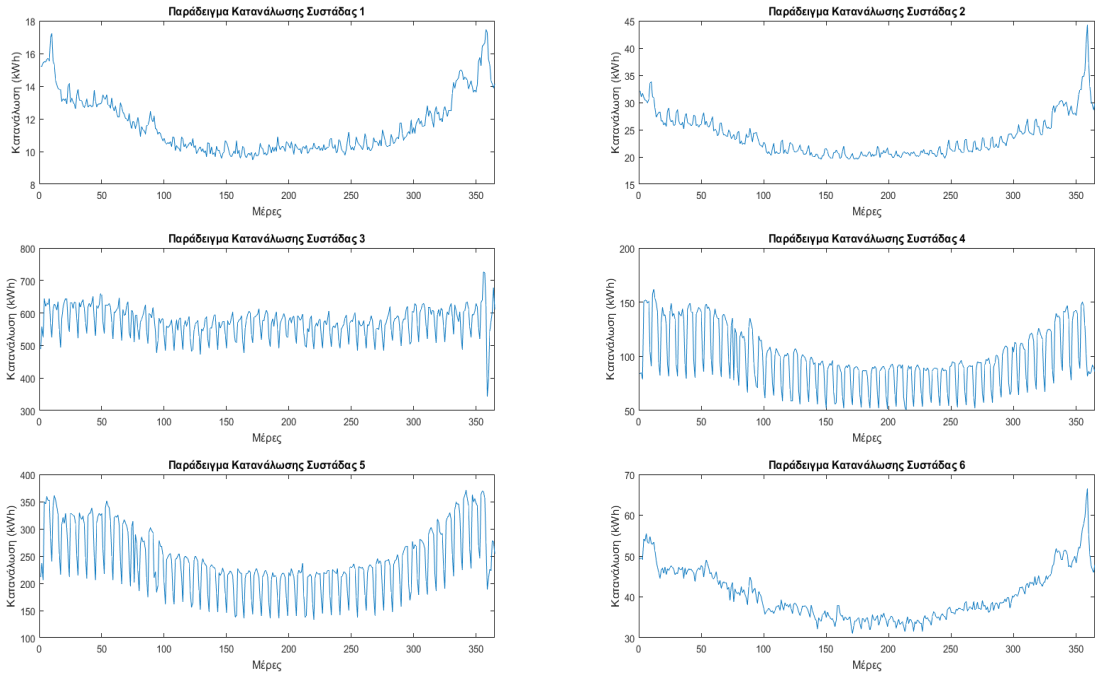
- Η συστάδα 2 θυμίζει σημαντικά λευκό θόρυβο, καθώς δεν παρατηρείται έντονη απόκλιση από την μέση τιμή της καμπύλης, ενώ παράλληλα υπάρχει έντονος βαθμός τυχαιότητας στις διακυμάνσεις με την κατανάλωση να αυξάνεται μόνο τον τελευταίο μήνα του έτους.

- Η συστάδα 3 εμφανίζει μια σχετικά ακανόνιστη αλλά φθίνουσα εν γένει πορεία. Ειδικότερα, οι ελάχιστες τιμές κατανάλωσης ξεκινούν την άνοιξη χωρίς να εμφανίζεται ανοδική πορεία μέχρι το τέλος του έτους.

- Η συστάδα 4 έχει εμφανώς αρχικά φθίνουσα τάση, ενώ μετά το καλοκαίρι ξεκινά να αυξάνεται η ημερήσια κατανάλωση ομαλά αρχικά και μετά βίαια.

- Η συστάδα 5 έχει πολύ έντονες και συνεχείς διακυμάνσεις, αλλά κρατά σχεδόν σταθερό μέσο όρο ανά τις ημέρες, καθώς η διακύμναση είναι έντονη αλλά γύρω από μια νοητή γραμμή με ελάχιστη κλίση. Παράλληλα, είναι εμφανές πως στους χειμερινούς μήνες έχουμε αισθητή αύξηση της κατανάλωσης.

- Η συστάδα 6 δείχνει πως στο ενδιάμεσο του έτους εμφανίζεται μείωση της κατανάλωσης, ενώ κοντά στον χειμώνα, όπου ξεκινά και τελειώνει η χρονοσειρά, παρατηρείται αύξησή της.



Σχήμα 3.2: Παραδείγματα χρονοσειρών συσταδοποίησης βάσει του ύψους της κατανάλωσης

Στο παραπάνω Σχήμα εμφανίζονται τα παραδείγματα των συστάδων που δημιουργήθηκαν βάσει του ύψους των ημερήσιων καταναλώσεων με τις εξής επισημάνσεις:

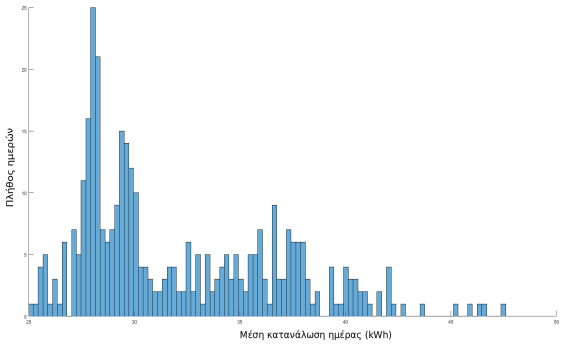
- Η συστάδα 1, που αποτελεί τη 2η μεγαλύτερη συστάδα, περιέχει τιμές που κυμαίνονται γενικώς γύρω στις 11.6 kWh με 2 κύριες αλλαγές στη μονοτονία από φθίνουσα σε αύξουσα.
- Η συστάδα 2 δεν παρουσιάζει κάποιο ιδιαίτερο χαρακτηριστικό, καθώς εμφανίζει εξαιρετικές ομοιότητες με τη συστάδα 1, με μόνες διαφορές την μικρότερη κλίση στις μονοτονίες και την υψηλότερη κατανάλωση.
- Η συστάδα 3 εμφανίζει πολύ ξεχωριστή συμπεριφορά, όντας καμπύλη μιας επιχείρησης με μεγάλες ενεργειακές απαιτήσεις που έχει μεγάλη και συνεχή ζήτηση ενέργειας σε όλη τη διάρκεια του έτους.
- Η συστάδα 4 εμπεριέχει καταναλωτές μικρομεσσαίων επιχειρήσεων με έντονες διακυμάνσεις και σχετικά μεγάλες καταναλώσεις.
- Η συστάδα 5 περιγράφει καταναλωτές επιχειρήσεων με έντονη διακύμανση της κατανάλωσης, ξεκινώντας με φθίνουσα πορεία και ακολουθώντας με ομαλή αύξουσα πορεία μετά το καλοκαίρι.
- Η συστάδα 6 περιλαμβάνει ένα μεγάλο μέρος των οικιακών καταναλωτών που έχουν προσγειωμένες τιμές ημερήσιας κατανάλωσης αλλά και μικρές διακυμάνσεις στη μονοτονία και στις μετρήσεις τους.

Ιστογράμματα Συχνοτήτων

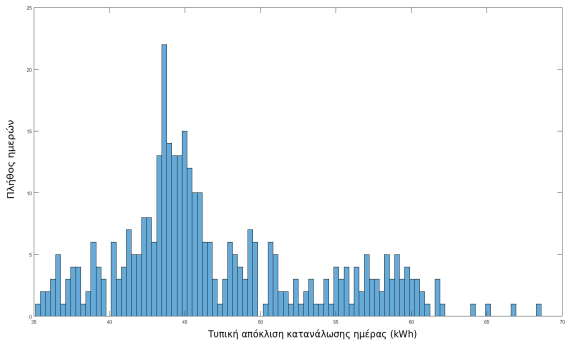
Για τη δημιουργία των ιστογραμμάτων απαιτούνται διανύσματα δεδομένων που επιλέχθηκαν να είναι ο μέσος όρος και η τυπική απόκλιση ως προς τις δύο διαστάσεις του πίνακα δεδομένων 4.500×365. Παίρνοντας το μέσο όρο και την τυπική απόκλιση της κάθετης συνιστώσας, δημιουργούνται δύο διανύσματα που αποτελούνται από τον μέσο όρο και την τυπική απόκλιση της κατανάλωσης όλων των πελατών ανά ημέρα. Αντίστοιχα, αν επαναληφθεί η διαδικασία για την οριζόντια συνιστώσα, εξάγεται ο μέσος όρος και η τυπική απόκλιση της ημερήσιας κατανάλωσης ανά πελάτη. Ο σκοπός ενός ιστογράμματος είναι να αναπαριστά γραφικά την κατανομή των δεδομένων με εξάρτηση από μια μεταβλητή. Το ιστόγραμμα χρησιμοποιείται ευρέως για να δώσει απάντηση στα παρακάτω ερωτήματα[8]:

1. Τι είδους κατανομή ακολουθεί ο πληθυσμός;
2. Πού τοποθετούνται τα δεδομένα στον οριζόντια άξονα;
3. Πόσο αραιά είναι;
4. Υπάρχει εμφανής συμμετρία ή κυρτότητα;
5. Υπάρχουν ανωμαλίες στα δεδομένα;

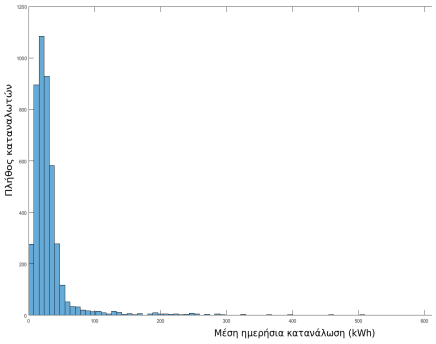
Εδώ διαχωρίζονται τα δεδομένα σε διανύσματα που αφορούν τις καταναλώσεις όλων των πελατών ανά ημέρα και τις ημερήσιες καταναλώσεις ανά πελάτη. Με αυτό τον τρόπο, παρατηρείται πώς διαχωρίζονται οι μέρες βάσει της ημερήσιας κατανάλωσης όλων των πελατών και οι καταναλωτές βάσει της ημερήσιας κατανάλωσής τους. Έτσι, μπορούμε να παρατηρήσουμε ποσοτικά πόσες kWh καταναλώνονται σε μία μέρα από όλους τους πελάτες, αλλά και πόσες kWh καταναλώνει κάθε πελάτης ανά ημέρα. Χρησιμοποιώντας τους μέσους όρους του πίνακα δεδομένων, δημιουργούνται δύο διανύσματα, το διάνυσμα μέσης κατανάλωσης ανά ημέρα και το διάνυσμα μέσης ημερήσιας κατανάλωσης ανά καταναλωτή. Παράλληλα, είναι ιδιαίτερα χρήσιμη η παρατήρηση της τυπικής απόκλισης των δεδομένων μεταξύ τους και του βαθμού συνέπειάς τους, παρακολουθώντας τα ιστογράμματα τυπικής απόκλισης.



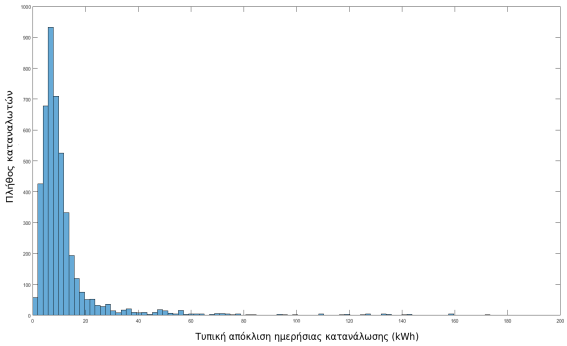
(α′) Διάνυσμα μέσης κατανάλωσης ανά ημέρα



(β′) Διάνυσμα τυπικής απόκλισης κατανάλωσης ανά ημέρα



(γ′) Διάνυσμα μέσης ημερήσιας κατανάλωσης ανά πελάτη



(δ′) Διάνυσμα τυπικής απόκλισης ημερήσιας κατανάλωσης ανά πελάτη

Σχήμα 3.3: Ιστογράμματα για καταναλώσεις

Από τα σχήματα 3.3α′ και 3.3β′ φαίνεται πως και τα δύο ιστογράμματα έχουν θετική λοξότητα σε σχέση με το μέσο όρο του δείγματος. Παρ′ όλα αυτά, υποτίθεται ότι η κατανομή του δείγματος προέρχεται από κανονική κατανομή πληθυσμού. Αντίστοιχα, τα ιστογράμματα των σχημάτων 3.3γ′ και 3.3δ′ δείχνουν επίσης θετική λοξότητα, αλλά με σημαντικά υψηλότερη κορυφή στο διάγραμμα, καθώς πρόκειται για πλήθος καταναλωτών.

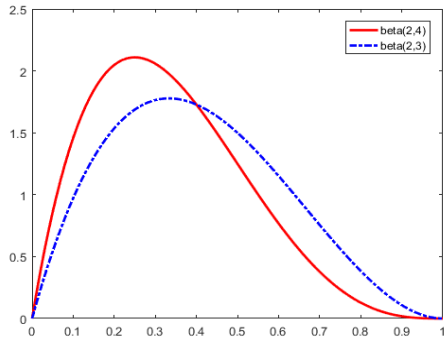
Σε αυτό το σημείο έχει νόημα να προσεγγιστούν οι ερωτήσεις που τέθηκαν παραπάνω.

Γίνεται, λοιπόν, σαφές πως τα δύο πρώτα ιστογράμματα έχουν μεγάλο εύρος και 2 κορυφές, ενώ τα επόμενα έχουν μικρό εύρος και μία μόνο κυριαρχούσα κορυφή. Παράλληλα, δεν επιβαρύνονται τα δεδομένα με ανωμαλίες ή ακραίες ομάδες με ιδιαίτερες καταναλωτικές συμπεριφορές. Ωστόσο, τα τελευταία 2 σχήματα προδίδουν την ύπαρξης καταναλωτών με μεγάλες ενεργειακές ανάγκες, αλλά λόγω του μικρού τους πλήθους δεν απαιτείται περαιτέρω εξερεύνηση προς τη συγκεκριμένη κατεύθυνση.

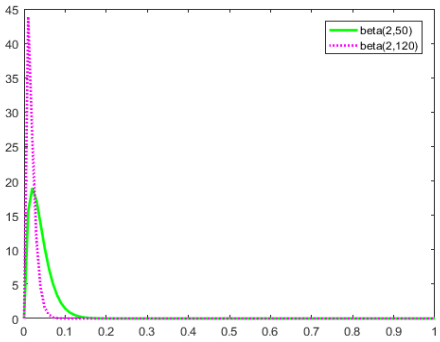
Γενικότερα, τα ιστογράμματα περιγράφονται από μη συμμετρικές καμπύλες με εξόγκωση προς τα αριστερά και μεγάλη ουρά προς τα δεξιά ($skewness > 0$). Για την προσέγγιση των κατανομών των ιστογραμμάτων χρησιμοποιήθηκε η κατανομή Βήτα, καθώς η συνάρτηση πυκνότητάς της είναι πολύ ευέλικτη στην αναπαράσταση μεγεθών και πιθανοτήτων [13]. Υπάρχουν δύο παράμετροι που θα συνεκτιμηθούν ταυτοχρόνως και θα καθορίσουν αν η κατανομή έχει επικρατούσα τιμή στο διάστημά της και αν αυτή είναι συμμετρική. Η κανονική Βήτα κατανομή παρέχει την πυκνότητα πιθανότητας της τιμής x στο διάστημα(0,1):

$$Beta(\alpha,\beta) : prob(x\|\alpha,\beta) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha,\beta)} \text{ όπου } B \text{ είναι η βήτα συνάρτηση}$$
$$B(\alpha,\beta) = \int_0^1 t^{\alpha-1}(1-t)^{\beta-1}dt$$

Για την προσέγγιση των σχημάτων 3.3α΄ και 3.3β΄ χρησιμοποιήθηκαν οι κατανομές $Beta(2,4)$ και $Beta(2,3)$, ενώ τα σχήματα 3.3γ΄ και 3.3δ΄ αντιστοιχίζονται με τις κατανομές $Beta(2,50)$ και $Beta(2,120)$, δημιουργώντας σε κάθε παράδειγμα μια επικρατούσα τιμή. Παρακάτω μπορεί να φανεί η αναπαράστασή τους στο Σχήμα 3.4.



(α΄) Προσέγγιση Βήτα κατανομής στα σχήματα 3.3α΄ και 3.3β΄



(β΄) Προσέγγιση Βήτα κατανομής στο σχήματα 3.3γ΄ και 3.3δ΄

Σχήμα 3.4: Εύρεση συνάρτησης πυκνότητας πιθανότητας με Βήτα κατανομή

Ένα ακόμη απαραίτητο στάδιο στη μελέτη ιστογραμμάτων θετικής λοξότητας είναι η ποσοτικοποίηση μετρικών που να συνοψίζουν τα δεδομένα. Για αυτό το στάδιο επιλέχθηκαν ο μέσος όρος, ο διάμεσος και η επικρατούσα τιμή. Τα αποτελέσματα για κάθε ιστόγραμμα μπορούν να φανούν στον Πίνακα 3.3.

Μέτρο	Σχήμα 3.3α΄	Σχήμα 3.3β΄	Σχήμα 3.3γ΄	Σχήμα 3.3δ΄
Μέσος Όρος	31.99	42.61	31.99	12.4111
Διάμεσος	29.82	40.24	23.85	8.34
Επικρατούσα Τιμή	24.71	30.44	23.50	9.95

Πίνακας 3.3: Ποσοτικά μέτρα περιγραφής ιστογραμμάτων

3.1.2 Μοντελοποίηση εποχιακών δεικτών

Για βαθύτερη κατανόηση των χρονοσειρών γίνεται εκτίμηση της εποχιακής και μη εποχιακής καταναλωτικής τάσης με τη χρήση παραμετρικών μοντέλων. Με αυτό τον τρόπο θα καταστεί δυνατή η παρατήρηση της επαναληψιμότητας και των μορφών των καταναλώσεων. Για να γίνει αυτό, χρησιμοποιείται αρχικά ο αλγόριθμος K-Means για την συσταδοποίηση των καταναλωτών σε τέσσερις συστάδες βάσει της μέσης ημερήσιας κατανάλωσης σε ένα έτος. Στη συνέχεια δημιουργείται ένα προφίλ κατανάλωσης για κάθε συστάδα, βρίσκοντας τον μέσο ημερήσιο όρο κατανάλωσης κάθε συστάδας. Χρειάστηκαν 2.000 καταναλωτές για αυτή την ανάλυση με τους περισσότερους (1.800) να ομαδοποιούνται σε δύο ομάδες, υποδεικνύοντας προφίλ οικιακών καταναλωτών. Για να είναι πιο ρεαλιστική η μελέτη, έγινε και προσομοίωση μη τεχνικών απωλειών στο 10% του πληθυσμού.

Ανάλυση Παλινδρόμησης

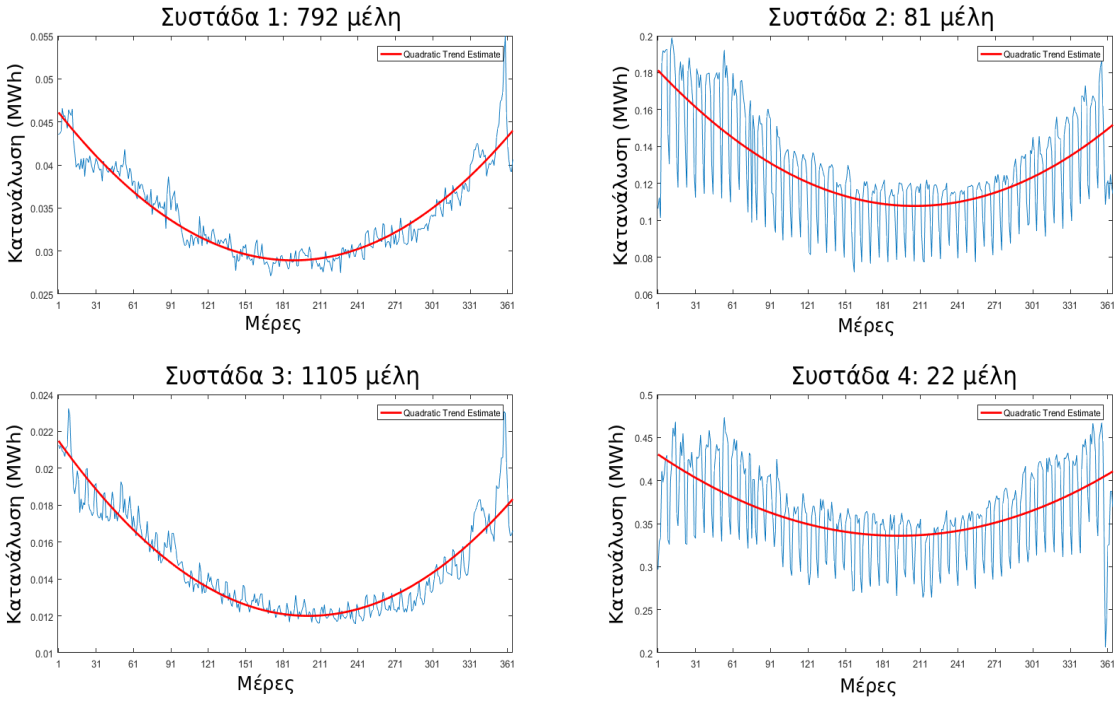
Σκοπός, λοιπόν, αυτού του μέρους είναι να γίνει στατιστική μελέτη του πολυωνυμικού μοντέλου στα δεδομένα μας και να εξεταστεί αν οι χρονοσειρές κάθε συστάδας μπορούν να περιγραφούν με πολυώνυμο δευτέρου βαθμού [14].

$$T_t = \beta_0 + \beta_1 t + \beta_2 t^2$$

Όπως φαίνεται στο Σχήμα 3.5 οι συστάδες μπορούν να χαρακτηριστούν από μια παραβολική καμπύλη με θετικό συντελεστή μεγιστοβάθμιου όρου.

- Η συστάδα 1 αποτελείται από 792 καταναλωτές και η παραβολική καμπύλη τάσης λαμβάνει ελάχιστη τιμή την 189η μέρα του έτους.
- Η συστάδα 2 αποτελείται από 81 καταναλωτές και η παραβολική καμπύλη τάσης λαμβάνει ελάχιστη τιμή την 206η μέρα του έτους.
- Η συστάδα 3 αποτελείται από 1105 καταναλωτές και η παραβολική καμπύλη τάσης λαμβάνει ελάχιστη τιμή την 201η μέρα του έτους.
- Η συστάδα 4 αποτελείται από 22 καταναλωτές και η παραβολική καμπύλη τάσης λαμβάνει ελάχιστη τιμή την 194η μέρα του έτους.

Εύκολα, λοιπόν, εξάγεται το συμπέρασμα πως οι οικιακοί καταναλωτές έχουν την τάση να έχουν πιο ομοιόμορφα κατανεμημένα την παραβολική καμπύλη, ενώ οι επιχειρήσεις έχουν μεγαλύτερο βαθμό τυχαιότητας και λιγότερο συμμετρική καμπύλη ως προς το ελάχιστο σημείο της.



Σχήμα 3.5: Εφαρμογή πολυωνύμου δευτέρου βαθμού

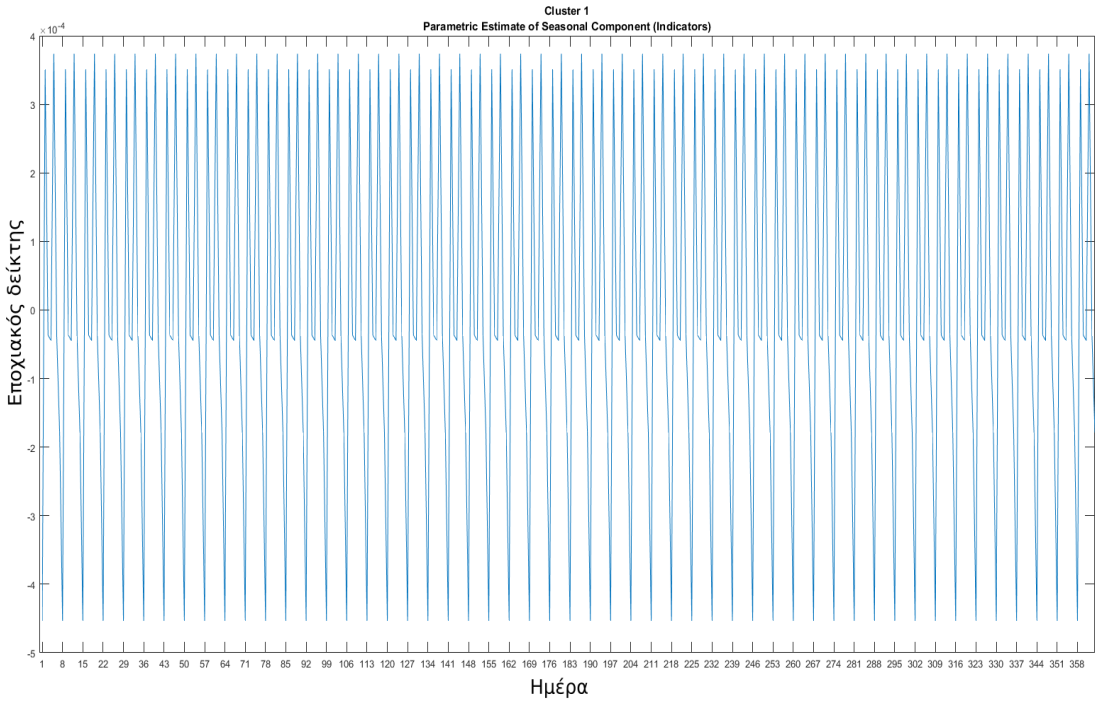
3.1.3 Εκτίμηση εποχιακών δεικτών

Αρχικά για την εκτίμηση των εποχιακών δεικτών απαιτείται η αφαίρεση του πολυώνυμου δευτέρου βαθμού από τις χρονοσειρές των ομάδων [10]. Δεδομένης της μικρής διάρκειας των καταναλώσεων (1 έτος), καθίσταται αδύνατη η εξαγωγή εποχιακών δεικτών ανά μήνα έτους ή ανά εποχή έτους. Για αυτό το λόγο, οι εποχιακοί δείκτες μεταφέρθηκαν ανά ημέρα της εβδομάδας ή ανά ημέρα του μήνα.

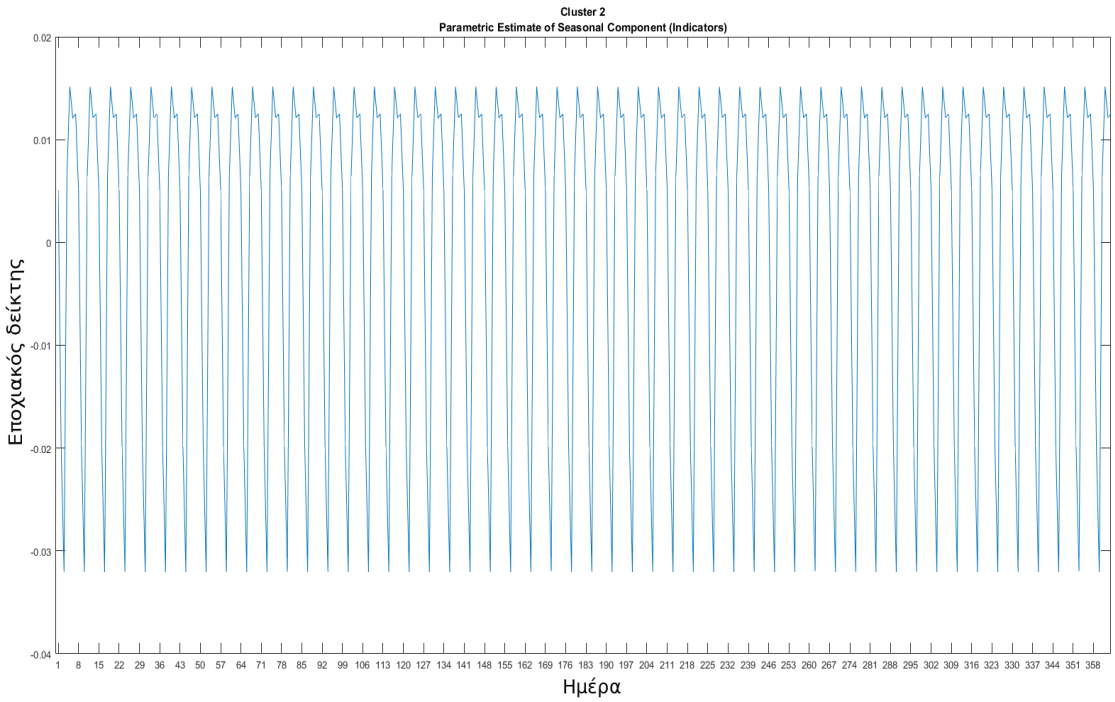
Δημιουργούνται δύο πίνακες με δυαδικά στοιχεία σαν ενδεικτικές μεταβλητές για κάθε ημέρα σε εβδομάδα ή μήνα στο έτος. Η πρώτη ενδεικτική μεταβλητή του πίνακα είναι 1 για την πρώτη μέρα της εβδομάδα ή του μήνα, αλλιώς 0. Η δεύτερη ενδεικτική μεταβλητή είναι 1 για τη δεύτερη μέρα της εβδομάδας ή του μήνα, αλλιώς 0. Στην πρώτη περίπτωση οι δείκτες αναφέρονται στις ημέρες κάθε εβδομάδας, ενώ στη δεύτερη στις ημέρες κάθε μήνα, δημιουργώντας 7 ή 30 δείκτες αντίστοιχα. Για να ολοκληρωθεί η διαδικασία, παλινδρομούνται οι χρονοσειρές χωρίς το πολυώνυμο βάσει των εποχιακών δεικτών. Για την εβδομαδιαία εποχιακότητα εμφανίζονται οι παρακάτω καμπύλες για κάθε ομάδα.

Εκτίμηση με διαστήματα ημέρας ανά εβδομάδα

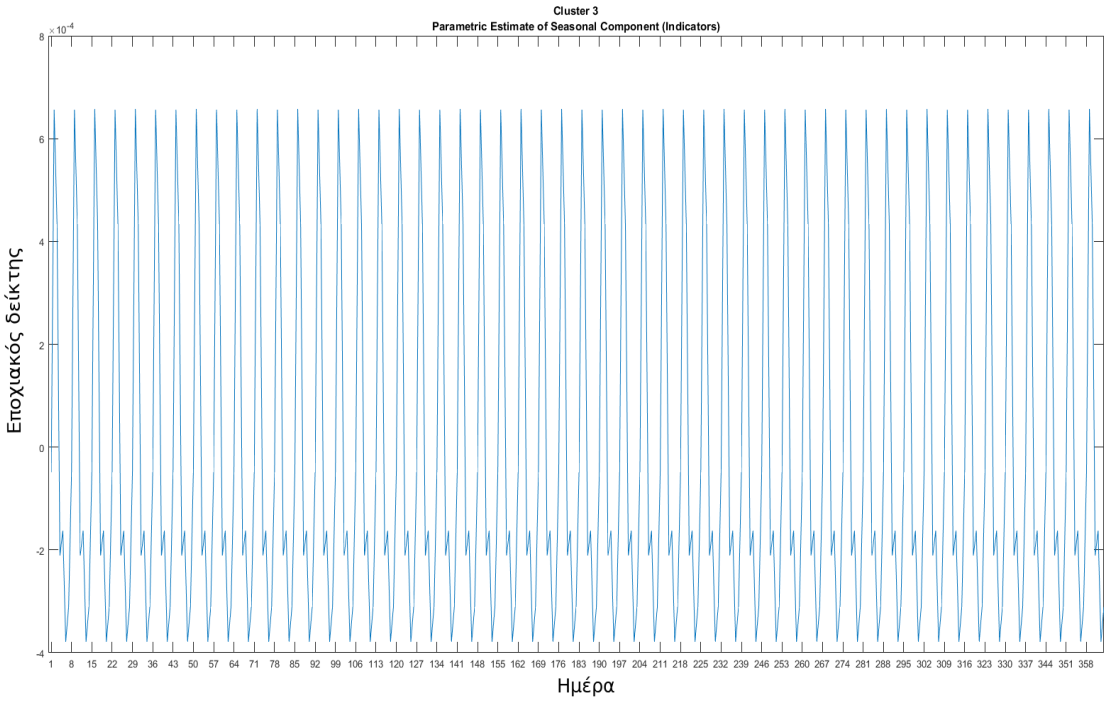
Από την εβδομαδιαία εποχιακότητα, λοιπόν, εύκολα κάποιος αντιλαμβάνεται πως ανάλογα με τον τύπο των καταναλωτών οι μέρες που παρατηρείται μέγιστη και ελάχιστη κατανάλωση διαφέρουν ριζικά. Η πρώτη μέρα του έτους για το έτος που μελετάται είναι Πέμπτη.



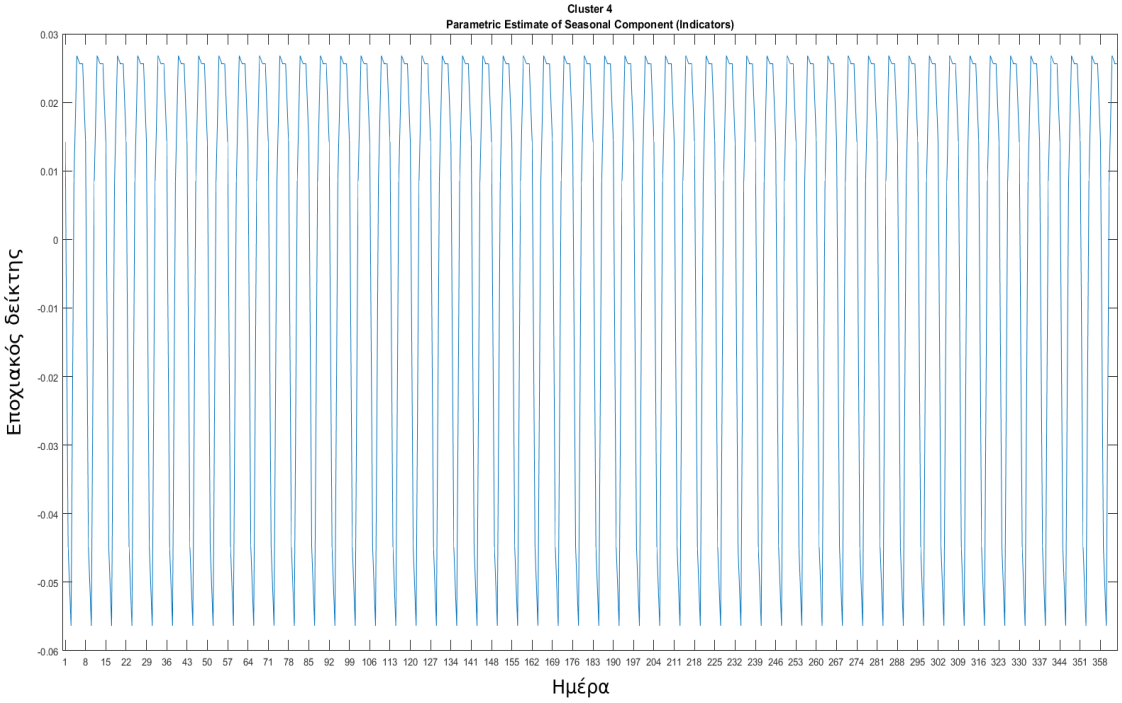
Σχήμα 3.6: Εβδομαδιαία εποχιακότητα ομάδας 1



Σχήμα 3.7: Εβδομαδιαία εποχιακότητα ομάδας 2



Σχήμα 3.8: Εβδομαδιαία εποχιακότητα ομάδας 3



Σχήμα 3.9: Εβδομαδιαία εποχιακότητα ομάδας 4

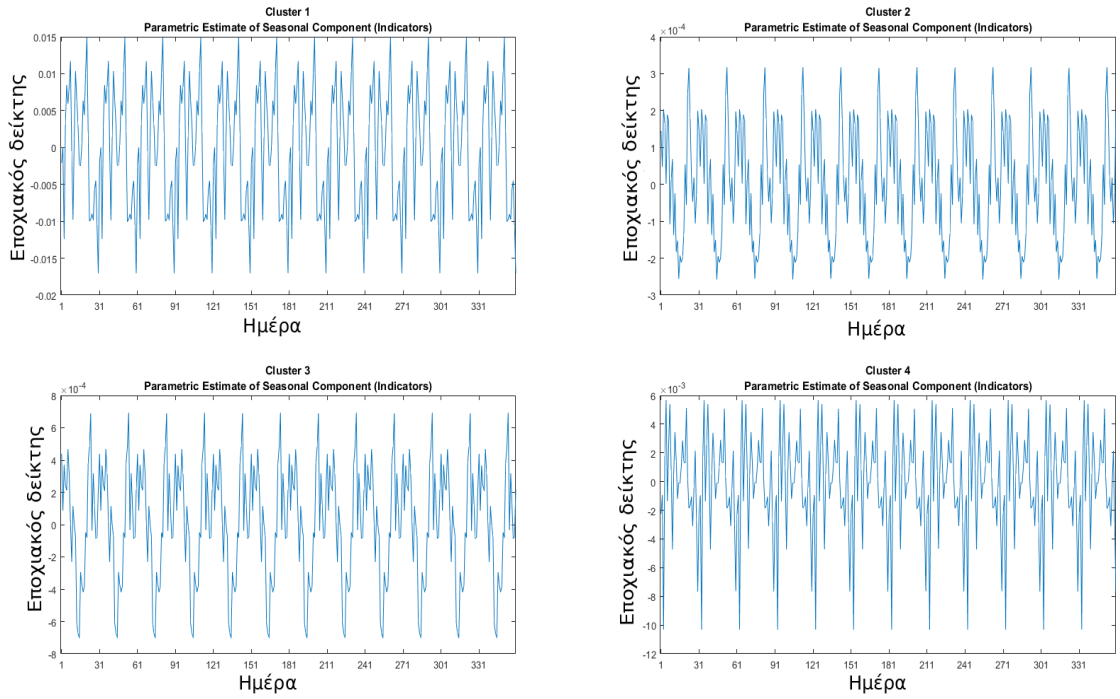
Ειδικότερα από τις χρονοσειρές μπορεί να παρατηρηθεί ότι:

- Για τους καταναλωτές συστάδας 1 (οικιακοί καταναλωτές) έχουμε ελάχιστες καταναλώσεις τις Πέμπτες.
- Για τους καταναλωτές συστάδας 2 (επιχειρήσεις) έχουμε ελάχιστες καταναλώσεις τα Σάββατα.
- Για τους καταναλωτές συστάδας 3 (οικιακοί καταναλωτές) έχουμε ελάχιστες καταναλώσεις τις Τρίτες.
- Για τους καταναλωτές συστάδας 4 (επιχειρήσεις) έχουμε ελάχιστες καταναλώσεις τα Σάββατα.

Ωστόσο, οι πληροφορίες που παράγονται σε αυτή τη δοκιμή δεν είναι ευανάγνωστες λόγω του πλήθους των ακμών στα σχήματα. Παρ΄ όλα αυτά, έχει ενδιαφέρον να παρατηρηθούν οι καμπύλες σε διάστημα ημέρας ανά μήνα που αναμένεται να παρέχουν πιο ευδιάκριτη μορφή.

Εκτίμηση σε διαστήματα ημέρας ανά μήνα

Το διάστημα ενός μήνα αφήνει μεγαλύτερα περιθώρια εποπτείας της χρονοσειράς, ενώ ταυτόχρονα δημιουργεί αποτελέσματα με μεγαλύτερη συνοχή. Από την άλλη πλευρά οι 12 μήνες του έτους δεν μπορούν να εξάγουν πολύ ασφαλή δεδομένα, αν συγκριθούν με τις 52 εβδομάδες.



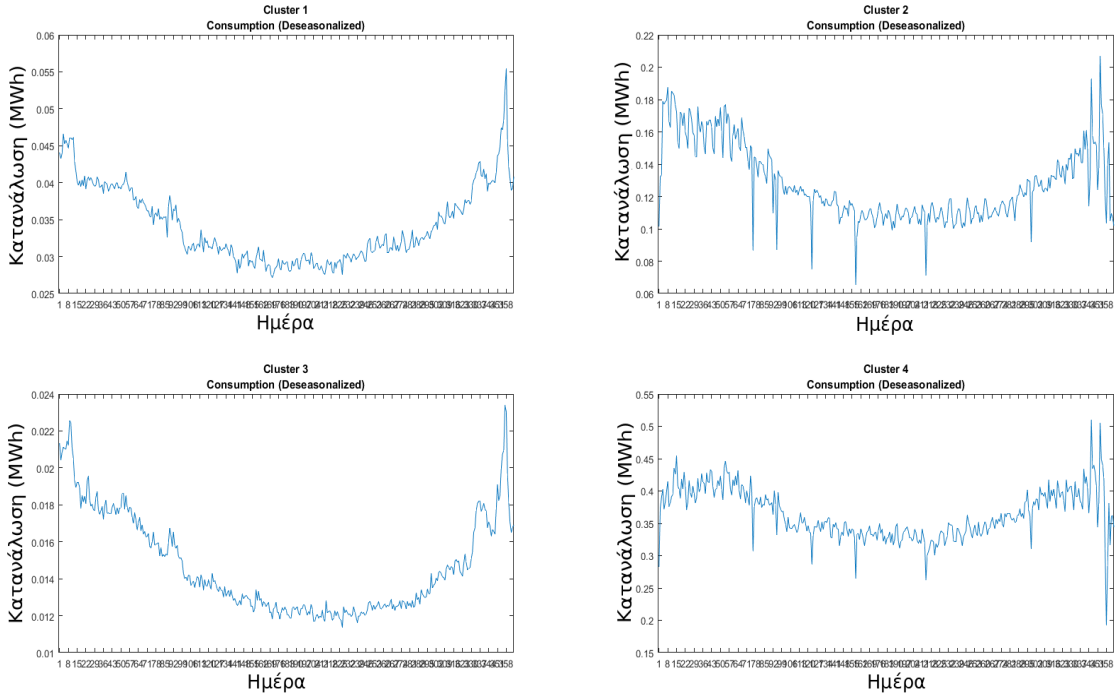
Σχήμα 3.10: Μηνιαία εποχιακότητα

Όπως επισημάνθηκε και παραπάνω και από την μηνιαία εποχιακότητα αντιλαμβανόμαστε πως ανάλογα με τον τύπο των καταναλωτών οι μέρες που εμφανίζεται μέγιστη και ελάχιστη κατανάλωση διαφέρουν σημαντικά. Ειδικότερα:

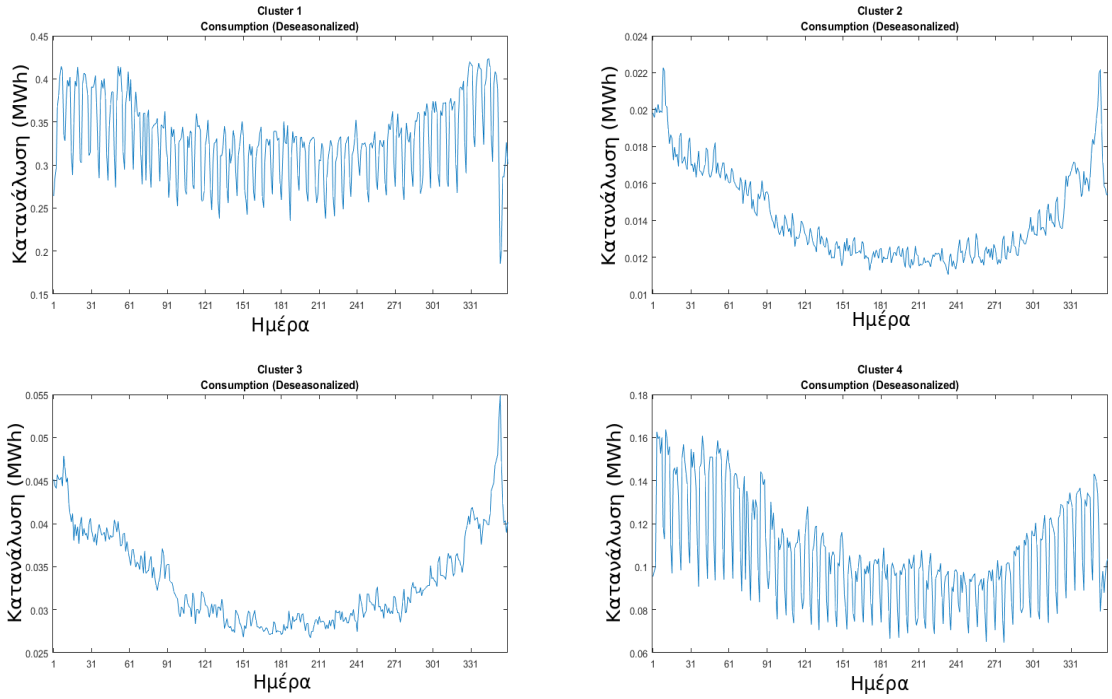
- Για τους καταναλωτές συστάδας 1 (επιχειρήσεις) έχουμε ελάχιστες καταναλώσεις στις 30 του μηνός.
- Για τους καταναλωτές συστάδας 2 (οικιακοί καταναλωτές) έχουμε ελάχιστες καταναλώσεις στις 15 του μηνός.
- Για τους καταναλωτές συστάδας 3 (οικιακοί καταναλωτές) έχουμε ελάχιστες καταναλώσεις στις 15 του μηνός.
- Για τους καταναλωτές συστάδας 4 (επιχειρήσεις) έχουμε ελάχιστες καταναλώσεις στις 3 του μηνός.

3.1.4 Αφαίρεση εποχιακών δεικτών

Σε αυτό το σημείο είναι σημαντικό να παρατηρηθεί η κατανάλωση χωρίς τους εποχιακούς δείκτες. Με αυτό τον τρόπο, καθίσταται ευκολότερη η θεώρηση της μορφής των κυματομορφών και η σύγκρισή τους με τις αρχικές καταναλώσεις του πρώτου μέρους. Αφαιρώντας τα εποχιακά χαρακτηριστικά, οι καμπύλες πλησιάζουν περισσότερο στην παραβολική συνάρτηση. Έτσι, η καταναλωτική τους τάση χωρίς τους εποχιακούς δείκτες γίνεται πιο έντονη και ευδιάκριτη.



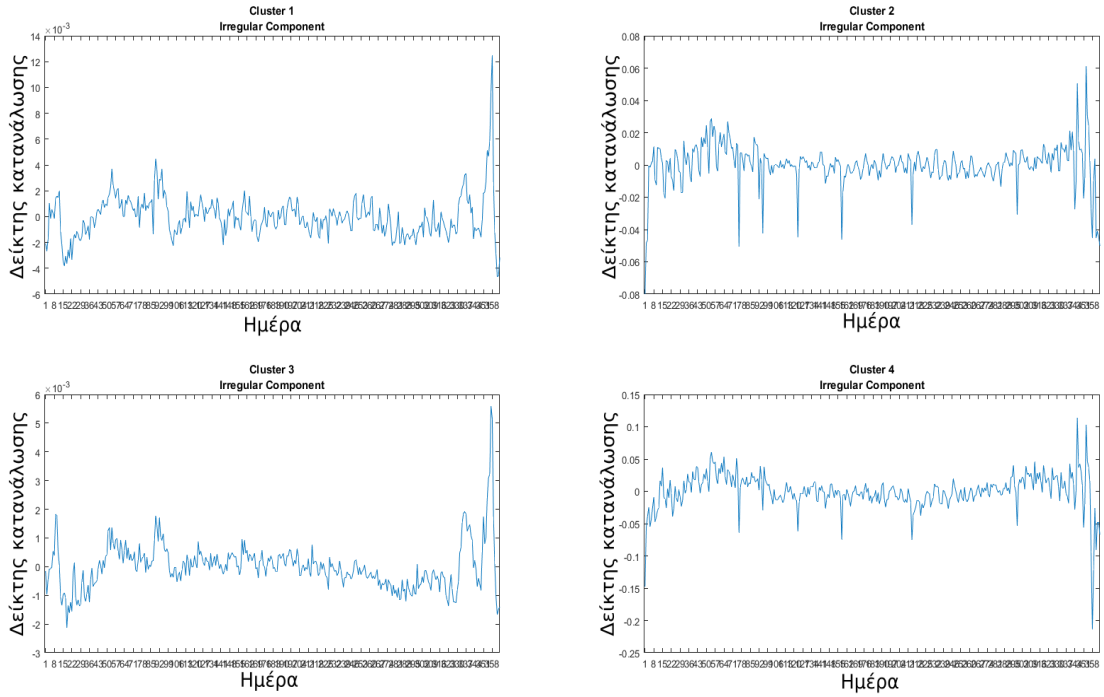
Σχήμα 3.11: Κατανάλωση χωρίς εποχιακούς δείκτες ανά εβδομάδα



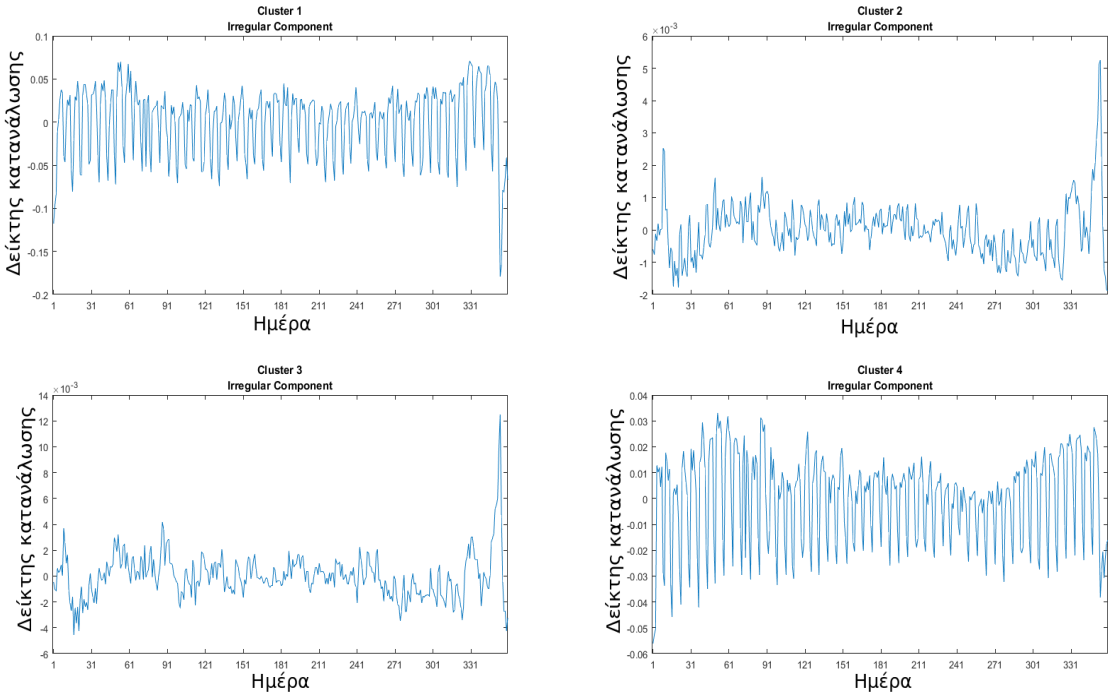
Σχήμα 3.12: Κατανάλωση χωρίς εποχιακούς δείκτες ανά μήνα

3.1.5 Εκτίμηση ακανόνιστης συνιστώσας

Τέλος, έχει ενδιαφέρουν να δούμε τον βαθμό της τυχαιότητας που έχουμε στις καταναλώσεις των συστάδων που δημιουργήθηκαν. Αυτό επιτυγχάνεται αφαιρώντας την εποχιακή χρονοσειρά και την καταναλωτική τάση της αρχικής χρονοσειράς. Με αυτό τον τρόπο, γίνεται σαφές ότι παρά την εποχιακότητα και την τάση οι χρονοσειρές έχουν αισθητό τυχαίο παράγοντα. Η αφαίρεση προκαλεί αλλαγές στο επίπεδο της χρονοσειράς, σταθεροποιώντας έτσι το μέσο όρο της. Γίνεται αντιληπτό πως έχουν μη προβλέψιμα πρότυπα τουλάχιστον με δεδομένα διάρκειας ενός έτους. Τέτοιου τύπου δεδομένα λέγενται στατικές χρονοσειρές.[11]



Σχήμα 3.13: Εκτίμηση ακανόνιστης συνιστώσας με εβδομαδιαία εποχιακότητα



Σχήμα 3.14: Εκτίμηση ακανόνιστης συνιστώσας με μηνιαία εποχιακότητα

Εξερεύνηση ημερών με χαμηλές καταναλώσεις

Για να αντληθούν περαιτέρω χαρακτηριστικά των χρονοσειρών χρειάστηκε η υλοποίηση αλγορίθμου με διπλή συσταδοποίηση. Σύμφωνα με τον αλγόριθμο, πρώτα συσταδοποιούνται οι καταναλωτές με βάση την ημερήσια κατανάλωση, εν συνεχεία για κάθε συστάδα δημιουργείται νέα ομαδοποίηση με βάση την ομοιότητα κάθε ημερήσιας κατανάλωσης. Με αυτό τον τρόπο, μπορεί να παρατηρηθεί ποιες μέρες όμοιων καταναλωτών έχουν παρόμοιες καταναλώσεις. Καθίσταται, έτσι, εφικτό να φιλτράρουμε από τα δεδομένα μας μέρες με χαμηλή κατανάλωση που γνωρίζουμε πως θα δυσκόλευαν το πρόβλημα της ταξινόμησης σε αληθή και αλλοιωμένα δεδομένα.

Για να επιτευχθεί τεχνικά το παραπάνω απαιτείται η συσταδοποίηση του πίνακα 2.000×365 οριζόντια, δημιουργώντας 4 συστάδες και για κάθε συστάδα γίνεται νέα συσταδοποίηση στον ανάστροφο πίνακα, δημιουργώντας 7 νέες συστάδες για κάθε συστάδα της πρώτης συσταδοποίησης. Συνολικά, λοιπόν, δημιουργήθηκαν $4 \times 7 = 28$ συστάδες για την εύρεση κοινών καταναλωτικών συνηθειών.

Τα αποτελέσματα του αλγορίθμου έδειξαν πως μόνο τα Σάββατα μιας αρχικής συστάδας εμφανίζουν έντονη ομοιότητα οικιακών καταναλώσεων. Οι Κυριακές κατά κύριο λόγο συσταδοποιούνται με την υπόλοιπη εβδομάδα δημιουργώντας την εβδομαδιαία τάση. Παράλληλα, παρατηρείται πως ανά περιόδους οι καταναλώσεις δημιουργούν νέες συστάδες, αφήνοντας μόνο τα Σάββατα να σπάνε τη συνεχόμενη συσταδοποίηση. Στον Πίνακα 3.4 φαίνεται πως ακόμη και στα Σάββατα δεν έχουμε απολύτως γεμάτες συστάδες.

Συστάδες Καταναλωτών				
Συστάδες Σαββάτου	Συστάδα 1	Συστάδα 2	Συστάδα 3	Συστάδα 4
Συστάδα 1	0	24	30	19
Συστάδα 2	9	11	0	15
Συστάδα 3	0	9	0	0
Συστάδα 4	42	0	0	0
Συστάδα 5	0	2	0	0
Συστάδα 6	0	4	0	7
Συστάδα 7	0	1	21	10

Πίνακας 3.4: Έλεγχος συσταδοποίησης Σαββάτου

Παρατηρήσεις

Τα εμφανή χαρακτηριστικά εποχιακότητας και η εφαρμογή πολυωνύμου δευτέρου βαθμού στις χρονοσειρές πιστοποιούν ότι τα μοντέλα πρόβλεψης χρονοσειρών μπορούν να χρησιμοποιηθούν με αποτελεσματικότητα. Τα μοντέλα αυτά παρέχουν τη δυνατότητα πρόβλεψης κατανάλωσης του πληθυσμού που επιλέχθηκε για κάποιο χρονικό διάστημα. Αν κάποιος καταναλωτής αποκλίνει σημαντικά από το μοντέλο, τότε θεωρείται πως εμφανίζει ύποπτη καταναλωτική συμπεριφορά. Η πληροφορία αυτή είναι χρήσιμη για ένα σύστημα ταξινόμησης για

τον διαχωρισμό κανονικών και ακανόνιστων χρονοσειρών.

3.2 Προεπεξεργασία και καθάρισμα δεδομένων

Πριν τις αρχικές δοκιμές των ταξινομητών απαιτείται η επιλογή της τελικής μορφής των δεδομένων που θα χρησιμοποιηθούν στο υπόλοιπο σύστημα. Για να μπορέσουν τα δεδομένα να είναι κατανοητά και ξεκάθαρα, χρειάζεται να οργανωθούν ανά ID μετρητή που είναι ξεχωριστός για κάθε πελάτη και εν συνεχεία να οργανωθούν σε συνεχείς χρονικές περιόδους. Λαμβάνοντας υπόψη ότι τα δεδομένα είχαν χρονικό παράθυρο λιγότερο από 2 έτη, επιλέχθηκε πως κάθε καταναλωτής θα πρέπει να έχει ένα γεμάτο έτος μετρήσεων για να μπει σε οποιαδήποτε δοκιμή.

Έτσι, όποιος καταναλωτής έχει πλήρη δεδομένα για όλα τα ημίωρα του έτους από την 1η Ιανουαρίου μέχρι και τη 31η Δεκέμβριου του 2010 εντάσσεται στο τελικό σύνολο δεδομένων. Σε αυτό το στάδιο κάθε καταναλωτής περιγράφεται από ένα διάνυσμα με 17.520 μετρήσεις. Δυστυχώς, ακόμη και για τα σημερινά δεδομένα ένας πίνακας αποτελούμενος από τόσες μετρήσεις για κάθε καταναλωτή είναι δύσκολος στη διαχείριση και χρονοβόρος στην επεξεργασία. Για να δοθεί λύση στο πρόβλημα αυτό, δημιουργήθηκαν δύο είδη πινάκων.

Το πρώτο είδος πίνακα περιέχει μεγάλο όγκο πληροφοριών, ώστε να είναι δυνατές λεπτομερείς επεξεργασίες των δεδομένων, αλλά είναι δύσχρηστος στις δοκιμές, καθώς απαιτεί μεγάλη υπολογιστική δύναμη για να συμπεριληφθεί σε περίπλοκες πράξεις πινάκων. Ειδικότερα, κάθε καταναλωτής περιγράφεται από ένα πίνακα που περιέχει τις μετρήσεις του ανά ημέρα σε ημίωρα ή ανά μήνα σε ώρες ή ανά εβδομάδα σε ώρες κ.ο.κ. Το δεύτερο είδος πίνακα είναι λιγότερο περιεκτικό, αλλά χειρίζεται πολύ πιο εύκολα και γρήγορα από τους αλγορίθμους που χρησιμοποιήθηκαν. Πιο συγκεκριμένα, κάθε καταναλωτής έχει ένα διάνυσμα που περιέχει τις τιμές κατανάλωσης ενός έτους σε ώρες, ημίωρα, μέρες, εβδομάδες ή και μήνες. Για να δοθεί ένα πρακτικό παράδειγμα των δύο ειδών πινάκων, ένας περιγραφικός πίνακας για 2.000 καταναλωτές με ανάλυση σε ώρες ανά μέρα έχει 730.000 γραμμές και 24 στήλες, ενώ ο αντίστοιχος πίνακας για υπολογισμούς έχει 2.000 γραμμές και 24 στήλες. Ουσιαστικά, ο περιγραφικός πίνακας είναι 365 φορές μεγαλύτερος και κρίνεται ακατάλληλος για περίπλοκες πράξεις πινάκων.

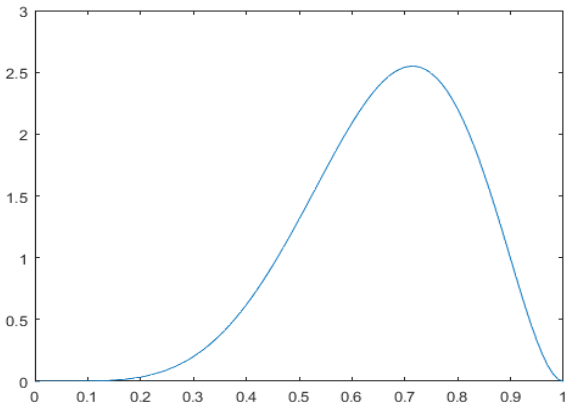
Τα προϊόντα της προεπεξεργασίας και του καθαρίσματος των δεδομένων είναι ένα διάνυσμα με τα ID των μετρητών, ένας πίνακας με ετήσια διανύσματα κατανάλωσης και ένας πίνακας τριών διαστάσεων που περιγράφει αναλυτικά την καταναλωτική συμπεριφορά των πελατών. Το διάνυσμα με τα ID των μετρητών χρησιμοποιείται για την αντιστοίχιση των πελατών με τις ετήσιες καταναλώσεις τους. Ο πίνακας διανυσμάτων κατανάλωσης χρησιμοποιείται για πολύπλοκες και επίπονες πράξεις, ενώ ο αναλυτικός πίνακας για λεπτομερή μελέτη και μικρή επεξεργασία.

3.3 Προσομοίωση απάτης

Δεδομένου ότι οι μετρήσεις που συλλέχθηκαν ήταν από αξιόπιστους καταναλωτές, θα πρέπει να μοντελοποιηθεί η συμπεριφορά με μη τεχνικές απώλειες. Σε αυτή τη διατριβή προσεγγίζεται η περίπτωση της επέμβασης στο μετρητή, κατά την οποία παράνομοι καταναλωτές αλλοιώνουν το σύστημα μέτρησης, για να αναφέρει μικρότερα ποσά. Αυτό μπορεί να συμβεί με χρήση μαγνήτη που παρεμβαίνει στο μετρητή. Παράλληλα, επιθέσεις στο σύστημα μέτρησης μπορούν να επιτευχθούν και με ηλεκτρονικά μέσα (Cyber attacks), αλλοιώνοντας τις τιμές. Σε κάθε περίπτωση, ο καταναλωτής εισάγεται μια μέρα στη διαδικασία της ρευματοκλοπής και ανάλογα με τον τρόπο παρέμβασης, αλλοιώνονται όλα ή μερικά από τα δεδομένα του με σταθερό ή μεταβλητό ρυθμό. Όπως γίνεται αντιληπτό, μπορεί να εισαχθεί μεγάλος βαθμός τυχαιότητας στη ρευματοκλοπή. Στην περίπτωση της φυσικής επίθεσης, είναι ευκολότερος ο προσδιορισμός της απάτης και σχετικά σταθερός ο βαθμός αλλοίωσης των δεδομένων, ενώ στις επιθέσεις με ηλεκτρονικά μέσα είναι δυνατό να παρεισφρήσουν πολλοί εξωτερικοί και άγνωστοι παράγοντες, που μπορεί να έχουν στόχο την απόκρυψη και ελαχιστοποίηση της κλοπής, ώστε να μην γίνεται εύκολα αντιληπτή.

3.3.1 Τύποι απάτης

Έτσι, μοντελοποιήθηκαν 4 συμπεριφορές που καθεμία εισάγει έναν διαφορετικό παράγοντα [23]. Για όλους τους τύπους απάτης θεωρείται ότι μια μέρα ο καταναλωτής εισάγεται στην ρευματοκλοπή και από εκείνη τη μέρα χρησιμοποιεί με διαφορετικούς ρυθμούς το σύστημα αλλοίωσης. Παράλληλα, για την ένταση της κλοπής χρησιμοποιούνται διαφορετικές κατανομές, για να επιλέγεται από αυτές η ένταση της επίθεσης. Η κατανομή Βήτα με παραμέτρους 6 και 3 (Σχήμα 3.15) θεωρήθηκε η πιο ρεαλιστική, καθώς έχει κορυφή στο 0.7 και σχετικά μεγάλο εύρος τιμών, εισάγοντας βαθμό τυχαιότητας, αλλά με κατεύθυνση τις έντονες επιθέσεις.



Σχήμα 3.15: Συνάρτηση πυκνότητας πιθανότητας Βήτα(6,3)

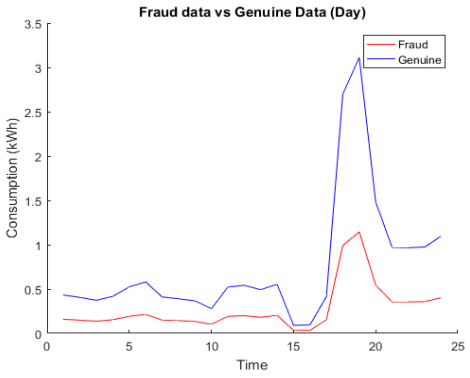
1. *Απώλειες Τύπου 1*: Μοντελοποιεί τον καταναλωτή που θα χρησιμοποιεί αδιάκοπα και μόνιμα το σύστημα αλλοίωσης μετρήσεων με την ίδια ένταση.

2. *Απώλειες Τύπου 2*: Μοντελοποιεί τον καταναλωτή που θα χρησιμοποιεί τυχαίες μέρες και για τυχαία διάρκεια μέσα στη μέρα σύστημα που αλλοιώνει τη μέτρηση με διαφορετική ένταση ανά ημέρα.

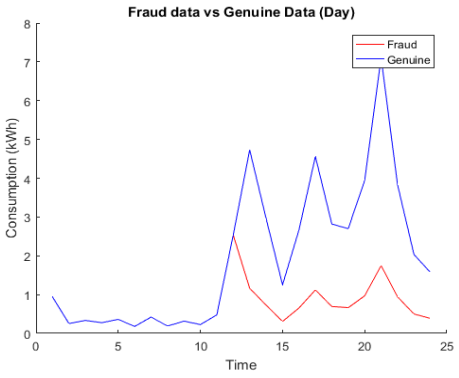
3. *Απώλειες Τύπου 3*: Μοντελοποιεί τον καταναλωτή που θα χρησιμοποιεί τυχαίες μέρες και για τυχαία διάρκεια μέσα στη μέρα σύστημα που αλλοιώνει τη μέτρηση με διαφορετική ένταση ανά ώρα για κάθε διάρκεια.

4. *Απώλειες Τύπου 4*: Μοντελοποιεί τον καταναλωτή που εκμεταλλεύεται την κυμαινόμενη χρέωση και αλλοιώνει τις τιμές του κατά τέτοιο τρόπο ώστε η μεγάλη κατανάλωση να μεταφέρεται τις ώρες μειωμένης χρέωσης.

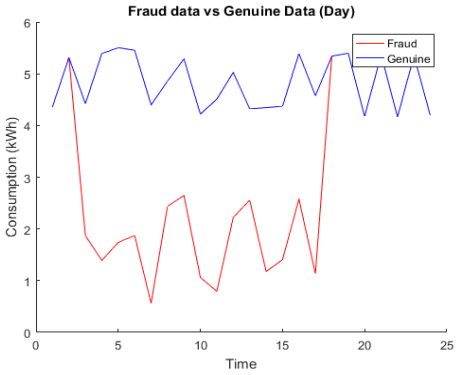
5. *Απώλειες Μικτών Τύπων*: Μοντελοποιεί το 70% με απώλειες τύπου 1, το 20% με απώλειες τύπου 2 και το 10% με απώλειες τύπου 3. Η παραπάνω ποσόστωση βασίζεται στο γεγονός πως ο ευκολότερος τύπος απώλειας συναντάται πολύ συχνότερα από τον πιο περίπλοκο.



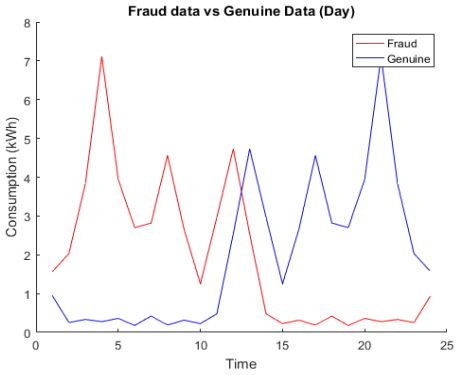
(α′) Απώλειες Τύπου 1



(β′) Απώλειες Τύπου 2



(γ′) Απώλειες Τύπου 3



(δ′) Απώλεια Τύπου 4

Σχήμα 3.16: Παραδείγματα απωλειών σε μια ημέρα

Κεφάλαιο 4

Αλγόριθμοι επιβλεπόμενης μάθησης

Στο παρόν κεφάλαιο γίνεται μια εξερεύνηση στους αλγορίθμους επιβλεπόμενης μάθησης. Αυτό επιτεύχθηκε με τη χρήση γραμμικών και μη γραμμικών ταξινομητών, διερευνώντας διαφορετικά δεδομένα εισόδου για κάθε περίπτωση. Η βιβλιοθήκη που χρησιμοποιήθηκε για τη γραμμική ταξινόμηση ονομάζεται LIBLINEAR και χαρακτηρίζεται από εξαιρετικές επιδόσεις σε προβλήματα με μεγάλα σετ δεδομένων. Αντίστοιχα, για τη μη γραμμική ταξινόμηση χρησιμοποιήθηκε η βιβλιοθήκη LIBSVM, η οποία αναγάγει τα δεδομένα εισόδου σε μεγαλύτερο χώρο διαστάσεων.

4.1 Θεωρία γραμμικής ταξινόμησης

Η βιβλιοθήκη LIBLINEAR υποστηρίζει δύο δημοφιλείς δυαδικά γραμμικούς ταξινομητές: τη λογιστική παλινδρόμηση (Logistic Regression) και τη γραμμική μηχανή υποστήριξης διανυσμάτων (linear SVM). Δεδομένου ενός σετ εκπαίδευσης $(\mathbf{x}_i, y_i)$, $i = 1, ..., l$, όπου $\mathbf{x}_i \in \mathbb{R}^n$ είναι ένα χαρακτηριστικό διάνυσμα και $y_i = \pm 1$ είναι οι ετικέτες, ένας γραμμικός ταξινομητής βρίσκει ένα διάνυσμα βαρών $\mathbf{w} \in \mathbb{R}^n$ επιλύοντας το ακόλουθο πρόβλημα:

$$min_{\mathbf{w}} f(\mathbf{w}) \equiv \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{i=1}^{l} \xi(y_i \mathbf{w}^T x_i)$$

όπου $\mathbf{w}^T\mathbf{w}/2$ είναι ο όρος ομαλοποίησης, $\xi(y_i\mathbf{w}^T x_i)$ είναι η συνάρτηση κόστους (loss function) και $C > 0$ είναι η παράμετρος ομαλοποίησης. Θεωρούμε τις συναρτήσεις κόστους στη λογιστική παλινδρόμηση (LR), στο L1-SVM, στο L2-SVM:

$$\xi_{LR}(y\mathbf{w}^T\mathbf{x}) = log(1 + exp(-y\mathbf{w}^T\mathbf{x}))$$
$$\xi_{L1}(y\mathbf{w}^T\mathbf{x}) = (max(0, 1 - y\mathbf{w}^T\mathbf{x}))$$
$$\xi_{L2}(y\mathbf{w}^T\mathbf{x}) = (max(0, 1 - y\mathbf{w}^T\mathbf{x}))^2$$

Σε μερικές περιπτώσεις, η συνάρτηση διακρίσεως του ταξινομητή περιλαμβάνει και έναν παράγοντα βάρους b. Η LIBLINEAR χειρίζεται αυτόν τον παράγοντα αυξάνοντας το διάνυσμα $\mathbf{w}$ και κάθε παράδειγμα $\mathbf{x}_i$ με μία επιπλέον διάσταση: $\mathbf{w}^T \leftarrow [\mathbf{w}^T, b]$, $\mathbf{x}_i^T \leftarrow [\mathbf{x}_i^T, B]$, όπου B

είναι μια σταθερά που ορίζεται από τον χρήστη. Η προσέγγιση για το L1-SVM και το L2-SVM είναι μέσω της μεθόδου coordinate descent. Για το LR και το L2-SVM, η LIBLINEAR υλοποιεί μια μέθοδο περιοχής εμπιστοσύνης Newton. Στη φάση των δοκιμών, εκτιμάται ένα μέλος των δεδομένων $\mathbf{x}$ σαν θετικό εάν $\mathbf{w}^T\mathbf{x} > 0$, και αρνητικό σε αντίθετη περίπτωση [25], [5].

Η μηχανή διανυσμάτων υποστήριξης εντάσσεται στο γενικότερο πλαίσιο της βελτιστοποίησης κυρτών συναρτήσεων και η προσέγγισή της έχει νόημα για όλους τους γραμμικούς ταξινομητές. Σε αδρές γραμμές, η διαδικασία εξελίσσεται σε τέσσερα κύρια βήματα:

1. Το πρόβλημα της εύρεσης του βελτίστου υπερεπιπέδου ξεκινά με μια δήλωση του προβλήματος στον πρωτεύοντα χώρο βαρών, ως ένα πρόβλημα βελτιστοποίησης με περιορισμούς.

2. Κατασκευάζεται η συνάρτηση Lagrange του προβλήματος.

3. Διατυπώνονται οι συνθήκες για τη βελτιστοποίηση της μηχανής.

4. Στήνεται το σκηνικό για την επίλυση του προβλήματος βελτιστοποίησης στο δυικό χώρο των πολλαπλασιαστών Lagrange.

Όπως προαναφέρθηκε, το πρωτεύον πρόβλημα ασχολείται με μια κυρτή συνάρτηση κόστους και γραμμικούς περιορισμούς. Δοθέντος ενός τέτοιου προβλήματος βελτιστοποίησης με περιορισμούς, είναι δυνατό να κατασκευάσουμε ένα άλλο πρόβλημα, το αποκαλούμενο δυικό του πρωτεύοντος. Αυτό το δεύτερο πρόβλημα έχει την ίδια βέλτιστη τιμή με το πρωτεύον, αλλά με τους πολλαπλασιαστές Lagrange να παρέχουν τη βέλτιστη λύση [31].

4.2 Εξερεύνηση γραμμικών ταξινομητών

Αρχικά έγινε μια εξερεύνηση των μεθόδων που παρέχει η LIBLINEAR για την επίλυση του προβλήματος. Λαμβάνοντας υπόψη 2.000 καταναλώσεις πελατών με ωριαίες μετρήσεις, επιλέχθηκε 10% ποσοστό ρευματοκλοπών για την προσομοίωση. Η βιβλιοθήκη που χρησιμοποιήθηκε περιλαμβάνει 7 διαφορετικούς συνδυασμούς ταξινομητών και συναρτήσεων κόστους, για να μπορούν να αντιμετωπιστούν όσο το δυνατόν περισσότερα προβλήματα. Παρ΄ όλα, αυτά οι μέθοδοι L1 είναι παλαιότερες εκδόσεις των L2 και αναμένεται να έχουν χειρότερα αποτελέσματα στις δοκιμές. Για την σφαιρική αντιμετώπιση του προβλήματος χρησιμοποιήθηκαν όλοι οι ταξινομητές που παρέχονται από τη βιβλιοθήκη σε κάθε τύπο απάτης. Παρακάτω παρατίθενται οι συνδυασμοί ταξινομητών και συναρτήσεων κόστους που δοκιμάστηκαν και τα αποτελέσματα σε κάθε τύπο απάτης.

1. L2 ομαλοποιημένη λογιστική παλινδρόμηση (πρωτεύον)

2. L2 ομαλοποιημένος ταξινομητής με L2 συνάρτηση κόστους διανυσμάτων υποστήριξης (δυικό)

3. L2 ομαλοποιημένος ταξινομητής με L2 συνάρτηση κόστους διανυσμάτων υποστήριξης (πρωτεύον)
4. L2 ομαλοποιημένη ταξινομητής με L1 συνάρτηση κόστους διανυσμάτων υποστήριξης (δυικό)
5. Ταξινόμηση διανυσμάτων υποστήριξης από Crammer και Singer
6. L1 ομαλοποιημένος ταξινομητής με L2 συνάρτηση κόστους διανυσμάτων υποστήριξης
7. L1 ομαλοποιημένη λογιστική παλινδρόμηση
8. L2 ομαλοποιημένη λογιστική παλινδρόμηση (δυικό)

Αρχικά έγινε μια δοκιμή σε 2.000 καταναλωτές με το 10% τους να έχει εισροή μη τεχνικών απωλειών. Τα διανύσματα των καταναλωτών είχαν 8.760 χαρακτηριστικά που αντιστοιχούν στις ώρες ενός έτους. Για να ευρεθεί η συνολική απόδοση όλων των γραμμικών ταξινομητών, έγιναν δοκιμές και στους τέσσερις τύπους απάτης (1, 2, 3, μικτός) και για αυτό και τα αποτελέσματα αναμένονται σχετικά χαμηλά. Ειδικότερα θα εξαχθεί ο μέσος όρος του F1 score και του Accuracy από τη δοκιμή κάθε αλγορίθμου και στους τέσσερις τύπους απάτης.

Χρησιμοποιώντας 70% του δείγματος για τις εκπαιδεύσεις κάθε ταξινομητή και 30% για τις προβλέψεις, πραγματοποιήθηκαν τέσσερις δοκιμές σε κάθε ένα από τους οκτώ συνολικά αλγορίθμους. Τα αποτελέσματα φαίνονται παρακάτω στον Πίνακα 4.1.

Αλγόριθμος	1	2	3	4	5	6	7	8
F1 score	23.92	31.99	30.19	28.67	32.66	29.28	20.43	24.04
Accuracy	91.36	90.41	90.46	90.56	90.15	90.37	91.43	91.35
Μέσος όρος	57.64	61.2	60.33	59.61	61.4	59,83	55.93	57.7

Πίνακας 4.1: Μέσος όρος Accuracy των δοκιμών

Εύκολα παρατηρείται από τον Πίνακα 4.1 πως η επίδοση των αλγορίθμων στο F1 score είναι περιορισμένη, υποδηλώνοντας δυσκολία στην ταξινόμηση. Αυτό οφείλεται στις κακές επιδόσεις των αλγορίθμων στις απάτες τύπου δύο, τρία και μικτού. Από την άλλη πλευρά τα αποτελέσματα του Accuracy είναι υποσχόμενα, αλλά πρέπει να ληφθεί υπόψη πως λόγω του χαμηλού ποσοστού κλοπών ένας κακός αλγόριθμος θα μπορούσε να προβλέπει πάντα αρνητικά και να είχε 90% Accuracy. Καθίσταται, λοιπόν, σαφές πως οι ταξινομητές έχουν μεγάλη δυσκολία να διαχωρίσουν τις προσομοιώσεις μη τεχικών απωλειών με μεγάλο τυχαίο παράγοντα από τις φυσιολογικές καταναλώσεις. Παρ΄ όλα αυτά, τα αποτελέσματα των δοκιμών στις απάτες τύπου 1 είναι ικανοποιητικά και δημιουργούν ανάγκη περαιτέρω ανάλυσης.

Η διαφορά της απόδοσης των γραμμικών ταξινομητών σε κάθε είδος απάτης και ειδικότερα σε σχέση με της κλοπές τύπου 1 υπήρξε ο λόγος εκκίνησης νέου κύκλου δοκιμών. Έγινε λοιπόν δοκιμή κάθε αλγορίθμου σε απάτες τύπου 1 με 10% ποσοστό ρευματοκλοπών. Το 70% των δεδομένων χρησιμοποιήθηκε για τις εκπαιδεύσεις των ταξινομητών και το υπόλοιπο 30% για τις προβλέψεις των αλγορίθμων.

Αλγόριθμος	DR	FPR	Accuracy	F1 score	BDR %
1	77.44	1.56	96.37	80.78	85
2	79.70	1.81	96.37	81.23	83
3	78.95	2.22	95.93	79.25	80
4	78.95	2.05	96.07	79.85	81
5	78.20	1.81	96.22	80.31	83
6	77.44	2.14	95.85	78.63	80
7	75.94	1.81	96.00	78.91	82
8	79.70	1.81	96.37	81.23	83

Πίνακας 4.2: Αποτελέσματα δοκιμής τύπου 1 χωρίς κανονικοποίηση

4.2.1 Παρατηρήσεις

Παρατηρώντας τους πίνακες αποτελεσμάτων εύκολα αποδεικνύεται η αρχική υπόθεση πως οι ταξινομητές και συναρτήσεις κόστους L2 έχουν καλύτερη συμπεριφορά ως προς την αντιμετώπιση του προβλήματος αναγνώρισης χρονοσειρών. Πιο συγκεκριμένα, για την τελική επιλογή του συνδυασμού μεθόδων επιλέχθηκαν δύο μετρικές για να καθορίσουν την επιλογή του καλύτερου πακέτου. Λήφθηκε υπόψη η μεταβολή της ευστοχίας (Accuracy) και παράχθηκε μέσος όρος για όλους τους τύπους. Παράλληλα, υπολογίστηκε μέσος όρος των δοκιμών με γνώμονα το καλύτερο F1 score, καθώς είναι μια αρκετά ζυγισμένη μετρική για τα προβλήματα ταξινόμησης. Βάσει, λοιπόν, του Πίνακα 4.1 την καλύτερη επίδοση έχει το πρωτεύον πρόβλημα που αποτελείται από L2 ομαλοποιημένο ταξινομητή με L2 συνάρτηση κόστους διανυσμάτων υποστήριξης, καθώς όπως μπορεί και να φανεί στον Πίνακα Α΄.6 του Παραρτήματο,ς η μηχανή διανυσμάτων υποστήριξης Crammer και Singer έχει καλύτερη επίδοση στους τύπους 2, 3 και στον μικτό. Αλλά, στην παρούσα φάση θα ασχοληθούμε με την απάτη τύπου 1.

4.3 Εξερεύνηση διαφορετικών τρόπων κανονικοποίησης

Το σκέλος της κανονικοποίησης των δεδομένων είναι ζωτικής σημασίας για κάθε σύστημα μηχανικής μάθησης. Η κανονικοποίηση των δεδομένων υλοποιείται, μειώνοντας το εύρος των τιμών σε οποιαδήποτε σχετικά μικρό εύρος. Συνηθέστερη πετυχημένη πρακτική είναι η αναγωγή των τιμών σε εύρος [0,1] ή [-1,1] με στόχο τη βελτίωση της επίδοσης και της ταχύτητας του αλγορίθμου.

Επιλέγοντας λοιπόν το σύνηθες δείγμα 2.000 καταναλωτών με ωριαίες μετρήσεις έτους και 10% ποσοστό καταναλωτών με μη τεχνικές απώλειες, οι αλγόριθμοι εκπαιδεύτηκαν με 70% του δείγματος και η πρόβλεψη επιτεύχθηκε με 30% για κάθε μέθοδο κανονικοποίησης.

Η βελτίωση της επίδοσης του αλγορίθμου επιτυγχάνεται σε μεγάλο βαθμό στη συγκεκριμένη περίπτωση από την κανονικοποίηση στο εύρος [0,1], βελτιώνοντας σε μικρό βαθμό τις μετρικές και μειώνοντας σχεδόν 10 φορές τον χρόνο εκτέλεσης της εκπαίδευσης. Στον Πίνακα 4.3 παρατίθενται τα αποτελέσματα των βέλτιστων ταξινομητών σε κάθε είδος κανονικοποίησης.

Κανονικοποίηση	DR	FPR	Accuracy	F1 score	BDR %	χρόνος εκπαίδευσης (s)
[0,1]	80.87	1.54	96.96	81.94	85	6.492741
[-1,1]	91.67	21.23	80.15	49.62	32	551.264250
-	79.70	1.81	96.37	81.23	83	58.246916

Πίνακας 4.3: Αποτελέσματα κανονικοποιήσεων

4.4 Εξερεύνηση χρονικής υποδιαίρεσης χρονοσειρών

Ολοκληρώνοντας την εξερεύνηση των ταξινομητών, απαιτείται να γίνει έλεγχος στις χρονικές υποδιαιρέσεις των χρονοσειρών. Για αυτό το σκοπό έγινε δοκιμή του πιο εύστοχου ταξινομητή σε 2.000 καταναλωτές με ποσοστό ρευματοκλοπών 10% και μόνο απάτες τύπου 1. Στη δοκιμή οι χρονοσειρές διαιρέθηκαν σε ημερήσιες, ωριαίες και ημίωρες μετρήσεις, λαμβάνοντας υπόψη όχι μόνο τις μετρικές ευστοχίας, αλλά και τον χρόνο εκτέλεσης της εκπαίδευσης κάθε ταξινομητή. Επιλέγοντας ως συνήθως 70% των δεδομένων για εκπαίδευση και το υπόλοιπο για πρόβλεψη, έγιναν δοκιμές για κάθε χρονική υποδιαίρεση.

Στον Πίνακα 4.4 εμφανίζεται, όπως αναμενόταν πως όσο αυξάνεται η συχνότητα των μετρήσεων τόσο πιο εύστοχος γίνεται ο ταξινομητής. Ωστόσο, ο χρόνος εκτέλεσης της εκπαίδευσης φαίνεται να επηρεάζεται έντονα από διαφορετικές χρονικές υποδιαιρέσεις με την ταξινόμηση με συχνότητα λήξης ανά ημέρα να είναι σημαντικά γρηγορότερη από τις υπόλοιπες, αλλά παρουσιάζεται σχετική δυσκολία στην αναγνώριση της απάτης.

Συχνότητα	DR	FPR	Accuracy	F1 score	BDR %	χρόνος εκπαίδευσης (s)
μέρες	81.62	2.55	95.85	79.86	78	0.069182
ώρες	82.88	2.16	96.22	82.59	81	4.152410
ημίωρα	81.08	1.66	96.44	83.33	84	12.169304

Πίνακας 4.4: Αποτελέσματα δοκιμής χρονικής υποδιαίρεσης

4.5 Θεωρία Μηχανών Διανυσμάτων Υποστήριξης

Για την ταξινόμηση με μηχανές διανυσμάτων υποστήριξης επιλέχθηκε η βιβλιοθήκη LIBSVM, η οποία προέρχεται από τους ίδιους δημιουργούς της LIBLINEAR. Σκοπός του SVM είναι η παραγωγή μοντέλων (βάσει των δεδομένων εκπαίδευσης), τα οποία προβλέπουν τα χαρακτηριστικά των δεδομένων δοκιμής βάσει μόνο των πληροφοριών που αντλούνται από τις τιμές των δεδομένων.

Ξεκινώντας από τα δεδομένα εκπαίδευσης έχουμε ζευγάρια παραδειγμάτων-δυαδικών χαρακτηριστικών $(\mathbf{x}_i, y_i)$,$i = 1, ..., l$ όπου $\mathbf{x}_i \in \mathbb{R}^n$ και $y \in \{1, -1\}^l$, ενώ οι μηχανές διανυσμάτων υποστήριξης (SVM) απαιτούν την λύση του παρακάτω προβλήματος βελτιστοποίησης:

$$min_{\mathbf{w},\mathbf{b},\xi} \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{i=1}^{l}\xi_i$$
$$\text{δεδομένου } y_i(\mathbf{w}^T\phi(\mathbf{x}_i) + b) \geq 1 - \xi_i,$$

$$\xi_i \geq 0$$

Εδώ τα διανύσματα εκπαίδευσης $\mathbf{x}_i$ ανάγονται σε μεγαλύτερο (ίσως άπειρο) χώρο διαστάσεων από τη συνάρτηση ϕ. Τα SVM βρίσκουν ένα γραμμικά διαχωρίσιμο υπερεπίπεδο με μέγιστο περιθώριο σε αυτό χώρο ανώτερων διαστάσεων. $C > 0$ είναι ο παράγοντας που θέτει ποινή στον παράγοντα λάθους (error term). Επιπροσθέτως, η σχέση $K(\mathbf{x}_i, \mathbf{x}_j) \equiv \phi(\mathbf{x}_i)^T\phi(\mathbf{x}_j)$ ονομάζεται συνάρτηση πυρήνα. Παρόλο που νέοι πυρήνες προτείνονται από ερευνητές, έχουν θεσπιστεί οι ακόλουθοι:

- Γραμμικός: $K(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{x}_i^T\mathbf{x}_j$.
- Πολυωνυμικός: $K(\mathbf{x}_i, \mathbf{x}_j) = (\gamma\mathbf{x}_i^T\mathbf{x}_j + r)^d$, $\gamma = \frac{1}{2\sigma^2} > 0$.
- RBF: $K(\mathbf{x}_i, \mathbf{x}_j) = exp(-\gamma\|\mathbf{x}_i - \mathbf{x}_j\|^2)$, $\gamma > 0$.
- Σιγμοειδής: $K(\mathbf{x}_i, \mathbf{x}_j) = tanh(\gamma\mathbf{x}_i^T\mathbf{x}_j + r)$.

Εδώ τα γ, r και d είναι παράμετροι των πυρήνων [22].

Χρησιμοποιώντας τη μέθοδο των πολλαπλασιαστών Lagrange μπορεί να διατυπωθεί το δυικό πρόβλημα για τα μη διαχωρίσιμα πρότυπα. Δοθέντος του δείγματος εκπαίδευσης $\{(\mathbf{x}_i, y_i)\}_{i=1}^N$, βρίσκονται οι πολλαπλασιαστές Lagrange $\{\alpha\}_{i=1}^N$ που μεγιστοποιούν την αντικειμενική συνάρτηση:

$$Q(\alpha) = \sum_{i=1}^N \alpha_i - \frac{1}{2}\sum_{i=1}^N\sum_{j=1}^N \alpha_i\alpha_j y_i y_j \mathbf{x}_i^T\mathbf{x}_j \text{ υπό τους περιορισμούς } \sum_{i=1}^N \alpha_i d_i = 0$$
$$0 \leq \alpha_i \leq C \text{ για } i = 1, 2, ..., N$$

όπου C είναι μια καθοριζόμενη από το χρήστη θετική παράμετρος [31].

4.5.1 Θεωρία επιλογής πυρήνα RBF

Γενικώς, ο πυρήνας δικτύου ακτινικής συνάρτησης βάσης (RBF) είναι μια λογική πρώτη επιλογή. Αυτός ο πυρήνας ανάγει μη γραμμικά τα δείγματα σε υψηλότερο χώρο διαστάσεων που μπορεί να διαχειριστεί την περίπτωση κατά την οποία η σχέση μεταξύ της τάξης και της τιμής είναι μη γραμμική. Επιπροσθέτως, ο γραμμικός πυρήνας είναι μια ειδική περίπτωση του RBF, καθώς ο γραμμικός πυρήνας με την παράμετρο ποινής $\check{C}$ έχει την ίδια επίδοση με τον RBF με δύο παραμέτρους (C,γ). Ακόμη, ο πυρήνας με σιγμοειδή συνάρτηση συμπεριφέρεται όπως με RBF για συγκεκριμένες παραμέτρους.

Ο δεύτερος λόγος είναι ο αριθμός των υπερπαραμέτρων, οι οποίες επηρεάζουν την πολυπλοκότητα της επιλογής μοντέλου. Ο πολυωνυμικός πυρήνας έχει περισσότερες υπερπαραμέτρους από τον RBF πυρήνα.

Τέλος, ο πυρήνας RBF έχει λιγότερες αριθμητικές δυσκολίες. Το χαρακτηριστικό κλειδί είναι πως το $0 < K_{ij} \leq 1$ είναι σταθερά του πολυωνυμικού πυρήνα του οποίου οι τιμές μπορούν να φτάνουν το άπειρο ($\gamma\mathbf{x}_i^T\mathbf{x}_j + r > 1$) ή το μηδέν ($\gamma\mathbf{x}_i^T\mathbf{x}_j + r < 1$) ενώ ο βαθμός είναι ήδη μεγάλος. Επίσης, πρέπει να σημειωθεί πως ο σιγμοειδής πυρήνας δεν είναι εφικτός με κάποιες παραμέτρους.

Υπάρχουν κάποιες περιπτώσεις όπου ο πυρήνας RBF δεν είναι κατάλληλος. Πιο συγκεκριμένα, όταν ο αριθμός των χαρακτηριστικών είναι πολύ μεγάλος, κάποιος θα μπορούσε να χρησιμοποιήσει τον γραμμικό πυρήνα [22].

4.6 Δοκιμή ταξινόμησης με Μηχανές Διανυσμάτων Υποστήριξης

Η προτεινόμενη διαδικασία που ακολουθείται από τους δημιουργούς του LIBSVM είναι η εξής:

- Μετατροπή των δεδομένων σε αναγνωρίσιμη μορφή με το πακέτο SVM.
- Κανονικοποίηση δεδομένων.
- Εξέταση του RBF πυρήνα.
- Χρήση cross-validation για την εύρεση των βέλτιστων παραμέτρων C και γ.
- Χρήση των βέλτιστων παραμέτρων C και γ για την εκπαίδευση των δεδομένων εκπαίδευσης.
- Δοκιμή.

Έχοντας τη διαδικασία αυτή υπόψη δοκιμάστηκαν επιτυχώς δύο διαφορετικά σενάρια ταξινόμησης. Στο πρώτο σενάριο ταξινομήθηκαν οι χρονοσειρές κάθε καταναλωτή βάσει της ετήσιας κατανάλωσης του και αναγνωρίζοντας κάθε τύπο κλοπής. Στο δεύτερο σενάριο χρησιμοποιήθηκε ο πυρήνας RBF και έγινε μια προσέγγιση στην αναγνώριση των ημερήσιων μη τεχνικών απωλειών ταξινομώντας σε πρώτη φάση τις ημέρες όλων των καταναλωτών και σε δεύτερη φάση κάθε καταναλωτή [22].

4.6.1 Δοκιμή χρονοσειρών χωρίς πυρήνα

Δεδομένης της ευστοχίας των γραμμικών ταξινομητών, θεωρήθηκε αναγκαία η δοκιμή του γραμμικού πυρήνα SVM. Παρ΄ όλα αυτά, η διαίσθηση δεν ήταν η μόνη κινητήριος δύναμη για την υλοποίηση αυτής της δοκιμής. Γενικότερα, αν ο αριθμός των μετρήσεων είναι μεγάλος, δεν απαιτείται να αναχθούν τα δεδομένα σε χώρο ανώτερων διαστάσεων [22]. Πρακτικά, αυτό σημαίνει πως η μη γραμμική αναγωγή δεν βελτιώνει την επίδοση του συστήματος. Ενώ είναι γενικώς αποδεκτό ότι ο πυρήνας RBF είναι τουλάχιστον καλύτερος από το γραμμικό, αυτή η δήλωση είναι αληθής, μόνο αφού έχουν επιλεχθεί οι παράμετροι (C,γ). Ένας γενικός κανόνας χρήσης του γραμμικού πυρήνα είναι η χρήση του όταν ο αριθμός των παραδειγμάτων (καταναλωτών) είναι μικρότερος ή σχετικός με τον αριθμό των χαρακτηριστικών (ωριαίες μετρήσεις έτους).

Αποτελέσματα δοκιμής

Η δοκιμή έγινε σε 4.500 καταναλωτές με 8.760 χαρακτηριστικά, ελέγχοντας αρχικά την επίδοση του συστήματος σε κάθε τύπο απάτης με ποσοστό ρευματοκλοπής 10%. Το 70% του δείγματος καταναλωτών χρησιμοποιήθηκε για την εκπαίδευση των ταξινομητών και το 30% για τις προβλέψεις τους. Οι αλγόριθμοι του LIBSVM αναμένεται να αντιμετωπίσουν δυσκολίες στους τύπους απάτης 2, 3 και μικτό όπως και οι υπόλοιποι γραμμικοί ταξινομητές.

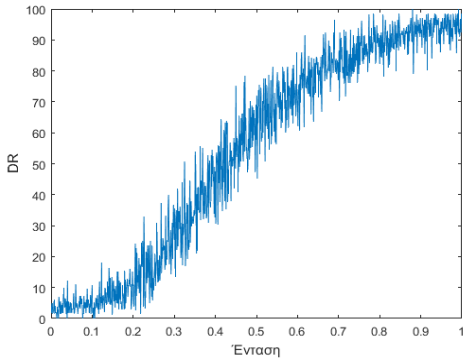
Στον Πίνακα 4.5 φαίνονται τα αποτελέσματα της δοκιμής. Γίνεται, λοιπόν, σαφές πως ο ταξινομητής μπορεί να αναγνωρίσει με αξιοπιστία μόνο τις απάτες τύπου 1, όπως και οι αντίστοιχοι ταξινομητές της LIBLINEAR. Ωστόσο, ακόμα και στα χαμηλότερα αποτελέσματα έχουμε ικανοποιητικό Accuracy, γεγονός που φανερώνει ότι ο ταξινομητής λειτουργεί όπως αναμενόταν.

Τύπος	DR	FPR	Accuracy	F1 score	BDR %	χρόνος εκτέλεσης (s)
1	81.43	1.24	96.96	84.76	88	10.188667
2	22.63	7.25	85.63	24.22	26	39.489221
3	23.78	10.36	82.67	22.52	20	39.648516
Μικτός	27.13	7.37	86.37	27.56	29	36.836504

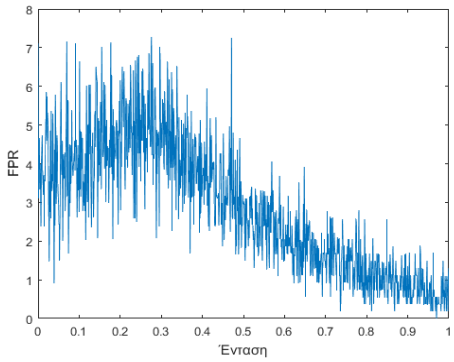
Πίνακας 4.5: Αποτελέσματα Γραμμικού SVM σε όλους τους τύπους απάτης

Για τη βαθύτερη κατανόηση της λειτουργίας του ταξινομητή απαιτείται η παρατήρηση της σχέσης των μετρικών με την ένταση κλοπής. Η ένταση κλοπής μαθηματικοποιείται σαν ένας παράγοντας που μπορεί να ποσοτικοποιήσει πόσο απέχουν τα αλλοιωμένα δεδομένα από τις πραγματικές μετρήσεις. Ουσιαστικά είναι ο συντελεστής υποδιαίρεσης των πραγματικών μετρήσεων.

Τα Σχήματα 4.1 δείχνουν πως ο ταξινομητής ξεκινά να βελτιώνεται αφότου η ένταση αυξηθεί πάνω από 30%, καθώς το DR αυξάνεται σχεδόν γραμμικά με την ένταση και το FPR μειώνεται σταθερά μετά από αυτό το σημείο. Εκεί που ο ταξινομητής έχει τη βέλτιστη απόδοση είναι στο εύρος [70%-90%], αφού η κλίση της καμπύλης σε αυτά τα σημεία είναι σημαντικά μικρότερη, γεγονός που υποδεικνύει σύγκλιση.



(α′) DR συναρτήσει της έντασης της κλοπής



(β′) FPR συναρτήσει της έντασης της κλοπής

Σχήμα 4.1: Επίπτωση της έντασης στα αποτελέσματα

4.6.2 Ημερήσια ταξινόμηση με πυρήνα RBF

Σε αυτή τη φάση δημιουργήθηκε η ανάγκη για εξαγωγή χαρακτηριστικών, ώστε να μειωθούν οι διαστάσεις των πινάκων και να επιταχυνθεί η διαδικασία. Παράλληλα, τα χαρακτηριστικά παρέχουν ένα επίπεδο αποπροσωποποίησης δημιουργώντας ένα αποτύπωμα της καταναλωτικής συνήθειας [15]. Μετρώντας τα αθροίσματα, τα ελάχιστα, τα μέγιστα και τους μέσους όρους των καθημερινών καταναλώσεων δημιουργείται ένας βασικός κορμός χαρακτηριστικών για κάθε καταναλωτή που μπορεί εύκολα να επεκταθεί και σε άλλα γραμμικά και μη εξαρτώμενα χαρακτηριστικά.

- *Μέγιστο και ώρα μεγίστου*
- *Ελάχιστο και ώρα ελαχίστου*
- *Άθροισμα κατανάλωσης ανά ημέρα*
- *Μέσος όρος, διακύμανση και τυπική απόκλιση ανά ημέρα*
- *Παράγοντας φορτίου, ελάχιστο προς μέση τιμή, ελάχιστο προς μέγιστο*
- *Επίδραση βραδινής κατανάλωσης*
- *Λοξότητα και Κύρτωση*

Η πρώτη δοκιμή του SVM έγινε με επιλογή 300 τυχαίων καταναλωτών μιας περιοχής με σκοπό να εκπαιδευτεί το σύστημα, ώστε να μπορεί να αναγνωρίζει ημέρες απάτης μέσα στο έτος. Η εκπαίδευση του ταξινομητή έγινε με τα ημερήσια χαρακτηριστικά για κάθε καταναλωτή. Τα δεδομένα διαχωρίζονται σε 2 τμήματα, το τμήμα της εκπαίδευσης που περιέχει 70% του δείγματος των καταναλωτών και το τμήμα της δοκιμής που περιέχει ένα ποσοστό της τάξης του 30% από το ίδιο δείγμα.

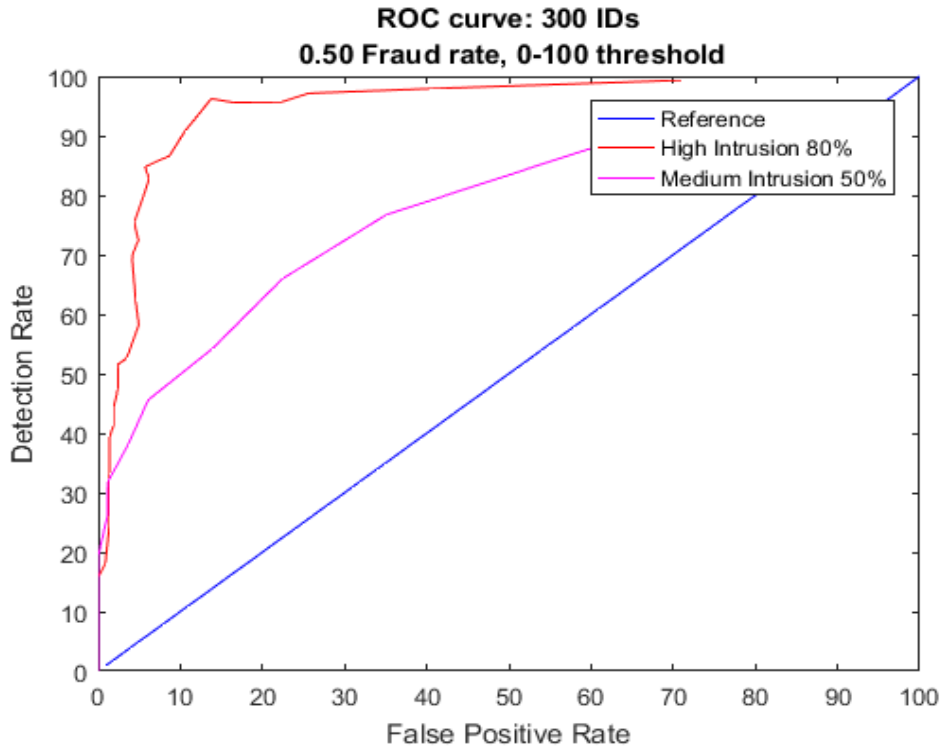
Ο ταξινομητής, λοιπόν, εκπαιδεύεται με ημερήσια χαρακτηριστικά κάθε καταναλωτή, αλλά θα πρέπει να αποφανθεί στο τέλος αν ο καταναλωτής έχει νοθεύσει τις μετρήσεις του ή όχι. Η

λύση δόθηκε εισάγοντας ένα όριο ημερών το οποίο, αν ο ταξινιμητής το προσπερνούσε, τότε ο καταναλωτής θεωρείται πως έχει αλλοιώσει τα δεδομένα του. Για να βρούμε την βέλτιστη τιμή αυτού του ορίου, χρησιμοποιήθηκαν ROC καμπύλες για να παρατηρηθεί η μεταβολή του DR και FPR, ενώ αλλάζει το όριο ημερών.

Αποτελέσματα δοκιμής

Ελέγχοντας τα αποτελέσματα του Πίνακα 4.6, παρατηρείται πως επιλέγοντας όριο στις 10 ημέρες επιτυγχάνεται ακρίβεια της τάξης του 95% στην εύρεση της απάτης, αλλά με σχετικά υψηλό ποσοστό λάθος συναγερμού της τάξης του 15% για τις έντονες απάτες. Αν χρειαστεί να ελαχιστοποιηθεί το FPR, θα πρέπει να επιλεχθεί μια μεγαλύτερη οριακή τιμή όπως το 14, που έχει ικανοποιητικό ποσοστό και στο DR που είναι της τάξης του 85% και στο FPR που είναι της τάξης του 8%. Οι απάτες που έγιναν με μικρότερη ένταση δεν γίνονται αντιληπτές από τον ταξινομητή που παράγει καμπύλη με παρόμοια κλίση με αυτή της ευθείας αναφοράς $y = x$.

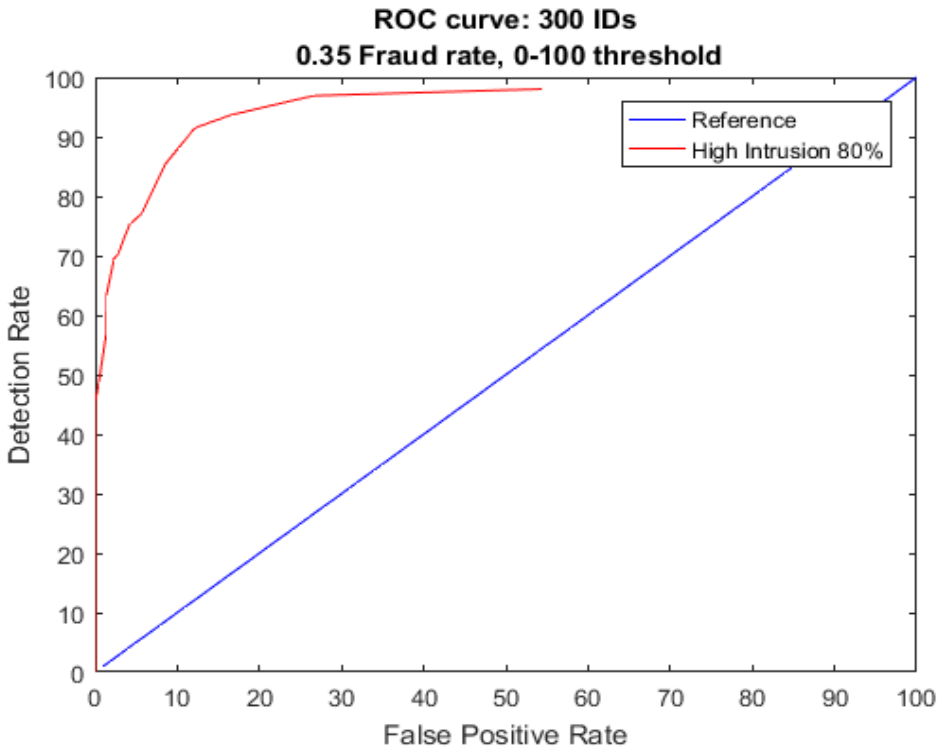
Αντίστοιχα, στον Πίνακα 4.7 φαίνεται πως η μείωση του ποσοστού απάτης (FR) επηρέασε το σύστημα, και ειδικότερα μείωσε το όριο στις 10 μέρες με DR=85% και FPR=9%. Ουσιαστικά φαίνεται πως το σύστημα χρειάζεται και άλλους καταναλωτές, ώστε να αποτυπωθούν και οι καμπύλες για χαμηλότερες εντάσεις διείσδυσης στα δεδομένα.



Σχήμα 4.2: Καμπύλη ROC για FR=0.50

300 IDs, 0.5 rate, 0-100 threshold				
Όριο (Μέρες)	DR (0.8)	FPR (0.8)	DR (0.5)	FPR (0.5)
2	97,917	40,385	76,712	35,0649
4	97,143	25,625	65,972	22,436
6	95,683	22,360	54,225	13,924
8	95,588	16,463	45,588	6,098
10	96,241	13,772	37,879	3,571
12	90,698	10,526	31,783	1,17
14	86,614	8,671	26,190	1,149
16	84,8	5,714	19,355	0
18	82,787	6,18	15,702	0
20	79,832	5,525	11,667	0

Πίνακας 4.6: Πίνακας επιλογής ορίου FR=0.5



Σχήμα 4.3: Καμπύλη ROC για FR=0.35

300 IDs, 0.35 rate, 0-100 threshold		
Όριο (Μέρες)	DR (0.8)	FPR (0.8)
2	95,192	30,612
4	91,176	23,232
6	89,691	17,734
8	85,567	12,808
10	85,106	8,738
12	84,444	5,238
14	79,545	2,830
16	75	2,830
18	68,235	2,791
20	63,529	2,791

Πίνακας 4.7: Πίνακας επιλογής ορίου FR=0.35

4.7 Σχόλια

Συνοψίζοντας, καθίσταται σαφές πως μπορεί να χρησιμοποιηθεί επιτυχώς επιβλεπόμενη μάθηση για τον εντοπισμό μη τεχνικών απωλειών. Οι γραμμικοί ταξινομητές μπορούν να αναγνωρίσουν αξιόπιστα και γρήγορα τον πρώτο τύπο απάτης, ενώ έχουν δυσκολία εντοπισμού στους υπόλοιπους τύπους. Παρ΄ όλα αυτά, χρησιμοποιώντας τον πυρήνα RBF, γίνεται εφικτή η αναγνώριση μη τεχνικών απωλειών αρχικά κάθε ημέρας και εν συνεχεία κάθε καταναλωτή. Γενικότερα, όμως, και οι δύο ομάδες ταξινομητών έχουν καλές επιδόσεις στον εντοπισμό ρευματοκλοπών με έντονη ένταση κλοπής, ενώ όσο μειώνεται η ένταση οι ταξινομητές δείχνουν μεγαλύτερη δυσκολία να διαχωρίσουν αλλοιωμένα από κανονικά δεδομένα.

Παράλληλα, γίνεται εμφανής η ανάγκη για σωστή επιλογή της δομής των δεδομένων εισόδου, καθώς κάθε ταξινομητής απαιτεί διαφορετική μεταχείριση. Οι γραμμικοί ταξινομητές απαιτούν πολλά χαρακτηριστικά (μετρήσεις), ενώ οι μη γραμμικοί μπορούν να λειτουργήσουν με πολύ λιγότερα. Αντίστοιχα, η κανονικοποίηση προσφέρει άμεση βελτίωση στα αποτελέσματα και επιταχύνει τη διαδικασία εκπαίδευσης σε μεγάλο βαθμό.

Κεφάλαιο 5

Συστήματα μη επιβλεπόμενης μάθησης

Ολοκληρώνοντας τον κύκλο των δοκιμών για τους αλγορίθμους επιβλεπόμενης μάθησης, δημιουργήθηκε η ανάγκη για περαιτέρω έρευνα σε διαφορετικούς αλγορίθμους. Οι αλγόριθμοι επιβλεπόμενης μάθησης έχουν ένα βασικό μειονέκτημα, όταν προσεγγίζεται ένα πραγματικό πρόβλημα. Αυτό είναι η δυσκολία εφαρμογής του αλγορίθμου, λόγω της έλλειψης των τάξεων των δεδομένων που απαιτεί ένα τέτοιο σύστημα για να εκπαιδευτεί. Η δυσκολία αυτή παρακάμπτεται, χρησιμοποιώντας μη επιβλεπόμενους ή ημι-επιβλεπόμενους αλγορίθμους που απαιτούν λίγα ή και κανένα ταξινομημένο παράδειγμα. Σε αυτό το κεφάλαιο θα προσεγγιστεί το πρόβλημα της ταξινόμησης καταναλωτών με νέα συστήματα που μπορούν να έχουν άμεσα χρήση στη λύση του πραγματικού προβλήματος, κάνοντας μια ανασκόπηση στις νέες δυσκολίες που προέκυψαν.

5.1 Εξαγωγή Χαρακτηριστικών

Στο παρόν μέρος θα γίνει παρουσίαση και ανάλυση των χαρακτηριστικών που χρησιμοποιήθηκαν στο μερικώς επιβλεπόμενο σύστημα, αλλά και στο μη επιβλεπόμενο σύστημα. Κάθε παράδειγμα μπορεί να περιγραφεί από ένα συνδυασμό τιμών που αναφέρονται ως μεταβλητές, χαρακτηριστικά, πεδία ή διαστάσεις. Οι τιμές αυτές μπορούν να είναι διαφορετικού τύπου όπως συνεχείς, δυαδικές ή κατηγορίες. Κάθε παράδειγμα μπορεί να αποτελείται μόνο από μια τιμή (μονοπαραγοντικό) ή και από περισσότερες (πολυπαραγοντικό). Στην περίπτωση των πολυπαραγοντικών παραδειγμάτων, όλες οι τιμές μπορεί να είναι ίδιου τύπου ή μπορεί να είναι ένας συνδυασμός διαφορετικών τύπων [30].
Παράλληλα, κάθε παράδειγμα μπορεί να οριστεί βάσει ακόμη δύο δομών ως προς τον ορισμό του προβλήματος [30].

1. *Τιμές Συσχετισμού*: τέτοιου είδους τιμές χρησιμοποιούνται για να περιγράψουν ένα γενικό πλαίσιο που χαρακτηρίζει ένα παράδειγμα. Στις χρονοσειρές, ο χρόνος είναι μια τιμή που παρέχει μια σχετικότητα, η οποία καθορίζει τη θέση ενός παραδείγματος σε

μια ολόκληρη ακολουθία. Μία τιμή γενικού πλαισίου είναι η μηνιαία κατανάλωση ενός κατοίκου.

2. *Συμπεριφορικές Τιμές*: είναι οι τιμές που δεν προδίδουν ένα γενικό πλαίσιο για κάποιο παράδειγμα ή κάποια σχετικότητα. Ένα τέτοιο παράδειγμα θα μπορούσε να είναι η ετήσια παραγωγή ενέργειας σε όλο τον κόσμο.

5.1.1 Φύση Χαρακτηριστικών

Το μερικώς επιβλεπόμενο και μη επιβλεπόμενο σύστημα απαιτούν εισόδους που να δίνουν τη δυνατότητα να διαχωρίζονται σε δύο κλάσεις οι καταναλωτές. Για να γίνει αυτό, απαιτείται η χρήση χαρακτηριστικών που αντιπροσωπεύουν την κλάση, αλλά και χαρακτηριστικών που προσδίδουν γενικότητα στο κάθε παράδειγμα. Με αυτό τον τρόπο, παρέχεται ένα περιθώριο στον αλγόριθμο, έτσι ώστε να μπορεί εύκολα να προσαρμόζεται σε καινούργια και ξεχωριστά παραδείγματα. Ένας απλοϊκός τρόπος να διαχωρίσουμε τα χαρακτηριστικά είναι σε γενικά και σε εξειδικευμένα χαρακτηριστικά για τον εντοπισμό κλοπής. Όλα τα παρακάτω χαρακτηριστικά αποτελούν τιμές συσχέτισης.

Γενικά χαρακτηριστικά

Τα πλεονέκτημα των γενικών χαρακτηριστικών είναι ότι βοηθούν στην κατάταξη του καταναλωτή σε σχέση με τους υπόλοιπους, ώστε να εξαχθούν πληροφορίες, όπως ο τύπος καταναλωτή (οικιακού ή βιομηχανικού) και το προφίλ κατανάλωσής του. Τέτοια χαρακτηριστικά, όμως, πρέπει να περιορίζονται σε αριθμό, καθώς ενδέχεται να δυσκολεύσουν τον διαχωρισμό με βάση το κριτήριο που θέτουμε, παρέχοντας μεγάλο παράγοντα γενίκευσης. Τέτοιου είδους χαρακτηριστικά είναι τα παρακάτω:

1. *Ετήσια μέση τιμή ημίωρου*: βρίσκεται ο μέσος όρος ημίωρου κάθε μέρας και ο ετήσιος μέσος όρος για όλες τις μέρες του έτους βρίσκεται.

2. *Ετήσια τυπική απόκλιση ημίωρου*: βρίσκεται η τυπική απόκλιση κάθε μέρας και για όλες τις μέρες του έτους ο ετήσιος μέσος όρος της τυπικής απόκλισης.

3. *Διαφορά Ετήσιου Ελάχιστου τάσης με όμοιους*: βάσει αυτού του χαρακηριστικού ορίζεται για όλους τους καταναλωτές το ελάχιστο της τάσης των χρονοσειρών τους και στη συνέχεια βρίσκεται η απόλυτη διαφορά σε ημέρες μεταξύ των συστάδων που δημιουργήθηκαν.

4. *Διαφορά μέσης τιμής με ομοίους*: με αυτό το χαρακτηριστικό βρίσκεται η διαφορά του ετήσιου μέσου όρου κάθε καταναλωτή με την ομάδα καταναλωτών που ανήκει.

5. *Διαφορά τυπικής απόκλισης με ομοίους*: με αυτό το χαρακτηριστικό βρίσκεται η διαφορά της ετήσιας τυπικής απόκλισης κάθε καταναλωτή με την ομάδα καταναλωτών που ανήκει.

Εξειδικευμένα χαρακτηριστικά

Τα εξειδικευμένα χαρακτηριστικά επικεντρώνονται στην όξυνση των διαφορών μεταξύ των καταναλωτών διαφορετικών κλάσεων. Λειτουργούν, λοιπόν, σαν οδηγοί για τον αλγόριθμο, ώστε να κάνουν πιο εμφανείς τις διαφορές των κλάσεων. Το πλεονέκτημά τους είναι ο παράγοντας εξειδίκευσης που παρέχουν στον αλγόριθμο, διευκολύνοντάς τον να αναγνωρίζει με διαφορετικούς τρόπους κάθε κλάση. Το μειονέκτημα είναι πως λόγω της εξειδικευμένης τους φύσης μπορεί να μην εφαρμόζονται απόλυτα από όλους τους καταναλωτές ή στη χειρότερη περίπτωση να περιγράφουν μια σπάνια συμπεριφορά που δεν ενδιαφερόμαστε να διαχωρίσουμε.

1. *Κινούμενος μέσος όρος μηνιαίου μέσου όρου*: πρόκειται για υπό συνθήκη χαρακτηριστικό, που αν παρατηρήσει κάποια σημαντική πτώση των καταναλώσεων, τότε ψάχνει για τη μέγιστη και την καταγράφει. Θεωρώντας ως min τον μήνα του ελαχίστου και c την κατανάλωση του αντίστοιχου i μήνα, ορίζεται η εξής φόρμουλα για αυτό το χαρακτηριστικό.

$$\bar{c_p} - \bar{c_a} = \frac{1}{k-1} \sum_{i=1}^{k} c_{m-i} - \frac{1}{w} \sum_{i=0}^{w} c_{m+i}$$

2. *Κινούμενος μέσος όρος μηνιαίας τυπικής απόκλισης*: πρόκειται για υπό συνθήκη χαρακτηριστικό, που αν παρατηρήσει κάποια σημαντική πτώση της τυπικής απόκλισης, τότε ψάχνει για την μέγιστη και την καταγράφει. Θεωρώντας ως min τον μήνα του ελαχίστου και std την τυπική απόκλιση της κατανάλωσης τον αντίστοιχο i μήνα, ορίζεται η εξής φόρμουλα για αυτό το χαρακτηριστικό.

$$\bar{std_p} - \bar{std_a} = \frac{1}{k-1} \sum_{i=1}^{k} std_{m-i} - \frac{1}{w} \sum_{i=0}^{w} std_{m+i}$$

3. *Συμμετρική διαφορά καταναλώσεων*: πρόκειται για υπό συνθήκη χαρακτηριστικό που παρατηρεί μια γενική συμπεριφορά όμοιων καταναλωτών ως προς τη χρονική στιγμή της ελάχιστης κατανάλωσης και ψάχνει για κάποια σημαντική πτώση της κατανάλωσης ανάμεσα σε δύο συμμετρικές χρονικές στιγμές με άξονα συμμετρίας την εκάστοτε χρονική στιγμή ελαχίστου. Θεωρώντας ως min την ημέρα του ελαχίστου και c την κατανάλωση της αντίστοιχης i ημέρας ορίζονται οι εξής φόρμουλες εισάγοντας σε αυτό το σημείο και την ευκλείδεια απόσταση.

$$\bar{c_p} - \bar{c_a} = \frac{1}{n} \sum_{i=1}^{n+1} c_{min-i} - \frac{1}{n} \sum_{i=0}^{n} c_{min+i}$$
$$||c_p|| - ||c_a|| = \sqrt{\sum_{i=1}^{n+1} (c_{min-i})^2} - \sqrt{\sum_{i=0}^{n} (c_{min+i})^2}$$

4. *Συμμετρική διαφορά τυπικής απόκλισης*: πρόκειται για υπό συνθήκη χαρακτηριστικό που παρατηρεί μια γενική συμπεριφορά όμοιων καταναλωτών ως προς τη χρονική στιγμή της ελάχιστης κατανάλωσης και ψάχνει για κάποια σημαντική πτώση της τυπικής απόκλισης ανάμεσα σε δύο συμμετρικές χρονικές στιγμές με άξονα συμμετρίας την εκάστοτε χρονική στιγμή ελαχίστου. Θεωρώντας ως min την ημέρα του ελαχίστου και std την τυπική απόκλιση της κατανάλωσης την αντίστοιχη i ημέρα, ορίζεται η εξής φόρμουλα για αυτό το χαρακτηριστικό.

$$\bar{std}_p - \bar{std}_a = \frac{1}{n}\sum_{i=1}^{n+1} std_{min-i} - \frac{1}{n}\sum_{i=0}^{n} std_{min+i}$$
$$||std_p|| - ||std_a|| = \sqrt{\sum_{i=1}^{n+1}(std_{min-i})^2} - \sqrt{\sum_{i=0}^{n}(std_{min+i})^2}$$

5. *Τμηματική διαφορά κατανάλωσης με όμοιους καταναλωτές*: πρόκειται για υπό συνθήκη χαρακτηριστικό που παρατηρεί μια γενική συμπεριφορά όμοιων καταναλωτών ως προς τη χρονική στιγμή της ελάχιστης κατανάλωσης και ψάχνει για κάποια σημαντική πτώση της κατανάλωσης ανάμεσα στον καταναλωτή και τους ομοίους του μετά την χρονική στιγμή της ελάχιστης κατανάλωσης. Πιο φορμαλιστικά, θεωρώντας τον όρο c_{cl} ως κατανάλωση μιας ομάδας και τον όρο c_{co} ως κατανάλωση ενός πελάτη, έχουμε την παρακάτω διαφορά μέσων όρων και νορμών των καταναλώσεων.

$$\bar{c_{cl}} - \bar{c_{co}} = \frac{1}{n}\sum_{i=1}^{n+1} c_{cl,min-i} - \frac{1}{n}\sum_{i=0}^{n} c_{co,min+i}$$
$$||c_{cl}|| - ||c_{co}|| = \sqrt{\sum_{i=1}^{n+1}(c_{cl,min-i})^2} - \sqrt{\sum_{i=0}^{n}(c_{co,min+i})^2}$$

6. *Τμηματική διαφορά τυπικής απόκλισης με όμοιους καταναλωτές*: πρόκειται για υπό συνθήκη χαρακτηριστικό που παρατηρεί μια γενική συμπεριφορά όμοιων καταναλωτών ως προς τη χρονική στιγμή της ελάχιστης κατανάλωσης και ψάχνει για κάποια σημαντική πτώση της τυπικής απόκλισης ανάμεσα στον καταναλωτή και τους όμοιούς του μετά την χρονική στιγμή της ελάχιστης κατανάλωσης. Πιο φορμαλιστικά θεωρώντας τον όρο std_{cl} ως την τυπική απόκλιση κατανάλωσης μιας ομάδας και τον όρο std_{co} την τυπική απόκλιση κατανάλωσης ενός πελάτη, έχουμε την παρακάτω διαφορά μέσων όρων και νορμών των τυπικών αποκλίσεων των καταναλώσεων.

$$\bar{std}_{cl} - \bar{std}_{co} = \frac{1}{n}\sum_{i=0}^{n} std_{cl,min+i} - \frac{1}{n}\sum_{i=0}^{n} std_{co,min+i}$$
$$||std_{cl}|| - ||std_{co}|| = \sqrt{\sum_{i=0}^{n}(std_{cl,min+i})^2} - \sqrt{\sum_{i=0}^{n}(std_{co,min+i})^2}$$

7. *Χρονική Διαφορά Ελαχίστου*: πρόκειται για υπό συνθήκη χαρακτηριστικό που εξερευνά το ελάχιστο χρονικό σημείο της τάσης της καμπύλης κάθε καταναλωτή. Με βάση την ομάδα που ανήκει κάθε καταναλωτής, υπολογίζεται η απόλυτη τιμή της χρονικής διαφοράς μεταξύ του ελαχίστου κάθε καταναλωτή με την ομάδα που ανήκει. Χρησιμοποιώντας ένα όριο για τη διαφορά αυτή, γίνεται αντιληπτή οποιαδήποτε έντονη διακύμανση του καταναλωτή με την ομάδα του και καταγράφεται σαν χαρακτηριστικό διαχωρισμού από την αναμενόμενη συμπεριφορά κατανάλωσης.

$$|t_{cl,min} - t_{co,min}|$$

5.1.2 Δοκιμή Χαρακτηριστικών με σταθερή απάτη

Αφού οριστούν τα χαρακτηριστικά που εκτιμάται ότι μπορούν να βοηθήσουν στον διαχωρισμό των κλάσεων, έπεται φυσικά η δοκιμή τους με έναν απλό τρόπο, έτσι ώστε να επιβεβαιωθεί ότι μπορούν να λειτουργήσουν όπως αναμένεται. Παράλληλα, η δοκιμή αυτή παρέχει μεγάλο όγκο πληροφοριών, αφού καθιστά εμφανή τα σημεία και τις προϋποθέσεις τα οποία τα χαρακτηριστικά έχουν μεγάλη ακρίβεια, αλλά και αυτά στα οποία εκεί που υστερούν.

Ο κώδικας της δοκιμής θεωρεί δεδομένη και σταθερή την ένταση κλοπής και την ημέρα που κάθε καταναλωτής ξεκινά να αλλοιώνει τις τιμές του. Ειδικότερα, επιλέχθηκαν 2.000 καταναλωτές με το ποσοστό καταναλωτών που αλλοιώνει τις μετρήσεις του να είναι 50%, η ένταση της κλοπής είναι της τάξης του 80% και η μέρα κλοπής ορίζεται η 182η. Δηλαδή, μετά από 6 μήνες κανονικής κατανάλωσης ο χρήστης εισάγει σύστημα αλλοίωσης της μέτρησής του. Δοκιμάζοντας ξεχωριστά τα χαρακτηριστικά διαχωρισμού, ελέγχουμε το όριο κάθε χαρακτηριστικού, έτσι ώστε να δώσει μεγαλύτερη ακρίβεια στις επιθέσεις δεδομένων υπό τις παραπάνω συνθήκες. Αν παρατηρηθούν τέτοια χαρακτηριστικά, ο καταναλωτής θεωρείται θετικός στην κλοπή. Αντίθετα, αν ο καταναλωτής δεν έχει την αναμενόμενη συμπεριφορά το χαρακτηριστικό δεν καταγράφει κάποια τιμή και ο καταναλωτής θεωρείται αρνητικός στην κλοπή. Αναλυτικότερα, για κάθε εξειδικευμένο χαρακτηριστικό λήφθηκαν τα παρακάτω αποτελέσματα:

1. *Κινούμενος μέσος όρος μηνιαίου μέσου όρου*
   Στην πρώτη δοκιμή δόθηκε έμφαση στη γενικότερη συμπεριφορά του χαρακτηριστικού ως προς το όριο που τίθεται κάθε φορά. Έτσι παρατηρείται εύκολα πως όταν το όριο είναι μεγάλο, ο διαχωρισμός παρουσιάζει χαμηλή ακρίβεια στον εντοπισμό με εξαιρετικά μικρό ποσοστό αστοχίας. Αντίθετα, αν το όριο χαμηλώσει αισθητά, χάνεται η έννοια του διαχωρισμού και ο αλγόριθμος θεωρεί θετικούς σε κλοπές σχεδόν όλους τους καταναλωτές.

Όριο	DR	FPR	BDR %	Accuracy	F1
0,8	44,8	1,4	97	71,7	61,29
0,7	98,7	1,9	98	98,4	98,4
0,6	99,3	3,6	97	97,85	97,88
0,5	99,8	7,5	93	96,15	96,15
0	99,9	91,5	52	54,2	68,57

Πίνακας 5.1: Δοκιμή 1ου χαρακτηριστικού

2. *Κινούμενος μέσος όρος μηνιαίας τυπικής απόκλισης*
   Αντίστοιχα, για παρόμοιες τιμές του ορίου με το προηγούμενο χαρακτηριστικό ο διαχωρισμός είναι εξαιρετικά εύστοχος και δεν αφήνει περιθώρια για αμφισβήτηση.

Όριο	DR	FPR	BDR %	Accuracy	F1
0,7	98,2	2,3	98	98,3	98,31
0,6	99,8	4,1	96	97,85	97,89
0,5	99,5	8,2	92	95,65	95,81

Πίνακας 5.2: Δοκιμή 2ου χαρακτηριστικού

3. *Συμμετρική διαφορά καταναλώσεων*
   Το συγκεκριμένο χαρακτηριστικό δεν δίνει αξιόπιστα αποτελέσματα, καθώς η συμμετρία που προκύπτει από τον χρησιμοποιούμενο τύπο απάτης κάνει το συγκεκριμένο χαρακτηριστικό να αποτυγχάνει σε αυτή τη δοκιμή. Παρ΄ όλα αυτά, το χαμηλό ποσοστό αστοχίας αφήνει δεύτερες σκέψεις, καθώς δεν επιβαρύνει αισθητά τα αποτελέσματα, αλλά βοηθά στη γενίκευση του τύπου κλοπής.

Όριο	DR	FPR	BDR %	Accuracy	F1
0,1	26,3	5,7	82	60,3	39,85

Πίνακας 5.3: Δοκιμή 3ου χαρακτηριστικού

Η δοκιμή συνεχίστηκε και με τις νόρμες των καταναλώσεων, παρουσιάζοντας ελάχιστη βελτίωση στο DR, ενώ αισθητά καλύτερα αποτελέσματα παρατηρούνται στο FPR που μειώθηκε ακόμη περισσότερο.

Όριο	DR	FPR	BDR %	Accuracy	F1
0,3	29	2,1	93	63,65	44,71

Πίνακας 5.4: Δοκιμή 3ου χαρακτηριστικού με νόρμες

4. *Συμμετρική διαφορά τυπικής απόκλισης*
   Αντίστοιχα συμπεράσματα ισχύουν και στη συμμετρική διαφορά τυπικής απόκλισης που οριακά ξεπερνά το 10% στο FPR. Η γενίκευση που προσφέρει ωστόσο το συγκεκριμένο χαρακτηριστικό είναι χρήσιμη, καθώς εν τέλει όλα τα χαρακτηριστικά θα ενώσουν τα δυνατά τους σημεία για να διαχωρίσουν απάτες με μεγαλύτερο τυχαίο παράγοντα.

Όριο	DR	FPR	BDR %	Accuracy	F1
0,1	38,9	10,2	79	64,35	52,18

Πίνακας 5.5: Δοκιμή 4ου χαρακτηριστικού

Σε αυτό το σημείο η μείωση του FPR είναι αρκετά σημαντικό ζήτημα που τελικώς επιτεύχθηκε με τις νόρμες που μπόρεσαν να μειώσουν το FPR, χωρίς να επηρεάσουν

αρνητικά το DR.

Όριο	DR	FPR	BDR %	Accuracy	F1
0,1	40,2	8,9	82	65,65	53,92

Πίνακας 5.6: Δοκιμή 4ου χαρακτηριστικού με νόρμες

5. *Τμηματική διαφορά κατανάλωσης με όμοιους καταναλωτές*
Σε αυτό το χαρακτηριστικό παρατηρείται σχετική αστοχία σε σχέση με τα πρώτα χαρακτηριστικά, υποδεικνύοντας ανάγκη για καλύτερη ρύθμιση του χαρακτηριστικού.

Όριο	DR	FPR	BDR %	Accuracy	F1
0,1	98,4	11,4	90	93,5	93,8
0,2	93,9	7	93	93,45	93,48
0,3	88,3	4,9	95	91,7	91,41

Πίνακας 5.7: Δοκιμή 5ου χαρακτηριστικού

Δεδομένης της διαφοράς των καταναλώσεων, με τη γενικευμένη κατανάλωση μιας ομάδας δημιουργείται η ανάγκη για κανονικοποίηση σε κάθε διάνυσμα κατανάλωσης. Με αυτό τον τρόπο επιτυγχάνονται πολύ καλύτερα αποτελέσματα.

Όριο	DR	FPR	BDR %	Accuracy	F1
0,3	98,9	5,4	95	96,75	96,82

Πίνακας 5.8: Δοκιμή 5ου χαρακτηριστικού με κανονικοποίηση

Όριο	DR	FPR	BDR %	Accuracy	F1
0,1	98,7	7	93	95,85	95,96
0,2	97,6	4,4	96	96,6	96,63

Πίνακας 5.9: Δοκιμή 5ου χαρακτηριστικού με κανονικοποίηση και νόρμες

6. *Τμηματική διαφορά τυπικής απόκλισης με όμοιους καταναλωτές*

Αντίστοιχη μεθοδολογία εφαρμόστηκε και σε αυτό το χαρακτηριστικό. Τα αποτελέσματα ήταν ικανοποιητικά, αλλά όχι αρκετά. Έτσι, χρησιμοποιήθηκε κανονικοποίηση, για μπορέσει να μειωθεί το FPR, ενώ αυξάνεται το DR.

Όριο	DR	FPR	BDR %	Accuracy	F1
0,1	99,4	16,1	86	91,65	92,25
0,2	95,7	8,3	92	93,7	93,82
0,3	89,8	6,2	94	91,8	91,63

Πίνακας 5.10: Δοκιμή 6ου χαρακτηριστικού

Όριο	DR	FPR	BDR %	Accuracy	F1
0,4	97	4,9	95	96,05	96,09
0,3	96,3	5,3	95	95,5	95,54

Πίνακας 5.11: Δοκιμή 6ου χαρακτηριστικού με κανονικοποίηση

Όριο	DR	FPR	BDR %	Accuracy	F1
0,2	98,5	4,5	96	97	97,04
0,1	99,2	5,3	95	96,95	97,02

Πίνακας 5.12: Δοκιμή 6ου χαρακτηριστικού με κανονικοποίηση και νόρμες

7. *Χρονική Διαφορά Ελαχίστου*

   Δοκιμάζοντας το μοναδικό χαρακτηριστικό που σχετίζεται με χρόνο και όχι με κατανάλωση, καθίσταται σαφές πως δεν δίνει περισσότερη διακριτική ικανότητα στις κλάσεις. Αντίθετα, παρέχει μεγάλη γενικότητα στον αλγόριθμο, δίνοντας μια ακόμη πληροφορία για τη συμπεριφορά κατανάλωσης, αλλά λόγω του αισθητά μεγάλου FPR αποτυγχάνει να διαχωρίσει.

Όριο	DR	FPR	BDR %	Accuracy	F1
0,1	85,4	94,7	47	45,35	60,98
0,2	74,9	81,7	48	46,6	58,38
0,3	18,9	56,4	25	31,25	21,56
0,4	13,7	34	29	39,85	18,55

Πίνακας 5.13: Δοκιμή 7ου χαρακτηριστικού με κανονικοποίηση

5.1.3 Δοκιμή Χαρακτηριστικών με μεταβλητή απάτη

Τέλος, έγινε μια τελική δοκιμή στα χαρακτηριστικά, αυτή τη φορά με μεγαλύτερο τυχαίο παράγοντα. Η ένταση της κλοπής καθορίζεται από μια Βήτα κατανομή με παραμέτρους 6 και 3, ενώ η ημέρα που ξεκινά η κλοπή επιλέγεται από μια κανονική κατανομή με παραμέτρους 182,5 και 56,1538. Σε κάθε καταναλωτή που έχει επιλεχθεί για κλοπή επιβάλλονται διαφορετικές τιμές των παραπάνω κατανομών κρατώντας όμως το γενικότερο πλαίσιο του τύπου της κλοπής που είδαμε προηγουμένως. Αυτό που έχει ενδιαφέρον να παρατηρηθεί σε αυτό το σημείο είναι το χαμηλό FPR, καθώς αναμένεται να χαμηλώσει σημαντικά η ακρίβεια λόγω της απλότητας του κανόνα διαχωρισμού.

Χαρακτ.	Όριο	DR	FPR	BDR %	Accuracy	F1
1	0,7	42,8	2,1	95	70,35	59,08
2	0,7	46,5	1,8	96	72,35	62,71
3	0,1	58	9,9	85	74,05	69,09
4	0,1	59,6	9,4	86	75,1	70,53
5	0,3	66,4	8,2	89	79,1	76,06
6	0,4	58,4	5,6	91	76,4	71,22
7	0,3	48,8	39,8	55	54,5	51,75

Πίνακας 5.14: Δοκιμή χαρακτηριστικών με τυχαίο παράγοντα

Χαρακτ.	Όριο	DR	FPR	BDR %	Accuracy	F1
3	0,3	47,9	2,9	95	72,65	63,65
4	0,1	64,7	10,5	86	77,1	73,86
5	0,1	77,5	8,8	90	84,35	83,2
6	0,1	75,7	7,9	91	83,9	82,46

Πίνακας 5.15: Δοκιμή χαρακτηριστικών με τυχαίο παράγοντα και νόρμες

Καθίσταται, λοιπόν, σαφές πως τα χαρακτηριστικά σε γενικές γραμμές έχουν χαμηλότερη ακρίβεια, αλλά κρατούν χαμηλό FPR κάτι που μας ενδιαφέρει περισσότερο σε αυτό το σημείο. Παράλληλα, τα χαρακτηριστικά 3 και 4 που είχαν απογοητευτικά αποτελέσματα στην προηγούμενη δοκιμή, σε αυτήν δείχνουν να βελτιώνονται αισθητά σε σχέση με την επίδοση των υπολοίπων. Αυτό μας πληροφορεί ότι η αρχική υπόθεσή μας για το αίτιο αστοχία τους ήταν αληθής. Παράλληλα, το χαρακτηριστικό 7 που είχε επίσης εξαιρετικά αποθαρρυντικά αποτελέσματα στην προηγούμενη δοκιμή, εξισορροπεί σε αυτή τη δοκιμή τη σχέση μεταξύ του DR και FPR, παρόλο που ακόμη φαίνεται ως το χαρακτηριστικό με τα χειρότερα αποτελέσματα.

5.2 Αλγόριθμοι συσταδοποίησης

Η συσταδοποίηση είναι από τους δημοφιλέστερους τύπου μη επιβλεπόμενης εκμάθησης. Σε αυτόν τον τύπο εκμάθησης, ο στόχος δεν είναι η ταξινόμηση των δεδομένων, αλλά απλά η εύρεση των ομοιοτήτων μεταξύ τους. Η υπόθεση είναι συνήθως πως οι συστάδες που ανακαλύπτονται θα ταιριάξουν σχετικά καλά με τη διαισθητική ταξινόμηση [12]. Ειδικότερα ένα σύνολο παρατηρήσεων (σημείων δεδομένων) διαμερίζεται σε φυσικές ομαδοποιήσεις, ή συστάδες (clusters), με τέτοιο τρόπο ώστε το μέτρο ομοιότητας μεταξύ οποιουδήποτε ζεύγους παρατηρήσεων που αντιστοιχίζεται σε κάθε συστάδα να ελαχιστοποιεί μια καθορισμένη συνάρτηση κόστους.

5.2.1 K-Means

Δοθέντος ενός συνόλου N παρατηρήσεων, ζητείται ο κωδικοποιητής C που αντιστοιχίζει αυτές τις παρατηρήσεις στις K συστάδες με τέτοιο τρόπο, ώστε μέσα σε μια συστάδα ο μέσος όρος του μέτρου ανομοιότητας των αντιστοιχισμένων παρατηρήσεων ως προς το κέντρο (μέσο) της συστάδας να ελαχιστοποιείται μέσω της συνάρτησης κόστους.

$$J(C) = \sum_{j=1}^{K} \sum_{C(i)=j} \|\mathbf{x}_i - \bar{\mu_j}\|^2$$

Με μαθηματικούς όρους, ο αλγόριθμος (K-Means) εξελίσσεται σε δύο βήματα:

1. Για ένα δεδομένο κωδικοποιητή C, η συνολική διακύμανση συστάδας ελαχιστοποιείται ως προς το σύνολο μέσων συστάδας $\{\bar{\mu_j}\}_{j=i}^{K}$, δηλαδή εκτελούμε την ακόλουθη ελαχιστοποίηση:

$$min_{\{\bar{\mu_j}\}_{j=i}^{K}} \sum_{j=i}^{K} \sum_{C(i)=j} \|\mathbf{x}_i - \bar{\mu_j}\|^2 \text{ για δεδομένο } C$$

2. Αφού υπολογιστούν οι βελτιστοποιημένοι μέσοι συστάδας $\{\bar{\mu_j}\}_{j=i}^{K}$, στη συνέχεια βελτιστοποιούμε τον κωδικοποιητή ως εξής:

$$C(i) = argmin_{1 \leq j \leq K} \|\mathbf{x}_i - \bar{\mu_j}\|^2$$

Ξεκινώντας από κάποια αρχική επιλογή κωδικοποιητή C, ο αλγόριθμος εναλλάσσεται μεταξύ αυτών των δύο βημάτων μέχρι να μην υπάρξει περαιτέρω αλλαγή στις αντιστοιχίσεις των συστάδων[31].

5.2.2 Fuzzy C-Means

Ο αλγόριθμος ασαφών C μέσων (FCM) είναι πολύ κοντά στη λογική του K-Means, αλλά εισάγει μια πιο πιθανοτική προσέγγιση. Επιλύει το πρόβλημα της ελαχιστοποίησης των αποστάσεων μέσα σε μια συστάδα και της μεγιστοποίησης των αποστάσεων μεταξύ των συστάδων με βάση το παρακάτω κριτήριο βελτιστοποίησης[33]:

$$J_m = \sum_{k=1}^{N} \sum_{i=1}^{n} (u_{ik})^m \|\mathbf{x}_k - v_i\|^2$$

Έστω ένα σύνολο διανυσμάτων δεδομένων $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_3\}$ με $x_k \in \mathbb{R}^p$ $(1 \leq k \leq n)$, τα οποία ομαδοποιούνται σε ασαφείς συστάδες.

1. Επιλογή αριθμού c των ασαφών συστάδων, της παραμέτρου m, των αρχικών τιμών για τα διανύσματα $\mathbf{v}_1, \mathbf{v}_2, ..., \mathbf{v}_c$ και της παραμέτρου c.

$$\text{όπου } \mathbf{v}_i = \frac{\sum_{i=1}^{n} (u_{ik})^m x_{ik}}{\sum_{i=1}^{n} (u_{ik})}$$

2. Χρήση της παρακάτω εξίσωσης για τον υπολογισμό των συναρτήσεων συμμετοχής u_{ik}

$$u_{ik} = \frac{1}{\sum_{j=1}^{c} (\frac{\|\mathbf{x}_k - \mathbf{v}_i\|}{\|\mathbf{x}_k - \mathbf{v}_j\|})}, \ (1 \leq k \leq n), \ (1 \leq i \leq c)$$

3. Βάσει της εξίσωσης του βήματος 1, γίνεται προσδιορισμός των νέων τιμών για τα κέντρα των ασαφών υποομάδων $\mathbf{v}_1^{new}, \mathbf{v}_2^{new}, ..., \mathbf{v}_c^{new}$

4. Αν $max_i\{\|\mathbf{v}_i - \mathbf{v}_i^{new}\|^2\} < \epsilon$ τότε ο αλγόριθμος σταματάει, αλλιώς θέτει $\mathbf{v}_i = \mathbf{v}_i^{new}$ και η ροή του πηγαίνει στο βήμα 2.

5.2.3 SOM

Εμπνευσμένοι από τα νευρωνικά δίκτυα, οι αυτο-οργανωμένοι χάρτες (SOM) χρησιμοποιούν ένα μηχανισμό ανταγωνισμού και συνεργασίας για να πετύχουν μη επιβλεπόμενη εκμάθηση. Στην κλασσική περίπτωση του SOM, ένα μέρος από κόμβους οργανώνεται σε γεωμετρικό σχήμα, συνήθως δισδιάστατο πλέγμα. Κάθε κόμβος σχετίζεται με ένα διάνυσμα βάρους με τις ίδιες διαστάσεις όπως η είσοδος. Ο σκοπός του SOM είναι να βρει μια χαρτογράφηση από τον υψηλό χώρο διαστάσεων της εισόδου σε δισδιάσταστη αναπαράσταση των κόμβων. Ένας τρόπος να χρησιμοποιηθεί για συσταδοποίηση είναι να αναμένεται τα αντικείμενα στον χώρο εισόδου να αναπαρασταθούν από τον ίδιο κόμβο, όπως σχηματίστηκαν στη συστάδα. Στη διάρκεια της εκπαίδευσης, κάθε αντικείμενο στην είσοδο αναπαριστάται στον χάρτη και αναγνωρίζεται ο κόμβος που ταιριάζει βέλτιστα. Τυπικά, όταν η είσοδος και τα διανύσματα βαρών κανονικοποιηθούν, για δείγμα εισόδου $x(t)$ ο νικητής δείκτης c ορίζεται κάτω από τη συνθήκη:

$$\text{για όλα } i, \ \|x(t) - m_c(t)\| \leq \|x(t) - m_i(t)\|$$

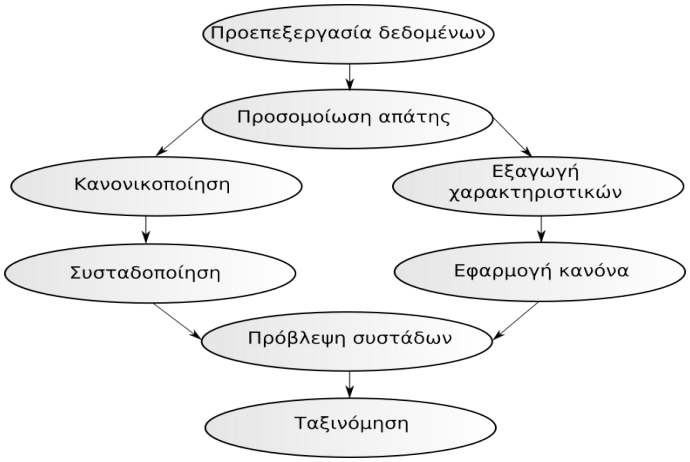
όπου t είναι το χρονικό βήμα στην εκπαίδευση, m_i είναι το διάνυσμα βάρους του i κόμβου. Μετά από αυτό, το διάνυσμα βάρους γύρω από τον βέλτιστο κόμβο $c = c(x)$ ανανεώνεται ως εξής:

$$m_i(t+1) = m_i(t) + \alpha h_{c(x),i}(x(t) - m_i(t))$$

όπου α είναι ο ρυθμός εκμάθησης και $h_{c(x),i}$ είναι η «γειτονική συνάρτηση», μια φθίνουσα συνάρτηση μεταξύ i και c κόμβων στο δίκτυο του χάρτη [1].

5.3 Συστατικά συστήματος μη επιβλεπόμενης μάθησης

Οι δοκιμές στην επιβλεπόμενη μάθηση έδειξαν πως η ταξινόμηση των συγκεκριμένων χρονοσειρών δεν είναι εύκολη διαδικασία. Για αυτό το λόγο, χρησιμοποιήθηκε συστοιχία μη επιβλεπόμενων αλγορίθμων για την ταξινόμηση των καταναλωτών. Ειδικότερα, εισήχθη ένα σύστημα με ταξινόμηση βάσει κανόνων, το οποίο συσταδοποιεί τα δεδομένα και εξάγει χαρακτηριστικά των χρονοσειρών και αποτελέσματα, λαμβάνοντας υπόψη τα παραπάνω. Η βέλτιστη δομή του συστήματος επιτεύχθηκε, όπως φαίνεται στο Σχήμα 5.1.



Σχήμα 5.1: Δομή μη επιβλεπόμενου ταξινομητή

Παρέχοντας περαιτέρω πληροφορίες για τα μέλη που απαρτίζουν το σύστημα προς έρευνα, έχουμε:

- *Προεπεξεργασία δεδομένων*: επιλέγονται και οργανώνονται τα δεδομένα σε συγκεκριμένους πίνακες και διανύσματα.
- *Προσομοίωση απάτης*: αλλοιώνονται οι μετρήσεις κάποιων καταναλωτών και ενημερώνονται οι προϋπάρχοντες πίνακες και διανύσματα.
- *Κανονικοποίηση*: κανονικοποιούνται οι ετήσιες χρονοσειρές κάθε καταναλωτή σε εύρος τιμών [-1,1].
- *Συσταδοποίηση*: συσταδοποιούνται οι καταναλωτές με βάση τις κανονικοποιημένες τιμές σε δύο συστάδες. Η μια συστάδα ομαλή και η άλλη ανώμαλη.
- *Εξαγωγή χαρακτηριστικών*: βάσει των χρονοσειρών δημιουργούνται ετήσια χαρακτηριστικά για κάθε καταναλωτή, προσπαθώντας να ανιχνευθεί ύποπτη συμπεριφορά.

- *Εφαρμογή κανόνα*: απενοχοποιούνται κάποιοι καταναλωτές που βρίσκονται στην ανώμαλη συστάδα λαμβάνοντας υπόψη το πλήθος των χαρακτηριστικών.
- *Πρόβλεψη συστάδων*: ορίζονται κλάσεις στις συστάδες με σεβασμό στον κανόνα.
- *Ταξινόμηση*: ταξινομούνται οι καταναλωτές και παράγονται τα τελικά αποτελέσματα και μετρικές.

5.3.1 Μεθοδολογία εξαγωγής αποτελεσμάτων

Η εξαγωγή αποτελεσμάτων διαδραματίζει μεγάλο ρόλο στην τελική απόδοση του αλγορίθμου, οπότε χρειάζεται ιδιαίτερη προσοχή στην τοποθέτηση δυαδικών χαρακτηριστικών. Η γενικότερη μεθοδολογία βασίζεται σε δύο σημαντικούς άξονες, καθώς η τομή των δύο είναι αυτή που εξάγει τα βέλτιστα αποτελέσματα. Αυτό γίνεται ξεκάθαρο παρατηρώντας τον πίνακα αποτελεσμάτων 5.16.

Ο πρώτος άξονας αποτελείται από την κανονικοποίηση και τη συσταδοποίηση. Κατά την διαδικασία της κανονικοποίησης, ο πίνακας με τις χρονοσειρές αναστρέφεται, κανονικοποιείται και αναστρέφεται για δεύτερη φορά για να αποκτήσει την ίδια μορφή με την αρχική, αλλά με εύρος τιμών [-1,1]. Με αυτό τον τρόπο, δίνεται έμφαση στη μορφή και όχι στα μεγέθη των χρονοσειρών. Έτσι, το σύστημα εκμεταλλεύεται το γεγονός ότι οι χρονοσειρές είναι αρκετά ομοιόμορφες ως προς το σχήμα. Σε επόμενη φάση εκτελείται συσταδοποίηση στα κανονικοποιημένα μεγέθη και διαχωρίζονται οι καταναλωτές σε μια μεγάλη συστάδα με αναμενόμενες μορφές και σε μια μικρή συστάδα με ακανόνιστες συμπεριφορές.

Ο δεύτερος άξονας αποτελείται από την εξαγωγή χαρακτηριστικών των χρονοσειρών και την εφαρμοφή του κανόνα. Η εξαγωγή δίνει τη δυνατότητα μέσω των χαρακτηριστικών διαχωρισμού να δημιουργηθούν ομάδες καταναλωτών που έχουν ύποπτες και αναμενόμενες μετρήσεις. Αν έχουμε έλλειψη χαρακτηριστικών, δηλαδή 0, πρακτικά σημαίνει πως ο καταναλωτής έχει αναμενόμενη συμπεριφορά. Στην αντίθετη περίπτωση ο καταναλωτής έχει αποκλίνουσα συμπεριφορά και θεωρείται ύποπτος. Εκεί εφαρμόζεται ο κανόνας που ορίζει πως αν ο καταναλωτής έχει λιγότερες από τρεις μετρήσεις στα χαρακτηριστικά διαχωρισμού, η συμπεριφορά του θεωρείται αναμενόμενη.

Για τη δοκιμή των κανόνων επιλέχθηκε δείγμα 2.000 καταναλωτών με ποσοστό διείσδυσης μη τεχνικών απωλειών στο 10%. Έγιναν συνολικά τρεις δοκιμές για τους κανόνες, με τον πρώτο να ελέγχει την απόδοση της συσταδοποίησης, τον δεύτερο να ελέγχει την απόδοση των χαρακτηριστικών και τον τρίτο να ελέγχει τον συνδυασμό των δύο κανόνων.

Κανόνας	DR	FPR	Accuracy	F1 score	BDR %
Συσταδ.	98.67	34.2	69.09	38.96	24
Χαρακτ.	87.78	6.72	92.73	70.73	59
Συνδ.	89.33	5.93	93.6	73.63	63

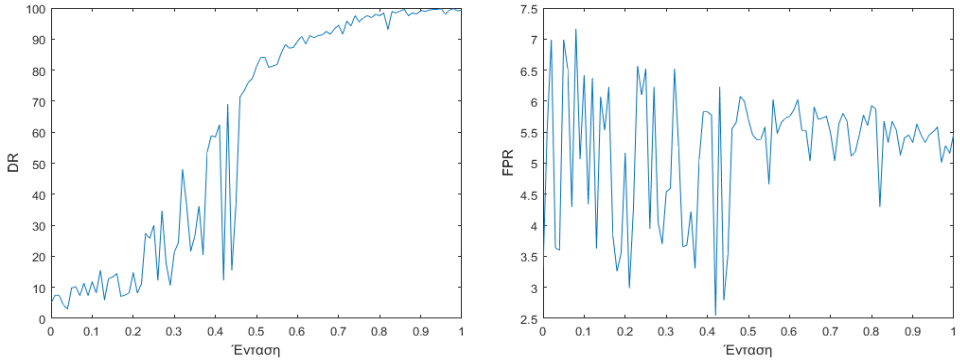
Πίνακας 5.16: Δοκιμή στους κανόνες

5.4 Δοκιμή συστήματος μη επιβλεπόμενης μάθησης

Για να επιβεβαιωθεί η ορθή και βέλτιστη λειτουργία του συστήματος, απαιτείται δοκιμή των παραμέτρων που το απαρτίζουν. Για να συμβεί αυτό, επιλέχθηκαν 4.500 καταναλωτές και αλλοιώθηκαν τα δεδομένα μόνο του 10%. Ο τύπος απάτης που χρησιμοποιήθηκε για την αλλοίωση των δεδομένων είναι ο πρώτος, καθώς φάνηκε πως είναι ιδιαίτερα πολύπλοκο ακόμη και για επιβλεπόμενο σύστημα να παράξει αξιόπιστα αποτελέσματα στους υπόλοιπους τύπους απάτης. Παράλληλα, έγινε μια ακόμη δοκιμή μη επιβλεπόμενου συστήματος εξερευνώντας τις δυνατότητες της ασαφούς συσταδοποίησης FCM.

5.4.1 Αποτελέσματα δοκιμής συστήματος

Παρατηρώντας το Σχήμα 5.2α΄, μπορούμε να παρατηρήσουμε πως έχουμε ομαλή αύξηση του DR μετά το 0.5, ενώ αντίστοιχα έχουμε ομαλή μείωση του FPR μετά το ίδιο σημείο. Πριν από το σημείο αυτό, οι κυματομορφές έχουν σχετική ασυνέπεια στα αποτελέσματα, κάνοντας βίαιες αλλαγές στις μετρικές τους. Πιο συγκεκριμένα, στο εύρος [0.4, 0.5] εμφανίζονται δύο μεγάλα πλήγματα στην επίδοση του συστήματος που προδίδουν πως υπό κάποιες συνθήκες το σύστημα δυσκολεύεται να ορίσει την κλοπή, χωρίς όμως να ενοχοποιεί αδίκως.



(α΄) DR συναρτήσει της έντασης της κλοπής

(β΄) FPR συναρτήσει της έντασης της κλοπής

Σχήμα 5.2: Επίπτωση της έντασης στα αποτελέσματα

Παράλληλα, αξίζει να σημειωθεί εδώ πως οι αλγόριθμοι συσταδοποίησης μπορούν να αλλάξουν σε κάποιο βαθμό τα χαρακτηριστικά του συστήματος και την επίδοσή του. Ως αποτέλεσμα, έγινε δοκιμή για τους διαφορετικούς αλγορίθμους που χρησιμοποιήθηκαν στην εξαγωγή των χαρακτηριστικών, καταλήγοντας σε αποτελέσματα για κάθε περίπτωση.

Αλγ.	DR	FPR	Accuracy	F1 score	BDR %
K-Means	86.44	5.43	93.76	73.47	64
SOM	89.11	5.23	94.2	75.45	65
Fuzzy	85.78	4.99	94.09	74.37	66

Πίνακας 5.17: Εξερεύνηση συσταδοποιήσεων χαρακτηριστικών στο μη-επιβλεπόμενο σύστημα

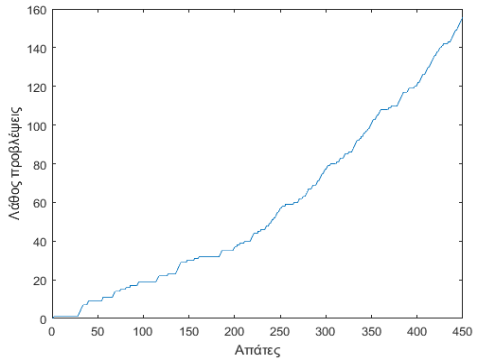
Γίνεται, λοιπόν, αντιληπτό πως οι αλγόριθμοι συσταδοποίησης στην εξαγωγή δεδομένων παίζουν σχετικά μικρό ρόλο, αφού τα αποτελέσματα έχουν πολύ μικρές αποκλίσεις μεταξύ τους. Αυτό ήταν κάτι αναμενόμενο βέβαια, καθώς μόνο δύο από τα οκτώ χαρακτηριστικά έχουν άμεση συσχέτιση με τη συσταδοποίηση.

5.4.2 Εξερεύνηση δυνατοτήτων FCM

Ο αλγόριθμος ασαφών C μέσων μέσα από τον παράγοντα ασάφιας δίνει τη δυνατότητα να εξερευνηθούν οι συστάδες και με διαφορετικούς τρόπους. Ειδικότερα, ο παράγοντας αυτός καθορίζει την επικάλυψη των συστάδων και στη συγκεκριμένη δοκιμή επιλέχθηκε παράγοντας ασάφειας 3 με το 1 να αντιστοιχεί σε συσταδοποίηση χωρίς επικαλύψεις. Παράλληλα, για να μπορέσει να διευκρινιστεί τελικά πού ανήκει κάθε παράδειγμα, παρέχεται μια τιμή για παράδειγμα με τη μεγαλύτερη από αυτή να υποδηλώνει μεγάλο βαθμό ομοιότητας του παραδείγματος με τη συστάδα.

Με αυτό το σκεπτικό δημιουργήθηκε μια δοκιμή με 4.500 καταναλωτές και 10% διείσδυση μη τεχνικών απωλειών κατά την οποία ταξινομούνται οι καταναλωτές χωρίς εξαγωγή χαρακτηριστικών. Ειδικότερα, γνωρίζοντας σε αδρές γραμμές το ποσοστό των απατών, τίθεται ένα όριο στο πλήθος που επιθυμεί κάποιος να ελέγξει. Ο αλγόριθμος βάσει αυτού του πλήθους επιλέγει το δείγμα των καταναλωτών που φαίνεται πιο σίγουρο ότι ανήκει στη συστάδα με ακανόνιστες μετρήσεις. Στην πράξη, αν από 450 κλοπές τεθεί ένα όριο στην εύρεση μόνο των 100, ο αλγόριθμος έχει τη δυνατότητα να αναγνωρίσει σωστά 81, ενώ λάθος 19, όπως φαίνεται και στο Σχήμα 5.3.

Πρακτικά, για να συμβεί αυτό ταξινομούνται σε αύξουσα σειρά οι καταναλωτές με το μεγαλύτερο ποσοστό συμμετοχής στην ακανόνιστη συστάδα. Έτσι, οι πιο ανώμαλες καταναλωτικές συμπεριφορές έχουν προτεραιότητα στον έλεγχο και καθίσταται δυνατό να οριστεί ένα όριο χαμηλότερο από τις συνολικές απώλειες για έλεγχό τους βάσει της ταξινόμησης, εμπνέοντας μεγαλύτερο βαθμό σιγουριάς.



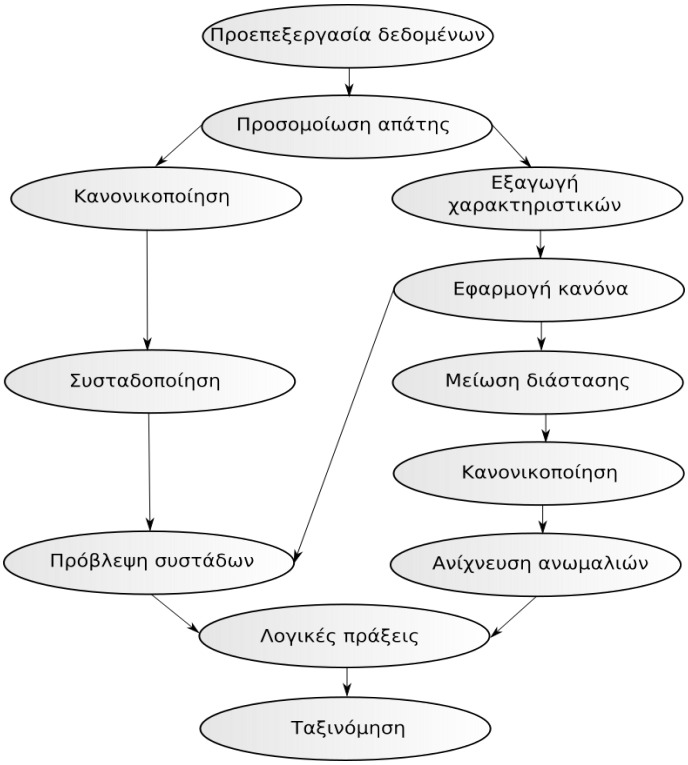
Σχήμα 5.3: Καμπύλη λάθος προβλέψεων με FCM

5.5 Συστατικά συστήματος ημι-επιβλεπόμενης μάθησης

Η ημι-επιβλεπόμενη προσέγγιση του προβλήματος επιτυγχάνεται εισάγοντας στο σύστημα μη επιβλεπόμενης μάθησης νέους αλγορίθμους. Με αυτό τον τρόπο αποκτάται η δυνατότητα εκπαίδευσης με ένα μικρό δείγμα καταναλωτών και των δύο τάξεων ή με μεγαλύτερο δείγμα καταναλωτών της αρνητικής τάξης. Έτσι, καλύπτονται και οι δύο δημοφιλέστερες προσεγγίσεις της ημι-επιβλεπόμενης μάθησης. Εν συνεχεία, βάσει του μοντέλου της εκπαίδευσης ταξινομούνται οι καταναλωτές. Παράλληλα, η προσθήκη νέων αλγορίθμων δίνει τη δυνατότητα εποπτείας των χαρακτηριστικών, αλλά και του μοντέλου που δημιουργήθηκε σε δισδιάστατο χώρο. Οπτικοποιούνται λοιπόν οι πληροφορίες και η εσωτερική λειτουργία του αλγορίθμου, ενώ παράλληλα παρέχεται η δυνατότητα εκπαίδευσης προτύπων.

Αναλυτικότερα, η δομή του αλγορίθμου αναπαρίσταται στο Σχήμα 5.4 ενώ αξίζει να γίνει μια εισαγωγή στα κομμάτια που απαρτίζουν το σύστημα:

- *Προεπεξεργασία δεδομένων*: επιλέγονται και οργανώνονται τα δεδομένα σε συγκεκριμένους πίνακες και διανύσματα.
- *Προσομοίωση απάτης*: αλλοιώνονται οι μετρήσεις κάποιων καταναλωτών και ενημερώνονται οι προϋπάρχοντες πίνακες και διανύσματα.
- *Κανονικοποίηση*: κανονικοποιούνται οι ετήσιες χρονοσειρές και τα χαρακτηριστικά κάθε καταναλωτή σε εύρος τιμών [-1,1] και [0,1] αντίστοιχα.
- *Συσταδοποίηση*: συσταδοποιούνται οι καταναλωτές με βάση τις κανονικοποιημένες τιμές σε δύο συστάδες. Η μια συστάδα ομαλή και η άλλη η ανώμαλη.
- *Εξαγωγή χαρακτηριστικών*: βάσει των χρονοσειρών δημιουργούνται ετήσια χαρακτηριστικά για κάθε καταναλωτή, με στόχο να ανιχνευθεί ύποπτη συμπεριφορά.
- *Εφαρμογή κανόνα*: απενοχοποιούνται κάποιοι καταναλωτές που βρίσκονται στην ανώμαλη συστάδα λαμβάνοντας υπόψη το πλήθος των χαρακτηριστικών .
- *Πρόβλεψη συστάδων*: προβλέπονται οι κλάσεις στις συστάδες με σεβασμό στον κανόνα.



Σχήμα 5.4: Δομή ημι-επιβλεπόμενου ταξινομητή

- *Μείωση διάστασης*: ο πολυδιάστατος χώρος των χαρακτηριστικών μειώνεται σε χώρο δύο διαστάσεων.

- *Ανίχνευση ανωμαλιών*: εκπαιδεύεται το μοντέλο πρόβλεψης βάσει των χαρακτηριστικών και βελτιστοποιούνται τα όρια ταξινόμησης.

- *Λογικές πράξεις*: εκτελούνται λογικές πράξεις μεταξύ των δυαδικών χαρακτηριστικών που προέρχονται από την πρόβλεψη συστάδων και την ανίχνευση ανωμαλιών.

- *Ταξινόμηση*: ταξινομούνται οι καταναλωτές και παράγονται τα τελικά αποτελέσματα και μετρικές.

5.5.1 Θεωρία αλγορίθμου μείωσης διάστασης

Το PCA είναι ένας μη επιβλεπόμενος αλγόριθμος γραμμικής μείωσης διάστασης που στοχεύει στην εύρεση μιας βάσης ή ενός συστήματος συντεταγμένων με περισσότερο νόημα για τα δεδομένα και λειτουργεί βάσει του πίνακα συνδιακύμανσης για την εύρεση ισχυρών χαρακτηριστικών.

Χρησιμοποιείται όταν χρειάζεται να αντιμετωπιστούν οι δυσκολίες των διαστάσεων σε δεδομένα με γραμμικές σχέσεις, καθώς ο μεγάλος αριθμός διαστάσεων (χαρακτηριστικών) μπορεί να δημιουργήσει θόρυβο. Το φαινόμενο αυτό επιδεινώνεται όταν τα χαρακτηριστικά έχουν διαφορετικές κλίμακες.

Αυτό επιτυγχάνεται μειώνοντας τις διαστάσεις δηλαδή χαρακτηριστικά. Η μείωση διάστασης έχει τα εξής θετικά αντίκτυπα με δεδομένες μερικές προϋποθέσεις.

- *Καλύτερη εποπτεία και μικρότερη πολυπλοκότητα*: όταν απαιτείται μια πιο ρεαλιστική εποπτεία των διαστάσεων και υπάρχουν πολλά χαρακτηριστικά σε ένα σετ δεδομένων και ειδικότερα όταν υπάρχει διαισθητική γνώση πως δεν απαιτούνται πολλά χαρακτηριστικά.
- *Καλύτερη οπτικοποίηση*: όταν είναι αδύνατο να έχουμε καλή οπτικοποίηση λόγω του πλήθους των διαστάσεων, χρησιμοποιείται PCA για να μειωθεί σε μια σκιά με δύο ή τρεις διαστάσεις.
- *Μείωση μεγέθους*: όταν υπάρχει μεγάλος όγκος δεδομένων και σκοπεύεται να χρησιμοποιηθούν χρονοβόροι αλγόριθμοι στα δεδομένα, χρειάζεται να ελαχιστοποιηθούν οι πλεονασμοί.
- *Διαφορετική οπτική*: όταν υποβόσκει ανάγκη να αυξηθεί η γνώση πάνω στα δεδομένα. Το PCA μπορεί να δώσει τους καλύτερους γραμμικά ανεξάρτητους και διαφορετικούς συνδυασμούς χαρακτηριστικών, ώστε να περιγραφούν διαφορετικά τα δεδομένα.

Η πρακτική υλοποίηση του PCA είναι εύκολη και συνοψίζεται σε τρία βήματα [17]:

1. Οργάνωση των δεδομένων σε πίνακα $m \times n$, όπου m είναι ο αριθμός των μετρήσεων (χαρακτηριστικών) και n ο αριθμός των δοκιμών.
2. Αφαίρεση του μέσου όρου από κάθε μέτρηση ή από κάθε σειρά.
3. Υπολογισμός SVD των ιδιοδιανυσμάτων της συνδιακύμανσης.

Η συνδιακύμανση μεταξύ δύο χαρακτηριστικών υπολογίζεται ως εξής:

$$\sigma_{jk} = \frac{1}{n-1} \sum_{i=1}^{N} (x_{ij} - \bar{x_j})(x_{ik} - \bar{x_k})$$

Η παραπάνω μπορεί να γενικευθεί σε υπολογισμό του πίνακα συνδιακύμανσης με την ακόλουθη εξίσωση πινάκων:

$$\Sigma = \frac{1}{n-1}((X - \bar{x})^T (X - \bar{x}))$$

όπου $\bar{x}$ είναι το διάνυσμα του μέσου όρου $\bar{x} = \sum_{k=1}^{n} x_i$

Υπάρχουν τρεις προσεγγίσεις οι οποίες αποδίδουν τα ίδια ιδιοδιανύσματα και ζευγάρια ιδιοτιμών:

- Ιδιοπαραγοντοποίηση του πίνακα συνδιακύμανσης μετά από κανονικοποίηση δεδομένων.
- Ιδιοπαραγοντοποίηση του πίνακα συσχέτισης.
- Ιδιοπαραγοντοποίηση του πίνακα συσχέτισης μετά από κανονικοποίηση δεδομένων.

Στην παρούσα εργασία παρ' όλα αυτά, χρησιμοποιείται παραγοντοποίηση ιδιόμορφων ιδιοδιανυσμάτων (SVD) για τη βελτίωση της υπολογιστικής επίδοσης [24].

5.5.2 Θεωρία αλγορίθμου ανίχνευσης ανωμαλιών

Ο αλγόριθμος που χρησιμοποιήθηκε για την ανίχνευση ανωμαλιών είναι βασισμένος στο Γκαουσιανό μοντέλο. Τέτοιες τεχνικές υποθέτουν πως τα δεδομένα δημιουργούνται από μια Γκαουσιανή κατανομή. Οι παράμετροι υπολογίζονται με εκτιμητές μέγιστης πιθανοφάνειας (MLE). Η απόσταση ενός παραδείγματος από το εκτιμώμενο μέσο είναι το αποτέλεσμα του ποσοστού ανωμαλίας. Τίθεται ένα όριο στα ποσοστά αυτά για να οριστούν οι ανωμαλίες [30].

Ερμηνεύοντας αυτή την τεχνική πιο φορμαλιστικά, θεωρούνται χαρακτηριστικά x_i που υποδεικνύουν ανώμαλα παραδείγματα. Για m παραδείγματα εκπαίδευσης και n χαρακτηριστικά ορίζονται τα δεδομένα εξόδου $\{x^{(1)}, ..., x^{(m)}\}$ που δημιουργούν τη μέση τιμή και διακύμανση κάθε χαρακτηριστικού $\mu_1, ..., \mu_n$,$\sigma_1^2, ..., \sigma_n^2$.

$$\mu_j = \frac{1}{m} \sum_{i=1}^{m} x_j^{(i)}$$
$$\sigma_j^2 = \frac{1}{m} \sum_{i=1}^{m} (x_j^{(i)} - \mu_j)^2$$

Δεδομένου ενός νέου παραδείγματος x, υπολογίζεται $p(x)$:

$$p(x) = \prod_{j=1}^{n} p(x_j; \mu_j, \sigma_j^2) = \prod_{j=1}^{n} \frac{1}{\sqrt{2\pi}\sigma_j} exp(-\frac{(x_j^{(i)} - \mu_j)^2}{2\sigma_j^2})$$

Η ανωμαλία λοιπόν ορίζεται αν $p(x) < \epsilon$.
Αντίστοιχα το ϵ είναι προϊόν της διαδικασίας βελτιστοποίησης του αλγορίθμου.

5.5.3 Μεθοδολογία εξαγωγής αποτελεσμάτων

Η μεθοδολογία που χρησιμοποιήθηκε σε αυτό το σύστημα έχει κάποια κοινά στοιχεία με τη μεθοδολογία του συστήματος μη επιβλεπόμενης μάθησης. Συνεπώς, τα αποτελέσματα προέρχονται από δύο βασικές συνιστώσες με την πρώτη να είναι η κανονικοποίηση των καταναλωτών ανά έτος και εν συνεχεία η συσταδοποίησή τους σε δύο συστάδες. Η μία συστάδα έχει αναμενόμενες καταναλωτικές συνήθειες και η άλλη αποτελείται από ασυνήθιστες. Η συστάδα με τις ασυνήθιστες μετρήσεις βελτιστοποιείται με την παρατήρηση των χαρακτηριστικών εξειδίκευσης και έτσι δημιουργείται η πρώτη πρόβλεψη του αλγορίθμου.

Παράλληλα, επεκτείνεται η δεύτερη συνιστώσα, για να αποκτηθεί και δεύτερη πρόβλεψη μέσω των χαρακτηριστικών. Ειδικότερα, τα χαρακτηριστικά περνούν από αλγόριθμο μείωσης διάστασης, για να γίνει εφικτή η εποπτεία των χαρακτηριστικών σε δισδιάστατο περιβάλλον. Εν συνεχεία, χρησιμοποιείται ο αλγόριθμος ανίχνευσης ανωμαλιών για την ολοκλήρωση της δεύτερης πρόβλεψης με δύο διαφορετικές μεθόδους:

- Η πρώτη μέθοδος εξάγει τον μέσο όρο και τη διακύμανση από τα μικτά δεδομένα εκπαίδευσης που χρησιμοποιούνται για την εύρεση της πυκνότητας της πολυμεταβλητής κανονικής κατανομής. Τα δεδομένα δοκιμής και τα δυαδικά χαρακτηριστικά τους χρησιμοποιούνται για τη βελτιστοποίηση του ορίου ταξινόμησης, για να εφαρμοστούν στα δεδομένα εκπαίδευσης. Το σύστημα αυτό αναφέρεται παρακάτω ως τυπικό, καθώς ο αλγόριθμος της ανίχνευσης ανωμαλιών λειτουργεί με τον προκαθορισμένο τρόπο.

- Η δεύτερη μέθοδος εκμεταλλεύεται τη γνώση που παράχθηκε από την πρώτη συνιστώσα εκπαιδεύοντας το μοντέλο μόνο με αρνητικά παραδείγματα που χρησιμοποιούνται για την εύρεση της πυκνότητας της πολυμεταβλητής κανονικής κατανομής στο ένα κομμάτι των δεδομένων δοκιμής. Το άλλο κομμάτι των δεδομένων δοκιμής και τα δυαδικά χαρακτηριστικά τους χρησιμοποιούνται για τη βελτιστοποίηση του ορίου ταξινόμησης που εφαρμόζεται στο πρώτο κομμάτι δεδομένων δοκιμής. Το σύστημα αυτό αναφέρεται ως εναλλακτικό, αφού τα δεδομένα χωρίζονται σε τρία κομμάτια αντί για δύο όπως συνηθίζεται.

Και οι δύο μέθοδοι εξάγουν δυαδικές προβλέψεις για τη δεύτερη συνιστώσα, ολοκληρώνοντας με αυτό τον τρόπο τις προβλέψεις του ταξινομητή. Δεδομένου ότι ο ταξινομητής πρέπει να έχει μια εκτίμηση, τα δύο δυαδικά χαρακτηριστικά εκτελούν μεταξύ τους απλές δυαδικές πράξεις που καταλήγουν στην τελική πρόβλεψη του αλγορίθμου.

5.6 Δοκιμή συστημάτων ημι-επιβλεπόμενης μάθησης

Σκοπός των αλγορίθμων ημι-επιβλεπόμενης μάθησης είναι να παραχθούν βελτιωμένα αποτελέσματα που να προσεγγίζουν τα αποτελέσματα της επιβλεπόμενης μάθησης. Αυτό, όμως, δεν είναι εύκολα εφικτό, καθώς οι συγκεκριμένοι αλγόριθμοι χρησιμοποιούν μόνο το 30% των δυαδικών χαρακτηριστικών, ενώ οι επιβλεπόμενοι αλγόριθμοι το 70%.

Και εδώ, όπως και στις άλλες δοκιμές, χρησιμοποιήθηκαν 4.500 καταναλωτές με 10% να έχουν αλλοιωμένες μετρήσεις. Η κανονικοποίηση των ετήσιων χρονοσειρών επιτεύχθηκε σε εύρος [-1,1], ενώ η κανονικοποίηση των χαρακτηριστικών σε εύρος [0,1].

5.6.1 Εξερεύνηση λογικών πράξεων στα ημι-επιβλεπόμενα συστήματα

Αρχικά, αξίζει να παρατηρηθεί ποια λογική πράξη στην εξαγωγή αποτελεσμάτων παρουσιάζει τα βέλτιστα αποτελέσματα. Η χρήση της OR αναμένεται να διευρύνει τα όρια του ταξινομητή, αλλά αν η δεύτερη συνιστώσα του ταξινομητή είναι εξαιρετικά εύστοχη, οι λάθος προβλέψεις δεν θα αυξηθούν σε μεγάλο βαθμό. Από την άλλη, η χρήση της AND αναμένεται να μειώσει τις λάθος προβλέψεις και να κάνει τον αλγόριθμο πιο προσεκτικό στην επιλογή της απάτης. Στον πίνακα παρακάτω οι παραπάνω υποθέσεις παίρνουν σάρκα και οστά.

Πύλη	DR	FPR	Accuracy	F1 score	BDR %
AND	72.01	2.68	94.76	73.52	75
OR	91.08	7.61	92.25	70.81	57

Πίνακας 5.18: Εξερεύνηση λογικών πράξεων στο τυπικό ημι-επιβλεπόμενο σύστημα

Πύλη	DR	FPR	Accuracy	F1 score	BDR %
AND	90.63	10.58	89.65	77.33	49
OR	98.11	28.59	76.38	60.73	28

Πίνακας 5.19: Εξερεύνηση λογικών πράξεων στο εναλλακτικό ημι-επιβλεπόμενο σύστημα

5.6.2 Εξερεύνηση συσταδοποιήσεων στα ημι-επιβλεπόμενα συστήματα

Βάσει των αποτελεσμάτων του F1 score και του Accuracy επιλέγεται η πύλη AND, καθώς δίνεται μεγαλύτερη βάση στη γενική απόδοση του αλγορίθμου από την απόλυτη ακρίβεια στον εντοπισμό των μη τεχνικών απωλειών (DR). Ειδικότερα, το υψηλότερο F1 score προδίδει πως υποβόσκει πολύ μικρό ποσοστό στη λάθος πρόβλεψη και ικανοποιητικό ποσοστό στον εντοπισμό της απάτης.

Στη συνέχεια, επιλέγεται να γίνει αναλυτική δοκιμή των συσταδοποιήσεων των χρονοσειρών και των χαρακτηριστικών. Δυστυχώς, η συσταδοποίηση SOM αδυνατεί να ολοκληρώσει τη συσταδοποίηση στις χρονοσειρές. Παρ΄ όλα αυτά, έγινε διεξοδική εξερεύνηση των συσταδοποιήσεων. Από προηγούμενες δοκιμές αναμένεται να μην παρατηρηθούν μεγάλες αποκλίσεις στα αποτελέσματα.

Συστ. χρ.	Συστ. χαρ.	DR	FPR	Accuracy	F1 score	BDR
K-Means	K-Means	71.09	2.00	95.49	74.64	0.80
K-Means	SOM	76.85	3.06	94.95	75.04	0.74
FCM	SOM	79.48	3.83	94.54	73.94	0.70
FCM	FCM	78.77	3.75	94.44	74.53	0.70

Πίνακας 5.20: Εξερεύνηση συσταδοποιήσεων στο τυπικό ημι-επιβλεπόμενο σύστημα

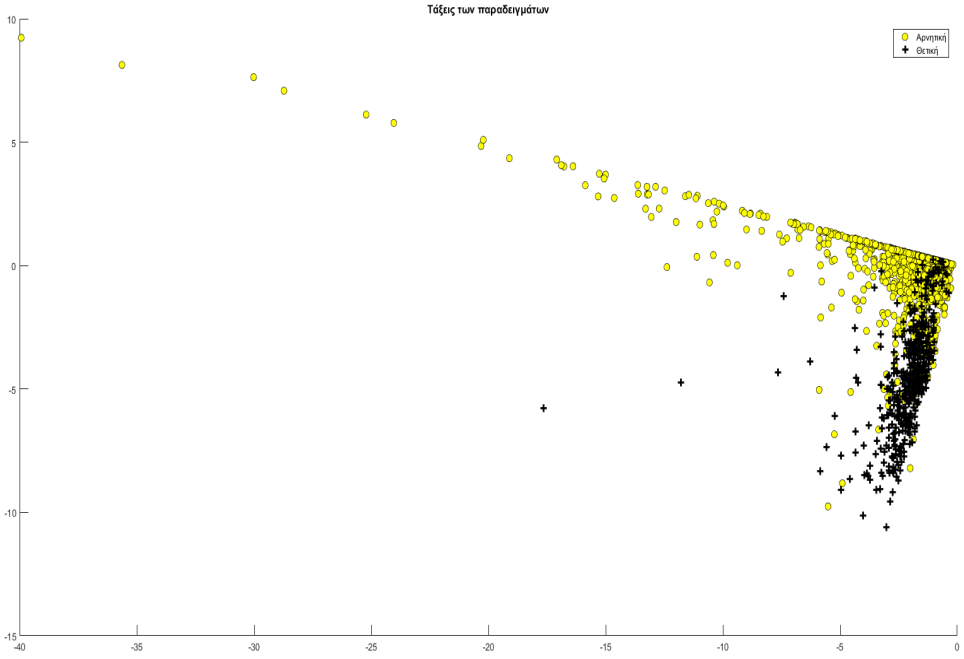
Συστ. χρ.	Συστ. χαρ.	DR	FPR	Accuracy	F1 score	BDR %
K-Means	K-Means	91.15	10.96	89.45	77.01	48
K-Means	SOM	91.89	10.21	90.21	78.92	50
FCM	SOM	91.90	9.04	91.13	79.37	53
FCM	FCM	89.32	9.23	90.51	76.98	52

Πίνακας 5.21: Εξερεύνηση συσταδοποιήσεων στο εναλλακτικό ημι-επιβλεπόμενο σύστημα

5.6.3 Εξερεύνηση μείωσης διάστασης στους ημι-επιβλεπόμενους αλγόριθμους

Σε αυτό το σημείο αξίζει να παρατηρήσουμε τους ορίζοντες που ανοίγει ο αλγόριθμος μείωσης διάστασης. Αρχικά, δίνει τη δυνατότητα να έχουμε εποπτεία σε όλο το σετ δεδομένων και επίσης να παρατηρήσουμε τα όρια που θέτει ο αλγόριθμος ανίχνευσης ανωμαλιών.

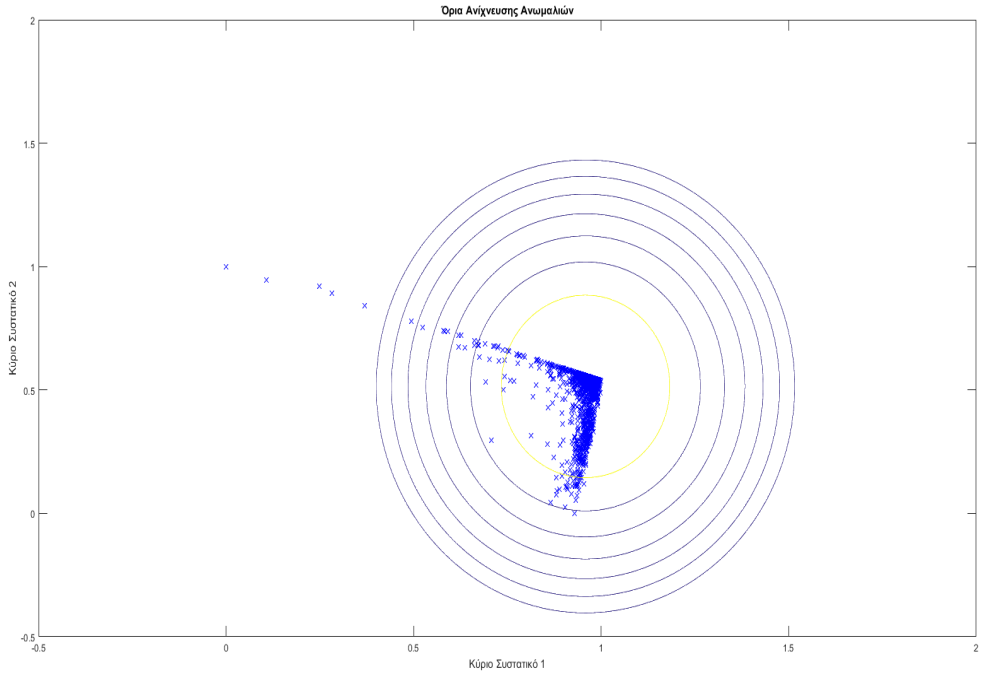
Στο Σχήμα 5.5 παρατηρείται η αποτύπωση που δημιουργεί η μείωση διάστασης στα χαρακτηριστικά των καταναλωτών. Τα κυκλικά κίτρινα σημεία είναι η αρνητική ομάδα, ενώ οι μαύροι σταυροί είναι η θετική ομάδα. Από την κατανομή της αρνητική τάξης εύκολα γίνεται αντιληπτό ότι οι περισσότεροι καταναλωτές βρίσκονται κοντά στο κέντρο των αξόνων. Αντίστοιχα, η θετική τάξη αποτυπώνεται κάτω από το κέντρο των αξόνων μαζί με λίγα αρνητικά παραδείγματα. Γίνεται, συνεπώς, αντιληπτό πως τα σύνολα δεν είναι πλήρως διαχωρίσιμα και για αυτό τον λόγο αναμένεται τα αποτελέσματα με μείωση διάστασης να είναι χειρότερα. Η δυσκολία αυτή ωστόσο παρακάμπτεται με την πρώτη συνιστώσα της ταξινόμησης.



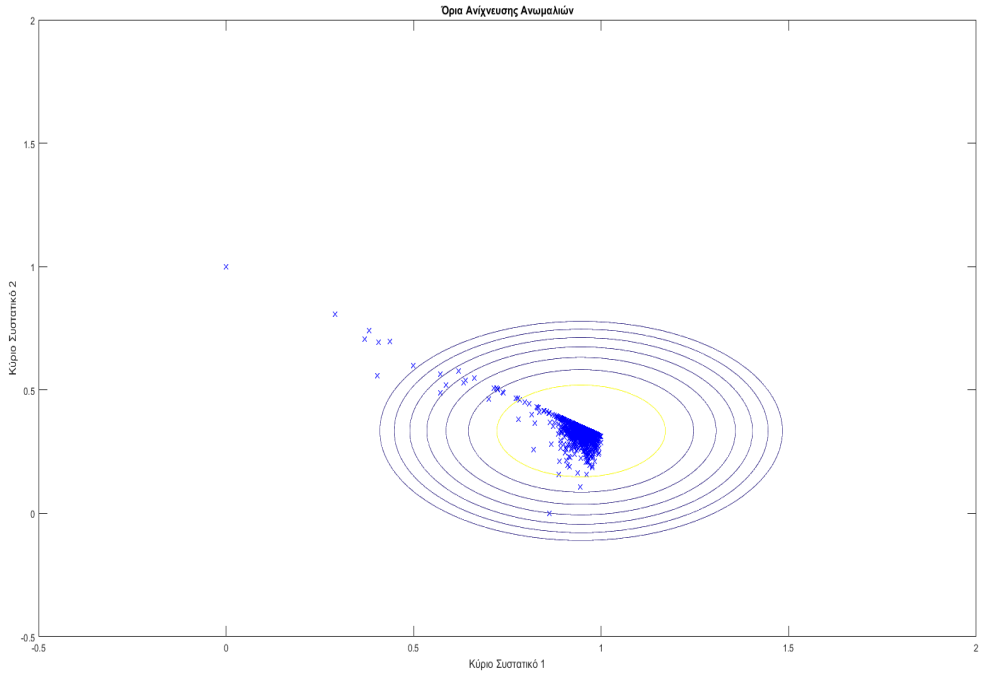
Σχήμα 5.5: Χαρακτηριστικά και τάξεις καταναλωτών

Στο Σχήμα 5.6 παρατηρούνται οι ισοϋψείς καμπύλες της Γκαουσιανής κατανομής. Το Σχήμα 5.6α΄ εμφανίζει την προσαρμογή στα δεδομένα του τυπικού αλγορίθμου, ενώ το Σχήμα 5.6β΄ αποδίδει την προσαρμογή στα δεδομένα του εναλλακτικού αλγορίθμου. Στην πρώτη περίπτωση, είναι εμφανές πως η οριοθέτηση της κατανομής γίνεται σε ένα στενό κύκλο, ενώ στη δεύτερη περίπτωση η οριοθέτηση γίνεται με πιο πλατή κύκλο. Λαμβάνοντας υπόψη τα παραπάνω, γίνεται για άλλη μια φορά σαφές πως η μείωση διαστάσεων δε βοηθά στη βελτιστοποίηση του αλγορίθμου άμεσα, αλλά δίνει μια άλλη οπτική των δεδομένων.

Η οριοθέτηση γίνεται μόνο στα δεδομένα δοκιμής που στον τυπικό αλγόριθμο είναι το 70% των δεδομένων, ενώ στον εναλλακτικό αλγόριθμο είναι το 35%. Παράλληλα, η υλοποίησή της έχει σαν είσοδο τον μέσο όρο της και τη διακύμανση που παρήχθησαν στη δημιουργία του μοντέλου. Η χάραξη των ισοϋψών καμπυλών γίνεται με τη βοήθεια ενός δισδιάστατου πλέγματος που ορίζεται παρατηρώντας τη διάταξη των δεδομένων στο επίπεδο. Εξάγοντας τις πιθανότητες της πολυμεταβλητής Γκαουσιανής κατανομής με είσοδο το πλέγμα και τον μέσο όρο και διακύμανση, δημιουργούνται 7 ομόκεντρα περιγράμματα. Η ακριβής οριοθέτηση ορίζεται από την διαδικασία βελτιστοποίησης του F1 score.



(α΄) Προσαρμογή κατανομής στα τυπικά δεδομένα



(β΄) Προσαρμογή κατανομής στα ενναλακτικά δεδομένα

Σχήμα 5.6: Ισοϋψείς Γκαουσιανής κατανομής

Για να αποδειχθεί και με αποτελέσματα η παραπάνω υπόθεση, γίνεται μια νέα σειρά δοκιμών με τον αλγόριθμο μείωσης διάστασης.

Σύστημα	Πύλη	DR	FPR	Accuracy	F1 score	BDR %
τυπικό	AND	79.01	2.51	95.59	78.65	78
εναλλακτικό	AND	0.00	0.00	80.28	-	-
εναλλακτικό	OR	87.85	11.79	88.14	73.44	45

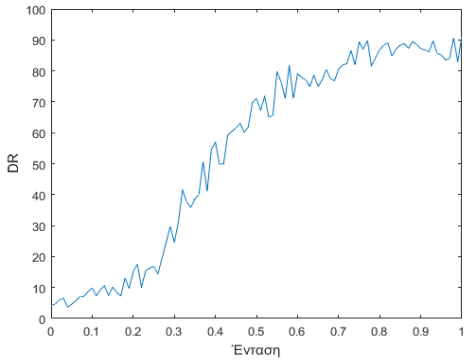
Πίνακας 5.22: Εξερεύνηση μείωσης διάστασης στους ημι-επιβλεπόμενους αλγορίθμους

5.6.4 Αποτελέσματα δοκιμής συστημάτων

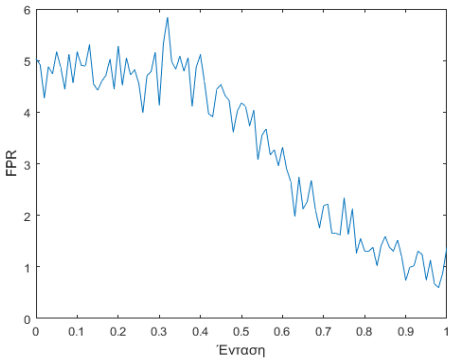
Για την τελική δοκιμή επιλέχθηκαν όλοι οι πιθανοί καταναλωτές και εισήχθη 10% ποσοστό κλοπών. Ένας ικανοποιητικός τρόπος να παρατηρηθεί η λειτουργία του αλγορίθμου είναι να δοκιμαστεί το σύστημα υπό διαφορετικές εντάσεις κλοπής. Έτσι επιλέχθηκαν τα συστήματα με K-Means, χωρίς μείωση διάστασης και πύλες AND που είχαν τα πιο συνεπή αποτελέσματα.

Στον τυπικό αλγόριθμο παρατηρούνται ομαλές καμπύλες με συνεχή αύξουσα πορεία στο DR και σχεδόν συνεχή φθίνουσα στο FPR. Παράλληλα, αξίζει να σημειωθεί πως ενώ το DR δεν φτάνει το 100% αλλά το 90%, το FPR φτάνει σε εξαιρετικά χαμηλά επίπεδα πράγμα που ουσιαστικά σημαίνει ότι πρόκειται για ένα αρκετά συμπαγές και έμπιστο σύστημα.

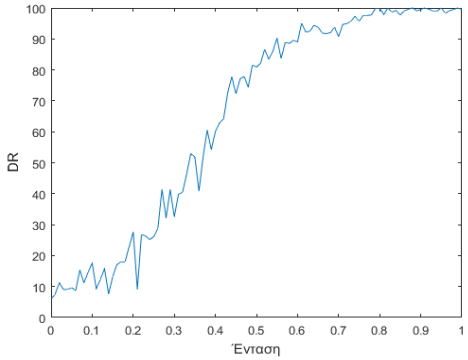
Από την άλλη πλευρά, ο εναλλακτικός αλγόριθμος έχει και αυτός ομαλή καμπύλη DR με αύξουσα κυρίως πορεία, αλλά η καμπύλη FPR έχει ιδιαίτερα περίεργη συμπεριφορά. Το σύστημα φαίνεται πως όσο η ένταση αυξάνει τη σιγουριά της πρόβλεψης ως προς την απάτη, παράλληλα αυξάνει σε τόσο μεγάλο ποσοστό και την ενοχοποίηση των καταναλωτών, επιλέγοντας πιο αυθαίρετα τη θετική κλάση, με αποτέλεσμα τη σταδιακή αύξηση και των δύο μετρικών.



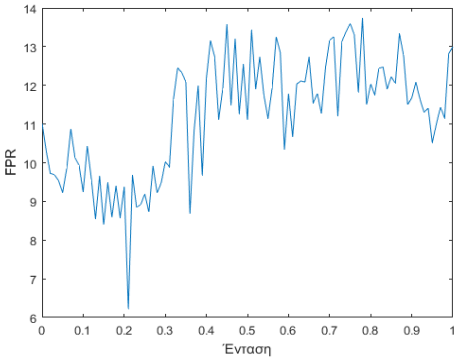
(α′) DR συναρτήσει της έντασης του τυπικού αλγορίθμου



(β′) FPR συναρτήσει της έντασης του τυπικού αλγορίθμου



(γ′) DR συναρτήσει της έντασης του εναλλακτικού αλγορίθμου



(δ′) FPR συναρτήσει της έντασης του εναλλακτικού αλγορίθμου

Σχήμα 5.7: Δοκιμή έντασης ημι-επιβλεπόμενων συστημάτων

5.7 Σχόλια

Στο παρόν κεφάλαιο έγινε μια διεξοδική αναζήτηση μη επιβλεπόμενων και ημι-επιβλεπόμενων συστημάτων με σκοπό την άμεση σύγκριση με τον αλγόριθμο επιβλεπόμενης μάθησης. Τα αποτελέσματα δείχνουν πως μπορεί να υπάρξει σύστημα που εντοπίζει μη τεχνικές απώλειες χωρίς καμία εκπαίδευση και χωρίς τη χρήση δυαδικών χαρακτηριστικών. Παράλληλα, η επίδοση του ημι-επιβλεπόμενου συστήματος δίνει τη δυνατότητα κατανόησης πως ακόμη και με λίγα δυαδικά χαρακτηριστικά ο μη επιβλεπόμενος αλγόριθμος μπορεί να γίνει πιο αξιόπιστος στην αναγνώριση απάτης και να βελτιώσει την επίδοσή του. Συγκρίνοντας τους ημι-επιβλεπόμενους αλγόριθμους, γίνεται φανερό πως η τυπική ανίχνευση ανωμαλιών είναι γενικά πιο προβλέψιμη και πιο εύστοχη.

Συνοψίζοντας, καθίσταται σαφές πως οι δύο συνιστώσες ταξινόμησης είναι απαραίτητες για την εξαγωγή ικανοποιητικών μετρικών και πως η αλληλεπίδρασή τους καθιστά την ταξινόμηση αξιόπιστη. Από την άλλη πλευρά, η μέθοδος συσταδοποίησης δεν έχει εξαιρετική σημασία για την απόδοση του συστήματος, ακόμη και όταν αλλάζει η μέθοδος και στους δύο άξονες ταξινόμησης. Τέλος, για την αύξηση εμπιστοσύνης στα συστήματα απαιτείται η χρήση της πύλης AND για την τελική δυαδική πράξη των ταξινομητών, καθώς επιτυγχάνει ικανοποιητικά

αποτελέσματα σε δύο πολύ σημαντικές μετρικές, το F1 score και το Accuracy.

Κεφάλαιο 6

Δυσκολίες στην εκπόνηση της διπλωματικής

Στην παρούσα διπλωματική αντιμετωπίστηκαν δυσκολίες που εν μέρει όρισαν τη μελλοντική πορεία του ζητήματος. Υπήρξαν δύο ειδών τεχνικές δυσκολίες στην ανίχνευση μη τεχνικών απωλειών σε ετήσια δεδομένα. Η πρώτη βασίζεται στο γεγονός ότι πρόκειται για χρονοσειρές διάρκειας ενός έτους, στις οποίες δεν μπορεί εύκολα να αποτυπωθεί μια αξιόπιστη καταναλωτική συμπεριφορά. Η δεύτερη σχετίζεται με την ευρεία χρήσης και δοκιμή πολλών ταξινομητών και την ανάγκη να λαμβάνονται υπόψη οι ιδιαίτερες απαιτήσεις του καθενός. Παράλληλα, πρέπει να καθοριστεί και ένα όριο στην αξιοπιστία των συστημάτων μηχανικής μάθησης, καθώς ένα αποτελεσματικό σύστημα πρέπει να έχει σιγουριά στον εντοπισμό του ζητούμενου συμβάντος και να ελαχιστοποιεί τα περιθώρια λάθους εκτίμησης.

Στην πιο ευρεία σφαίρα του ζητήματος, τίθενται θέματα προστασίας της ιδιωτικότητας των καταναλωτών. Από την άλλη, τους δίνεται η δυνατότητα ανωνυμοποίησης των δεδομένων τους [6], γεγονός που δυσκολεύει σε μεγάλο βαθμό την εξόρυξη δεδομένων σε επόμενα στάδια. Με την ύπαρξη των έξυπνων μετρητών ανοίγεται ένα παράθυρο που εκθέτει τις προσωπικές δραστηριότητες σε οποιονδήποτε έχει πρόσβαση σε καταναλωτικές πληροφορίες. Οι τεράστιες δυνατότητες που ανοίγονται στην αναλυτική μελέτη χρονοσειρών εγείρουν ζητήματα προστασίας των προσωπικών δεδομένων.

6.1 Τεχνικά εμπόδια

Η αντιμετώπιση τεχνικών θεμάτων πάντα απαιτεί λεπτομερή ανάλυση της δυσκολίας και λήψεις αποφάσεων. Η έκταση των δεδομένων αποδείχθηκε σχετικά μικρή, καθώς τα συστήματα δεν είχαν τη δυνατότητα παρατήρησης των καταναλωτικών συνηθειών σε μεγάλο βάθος χρόνου. Το συγκεκριμένο πρόβλημα γεννά νέες δυσκολίες και μπορεί να προκαλέσει την αναξιοπιστία του συστήματος σε δεδομένα άλλων χρονικών περιόδων. Τέλος, αξίζει να ληφθεί υπόψη πως η διαδικασία εύρεσης και επεξεργασίας δεδομένων και χαρακτηρισμών τους είναι εξαιρετικά επίπονη και απαιτεί εμπιστοσύνη στην πηγή τους.

6.1.1 Έλλειψη μακροχρόνιων δεδομένων

Για να μπορέσει να αντιμετωπιστεί το ζήτημα των μη τεχνικών απωλειών με μακροπρόθεσμο ορίζοντα, απαιτείται η βαθιά κατανόηση της συχνότητας των προτύπων και των στιγμιοτύπων των χρονοσειρών. Με αυτό τον τρόπο, αναλύονται σε βάθος οι καταναλωτικές συνήθειες και γνωστοποιούνται οι μεταβλητές που τις επηρεάζουν. Τα δεδομένα της παρούσας εργασίας αφορούσαν χρονικό διάστημα που δεν ξεπερνούσε τα δύο έτη. Με τέτοιο εύρος μετρήσεων ήταν λοιπόν λογικό να περιοριστούν οι δοκιμές σε ενός έτους.

Εκεί που εγείρεται η σημαντική δυσκολία είναι το γεγονός ότι οι καταναλωτές ταξινομούνται με ένα και μόνο έτος αναφοράς. Ειδικότερα, τα συστήματα χρησιμοποιούν τις γενικές καταναλωτικές συνήθειες του έτους για να ταξινομήσουν κάθε καταναλωτή με αυτά τα κριτήρια. Η πιο ασφαλής προσέγγιση, για να κριθεί ένα έτος ύποπτο, θα απαιτούσε να υπάρχει μεγάλο χρονικό παράθυρο κατανάλωσης, ώστε να μπορεί εύκολα κάποιος να παρατηρήσει μια ασυνήθιστη τάση των δεδομένων. Έτσι, θα μπορούσαν να οργανωθούν ευκολότερα οι καταναλωτές σε ομάδες που θα είχαν μια γενικότερη ομοιότητα ως προς τις καταναλωτικές συνήθειες.

6.1.2 Έλλειψη παραδειγμάτων

Παράλληλα, έχει νόημα να παρατηρηθεί πως το δείγμα των καταναλωτών δεν είναι τελείως αντιπροσωπευτικό ως προς τη δυνατότητα γενίκευση σε μεγαλύτερο πληθυσμό. Ειδικότερα, οι 4.500 καταναλωτές θα μπορούσαν να είχαν πολύ διαφορετικές συνήθειες, αν ζούσαν σε διαφορετική τοποθεσία, άρα και διαφορετικές χρονοσειρές που θα εξετάζονταν διαφορετικά, αν απέκλιναν σημαντικά από τις υπάρχουσες. Το πρόβλημα εντείνεται, παρακολουθώντας την ομοιογένεια των τύπων των καταναλωτών. Εμφανίζεται, δηλαδή, μια κυρίαρχη ομάδα που έχει σχετική ομοιογένεια μεταξύ της και αποτελείται από νοικοκυριά και οικιακούς χρήστες. Στην ομάδα αυτή ανήκουν τουλάχιστον τα τρία τέταρτα του δείγματος, γεγονός που έχουν εκμεταλλευτεί τα συστήματα ταξινόμησης, αλλά αίρονται ερωτήματα για το υπόλοιπο ένα τέταρτο του πληθυσμού, το οποίο στελεχώνεται από καταναλωτές με υψηλές ενεργειακές απαιτήσεις, δηλαδή από μικρομεσαίες επιχειρήσεις. Αυτό το μικρό δείγμα δεν μπορεί να εξάγει εύκολα μια γενικευμένη συμπεριφορά που να εκφράζει όλο το σύνολο, καθώς κάθε επιχείρηση ανάλογα με τις ανάγκες της προσαρμόζει τη λειτουργία της. Το αποτέλεσμα είναι να έχουμε ένα ικανοποιητικό πλήθος ομοιόμορφων καταναλωτών που εξάγουν όμοια χαρακτηριστικά και ένα μικρό υποσύνολο των δεδομένων με επιχειρήσεις, που έχουν μεγάλες και αδιευκρίνιστες ανάγκες.

6.1.3 Δυσκολία επιλογής μετρικών

Στην παρούσα διπλωματική χρησιμοποιήθηκε πλήθος αλγορίθμων μηχανικής μάθησης με καθένα να έχει τα δικά του ιδιαίτερα χαρακτηριστικά. Δημιουργήθηκε, λοιπόν, η ανάγκη σύγκρισης των αλγορίθμων βάσει κάποιων απόλυτων μετρικών για την τελική αξιολόγησή τους. Ειδικότερα, οι επιβλεπόμενοι αλγόριθμοι χρησιμοποιούν 70% των δεδομένων για εκπαίδευση και το 30% για προβλέψεις, οι μη επιβλεπόμενοι δεν χρησιμοποιούν εκπαίδευση για τη δημιουργία μοντέλου πρόβλεψης, ενώ οι ημι-επιβλεπόμενοι χρησιμοποιούν 70% για την

εξαγωγή του στατιστικού μοντέλου και την πρόβλεψη και 30% για τη βελτιστοποίηση του μοντέλου. Όπως γίνεται αντιληπτό, οι προβλέψεις γίνονται σε διαφορετικά δείγματα των πληθυσμών, δημιουργώντας απαίτηση για αξιόπιστες μετρικές.

Τα DR και FPR μπορούν γρήγορα να δώσουν μια πρώτη αίσθηση για την ευστοχία του αλγορίθμου, αλλά λόγω της ευαισθησίας του προβλήματος δεν πρέπει να θεωρούνται οι κύριες μετρικές. Αυτό οφείλεται στο γεγονός ότι ένας αλγόριθμος με πολύ υψηλό DR μπορεί να αναγνωρίσει τις κλοπές, αλλά αν έχει FPR που ξεπερνά το 5%, οι προβλέψεις δεν θεωρούνται εντελώς αξιόπιστες, καθώς εισάγεται μεγάλο περιθώριο λάθους. Ένας τρόπος να αποτυπωθεί η σχέση μεταξύ του DR και FPR είναι το F1 score, που έχει εξάρτηση και από τις δύο μετρικές και εξάγει ικανοποιητικά αποτελέσματα μόνο με χαμηλό FPR. Παράλληλα, ένας γενικότερος τρόπος να εξεταστεί η ταξινόμηση είναι με την ευστοχία Accuracy που πρέπει να βρίσκεται πάντα πάνω από το 90% και περιγράφει τη γενικότερη πρόβλεψη του συστήματος. Όταν οι αλγόριθμοι έχουν παρόμοιες αυτές τις μετρικές, αξίζει να ελεγχθεί το BDR που προσφέρει μια πιθανοτική προσέγγιση, ορίζοντας την πιθανότητα πραγματικής κλοπής, δεδομένου ότι προβλέφθηκε.

6.1.4 Εύρεση αξιόπιστων δυαδικών χαρακτηρισμών

Ένα σημαντικός παράγοντας που δεν πρέπει να αμεληθεί είναι η αξιοπιστία και η προέλευση των δυαδικών χαρακτηριστικών των χρονοσειρών. Στην παρούσα εργασία δεν απαιτήθηκε να ευρεθούν τέτοια δεδομένα, καθώς προσομοιώθηκαν οι απάτες. Στην περίπτωση όμως που τα δεδομένα έρχονται με δυαδικούς χαρακτηρισμούς από ένα φορέα, απαιτείται έλεγχος στη μεθοδολογία εξαγωγής των χαρακτηριστικών. Η εγκυρότητα των δυαδικών αυτών διανυσμάτων είναι καίριας σημασίας για την εκπαίδευση και τον έλεγχο του συστήματος, καθώς αποτελεί τη βάση της υλοποίησης των αλγορίθμων και την κινητήριο δύναμη των αλγορίθμων βελτιστοποίησης. Επιπρόσθετα αξίζει να σημειωθεί πως θα μπορούσε να δημιουργηθεί ένα σύστημα με ανατροφοδότηση των φυσικών ελέγχων για τη δημιουργία αξιόπιστων δυαδικών χαρακτηριστικών.

6.2 Ασφάλεια Καταναλωτών

Η εισαγωγή των έξυπνων μετρητών στην καθημερινότητά μας δίνει τη δυνατότητα να διερευνηθούν σε βάθος οι καταναλώσεις ενέργειας και διευκολύνει την επικοινωνία των δεδομένων με εγκεκριμένους φορείς. Αυτή όμως η πραγματικότητα έχει και μια σκοτεινή πτυχή που αντιμετωπίζεται στις περισσότερες μελέτες μεγάλης κλίμακας δεδομένων. Οι προσωπικές πληροφορίες των πελατών είναι εκτεθειμένες σε ένα δίκτυο αμφίδρομης επικοινωνίας καταναλωτών και παρόχων, ενώ ανά πάσα στιγμή κάποιος εργαζόμενος μπορεί να ανατρέξει σε αυτές και να της εκμεταλλευτεί για προσωπικούς λόγους.

Η σημερινή τεχνολογία των έξυπνων μετρητών που βασίζονται στο NALM αλγόριθμο, παρέχει τρόπους να αναγνωρίζονται συσκευές σε λειτουργία ακόμη και όταν οι μετρήσεις αφορούν ένα σύνολο νοικοκυριών. Έτσι, κάποιος κακόβουλος χρήστης θα μπορούσε να

αντλήσει δεδομένα για το πρόγραμμα των νοικοκυριών, τα είδη των συσκευών τους και τις ανάγκες τους. Ένας τρόπος να αντιμετωπιστεί αυτό το θέμα είναι η διαχείριση της ενεργειακής χρήσης μέσα στο σπίτι, πριν συλλεχθούν τα δεδομένα του μετρητή [6].

Γίνεται λοιπόν σαφές πως οι έξυπνοι μετρητές χωρίς κάποιο σύστημα ανωνυμοποίησης πλήττουν την ιδιωτικότητα των καταναλωτών και εγείρουν θέματα ασφαλείας. Η έρευνα προς αυτή την κατεύθυνση υπερβαίνει το πλαίσιο αυτής τη διπλωματική εργασίας, αλλά ήδη προτείνονται νέοι αλγόριθμοι και δικτυακές δομές, για να μπορέσει να συμβαδίσει η προστασία της ιδιωτικότητας με την αποτελεσματικότητα των ερευνών.

Κεφάλαιο 7

Συμπεράσματα και δυνατότητες μελλοντικής επέκτασης

Το κεφάλαιο αυτό συνοψίζει όλη τη γνώση που δημιουργήθηκε από τη μελέτη των αλγορίθμων και την εξαγωγή αποτελεσμάτων. Παράλληλα, παρατηρώντας από ένα ευρύτερο πεδίο το θέμα δημιουργείται μια άλλη οπτική στην αντιμετώπιση του ζητήματος του εντοπισμού των μη τεχνικών απωλειών. Παρατηρούνται τα πλεονεκτήματα και τα μειονεκτήματα κάθε συστήματος, καθώς δίνεται βάση στο εύρος του πεδίου εφαρμογής του καθενός και στις δυνατότητές του. Τέλος διατυπώνονται κάποιες επισημάνσεις που έχουν ως κύριο μέλημα τη βελτιστοποίηση των συστημάτων.

7.1 Σύγκριση αποτελεσμάτων

Κάνοντας μια επισκόπηση στα αποτελέσματα, εύκολα παρατηρείται πως ο επιβλεπόμενος αλγόριθμος έχει την καλύτερη σχέση μεταξύ ποσοστού ευστοχίας στην εύρεση DR και ποσοστού λάθος προβλέψεων FPR. Αυτό ήταν αναμενόμενο από τα πρώτα στάδια της διπλωματικής, καθώς ο επιβλεπόμενος αλγόριθμος είναι ευρέως μελετημένος και είναι κοινά αποδεκτή η αποτελεσματικότητά του σε τέτοιου είδους δεδομένα. Παράλληλα, παρατηρείται πως ο αλγόριθμος μη επιβλεπόμενης μάθησης έχει υψηλότερο DR αλλά και υψηλότερο FPR στα όρια της κόκκινης γραμμής, που ορίστηκε στο 5%. Με αυτό το σκεπτικό δημιουργήθηκε το ημι-επιβλεπόμενο σύστημα, ώστε να χαμηλώσει το FPR και να παραχθούν πιο σίγουρες προβλέψεις.

Τίθεται, λοιπόν, σαν άξονας αναφοράς ο επιβλεπόμενος αλγόριθμος που από τη μία έχει τα καλύτερα αποτελέσματα από τα άλλα συστήματα, αλλά από την άλλη είναι ο λιγότερο εφαρμόσιμος σε πραγματικά προβλήματα, λόγω της ανάγκης ύπαρξης δυαδικών χαρακτηριστικών. Συγκρίνοντας τα συστήματα με τον επιβλεπόμενο αλγόριθμο παρατηρούνται τα εξής:

- Το μη επιβλεπόμενο σύστημα κατέχει το σημαντικότερο πλεονέκτημα, που είναι η ευρεία και άμεση εφαρμογή του σε υπάρχοντα προβλήματα. Αυτό συμβαίνει, καθώς δεν απαιτεί κανενός είδους εκπαίδευση, αλλά μόνο εφαρμογή συμπαγών και αξιόπιστων κανόνων που να διαχωρίζουν τις δύο κλάσεις. Παράλληλα, λόγω της έλλειψης εκπαίδευσης, η

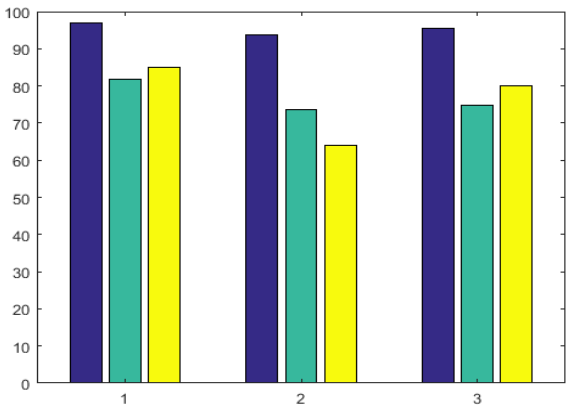
δημιουργία του μοντέλου διαχωρισμού γίνεται ταχύτατα. Στα μειονεκτήματα του αλγορίθμου εντάσσεται η οριακή του γενική απόδοση. Το FPR είναι ακριβώς στα όρια ανοχής που ορίστηκαν (5%), ενώ το Accuracy είναι λίγο χαμηλότερα από τα επιθυμητά επίπεδα (95%).

- Το ημι-επιβλεπόμενο σύστημα απαιτεί μόνο μικρό ποσοστό καταναλωτών για βελτιστοποίηση του ορίου επιλογής, γεγονός που το καθιστά πιο εύκολα εφαρμόσιμο από το επιβλεπόμενο σύστημα, αλλά λιγότερο εφαρμόσιμο από το μη επιβλεπόμενο σύστημα. Ουσιαστικά ο ημι-επιβλεπόμενος αλγόριθμος λειτουργεί σαν ενδιάμεση λύση σε όλα τα κριτήρια που έχουν τεθεί. Ειδικότερα, η απόδοσή του είναι βελτιωμένη με εξαιρετικά χαμηλό FPR και βελτιωμένο Accuracy. Παρόλο που το σύστημα φαίνεται να έχει σχετικά χαμηλό DR, αυτό εξισορροπείται από τη σιγουριά της πρόβλεψης που δίνεται από το BDR. Πιο συγκεκριμένα, όταν το σύστημα προβλέπει απάτη, είναι 80% σίγουρο ότι πρόκειται για απάτη, ποσοστό που προσεγγίζει σε μεγάλο βαθμό το επιβλεπόμενο σύστημα.

Σύστημα	DR	FPR	Accuracy	F1	BDR
επιβλεπόμενο	80.87	1.54	96.96	81.94	0.85
μη-επιβλεπόμενο	86.44	5.43	93.76	73.47	0.64
ημί-επιβλεπόμενο	71.09	2.00	95.49	74.64	0.80

Πίνακας 7.1: Σύγκριση συστημάτων

Για την οπτικοποίηση αυτών των παρατηρήσεων δημιουργήθηκε ένα γράφημα δίνοντας βάση στο Accuracy, F1 score και BDR. Καθίσταται λοιπόν σαφές, πως η ημι-επιβλεπόμενη μάθηση βρίσκεται ακριβώς ανάμεσα στις επιδόσεις και στη χρηστικότητα του επιβλεπόμενου και μη επιβλεπόμενου συστήματος.



Σχήμα 7.1: Σύγκριση συστημάτων

7.2 Συμπερασματικές σημειώσεις

Συνοψίζοντας χρήσιμες πληροφορίες που παρήχθησαν από αυτή τη μελέτη, γίνεται σαφές πως η ανίχνευση μη τεχνικών απωλειών με αλγορίθμους και συστήματα μηχανικής μάθησης είναι εφικτή και μάλιστα με υποσχόμενα αποτελέσματα. Οι γραμμικοί ταξινομητές μπορούν με μεγάλη επιτυχία να εντοπίσουν με ένα έτος εκπαίδευσης αν έχει εγκατασταθεί σύστημα αλλοίωσης των μετρήσεων. Η συσταδοποίηση των κανονικοποιημένων χρονοσειρών έχει επίσης πολύ καλά αποτελέσματα στον διαχωρισμό του κυρίως πληθυσμού από τον πληθυσμό καταναλωτών με ασυνήθιστες μετρήσεις. Το γεγονός αυτό είναι και ο λόγος που η συσταδοποίηση αποτέλεσε το βασικό συστατικό του μη επιβλεπόμενου και ημι-επιβλεπόμενου συστήματος. Χρησιμοποιήθηκαν δύο ειδών κανονικοποιήσεις, και διαπιστώθηκε πως η κανονικοποίηση σε εύρος [-1,1] ταιριάζει στις χρονοσειρές, ενώ σε εύρος [0,1] σε χαρακτηριστικά διαχωρισμού με αραιούς πίνακες. Τέλος, καθίσταται σαφές πως η σύνδεση και αλληλεπίδραση διαφορετικών αλγορίθμων για τη δημιουργία μιας τελικής ταξινόμησης μπορεί να λειτουργήσει ικανοποιητικά, παρόλο που υπάρχουν περιθώρια δοκιμών και βελτιστοποιήσεων.

7.3 Μελλοντική επέκταση

Καταλήγοντας στην παρούσα διπλωματική αναλύθηκε ευρύ φάσμα αλγορίθμων μηχανικής μάθησης με επιτυχία, αλλά το ταξίδι για την βελτιστοποίηση συστημάτων μηχανικής μάθησης δεν έχει τέλος. Σε αυτό το σημείο αξίζει να αναφερθούν οι μελλοντικές επεκτάσεις της παρούσας έρευνας που πιστεύεται πως θα μπορούσαν να εξάγουν όμοια ή και καλύτερα αποτελέσματα στο πρόβλημα της ταξινόμησης. Πιο συγκεκριμένα, τα συστήματα που δημιουργήθηκαν θα μπορούσαν να συμπεριλάβουν τα εξής:

- *Εφαρμογή ταξινόμησης σε περισσότερες από δύο κλάσεις.* Όλες οι ταξινομήσεις που δοκιμάστηκαν σε αυτή τη διπλωματική έγιναν σε δύο κλάσεις, την αρνητική και τη θετική. Παρ΄ όλα αυτά, τα αποτελέσματα έδειξαν ότι υπάρχουν καταναλωτές με ακανόνιστες χρονοσειρές που δεν έχουν ομοιότητες ούτε με τις πραγματικές χρονοσειρές (αρνητική κλάση), αλλά ούτε και με τις προσομοιωμένες απάτες (θετική κλάση). Για αυτό τον λόγο, θα είχε νόημα να δημιουργηθούν και άλλες κλάσεις που να ενδεικνύουν ιδιαίτερη καταναλωτική συμπεριφορά, αλλά όχι ρευματοκλοπή. Έτσι, θα ήταν εφικτή η περαιτέρω μείωση των εσφαλμένων προβλέψεων.

- *Εξερεύνηση τεχνικών ανίχνευσης ανωμαλιών.* Στην ανίχνευση ανωμαλιών στον ημι-επιβλεπόμενο αλγόριθμο χρησιμοποιήθηκε παραμετρική τεχνική Γκαουσιανού μοντέλου, καθώς είναι η συνηθέστερη τεχνική. Παρ΄ όλα αυτά, υπάρχουν ενδείξεις από τους γραμμικούς ταξινομητές πως οι παλινδρομήσεις των χρονοσειρών οδηγούν σε αξιόλογα αποτελέσματα. Θα είχε λοιπόν νόημα να δοκιμαστεί ανίχνευση ανωμαλιών βάσει του μοντέλου παλινδρόμησης. Παράλληλα, θα είχε ενδιαφέρον η προσέγγιση του αλγορίθμου από τη μη παραμετρική σκοπιά βάσει των ιστογραμμάτων και των συναρτήσεων πυρήνων, καθώς ήδη στην παρούσα διπλωματική υπάρχουν ενδείξεις με υποσχόμενα αποτελέσματα στο κεφάλαιο 3.1.1 και 4.6.2 αντίστοιχα.

- *Ταξινόμηση βάσει πρόβλεψης μελλοντικής χρονοσειράς.* Η ύπαρξη δεδομένων με μεγαλύτερο χρονικό ορίζοντα θα καθιστούσε δυνατή την καλύτερη κατανόηση των μεταβλητών που επηρεάζουν το επίπεδο της κατανάλωσης. Με αυτό τον τρόπο, θα μπορούσε κάθε καταναλωτής να αποκτά μια πρόβλεψη της κατανάλωσής του για το επόμενο έτος με μικρή απόκλιση από την πραγματική του κατανάλωση. Στην περίπτωση που η πρόβλεψη απέκλινε σημαντικά από την καταγραφείσα κατανάλωση, ο καταναλωτής θα θεωρούνταν ύποπτος. Η προσέγγιση αυτή θεωρείται υποσχόμενη, όπως έδειξε η στατιστική μελέτη που έγινε στο κεφάλαιο 3.1.2.

Βιβλιογραφία

[1] Osama Abu Abbas. Comparisons between data clustering algorithms. pages 3–4, 2008. Yarmouk University.

[2] P. Antmann. Reducing technical and non-technical losses in the power sector. In *Transmission and Distribution Conference and Exposition*, pages 24–26. World Bank, 2009.

[3] S. Axelsson. The base-rate fallacy and its implications for the difficulty of intrusion detection. In *CCS '99 Proceedings of the 6th ACM conference on Computer and communications security*, pages 1–7. Computer and Communications Security, 1999.

[4] Jason Brownlee. A tour of machine learning algorithms, 2013. Accessed: 5 August 2017.

[5] Y. Zhu C.-Y. Hsia and Chih-Jen Lin. A study on trust region update rules in newton methods for large-scale linear classification. Technical report, JMLR, 2017.

[6] Costas Efthymiou and Georgios Kalogridis. Smart grid privacy via anonymization of smart metering data. In *Smart Grid Communications*, pages 2–4. IEEE, 2010.

[7] ERGEG. *Smart Metering with a Focus on Electricity Regulation*, 2007. E07-RMF-04-03.

[8] James J. Filliben and Alan Heckert. Nist/sematech ehandbook of statistical methods. Accessed: 25 August 2017.

[9] Commission for Energy Regulation. General information. Accessed: 24 August 2017.

[10] Gregory C. Reinsel George E. P. Box, Gwilym M. Jenkins and Greta M. Ljung. *Time Series Analysis: Forecasting and Control*. Wiley, 2016.

[11] Rob J. Hyndman and George Athanasopoulos. Forecasting: principles and practice, 2012. Accessed: 5 August 2017.

[12] Paul Johnson and Matt Beverlin. Machine learning, part ii: Supervised and unsupervised learning. Accessed: 1 September 2017.

[13] Paul Johnson and Matt Beverlin. Beta distribution, 2013.

[14] Mathworks. Parametric trend estimation, 2017. Accessed: 4 August 2017.

[15] G. Messinis and A. Dimeas. Utilizing smart meter data for electricity fraud. In *First South East European Region CIGRE Conference*, pages 2–4. CIGRE, 2014.

[16] Mkhwanazi. Electricity as a birthright and the problem of non-payment. In *Third Annual South Africa Revenue Protection Conference*, 1999.

[17] Andrew Ng. Principal components analysis. CS229 Lecture notes, 7 2014. Stanford University.

[18] Kiambang Nik. Tenaga out to short-circuit electricity thefts. 1 1999.

[19] Oracle. Data mining concepts. Accessed: 24 August 2017.

[20] J. F. G. Cobben P. Kadurek, J. Blom and W.L.Kling. Theft detection and smart metering practices and expectations in the netherlands. *Innovative Smart Grid Technologies Conference Europe*, pages 1–2, 2010. IEEE.

[21] J. F. G. Cobben P. Kadurek, J. Blom and W.L.Kling. Theft detection and smart metering practices and expectations in the netherlands. In *Innovative Smart Grid Technologies Conference Europe*, page 1. IEEE, 2010.

[22] Nasim Arianpoo Paria Jokar and Victor C. M. Leung. A practical guide to support vector classification. 7:1–3 12–16, 2003. University of Freiburg.

[23] Nasim Arianpoo Paria Jokar and Victor C. M. Leung. Electricity theft detection in ami using customers' consumption patterns. *Innovative Smart Grid Technologies Conference Europe*, 7:216–226, 2016. IEEE.

[24] Plotly. Principal component analysis in python. Accessed: 4 September 2017.

[25] Cho-Jui Hsieh Xiang-Rui Wang Rong-En Fan, Kai-Wei Chang and Chih-Jen Lin. Liblinear: A library for large linear classification. *Journal of Machine Learning Research*, 9:1871–1874, 2008.

[26] A. Naveen S. De, R. Anand and S. Moinuddin. E-metering solution for checking energy thefts and streamlining revenue collection in india. In *Transmission and Distribution Conference and Exposition*, pages 654–658. IEEE, 2003.

[27] Jon Shlens. *A Tutorial on Principal Component Analysis*. PhD thesis, Princeton University, 1993.

[28] Thomas B. Smith. Electricity theft: a comparative analysis. *Energy Policy*, 32(18):2067–2076, 2004.

[29] TACIS. *Improving Residential Electricity Services*, 1998. Tacis Technical Dissemination Project.

[30] V. Kumar V. Chandola, A. Banerjee. Anomaly detection: A survey. Technical report, ACM Computing Surveys, 2009.

[31] Simon Haykin. *Νευρωνικά Δίκτυα και Μηχανική Μάθηση*. Παπασωτηρίου, Αθήνα, 2010.

[32] ΔΕΗ. *Το Κόστος των Ρευματοκλοπών*, 5 2017. Δελτίο τύπου 552017.

[33] Αντώνης Νείρου. Ανάπτυξη μεθόδων ασαφούς συσταδοποίησης για τη μοντελοποίηση νευρωνικών δικτύων συναρτήσεων ακτινικής βάσης. παγες 54–59, 2011. Πανεπιστήμιο Αιγαίου.

[34] Θοδωρής Παναγούλης. «Εγχειρίδιο» από τη ΡΑΕ για την αντιμετώπιση των όλο και περισσότερων ρευματοκλοπών. Accessed: 6 August 2017.

[35] ΡΑΕ. *Εγχειρίδιο Ρευματοκλοπών σε εφαρμογή της παραγράφου 23 του άρθρου 95 του Κώδικα Διαχείρισης Δικτύου Διαχείρισης Διανομής Ηλεκτρικής Ενέργειας*, 5 2017. Εφημερίδα της κυβερνήσεως της Ελληνικής Δημοκρατίας.

Παράρτημα Α΄

Αναλυτικά αποτελέσματα γραμμικών ταξινομητών

Συνδυασμός	DR	FPR	Accuracy	F1 score	BDR
1	94.66	35.93	67.04	35.79	0.23
2	93.89	34.62	68.15	36.39	0.23
3	92.37	39.70	63.41	32.88	0.21
4	91.67	21.23	80.15	49.62	0.32
5	93.13	34.21	68.44	36.42	0.23
6	91.60	35.11	67.48	35.35	0.22
7	93.89	34.29	68.44	36.61	0.23
8	94.66	35.93	67.04	35.79	0.23

Πίνακας Α΄.1: Αποτελέσματα δοκιμής τύπου 1 κανονικοποίηση [-1,1]

Συνδυασμός	DR	FPR	Accuracy	F1 score	BDR
1	75.65	1.38	96.67	79.45	0.86
2	80.00	1.46	96.96	81.78	0.86
3	80.00	1.46	96.96	81.78	0.86
4	80.87	1.54	96.96	81.94	0.85
5	81.74	1.94	96.67	80.69	0.82
6	77.39	1.54	96.67	79.82	0.85
7	65.22	1.62	95.56	71.43	0.82
8	75.65	1.46	96.59	79.09	0.85

Πίνακας Α΄.2: Αποτελέσματα δοκιμής τύπου 1 κανονικοποίηση [0,1]

Συνδυασμός	DR	FPR	Accuracy	F1 score	BDR
1	4.65	0.82	90.15	8.28	0.39
2	12.40	3.77	88.22	16.75	0.27
3	10.08	3.19	88.52	14.36	0.26
4	9.30	2.87	88.74	13.64	0.26
5	13.18	4.01	88.07	17.44	0.27
6	8.53	3.44	88.15	12.09	0.22
7	0.78	0.41	90.15	1.48	0.17
8	4.65	0.82	90.15	8.28	0.39

Πίνακας Α΄.3: Αποτελέσματα δοκιμής τύπου 2 με κανονικοποίηση [0,1]

Συνδυασμός	DR	FPR	Accuracy	F1 score	BDR
1	2.05	0.83	88.67	3.77	0.21
2	9.59	2.82	87.70	14.43	0.27
3	8.22	2.66	87.70	12.63	0.25
4	8.22	2.33	88.00	12.90	0.28
5	10.27	3.16	87.48	15.08	0.26
6	8.90	2.41	88.00	13.83	0.29
7	0.68	0.50	88.81	1.31	0.13
8	2.05	0.83	88.67	3.77	0.21

Πίνακας Α΄.4: Αποτελέσματα δοκιμής τύπου 3 με κανονικοποίηση [0,1]

Συνδυασμός	DR	FPR	Accuracy	F1 score	BDR
1	1.55	0.74	89.93	2.86	0.19
2	10.85	3.03	88.74	15.56	0.28
3	10.08	3.03	88.67	14.53	0.27
4	5.43	2.70	88.52	8.28	0.18
5	13.18	3.69	88.37	17.80	0.28
6	8.53	2.87	88.67	12.57	0.25
7	0.00	0.25	90.22	NαN	0.00
8	1.55	0.66	90.00	2.88	0.21

Πίνακας Α΄.5: Αποτελέσματα δοκιμής μικτών τύπων με κανονικοποίηση [0,1]

μικρός	3	2	1
89.9300	88.6700	90.1500	96.6700
88.7400	87.7000	88.2200	96.9600
88.6700	87.7000	88.5200	96.9600
88.5200	88.0000	88.7400	96.9600
88.3700	87.4800	88.0700	96.6700
88.6700	88.0000	88.1500	96.6700
90.2200	88.8100	90.1500	96.5600
90.0000	88.6700	90.1500	96.5900

Πίνακας Α΄.6: Πίνακας Accuracy

1	2	3	μικτός
80.7800	8.2800	3.7700	2.8600
81.2300	16.7500	14.4300	15.5600
79.2500	14.3600	12.6300	14.5300
79.8500	13.6400	12.9000	8.2800
80.3100	17.4400	15.0800	17.8000
78.6300	12.0900	13.8300	12.5700
78.9100	1.4800	1.3100	0
81.2300	8.2800	3.7700	2.8800

Πίνακας Α΄.7: Πίνακας F1 score

Γλωσσάριο

Ελληνικός όρος	Αγγλικός όρος
στιβαρότητα	robustness
κινητοί μέσοι όροι	moving averages
επαναδειγματοληψία	resampling
δειγματοληψία προς τα πάνω	upsampling
δειγματοληψία προς τα κάτω	downsampling
βάση σύγκρισης	benchmark
εκθετική εξομάλυνση	exponential smoothing
γραμμές Θ	theta lines
μηχανική μάθηση	machine learning
ανάλυση συστάδων	cluster analysis
συστάδα	cluster
συσταδοποίηση	clustering
υπερπροσαρμογή	overfitting
περιηγητής	browser