Εφαρμογές της τεχνητής νοημοσύνης στην έξυπνη πόλη

Φοιτητής: Καρποντίνης Δημήτρης

Επιβλέποντες Καθηγητές:
Λουλάκης Μιχαήλ
Παπαπαντολέων Αντώνης

12 Απριλίου 2021

Περιεχόμενα

1 Εισαγωγή

1.1 Ευχαριστίες

Για την συγγραφή της διπλωματικής αυτής εργασίας καταλυτική αποτελέσε η συνεισφορά του επικεφαλή του τμήματος έρευνας και ανάπτυξης της εταιρίας Protergia κ. Νικολόπουλου Βασίλη.

Η συμμετοχή μου στο τρέχων project δεν θα ήταν εφικτή χωρίς τις συστάσεις από τους υπεύθυνους για την εργασία αυτή καθηγητές κ.Λουλάκη Μιχαήλ και κ.Παπαπαντολέων Αντώνη.

Για την διαμόρφωση των μεθόδων ανάλυσης καθώς και την επεξήγησή τους, αρμόδιος ήταν ο υπεύθυνος τμήματος πωλήσεων δημοσίου τομέα κ.Σπύρου Ιωάννης.

Επίσης σημαντική για την κατανοήση και επίλυση των τεχνικών ζητημάτων κατά την επεξεργασία δεδομένων, αποτέλεσε η συμβολή του Project Manager κ.Μπάμπη Νίκου.

Τέλος θα ήθελα να ευχαριστήσω τόσο την οικογενειά μου, όσο και τους φίλους μου για την υποστήριξη και συμπαράσταση τους τα τελευταία χρόνια, ειδικά κατά την δύσκολη αυτή περίοδο της πανδημίας.

1.2 Εισαγωγικές έννοιες

Στην εργασία αυτή, θα παρουσιαστεί η μελέτη που ολοκληρώθηκε στα πλαίσια της συνεργασίας με την εταιρία Protergia.

Το πρώτο μέρος της εργασίας θα επικεντρωθεί στην θεωρητική ανάλυση ενεργειακών και δημοσιονομικών δεδομένων με σκοπό την επίτευξη στόχων που είναι συνιφασμένη με την ανάπτυξη του πελάτη (Δήμος Τρικκαίων) προς την κατεύθυνση της έξυπνης πόλης. Οι στόχοι αυτοί θα παρουσιαστούν περιληπτικά στην εισαγωγή. Τα θεωρητικά σύνολα δεδομένων, οι προτεινόμενοι αλγόριθμοι καθώς και η ροή τους θα παρουσιαστούν στην ενότητα ″Θεωρητικά εργαλεία″.

Στη συνέχεια, στην ενότητα ″Μεθοδολογίες ανάλυσης δεδομένων″ θα επεξηγηθούν αναλυτικά τα διαθέσιμα δεδομένα, οι μεθοδολογίες ανάλυσης που επιλέχθηκαν, αλλά και η απαραίτητη προεπεξεργασία για την υλοποίησή τους.

Τέλος στην ενότητα ″Τελικά αποτελέσματα″, θα παρατεθούν και θα ερμηνευτούν τα τελικά αποτελέσματα των αλγορίθμων ενώ θα προταθούν πιθανές επεκτάσεις, με σκοπό την βελτίωση των μεθόδων. Τα τελικά αποτελέσματα θα παρουσιαστούν στο Δήμο, με στόχο την ανακάλυψη ευρημάτων και την αντιμετώπιση προβλημάτων που είναι συνυφασμένα με την διαμόρφωση μιας έξυπνης πόλης.

Στη συνέχεια παρουσιάζονται οι στόχοι του δήμου για την επίτευξη των οποίων υλοποιείται η ανάλυση των θεωρητικών δεδομένων.

- Έξυπνος φωτισμός:
  Με το σύστημα έξυπνου φωτισμού, ο δήμος αποζητά να ελέγξει και να βελτιώσει την ποιότητα του δημοτικού ηλεκτροφωτισμού και να επιτύχει εξοικονόμηση ενέργειας μέσω της αντικατάστασης προϋπάρχοντων φωτιστικών συστημάτων με συστήματα LED. Επίσης, ενσωματώνει αισθητήρες κίνησης στον παρόδιο ηλεκτροφωτισμό, ώστε να αυξομειώνεται η ένταση του φωτισμού (dimming) σε συνάρτηση με την κυκλοφορία οχημάτων και πολιτών και επομένως να επιτευχθεί περαιτέρω εξοικονόμηση.

- Έξυπνη στάθμευση:
  Το σύστημα έξυπνης στάθμευσης στοχεύει στην εύρεση, την απεικόνιση και τον έλεγχο οριοθετημένων θέσεων στάθμευσης στο κέντρο της πόλης. Το πρόγραμμα θα ενημερώνει τους πολίτες σε πραγματικό χρόνο για τη διαθεσιμότητα θέσεων με εφαρμογή για κινητά τηλέφωνα, αλλά και από πληροφοριακές πινακίδες σε καίρια σημεία της πόλης.
  Επίσης, θα παρέχεται στη δημοτική αστυνομία, ενημέρωση για περιπτώσεις παράνομης στάθμευσης.

- Σύστημα παρακολούθησης περιβαλοντικών συνθηκών:
  Με χρήση περιβαλλοντικών αισθητήρων θα εκτιμάται η ποιότητα της ατμόσφαιρας και θα αξιολογείται ο πιθανός αντίκτυπος στη δημόσια υγεία.

- Σύστημα παρακολούθησης κάδων:
  Το σύστημα αυτό στοχεύει στην καλύτερη διαχείριση των απορριματοφόρων οχημάτων και την αποφυγή των φαινομένων υπερχύλισης των κάδων ανακύκλωσης.
  Για τον λόγο αυτό θα χρησιμοποιούνται αισθητήρες μέτρησης του όγκου των απορριμμάτων σε κάδους ανακύκλωσης και θα εκτιμάται η ταχύτητα πλήρωσης αυτών ανά περιοχή στην πόλη.

- Σύστημα παρακολούθησης υδρομέτρων:
  Το σύστημα αυτό αποζητά την μελέτη της ποιότητας των υδρομέτρων στην πόλη κάθως και την αντιμετώπιση προβλημάτων που συσχετίζονται με αυτή. Για τον λόγο αυτό χρησιμοποιούνται αισθητήρες για τον εξ αποστάσεως έλεγχο κατανάλωσης και πιθανών διαρροών (τηλεμετρία).

- Σύστημα παρακολούθησης καταναλώσεων ενέργειας:
  Με ανάλυση των δεδομένων δημοτικής κατανάλωσης θα εντοπίζονται και θα αναβαθμίζονται ενεργοβόρα κτήρια και φωτιστικά σώματα με σκοπό την ενεργειακή επάρκεια και απόδοση του δήμου.
  Ταυτόχρονα το σύστημα θα ελέγχει για πιθανές ζημιές και κλοπές, ενώ θα προτείνει πιθανές δράσης για την αντιμετώπιση αυτών.

- Καλλιέργεια ακρίβειας:
  Ο Δήμος εγκαθιστά αιθητήρες σε μικρές αγροτικές εκτάσεις, με στόχο την αυτόματη φροντίδα και τηλεπαρακολούθησή τους.

Στη συνέχεια, εξηγείται το περιεχόμενο των τριών βασικών υποενοτήτων της επόμενης ενότητας, η χρησιμότητα τους, η σύνδεση μεταξύ τους καθώς και τα εργαλεία που χρησιμοποιήθηκαν για την διεκπεραίωσή τους.

- Σημασιολογικό διάγραμμα:
  Πριν την οποιαδήποτε ανάλυση των δεδομένων, σημαντική αποτελεί η κατανοήση των συσχετίσεων μεταξύ των συνόλων που θα μας δωθούν. Κάτι τέτοιο μας δίνει μια αρχική εικόνα για τον τρόπο αξιοποίησης των δεδομένων μας και επομένως τους πιθανούς αλγορίθμους που θα μας προσφέρουν μια ποιοτική ανάλυση, εκμεταλευόμενοι τις μεταξύ τους συσχετίσεις.

  Η παραπάνω διεργασία θα γίνει με τη χρήση ενός γράφου,ο οποίος θα κατασκευαστεί με την χρήση του προγράμματος gephi και θα αναπαραστίσει το σημασιολογικό μας διάγραμμα[2].

- Διάγραμμα αξίας αλγορίθμων:
  Έχοντας πλέον αποκτήσει μια πρώτη εικόνα των συσχετίσεων των δεδομένων μας, ερχόμαστε τώρα στην πρόταση των αλγορίθμων με βάση τους οποίους θα υλοποιηθεί η επιθυμητή ανάλυση. Όπως αναφέρθηκε και πριν το σημασιολογικό διάγραμμα μας παρέχει ένα σύνολο πιθανών αλγορίθμων, μέσω των συσχετίσεων που παρουσιάζονται μεταξύ των συνόλων.

  Παρόλα αυτά έννοιες όπως το υπολογιστικό κόστος, ο απαιτούμενος αριθμός δεδομένων για την υλοποιήση και ο συνολικός χρόνος που μας έχει δωθεί μέχρι την ολοκλήρωση του Project πρέπει επίσης να ληφθούν σοβαρά υπόψη πρωτού ένας αλγόριθμος ενταχθεί στο διάγραμμα.

  Τέλος, αφότου όλα τα παραπάνω έχουν ληφθεί υπόψη, καταγράφουμε λεπτομερώς την αξία που θα μας παρέχει η υλοποίηση των τελικά επιλεγόμενων αλγορίθμων. Η αξία αυτή, πηγάζει από το τρόπο χρήσης των αποτελεσμάτων των αλγορίθμων, με σκοπό την επίτευξη των στόχων που έχουν αναφερθεί παραπάνω(usability).

- Διάγραμμα αλγοριθμικής ροής:
  Αφού στο σημείο αυτό έχει καθοριστεί το σύνολο των αλγορίθμων που θα χρησιμοποιηθούν, ερχόμαστε τώρα στην επεξήγησή τους με χρήση διαγράμματος.

  Συγκεκριμένα κάθε αλγόριθμος που αποφασίστηκε να χρησιμοποιηθεί, θα παρουσιάζεται σε βήματα καθένα από τα οποία θα καταγράφεται εντός ενός σχήματος. Τα βήματα της αρχής και του τέλους θα εμφανίζονται στο εσωτερικό κύκλων. Τα βήματα που απαιτούν δεδομένα από τον χρήστη θα εμφανίζονται στο εσωτερικό παραλληλογράμων.

  Τα βήματα όπου γίνεται έλεγχος μιας λογικής έκφρασης θα εμφανίζονται στο εσωτερικό ρόμβων. Τέλος τα βήματα που εκφράζουν την εκτέλεση κάποιας διαδικασίας θα εμφανίζονται στο εσωτερικό ορθογωνίων παραλληλογράμων. Η χρησιμότητα της παρουσίασης του αλγορίθμου με αυτό τον τρόπο πηγάζει από την απλότητα εξήγησης που ένα τέτοιο διάγραμμα παρέχει.

2 Θεωρητικά εργαλεία

2.1 Σημασιολογικό διάγραμμα

Όπως αναφέρθηκε και πριν,καταλυτικό βήμα για την ποιοτική ανάλυση των δεδομένων που θα μας δωθούν είναι η εύρεση των συσχετίσεων μεταξύ τους. Κάτι τέτοιο θα επιτευχθεί με την χρήση του σημασιολογικού διαγράμματος.
Το σημασιολογικό διάγραμμα αποτελεί έναν γράφο οι κόμβοι του οποίου χωρίζονται σε δύο είδη.

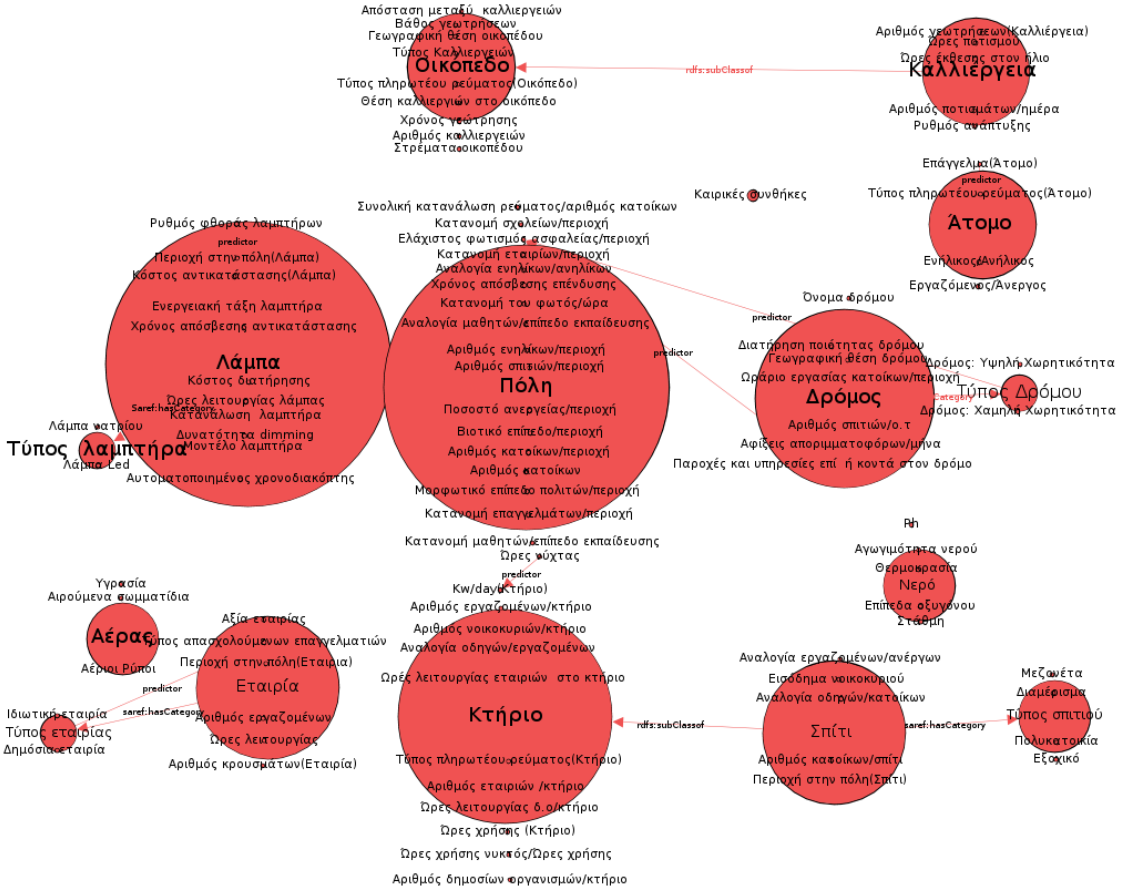
Το πρώτο εκφράζει το σύνολο των δεδομένων που θα μας δωθεί, ενώ το δεύτερο τις γενικευμένες οντολογίες στις οποίες αυτά ανήκουν. Κάθε σύνολο περιγράφεται από έναν κόμβο χρώματος πράσινου. Μέσω των ακμών που συνδέουν τα σύνολα, εκφράζονται οι συσχέτισεις που παρουσιάζονται μεταξύ τους, με την κάθε ακμή να αναγράφει το είδος της αντίστοιχης συσχέτισης.
Παρακάτω δίνεται το αναφερόμενο διάγραμμα:

Καθένα από τα σύνολα δεδομένων ανήκει εννοιολογικά σε μια ευρύτερη κλάση ή οντολογία. Κάθε οντολογία περιγράφεται με έναν κόμβο χρώματος κόκκινου. Ακόμα κάθε οντολογία περιέχει ένα σύνολο από χαρακτηριστικά γνωρίσματα (features) που μπορεί να μας φανούν χρήσιμα στην περαιτέρω ανάλυση των δεδομένων μας [1].

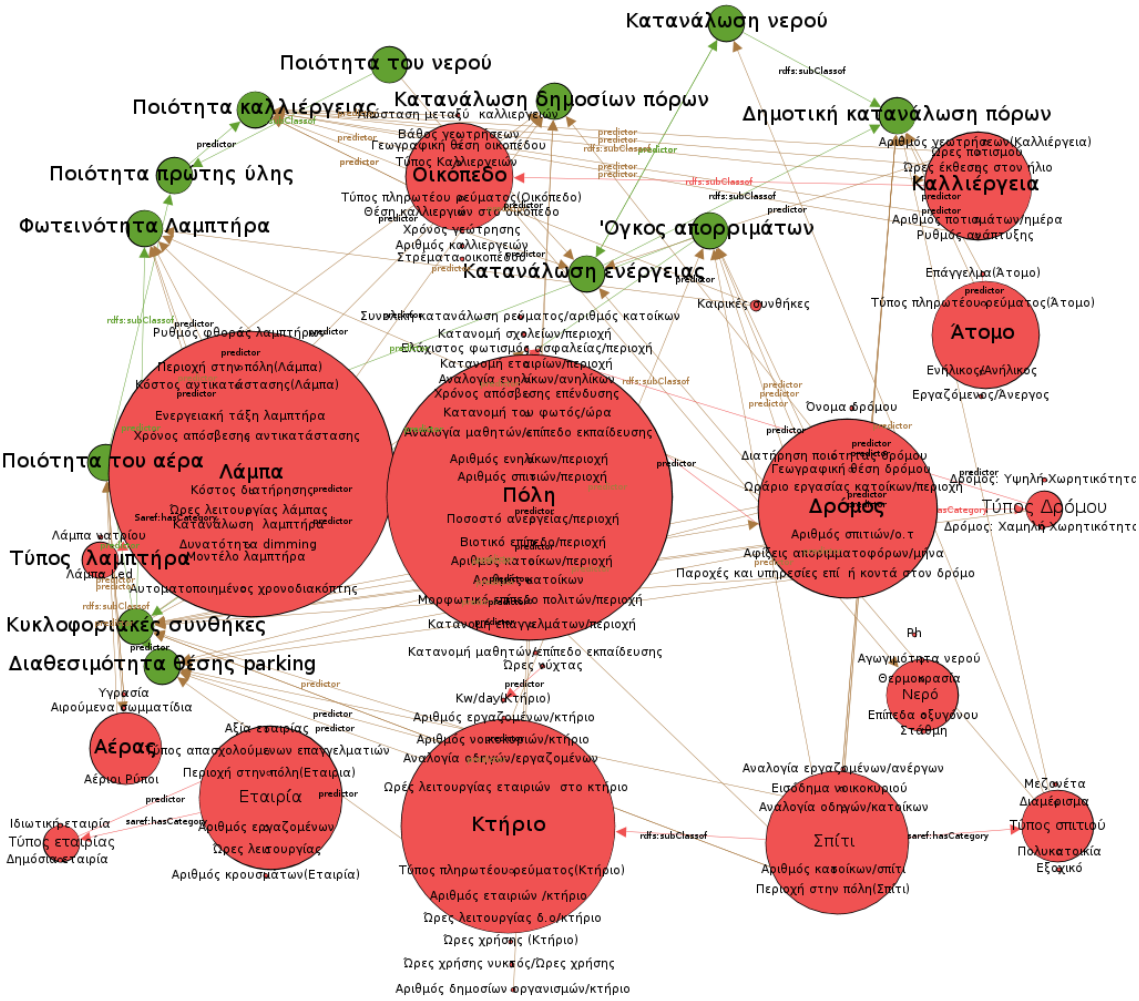
Παρότι δεν έχουμε καμία βεβαιότητα πως κάποιο από τα γνωρίσματα θα μας δωθεί στην πράξη, η προσθήκη τους κατά την διαμόρφωση του σημασιολογικού διαγράμματος είναι σημαντική. Ο λόγος για το παραπάνω είναι, πως παρότι τα ίδια τα χαρακτηριστικά μπορεί να μην μας δωθούν, η προσθήκη τους στο διάγραμμα μπορεί να εμφανίσει ″κρυφές″ συσχετίσεις μεταξύ των δεδομένων μας.

Με άλλα λόγια, η χρήση των οντολογίων και των χαρακτηριστικών τους, μπορεί να εμπλουτίσει το σύνολο των δεδομένων μας με περαιτέρω νόημα ακόμα και αν οι πληροφορία που παρέχουν, δεν μας δωθεί από τον Δήμο. Το νόημα αυτό παρουσιάζεται στο διάγραμμα με την χρήση ακμών που συνδέουν τις οντολογίες και τα χαρακτηριστικά τους με τα σύνολα δεδομένων. Ακόμα συσχετίσεις διαμορφώνονται και μεταξύ των οντολογιών, οι οποίες επίσης παρουσιάζονται μέσω ακμών.

Το δεύτερο αυτό είδος κόμβων μαζί με τις συσχετίσεις τους, παρουσιάζονται στο παρακάτω διάγραμμα:


Απόσταση μεταξύ καλλιεργειών
Βάθος γεωτρήσεων
Γεωγραφική θέση οικοπέδου
Τύπος Καλλιεργειών
Οικόπεδο
Τύπος πληρωτέου ρεύματος(Οικόπεδο)
Θέση καλλιεργιών στο οικόπεδο
Χρόνος γεώτρησης
Αριθμός καλλιεργειών
Στρέμματα οικοπέδου
rdfs:subClassof
Αριθμός γεωτρήσεων(Καλλιέργεια)
Ώρες πατισμού
Καλλιέργεια
Ώρες έκθεσης στον ήλιο
Αριθμός ποτισμάτων/ημέρα
Ρυθμός ανάπτυξης
Επάγγελμα(Άτομο)
predictor
Τύπος πληρωτέου ρεύματος(Άτομο)
Άτομο
Ενήλικος/Ανήλικος
Εργαζόμενος/Άνεργος
Καιρικές συνθήκες
Συνολική κατανάλωση ρεύματος/αριθμός κατοίκων
Κατανομή σχολείων/περιοχή
Ελάχιστος φωτισμός ασφαλείας/περιοχή
Κατανομή εταιρίων/περιοχή
Αναλογία ενηλίκων/ανηλίκων
Χρόνος απόσβεσης επένδυσης
Κατανομή του φωτός/ώρα
Αναλογία μαθητών/επίπεδο εκπαίδευσης
Ρυθμός φθοράς λαμπτήρων
predictor
Περιοχή στην πόλη(Λάμπα)
Κόστος αντικατάστασης(Λάμπα)
Ενεργειακή τάξη λαμπτήρα
Χρόνος απόσβεσης αντικατάστασης
Λάμπα
Κόστος διατήρησης
Ώρες λειτουργίας λάμπας
Saref:hasCategory
Κατανάλωση λαμπτήρα
Λάμπα κατ/ίου
Δυνατότητα dimming
Τύπος λαμπτήρα
Μοντέλο λαμπτήρα
Λάμπα Led
Αυτοματοποιημένος χρονοδιακόπτης
Αριθμός ενηλίκων/περιοχή
Αριθμός σπιτιών/περιοχή
Πόλη
Ποσοστό ανεργείας/περιοχή
Βιοτικό επίπεδο/περιοχή
Αριθμός κατοίκων/περιοχή
Αριθμός κατοίκων
Μαρφωτικό επίπεδο πολιτών/περιοχή
Κατανομή επαγγελμάτων/περιοχή
predictor
Όνομα δρόμου
predictor
Διατήρηση ποιότητας δρόμου
Γεωγραφική θέση δρόμου
Ωράριο εργασίας κατοίκων/περιοχή
Δρόμος
Αριθμός σπιτιών/ο.τ
Αφίξεις απορριμματοφόρων/μήνα
Παροχές και υπηρεσίες επί ή κοντά στον δρόμο
Τύπος Δρόμου
Δρόμος: Υψηλή Χωρητικότητα
Δρόμος: Χαμηλή Χωρητικότητα
Κατανομή μαθητών/επίπεδο εκπαίδευσης
Ώρες νύχτας
predictor
Kw/day(κτήριο)
Αριθμός εργαζομένων/κτήριο
Αριθμός νοικοκυριών/κτήριο
Αναλογία οδηγών/εργαζομένων
Ώρες λειτουργίας εταιριών στο κτήριο
Κτήριο
Τύπος πληρωτέου ρεύματος(Κτήριο)
Αριθμός εταιριών /κτήριο
Ώρες λειτουργίας δ.ο/κτήριο
Ώρες χρήσης (Κτήριο)
Ώρες χρήσης νυκτός/Ώρες χρήσης
Αριθμός δημοσίων οργανισμών/κτήριο
Ph
Αγωγιμότητα νερού
Θερμοκρασία
Νερό
Επίπεδα οξυγόνου
Σταθμή
Υγρασία
Αιρούμενα σωματίδια
Αέρας
Αέριοι Ρύποι
Αξία εταιρίας
Τύπος απασχολούμενων επαγγελματιών
Περιοχή στην πόλη(Εταιρία)
Εταιρία
predictor
Ιδιωτική εταιρία
saref:hasCategory
Τύπος εταιρίας
Δημόσια εταιρία
Αριθμός εργαζομένων
Ώρες λειτουργίας
Αριθμός κρουσμάτων(Εταιρία)
Αναλογία εργαζομένων/ανέργων
Εισόδημα νοικοκυριού
Αναλογία οδηγών/κατοίκων
Σπίτι
rdfs:subClassof
Αριθμός κατοίκων/σπίτι
Περιοχή στην πόλη(Σπίτι)
saref:hasCategory
Μεζονέτα
Διαμέρισμα
Τύπος σπιτιού
Πολυκατοικία
Εξοχικό

Έχοντας πλέον αναλύσει τις δύο κατηγορίες κόμβων καθώς και τις σχέσεις που αναπτύσσονται μεταξύ τους, παραθέτουμε το τελικό διάγραμμα:



2.2 Διάγραμμα αξίας αλγορίθμων

Οι σημασιολογικές συσχετίσεις μεταξύ των συνόλων δεδομένων και οντολογιών που εμφανίζονται παραπάνω, θα ελεγχθούν με την εφαρμογή χρήσιμων για την περαιτέρω ανάλυση αλγορίθμων. Παρακάτω παρουσιάζουμε ένα τέτοιο σύνολο αλγορίθμων:

Πίνακας 1: Χρήσιμοι αλγόριθμοι

Όνομα αλγορίθμου
Regression algorithms
Linear regression
Polynomial regression
Quantile regression
Logistic regression
ARMA
Classification algorithms
Desicion Trees
Random Forest
Clustering algorithms
K-Means
Hierarchical Clustering
Spectral Clustering
Other algorithms
PCA
ANOVA
Factor Analysis
Corellation
Simplex

(α΄) Αλγόριθμοι παλινδρόμησης

(i) Γραμμική παλινδρόμηση:
Ο αλγόριθμος αυτός αποτελεί έναν από τους πιο στοιχειώδεις αλγόριθμους που μπορεί κανείς να χρησιμοποιήσει στα πλαίσια της μελέτης ενός συνόλου δεδομένων.
Οι αυστηρές υποθέσεις του για την γραμμική σχέση μεταξύ των επεξηγηματικών μεταβλητών (predictors) και της μεταβλητής απόκρισης (response) καθιστά συχνά την εκτίμηση της απόκρισης από το μοντέλο μη ικανοποιητική.

Για τον λόγο αυτό σε περίπτωση που οι υποθέσεις του μοντέλου φαίνεται να μην ικανοποιούνται χρησιμοποιούμε το μοντέλο της πολυωνυμικής παλινδρόμησης.
Παρόλα αυτά η εφαρμόγη του παραμένει ιδιαίτερα σημαντική, καθώς μας προσφέρει μια αρχική εκτίμηση των στατιστικά σημαντικών επεξηγηματικών μεταβλητών για την μεταβλητή απόκρισης.

Η σχέση της μεταβλητής απόκρισης με τις επεξηγηματικές μεταβλητές για τον συγκεκριμένο αλγόριθμο δίνεται ως:

$$y = X\beta + \epsilon$$

Όπου:

- $y \in R^{n\times 1},\ n\ \in\ \mathbb{N}$
- $\beta \in R^{k\times 1},\ k \in \mathbb{N}$
- $X \in R^{n\times p},\ p = k + 1$
- $\sigma \in \mathbb{R}$
- $\epsilon \sim N(0, \sigma^2\ I_n)$

Με την εφαρμογή της μεθόδου ελαχίστων τετραγώνων το β παίρνει την μορφή:

$$\beta = (X'\ X)^{-1}\ (X'\ y)$$

(ii) Πολυωνυμική παλινδρόμηση:
Όμοια με την γραμμική παλινδρόμηση, ο παραπάνω αλγόριθμος προσαρμόζεται με την μέθοδο των ελαχίστων τετραγώνων. Αντίθετα όμως από αυτή, η υποθέση της γραμμικότητας μεταξύ απόκρισης και επεξηγηματικών μεταβλητών καθώς και η υπόθεση της μη αλληλεπίδρασης των επεξηγηματικών μεταβλητών μεταξύ τους δεν ισχύουν. Επομένως η πολυωνυμική παλινδρόμηση αποτελεί μια ιδιαίτερα ευέλικτη μέθοδο, ικανή να προσαρμοστεί ικανοποιητικά σε ένα μεγάλο αριθμό συνόλων δεδομένων.

Στα πλαίσια λοιπόν της μελέτης μας ο παραπάνω αλγόριθμος θα χρησιμοποιηθεί τόσο για την εύρεση των στατιστικά σημαντικών επεξηγηματικών μεταβλητών, όσο και για μια πρώτη εκτίμηση της τιμής της απόκρισης, για δεδομένες τίμες των επεξηγηματικών μεταβλητών.

Η σχέση της μεταβλητής απόκρισης με τις επεξηγηματικές μεταβλητές για τον συγκεκριμένο αλγόριθμο δίνεται ως:

$$y = \beta_0 + \sum_{l=1}^{m} \sum_{i=1}^{n} (\beta_i \ x_i)^l + \epsilon$$

Όπου:

- $\epsilon \sim N(0, \sigma^2)$
- $n, m \ \in \mathbb{N}$
- $\sigma \in \mathbb{R}$
- $y, x_i \in \mathbb{R} \ \forall \ i \ \in \ \{1, 2, \ldots, n\}$

(iii) Ποσοστιαία παλινδρόμηση:
Όπως είναι γνωστό, για δεδομένες τιμές των επεξηγηματικών μεταβλητών η απόκριση ακολούθει μια άγνωστη σε εμάς κατανομή. Οι δύο προηγούμενες μέθοδοι μας παρέχουν μια εκτίμηση για την μέση τιμή της εν λόγω κατανομής. Αντίθετα, ο αλγόριθμος της ποσοστιαίας παλινδρόμησης εκτιμά κάποιο προκαθορισμένο ποσοστιαίο σημείο της κατανομής αυτής.

Η χρησιμότητα της μεθόδου αυτής έγκειται στην εύρεση ακραίων τιμών στο σύνολο των δεδομένων μας μέσω της σύγκρισής τους με τα ποσοστιαία σημεία της κατανομής. Κάτι τέτοιο μας επιτρέπει να εντοπίσουμε "αφύσικες" τιμές στην κατανάλωση ενέργειας, η περαιτέρω μελέτη των οποίων μπορεί να οδηγήσει σε σημαντικά συμπεράσματα.

Η εκτίμηση του τ-ποσοστιαίου σημείου $(Q_\tau(y))$, για δεδομένες τιμές των επεξηγηματικών μεταβλητών $X_1 = x_1, X_2 = x_2, \ldots, X_p = x_p, \ p \in \mathbb{N}$, δίνεται από το τύπο:

$$Q_\tau(y) = \hat{\beta_0}(\tau) + \hat{\beta_1}(\tau) \ x_1 + \cdots + \hat{\beta_p}(\tau) \ x_p$$

Με τους συντελεστές $\hat{\beta_0}, \hat{\beta_1}, \ldots, \hat{\beta_p}$ να προκύπτουν από την ελαχιστοποίηση της συνάρτησης Median Absolute Deviation (M.A.D):

$$M.A.D(\beta_0, \beta_1, \ldots, \beta_p) = \frac{1}{n} \sum_{i=1}^{n} \rho_\tau(y - \beta_0 - \beta_1 \ x_1 - \cdots - \beta_p \ x_p)$$

Όπου:

$$\rho_\tau(u) = \tau \max(u, 0) + (1 - \tau) \max(-u, 0)$$

(iv) Λογιστική παλινδρόμηση:
Στις μεθόδους που αναφέραμε μέχρι τώρα χρησιμοποιούσαμε ένα σύνολο επεξηγηματικών μεταβλητών με σκόπο την εκτίμηση μίας ποσοτικής απόκρισης.
Πολλές φορές όμως ερχόμαστε αντιμέτωποι με προβλήματα εκτίμησης μιας ποιοτικής απόκρισης. Στην ειδική περίπτωση όπου η απόκριση αυτή έχει δύο μονάχα κατηγορίες, χρήσιμος αποτελεί ο αλγόριθμος της λογιστικής παλινδρόμησης.

Η λογιστική παλινδρόμηση εκτιμά την πιθανότητα μια ποιοτική μεταβλητή να ανήκει σε μια κατηγορία για δεδομένες τιμές των επεξηγηματικών μεταβλητών.
Δηλαδή εκτιμά την πιθανότητα:

$$P[Y = k | X = x], \ k = 0, 1 \ x \in \mathbb{R}$$

Όπου X είναι το διάνυσμα των επεξηγηματικών μεταβλητών και Y η μεταβλητή απόκρισης.

Στη συνέχεια για την δεδομένη τιμή του X κατηγοριοποιούμε την Y στην κλάση με την μεγαλύτερη εκτιμούμενη πιθανότητα.

Με βάση αυτή τη μέθοδο, η πιθανότητα κατηγοριοποίησης μιας παρατήρησης με τιμή $x \in \mathbb{R}^k,\ k \in \mathbb{N}$ στην κατηγορία 1 ($P[Y = 1|X = x] = p_x$) είναι:

$$p_x = \frac{e^{\beta_0+\beta_1 x_1+\cdots+\beta_k x_k}}{1 + e^{\beta_0+\beta_1 x_1+\cdots+\beta_k x_k}}$$

Με τους συντελεστές $\beta_i\ i \in \{0, 1, \ldots, k\}$ να εκτιμόνται μέσω της μεθόδου ελαχίστων τετραγώνων.

Ο αλγόριθμος αυτός θα χρησιμοποιηθεί για την κατηγοριοποιήση των καταναλωτών ως προς τις καταναλωτικές τους συνήθειες. Κάτι τέτοιο, μας επιτρέπει να δημιουργούμε στοχευμένα προϊόντα, λαμβάνοντας υπόψη τις συνήθειες και ανάγκες των καταναλωτών.

(v) A.R.M.A

Πολλές φορές στα πλαίσια της πρόβλεψης της μεταβλητής απόκρισης, σημαντικός παράγοντας αποτελεί η τιμή της μεταβλητής σε παρελθοντικούς χρόνους. Στις περιπτώσεις αυτές χρήσιμο είναι το μοντέλο A.R (Autoregressive model).
Ο τύπος του μοντέλου τάξης p (A.R(p)) είναι:

$$Y_t = c + \sum_{i=1}^{p} \varphi_i Y_{t-i} + \epsilon_t$$

Όπου:

- Y_t η απόκριση στο χρόνο t
- c σταθερά
- $p \in \mathbb{N}$ το πλήθος των παρελθοντικών τιμών της απόκρισης που λαμβάνουμε υπόψη.
- $\varphi_i\ \forall\ i\ \in 1, 2, \ldots, p$ παραμετροι του μοντέλου
- ϵ_t λευκός ήχος στο χρόνο t

Από την άλλη, αν παρατηρήσουμε μεταβολή στον μέσο όρο της απόκρισης σαν συνάρτηση του χρόνου, τότε συνιστάται η εφαρμογή της μεθόδου M.A (Moving Average).

Ο τύπος του μοντέλου τάξης q (M.A(q)) είναι:

$$Y_t = \mu + \epsilon_t + \sum_{j=1}^{q} \theta_j \epsilon_{t-j}$$

Όπου:

- Y_t η απόκριση στο χρόνο t
- μ η μέση τιμής της απόκρισης στο χρόνο t
- $q \in \mathbb{N}$ το πλήθος των παρελθοντικών τιμών του λευκού ήχου που λαμβάνουμε υπόψη.
- $\theta_j\ \forall\ j\ \in 1, 2, \ldots, q$ παραμετροι του μοντέλου

- ϵ_t λευκός ήχος στο χρόνο t

Έτσι το τελικό μοντέλου A.R.M.A(p,q) δίνεται από τον τύπο:

$$Y_t = c + \sum_{i=1}^{p} \varphi_i Y_{t-i} + \epsilon_t + \mu + \epsilon_t + \sum_{j=1}^{q} \theta_j \epsilon_{t-j}$$

Οι παράμετροι του τελικού μοντέλου υπολογίζονται, μέσω της επαναληπτικής μεθόδου Box-Jenkings.

Η χρησιμότητα αυτής της μεθόδου έγκειται στην δυνάτοτητα πρόβλεψης της κατανάλωσης ενέργειας, δεδόμενης της σταθερότητας εξωτερικών παραγόντων κατά την περίοδο αυτής.

Στη συνέχεια παρουσιάζουμε πίνακα χρησιμότητας των αλγορίθμων παλινδρόμησης που παρουσιάστηκαν παραπάνω:

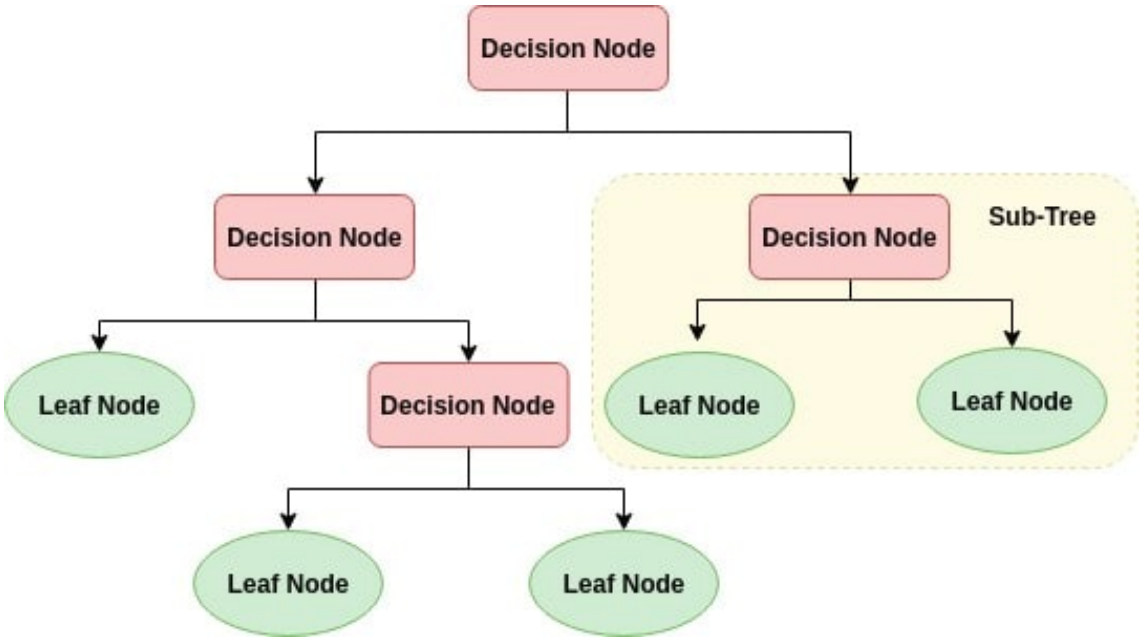
Πίνακας 2: Χρησιμότητα αλγορίθμων παλινδρόμησης

Όνομα Αλγορίθμου	Χρησιμότητα
Γραμμική παλινδρόμηση	Εύρεση στατιστικά σημαντικών παραγόντων για την κατανάλωση ενέργειας
Πολυωνιμική παλινδρόμιση	Εύρεση σημαντικών παραγόντων και πρόβλεψη της κατανάλωσης
Ποσοστιαία παλιδρόμηση	Εύρεση ακραίων παρατηρήσεων κατανάλωσης
Λογιστική παλινδρόμηση	Πρόβλεψη καταναλωτικών συμπεριφορών και δημιουργία νέων προϊόντων.
A.R.M.A	Πρόβλεψη της κατανάλωσης

(β΄) Αλγόριθμοι κατηγοριοποίησης:
Είδαμε προηγουμένως το πρόβλημα της πρόβλεψης μιας ποιοτικής απόκρισης ή το πρόβλημα της κατηγοριοποίησης (Classification problem). Στη συνέχεια παραθέτουμε δύο από τους σημαντικότερους αλγορίθμους που χρησιμοποιούνται σε τέτοια προβλήματα.

(i) Δέντρο απόφασης:
Ένα δέντρο απόφασης απαρτίζεται από κλαδία και φύλλα. Στα προβλήματα κατηγοριοποιήσης τα κλαδιά συμβολίζουν τις συνθήκες διαχωρισμού των δεδομένων, ενώ τα φύλλα συμβολίζουν υποσύνολα των δεδομένων, τα οποία και έχουν διαμορφωθεί από τους διαχωρισμούς, με μοναδική για το κάθε φύλλο προβλεπόμενη απόκριση.
Οι κόμβοι που δεν αποτελούν φύλλα(εσωτερικοί κόμβοι ή κόμβοι απόφασης) συμβολίζουν υποσύνολα, στα οποία απαιτείται περαιτέρω διαχωρισμός, με σκοπό την εύστοχη πρόβλεψη της απόκρισης από το μοντέλο.

Παραθέτουμε παρακάτω ένα παράδειγμα δέντρου απόφασης:



Με τον τρόπο αυτό διαχωρίζουμε το σύνολο των δεδομένων μας σε υποσύνολα μέχρις ότου όλα τα δεδομένα που βρίσκονται εντός ενός υποσυνολου να διαθέτουν την ίδια τιμή απόκρισης ή μέχρι ο περαιτέρω διαχωρισμός να μην επιφέρει σημαντικό κέρδος στην ικανότητα πρόβλεψης του μοντέλου.

Η χρησιμότητα αυτού του αλγορίθμου έγκειται στην κατηγοριοποιήση των καταναλωτών ως προς τις καταναλωτικές τους συνήθειες. Κάτι τέτοιο, μας επιτρέπει να δημιουργούμε εύστοχα προϊόντα, λαμβάνοντας υπόψη τις συνήθειες και ανάγκες των καταναλωτών.

(ii) Τυχαίο δάσος:
Η μέθοδος αυτή χρησιμοποιεί ένα σύνολο από δέντρα απόφασης το καθένα από τα οποία εκπαιδεύεται ανεξάρτητα των υπολοίπων(″παράλληλη εκμάθηση″), με το μοντέλο να λαμβάνει υπόψη την έξοδο του κάθε δέντρου κατά την τελική εκτίμηση. Η παραπάνω διαδικασία εκμάθησης ονομάζεται bagging(bootstrap aggregation), ενώ το σύνολο των δέντρων απόφασης χαρακτηρίζονται διαφορετικά ως weak learners.

Ο αλγόριθμος αρχικά, διαχωρίζει το αρχικό σύνολο των δεδομένων, καθώς και το σύνολο των επεξηγηματικών μεταβλητών σε ισοπληθή υποσύνολα. Με στόχο την εκπαίδευση των δέντρων, δίνεται ως όρισμα στο καθένα, διαφορετικό υποσύνολο παρατηρήσεων και επεξηγηματικών μεταβλητών. Με τον τρόπο αυτό, λαμβάνουμε πλήθος διαφορετικών ασθενών μαθητών(weak learners), που καλούνται να κατηγοριοποιήσουν τα δεδομένα.

Τέλος, μια παρατήρηση κατηγοριοποιείται στην κλάση στην οποία η πλειοψηφία των δέντρων την έχει τοποθετήσει. Η μέθοδος αυτή είναι συχνά πιο αποτελεσματική από την προηγούμενη όσον αφορά την ικανότητα σωστής κατηγοριοποίησης των παρατηρήσεων. Όπως και προηγουμένως, ο αλγόριθμος θα χρησιμοποιείται με σκοπό την κατηγοριοποιήση των καταναλωτών ως προς τις καταναλωτικές τους συνήθειες.

Στο σημείο αυτό παραθέτουμε τον πίνακα χρησιμότητας για τους αλγόριθμους κατηγοριοποίησης:

Πίνακας 3: Χρησιμότητα αλγορίθμων κατηγοριοποίησης

Όνομα αλγορίθμου	Χρησιμότητα
Δέντρο απόφασης	Πρόβλεψη καταναλωτικών συμπεριφορών και δημιουργία νέων προϊόντων
Τυχαίο δάσος	Πρόβλεψη καταναλωτικών συμπεριφορών και δημιουργία νέων προϊόντων

(γ′) Αλγόριθμοι συσταδοποίησης:
Μέχρι στιγμής κάθε αλγόριθμος που έχουμε μελετήσει λάμβανε δύο είδη μεταβλητών, τις επεξηγηματικές και την μεταβλητή απόκρισης. Υπάρχουν όμως και αλγόριθμοι, η είσοδος των οποίων δεν διαχωρίζεται σε κατηγορίες. Μέθοδοι αυτής της μορφής ονομάζονται unsupervised learning methods μιας και η εκπαίδευση δεν ″επιβλέπεται″ από κάποια μεταβλητή απόκρισης.

Μια από τις σημαντικότερες κατηγορίες unsupervised μεθόδων αποτελούν οι μέθοδοι συσταδοποίησης, σκοπός των οποίων είναι η εύρεση προτύπων στα δεδομένα αλλά και η κατηγοριοποίση τους.

Στη συνέχεια αναφέρουμε τους σημαντικότερους αλγορίθμους αυτής της κατηγορίας καθώς και την χρησιμότητά τους στην περίπτωσή μας.

(i) KMeans:
Όπως αναφέραμε και προηγουμένως σκοπός των μεθόδων συσταδοποίησης είναι η εύρεση προτύπων στα δεδομένα. Ο αλγόριθμος KMeans επιτυγχάνει το παραπάνω με την ομαδοποίηση των δεδομένων σε έναν προκαθορισμένο αριθμό κλάσεων.

Ο τρόπος ταξινόμησης των δεδομένων καθορίζεται από την απόσταση των κεντροειδών των ομάδων από αυτά. Η αρχική επιλογή των κεντροειδών διαφέρει ανάλογα με το πρόβλημα. Έχοντας πλέον καθορίσει τις ομάδες στις οποίες χωρίζεται το σύνολο των δεδομένων μας,

μπορούμε να ελέγξουμε για την αντιστοχία των προτύπων των δεδομένων στο εσωτερικό των ομάδων (pattern matching).

Η χρησιμότητα του αλγορίθμου στην περίπτωση μας πηγάζει από την εύρεση ασυνήθιστων παρατηρήσεων, μέσω της σύγκρισης των δεδομένων με το κεντροειδές της ομάδας στην οποία ανήκουν (anomaly detection). Η διαδικασία αυτή θα μας οδηγήσει σε σημαντικά ευρήματα, η περαιτέρω μελέτη των οποίων αποτελεί σημαντική για την καλύτερη κατανόηση του συνόλου μας και την λήψη αποφάσεων. Συγκεκριμένα η κατηγοριοποίηση των δεδομένων κατανάλωσης και η εύρεση ασυνήθιστων παρατηρήσεων μπορεί να οδηγήσει στην ανακάλυψη και αντιμετώπιση υποκλοπής ρεύματος.

(ii) Hierarchical clustering:
Όπως υποδηλώνει και το όνομά της η μέθοδος αυτή χρησιμοποιείται για την δημιουργία μιας ιεραρχίας ομάδων. Η μέθοδος έχει δύο διαφορετικές στρατηγικές, με βάση τις οποίες δημιουργεί την εν λόγω ιεραρχία:

(1) Agglomerative strategy:
Η στρατηγική αυτή αποτελεί μια bottom up προσέγγιση για την διαμόρφωση της ιεραρχίας. Συγκεκριμένα ξεκινάμε με την δημιουργία μιας ομάδας για κάθε στοιχείο του συνόλου. Στη συνέχεια με χρήση μιας μετρκής απόστασης, αποφασίζουμε ποια σημεία βρίσκονται αρκετά κοντά και τα ταξινομούμε στην ίδια ομάδα. Η μέθοδος συνεχίζει, μέχρι να καταλήξουμε στον προκαθορισμένο αριθμό ομάδων.

(2) Divisive strategy
Στην στρατηγική αυτή ξεκινάμε με την δημιουργία μιας ομάδας για όλα τα στοιχεία. Στη συνέχεια με χρήση μιας μετρικής και ενός κριτηρίου διάσπασης διαχωρίζουμε την αρχική ομάδα στα δύο. Η διαδικασία συνεχίζεται μέχρις ότου να καταλήξουμε στον προκαθορισμένο αριθμό ομάδων.

Με την δημιουργία μιας ιεραρχίας ομάδων είναι εφικτό να διακρίνουμε τις ομάδες με την μεγαλύτερη κατανάλωση. Η πληροφορία αυτή είναι ιδιαίτερα χρήσιμη, καθώς μας επιτρέπει να σχεδιάσουμε στοχευμένες διαφημιστικές καμπάνιες, με σκοπό την εκπαίδευση των πολιτών ως προς την αξία της ενεργειακής οικονομίας και απόδοσης.

(iii) Spectral clustering:
Ο αλγόριθμος αυτός χρησιμοποιεί τις ιδιοτιμές και τα ιδιοδιανύσματα του πίνακα ομοιότητας (Affinity matrix) ή ενός πίνακα που παράγεται από αυτόν με σκοπό των διαχωρισμό των δεδομένων σε ομάδες.

Προκειμένου να πετύχει το παραπάνω η μέθοδος θεωρεί το σύνολο των δεδομένων σαν έναν μη συμπαγή γράφο. Ο γράφος αυτός έχει ως κόμβους του τα δεδομένα (data points) και ως ακμές τις συσχετίσεις των δεδομένων που επάγονται από το επιλεγμένο μέτρο ομοιότητας (similarity measure). Κάθε κόμβος συσχετίζεται με τους υπόλοιπους με βάρος το μέτρο ομοιότητας μεταξύ των δύο κόμβων.

Η χρησιμότητα του αλγορίθμου προκύπτει από την ικανότητά του να βρίσκει στατιστικά σημαντικές ομαδοποιήσεις στο σύνολο των δεδομένων. Συγκεκριμένα ο αλγόριθμος θα χρησιμοποιηθεί με σκοπό την εύρεση σημαντικών γεωγραφικών ομάδων μεταξύ των ″αφύσικων″ ενεργειακών δεδομένων.

Δηλαδή έχοντας ήδη βρεί με χρήση της μεθόδου pattern matching ″αφύσικα″ ενεργειακά δεδομένα, ο αλγόριθμος θα επειχηρήσει να ελέγξει για γεωγραφικές περιοχές στην πόλη

στις οποίες συγκεντρώνονται τα δεδομένα αυτά. Με τον τρόπο αυτό ενδέχεται να ευρεθούν γειτονίες στις οποίες παρουσιάζεται νοοτροπία κλοπής ή ακόμα και πιθανές βλάβες στις εγκαταστάσεις ρεύματος.

Η πληροφορία αυτή αποτελεί ιδιαίτερα σημαντική καθώς μας επιτρέπει να διαμορφώσουμε στοχευμένα δρώμενα κατά περιοχή, με στόχο την ενημέρωση των πολιτών ως προς τα προβλήματα της υποκλοπής ρεύματος και του βανδαλισμού. Επίσης οι ανακάλυψη βλαβών σε συγκεκριμένες περιοχές της πόλης, μας οδηγεί σε περαιτέρω μελέτη για την ύπαρξη και αντικατάσταση καιρισμένου και ελαττωματικού εξοπλισμού.

Στο σημείο αυτό παραθέτουμε το διάγραμμα χρησιμότητας για τους αλγόριθμους ομαδοποίησης :

Πίνακας 4: Χρησιμότητα αλγορίθμων συσταδοποίησης:

Όνομα αλγορίθμου	Χρησιμότητα
KMeans	Ανακάλυψη και αντιμετώπιση υποκλοπής ρεύματος.
Hierarchical clustering	Σχεδιασμός δρώμενων και διαφημίσεων με στόχο την ενεργειακή απόδοση.
Spectral clustering	Σχεδιασμός δρώμενων ενάντια στην υποκλοπή και τον βανδαλισμό.

2.3 Διάγραμμα αλγοριθμικής ροής

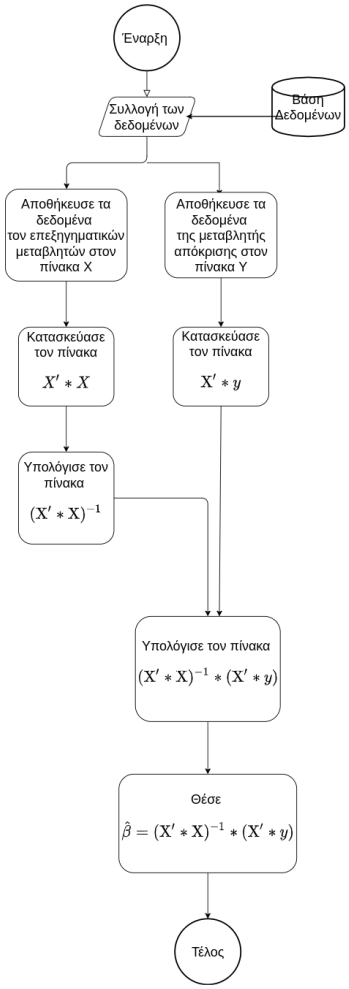
Έχοντας πλέον αναφέρει τους αλγορίθμους που θα χρησιμοποιηθούν καθώς και την χρησιμότητα αυτών, μεταβαίνουμε στην διαγραμματική τους παρουσίαση. Επαναλαμβάνουμε επιγραμματικά την δομή του διαγράμματος.

Κατά την διαγραμματική παρουσιάση του αλγορίθμου θα χρησιμοποιθούν 4 διαφορετικά σχήματα:

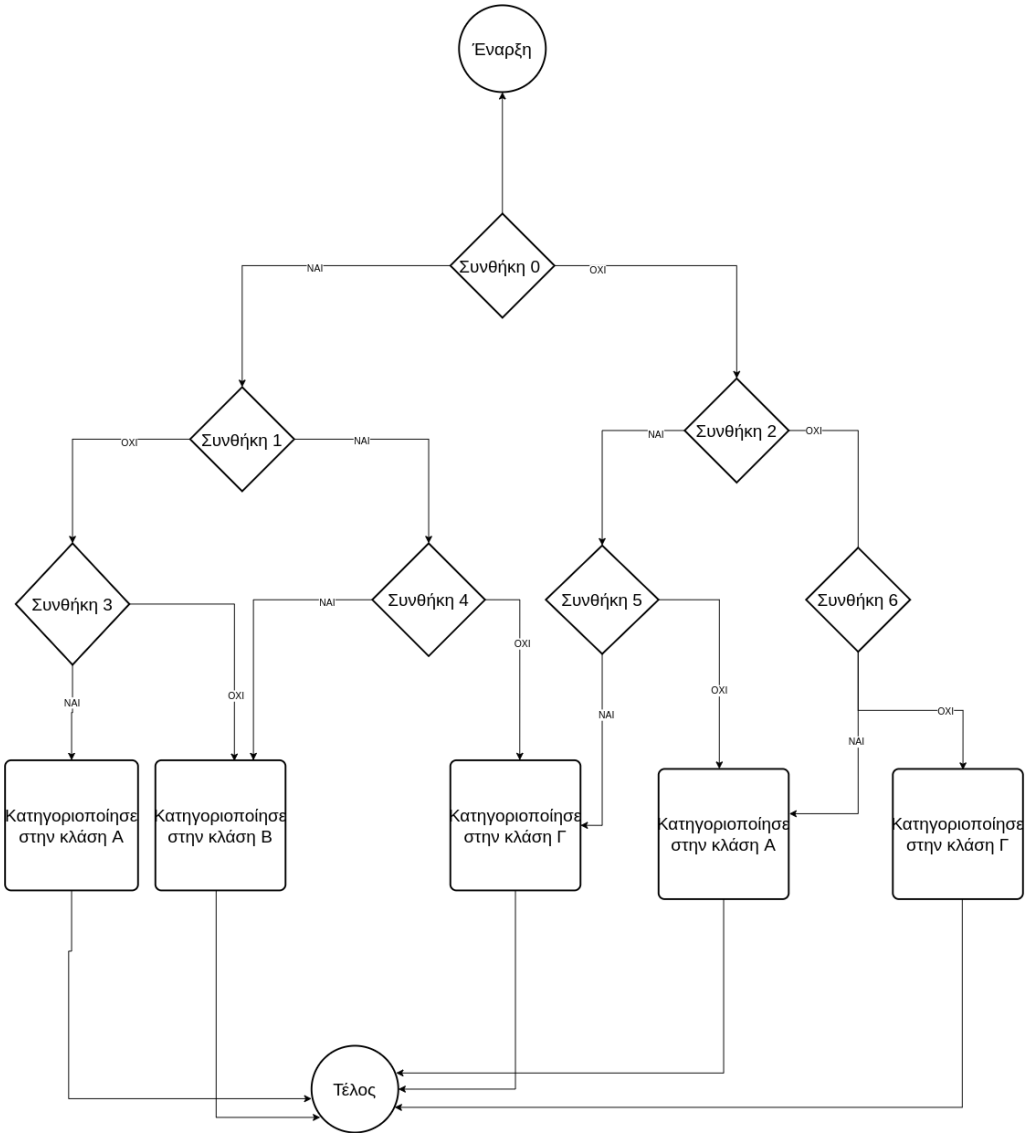
- Κύκλος ή έλλειψη: Περιγράφει το σημείο έναρξης/τερματισμού του αλγορίθμου.
- Παραλληλόγραμμο: Περιγράφει το σημείο εισαγωγής δεδομένων από τον χρήστη.
- Ρόμβος: Περιγράφει την λήψη μιας απόφασης.
- Ορθογώνιο παραλληλόγραμμο: Περιγράφει την εκτέλεση κάποιας διαδικασίας.

Παρακάτω παραθέτουμε τα διαγράμματα αλγοριθμικής ροής για τις μεθόδους που επιλέξαμε.

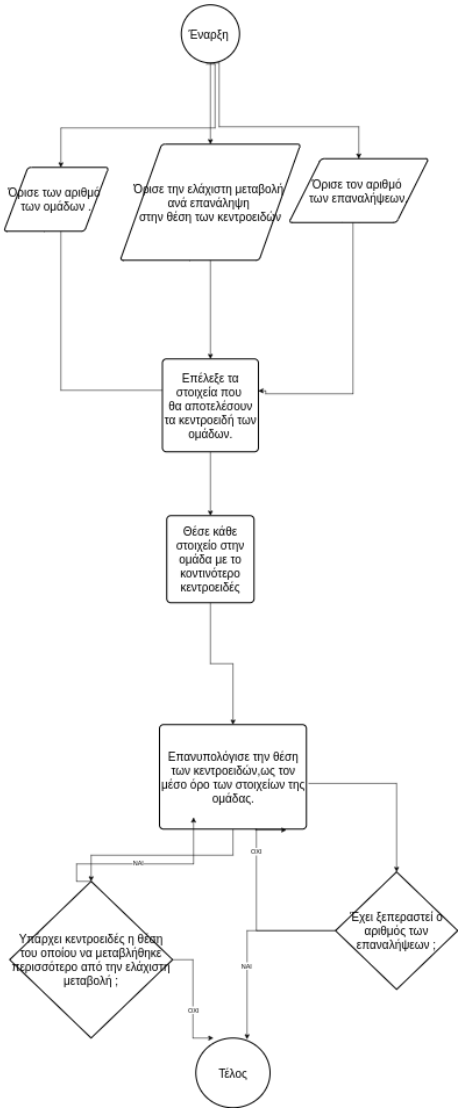
- Αλγόριθμοι παλινδρόμησης:
  - Γραμμική/Πολυωνυμική παλινδρόμηση:



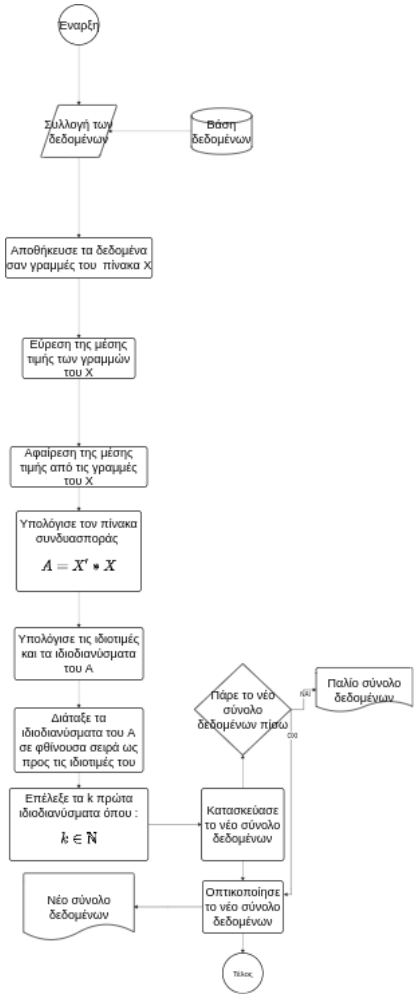
- Αλγόριθμοι κατηγοριοποίησης:
  - Δέντρο απόφασης:



- Αλγόριθμοι συσταδοποίησης:
  - K-means



- Λοιποί Αλγόριθμοι:
  - PCA



Με την υλοποίηση των αλγορίθμων που παρουσιάστηκαν παραπάνω αναμένουμε την εύρεση σημαντικών αποτελεσμάτων. Από τα ευρήματα αυτά θα προκύψει πιθανό όφελος για τον δήμο, μέσω της χρήσης τους για την επίτευξη των στόχων που αναφέρθηκαν στην εισαγωγή.

Αναφέρουμε στο σημείο αυτό πως σε αντίθεση με τους υπόλοιπους αλγορίθμους, ο αλγόριθμος PCA δεν παρέχει άμεσο όφελος στον Δήμο. Παρόλα αυτά η εφαρμογή του καθιστά την ανάλυση των δεδομένων πιο ποιοτική και επομένως τα ευρήματα που προσφέρονται πιο αξιόπιστα.

Στον παρακάτω πίνακα παραθέτουμε τα πιθανά ευρήματα και οφέλη για τους τέσσερις αλγορίθμους που παρουσιάστηκαν παραπάνω:

Πιθανό όφελος ανάλυσης		
Όνομα αλγορίθμου	Πιθανά ευρήματα	Πιθανό όφελος
1. Γραμμική/Πολυωνιμική παλινδρόμηση		
• Γραμμική • Πολυωνιμική	Εύρεση σημ. παραγ. για την πρόβλ. κατ. Εύρεση σημ. παραγ., πρόβλεψη καταν.	Διερεύνηση σημαντ. παραγόντων Εύρεση "αφύσικων" καταν.
2. Δέντρο απόφασης/Τυχαίο δάσος		
• Δέντρο απόφασης • Τυχαίο Δάσος	Ταξινόμηση πολιτών(καταναλ. συνήθειες) Ταξινόμηση πολιτών(καταναλ. συνήθειες)	Διαμόρφωση νέων υπηρεσιών για την ικανοποίηση των πολιτών Εύρεση πιθανής υποκλοπής
3. K-means		
• K-means	Εύρεση προτύπων κατανάλωσης	Εύρεση πιθανής υποκλοπής
4. PCA		
• PCA	Εύρεση σημαντικού υποσυνόλου επεξ. μεταβ.	Ποιοτικότερη ανάλυση

2.4 Συμπεράσματα θεωρητικής μελέτης

Στο σημείο αυτό έχουμε ολοκληρώσει την θεωρητική ανάλυση των ενεργειακών και δημοσιονομικών δεδομένων.
Υπενθυμίζουμε πως κατά την διαδικασία αυτή:

(α΄) Εμπλουτίσαμε νοηματικά τα ως προς ανάλυση σύνολα δεδομένων.

(β΄) Μελετήσαμε τους χρήσιμους, δεδομένων των περιορισμών και στόχων αλγορίθμους.

(γ΄) Καταγράψαμε την πιθανή αξία που θα έχουν τα ευρήματα αυτών.

(δ΄) Παραθέσαμε την αλγοριθμική ροή των προτεινώμενων για μια αρχική ανάλυση αλγορίθμων.

Από την παραπάνω διεργασία ο δήμος αποκτά ένα πρακτικό κέρδος το οποίο παράγεται από την λήψη αποφάσεων και υλοποίηση συστημάτων που επιτρέπουν την επίτευξη των στόχων που αυτός έθεσε. Συγκεκριμένα η κατηγοριοποίηση των συνόλων, η εύρεση προτύπων σε αυτά και η κατανόηση των

συσχετίσεων μεταξύ τους μας επίτρεπει να δημιουργούμε συστήματα πρόβλεψης, συστήματα αξιοποίησης και εξοικονόμησης και τέλος συστήματα αποφάσεων που θα αποτελούν καταλυτικής σημασίας για την επίτευξη των επιθυμητών στόχων.

Συμπεραίνουμε λοιπόν, πως η υλοποίηση των μεθόδων που αναφέρθηκαν παραπάνω θα επιτρέψει την εξοικονόμηση ενέργειας από τον δήμο, την καλύτερη αξιοποίηση των πόρων που αυτός διαθέτει, την εύρεση και αντιμετώπιση φαινομένων υποκλοπής και εγκληματικότητας και τέλος την δημιουργία δρώμενων με στόχο την ενημέρωση και εκπαίδευση των πολίτων στα ζητήματα που απασχολούν μια έξυπνη πόλη.

3 Μεθοδολογίες ανάλυσης δεδομένων

3.1 Μορφή διαθέσιμων δεδομένων

Κατά την περίοδο συγγραφής της διπλωματικής αυτής εργασίας, τα διαθέσιμα σύνολα δεδομένων που μας έχουν δωθεί από τον φορέα, περιορίζονται στην κατανάλωση ενέργειας του δήμου για τα έτη 2015-2018. Για τον λόγο αυτό θα επικεντρωθούμε στην μελέτη αυτών των συνόλων δεδομένων για το υπόλοιπο της εργασίας. Συγκεκριμένα, θα εστιάσουμε στις μεθοδολογίες ανάλυσης που χρησιμοποιήθηκαν, καθώς και στα τελικά αποτελέσματα που παράχθηκαν κατά την υλοποίησή τους.

Στο σημείο αυτό θα καταγράψουμε την μορφή των δεδομένων ενέργειας που μας δώθηκαν από τον πελάτη. Τα δεδομένα δώθηκαν σε μορφή xls. Κάθε γραμμή εκπροσωπεί έναν διαφορετικό λογαριασμό και κάθε στήλη μας παρέχει πληροφορίες για κάποιο χαρακτηριστικό του λογαριασμού.
Σκοπός μας λοιπόν είναι να περιγράψουμε περιληπτικά το περιεχόμενο των στήλων, που αποτελούν καθοριστικής σημασίας για την περαιτέρω επεξεργασία.

- Αριθμός Παροχής:

  Ο αριθμός ταυτοποίησης της κάθε παροχής. Με τον όρο παροχή αναφερόμαστε στο συμβόλαιο μεταξύ του πελάτη και της εταιρίας παροχής ηλεκτρικής ενέργειας. Ο αριθμός αυτός είναι μοναδικός για κάθε συμβόλαιο.

- Έτος: Το έτος έκδοσης του λογαριασμού

- Μήνας: Ο μήνας έκδοσης του λογαριασμού

- Τιμολόγιο:

  Το τιμολόγιο με βάση το οποίο χρεώνεται ο πελάτης. Το τιμολόγιο είναι μοναδικό για κάθε παροχή. Τα τιμολόγια που έχουμε στην διάθεση μας είναι:

  1. Γ21:

     Η εύρεση της κατανάλωσης για παροχές τέτοιου τιμολογίου γίνεται χειροκίνητα.
     Δηλαδή, για την καταμέτρηση της κατανάλωσης απαιτείται η φυσική παρουσία υπαλλήλου του ΔΕΔΔΗΕ. Το τιμολόγιο αυτό απευθύνεται σε επιχειρήσεις και οργανισμούς, οι εγκαταστάσεις των οποίων δεν ξεπερνούν ένα δεδομένο όριο τετραγωνικών μέτρων.

  2. Γ22:

     Η εύρεση της κατανάλωσης για παροχές τέτοιου τιμολογίου γίνεται χειροκίνητα.
     Το τιμολόγιο αυτό απευθύνεται σε επιχειρήσεις και οργανισμούς, οι εγκαταστάσεις των οποίων ξεπερνούν ένα δεδομένο όριο τετραγωνικών μέτρων.

3. Γ23:

   Η εύρεση της κατανάλωσης για παροχές τέτοιου τιμολογίου γίνεται μέσω τηλεμέτρησης. Δηλαδή το ρολόι της υπάρχουσας παροχής είναι ενσωματομένο με chip το οποίο καταγράφει και στέλνει την κατανάλωση του ρολογιού στο τέλος κάθε μήνα.
   Το τιμολόγιο αυτό απευθύνεται σε επιχειρήσεις.
   Η ιδιαιτερότητα του επικεντρώνεται στον διαχωρισμό τιμολόγησης σε πρωϊνή και νυχτερινή, με την νυχτερινή τιμολόγηση να αποτελεί την φθηνότερη από τις δύο.

4. ΦΟΠ:

   Τιμολόγιο για παρόχες δημοσίου φωτισμού. Η εύρεση της κατανάλωσης για παροχές τέτοιου τιμολογίου γίνεται χειροκίνητα.

5. ΑΓΡΟ:

   Τιμολόγιο για αγροτικές εκτάσεις. Αποτελεί το φθηνότερο τιμολόγιο.
   Η εύρεση της κατανάλωσης για παροχές τέτοιου τιμολογίου γίνεται μέσω τηλεμέτρησης.

- Όνομα Πελάτη
- Όνομα Οδού
- Αριθμός Οδού
- Πόλη/Περιοχή στην πόλη
- Ημερομηνία Έκδοσης Λογαριασμού
- Ημερομηνία Τελευταίας Καταμέτρησης:

  Η ημερομηνία στην οποία έγινε η τελευταία καταγραφή της κατανάλωσης για την εν λόγω παροχή.

- Ημερομηνία Προηγούμενης Καταμέτρησης:

  Η ημερομηνία στην οποία έγινε η αμέσως προηγούμενη καταγραφή της κατανάλωσης για την εν λόγω παροχή.
  Η περίοδος μεταξύ των ημερομηνιών προηγούμενης και τελευταίας καταμέτρησης θα αναφέρεται ως περίοδος καταμέτρησης.

- Ημέρες Κατανάλωσης:

  Η τιμή του πεδίου αυτού προκύπτει από την εύρεση των ημερών που μεσολάβησαν μεταξύ των ημερομηνιών προηγούμενης και τελευταίας καταμέτρησης.

- Κατανάλωση ενέργειας($Kw\ h$)

- Σύνολο Τρέχοντα Μήνα:

  Το συνολικό κόστος ενέργειας και διαχείρησης της παροχής για τον μήνα έκδοσης του λογαριασμού. Παρότι το κόστος αντιστοιχεί στον μήνα έκδοσης, η τιμή του αφορά την συνολική κατανάλωση, κατά τις καταμετρούμενες ημέρες αυτής της περιόδου.

- Τύπος Λογαριασμού:

  Οι λογαριασμοί χωρίζονται στις παρακάτω κατηγορίες:

  1. Εκκαθαριστικός(ΕΚΚΑΘ):

     Οι λογαριασμοί αυτής της μορφής, έχουν προκύψει από καταμέτρηση του ρολογιού της υπάρχουσας παροχής, είτε από υπάλληλο του ΔΕΔΔΗΕ είτε από τηλεμέτρηση.
     Στη συνέχεια της ανάλυσης, θα επικεντρωθούμε αποκλειστικά σε λογαριασμούς αυτής της μορφής, καθώς μας παρέχουν ακριβή πληροφορία για την κατανάλωση της παροχής.

  2. Έναντι(ΕΝΑΝΤ):

     Σε περίπτωση που η καταμέτρηση του ρολογιού δεν είναι εφικτή μετά από μια προκαθορισμένη αναμενόμενη περίοδο, η κατανάλωση της παροχής εκτιμάται. Η εκτίμηση αυτή προκύπτει από ιστορικά δεδομένα της παροχής ή και παροχών ίδιου τιμολογίου για αυτόν τον μήνα. Για τον λόγο αυτό λογαριασμοί αυτής της μορφής δεν θα ληφθούν υπόψη για την περαιτέρω επεξεργασία.

  3. Τελικός(ΤΕΛΙΚ):

     Σε περίπτωση που ο πελάτης έχει ζητήσει τερματισμό της παροχής, προσφέρεται ένας λογαριασμός τελευταίας κατανάλωσης. Ένας τελικός λογαριασμός μπορεί να έχει προκύψει είτε με καταμέτρηση της κατανάλωσης είτε με εκτίμηση. Έτσι η χρήση του θα εξαρτηθεί από την παραπάνω πληροφορία.

  4. Έκτακτος(ΕΚΤΑΚ):

     Οι λογαριασμοί αυτής της μορφής εκδίδονται εκτός της προγραμματισμένης περιόδου, συνήθως μετά από αμφισβήτηση της καταμετρούμενης κατανάλωσης από τον ιδιοκτήτη της παροχής.

3.2 Προεπεξεργασία δεδομένων

Έχοντας περιγράψει τα καθοριστικά για την ανάλυσή μας πεδία, αναφέρουμε ορισμένες διαδικασίες προεπεξεργασίας, που αποτελούν απαραίτητες για την περαιτέρω ανάλυση.

Αρχικά, θα διαμορφώσουμε μια βάση ιστορικών δεδομένων, με σκοπό την εκπαίδευση μοντέλων μηχανικής μάθησης και τον εντοπισμό αρχικών ευρημάτων στα δεδομένα κατανάλωσης που αφορούν τον συγκεκριμένο πελάτη. Προκειμένου τα συμπεράσματα που θα προκύψουν από τα ιστορικά δεδομένα να είναι χρήσιμα για την ανάλυση, η βάση θα περιέχει μονάχα τα δεδομένα των δύο προηγούμενων ετών. Με τον τρόπο αυτό, θα διαμορφωθεί μια επαρκής και σχετική βάση σύγκρισης, για τα μηνιαία δεδομένα που θα προσφέρονται από και προς τον πελάτη.

Ένα άλλο σημαντικό βήμα προεπεξεργασίας αφορά την επιλογή του τύπου λογαριασμού που θα χρησιμοποιήσουμε. Όπως έχουμε ήδη αναφέρει οι δύο κύριοι τύποι λογαριασμού είναι ο εκκαθαριστικός και ο έναντι. Σε αντίθεση με έναν εκκαθαριστικό λογαριασμό στον οποίο έχουμε καταμέτρηση της κατανάλωσης, ένας έναντι προσφέρει εκτιμούμενα δεδομένα.

Συγκεκριμένα, σε περίπτωση έναντι λογαριασμού, η κατανάλωση εκτιμάται με χρήση ιστορικών δεδομένων, ενώ στην περίπτωση εκκαθαριστικού, έχουμε καταμέτρηση της πραγματικής κατανάλωσης για την δεδομένη περίοδο. Για τον λόγο αυτό θα περιοριστούμε στα δεδομένα εκκαθαριστικών λογαριασμών, καθώς αυτά αποτελούν πιο αξιόπιστα για το δεδομένο έτος.

Πέραν όμως από την χρήση του τύπου των λογαριασμών για την αξιολόγηση, θα επικεντρωθούμε επίσης και στην ποιότητα της κάθε παροχής. Για κάθε παροχή ο συνολικός ετήσιως αριθμός εκκαθαρίσεων διαφέρει, με τον μέγιστο αριθμό να αποτελεί 12 και τον ελάχιστο 1.

Όπως είναι λοιπόν κατανοητό η αυξημένη συχνότητα εκκαθαρίσεων μιας παροχής μας προσφέρει μεγαλύτερη πληροφορία για τις αυξομειώσεις της κατανάλωσης κατα τους διαφορετικούς μήνες και εποχές. Έτσι, η πληθώρα των εκκαθαρίσεων αυξάνει την αξία των δεδομένων, καθώς επιτρέπει την καλύτερη αντιπροσώπευση της κατανάλωσης στον άξονα του χρόνου. Με σκοπό λοιπόν την ποιοτικότερη ανάλυση, θα αγνοήσουμε παροχές των οποίων η ετήσια συχνοτήτα εκκαθαρίσεων καθιστάται χαμηλή.

Ο διαχωρισμός αυτός θα γίνει με τον παρακάτω τρόπο:

- **Υψηλής Ποιότητας:** 7 - 12 εκκαθαριστικοί λογαριασμοί ετησίως.
- **Μέσης Ποιότητας:** 3 - 6 εκκαθαριστικοί λογαριασμοί ετησίως.
- **Χαμηλής Ποιότητας:** 0 - 2 εκκαθαριστικοί λογαριασμοί ετησίως.

Η γενική αρχή είναι, ότι όσο μεγαλύτερη η πυκνότητα των δεδομένων κατανάλωσης στον άξονα του χρόνου (άρα και του αριθμού εκκαθαριστικών λογαριασμών), τόσο ακριβέστερη η παρακολούθηση της διακύμανσης της κατανάλωσης, κάτι που με τη σειρά του επιτρέπει τον έγκαιρο εντοπισμό σημαντικών - ″μη κανονικών″ μεταβολών σε αυτήν.

Το τελικό βήμα προεπεξεργασίας αφορά την δημιουργία του πεδιού $kw\ h/day$ αλλά και την αναγωγή αυτού σε μηνιαία βάση. Όπως υποδηλώνεται από το όνομα του, το πεδίο αυτό προκύπτει διαιρώντας την κατανάλωση ενέργειας υπολογισμένη σε κιλοβατώρες, με τις ημέρες κατανάλωσης για δεδομένη περίοδο καταμέτρησης.

Με την παραπάνω διεργασία γίνεται εφικτή η δυνατότητα εύστοχης σύγκρισης της κατανάλωσης μεταξύ κοινών χρονικών περιόδων. Σε περίπτωση που η συχνότητα εκκαθαρίσεων της παροχής δεν είναι μηνιαία, απαιτείται για την διεξαγωγή της σύγκρισης, η προαναφερόμενη αναγωγή.

Η διαδικασία της αναγωγής εξηγείται παρακάτω με την χρήση υποθετικού παραδείγματος:

Ημερομ.Τελευτ.Καταμ	Ημερομ.Προηγ.Καταμ	Ημέρες Καταναλ.	Καταναλ.Ενέργειας	Τύπος Λογαρ.
24/01/2019	27/09/2018	120,00	3298,00	ΕΚΚΑΘ.
25/05/2019	25/01/2019	121,00	2728,00	ΕΚΚΑΘ.
25/09/2019	26/05/2019	123,00	2383,00	ΕΚΚΑΘ.

Όπως παρατηρούμε παραπάνω, για την συγκεκριμένη παροχή έχουν εκδοθεί τρεις εκκαθαριστικοί για το έτος 2019, με αποτέλεσμα να χαρακτηρίζεται ως μέσης ποιότητας με βάση τον παραπάνω διαχωρισμό.

Η μηνιαία αναγωγή του πεδίου $kw\ h/day$ για τον μήνα Μάιο υπολογίζεται ως εξής:

- Η περίοδος εκκαθάρισης που περιλαμβάνει τις 25 από τις 31 ημέρες του μήνα έχει κατανάλωση :

$$2728(kw\ h)/121(days) = 22,55(kw\ h/day)$$

- Η περίοδος εκκαθάρισης που περιλαμβάνει τις 6 από τις 31 ημέρες του μήνα έχει κατανάλωση:

$$2383(kw\ h)/123(days) = 19,37(kw\ h/day)$$

Επομένως η συνολική κατανάλωση για τον Μαΐο του 2019 είναι:

$$25(days) * 22.55(kw\ h/day) + 6(days) * 19.37(kw\ h/day) = 679.97(kw\ h)$$

Τέλος ο δείκτης $kw\ h/day$ για τον Μαΐο του 2019 ισούται με:

$$679,97(kw\ h)/31(days) = 21,93(kw\ h/day)$$

Σε περίπτωση που ένας μήνας περιέχεται πλήρως στην περίοδο εκκαθάρισης, η μηνιαία αναγωγή προκύπτει ως το γινόμενο των ημερών του συγκεκριμένου μήνα με τον δείκτη $kw\ h/day$ για την περίοδο εκκαθάρισης. Με την ολοκλήρωση της προεπεξεργασίας αναμένεται να προκύψουν 24 μηνιαία ιστορικά δεδομένα κατανάλωσης ανά παροχή, όπως και αντίστοιχες τιμές του δείκτη $kw\ h/day$, σαν αποτέλεσμα των 12 μηνιαίων καταναλώσεων ανά έτος για τα δύο προηγούμενα έτη.

3.3 Μέθοδοι ανάλυσης δεδομένων

Στο σημείο αυτό, ερχόμαστε πλέον στο κύριο αντικείμενο του συγκεκριμένου κεφαλαίου, την επεξήγηση των μεθολογίων ανάλυσης.

Για κάθε μέθοδο, θα επεξηγήσουμε την διαδικασία ανάλυσης, τα αξιοποιήσιμα πεδία, καθώς και το επιθυμητό τελικό αποτέλεσμα.

Οι μέθοδοι χωρίζονται σε δύο βασικές κατηγορίες:

- Ανάλυση ανά παροχή στον άξονα του χρόνου.
  Για μεθόδους αυτής της μορφής τα βασικά αξιοποιήσιμα πεδία αποτελούν ο αριθμός παροχής και οι ημέρες προηγούμενης, τελευταίας καταμέτρησης.
- Ανάλυση μεταξύ παροχών σε κοινό χρόνο.
  Για μεθόδους αυτής της μορφής τα βασικά αξιοποιήσιμα πεδία αποτελούν οι ημέρες προηγούμενης, τελευταίας καταμέτρησης.

Το βασικότερο πεδίο αξιοποίησης για κάθε κατηγορία, αποτελεί η κατανάλωση ενέργειας.
Το πεδίο αυτό, θα χρησιμοποιηθεί ως επί το πλείστον για την σύγκριση της κατανάλωσης μεταξύ διαφορετικών μηνών και περιόδων καταμέτρησης.

Προκειμένου αυτή η σύγκριση να είναι εύστοχη, το πεδίο θα επεξεργαστεί περαιτέρω.
Συγκεκριμένα, η κατανάλωση θα διαίρεθει με τις ημέρες κατανάλωσης της συγκεκριμένης περιόδου, κατασκευάζωντας έτσι ένα νέο πεδίο που θα εκφράζει την μέση ημερήσια κατανάλωση σε κιλοβατώρες ($kw\ h/day$), για την δεδομένη περίοδο καταμέτρησης.

Με τον τρόπο αυτό οι διαφορές στην κατανάλωση μεταξύ μηνών, που οφείλονται στον διαφορετικό αριθμό ημερών του καθενός δεν θα λαμβάνονται υπόψη. Αντίστοιχα οι διαφορές στην κατανάλωση μεταξύ διαφορετικών περιόδων, που οφείλονται στον διαφορετικό αριθμό ημερών κατανάλωσης της καθεμίας, δεν θα λαμβάνεται υπόψη. Οι διαφορές αυτές αποτελούν αναμενόμενες και επομένως η επισήμανσή τους κατά την ολοκλήρωση της ανάλυσης θεωρείται μη επιθυμητή.

Σε περίπτωση που η συχνότητα εκκαθαρίσεων για μια παροχή δεν είναι μηνιαία, η διεξαγωγή των παραπάνω συγκρίσεων απαιτεί περαιτέρω επεξεργασία, όπως αυτή έχει περιγραφεί στο προηγούμενο κεφάλαιο.

Πρωτού εμβαθύνουμε στις ιδιαιτερότητες της κάθε μεθόδου παρουσιάζουμε μια συνοπτική λίστα αυτών, διαχωρισμένη στις δύο προαναφερόμενες κατηγορίες.

A : Ανάλυση ανά παροχή στον άξονα του χρόνου

- $A1$: Σύγκριση του δείκτη $kw\ h/day$ για δεδομένο μήνα μεταξύ του τρέχοντος και προηγούμενου έτους.
- $A2$: Υπολογισμός απλού κινητού μέσου όρου 12 μηνών (KMO12) για την κατανάλωση ενέργειας.
- $A3$: Υπολογισμός του δείκτη z-score για όλες τις διαθέσιμες μηνιαίες τιμές του δείκτη $kw\ h/day$.

B : Ανάλυση μεταξύ παροχών σε κοινό χρόνο

$B1$: Εύρεση παροχών με μηδενική κατανάλωση ενέργειας.

$B2$: Εύρεση παροχών με ανεπαρκή αριθμό εκκαθαρίσεων.

$B3$: Εύρεση μοτίβου κατανάλωσης για παροχές ίδιου τιμολογίου.

Έχοντας καταγράψει της διαφορετικές μεθόδους που θα υλοποιήσουμε, επικεντρωνόμαστε πλέον στην λεπτομερή επεξήγησή τους.

Μέθοδος $A1$: Κατά την μεθοδολογία αυτή, θα συγκρίνουμε τον δείκτη $kw\ h/day$ για δεδομένο μήνα του τρέχοντος έτους με τον αντίστοιχο μήνα του προήγουμενου έτους. Η σύγκριση αυτή θα γίνει με την εύρεση της ποσοστιαίας διαφοράς μεταξύ των δύο δεικτών.

Η υπολογισμένη πληροφορία θα αποθηκεύεται σε πίνακα παροχών μαζί με τον μήνα, την περίοδο και τον δείκτη $kw\ h/day$ του μήνα για το τρέχον και το προηγούμενο έτος. Στη συνέχεια θα διαμορφωθεί ιστόγραμμα για τον δεδομένο αυτό μήνα, που θα παρουσίαζει την κατανομή των ποσοστών απόκλισης για τις διάφορες παροχές, με χρήση διαδοχικών διαστημάτων.

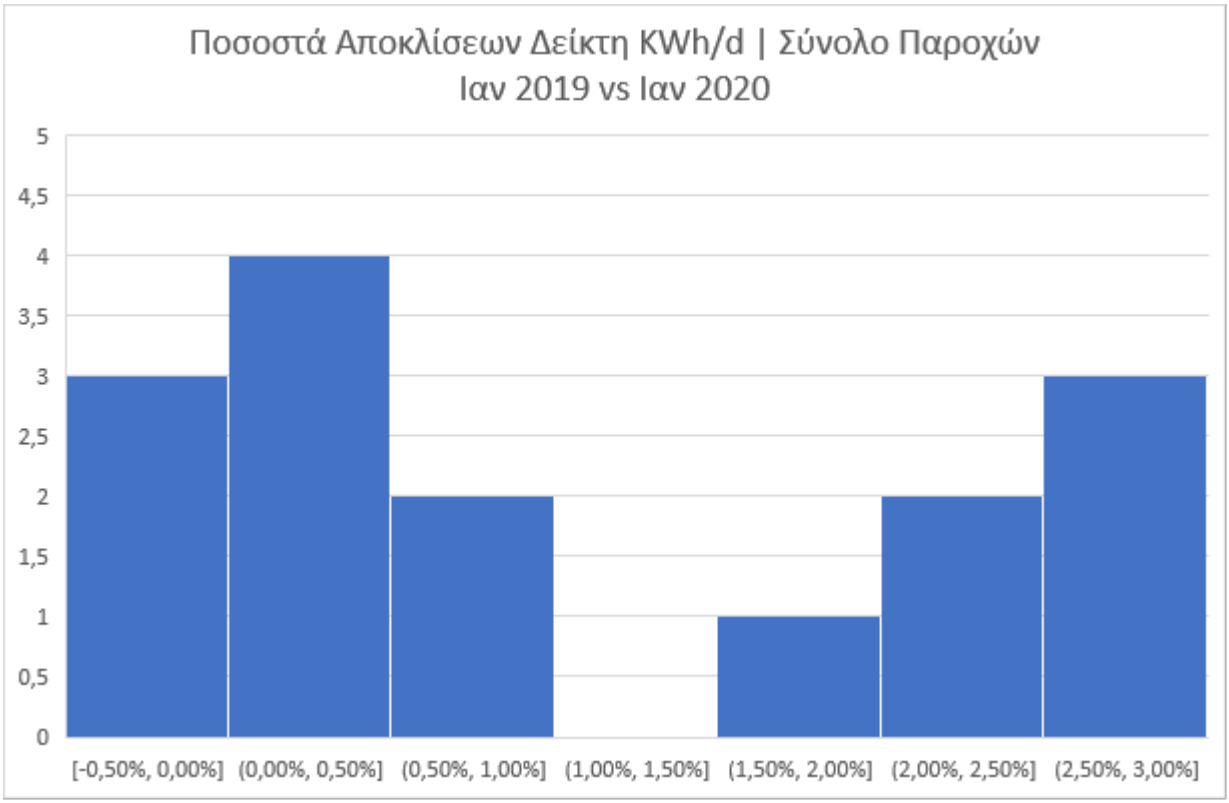
Ακολουθεί πίνακας υποθετικών δεδομένων 15 παροχών, για τον δείκτη $kw\ h/day$ του μήνα

Ιανουαρίου, για τα έτη 2019, 2020:

Αριθμός Παροχής	$kw\ h/day$ (Γεν.2019)	$kw\ h/day$ (Γεν.2020)	Ποσοστιαία Απόκλιση(%)
1	22,00	22,11	0,50%
2	30,00	30,15	0,50%
3	34,00	34,17	0,50%
4	45,00	45,23	0,50%
5	12,00	12,08	0,70%
6	45,00	45,32	0,70%
7	35,00	35,63	1,80%
8	67,00	68,41	2,10%
9	87,00	88,83	2,10%
10	98,00	100,65	2,70%
11	13,00	13,35	2,70%
12	43,00	44,16	2,70%
13	47,00	46,77	-0,50%
14	81,00	80,60	-0,50%
15	18,00	17,91	-0,50%

Μέσω του διαγράμματος αυτού ο πελάτης μπορεί να έχει μια καθολική και εύκολη εποπτεία των ποσοστιαίων αποκλίσεων των καταναλώσεων μεταξύ ενός μήνα του τρέχοντος έτους και του αντίστοιχου για το προηγούμενο. Με τον τρόπο, θα είναι η εφικτή η άμεση εύρεση παροχών με σημαντική απόκλιση, καθώς και η περαιτέρω διερεύνηση αιτιών της.

Στη συνέχεια παραθέτουμε το υποθετικό ιστόγραμμα που αντιστοιχεί στα παραπάνω δεδομένα.



Από το παραπάνω διάγραμμα γίνεται κατανοητο, ότι στη συγκριτική ανάλυση του μήνα Ιανουαρίου με τον αντίστοιχο περσινό προέκυψαν 3 παροχές με απόκλιση μεταξύ [2,50% - 3,00%], οι οποίες εύκολα μπορούν να εντοπιστούν από τον πίνακα.

Οι συγκεκριμένες παροχές έχουν προτεραιότητα, για την περαιτέρω διερεύνηση αιτιών των αποκλίσεων, αλλά και την υλοποίηση σχεδίων για την επιδιόρθωσή τους από τον πελάτη.
Στην μηνιαία αναφορά που θα παραδίδεται στον πελάτη θα περιέχεται το ιστόγραμμα και ο πίνακας παροχών που παρουσιάστηκαν παραπάνω.

Με σκοπό την εύκολη επισκόπηση των σημαντικότερων αποκλίσεων από τον δήμο, το πλήθος των στοιχείων του πίνακα θα περιορίζεται από τον μέγιστο αριθμό παροχών τον οποίο ο δήμος επιθύμει να παρακολουθεί σε μηνιαία βάση. Συγκεκριμένα δεδομένου ότι ο μέγιστος αυτός αριθμός είναι n, ο πίνακας θα περιέχει το πολύ τις n/2 παροχές με την μεγαλύτερη απόκλιση και το πολύ τις n/2 παροχές με την μικρότερη.

Μέθοδος $A2$: Κατά την μεθολογία αυτή θα υπολογίσουμε τον κινούμενο μέσο όρο 12 μηνών (ΚΜΟ12), με σκοπό την εύρεση πιθανών μακροχρόνιων τάσεων στην κατανάλωση δεδομένης παροχής.

Η υπολογισμένη ποσότητα θα αποθηκευτεί σε πίνακα παροχών σε συνδυασμό με τις μηνιαίες τιμές του δείκτη $kw\ h/day$ για το τρέχων και το προηγούμενο έτος, τον μήνα, την περίοδο υπολογισμού και την ποσοστιαία απόκλιση για τον δείκτη ΚΜΟ12 μεταξύ πρώτου και τελευταίου μήνα της τελευταίας δωδεκάδας μηνών που βρίσκονται στον πίνακα παροχών.

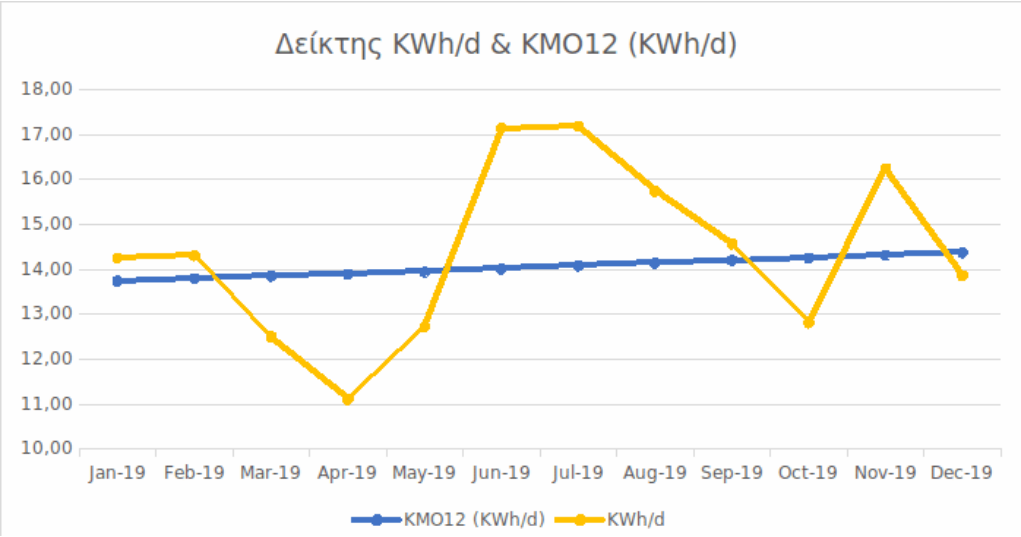
Στη συνέχεια θα διαμορφωθεί γράφημα που θα αναπαρίστα τις τιμές του δείκτη ΚΜΟ12 για τους τελευταίους 12 μήνες.

Ο κινητός μέσος όρος προκύπτει από τον υπολογισμό ενός συνόλου μέσων όρων των τιμών ενός μεγέθους, διατηρώντας ένα σταθερό πλήθος δεδομένων στο κάθε άθροισμα, στην περίπτωση μας 12 μήνες. Κάθε μέσος όρος σε αυτό το σύνολο θα προκύπτει λαμβάνοντας κάθε φορά υπόψη την πιο πρόσφατη τιμή των δεδομένων και αγνοώντας την παλαιότερη.

Ο συγκεκριμένος δείκτης θα υπολογίζεται αρχικά για τα ιστορικά δεδομένα που συλλέχθηκαν κατά την διαδικασία της προεξεργασίας. Στη συνέχεια, με την προσθήκη μηνιαίων δεδομένων κατανάλωσης σε κάθε νέο εκκαθαριστικό λογαριασμό της παροχής, ο δείκτης θα υπολογίζεται και θα παρακολουθείται ανά μήνα για τον έλεγχο της τάσης.

Ακολουθεί πίνακας υποθετικών τιμών κατανάλωσης για μια παροχή, ο δείκτης $kw\ h/day$ και ο αντίστοιχος δείκτης ΚΜΟ12 για τις τιμές του πεδίου $kw\ h/day$ όπως και η διαγραμματική τους απεικόνιση:

Μήνας	Κατανάλωση	$kw\ h/day$	ΚΜΟ12($kw\ h/day$)
Ιαν-19	441,68	14,25	13,74
Φεβ-19	400,50	14,30	13,79
Μαρ-19	387,48	12,50	13,84
Απρ-19	333,20	11,11	13,89
Μαϊ-19	394,26	12,72	13,94
Ιουν-19	513,80	17,13	14,01
Ιουλ-19	532,45	17,18	14,07
Αυγ-19	487,74	15,73	14,14
Σεπ-19	436,80	14,56	14,19
Οκτ-19	396,97	12,81	14,24
Νοε-19	487,20	16,24	14,31
Δεκ-19	429,48	13,85	14,36



Από το παραπάνω διάγραμμα φαίνεται καθαρά η αυξητική τάση της κατανάλωσης της συγκεκριμένης παροχής μέσω του δείκτη KMO12, παρά τις εποχικές της διακυμάνσεις που παρουσιάζονται μέσω του δείκτη $kw\ h/day$. Οι διακυμάνσεις αυτές αναμένεται να οφείλονται στις κλιματολογικές αλλαγές που παρουσιάζονται μεταξύ διαφορετικών μηνών και την σχετική με αυτές χρήση κλιματιστικών μονάδων.

Στην μηνιαία αναφορά προς τον δήμο θα παραδίδεται ο πίνακας παροχών που αναφέρθηκε παραπάνω για παροχές που υποδεικνύουν μακροχρόνια ανωδική ή πτωτική τάση, καθώς και γράφημα με τα δεδομένα του δείκτη KMO12 για τους τελευταίους δώδεκα μήνες.

Μια παροχή θα θεωρείται πως παρουσιάζει μακροχρόνια ανωδική ή πτωτική τάση σε περίπτωση που η ποσοστιαία απόκλιση που καταγράφεται στον πίνακα ξεπερνά κατά απόλυτη τιμή μια συμφωνημένη με τον πελάτη ποσοστιαία τιμή. Συγκεκριμένα σε περίπτωση που η απόκλιση είναι μεγαλύτερη από την τιμή, η παροχή θα θεωρείται πως εμφανίζει μακροχρόνια ανωδική τάση, ενώ σε περίπτωση που η απόκλιση είναι μικρότερη από το μείον της τιμής, η παροχή θα θεωρείται πως εμφανίζει μακροχρόνια πτωτική τάση.

Χρησιμοποιώντας την παραπάνω πληροφορία ο πελάτης θα οδηγηθεί σε περαιτέρω διερεύνηση των αιτιών που προκαλούν τις εν λόγω μακροχρόνιες τάσεις, καθώς και σε πιθανές διορθωτικές ενέργειες με σκοπό την εξομάλυνσή τους.

Μέθοδος $A3$: Κατά την μεθοδολογία αυτή θα υπολογιστεί το z-score για τις μηνιαίες τιμές του δείκτη $kw\ h/day$, με στόχο την εύρεση ακραίων τιμών κατανάλωσης.

Ο δείκτης z-score προκύπτει μέσω του τύπου:

$$z_{score} = \frac{(x - \mu^*)}{(\sigma^*/\sqrt{n})}$$

Όπου το x εκφράζει την κατανάλωση για δεδομένο μήνα και παροχή, το μ^* εκφράζει την μέση τιμή της τυχαίας μεταβλητής για την μηνιαία καταναλώση, το σ^* την διασπορά της και το n το πλήθος των καταγεγραμένων μηνιαίων καταναλώσεων.

Με σκοπό τον υπολογισμό του z-score και δεδομένου πως η πραγματική μέση τιμή και διασπορά της κατανάλωσης δεν είναι γνωστή, η παραπάνω ποσότητα θα εκτιμηθεί ως:

$$z_{score} \approx \frac{(x - \overline{x_n})}{s/\sqrt{n}}$$

Όπου $\overline{x_n}$ η αμερόληπτη εκτιμήτρια της μέσης τιμής και s η αμερόληπτη εκτιμήτρια της τυπικής απόκλισης, με τύπους:

- $$\overline{x_n} = \sum_{k=1}^{n} \frac{x_n}{n}$$
- $$s = \sqrt{\sum_{k=1}^{n} \frac{(x_n - \overline{x_n})^2}{n-1}}$$

Ακόμα, από το κεντρικό οριακό θεώρημα έχουμε ότι:

$$\frac{(\overline{X_m} - \mu)}{(\sigma/\sqrt{m})} \xrightarrow[m\to\infty]{d} N(0,1)$$

Η διαφορά μεταξύ των ποσοτήτων X και x είναι πως το πρώτο εκφράζει την κατανάλωση ως τυχαία μεταβλητή, ενώ το τελευταίο ως συγκεκριμένη τιμή.

Όπως έχουμε ήδη αναφέρει η ποσότητα $kw\ h/day$ εκφράζει την μέση ημερήσια κατανάλωση σε κιλοβατώρες και άρα μπορεί να γραφτεί στην μορφή:

$$\frac{kw\ h}{day} = \sum_{k=1}^{n} \frac{x_n}{n}$$

Επομένως για την τυχαία μεταβλητή $Kw\ H/Day$ μέσω του Κ.Ο.Θ καταλήγουμε πως:

$$\frac{(\frac{Kw\ H}{Day} - \mu)}{\sigma/\sqrt{n}} \xrightarrow[n\to\infty]{d} N(0,1)$$

Όπου μ η μέση τιμή της τυχαίας μεταβλητής $Kw\ H/Day$ και σ η διασπορά της.

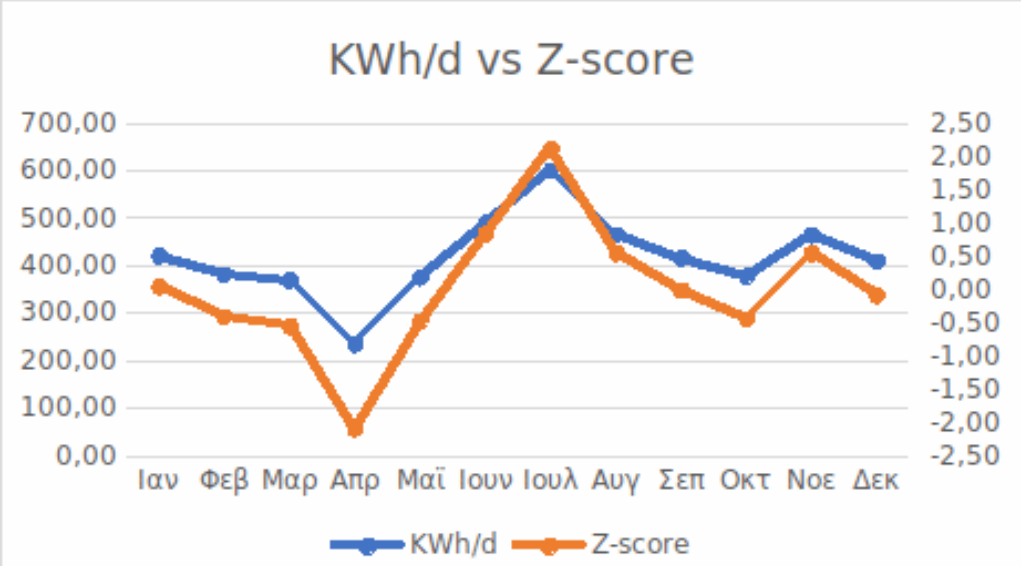
Ορίζουμε την τυχαία μεταβλητή $Z_{score} = \dfrac{(\frac{Kw\ H}{Day} - \mu)}{\sigma/\sqrt{n}}$

Έτσι για την τυχαία μεταβλητή Z_{score} και για αρκετά μεγάλο δείγμα ισχύει ότι $Z_{score} \dot{\sim} N(0,1)$

Τέλος, από την θεωρία γνωρίζουμε ότι κάθε τιμή της παραπάνω τ.μ. τ.ω $|z_{score}| > 2$ ξεπερνά το 95% των παρατηρήσεων. Για τον λόγο αυτό οι μηνιαίες τιμές του δείκτη $kw\ h/day$ για τις οποίες ικανοποιείται η παραπάνω συνθήκη, θα υποδηλώνουν ακραία κατανάλωση.

Ακολουθεί πίνακας παροχών υποθετικών δεδομένων με τις μηνιαίες τιμές των δεικτών $kw\ h/day$, z-score και διαγραμματική απεικόνιση του δείκτη $kw\ h/day$ για δεδομένη παροχή:

Μήνας	$kw\ h/day$	z-score
Ιαν	420.65	0.04
Φεβ	381,43	-0,41
Μαρ	369,03	-0,55
Απρ	235,00	-2,09
Μαϊ	375,48	-0,48
Ιουν	489,33	0,83
Ιουλ	600,00	2,11
Αυγ	464,52	0,55
Σεπ	416,00	-0,01
Οκτ	378,06	-0,45
Νοε	464,00	0,54
Δεκ	409,03	-0,09



Όπως παρατηρούμε από την διαγραμματική απεικόνιση και τον πίνακα, για τους μήνες Απρίλιο και Ιούλιο οι κιλοβατώρες έχουν σημαντική απόκλιση από την κεντρική τάση των υπόλοιπων τιμών.

Έχοντας πλέον την πληροφορία των z-score που συλλέχθηκε στο παραπάνω πίνακα για κάθε παροχή, θα διαμορφωθεί αντίστοιχος πίνακας για κάθε μήνα.

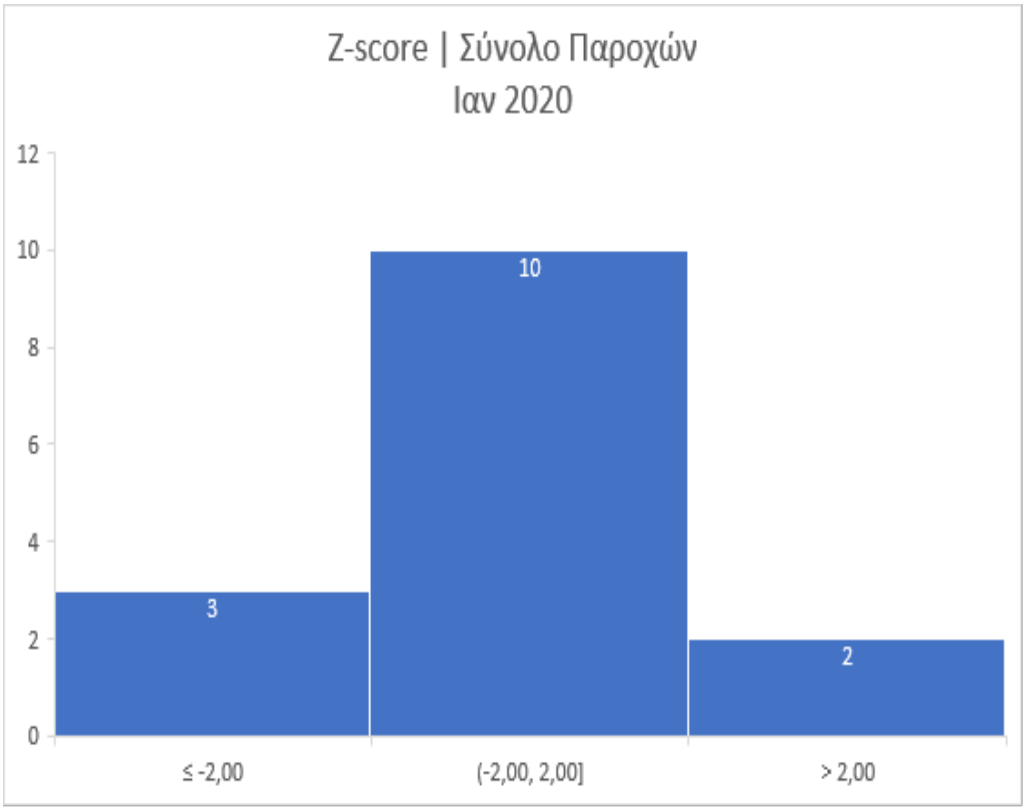
Οι τιμές z-score του πίνακα αυτού θα αναπαριστόνται σε ιστόγραμμα τρίων κλάσεων της μορφής:

$$\{z_{score} \leqslant -2\}\,,\ \{z_{score} \in (-2, 2]\}\,,\ \{z_{score} > 2\}$$

Με τον τρόπο αυτό ο πελάτης θα έχει ευκολή εποπτεία των παροχών οι οποίες παρουσιάζουν σημαντική απόκλιση, επιτρέποντάς του την διερεύνηση πιθανών αιτιών αλλά και τον σχεδιασμό δράσεων με σκοπό την εξομάλυνσή τους.

Ακολουθεί πίνακας 15 υποθετικών παροχών που περιχέχει τους δείκτες *kw h/day*, z-score για τον μήνα Γενάρη του 2020, καθώς και το αντίστοιχο ιστόγραμμα:

Αριθμός Παροχής	*kw h/day*	z-score
1	22,00	0,68
2	30,00	1,20
3	34,00	2,50
4	45,00	-0,40
5	12,00	-1,10
6	45,00	2,10
7	35,00	0,50
8	67,00	-2,22
9	87,00	-0,80
10	98,00	-2,90
11	13,00	-2,20
12	43,00	1,30
13	47,00	1,90
14	81,00	-0,88
15	18,00	2,00



Το συγκεκριμένο διάγραμμα υποδεικνύει ότι τον μήνα Γενάρη του 2020 εμφανίστηκαν 5 παροχές εκτός ορίων. Στις παροχές αυτές θα δωθεί ιδιαίτερη προτεραιτότητα, με σκοπό τον εντοπισμό πιθανών προβλημάτων.

Στην μηνιαία αναφορά προς τον πελάτη θα παραδωθεί πίνακας παροχών και ιστόγραμμα παρόμοιας μορφής με τα παραπάνω για τις παροχές με ακραία τιμή z-score.

Μέθοδος $B1$: Κατα την μεθοδολογία αυτή θα επικεντρωθούμε στην εύρεση παροχών με μηδενική κατανάλωση κατά την τελευταία ή και τις προηγούμενες περιόδους καταμέτρησης. Η συλλεγόμενη πληροφορία θα αποθηκευτεί σε πίνακα παροχών σε συνδυασμό με την αθροιστική δίαρκεια μηδενικής κατανάλωσης, το αθροιστικό κόστος και την πληροφορία νέας εισόδου στην λίστα παροχών.

Η αθροιστική διάρκεια μηδενικής κατανάλωσης θα προκύψει με άθροιση των ημερών κατανάλωσης για κάθε περίοδο με μηδενική κατανάλωση.

Το αθροιστικό κόστος θα προκύψει με άθροιση των τιμών του πεδίου ″Σύνολο τρέχοντα μήνα″ για κάθε περίοδο με μηδενική κατανάλωση. Όπως έχει ήδη εξηγηθεί το πεδίο αυτό εκφράζει το κόστος της συγκεκριμένης παροχής για τον μήνα έκδοσης του λογαριαμού. Παρότι η κατανάλωση των παροχών που μας απασχολούν είναι μηδενική, το συγκεκριμένο κόστος θα είναι μη μηδενικό, καθώς σε αυτό, συμπεριλαμβάνεται και το κόστος διαχείρησης της παροχής από τον δήμο.

Τέλος η πληροφορία νέας εισόδου θα καταχωρείται λαμβάνοντας υπόψη έαν η παροχή εντάχθηκε στην εν λόγω λίστα την τελευταία περίοδο εκκαθάρισης ή πρωτύτερα.

Σκοπός της παραπάνω διεργασίας είναι η εύρεση από τον πελάτη ανενεργών παροχών και άλλων χρήσιμων ευρυμάτων, που συσχετίζονται με την ένδειξη μηδενικής κατανάλωσης.

Ακολουθεί πίνακας 15 υποθετικών παροχών που περιέχει την παραπάνω αναφερόμενη πληροφορία καθώς και το αντίστοιχο ιστόγραμμα για την αθροιστική διάρκεια μηδενικής κατανάλωσης:

Αριθμός Παροχής	Νέα Είσοδος	Αθροιστική Διάρκεια Μηδεν.Κατανάλ.	Αθροιστικό Κόστος
1	ΌΧΙ	364 (Ημέρες)	13 €
2	ΌΧΙ	362(Ημέρες)	77 €
3	ΌΧΙ	238(Ημέρες)	8 €
4	ΌΧΙ	244(Ημέρες)	6 €
5	ΌΧΙ	363(Ημέρες)	13 €
6	ΌΧΙ	364(Ημέρες)	21 €
7	ΌΧΙ	363(Ημέρες)	12 €
8	ΌΧΙ	365(Ημέρες)	81 €
9	ΌΧΙ	364(Ημέρες)	12 €
10	ΝΑΙ	120(Ημέρες)	4 €
11	ΝΑΙ	127(Ημέρες)	5 €
12	ΌΧΙ	364(Ημέρες)	21 €
13	ΌΧΙ	237(Ημέρες)	8 €
14	ΌΧΙ	247(Ημέρες)	9 €
15	ΌΧΙ	366(Ημέρες)	14 €

Από το παραπάνω ιστόγραμμα εύκολα παρατηρούμε πως υπάρχουν 9 παροχές που παρουσιάζουν μηδενική κατανάλωση για περίοδο μεγαλύτερη του ενός έτος, οπότε πιθανώς να πρόκειται για μια μόνιμη και χρονίζουσα κατάσταση που πρέπει διερευνηθεί κατά προτεραιότητα από τον δήμο.

Ακόμα παρατηρούμε από τον πίνακα παροχών, 2 παροχές οι οποίες εντάχθηκαν για πρώτη φορά στη λίστα και 4 παροχές οι οποίες εξακολουθούν να έχουν μηδενική κατανάλωση. Ο πελάτης θα πρέπει να αξιολογήσει κατά πόσο η μηδενική κατανάλωση των 6 αυτών παροχών οφείλεται σε κάποιου είδους εποχικότητα της χρήσης τους ή αν υπάρχει αιτία που θα οδηγήσει σε χρονίζουσα κατάσταση αν δεν αντιμετωπιστεί εγκαίρως.

Στην μηνιαία αναφορά θα παραδίδεται στον δήμο πίνακας παροχών και ιστόγραμμα, παρόμοιας μορφής με τα παραπάνω.

Μέθοδος $B2$: Κατα την μεθοδολογία αυτή θα στοχεύσουμε στην εύρεση παροχών με ανεπαρκή κατανάλωση.

Όπως έχει επισημανθεί παραπάνω, όσο μεγαλύτερη η πυκνότητα των δεδομένων κατανάλωσης στον άξονα του χρόνου (άρα και του αριθμού εκκαθαριστικών λογαριασμών), τόσο ακριβέστερη η παρακολούθηση της διακύμανσης της κατανάλωσης κάτι που με την σειρά κάνει εφικτό τον έγκαιρο εντοπισμό σημαντικών-μη κανονικών μεταβολών σε αυτήν.

Υπό κανονικές συνθήκες εκδίδονται για μια παροχή κατελάχιστο 3 εκκαθαριστικοί λογαριασμοί σε ένα ημερολογιακό έτος, κάτι που συνεπάγεται, πως ο μέγιστος αριθμός ημερών μεταξύ των ημερομηνιών έκδοσης δύο διαδοχικών εκκαθαριστικών λογαριασμών θα πρέπει να είναι $365/3 \approx 123$ ημέρες.

Για τον λόγω αυτό όταν για μια παροχή δεν έχει εκδοθεί εκκαθαριστικός λογαριασμός για διάστημα που υπερβαίνει τις 130 ημέρες, αυτό θα αποτελεί ένδειξη ανεπαρκούς εκκαθάρισης κατανάλωσης.

Παρακάτω παρατίθενται υποθετικά δεδομένα μιας παροχής τιμολογίου Γ21:

Ημερ.Τελ.Καταμετ.	Ημερ.Προ.Καταμετ	Ημερ. Καταναλ.	Καταν.Ενεγ.(ΩΧΒ)	Συνολ.Τρεχ.Μήνα	Τύπος Λογ.
28/2/2019	27/10/2018	125	4969	575	ΕΚΚΑΘ
25/4/2019	1/3/2019	56	1216	112	ΕΝΑΝΤ
29/6/2019	26/4/2019	65	1411	229	ΕΝΑΝΤ
28/8/2019	30/6/2019	60	1750	281	ΕΝΑΝΤ
26/10/2019	1/3/2019	240	8298	907	ΕΚΚΑΘ
26/12/2019	27/10/2019	61	2425	244	ΕΝΑΝΤ

Το πρόβλημα εντοπίζεται στην 3η διαδοχική έναντι τιμολόγηση για το διάστημα 30/6/2019 – 28/8/2019, μέχρι την οποία το άθροισμα του πεδίου ″Ημέρες Κατανάλωσης″ για τους διαδοχικούς έναντι λογαριασμούς ανέρχεται στις 181 ημέρες.

Αφού ο αριθμός αυτός ξεπερνά τις 130 ημέρες, η παροχή θα ενταχθεί στην πίνακα παροχών ανεπαρκείς εκκαθάρισης σε συνδυασμό με την πληροφορία νέας εισόδου αλλά και την αθροιστική διάρκεια διαδοχικών έναντι λογαριασμών, μετρημένη σε ημέρες.

Στο πεδίο νέας εισόδου θα καταχωρείται η τιμή ″ΝΑΙ″ σε περίπτωση που στον προηγούμενο διαδοχικό έναντι λογαριασμό η αθροιστική διάρκεια έχει ξεπεράσει τις 130 ημέρες και αντιθέτως θα καταχωρείται η τιμή ″ΟΧΙ″.Στην ειδική περίπτωση που η αθροιστική διάρκεια ξεπερνά τις 130 ημερες από τον πρώτο έναντι λογα ριασμό στο πεδίο νέας εισόδου θα καταχωρείται η τιμή ″ΝΑΙ″.

Αξίζει να αναφέρουμε εδω, πως η συγκεκριμένη μέθοδος αποτελεί η μόνη στην οποία λαμβάνουμε υπόψη την πληροφορία που μας προσφέρεται από τους έναντι λογαριασμούς.

Με την ένταξη της παροχής, ο πίνακας θα έχει την παρακάτω μορφή:

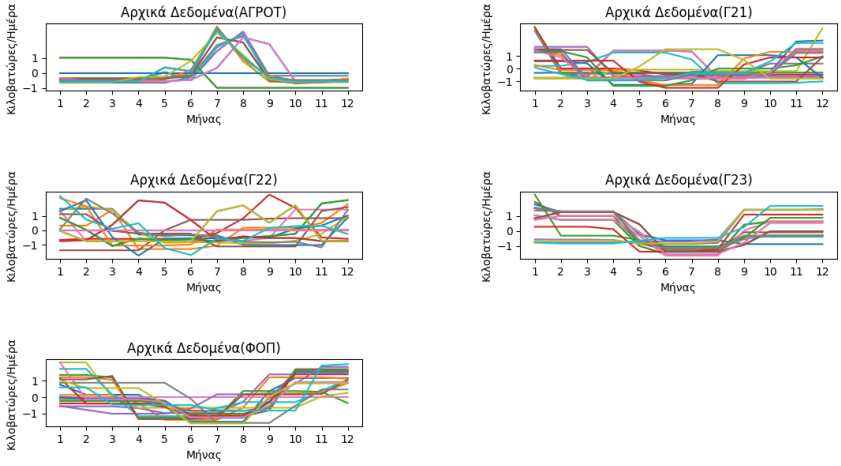
Αρ.Παροχής	Αθροιστ.Διάρκεια Διαδοχ.Έναντι Λογαρ.	Νέα Είσοδος
1	181	ΝΑΙ

Στην μηνιαία αναφορά προς τον Δήμο θα παραδίδεται ο παραπάνω πίνακας παροχών.

Μέθοδος $B3$: Κατα την μεθοδολογία αυτή θα επικεντρωθούμε στην εύρεση κοινού μοτίβου κατανάλωσης για παροχές ίδιου τιμολογίου.

Η αρχική υπόθεση επί της οποίας βασίζεται η μέθοδος είναι, πως οι παροχές ίδιου τιμολογίου παρουσιάζουν ένα κοινό μοτίβο κατανάλωσης στα πλαίσια ενός ημερολογιακού έτους.
Το μοτίβο αυτό θα παρουσιαστεί παρακάτω, μέσω της αναπαράστασης των ετήσιων χρονοσειρών κατανάλωσης, κάθε παροχής δεδομένου τιμολογίου, σε κοινή γραφική απεικόνιση.

Παρότι το μοτίβο των παροχών ίδιου τιμολογίου αναμένετεται να είναι παρόμοιο, η κλίμακα της κατανάλωσης για κάθε παροχή διαφέρει εμφανώς. Επομένως, επιλέγουμε για διαισθητικούς λόγους να κανονικοποιήσουμε την μέση τιμή και διασπορά της κάθε χρονοσειράς.



Όπως παρατηρούμε, από την παραπάνω απεικόνιση παρουσίαζεται πράγματι ένα κοινό μοτίβο για παροχές ίδιου τιμολογίου. Συγκεκριμένα για παροχές τιμολογίων ΑΓΡΟΤ, Γ23 το μοτίβο κατανάλωσης εμφανίζεται χωρίς σημαντικές μετατοπίσεις μεταξύ των παροχών στον άξονα του χρόνου. Αντίθετα, για παροχές των υπολοίπων τιμολογίων παρότι η ομοιότητα μοτίβου είναι εμφανής, η θέση των σημείων καμπής στον άξονα του χρόνου, διαφέρει ανά παροχή.

Αξίζει να αναφέρουμε, πως για τα δύο τιμολόγια στα οποία το μοτίβο δεν παρουσιάζει σημαντικές μετατοπίσεις μεταξύ παροχών, η καταμέτρηση της κατανάλωσης έχει γίνει μέσω τηλεμέτρησης. Η συμπεριφορά αυτή δικαιολογείται, από την χρονική ακρίβεια που χαρακτηρίζει τις ηλεκτρονικές καταμετρήσεις.

Συγκεκριμένα, η μηνιαία καταμέτρηση της κατανάλωσης επιτρέπει την ακριβή τοποθέτηση των ετήσιων αυξομειώσεων στον άξονα του χρόνου,προσφέρωντας μια αντιπροσωπευτική εικόνα του πραγματικού μοτίβου, για κάθε παροχή.

Στο σημείο αυτό, θα ερμηνεύσουμε τα μοτίβα που παρουσιάζονται για κάθε τιμολόγιο:

- ΑΓΡΟΤ:

  Στην περίπτωση του αγροτικού τιμολογίου παρατηρούμε μια σταθερά μειωμένη κατανάλωση κατά τους μήνες του φθινοπόρου και του χειμώνα. Από την άλλη παρατηρούμε μια σημαντική αύξηση της κατανάλωσης κατά την άνοιξη και το καλοκαίρι.

Το μοτίβο αυτό ερμηνεύεται από την αυξημένη δραστηριότητα των αγροτών τους θερμότερους μήνες, κατά τους οποίους ευδοκιμούν οι πειρσσότερες καλλιέργειες.

- Γ22:

  Στο τιμολόγιο τύπου Γ22 παρατηρούμε μια κυματοειδή καμπύλη κατανάλωσης, η οποία ακολουθεί της κλιματικές αλλαγές μεταξύ των εποχών ενός έτους. Συγκεκριμένα κατά το πρώτο τετράμηνο του έτους, δεν παρουσιάζεται αναγκαιότητα χρήσης κλιματιστικών μονάδων, λόγω της σχετικά χαμηλής θερμοκρασίας, κατί που οδηγεί στην μείωση της κατανάλωσης.

  Στη συνέχεια για το δεύτερο τετράμηνο του έτους, παρατηρούμε μια αύξηση της κατανάλωσης που όπως και πριν αναμένεται να συσχετίζεται με την αντίστοιχη αύξηση της θερμοκρασίας. Το μοτίβο αυτό συνεχίζεται μέχρι το τέλος του έτους.

- ΦΟΠ:

  Για το τιμολόγιο που αφορά δημόσιο φωτισμό παρατηρούμε μια καμπυλοειδή κατανάλωση που ελαχιστοποιείται κατά τους καλοκαιρινούς μήνες. Η συμπεριφορά αυτή εξηγείται συσχετίζοντας τον οδοφωτισμό με τις αυξομειώσεις των ωρών της νύχτας μεταξύ των μηνών ενός έτους.

  Συγκεκριμένα παρατηρούμε πως για το πρώτο εξάμηνο του έτους η κατανάλωση ακολουθεί πτωτική τροχία, ενώ για το δεύτερο ανωδική, σε συσχέτιση με την καμπύλη των ωρών της νύχτας ανά μήνα.

  Οι επίπεδες περιοχές της καμπύλης οφείλονται στις τετράμηνες περιόδους εκκαθάρισης που συνηθίζονται για παροχές τύπου ΦΟΠ. Αν υπήρχαν ετησίως 12 μηνιαία δεδομένα κατανάλωσης, δηλαδή καταμέτρηση της κατανάλωσης για κάθε μήνα ξεχωριστά, αναμένουμε ταύτιση των δύο προαναφερόμενων καμπυλών.

- Γ23:

  Για το τιμολόγιο αυτό παρατηρείται μοτίβο κατανάλωσης παρόμοιο με το μοτίβο οδοφωτισμού. Η συμπεριφορά αυτή είναι μη αναμενώμενη, με πιθανή ερμηνεία την σκόπιμη δήλωση παροχών οδοφωτισμού ως παροχές Γ23. Όπως έχουμε ήδη εξηγήσει η παροχές τύπου Γ23 τιμολογούνται διαφορετικά για της πρωινές και νυχτερινές ώρες της ημέρας, με την νυχτερινή τιμολόγηση να αποτελεί την φθηνότερη από τις δύο.

  Με σκοπό λοιπόν την οικονομικότερη χρέωση λαμπτήρων δημοσίου φωτισμού, πιθανολογούμε σκόπιμη από τον δήμο δήλωση των εν λόγω παροχών ως τύπου ΦΟΠ.

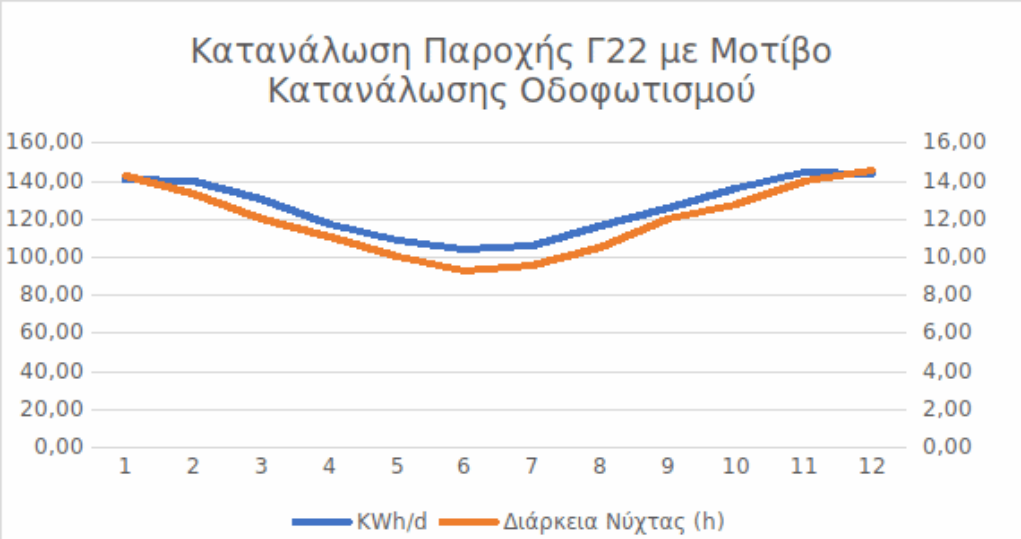
- Γ21:

  Όπως και παραπάνω, για το τιμολόγιο αυτό παρουσιάζεται μοτίβο παρόμοιο με παροχές τύπου ΦΟΠ. Η συμπεριφορά αυτή είναι μη αναμεωμενη και η αιτιολόγησή της μη εμφανής.

Για παροχές με τύπο τιμολογίου Γ21, Γ23 αναμένουμε μοτίβο κατανάλωσης παρόμοιο με αυτό που εμφανίζεται για παροχές τύπου Γ22. Η απόκλιση από την θεωρητικά αναμενώμενη συμπεριφορά θα αναφερθεί στον δήμο σε συνδυασμό με τα υπόλοιπα αρχικά αποτελέσματα που θα παραθέσουμε παρακάτω.

Γενικότερα, αποκλίσεις από το αναμενόμενο μοτίβο της καμπύλης κατανάλωσης ενός τύπου τιμολογίου χωρίς επαρκή αιτιολόγηση, αποτελούν ενδεχομένως αξιοποιήσιμα ευρήματα.

Για παράδειγμα, στο παρακάτω υποθετικό διάγραμμα καταναλώσεων μιας παροχής Γ22 η καμπύλη φαίνεται να αποκλίνει από το αναμενόμενο μοτίβο μιας παροχής Γ22, όπως αυτό παρουσιάστηκε παραπάνω, ενώ είναι πιο κοντά στο μοτίβο μιας παροχής οδοφωτισμού(ΦΟΠ).



Σκοπός λοιπόν της συγκεκριμένης μεθόδου είναι η αναγνώριση των διαφορετικών μοτίβων κατανάλωσης, αλλά και η συλλογή παροχών που αποκλίνουν από αυτά.

Με χρήση μοντέλου μηχανικής μάθησης, θα οριστεί το αναμενόμενο μοτίβο για παροχές που ανήκουν στο ίδιο τιμολόγιο και θα απομονωθούν όλες οι παροχές των οποίων το πραγματικό μοτίβο αποκλίνει από το αναμενόμενο. Οι παροχές αυτές θα ενταχθούν σε πίνακα παροχών, σε συνδυασμό με τα πεδία ″δηλωμενο τιμολόγιο″ και ″προβλεμπόμενο τιμολόγιο″.
Στη συνέχεια θα διαμορφώνεται το γράφημα με την ετήσια κατανάλωση, των παροχών του πίνακα,ώστε να εξεταστεί η πιθανή αντιστοίχισή τους με το προβλεπόμενο για αυτές τιμολόγιο.

Το παραδοτέο στην αναφορά προς τον δήμο, θα είναι ο σχετικός πίνακας παροχών
και οι καμπύλες κατανάλωσής τους. Σε αντίθεση με τις προηγούμενες μεθόδους, η αναφορά για την συγκεκριμενη μέθοδο θα παραδίδεται ετησίως, καθώς είναι απαραίτητη για την υλοποίηση της η καταγραφή της κατανάλωσης και για τους δώδεκα μήνες ανά παροχή.

Έχοντας πλέον καταγράψει τις μεθοδολογίες ανάλυσης που θα υλοποιηθούν στην συνέχεια, οδηγούμαστε στην συνοπτική παρουσίαση αυτών σε συνδυασμό με τα πιθανά ευρήματα και οφέλη τους.

Πιθανό όφελος ανάλυσης		
Όνομα μεθόδου	Πιθανά ευρήματα	Πιθανό όφελος
A) Ανάλυση ανά παροχή στον άξονα του χρόνου		
• A1 • A1	Σημαντική ποσοστιαία διαφορά μεταξύ ενός μήνα και του αντίστοιχου περσινού Σημαντική ποσοστιαία διαφορά μεταξύ ενός μήνα και του αντίστοιχου περσινού	Έγκαιρη αντιμετώπιση προβληματικής κατάστασης Επαλήθευση αποτελεσματικότητας μέτρων περιστολής
• A2 • A2	Ύπαρξη μακροχρόνιας καταναλωτικής τάσης Ύπαρξη μακροχρόνιας καταναλωτικής τάσης	Έγκαιρη αντιμετώπιση προβληματικής κατάστασης Επαλήθευση αποτελεσματικότητας μέτρων περιστολής
• A3 • A3	Ακραία τιμή κατανάλωσης σε σχέση με την κεντρική τάση Ακραία τιμή κατανάλωσης σε σχέση με την κεντρική τάση	Έγκαιρη αντιμετώπιση προβληματικής κατάστασης Επαλήθευση αποτελεσματικότητας μέτρων περιστολής
B) Ανάλυση μεταξύ παροχών σε κοινό χρόνο		
• B1 • B1 • B1	Ανενεργείς παροχές και σχετική δαπάνη Ανενεργείς παροχές και σχετική δαπάνη Ανενεργείς παροχές και σχετική δαπάνη	Εκκαθ. της κατανάλ. για την περίοδο πιθανής βλάβης Εξοικονόμηση από την πιθανή κατάργηση ανενεργής παροχής Επίλυση πιθανής βλάβης του μετρητή ενέργειας
• B2 • B2	Παροχές με ανεπαρκή ετήσια εκκαθάριση Παροχές με ανεπαρκή ετήσια εκκαθάριση	Αποφυγή συσσωρ. καταναλ. λόγω εκτενής περιόδου μη εκκαθ. Εμπλουτισμός ετήσιων δεδομένων εκκαθάρισης
• B3 • B3 • B3	Πραγματικό μοτίβο κατανάλωσης δεδομένου τιμολογίου Σημαντ. απόκλ. μεταξύ αναμενόμ./πραγματ. μοτίβου παροχής δεδομ. τιμολ. Σημαντ. απόκλ. μεταξύ αναμενόμ./πραγματ. μοτίβου παροχής δεδομ. τιμολ.	Κατάρτηση σχεδίων εξοικονόμησης ενέργειας ανά τιμολόγιο Οικονομικό όφελος από την αλλαγή τιμολογίου παροχής Καταγραφή πραγματικού πλήθους παροχών ανά τιμολόγιο

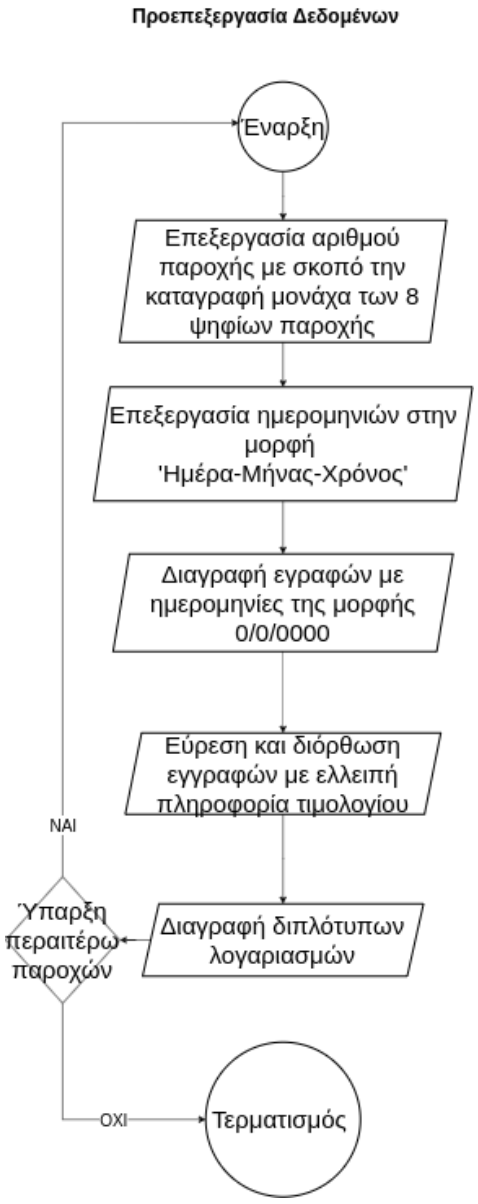
3.4 Αλγοριθμική ροή μεθόδων

Πρώτου παραθέσουμε τα τελικά αποτελέσματα θα αναπαραστήσουμε γραφικά την ροή των παραπάνω μεθόδων. Η διαγραμματική ροή εκφράζει σχηματικά τις διεργασίες που υλοποιήθηκαν, με σκοπό την επίτευξη της ανάλυσης της δεδομένης μεθόδου.

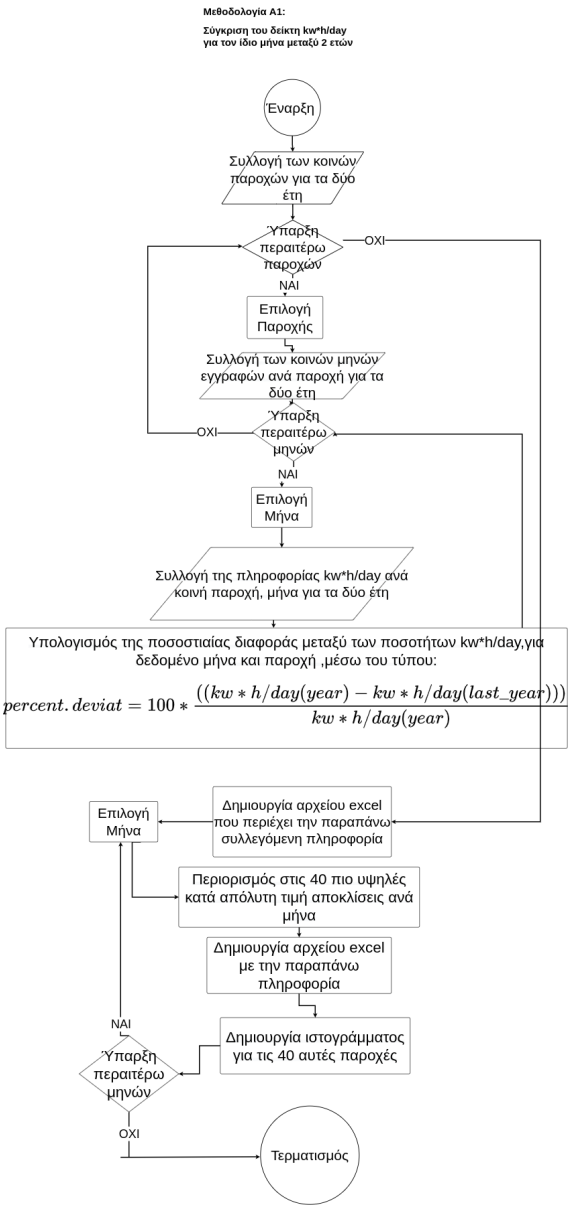
Τα σχήματα που θα χρησιμοποιηθούν και οι σημασία τους δίνεται παρακάτω:

- Κύκλος ή Έλλειψη: Έναρξη/Τερματισμός αλγορίθμου
- Ορθογώνιο Παραλληλόγραμμο: Εκτέλεση διαδικασίας
- Παραλληλόγραμμο: Εισαγωγή Δεδομένων

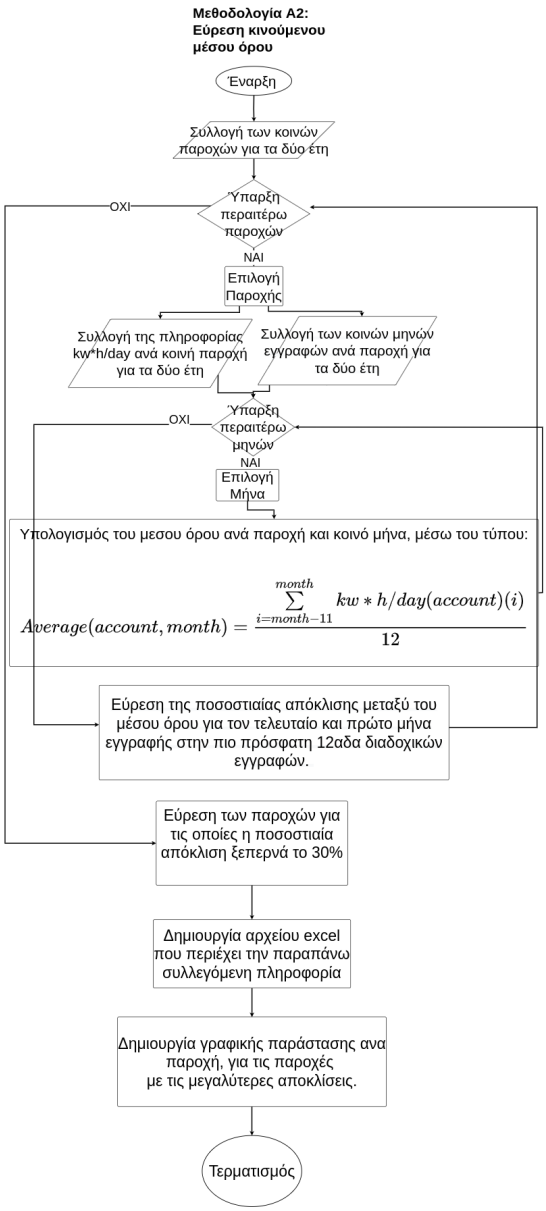
- Προεπεξεργασία:



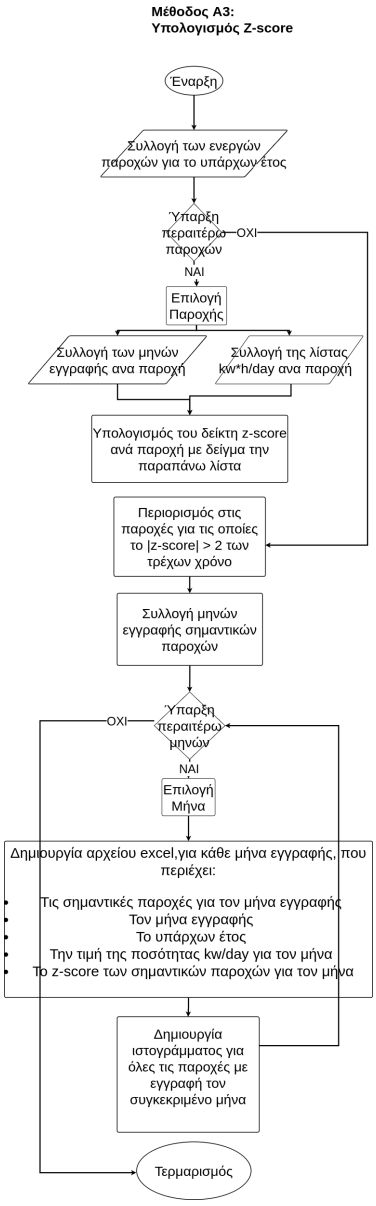
- Μέθοδος Α1:


Μεθοδολογία Α1:
Σύγκριση του δείκτη kw*h/day
για τον ίδιο μήνα μεταξύ 2 ετών
Έναρξη
Συλλογή των κοινών παροχών για τα δύο έτη
Ύπαρξη περαιτέρω παροχών
ΟΧΙ
ΝΑΙ
Επιλογή Παροχής
Συλλογή των κοινών μηνών εγγραφών ανά παροχή για τα δύο έτη
Ύπαρξη περαιτέρω μηνών
ΟΧΙ
ΝΑΙ
Επιλογή Μήνα
Συλλογή της πληροφορίας kw*h/day ανά κοινή παροχή, μήνα για τα δύο έτη
Υπολογισμός της ποσοστιαίας διαφοράς μεταξύ των ποσοτήτων kw*h/day,για δεδομένο μήνα και παροχή ,μέσω του τύπου:
$percent.deviat = 100 * \frac{((kw*h/day(year) - kw*h/day(last\_year)))}{kw*h/day(year)}$
Δημιουργία αρχείου excel που περιέχει την παραπάνω συλλεγόμενη πληροφορία
Επιλογή Μήνα
Περιορισμός στις 40 πιο υψηλές κατά απόλυτη τιμή αποκλίσεις ανά μήνα
Δημιουργία αρχείου excel με την παραπάνω πληροφορία
Δημιουργία ιστογράμματος για τις 40 αυτές παροχές
ΝΑΙ
Ύπαρξη περαιτέρω μηνών
ΟΧΙ
Τερματισμός

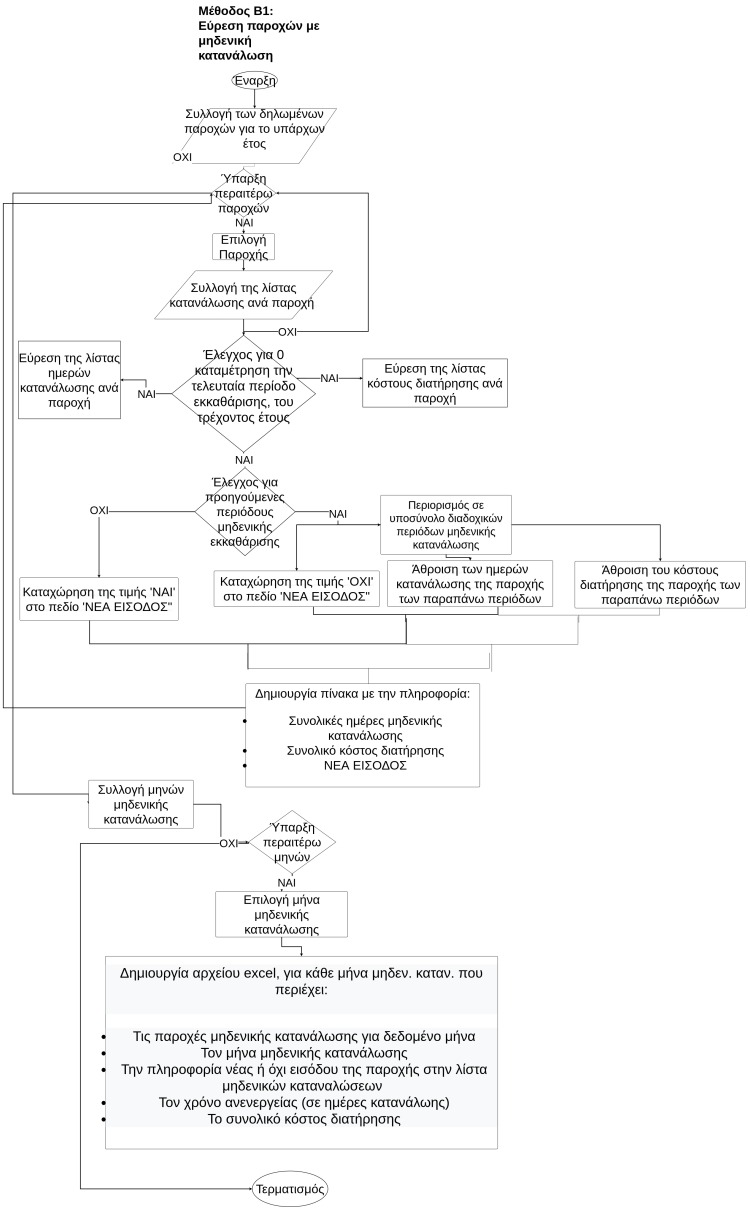
- Μέθοδος Α2:


Μεθοδολογία Α2:
Εύρεση κινούμενου
μέσου όρου
Έναρξη
Συλλογή των κοινών
παροχών για τα δύο έτη
Ύπαρξη
περαιτέρω
παροχών
ΟΧΙ
ΝΑΙ
Επιλογή
Παροχής
Συλλογή της πληροφορίας
kw*h/day ανά κοινή παροχή
για τα δύο έτη
Συλλογή των κοινών μηνών
εγγραφών ανά παροχή για
τα δύο έτη
Ύπαρξη
περαιτέρω
μηνών
ΟΧΙ
ΝΑΙ
Επιλογή
Μήνα
Υπολογισμός του μεσου όρου ανά παροχή και κοινό μήνα, μέσω του τύπου:
$$Average(account, month) = \frac{\sum_{i=month-11}^{month} kw * h/day(account)(i)}{12}$$
Εύρεση της ποσοστιαίας απόκλισης μεταξύ του
μέσου όρου για τον τελευταίο και πρώτο μήνα
εγγραφής στην πιο πρόσφατη 12αδα διαδοχικών
εγγραφών.
Εύρεση των παροχών για
τις οποίες η ποσοστιαία
απόκλιση ξεπερνά το 30%
Δημιουργία αρχείου excel
που περιέχει την παραπάνω
συλλεγόμενη πληροφορία
Δημιουργία γραφικής παράστασης ανα
παροχή, για τις παροχές
με τις μεγαλύτερες αποκλίσεις.
Τερματισμός

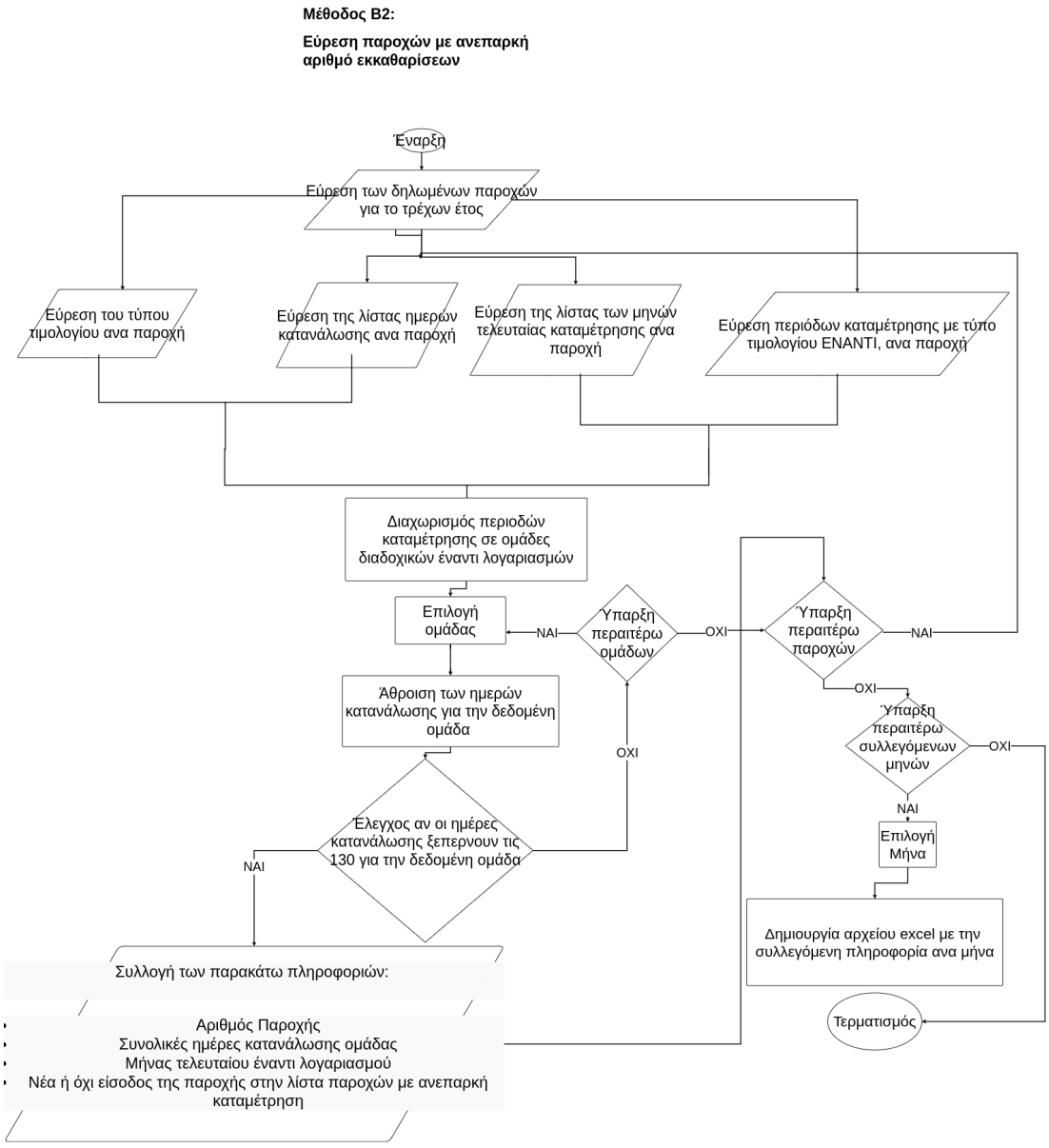
- Μέθοδος Α3:


Μέθοδος Α3:
Υπολογισμός Z-score
Έναρξη
Συλλογή των ενεργών παροχών για το υπάρχων έτος
Υπάρξη περαιτέρω παροχών
ΟΧΙ
ΝΑΙ
Επιλογή Παροχής
Συλλογή των μηνών εγγραφής ανα παροχή
Συλλογή της λίστας kw*h/day ανα παροχή
Υπολογισμός του δείκτη z-score ανά παροχή με δείγμα την παραπάνω λίστα
Περιορισμός στις παροχές για τις οποίες το |z-score| > 2 των τρέχων χρόνο
Συλλογή μηνών εγγραφής σημαντικών παροχών
Υπάρξη περαιτέρω μηνών
ΟΧΙ
ΝΑΙ
Επιλογή Μήνα
Δημιουργία αρχείου excel,για κάθε μήνα εγγραφής, που περιέχει:
Τις σημαντικές παροχές για τον μήνα εγγραφής
Τον μήνα εγγραφής
Το υπάρχων έτος
Την τιμή της ποσότητας kw/day για τον μήνα
Το z-score των σημαντικών παροχών για τον μήνα
Δημιουργία ιστογράμματος για όλες τις παροχές με εγγραφή τον συγκεκριμένο μήνα
Τερμαρισμός

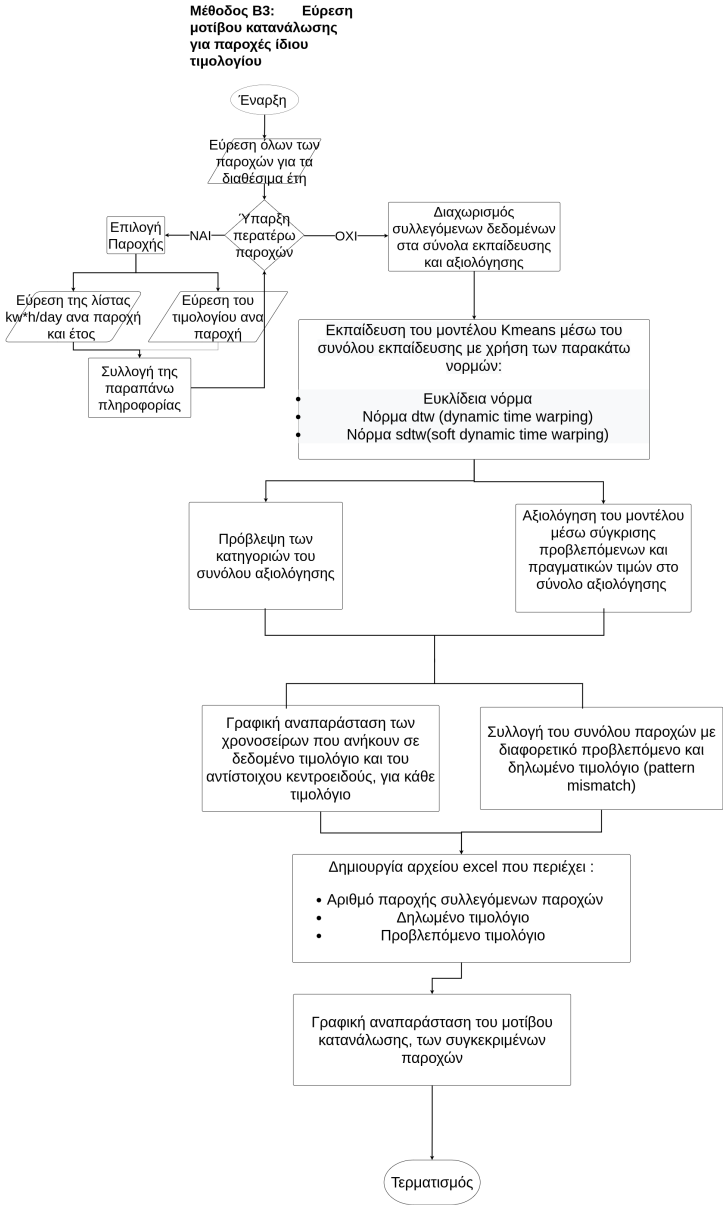
- Μέθοδος Β1:


Μέθοδος Β1: Εύρεση παροχών με μηδενική κατανάλωση
Έναρξη
Συλλογή των δηλωμένων παροχών για το υπάρχων έτος
ΟΧΙ
Ύπαρξη περαιτέρω παροχών
ΝΑΙ
Επιλογή Παροχής
Συλλογή της λίστας κατανάλωσης ανά παροχή
ΟΧΙ
Έλεγχος για 0 καταμέτρηση την τελευταία περίοδο εκκαθάρισης, του τρέχοντος έτους
ΝΑΙ
Εύρεση της λίστας κόστους διατήρησης ανά παροχή
ΝΑΙ
Εύρεση της λίστας ημερών κατανάλωσης ανά παροχή
ΝΑΙ
Έλεγχος για προηγούμενες περιόδους μηδενικής εκκαθάρισης
ΝΑΙ
Περιορισμός σε υποσύνολο διαδοχικών περιόδων μηδενικής κατανάλωσης
ΟΧΙ
Καταχώρηση της τιμής 'ΝΑΙ' στο πεδίο 'ΝΕΑ ΕΙΣΟΔΟΣ"
Καταχώρηση της τιμής 'ΟΧΙ' στο πεδίο 'ΝΕΑ ΕΙΣΟΔΟΣ"
Άθροιση των ημερών κατανάλωσης της παροχής των παραπάνω περιόδων
Άθροιση του κόστους διατήρησης της παροχής των παραπάνω περιόδων
Δημιουργία πίνακα με την πληροφορία:
Συνολικές ημέρες μηδενικής κατανάλωσης
Συνολικό κόστος διατήρησης
ΝΕΑ ΕΙΣΟΔΟΣ
Συλλογή μηνών μηδενικής κατανάλωσης
ΟΧΙ
Ύπαρξη περαιτέρω μηνών
ΝΑΙ
Επιλογή μήνα μηδενικής κατανάλωσης
Δημιουργία αρχείου excel, για κάθε μήνα μηδεν. καταν. που περιέχει:
Τις παροχές μηδενικής κατανάλωσης για δεδομένο μήνα
Τον μήνα μηδενικής κατανάλωσης
Την πληροφορία νέας ή όχι εισόδου της παροχής στην λίστα μηδενικών καταναλώσεων
Τον χρόνο ανενεργείας (σε ημέρες κατανάλωσης)
Το συνολικό κόστος διατήρησης
Τερματισμός

- Μέθοδος Β2:


Μέθοδος Β2:
Εύρεση παροχών με ανεπαρκή αριθμό εκκαθαρίσεων
Έναρξη
Εύρεση των δηλωμένων παροχών για το τρέχων έτος
Εύρεση του τύπου τιμολογίου ανα παροχή
Εύρεση της λίστας ημερών κατανάλωσης ανα παροχή
Εύρεση της λίστας των μηνών τελευταίας καταμέτρησης ανα παροχή
Εύρεση περιόδων καταμέτρησης με τύπο τιμολογίου ΕΝΑΝΤΙ, ανα παροχή
Διαχωρισμός περιοδών καταμέτρησης σε ομάδες διαδοχικών έναντι λογαριασμών
Επιλογή ομάδας
ΝΑΙ
Υπαρξη περαιτέρω ομάδων
ΟΧΙ
Ύπαρξη περαιτέρω παροχών
ΝΑΙ
ΟΧΙ
Ύπαρξη περαιτέρω συλλεγόμενων μηνών
ΟΧΙ
ΝΑΙ
Επιλογή Μήνα
Άθροιση των ημερών κατανάλωσης για την δεδομένη ομάδα
ΟΧΙ
Έλεγχος αν οι ημέρες κατανάλωσης ξεπερνουν τις 130 για την δεδομένη ομάδα
ΝΑΙ
Δημιουργία αρχείου excel με την συλλεγόμενη πληροφορία ανα μήνα
Τερματισμός
Συλλογή των παρακάτω πληροφοριών:
Αριθμός Παροχής
Συνολικές ημέρες κατανάλωσης ομάδας
Μήνας τελευταίου έναντι λογαριασμού
Νέα ή όχι είσοδος της παροχής στην λίστα παροχών με ανεπαρκή καταμέτρηση

- Μέθοδος Β3:


Μέθοδος Β3: Εύρεση μοτίβου κατανάλωσης για παροχές ίδιου τιμολογίου
Έναρξη
Εύρεση όλων των παροχών για τα διαθέσιμα έτη
Υπαρξη περατέρω παροχών
ΝΑΙ
ΟΧΙ
Επιλογή Παροχής
Διαχωρισμός συλλεγόμενων δεδομένων στα σύνολα εκπαίδευσης και αξιολόγησης
Εύρεση της λίστας kw*h/day ανα παροχή και έτος
Εύρεση του τιμολογίου ανα παροχή
Συλλογή της παραπάνω πληροφορίας
Εκπαίδευση του μοντέλου Kmeans μέσω του συνόλου εκπαίδευσης με χρήση των παρακάτω νορμών:
Ευκλίδεια νόρμα
Νόρμα dtw (dynamic time warping)
Νόρμα sdtw(soft dynamic time warping)
Πρόβλεψη των κατηγοριών του συνόλου αξιολόγησης
Αξιολόγηση του μοντέλου μέσω σύγκρισης προβλεπόμενων και πραγματικών τιμών στο σύνολο αξιολόγησης
Γραφική αναπαράσταση των χρονοσείρων που ανήκουν σε δεδομένο τιμολόγιο και του αντίστοιχου κεντροειδούς, για κάθε τιμολόγιο
Συλλογή του συνόλου παροχών με διαφορετικό προβλεπόμενο και δηλωμένο τιμολόγιο (pattern mismatch)
Δημιουργία αρχείου excel που περιέχει :
Αριθμό παροχής συλλεγόμενων παροχών
Δηλωμένο τιμολόγιο
Προβλεπόμενο τιμολόγιο
Γραφική αναπαράσταση του μοτίβου κατανάλωσης, των συγκεκριμένων παροχών
Τερματισμός

4 Τελικά αποτελέσματα

4.1 Τελικά αποτελέσματα

Στο σημείο αυτό παραθέτουμε τα τελικά αποτελέσματα της ανάλυσης μας.
Τα αποτελέσματα θα παρουσιαστούν ξεχωριστά για κάθε μεθοδολογία σε συνδυασμό με πιθανές ερμηνίες τους. Τέλος, θα προτείνουμε τρόπους επέκτασης των μεθοδολογιών, με σκοπό την καλύτερη επίτευξη των αποτελεσμάτων αλλά και την ανακάλυψη περαιτέρω ευρημάτων.

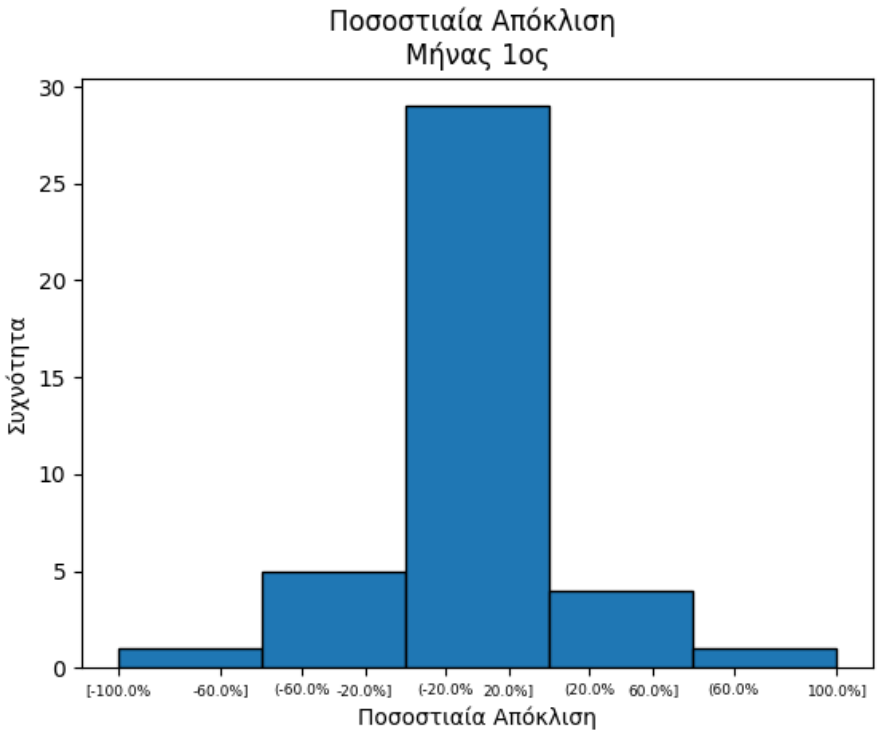
Όπως έχει εξηγηθεί και παραπάνω τα αποτελέσματα των μεθόδων μας είναι μηνιαία, δημιουργώντας έτσι, ειδικά σε μεθόδους ανάλυσης ανά παροχή, ένα εκτενές σύνολο αποτελεσμάτων. Για τον λόγο αυτό, η παρουσίαση θα είναι περιληπτική, με επίκεντρο χαρακτηριστικές περιπτώσεις.

- Μέθοδος A1: Όπως έχουμε ήδη αναφέρει τα αποτελέσματα για αυτή την μεδόθου αποτελούν ο πίνακας παροχών με τις σημαντικότερες αποκλίσεις για κάθε μήνα, καθώς και το ιστόγραμμα των αποκλίσεων τους. Το ιστόγραμμα αναμένουμε να προσομοιάζει την σ.π.π. της κανονικής κατανομής, με τις περισσότερες αποκλίσεις να βρίσκονται στην ομάδα που περιέχει το 0.

  Παρακάτω παραθέτουμε τρία ιστογράματα σε συνδυασμό με τους αντίστοιχους πίνακες για τον εν λόγω μήνα.

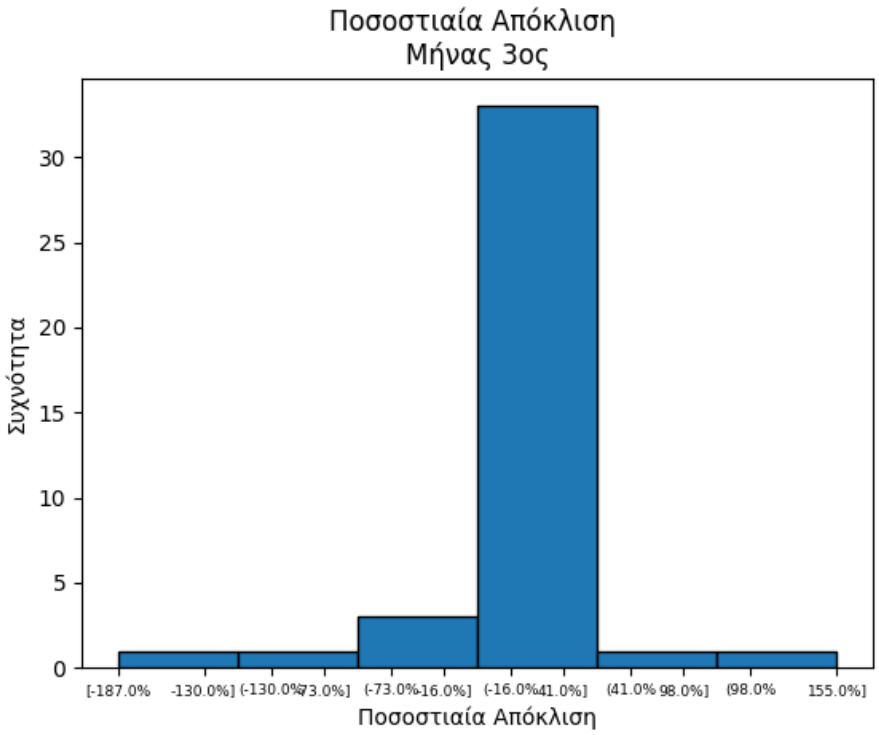
  – Ιανουάριος 2017-2018

Account	Period	Month	Kw/day(2017)	Kw/day(2018)	Deviation
41046018	2017-2018	1	18.420	18.100	-1.769
41058306	2017-2018	1	27.906	35.361	21.083
41089028	2017-2018	1	17.908	17.958	0.283
41058310	2017-2018	1	9.159	8.350	-9.693
41058312	2017-2018	1	22.000	24.864	11.519
41058313	2017-2018	1	21.084	22.624	6.807
41082889	2017-2018	1	1.058	0.000	-100
41078795	2017-2018	1	18.824	21.725	13.355
41029644	2017-2018	1	9.807	10.892	9.961
41078796	2017-2018	1	5.017	5.500	8.785
41089033	2017-2018	1	7.466	7.025	-6.274
41078799	2017-2018	1	15.459	16.145	4.25
41046032	2017-2018	1	28.504	27.860	-2.314
41058320	2017-2018	1	4.402	5.000	11.966
41056274	2017-2018	1	0.000	0.000	0
41074706	2017-2018	1	4.008	4.057	1.209
41043988	2017-2018	1	64.992	59.085	-9.997
41058326	2017-2018	1	0.000	0.000	0
41029656	2017-2018	1	0.000	12.717	100
41043994	2017-2018	1	4.900	4.737	-3.435
41076706	2017-2018	1	49.891	34.443	-44.852
41076707	2017-2018	1	5.538	5.484	-0.989
41076709	2017-2018	1	8.807	6.262	-40.631
41043945	2017-2018	1	8.908	7.339	-21.373
41043946	2017-2018	1	9.496	7.788	-21.926
41029611	2017-2018	1	48.723	48.200	-1.084
41043947	2017-2018	1	24.832	24.110	-2.994
41062378	2017-2018	1	21.449	22.244	3.572
41015278	2017-2018	1	17.835	20.933	14.802
41005039	2017-2018	1	4.957	9.757	49.191
41046000	2017-2018	1	16.345	20.927	21.897
41070574	2017-2018	1	32.408	21.959	-47.586
41078765	2017-2018	1	0.000	0.000	0
41089007	2017-2018	1	4.756	4.508	-5.503
41082868	2017-2018	1	12.909	21.008	38.552
41017334	2017-2018	1	0.000	0.000	0
41082870	2017-2018	1	0.000	0.000	0
41078776	2017-2018	1	45.210	40.492	-11.653
41070585	2017-2018	1	38.684	36.029	-7.371
41005051	2017-2018	1	33.067	36.227	8.724


Ποσοστιαία Απόκλιση
Μήνας 1ος
Συχνότητα
30
25
20
15
10
5
0
[-100.0%
-60.0%]
(-60.0%
-20.0%]
(-20.0%
20.0%]
(20.0%
60.0%]
(60.0%
100.0%]
Ποσοστιαία Απόκλιση

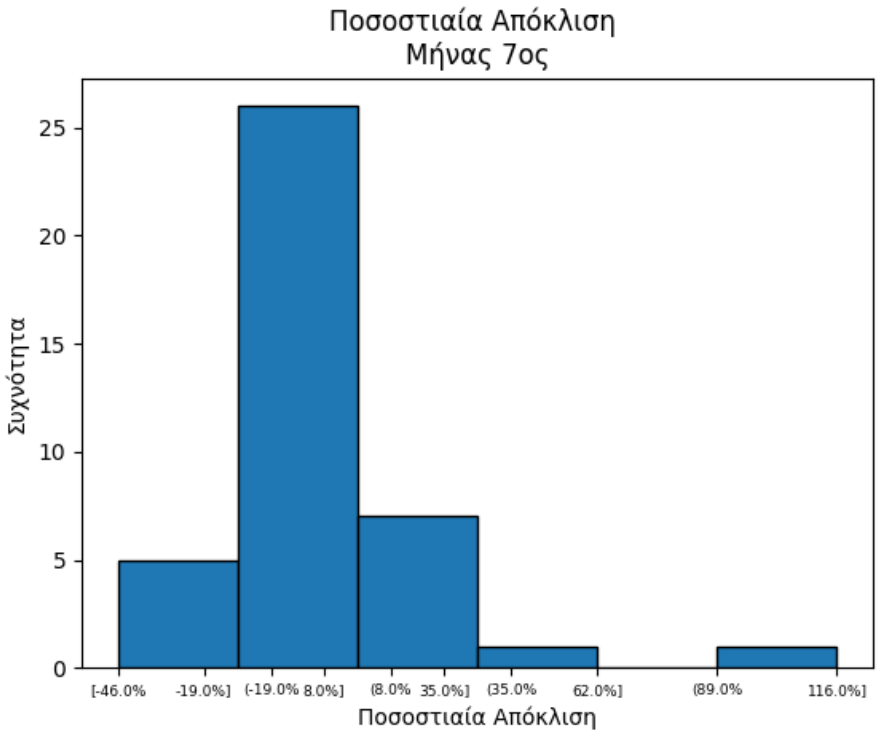
– Μάρτιος 2017-2018

Account	Period	Month	Kw/day(2017)	Kw/day(2018)	Deviation
41046018	2017-2018	3	17.042	16.917	-0.738
41058306	2017-2018	3	24.457	27.056	9.609
41089028	2017-2018	3	14.539	17.166	15.304
41058310	2017-2018	3	8.263	7.883	-4.812
41058312	2017-2018	3	22.000	24.864	11.519
41058313	2017-2018	3	21.084	22.624	6.807
41082889	2017-2018	3	0.049	2.792	98.238
41078795	2017-2018	3	14.846	17.878	16.962
41029644	2017-2018	3	8.261	8.101	-1.976
41078796	2017-2018	3	4.074	5.868	30.583
41089033	2017-2018	3	6.499	6.206	-4.729
41078799	2017-2018	3	14.534	15.169	4.19
41046032	2017-2018	3	26.267	26.265	-0.006
41058320	2017-2018	3	3.504	3.435	-1.993
41056274	2017-2018	3	0.000	0.000	0
41074706	2017-2018	3	3.145	3.397	7.405
41043988	2017-2018	3	58.342	53.588	-8.872
41058326	2017-2018	3	0.000	0.000	0
41029656	2017-2018	3	24.818	8.648	-186.984
41043994	2017-2018	3	4.470	4.220	-5.925
41076706	2017-2018	3	5.640	10.246	44.958
41076707	2017-2018	3	4.107	4.191	2.02
41076709	2017-2018	3	6.679	4.724	-41.391
41043945	2017-2018	3	6.495	5.334	-21.752
41043946	2017-2018	3	8.766	6.903	-26.986
41029611	2017-2018	3	38.689	37.589	-2.926
41043947	2017-2018	3	21.621	20.858	-3.659
41062378	2017-2018	3	17.195	16.411	-4.777
41015278	2017-2018	3	14.331	16.667	14.015
41005039	2017-2018	3	8.496	8.258	-2.876
41046000	2017-2018	3	15.687	19.734	20.504
41070574	2017-2018	3	15.544	15.645	0.643
41078765	2017-2018	3	0.000	0.000	0
41089007	2017-2018	3	3.893	3.741	-4.049
41082868	2017-2018	3	12.909	21.008	38.552
41017334	2017-2018	3	0.039	0.000	-100
41082870	2017-2018	3	0.000	0.000	0
41078776	2017-2018	3	36.063	32.438	-11.174
41070585	2017-2018	3	33.168	30.678	-8.117
41005051	2017-2018	3	35.463	32.642	-8.643


Ποσοστιαία Απόκλιση
Μήνας 3ος
Συχνότητα
0
5
10
15
20
25
30
[-187.0%
-130.0%]
(-130.0%
-73.0%]
(-73.0%
-16.0%]
(-16.0%
41.0%]
(41.0%
98.0%]
(98.0%
155.0%]
Ποσοστιαία Απόκλιση

– Ιούλιος 2017-2018

Account	Period	Month	Kw/day(2017)	Kw/day(2018)	Deviation
41046018	2017-2018	7	14.135	16.715	15.433
41058306	2017-2018	7	26.275	23.483	-11.888
41089028	2017-2018	7	14.645	20.104	27.153
41058310	2017-2018	7	6.950	5.645	-23.133
41058312	2017-2018	7	18.137	17.902	-1.311
41058313	2017-2018	7	16.536	16.870	1.979
41082889	2017-2018	7	1.618	4.057	60.125
41078795	2017-2018	7	16.255	20.512	20.754
41029644	2017-2018	7	8.931	8.294	-7.676
41078796	2017-2018	7	4.258	3.388	-25.694
41089033	2017-2018	7	5.766	5.334	-8.112
41078799	2017-2018	7	12.430	12.439	0.075
41046032	2017-2018	7	21.687	32.585	33.445
41058320	2017-2018	7	3.400	3.430	0.867
41056274	2017-2018	7	0.000	0.000	0
41074706	2017-2018	7	3.074	3.430	10.361
41043988	2017-2018	7	49.740	48.059	-3.497
41058326	2017-2018	7	0.000	0.000	0
41029656	2017-2018	7	8.058	6.610	-21.896
41043994	2017-2018	7	3.891	2.665	-46.001
41076706	2017-2018	7	2.588	24.165	89.291
41076707	2017-2018	7	4.303	4.270	-0.793
41076709	2017-2018	7	5.442	5.541	1.783
41043945	2017-2018	7	2.707	2.874	5.829
41043946	2017-2018	7	7.610	5.811	-30.959
41029611	2017-2018	7	40.416	45.334	10.849
41043947	2017-2018	7	18.308	17.223	-6.299
41062378	2017-2018	7	17.319	15.298	-13.217
41015278	2017-2018	7	15.387	13.174	-16.799
41005039	2017-2018	7	7.190	7.700	6.622
41046000	2017-2018	7	14.597	15.418	5.324
41070574	2017-2018	7	19.198	17.298	-10.989
41078765	2017-2018	7	0.000	0.000	0
41089007	2017-2018	7	3.778	3.533	-6.932
41082868	2017-2018	7	10.797	15.536	30.505
41017334	2017-2018	7	0.000	0.000	0
41082870	2017-2018	7	0.000	0.000	0
41078776	2017-2018	7	34.247	32.249	-6.195
41070585	2017-2018	7	28.893	25.634	-12.711
41005051	2017-2018	7	31.942	30.967	-3.15



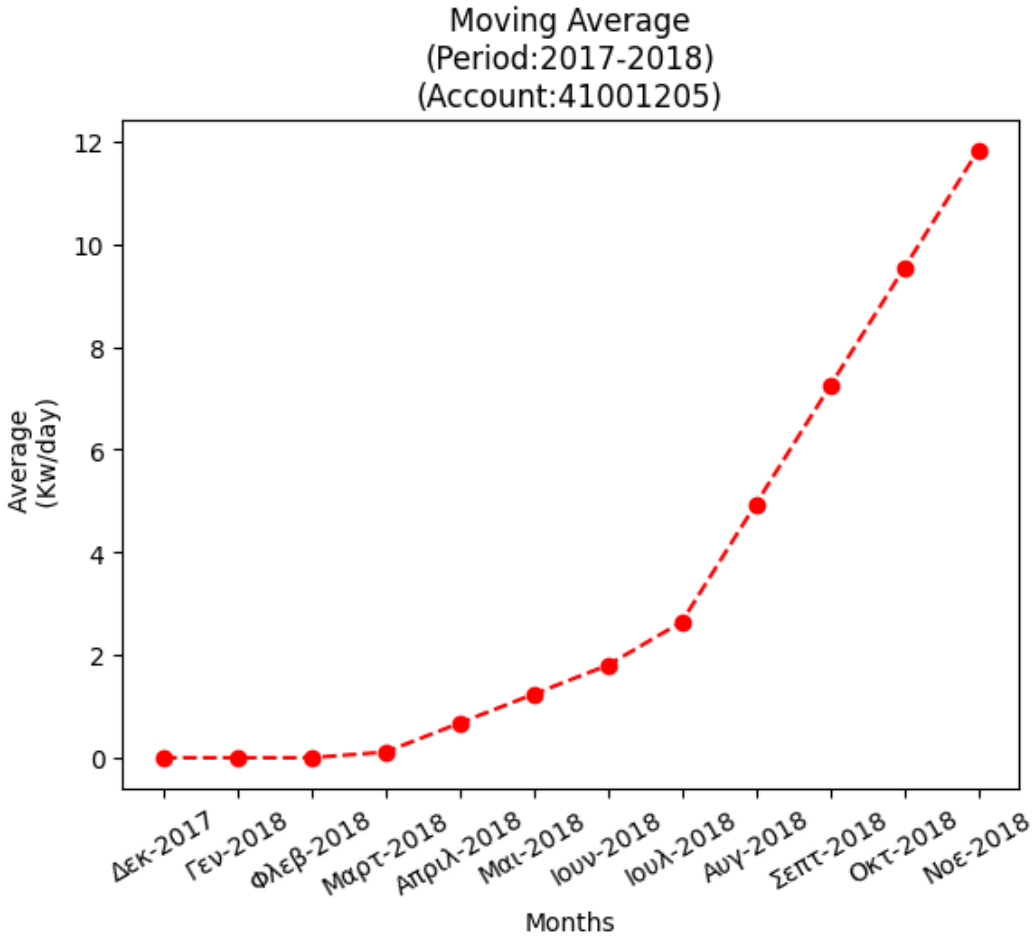
Σε κάθε περίπτωση παρατηρούμε πως ο αριθμός των περισσότερων αποκλίσεων ανήκει στην ομάδα που περιέχει το 0. Παρόλα αυτά παρουσιάζεται ένα σύνολο σημαντικών αποκλίσεων που πρέπει να ληφθεί υπόψη. Συγκεκριμένα οι αποκλίσεις που ανήκουν στις ακραίες ομάδες των ιστογραμμάτων πρέπει να διερευνηθούν περαιτέρω από τον πελάτη με σκοπό την κατανοήση των αιτιών σημαντικής απόκλισης, αλλά και την δυνατότητα επίλυσής της.

- Μέθοδος Α2: Για την μέθοδο αυτή παραθέτουμε τον πίνακα παροχών που παρουσιάστηκε στο σχετικό κεφάλαιο, καθώς και χαρακτηριστικές γραφικές παραστάσεις του KMO, για παροχές των οποίων η ποσοστιαία απόκλιση ξεπερνά το 40%. Υπενθυμίζουμε πως η ποσοστιαία απόκλιση προκύπτει από τους μέσους όρους του πρώτου και τελευταίου μήνα της πιο πρόσφατης 12άδας εγγραφών.

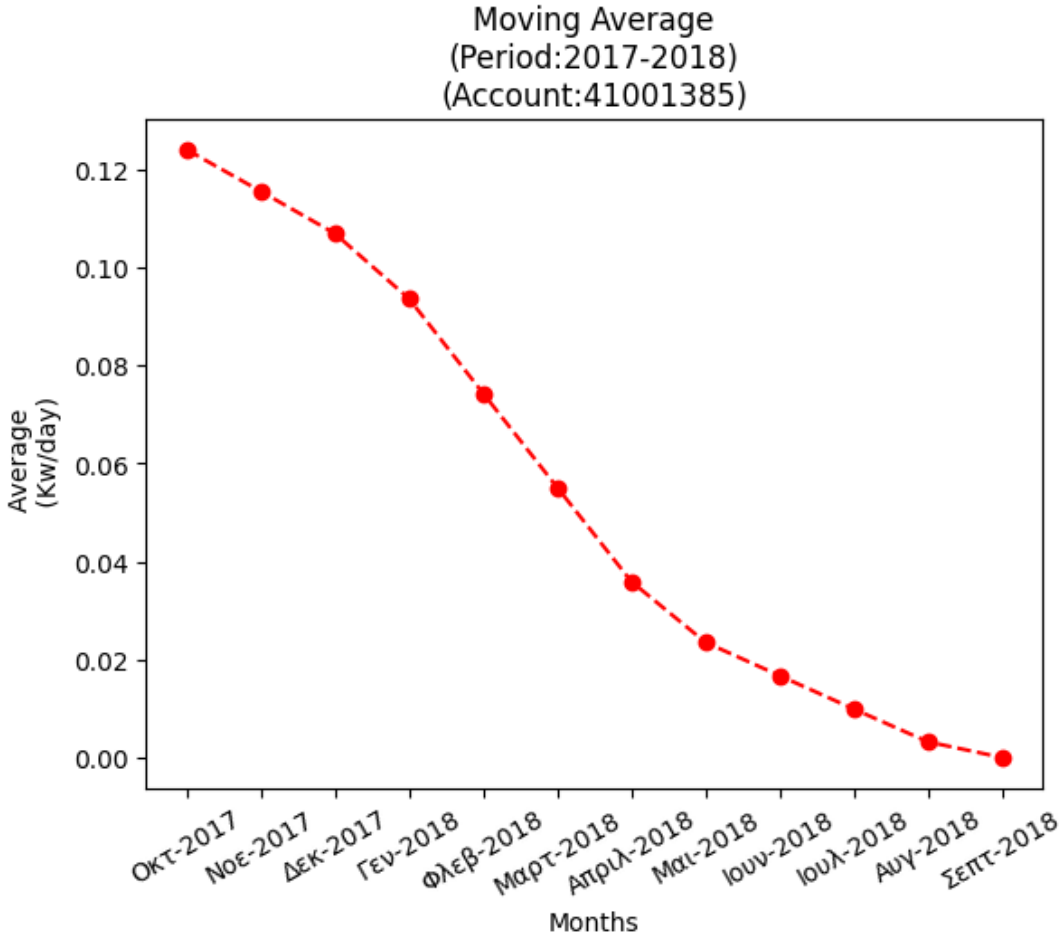
1

Account	Period	Kw/day(Last_Year)	Kw/day(Year)	Month	Average	Deviation
41001205	2016-2017	0	0	12	0	100
41001205	2017-2018	0	0	1	0	100
41001205	2017-2018	0	0	2	0	100
41001205	2017-2018	0	1.3092	3	0.1091	100
41001205	2017-2018	0	6.7642	4	0.6728	100
41001205	2017-2018	0	6.7642	5	1.2365	100
41001205	2017-2018	0	6.7642	6	1.8002	100
41001205	2017-2018	0	10.1162	7	2.6432	100
41001205	2017-2018	0	27.5462	8	4.9387	100
41001205	2017-2018	0	27.5462	9	7.2342	100
41001205	2017-2018	0	27.5462	10	9.5297	100
41001205	2017-2018	0	27.5462	11	11.8252	100
41001213	2016-2017	0.8417	0.3226	10	0.4162	-52.0293
41001213	2016-2017	0.8417	0.3226	11	0.3729	-52.0293
41001213	2016-2017	0.8417	0.3226	12	0.3297	-52.0293
41001213	2017-2018	0.7597	0.3184	1	0.2929	-52.0293
41001213	2017-2018	0.3333	0.2583	2	0.2867	-52.0293
41001213	2017-2018	0.3333	0.2583	3	0.2804	-52.0293
41001213	2017-2018	0.3333	0.2583	4	0.2742	-52.0293
41001213	2017-2018	0.3262	0.2573	5	0.2684	-52.0293
41001213	2017-2018	0.2231	0.2417	6	0.27	-52.0293
41001213	2017-2018	0.2231	0.2417	7	0.2715	-52.0293
41001213	2017-2018	0.2231	0.2417	8	0.273	-52.0293
41001213	2017-2018	0.2331	0.2417	9	0.2738	-52.0293
41001218	2016-2017	29.025	19.6833	12	22.9799	-34.6937
41001218	2017-2018	29.025	19.6833	1	22.2014	-34.6937
41001218	2017-2018	29.025	19.6833	2	21.423	-34.6937
41001218	2017-2018	28.4901	19.3269	3	20.6594	-34.6937
41001218	2017-2018	20.7339	16	4	20.2649	-34.6937
41001218	2017-2018	20.7339	16	5	19.8704	-34.6937
41001218	2017-2018	20.7339	16	6	19.4759	-34.6937
41001218	2017-2018	20.7339	15.9868	7	19.0803	-34.6937
41001218	2017-2018	21.6833	15.5917	8	18.5727	-34.6937
41001218	2017-2018	21.6833	15.5917	9	18.065	-34.6937
41001218	2017-2018	21.6833	15.5917	10	17.5574	-34.6937
41001218	2017-2018	21.55	15.5917	11	17.0609	-34.6937
41001238	2016-2017	14.4016	13.0992	10	12.8338	52.5788
41001238	2016-2017	14.4016	13.0992	11	12.7253	52.5788
41001238	2016-2017	14.4016	13.0992	12	12.6167	52.5788
41001238	2017-2018	14.4016	10.8103	1	12.3175	52.5788
41001238	2017-2018	14.4016	9.874	2	11.9402	52.5788
41001238	2017-2018	14.4016	9.874	3	11.5629	52.5788
41001238	2017-2018	14.4016	9.874	4	11.1855	52.5788
41001238	2017-2018	11.5906	40.4877	5	13.5936	52.5788
41001238	2017-2018	10.252	51.136	6	17.0006	52.5788
41001238	2017-2018	10.252	51.136	7	20.4076	52.5788
41001238	2017-2018	10.252	51.136	8	23.8146	52.5788
41001238	2017-2018	12.1501	51.136	9	27.0635	52.5788
41001264	2016-2017	24.7154	10.3333	12	24.2626	-91.7021
41001264	2017-2018	24.7154	10.3333	1	23.0641	-91.7021
41001264	2017-2018	24.7154	10.3333	2	21.8656	-91.7021
41001264	2017-2018	26.0513	15.6891	3	21.0021	-91.7021
41001264	2017-2018	26.3719	16.719	4	20.1977	-91.7021
41001264	2017-2018	26.3719	16.719	5	19.3933	-91.7021
41001264	2017-2018	26.3719	16.719	6	18.5888	-91.7021
41001264	2017-2018	27.9897	11.6115	7	17.224	-91.7021
41001264	2017-2018	28.3008	10.8548	8	15.7702	-91.7021
41001264	2017-2018	28.3008	10.8548	9	14.3163	-91.7021
41001264	2017-2018	28.3008	10.8548	10	12.8625	-91.7021
41001264	2017-2018	13.3279	10.8548	11	12.6564	-91.7021
41001271	2016-2017	23.1475	1.8512	10	11.9249	46.4245

KMO



Η ραγδαία αύξηση στην κατανάλωση πιθανώς να οφείλεται σε υποκλοπή ρεύματος, με την διερεύνηση των πραγματικών αιτιών να διεξαχθεί από τον Δήμο κατά την παράδωση των εν λόγω αποτελεσμάτων.



Η συμπεριφορά της συγκεκριμένης παροχής τιμολογίου ΦΟΠ,έχει επιβεβαιωθεί από τον Δήμο πως προκύπτει λόγω αντικατάστασης των λαμπτήρων με δίοδο εκπομπής φωτός (LED).

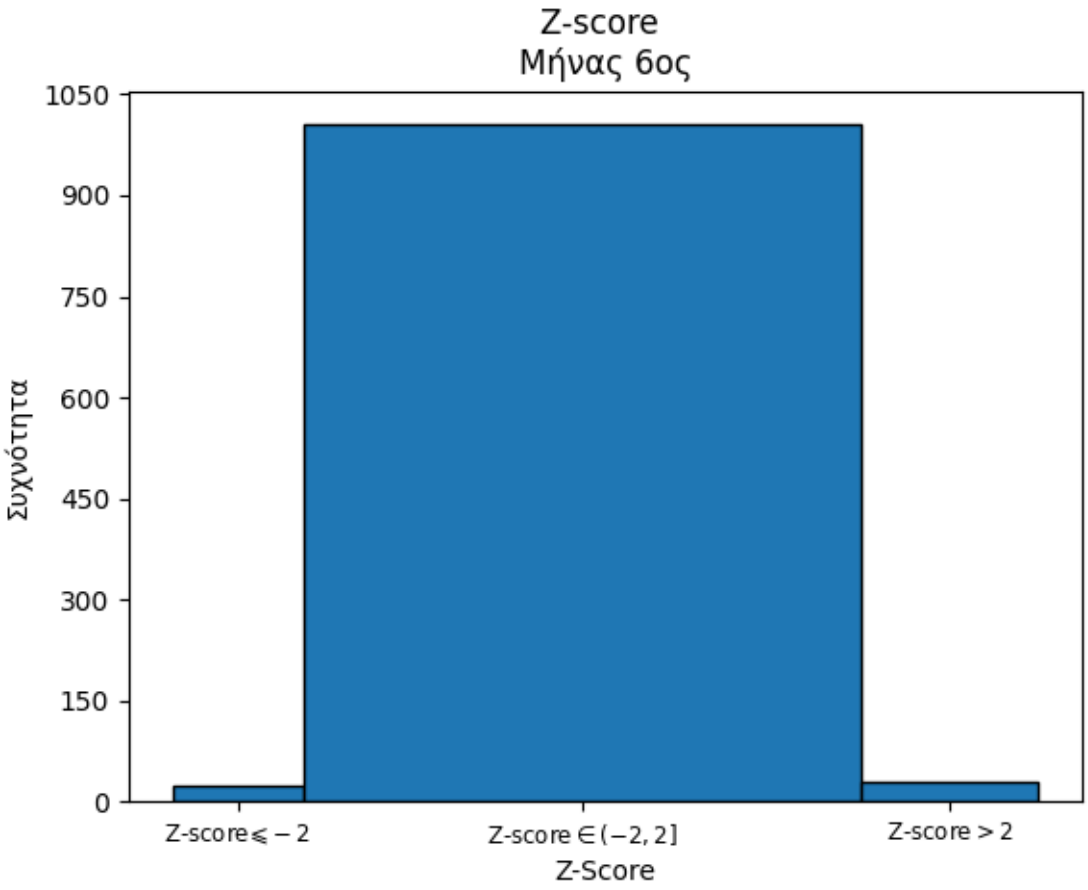
- Μέθοδος Α3: Για την μεθοδολογία αυτή, τα αποτελέσματα είναι ο πίνακας των παροχών με ακραία τιμή z-score για δεδομένο μήνα και το αντίστοιχο ιστόγραμμα για όλες τις παροχές με εγγραφή στο μήνα. Το ιστόγραμμα αναμένουμε να περιέχει την πλειοψηφία των τιμών στην ομάδα $z_{score} \in (-2, 2]$, με το σύνολο των τιμών που περιέχονται στις άλλες δύο κατηγορίες να αποτελούν προτεραιότητας για την ανακάλυψη πιθανών ευρημάτων.

  – Ιούνιος 2018

1

Account	Month	Year	Kw/day	Z-score
41000324	6	2018	54.837	2.47
41001233	6	2018	5.451	-2.734
41001238	6	2018	51.136	2.959
41001253	6	2018	29.423	-2.044
41001293	6	2018	9.525	2.28
41001357	6	2018	1.177	2.849
41001362	6	2018	5.699	2.268
41001370	6	2018	2.398	-2.065
41001388	6	2018	2.301	-2.437
41001395	6	2018	16.459	3.284
41001396	6	2018	59.12	2.134
41001435	6	2018	65.115	2.204
41031287	6	2018	14.738	-2.167
41037454	6	2018	11.403	2.464
41037455	6	2018	0.839	2.498
41039391	6	2018	4.452	-2.527
41039408	6	2018	8.21	3.027
41055074	6	2018	15.645	2.131
41055151	6	2018	4.975	3.064
41057388	6	2018	7.459	2.251
41059060	6	2018	5.608	-2.57
41062512	6	2018	10.073	-2.644
41062514	6	2018	10.048	-2.181
41062858	6	2018	42.965	-2.071
41064712	6	2018	9.117	2.056
41064930	6	2018	25.225	-2.205
41065299	6	2018	10.041	-2.518
41066650	6	2018	76.299	2.32
41067712	6	2018	6.16	2.573
41067787	6	2018	4.792	-2.417
41069616	6	2018	29.089	2.155
41071422	6	2018	11.754	-2.396
41072491	6	2018	11.702	2.42
41073142	6	2018	8.952	2.122
41073727	6	2018	5.463	-2.048
41075797	6	2018	11.411	2.26
41075943	6	2018	3.504	-2.146
41077057	6	2018	2.244	-2.028
41077140	6	2018	0.636	-2.137
41079071	6	2018	6.421	2.458
41079784	6	2018	12.844	-2.198
41081551	6	2018	15.342	-2.032
41082298	6	2018	3.603	2.16
41082889	6	2018	4.057	2.088
41082972	6	2018	1.37	-2.086
41083419	6	2018	0.702	-2.948
41084641	6	2018	25.234	2.445
41085672	6	2018	11.113	2.299
41086528	6	2018	8.123	2.236
41088992	6	2018	7.419	2.35

Z-scores-Μήναςδος

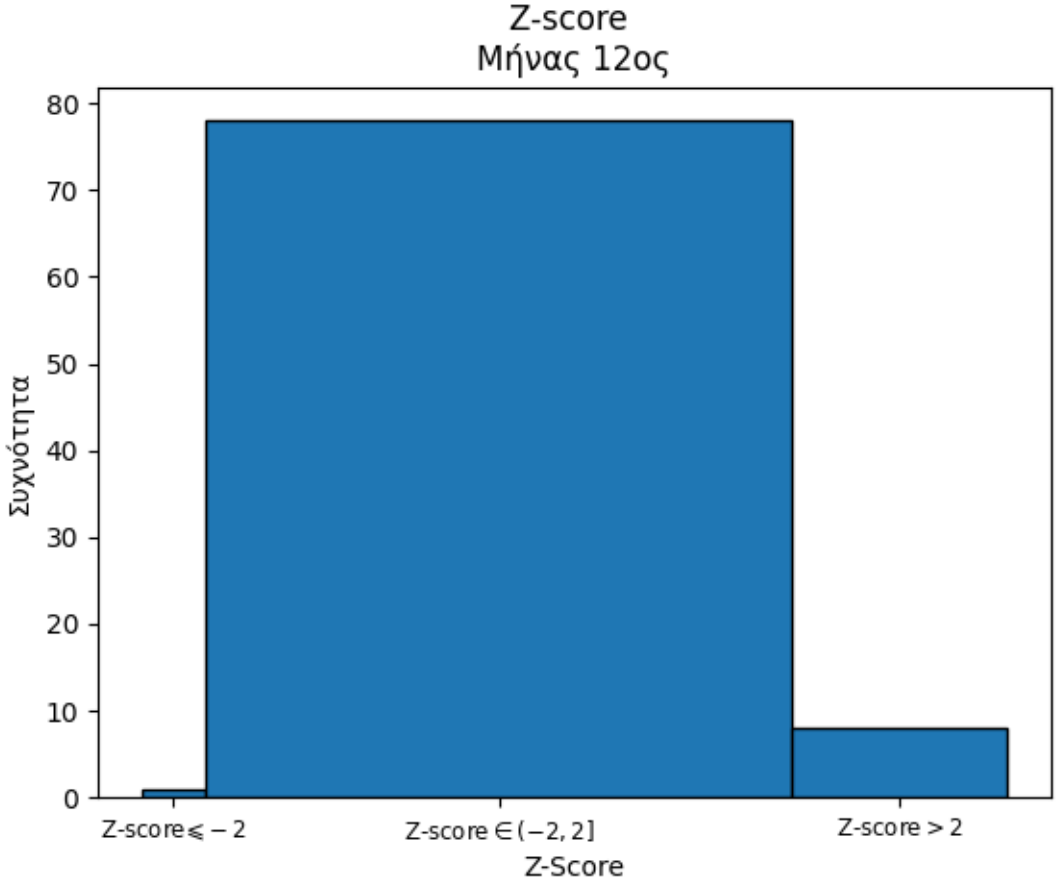

Z-score
Μήνας 6ος
1050
900
750
600
450
300
150
0
Συχνότητα
Z-score ⩽ −2
Z-score ∈ (−2, 2]
Z-score > 2
Z-Score

– Δεκέμβριος 2018

1

Account	Month	Year	Kw/day	Z-score
41001394	12	2018	46.529	2.967
41012677	12	2018	8.692	2.381
41066331	12	2018	6.684	3.052
41072693	12	2018	9.017	-2.445
41077068	12	2018	149.065	3.483
41079784	12	2018	62.672	2.048
41085515	12	2018	21.633	2.496
41086128	12	2018	8.797	2.381
41087267	12	2018	2.057	2.538

Z-scores-Μήνας12ος



Όπως αναμέναμε η πλειοψηφία των παρατηρήσεων βρίσκονται και στις δύο περιπτώσεις στην ομάδα $z_{score} \in (-2, 2]$, με τις ομάδες των ακραίων καταναλώσεων να περιέχουν το πολύ το 10% των συνολικών παρατηρήσεων. Το διάγραμμα λοιπόν επιτρέπει στο Δήμο, να διακρίνει το πλήθος των παροχών με σημαντική απόκλιση, από τα συνολικά συλλεγόμενα δεδομένα όλων των διαθέσιμων ετών. Στη συνέχεια με χρήση του αντίστοιχου πίνακα, είναι δυνατή η εύρεση των συγκεκριμένων παροχών με ακραίες τιμές z_{score}, κάτι που με την σειρά του επιτρέπει την διερεύνηση αιτιών, για την εν λόγω απόκλιση.

- Μέθοδος Β1: Κατά την μεθοδολογία αυτή αποσκοπούμε στην εύρεση παροχών με μηδενική κατανάλωση. Τα αιτία για ένα τέτοιο φαινόμενο περιλαμβάνουν μεταξύ άλλων την ανενεργεία παροχής καθώς και την βλάβη του ρολογιού καταμέτρησης. Παρόλα αυτά όπως έχει ήδη αναφερθεί η ανακάλυψη των πραγματικών αιτιών θα βρεθεί από τον αρμόδιο τεχνικό στον Δήμο.

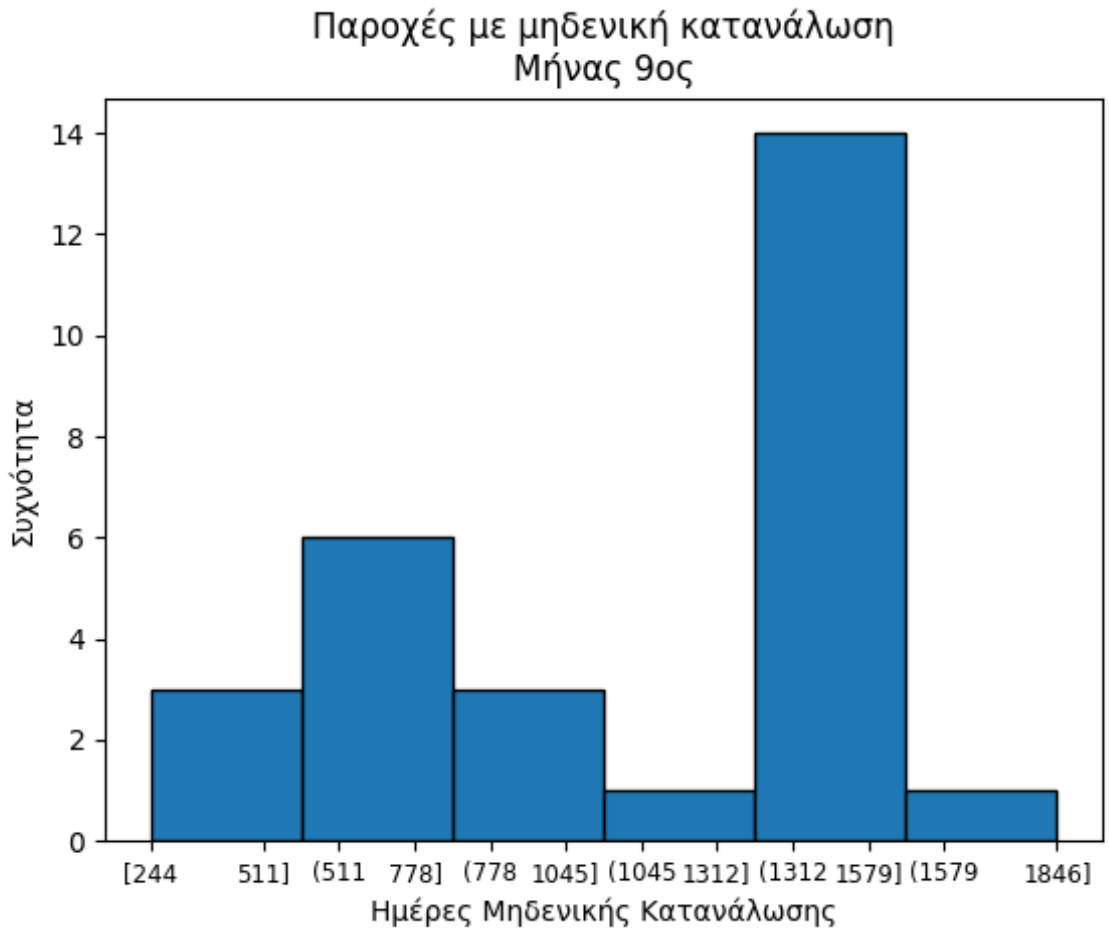
  Παρακάτω παραθέτουμε τρεις περιπτώσεις μηνών για το έτος 2018 με τον πίνακα παροχών και το αντίστοιχο ιστόγραμμα για κάθε μήνα.

  – Σεπτέμρβης 2018

1

Account	Month	Duration	Total_Cost	New_Entry
41001208	9	1336	45.59	OXI
41001239	9	972	30	OXI
41001344	9	1097	-92	OXI
41001359	9	1336	45.59	OXI
41001385	9	363	11	OXI
41004891	9	1339	79.43	OXI
41006400	9	1337	298.78	OXI
41016421	9	1460	70.41	OXI
41058326	9	1466	49.17	OXI
41064935	9	728	19	OXI
41071471	9	971	47	OXI
41071529	9	1462	91.26	OXI
41071896	9	1464	53.83	OXI
41073489	9	606	17	OXI
41074125	9	365	7	OXI
41074212	9	1582	70	OXI
41075713	9	1464	43.38	OXI
41077803	9	604	34	OXI
41078149	9	853	12	OXI
41078890	9	606	21	OXI
41079448	9	728	-21	OXI
41080076	9	1339	43.41	OXI
41081058	9	1461	51.04	OXI
41082294	9	244	-64	OXI
41085171	9	1459	54.42	OXI
41086227	9	730	12	OXI
41086680	9	1460	54.18	OXI
41088190	9	1464	49.05	OXI

Παροχές_Μηδενικής_Κατανάλωσης9ος

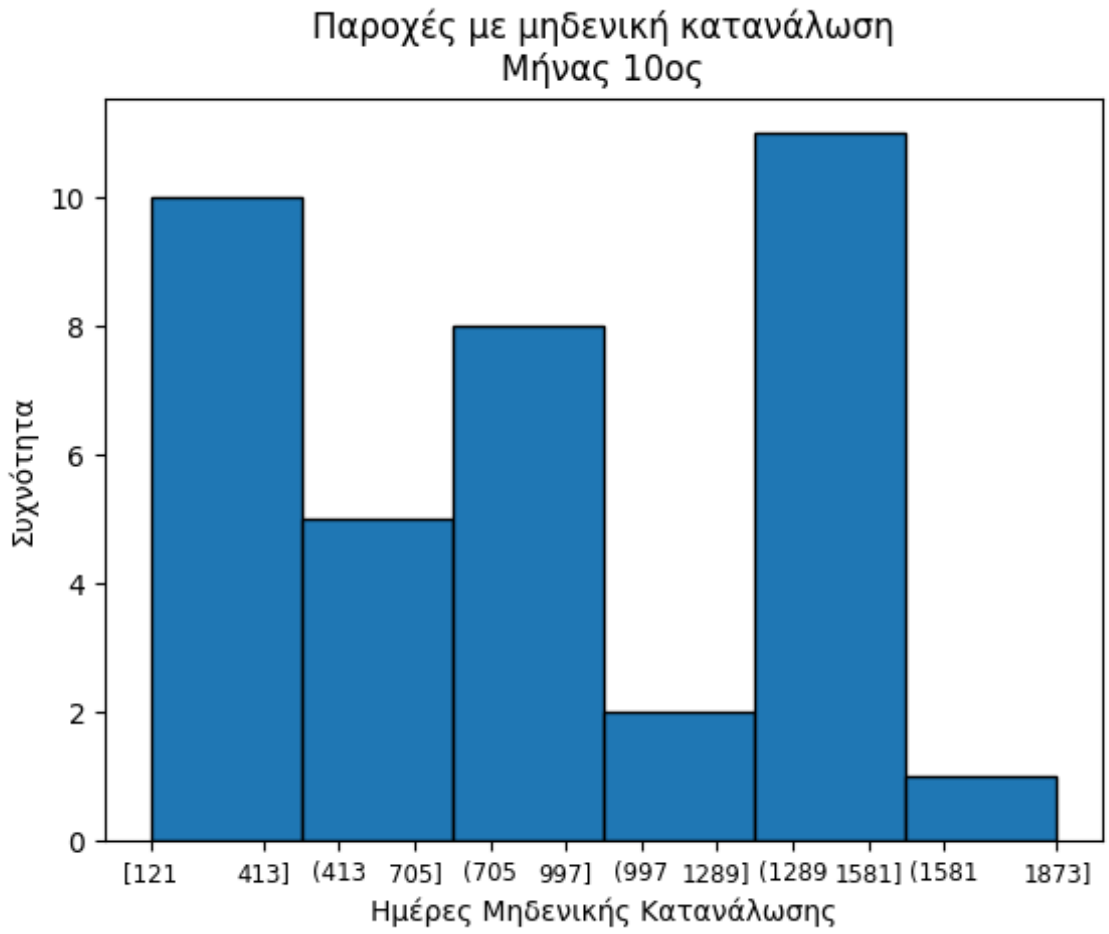

Παροχές με μηδενική κατανάλωση
Μήνας 9ος
Συχνότητα
0
2
4
6
8
10
12
14
[244 511]
(511 778]
(778 1045]
(1045 1312]
(1312 1579]
(1579 1846]
Ημέρες Μηδενικής Κατανάλωσης

– Οκτώβρης 2018

1

Account	Month	Duration	Total_Cost	New_Entry
41001341	10	485	-103	OXI
41001374	10	244	7	OXI
41017334	10	483	16	OXI
41017399	10	610	21	OXI
41017534	10	1338	47.51	OXI
41017594	10	1338	47.51	OXI
41018484	10	1339	-894.17	OXI
41018839	10	364	144	OXI
41055364	10	1339	79.34	OXI
41057084	10	852	-119	OXI
41062862	10	1092	180	OXI
41064949	10	242	10	OXI
41065900	10	1459	54.46	OXI
41070482	10	974	28	OXI
41070801	10	729	-2	OXI
41071257	10	849	30	OXI
41071667	10	121	-1	NAI
41072017	10	1582	315.87	OXI
41073379	10	1331	-1303	OXI
41073381	10	1093	32	OXI
41073382	10	1338	48.53	OXI
41075584	10	121	6	NAI
41076023	10	729	10	OXI
41076024	10	121	-1	NAI
41076025	10	364	2	OXI
41077947	10	123	0	NAI
41078356	10	121	-9	NAI
41081115	10	121	-11	NAI
41081409	10	1341	48.6	OXI
41082493	10	849	29	OXI
41084114	10	725	44	OXI
41086865	10	1462	47.56	OXI
41087823	10	1459	86.22	OXI
41089784	10	608	-138	OXI
41305647	10	488	-2	OXI
41305674	10	1345	51.45	OXI
41324520	10	855	25	OXI

Παροχές_Μηδενικής_Κατανάλωσης10ος

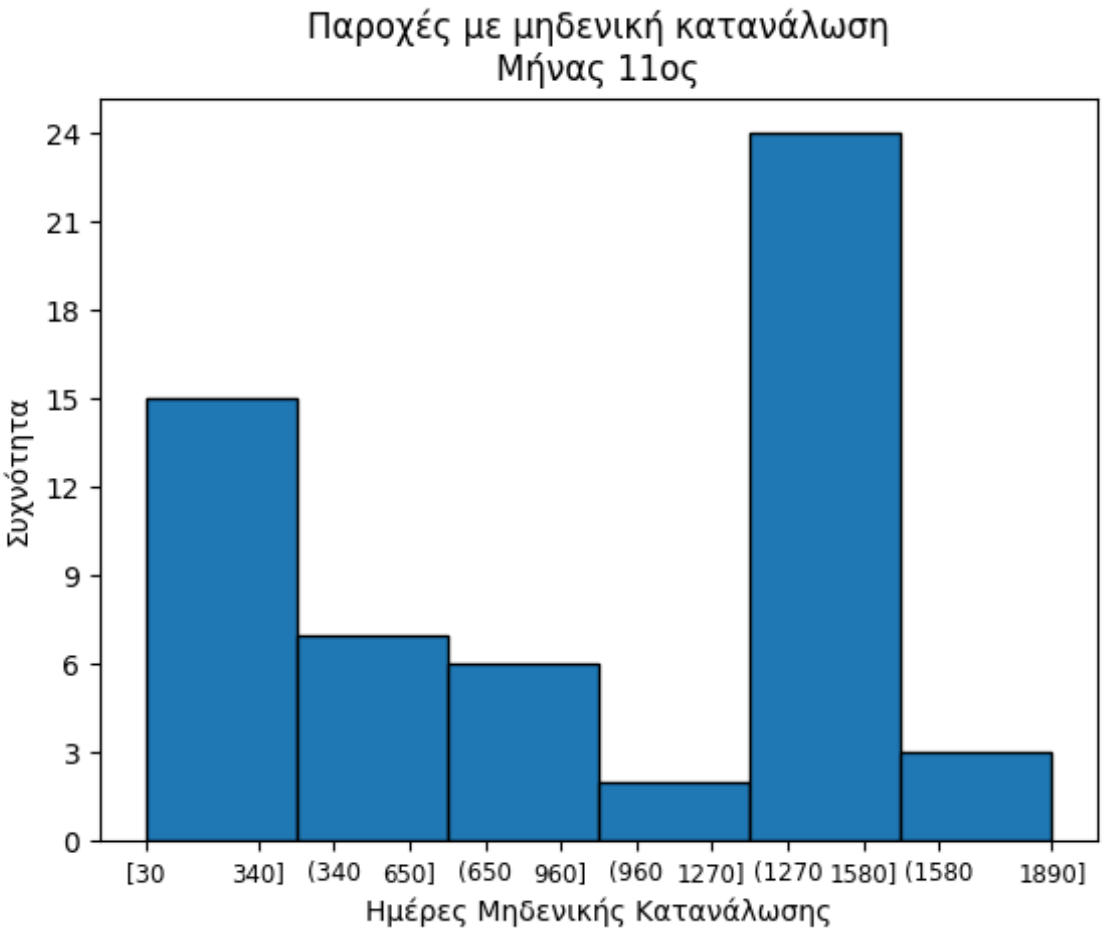

Παροχές με μηδενική κατανάλωση
Μήνας 10ος
Συχνότητα
0
2
4
6
8
10
[121 413]
(413 705]
(705 997]
(997 1289]
(1289 1581]
(1581 1873]
Ημέρες Μηδενικής Κατανάλωσης

– Νοέμβρης 2018

1

Account	Month	Duration	Total_Cost	New_Entry
41001351	11	850	-321	OXI
41001364	11	1581	51.15	OXI
41003164	11	849	29	OXI
41004047	11	1577	261.68	OXI
41004123	11	119	4	NAI
41004162	11	608	18	OXI
41004478	11	1461	44.35	OXI
41005683	11	1457	50.16	OXI
41029514	11	123	3	NAI
41030191	11	1339	77	OXI
41038249	11	1458	45.1	OXI
41039151	11	728	59	OXI
41039175	11	244	9	OXI
41045832	11	120	-63	NAI
41046990	11	30	7	NAI
41047261	11	91	14	OXI
41047556	11	1460	79.83	OXI
41056274	11	1408	927.61	OXI
41059151	11	30	13	NAI
41059209	11	1580	-406.27	OXI
41064146	11	1241	197	OXI
41064822	11	1459	52.1	OXI
41064848	11	1458	51.23	OXI
41065182	11	1461	62.75	OXI
41067844	11	1460	52.24	OXI
41068628	11	56	9	OXI
41068748	11	61	10	OXI
41069952	11	1406	926.3	OXI
41070840	11	245	-394	OXI
41071234	11	1584	-1190	OXI
41071982	11	1458	318.61	OXI
41072157	11	495	-428	OXI
41072340	11	91	14	OXI
41074037	11	483	17	OXI
41076177	11	91	-1004	OXI
41076309	11	849	48	OXI
41077475	11	1343	-169	OXI
41078262	11	609	-9	OXI
41078765	11	516	270	OXI
41079178	11	1216	-44	OXI
41079591	11	1578	81.94	OXI
41080059	11	1458	50.09	OXI
41080291	11	1457	51.03	OXI
41081732	11	1578	50.27	OXI
41082298	11	120	-5	NAI
41082572	11	1578	50.27	OXI
41082870	11	728	22	OXI
41083430	11	124	0	NAI
41083431	11	364	11	OXI
41084391	11	1578	51.27	OXI
41084666	11	1458	53.1	OXI
41086002	11	121	-120	NAI
41087358	11	852	28	OXI
41089000	11	1458	51.1	OXI
41089781	11	604	101	OXI
41314840	11	1410	602.08	OXI
94308701	11	1495	1743.09	OXI

Παροχές_Μηδενικής_Κατανάλωσης11ος



Όπως παρατηρούμε και από τις τρεις παραπάνω γραφικές, τα όρια των ομάδων των ιστογραμμάτων ξεπερνούν σχετικά άμεσα τις 365 ημέρες, κάτι που υποδηλώνει μακροχρόνια μηδενική κατανάλωση, για τις συγκερκιμενες παροχές. Ακόμα, η μέγιστη συχνότητα παρουσιάζεται κάθε φορά σε μια από τις τελευταίες ομάδες του ιστογράμματος, πράγμα που περαιτέρω ενισχύει την παραπάνω υπόθεση.

Από τις δύο αυτές λοιπόν παρατηρήσεις, συμπεραίνουμε έλλειψη παρελθοντικών ελεγχών ή και μέτρων αντιμετώπισης φαινομένων μηδενικής κατανάλωσης, που θα επέτρεπαν την διάγνωση και εξομάλυνση των παραπάνω ευρυμάτων.

- Μέθοδος B2: Κατά την μεθοδολογία αυτή αποσκοπούμε στην εύρεση παροχών με ανεπαρκή ετήσια εκκαθάριση. Το σύνολο των αποτελεσμάτων περιορίζεται στον πίνακα παροχών όπως αυτός παρουσιάστηκε στο σχετικό κεφάλαιο.

  Παρακάτω παραθέτουμε τρία χαρακτηριστικά παραδείγματα των συγκεκριμένων αποτελεσμάτων.

Account	Time	Month	New_Entry
41001345	180	3	NAI
41016421	178	3	NAI
41035651	179	3	NAI
41073239	177	3	NAI
41084663	177	3	NAI
41090182	178	3	NAI

Account	Time	Month	New_Entry
41001077	181	7	NAI
41070719	180	7	NAI
41077558	181	7	NAI
41082573	182	7	NAI
41083370	180	7	NAI
41089135	180	7	NAI

Account	Time	Month	New_Entry
41009231	179	8	NAI
41017711	183	8	NAI
41018166	186	8	NAI
41058369	183	8	NAI
41069716	179	8	NAI
41073384	182	8	NAI
41079097	180	8	NAI
41080623	182	8	NAI
41081673	182	8	NAI

- Μέθοδος Β3: Κατά την μεθοδολογία αυτή επικεντρωνόμαστε στην εύρεση των κοινών μοτίβων κατανάλωσης για παροχές ίδιου τιμολογίου, καθώς και την συλλογή παροχών που αποκλίνουν από το αναμενόμενο μοτίβο.

  Όπως έχουμε ήδη εξηγήσει με σκοπό την εύρεση των κοινών μοτίβων εκπαιδεύσαμε ένα μοντέλο μηχανικής μάθησης με χρήση του αλγορίθμου K-means.

  Ο αλγόριθμος ακολουθεί τα παρακάτω βήματα:

Βήμα 1 : Αρχική επιλογή των κ κεντροειδών με τυχαίο τρόπο από το σύνολο εκπαίδευσης.

Βήμα 2 : Υπολογισμός της απόστασης μεταξύ κάθε χρονοσειράς από τα κεντροειδή.

Βήμα 3 : Ανάθεση κάθε χρονοσειράς στην ομάδα, η απόσταση με το κεντροειδές της οποίας είναι ελάχιστη.

Βήμα 4 : Επανυπολογισμός των κεντροειδών ως τον μέσο όρο των σημείων που ανατέθηκαν στην αντίστοιχη ομάδα.

Βήμα 5 : Επανάληψη των βημάτων 2-4 μέχρι το μέσο τετραγωνικό σφάλμα για κάθε ομάδα να μην ξεπερνά κάποιο προκαθορισμένο tolerance ή μέχρι ο αριθμός των επαναλήψεων να ξεπεράσει έναν προκαθορισμένη τιμή.

Το μοντέλο εκπαιδεύτηκε τρεις φορές με χρήση καθε φορά διαφορετικής μετρικής, που χρησιμοποιείται για τον υπολογισμό της απόστασης μεταξύ των χρονοσειρών.

Οι μετρικές που χρησιμοποιήθηκαν είναι οι παρακάτω:

- Ευκλείδια μετρική:

$$d(x,y) = \sqrt{\sum_{i=1}^{n} |x_i - y_i|^2}$$

  Όπου $x = (x_1, x_2, \ldots, x_n),\ y = (y_1, y_2, \ldots, y_n)$ χρονοσειρές

- Συνάρτηση δυναμικής χρονικής παραμόρφωσης (dynamic time warping) [3]:

$$DTW(x,y) = \min_{\pi \in \Pi} \sqrt{\sum_{(i,j)\in\pi} |x_i - y_j|^2}$$

  Όπου:

  * $x = (x_1, x_2, \ldots, x_n),\ y = (y_1, y_2, \ldots, y_m)$ χρονοσειρές
  * $\pi = ((i_1, j_1), (i_2, j_2), \ldots, (i_K, j_K))$ υποσύνολο ζεύγους δεικτών
  * Π σύνολο υποσυνόλων δεικτών με τις ιδιότητες:

    1 $\forall\ \pi \in P\ \pi_o = (0,0), \pi_K = (n-1, m-1)$

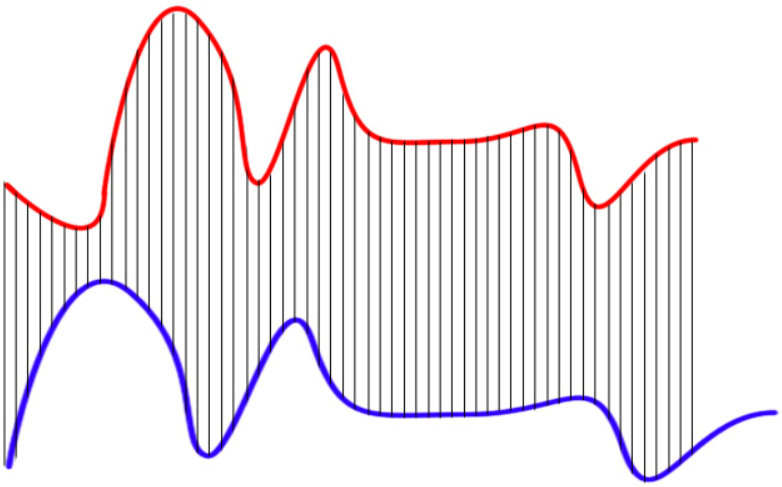
    2 $\forall\ k > 0\ \pi_k = f(\pi_{k-1})$ τ.ω:

    · $i_{k-1} \leqslant i_k \leqslant i_{k-1} + 1$

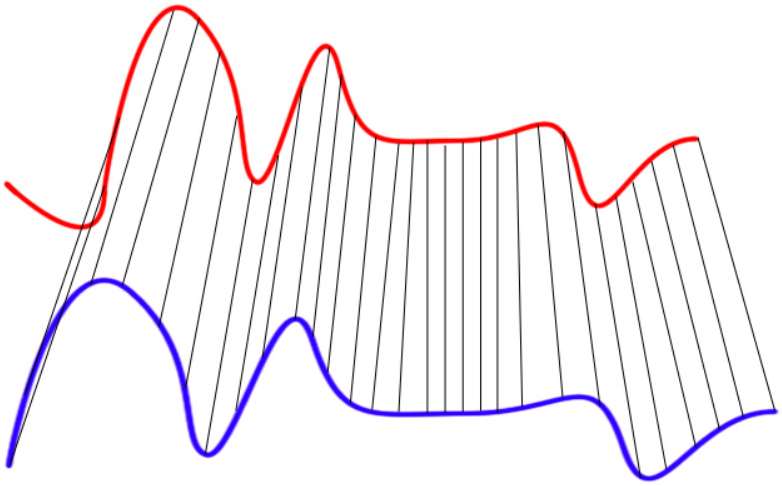
    · $j_{k-1} \leqslant j_k \leqslant j_{k-1} + 1$

  Συνοπτικά, η παραπάνω συνάρτηση αντιστοιχεί κάθε στοιχείο της χρονοσειράς x στο κοντινότερο στοιχείο της χρονοσειράς y. Παρότι η συγκεκριμένη συνάρτηση χρησιμοποιείται με σκοπό την μέτρηση της απόστασης μεταξύ δύο χρονοσειρών, δεν αποτελεί μετρική, καθώς ισχύει $DTW(x,y) \neq DTW(y,x)$.

Η αντιστοίχιση παρουσιάζεται στο παρακάτω υποθετικό γράφημα:



Euclidean Matching



Dynamic Time Warping Matching

Η απεικόνιση αυτή επωφελεί την εκπαίδευση είτε σε περιπώσεις όπου το πλήθος των στοιχείων μεταξύ των χρονοσειρών διαφέρει, είτε όταν υπάρχει μετατόπιση στον άξονα του χρόνου μεταξύ των χρονοσειρών.
Καθώς στην περίπτωσή μας κάθε χρονοσειρά περιέχει δώδεκα τιμές για κάθε μήνα του έτους, δεν αναμένουμε σημαντική βετλίωση στο τελικό μοντέλο,σε σύγκριση με το παραπάνω.

– Συνάρτηση απαλής δυναμικής χρονικής παραμόρφωσης (soft dynamic time warping) [4]:

Μπορεί να δειχθεί πως η συνάρτηση DTW δεν είναι κυρτή.
Η κυρτότητα καθιστά την διαδικασία βελτιστοποίησης γρηγορότερη καθώς μεταξύ άλλων δεν υπάρχει διαχωρισμός μεταξύ τοπικού και ολικού ελαχίστου.
Ακόμα, η dynamic time warping είναι μη παραγωγίσιμη λόγω της συνάρτησης $minimum$, κάτι που επίσης αποτελεί καθοριστικό για την διαδικασία βελτιστοποίησης.

Με σκοπό την επίλυση των ζητημάτων αυτών ορίζεται η απεικόνιση soft dynamic time warping μέσω του τύπου:

$$sdtw(x,y) = soft\text{–}min^{\gamma}_{\pi\in\Pi}\sqrt{\sum_{(i,j)\in\pi}|x_i - y_j|^2}$$

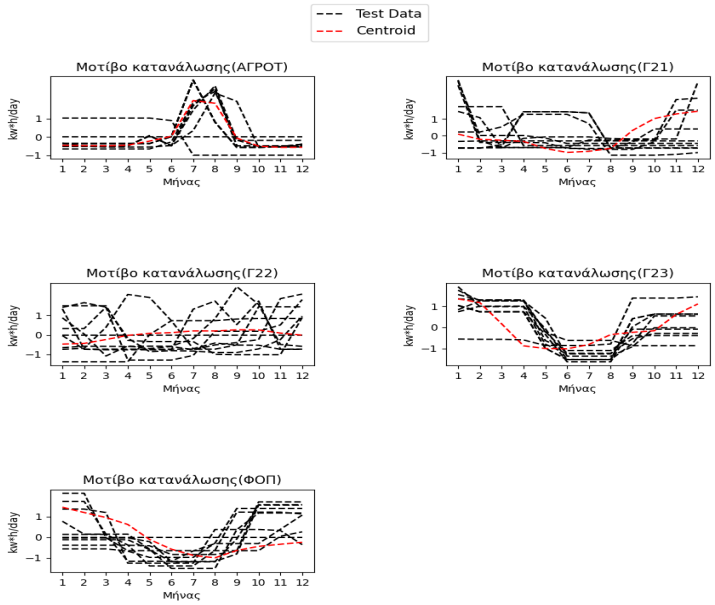
Όπου:

$$soft\text{–}min^{\gamma}(\pi_1,\pi_2,\ldots,\pi_K) = -\gamma\log(\sum_{i=1}^{K} e^{-\frac{\pi_i}{\gamma}})$$

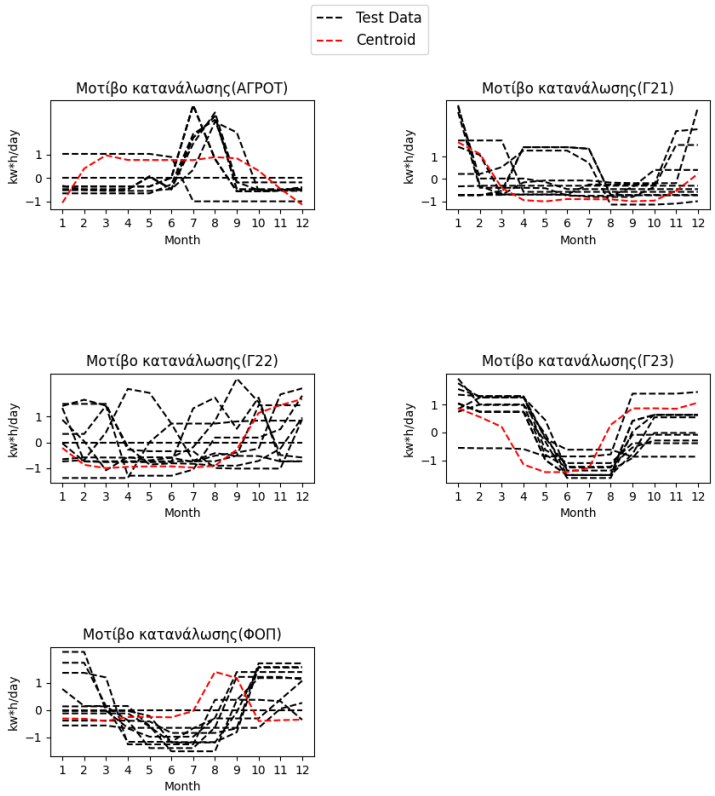
Συμβουλευόμενοι την βιβλιογραφία καταχωρούμε στο γ την τιμή 1.
Περιληπτικά αναφέρουμε πως καθώς το $\gamma \to \infty$, η απεικόνιση συγκλίνει σε μια κυρτή συνάρτηση.

Παρακάτω παραθέτουμε τις γραφικές των κεντροειδών για κάθε μοντέλο και τιμολόγιο, σε συνδυασμό με τις χρονοσειρές καταναλώσεων για παροχές του δεδομένου τιμολογίου.

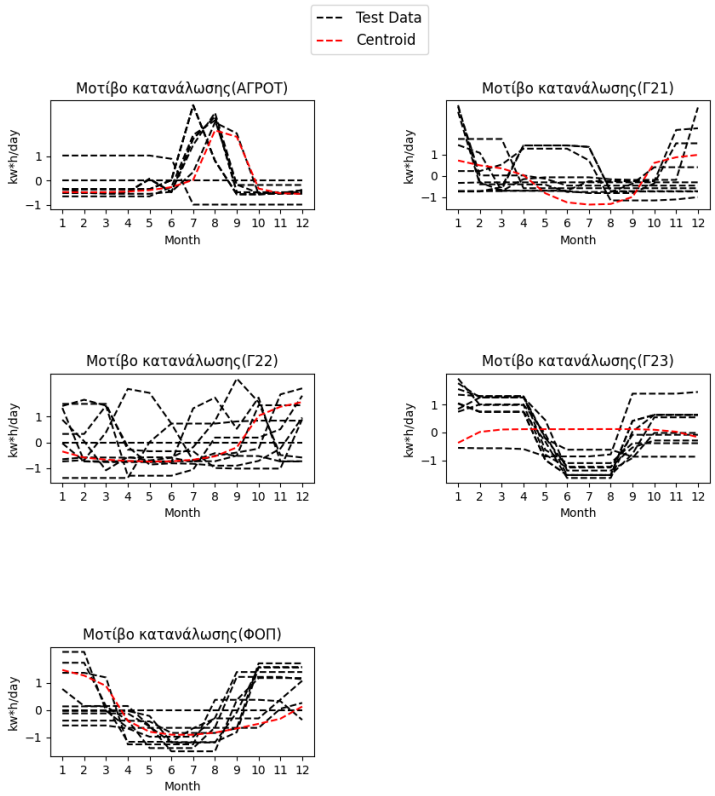
– Ευκλίδεια μετρική:



– Συνάρτηση DTW:



– Συνάρτηση SDTW:



Παραπάνω παρατηρούμε πως το μοντέλο που εκπαιδεύτηκε με χρήση της ευκλείδιας μετρική είναι πιο εύστοχο στην εύρεση του μοτίβου που αφορά τα τιμολόγια ΑΓΡΟΤ, Γ22.
Επίσης το μοντέλο που εκπαιδεύτηκε με την μετρική dynamic time warping, αποτελεί πιο εύστοχα για τα τιμολόγια Γ21, Γ23. Τέλος για την μετρική soft dynamic time warping το μοντέλο παρατηρείται να εκτιμά καλύτερα το μοτίβο του τιμολογίου ΦΟΠ.

Τα παραπάνω συμπεράσματα ενισχύονται και από τον πίνακα των μετρικών για τα τρία μοντέλα.

	ΑΓΡΟΤ	Γ21	Γ22	Γ23	ΦΟΠ
Precision_Euclid:	0.8706	0.1632	0.4133	0.0267	0.1771
Recall_Euclid:	0.7988	0.2334	0.3501	0.0221	0.173
Precision_DBA:	0	0.2634	0.2035	0.3087	0.0348
Recall_DBA:	0	0.3286	0.2254	0.2918	0.051
Precision_SDTW:	0.8113	0.1605	0.1987	0.032	0.2806
Recall_SDTW:	0.7988	0.163	0.1985	0.0221	0.3675

Ο παραπάνω πίνακας περιέχει τις μετρικές precision, recall :

- $precision = \frac{t_p}{t_p + f_p}$, όπου t_p εκφράζει τον αριθμό των χρονοσειρών που καταχωρήθηκαν ορθά σε ένα συγκεκριμένο τιμολόγιο από το μοντέλο, ενώ f_p εκφράζει τον αριθμό των χρονοσειρών που καταχωρήθηκαν λανθασμένα στο τιμολόγιο. Η μετρική λοιπόν εκφράζει το ποσοστό των χρονοσειρών που έχουν καταχωρηθεί σωστά στο εν λόγω τιμολόγιο.

- $recall = \frac{t_p}{t_p + f_n}$, όπου f_n εκφράζει τον αριθμό των στοιχείο που λανθασμένα δεν έχουν καταχωρηθεί σε ένα συγκεκριμένο τιμολόγιο. Η μετρική εκφράζει το ποσοστό ανάκλησης των χρονοσείρων που αρχικά άνηκαν στο τιμολόγιο.

Λόγω της μορφής του προβλήματος που αντιμετωπίζουμε μας είναι γνωστό, πως μια αρχικά δηλωμένη παροχή σε κάποιο τιμολόγιο ενδέχεται να έχει δηλωθεί λανθασμένα. Για το λόγω αυτό το ποσοστό ορθής καταχώρησης αποτελεί πιο σημαντικό από το ποσοστό ανάκλησης, καθώς ένα υψηλό score για το δεύτερο μπορεί να υποδηλώνει αδυναμία του μοντέλου να εντοπίζει χρονοσείρες που δεν ανήκουν στην πράξη στο εν λόγω τιμολόγιο.

Αν και οι δύο αυτές ποσότητες χαρακτηρίζονται ως μετρικές στην βιβλιογραφία, δεν θα πρέπει να συγχέονται με την συνάρτηση που ικανοποίει τις ιδιότητες μιας μετρικής, όπως αυτή ορίζεται μαθηματικά.

Ολοκληρώνοντας την εκπαίδευση του μοντέλου, ερχόμαστε στην πρόβλεψη του πραγματικού τιμολογίου για τις διάφορες παροχές. Η διαδικάσία αυτή προϋποθέτει την αποθήκευση ενός τελικού μοντελού μεταξύ των τριών που εκπαιδεύτηκαν, το οποίο και θα κληθεί για την εύρεση των παροχών που αποκλίνουν από το αναμενώμενο μοτίβο.

Για λόγους που εξηγήθηκαν παραπάνω, η επιλογή θα προκύψει λαμβάνοντας υπόψη την μέση τιμή της μετρικής precision για τα διαφορετικά τιμολόγια. Η μέγιστη τιμή της μετρικής βρίσκεται μέσω του πίνακα μετρικών για το μοντέλο που εκπαιδεύτηκε με χρήση της ευκλείδιας νόρμας.

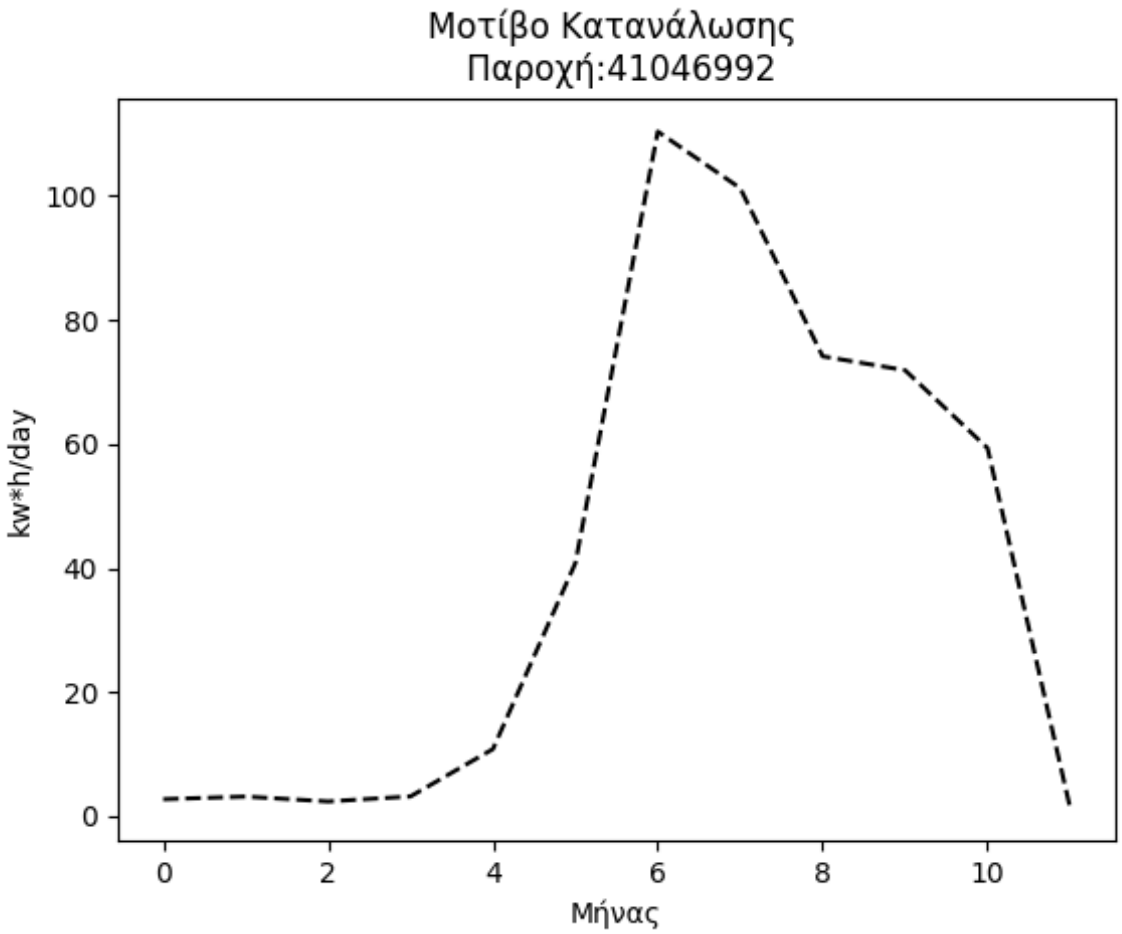
Παρακάτω παραθέτουμε των πίνακα αποκλινουσών παροχών καθώς και χαρακτηριστικά αποτελέσματα των καταναλώσεών τους.

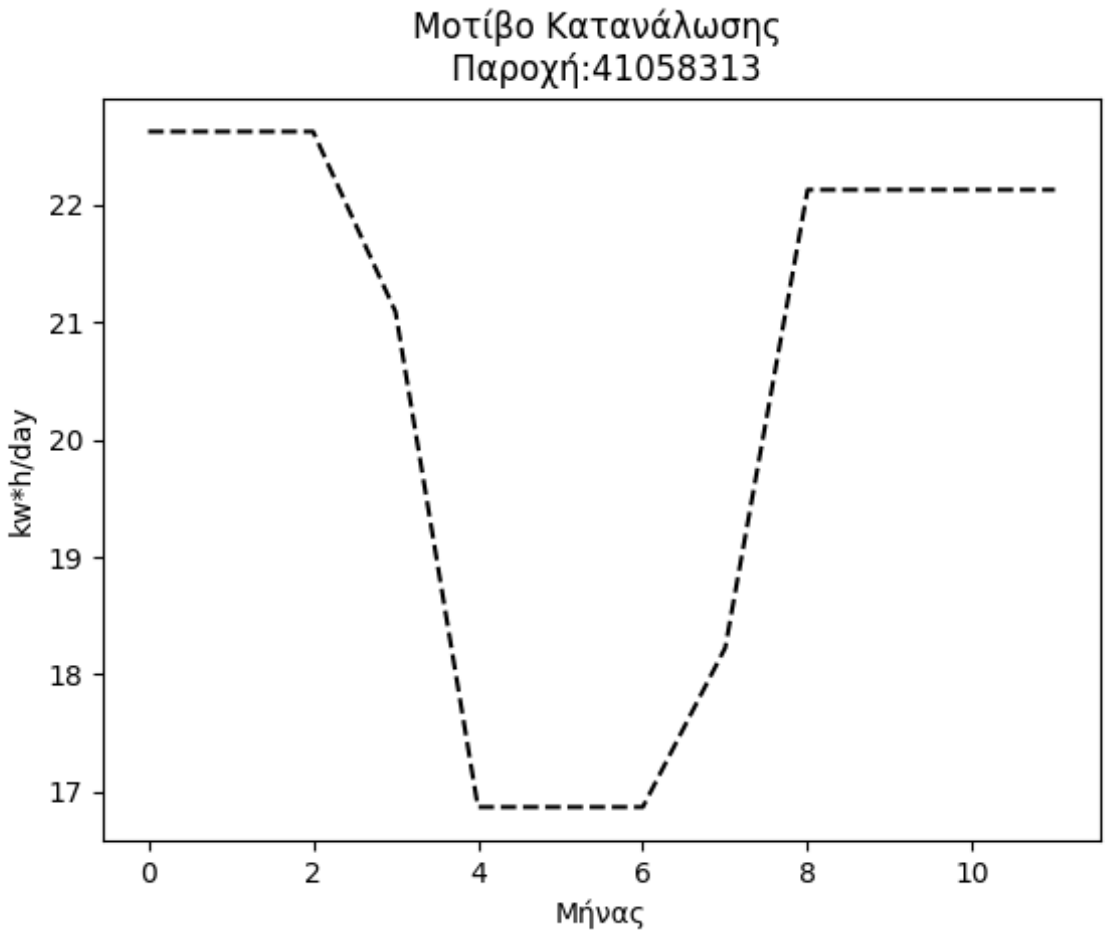
1

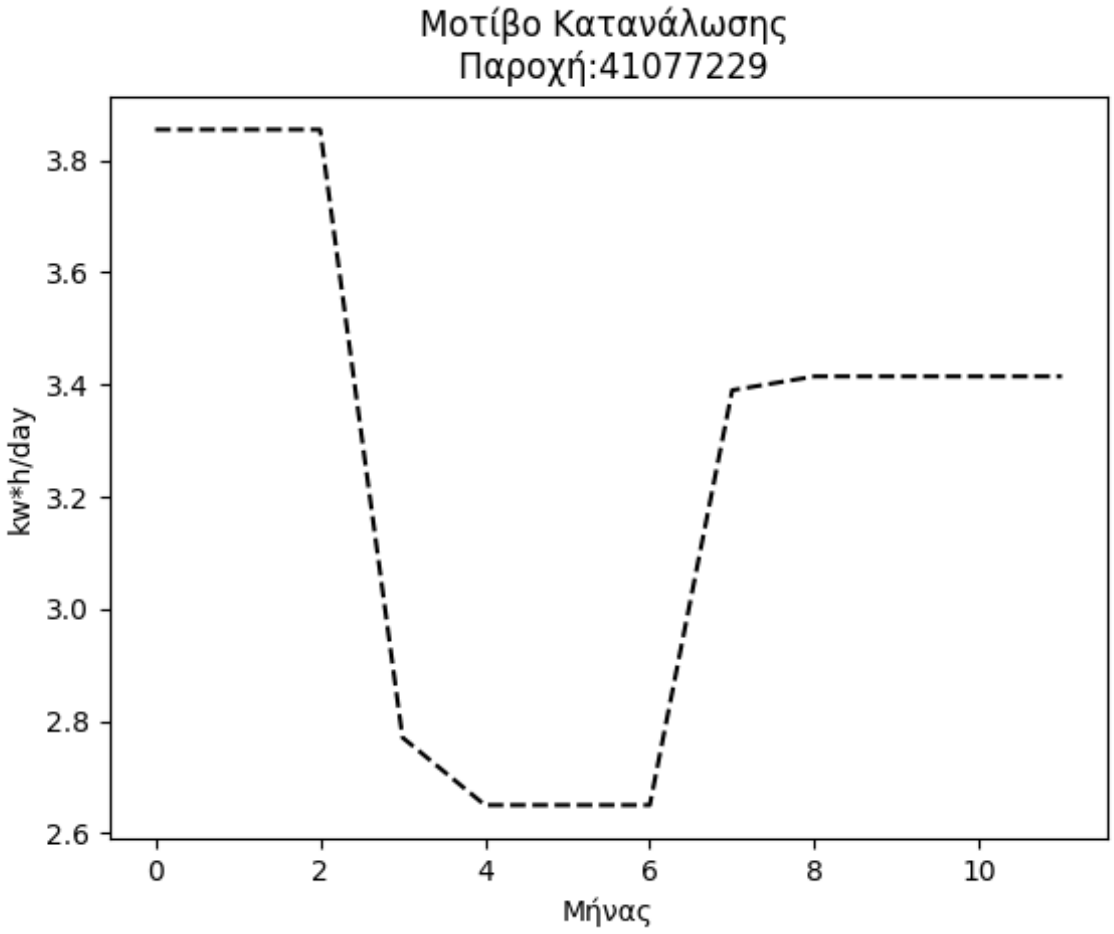
Παροχή	Δηλωμένο Τιμολόγιο	Εκτιμώμενο Τιμολόγιο
41001200	ΦΟΠ	Γ23
41001212	ΦΟΠ	Γ23
41001216	ΦΟΠ	Γ23
41001390	ΦΟΠ	Γ21
41001391	ΦΟΠ	Γ23
41001392	ΦΟΠ	Γ21
41001394	ΦΟΠ	Γ21
41001395	ΦΟΠ	ΑΓΡΟΤ
41001396	ΦΟΠ	ΑΓΡΟΤ
41001397	ΦΟΠ	Γ22
41001398	ΦΟΠ	Γ22
41001399	ΦΟΠ	Γ23
41001400	ΦΟΠ	Γ21
41001401	ΦΟΠ	Γ23
41008225	Γ21	Γ22
41008226	Γ21	Γ22
41008227	Γ21	Γ22
41008228	Γ21	Γ22
41008283	ΦΟΠ	Γ23
41008303	ΦΟΠ	Γ23
41008336	Γ21	Γ23
41012677	Γ21	Γ22
41012678	ΦΟΠ	Γ23
41012716	Γ21	Γ22
41012717	Γ21	Γ22
41016379	Γ21	Γ22
41016380	Γ21	Γ22
41016496	ΦΟΠ	Γ23
41016506	ΦΟΠ	Γ23
41016555	Γ21	ΑΓΡΟΤ
41016560	Γ21	ΦΟΠ
41016562	ΦΟΠ	Γ21
41016564	ΦΟΠ	Γ23
41019357	Γ21	ΑΓΡΟΤ
41029869	Γ22	Γ23
41036955	Γ22	ΦΟΠ
41046992	Γ22	ΑΓΡΟΤ
41047006	Γ22	Γ21
41056129	ΦΟΠ	Γ23
41056848	ΦΟΠ	Γ23
41058313	ΦΟΠ	Γ23
41058366	ΦΟΠ	Γ23
41062152	Γ21	ΑΓΡΟΤ
41063659	Γ22	Γ23
41064711	ΦΟΠ	Γ23
41065766	ΦΟΠ	Γ21
41066331	ΦΟΠ	Γ21
41067787	ΦΟΠ	Γ23
41069636	Γ21	Γ23
41069719	ΦΟΠ	Γ21
41071963	ΦΟΠ	Γ22
41072027	ΦΟΠ	Γ22
41072339	ΦΟΠ	Γ22
41072602	ΦΟΠ	Γ22
41072603	ΦΟΠ	Γ23
41073142	Γ21	Γ22
41074074	ΦΟΠ	Γ23
41074123	ΦΟΠ	Γ22
41074388	ΦΟΠ	Γ23
41075710	ΦΟΠ	Γ23
41077068	Γ22	Γ21

Αποκλίνουσες_Παροχές

41077229	ΦΟΠ	Γ23
41077451	ΦΟΠ	Γ21
41077879	Γ21	ΦΟΠ
41079306	Γ22	ΦΟΠ
41079538	Γ21	ΦΟΠ
41079784	ΦΟΠ	Γ21
41079914	Γ21	Γ23
41080004	ΦΟΠ	Γ21
41082868	ΦΟΠ	Γ23
41083380	ΦΟΠ	Γ21
41083419	Γ21	Γ23
41085515	ΦΟΠ	Γ21
41085928	Γ21	Γ23
41086128	ΦΟΠ	Γ21
41086535	Γ21	Γ22
41086883	Γ21	Γ23
41086884	Γ21	ΦΟΠ
41086885	Γ21	ΦΟΠ
41087267	Γ21	Γ22
41087274	Γ21	Γ22
41087825	Γ21	ΦΟΠ


Μοτίβο Κατανάλωσης
Παροχή:41046992
kw*h/day
Μήνας
0
20
40
60
80
100
0
2
4
6
8
10


Μοτίβο Κατανάλωσης
Παροχή:41058313
kw*h/day
Μήνας
22
21
20
19
18
17
0
2
4
6
8
10



Για την πρώτη παροχή η γραφική καθώς και το μοντέλο υποδηλώνουν ως πραγματικό τιμολόγιο το ΑΓΡΟΤ.

Για την δεύτερη και τρίτη παροχή, παρότι η γραφική υποδηλώνει σαν πραγματικό τιμολόγιο το ΦΟΠ το μοντέλο εκτιμά πως η παροχή είναι τύπου Γ23. Φαινόμενα όπως το παραπάνω απαιτούν περαιτέρω διερεύνηση και προτεινόμενες λύσεις θα δωθούν στο κεφάλαιο των επεκτάσεων.

Χρησιμοποιώντας λοιπόν τόσο των πίνακα παροχών όσο και τις γραφικές, ο Δήμος μπορεί με ευκολία να επιβεβαιώσει την αντιστοίχιση της αποκλίνουσας παροχής στο προβλεπόμενο από το μοντέλο τιμολόγιο, καθώς και να διερευνήσει αιτίες λανθασμένης αρχικής δήλωσης τιμολογίου.

4.2 Επεκτάσεις

Υπενθυμίζουμε εδώ, πως για τα τιμολόγια Γ21, Γ23 το αναμενώμενο θεωρητικό μοτίβο διαφέρει από το μοτιβό που μας προσφέρουν τα δεδομένα. Συγκεκριμένα αναφέραμε πως το μοτίβο μεταξύ των τιμολογίων ΦΟΠ, Γ21, Γ23 παρουσιάζεται να είναι παρόμοιο.

Σε περίπτωση λοιπόν που η υπόθεση σκόπιμης δήλωσης παροχών ΦΟΠ ως Γ21,Γ23 επιβεβαιωθεί, μια πιθανή μελλοντική δράση με σκοπό την βελτίωση του τελικού μοντέλου αποτελεί την συγχώνευση των τριών κατηγοριών σε μία. Με τον τρόπο αυτό αναμένουμε βελτίωση της μετρικής για την δεδομένη κατηγορία λόγω αύξησης του αντίστοιχου δείγματος.

Επίσης, λαμβάνοντας υπόψη την ικανότητα των τρίων μοντέλων να εκτιμούν καλύτερα διαφορετικά τιμολόγια, χρήσιμη θα ήταν η κατασκευή ενός meta-learning μοντέλου με σκοπό την τελική καταχώρηση της κάθε παροχής στο κατάλληλο τιμολόγιο. Δίνωντας διαφορετικό βάρος σε κάθε μοντέλο, ανάλογα με την τιμή της μετρικής precision για τα δίαφορα τιμολόγια, θα είναι εφικτή η συμβολή και των τρίων κατά την τελική καταχώρηση μιας παροχής.

Τέλος, πιθανή αποτελεί η χρήση διαφορετικού αλγορίθμου για την εκπαίδευση του μοντέλου με σκοπό την σύγκριση αποτελεσμάτων με τον αλγόριθμο K-Means.

5 Παράρτημα:Clustering(θεωρητικές έννοιες)

Στο κεφάλαιο αυτό θα εξηγήσουμε τις θεωρητικές έννοιες που διέπουν την διαδικασία του Clustering [5]. Συγκεκριμένα, θα ορίσουμε αυστηρά τα εργαλεία που χρησιμοποιήθηκαν κατά την μέθοδο B3, με τελικό στόχο την απόδειξη της σύγκλισης του αλγορίθμου K-Means.

Η διαδικασία Clustering αποτελεί μια από τις πιο διαδεδομένες μεθόδους στην ανάλυση δεδομένων και στοχεύει στην ομαδοποιήση όμοιων στοιχείων αλλά και τον διαχωρισμό στοιχείων με εμφανείς διαφορές.

Με βάση τον τύπο εισόδου και εξόδου, η παραπάνω διαδικασία διαχωρίζεται σε κατηγορίες. Παρακάτω παραθέτουμε την πιο συνιθισμένη διαδικάσια, στην οποία βασίζεται και ο αλγόριθμος
K-Means.

Έστω σύνολο $\mathbb{X}$ εφοδιασμένο με μια μετρική συνάρτηση ή μια συνάρτηση ομοιότητας:

Μια συνάρτηση $d : \mathbb{X} \times \mathbb{X} \to \mathbb{R}_+$ ορίζεται ως μετρική συνάρτηση ανν:

- $d(x, y) = 0 \Leftrightarrow x = y$
- $\forall\ x, y \in \mathbb{X}\ \ d(x, y) = d(y, x)$
- $\forall\ x, y, z \in \mathbb{X}\ \ d(x, y) \leqslant d(x, z) + d(y, z)$

Μια συνάρτηση $s : \mathbb{X} \times \mathbb{X} \to [0, 1]$ ορίζεται ως συνάρτηση ομοιότητας ανν:

- $\forall\ x, y \in \mathbb{X}\ \ s(x, y) = s(y, x)$
- $\forall\ x \in \mathbb{X}\ \ s(x, x) = 1$

Είσοδος:

- Ένα σύνολο $\mathbb{X}$ εφοδιασμένο με μετρική ή συνάρτηση ομοιότητας.
- $k \in\ \mathbb{N}$ που εκφράζει τον αριθμό των ομάδων στις οποίες θα διαχωρίζονται τα στοιχεία του $\mathbb{X}$

Έξοδος:

Ένας διαχωρισμός του συνόλου $\mathbb{X}$ σε υποσύνολα του. Δηλαδή ένα σύνολο υποσυνόλων του $\mathbb{X}$:

$C = \{C_1, C_2, \ldots, C_k\}$, τ.ω $\ \mathbb{X} = \bigcup_{i=1}^{k} C_i,\ C_i \cap C_j = \emptyset$ για $i \neq j$

Η έξοδος του αλγορίθμου K-Means που χρησιμοποίησαμε αποτελεί ίδια με την παραπάνω.

Άλλες έξοδοι για clustering αλγορίθμους είναι:

- Μια συνάρτηση $g : \mathbb{X} \to \mathbb{R}^k$ και ο διαχωρισμός C τ.ω:
  $x \overset{g}{\mapsto} g(x) = (p_1(x), p_2(x), \ldots, p_k(x))$ όπου:
  $p_i(x) = P[x\ \in\ C_i],\ C_i \in C$
- Το C και έναν γράφο, οι κόμβοι του οποίου εκφράζουν υποσύνολα του $\mathbb{X}$,
  με τα φύλλα να περιέχουν καθένα από τα σύνολα $C_i \in C\ \forall\ i \in \{1, 2, \ldots, k\}$
  και τον κορμό το αρχικό σύνολο $\mathbb{X}$. Ο γράφος αυτός ονομάζεται clustering dendogram.

Στη συνέχεια παρουσιάζουμε μια από τις πιο διαδεδομένες κατηγορίες clustering αλγορίθμων.

Η μέθοδος Linkage-Based Clustering αποτελεί μια από τις πιο απλές και διάσημες μεθόδους κατηγοριοποίησης. Κατά την μέθοδο αυτή ακολουθείται μια αλληλουχία βημάτων κατά την οποία συνδέονται τα στοιχεία του συνόλου $\mathbb{X}$ σε ομάδες. Στο πρώτο βήμα, κάθε στοιχείο του $\mathbb{X}$ ανήκει σε διαφορετική ομάδα, ενώ επαναληπτικά συνδέονται οι ομάδες που βρίσκονται κοντά μεταξύ τους.

Σε περίπτωση απουσίας συνθήκης εξόδου από τον αλγόριθμο, στο τέλος έχει διαμορφωθεί μια ομάδα, οι οποία και περιέχει κάθε στοιχείο του συνόλου. Προφανώς, τόσο η αρχική όσο και η τελική κατηγοριοποίηση των στοιχείων αποτελεί τετριμένη και επομένως απαιτείται η ύπαρξη μιας συνθήκης εξόδου με σκοπό την ουσιώδη κατηγοριοποίησή τους.

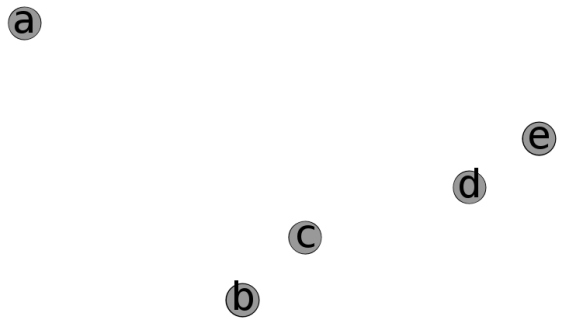
Πρώτα όμως πρέπει να αποφασίσουμε τον τρόπο μέτρησης της απόστασης μεταξύ των στοιχείων του $\mathbb{X}$, δηλαδή την μετρική συνάρτηση που θα χρησιμοποιήσουμε και στην συνέχεια τον τρόπο μέτρησης της απόστασης μεταξύ των ομάδων. Η μέτρηση της απόστασης μεταξύ των ομάδων ή γενικότερα των υποσυνόλων του $\mathbb{X}$ γίνεται μέσω επέκτασης της μετρικής συνάρτησης που έχει επιλεχθεί. Οι συναρτήσεις αυτές δεν αποτελούν μετρικές.

Παρακάτω παραθέτουμε ορισμένες επιλογές:

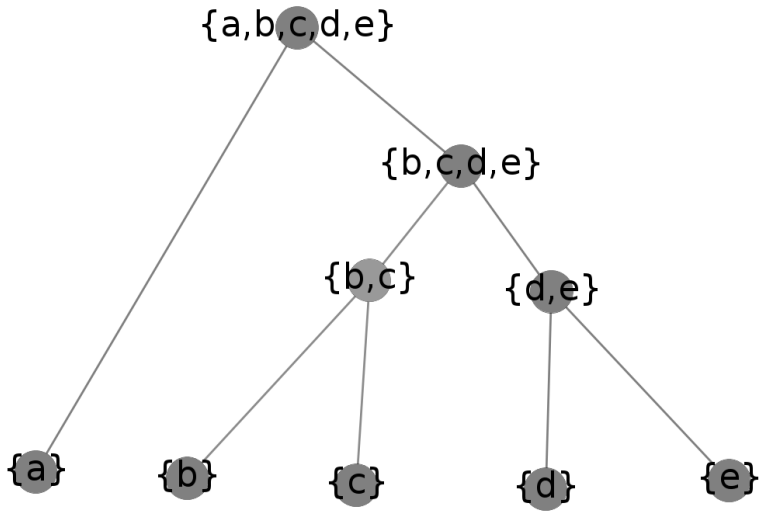
- $D(A,B) \coloneqq \min\{d(x,y) : x \in A,\ y \in B,\ A,B \subset \mathbb{X}\}$
- $D(A,B) \coloneqq \dfrac{1}{|A||B|} \displaystyle\sum_{x\in A,\ y\in B} d(x,y)\ A,B \subset \mathbb{X}$
- $D(A,B) \coloneqq \max\{d(x,y) : x \in A,\ y \in B,\ A,B \subset \mathbb{X}\}$

Όπως έχουμε ήδη αναφέρει χωρίς την χρήση μιας συνθήκης σταματημού, η κατηγορία αλγορίθμων Linkage-Based Clustering παραγεί μια τελική τετριμένη ομάδα. Το αποτέλεσμα του αλγορίθμου στην περίπτωση αυτή μπορεί να αναπαρασταθεί με την χρήση ενός δενδρογράμματος.

Για παραδείγμα αν ο χώρος $\mathbb{X}$ είναι: $\mathbb{X} = \{a,b,c,d,e\} \subset \mathbb{R}^2$, με την θέση των σημείων στον $\mathbb{R}^2$ να είναι:



Τότε η αλγοριθμική διαδικασία, με χρήση της μετρικής
$D(A, B) = \min\{d(x, y) : x \in A,\ y \in B,\ A, B \subset \mathbb{X}\}$, παρατίθεται στο παρακάτω δενδρόγραμμα:



Με σκοπό των διαχωρισμό του χώρου σε παραπάνω από ένα σύνολα, απαραίτητη αποτελεί η χρήση μιας συνθήκης σταματημού.

Κοινά κριτηρία σταματημού αποτελούν:

- Προκαθορισμένος αριθμός $k \in \mathbb{N}$ ομάδων clusters.
  Ο αλγοριθμός σταματά, μόλις ο αριθμός των ομάδων ισούται με k.
- Προκαθορισμένο άνω φράγμα απόστασης $r \in \mathbb{R}_+$.
  Συχνά $r = \alpha \max\{d(x, y)\ x, y \in \mathbb{X}\},\ a < 1$. Ο αλγόριθμος σταματά, όταν όλες οι αποστάσεις ανάμεσα στις ομάδες ξεπερνούν το r.

Μια άλλη δημοφιλής προσέγγιση για την clustering διαδικασία προκύπτει μέσω της χρήσης μιας συνάρτησης κόστους, με όρισμα ένα διαχώρισμο. Σκοπός του αλγορίθμου, αποτελεί την ευρεση του διαχωρισμού για τον οποίο η εν λόγω συνάρτηση θα ελαχιστοποιείται.

Με τον τρόπο αυτό το πρόβλημα διαχωρισμού ανάγεται σε πρόβλημα βελτιστοποίησης, με την αντικειμενική συνάρτηση να ορίζεται μέσω του πιθανού διαχωρισμού $C = (C_1, C_2, \ldots, C_k)$ και του μετρικού χώρου (X, d) και να καταλήγει στο $\mathbb{R}_+$.

Για πολλούς clustering αλγορίθμους, μπορεί να δειχθεί, πως το πρόβλημα βελτιστοποίησης αποτελεί NP-Hard ή ακόμα και κατά προσέγγιση NP-Hard για δεδομένους περιορισμούς.
Για τον λόγω αυτό, με σκοπό την εκπαίδευση του μοντέλου μηχανικής μάθησης, αρκούμαστε στην εκτέλεση ενός προσεγγιστικού αλγορίθμου έναντι της υλοποίησης της πραγματικής λύσης του προβλήματος βελτιστοποίησης.

Παρακάτω παραθέτουμε ορισμένες αντικειμενικές συναρτήσεις αυτής της μορφής:

- Αντικειμενική συνάρτηση K-Means:

  Κατά τον αλγόριθμο K-Means το σύνολο δεδομένων διαχωρίζεται σε ξένα υπόσυνολα του $\mathbb{X}$ (C_i), το καθένα από τα οποία εκπροσωπείται από ένα κεντροειδές μ_i.
  Υποθέτουμε πως το σύνολο $\mathbb{X}$ περιέχεται σε έναν μετρικό χώρο $(\mathbb{X}', d)$ $(\mathbb{X} \subset \mathbb{X}')$ και ακόμα πως $\mu_i \in \mathbb{X}' \ \ \forall\, i \in \{1, 2, \ldots, k\}$.

  Η αντικειμενική συνάρτηση μετρά το άθροισμα των τετραγώνων της απόστασης κάθε σημείου από το κεντροειδές της ομάδας στην οποία κατηγοριοποιείται.

  Το κεντροειδές κάθε ομάδας (C_i) ορίζεται ως:

  $$\mu_i(C_i) = \underset{\mu \in \mathbb{X}'}{\operatorname{argmin}} \sum_{x \in C_i} d(x, \mu)^2$$

  Τότε η αντικειμενική συνάρτηση παίρνει την μορφή:

  $$G_{k-means}(k, (\mathbb{X}, d), (C_1, C_2, \ldots, C_k) = \sum_{i=1}^{k} \sum_{x \in C_i} d(x, \mu_i(C_i))^2$$

  Διαφορετικά:

  $$G_{k-means}(k, (\mathbb{X}, d), (C_1, C_2, \ldots, C_k) = \min_{\mu_1, \mu_2, \ldots, \mu_k \in \mathbb{X}'} \sum_{i=1}^{k} \sum_{x \in C_i} d(x, \mu_i)^2$$

  Μπορεί να δειχθεί, πως το παραπάνω πρόβλημα ελαχιστοποίησης επιλύεται από τις μέσες τιμές των στοιχείων που ανήκουν στις ομάδες $C_i\big(\mu_i(C_i)\big) \ \forall\ i \ \in \ \{1, 2, \ldots, k\}$.

Έστω:

- $\mathbb{X}, \mathbb{X}' \subset \mathbb{R}^2,\ d(x,y) = \|x-y\|_2$
- $\mu_1, \mu_2, \ldots, \mu_k \in \mathbb{X}',\quad F(\mu_1, \mu_2, \ldots, \mu_k) \coloneqq \sum_{i=1}^{k} \sum_{x \in C_i} \|x - \mu_i\|_2^2$

Τότε:

$$= \sum_{i=1}^{k} \sum_{x \in C_i} \left[\big(x_1 - \mu_{i_1} + \mu_{i_1}(C_i) - \mu_{i_1}(C_i)\big)^2 + \big(x_2 - \mu_{i_2} + \mu_{i_2}(C_i) - \mu_{i_2}(C_i)\big)^2 \right]$$

$$= \sum_{i=1}^{k} \sum_{x \in C_i} \left[\big(x_1 - \mu_{i_1}(C_i)\big)^2 + 2\big(x - \mu_{i_1}(C_i)\big)\big(\mu_{i_1}(C_i) - \mu_{i_1}\big) + \big(\mu_{i_1}(C_i) - \mu_{i_1}\big)^2 \right]$$
$$+ \sum_{i=1}^{k} \sum_{x \in C_i} \left[\big(x_2 - \mu_{i_2}(C_i)\big)^2 + 2\big(x - \mu_{i_2}(C_i)\big)\big(\mu_{i_2}(C_i) - \mu_{i_2}\big) + \big(\mu_{i_2}(C_i) - \mu_{i_2}\big)^2 \right]$$

$$= \sum_{i=1}^{k} \sum_{x \in C_i} \left[\big(x_1 - \mu_{i_1}(C_i)\big)^2 + \big(x_2 - \mu_{i_2}(C_i)\big)^2 + \big(\mu_{i_1}(C_i) - \mu_{i_1}\big)^2 + \big(\mu_{i_2}(C_i) - \mu_{i_2}\big)^2 \right]$$
$$+ \sum_{i=1}^{k} \sum_{x \in C_i} \left[2\big(x - \mu_{i_1}(C_i)\big)\big(\mu_{i_1}(C_i) - \mu_{i_1}\big) + 2\big(x - \mu_{i_2}(C_i)\big)\big(\mu_{i_2}(C_i) - \mu_{i_2}\big) \right]$$

$$= \sum_{i=1}^{k} \sum_{x \in C_i} \left[\big(x_1 - \mu_{i_1}(C_i)\big)^2 + \big(x_2 - \mu_{i_2}(C_i)\big)^2 + \big(\mu_{i_1}(C_i) - \mu_{i_1}\big)^2 + \big(\mu_{i_2}(C_i) - \mu_{i_2}\big)^2 \right]$$
$$+ \sum_{i=1}^{k} \left[2\big(|C_i|\ \mu_{i_1}(C_i) - |C_i|\ \mu_{i_1}(C_i)\big)\big(\mu_{i_1}(C_i) - \mu_{i_1}\big) + 2\big(|C_i|\ \mu_{i_2}(C_i) - |C_i|\ \mu_{i_2}(C_i)\big)\big(\mu_{i_2}(C_i) - \mu_{i_2}\big) \right]$$

$$= \sum_{i=1}^{k} \sum_{x \in C_i} \left[\big(x_1 - \mu_{i_1}(C_i)\big)^2 + \big(x_2 - \mu_{i_2}(C_i)\big)^2 + \big(\mu_{i_1}(C_i) - \mu_{i_1}\big)^2 + \big(\mu_{i_2}(C_i) - \mu_{i_2}\big)^2 \right]$$

$$= \sum_{i=1}^{k} \sum_{x \in C_i} \left[\big(x_1 - \mu_{i_1}(C_i)\big)^2 + \big(x_2 - \mu_{i_2}(C_i)\big)^2 \right] + \sum_{i=1}^{k} \sum_{x \in C_i} \left[\big(\mu_{i_1}(C_i) - \mu_{i_1}\big)^2 + \big(\mu_{i_2}(C_i) - \mu_{i_2}\big)^2 \right]$$

$$\geqslant \sum_{i=1}^{k} \sum_{x \in C_i} \left[\big(x_1 - \mu_{i_1}(C_i)\big)^2 + \big(x_2 - \mu_{i_2}(C_i)\big)^2 \right]$$

Άρα πράγματι:
$\big(\mu_1(C_1), \mu_2(C_2), \ldots, \mu_k(C_k)\big) = \underset{\mu_1, \mu_2, \ldots, \mu_k \in \mathbb{X}'}{\operatorname{argmin}} F(\mu_1, \mu_2, \ldots, \mu_k)$ □

- Αντικειμενική συνάρτηση K-Medoids:

  Η αντικειμενική αυτή συνάρτηση είναι παρόμοια με την K-Means με την διαφορά πως τα κεντροειδοί $\mu_i \in \mathbb{X} \, \forall \, i \in \{1, 2, \ldots, k\}$
  Έτσι έχουμε:

$$G_{k-medoids}(k, (\mathbb{X}, d), (C_1, C_2, \ldots, C_k) = \min_{\mu_1, \mu_2, \ldots, \mu_k \in \mathbb{X}} \sum_{i=1}^{k} \sum_{x \in C_i} d(x, \mu_i)^2$$

- Αντικειμενική συνάρτηση K-Median:

$$G_{k-median}(k, (\mathbb{X}, d), (C_1, C_2, \ldots, C_k) = \min_{\mu_1, \mu_2, \ldots, \mu_k \in \mathbb{X}} \sum_{i=1}^{k} \sum_{x \in C_i} d(x, \mu_i)$$

Οι παραπάνω αντικειμενικές συναρτήσεις χρησιμοποιούνται στην επίλυση clustering προβλημάτων,στόχος των οποίων αποτελεί η εύρεση των κεντροειδών κάθε ομάδας.
Συγκεκριμένα οι λύσεις για τα αντίστοιχα clustering προβλήματα καθορίζονται από ένα σύνολο κεντροείδων, με την διαδικασία clustering να κατηγοριοποιεί κάθε στοιχείο του $\mathbb{X}$ στο κοντινότερο, με βάση την μετρική συνάρτηση, κεντροειδές.

Έτσι η γενικευμένη αντικειμένικη συνάρτηση, για προβλήματα αυτής της μορφής γράφεται ως:

$$G(k, (\mathbb{X}, d), (C_1, C_2, \ldots, C_k) = \min_{\mu_1, \mu_2, \ldots, \mu_k \in \mathbb{X}'} \sum_{i=1}^{k} \sum_{x \in C_i} f\big(d(x, \mu_i)\big)$$

Όπου $f : \mathbb{R}_+ \to \mathbb{R}_+$ μονότονη συνάρτηση και $\mathbb{X} \subset \mathbb{X}'$.

Σε άλλες περιπτώσεις προβλημάτων κατηγοριοποίησης, η εύρεση του κεντροειδούς δεν αποτελεί απαραίτητη για την ταξινόμηση των στοιχείων του $\mathbb{X}$.
Μια συνάρτηση που αντιστοιχεί σε ένα τέτοιο πρόβλημα αποτελεί την συνάρτηση SOD(sum of incluster distances):

$$G_{SOD}(k, (\mathbb{X}, d), (C_1, C_2, \ldots, C_k) = \min_{\mu_1, \mu_2, \ldots, \mu_k \in \mathbb{X}'} \sum_{i=1}^{k} \sum_{x, y \in C_i} d(x, y)$$

Κατά το πρόβλημα αυτό οι ομάδες διαμορφωνονται με τέτοιο τρόπο, ώστε οι αποστάσεις μεταξύ των σημείων εντός της κάθε ομάδας να είναι ελάχιστες.

Όπως έχουμε ήδη αναφέρει, για πολλά clustering προβλημάτα, η πραγματική λύση της διαδικασίας ελαχιστοποίησης αποτελεί υπολογιστικά μη εφικτή. Συγκεκριμένα μπορεί να δειχθεί, πως το πρόβλημα που αντιστοιχεί στην αντικειμενική συνάρτηση K-Means, αποτελεί NP-Hard ή και NP-Hard κατά προσέγγιση για δεδομένους περιορισμούς.

Για τον λόγω αυτό, η λύση του εν λόγω προβλήματος προσεγγίζεται μέσω του παρακάτω απλοποιημένου αλγορίθμου:

- Είσοδος: $\mathbb{X} \subset \mathbb{R}^n$, αριθμός των ομάδων $k \in \mathbb{N}$, μετρική συνάρτηση d
- Αρχικοποίηση: Επέλεξε τύχαια τα αρχικά κεντροειδοί από το στοιχεία του $\mathbb{X}$
- Επανέλαβε μέχρι την σύγκλιση:
  - $\forall\ i\ \in \{1, 2, \ldots, k\}\ C_i \coloneqq \{x\ \in \mathbb{X} : i = \underset{j \in [k]}{\operatorname{argmin}} \|x - \mu_j\|\}$
  - $\forall\ i\ \in \{1, 2, \ldots, k\}\ \mu_i \coloneqq \frac{1}{|C_i|} \sum_{x \in C_i} x$

Στο σημείο αυτό παραθέτουμε την απόδειξη σύγκλισης του παραπάνω αλγορίθμου.

Θεώρημα 1
Έστω $(C_1, C_2, \ldots, C_k)$ διαχωρισμός του $\mathbb{X}$, $G_{k-means}$ αντικειμενική συνάρτηση,
$a_n = G^n_{k-means}\big(k, (\mathbb{X}, d), (C^n_1, C^n_2, \ldots, C^n_k)\big)$ η τιμή της αντικειμενικής συνάρτησης στην n-οστή επανάληψη του αλγορίθμου και d η ευκλίδεια μετρική.

Τότε:

$$0 \leqslant a_n \leqslant a_{n-1} \ \forall \ n \in \mathbb{N}$$

Απόδειξη.

Η αντικειμενική συνάρτηση γράφεται στην μορφή:

$$G_{k-means}(k, (\mathbb{X}, d), (C_1, C_2, \ldots, C_k) = \sum_{i=1}^{k} \sum_{x \in C_i} \|x - \mu_i(C_i)\|_2^2$$

Η τιμή της αντικειμενικής συνάρτησης στην n-οστή επανάληψη του αλγορίθμου γράφεται ως:

$$G^n_{k-means}(k, (\mathbb{X}, d), \big(C^n_1, C^n_2, \ldots, C^n_k)\big) = \sum_{i=1}^{k} \sum_{x \in C^n_i} \|x - \mu_i(C^n_i)\|_2^2$$

Όπου:

$$\mu_i(C_i) = \underset{\mu \in \mathbb{X}'}{\operatorname{argmin}} \sum_{x \in C_i} \|x - \mu\|_2^2 = \frac{1}{|C_i|} \sum_{x \in C_i} x$$

Έστω $n \in \mathbb{N}$ η τρέχουσα επανάληψη του αλγορίθμου και $(C^{n-1}_1, C^{n-1}_2, \ldots, C^{n-1}_k)$ ο διαχωρισμός που προκύπτει από την προηγούμενη επανάληψη του αλγορίθμου.

Θέτω $\mu^{n-1}_i = \mu_i(C^{n-1}_i) \ \forall \ i \ \in \ \{1, 2, \ldots, k\}$

Τότε από τον ορισμό της συνάρτησης $G_{k-means}$ έπεται ότι:

$$a_n \leqslant \sum_{i=1}^{k} \sum_{x \in C^n_i} \|x - \mu^{n-1}_i\|^2 \tag{1}$$

Αφού $\mu^n_i = \underset{\mu \in \mathbb{X}'}{\operatorname{argmin}} \sum_{x \in C^n_i} \|x - \mu\|_2^2$

Επιπλέον, από τον ορισμό του διαχωρισμού $(C^n_1, C^n_2, \ldots, C^n_k)$ προκύπτει ότι:

$$(C^n_1, C^n_2, \ldots, C^n_k) = \underset{(C_1, C_2, \ldots, C_k)}{\operatorname{argmin}} \sum_{i=1}^{k} \sum_{x \in C_i} \|x - \mu^{n-1}_i\|^2$$

Επομένως:

$$\sum_{i=1}^{k} \sum_{x \in C_i^n} \|x - \mu_i^{n-1}\|^2 \leqslant \sum_{i=1}^{k} \sum_{x \in C_i^{n-1}} \|x - \mu_i^{n-1}\|^2$$

Δηλαδή:

$$\sum_{i=1}^{k} \sum_{x \in C_i^n} \|x - \mu_i^{n-1}\|^2 \leqslant a_{n-1} \tag{2}$$

Έτσι τελικά από (1), (2) έχουμε:

$$0 \leqslant a_n \leqslant \sum_{i=1}^{k} \sum_{x \in C_i^n} \|x - \mu_i^{n-1}\|^2 \leqslant a_{n-1}$$

$$0 \leqslant a_n \leqslant a_{n-1}$$

Η ακολουθία των τιμών της αντικειμενικής συνάρτησης είναι φθίνουσα και κάτω φραγμένη.
Ακόμα $a_n \in \mathbb{R}\ \forall\ n \in \mathbb{N}$.
Από γνωστό θεώρημα η ακολουθία συγκλίνει. □

6 Επίλογος

Στο σημείο αυτό ολοκληρώνεται η τρέχουσα διπλωματική εργασία. Τα αποτελέσματα που παρατέθηκαν παραπάνω θα παραδωθούν στον Δήμο Τρικκαίων για αξιολόγηση και ανακάλυψη ευρημάτων.

Σε περίπτωση διάθεσης από τον πελάτη των επιθυμητών δεδομένων, η ανάλυση θα επεκταθεί περαιτέρω, χρησιμοποιώντας τις σημασιολογικές συσχετίσεις μεταξύ των συνόλων που παρουσιάστηκαν στην δεύτερη ενότητα. Ακόμα, οι μέθοδοι που αφορούν τα δεδομένα κατανάλωσεις θα εμπλουτιστούν, λαμβάνοντας υπόψη σχόλια και παρατηρήσεις από τον πελάτη για την χρησιμότητά τους, αλλά και τις πιθανές μελλοντικές του ανάγκες.

Τελος για την μεθοδολογία B3, που αφορά την εύρεση των κοινών μοτίβων κατανάλωσης, θα συζητηθούν μελλοντικές δράσεις βελτίωσεις του τελικού μοντέλου, μεταξύ των οποίων θα περιλαμβάνονται και οι προτάσεις που παρατέθηκαν στο κεφάλαιο ″επεκτάσεις″.

Βιβλιογραφία

[1] Szilagyi, Ioan & Wira, Patrice. (2016). *Ontologies and Semantic Web for the Internet of Things - a survey. 6949-6954.* 10.1109/IECON.2016.7793744.

[2] Fotopoulou, Eleni & Zafeiropoulos, Anastasios & Terroso-Saenz, Fernando & Şimşek, Umutcan & González Vidal, Aurora & Tsiolis, George & Gouvas, Panagiotis & Liapis, Paris & Fensel, Anna & Skarmeta, Antonio. (2017). *Providing Personalized Energy Management and Awareness Services for Energy Efficiency in Smart Buildings.* Sensors. 17. 2054. 10.3390/s17092054.

[3] H. Sakoe and S. Chiba
Dynamic programming algorithm optimization for spoken word recognition
IEEE Transactions on Acoustics, Speech, and Signal Processing vol. 26, no. 1, pp. 43-49, February 1978, doi: 10.1109/TASSP.1978.1163055.

[4] Marco Cuturi, Mathieu Blondel *Soft-DTW: a Differentiable Loss Function for Time-Series* 2017
`https://arxiv.org/abs/1703.01541v2`

[5] S. Shalev-Shwartz and S. Ben-David, *Understanding Machine Learning: From Theory to Algorithms.* Cambridge University Press, 2014.