ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

Διατμηματικό Πρόγραμμα Μεταπτυχιακών Σπουδών:
«ΠΑΡΑΓΩΓΗ ΚΑΙ ΔΙΑΧΕΙΡΙΣΗ ΕΝΕΡΓΕΙΑΣ»

ΑΝΑΛΥΣΗ, ΚΑΤΗΓΟΡΙΟΠΟΙΗΣΗ ΚΑΙ ΠΡΟΒΛΕΨΗ ΦΟΡΤΙΟΥ ΚΑΤΑΝΑΛΩΤΩΝ ΕΛΛΗΝΙΚΗΣ ΑΓΟΡΑΣ ΗΛΕΚΤΡΙΚΗΣ ΕΝΕΡΓΕΙΑΣ

Μεταπτυχιακή Διπλωματική Εργασία
του
ΕΥΑΓΓΕΛΟΥ Π . ΒΟΥΤΣΙΝΑ

Επιβλέπων: Νικόλαος Χατζηαργυρίου, Καθηγητής Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών

Αθήνα, Ιούλιος 2021

ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

Διατμηματικό Πρόγραμμα Μεταπτυχιακών Σπουδών:
«ΠΑΡΑΓΩΓΗ ΚΑΙ ΔΙΑΧΕΙΡΙΣΗ ΕΝΕΡΓΕΙΑΣ»

ΑΝΑΛΥΣΗ, ΚΑΤΗΓΟΡΙΟΠΟΙΗΣΗ ΚΑΙ ΠΡΟΒΛΕΨΗ ΦΟΡΤΙΟΥ ΚΑΤΑΝΑΛΩΤΩΝ ΕΛΛΗΝΙΚΗΣ ΑΓΟΡΑΣ ΗΛΕΚΤΡΙΚΗΣ ΕΝΕΡΓΕΙΑΣ

Μεταπτυχιακή Διπλωματική Εργασία
του

ΕΥΑΓΓΕΛΟΥ Π . ΒΟΥΤΣΙΝΑ

Επιβλέπων: Νικόλαος Χατζηαργυρίου, Καθηγητής Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 7η Ιουλίου 2021.

.....	.....	.....
Ν. Χατζηαργυρίου	Σ. Παπαθανασίου	Γ. Τσεκούρας
Καθηγητής Ε.Μ.Π.	Καθηγητής Ε.Μ.Π.	Επίκουρος Καθηγητής ΠΑ.Δ.Α.

Αθήνα, Ιούλιος 2021

.....
Ευάγγελος Π. Βουτσινάς
Διπλωματούχος Ηλεκτρολόγος Μηχανικός και τεχνολογίας Υπολογιστών Ε.Μ.Π.

Copyright © Ευάγγελος Π. Βουτσινάς, 2021
Με επιφύλαξη παντός δικαιώματος. All rights reserved.

Απαγορεύεται η αντιγραφή, αποθήκευση και διανομή της παρούσας εργασίας, εξ ολοκλήρου ή τμήματος αυτής, για εμπορικό σκοπό. Επιτρέπεται η ανατύπωση, αποθήκευση και διανομή για σκοπό μη κερδοσκοπικό, εκπαιδευτικής ή ερευνητικής φύσης, υπό την προϋπόθεση να αναφέρεται η πηγή προέλευσης και να διατηρείται το παρόν μμήνυμα. Ερωτήματα που αφορούν τη χρήση της εργασίας για κερδοσκοπικό σκοπό πρέπει να απευθύνονται προς τον συγγραφέα. Οι απόψεις και τα συμπεράσματα που περιέχονται σε αυτό το έγγραφο εκφράζουν τον συγγραφέα και δεν πρέπει να ερμηνευθεί ότι αντιπροσωπεύουν τις επίσημές θέσεις του Εθνικού Μετσόβιου Πολυτεχνείου.

Το σύνολο της εργασίας αποτελεί πρωτότυπο έργο, παραχθέν από τον Ευάγγελο Βουτσινά, και δεν παραβιάζει δικαιώματα τρίτων καθ' οιονδήποτε τρόπο.
Υλικό που περιέχεται στην εργασία, το οποίο δεν έχει παραχθεί από τον ίδιο, είναι ευδιάκριτο και αναφέρεται ρητώς εντός του κειμένου της εργασίας ως προϊόν εργασίας τρίτου, σημειώνοντας με παρομοίως σαφή τρόπο τα στοιχεία ταυτοποίησής του, ενώ παράλληλα βεβαιώνεται πως στην περίπτωση χρήσης αυτούσιων γραφικών αναπαραστάσεων, εικόνων, γραφημάτων κλπ., ο συγγραφέας έχει λάβει τη χωρίς περιορισμούς άδεια του κατόχου των πνευματικών δικαιωμάτων για τη συμπερίληψη και επακόλουθη δημοσίευση του υλικού αυτού.

Μεταπτυχιακή Εργασία: **«Ανάλυση, κατηγοριοποίηση και πρόβλεψη φορτίου των καταναλωτών Μέσης Τάσης της Ελληνικής Αγοράς Ηλεκτρικής Ενέργειας»**

Φοιτητής: **Ευάγγελος Π. Βουτσινάς**

Επιβλέπων: **Νικόλαος Χατζηαργυρίου, Καθηγητής Σχολής Η.Μ.Μ.Υ. Ε.Μ.Π.**

Ακαδημαϊκό έτος: **2020-2021**

Περίληψη

Η παρούσα διπλωματική εργασία πραγματεύεται τη μελέτη της κατηγοριοποίησης και της βραχυπρόθεσμης πρόβλεψης στα Συστήματα Ηλεκτρικής Ενέργειας, καθώς και την υλοποίηση τους με Python για πελάτες Μέσης Τάσης του Ελληνικού Διασυνδεδεμένου Δικτύου Διανομής Ηλεκτρικής Ενέργειας. Εκπονήθηκε στο Διατμηματικό Πρόγραμμα Μεταπτυχιακών Σπουδών «Παραγωγή και Διαχείριση Ενέργειας» του Εθνικού Μετσόβιου Πολυτεχνείου με επιβλέποντα καθηγητή τον κ. Νικόλαο Χατζηαργυρίου.

Η εργασία δίνει λύση στο πρόβλημα της υλοποίησης πρόβλεψης όταν δεν είναι γνωστά τα δεδομένα της χρονοσειράς ζήτησης φορτίου ενός καταναλωτή. Η απάντηση βρίσκεται στην υλοποίηση κατηγοριοποίησης σε ομάδες καταναλωτών με γνωστές χρονοσειρές και ταξινόμηση του άγνωστου σε αυτές. Πρακτικά με την κατηγοριοποίηση που υλοποιείται, ένας πελάτης με γνωστό μόνο το τυπικό προφίλ κατανάλωσης φορτίου του μπορεί να καταχωρηθεί στην αντίστοιχη ομάδα που είναι πιο αντιπροσωπευτική για αυτόν και να γίνει εκτίμηση της ζήτησης φορτίου του.

Αρχικά, παρουσιάζονται οι μέθοδοι αναγνώρισης προτύπων που χρησιμοποιούνται και η σημασία τους για το σύστημα ηλεκτρικής ενέργειας. Στη συνέχεια, αναλύονται οι κυριότερες τεχνικές κατηγοριοποίησης, καθώς και οι δείκτες αξιολόγησης που χρησιμοποιούνται για τη σύγκρισή τους. Ακόμη, καταγράφονται οι σημαντικότερες μεθοδολογίες που ακολουθούνται για την υλοποίηση της βραχυπρόθεσμης πρόβλεψης και αναλύονται τα τεχνητά νευρωνικά δίκτυα που θα χρησιμοποιηθούν στην παρούσα εργασία.

Έπειτα, περιγράφεται η διαδικασία που ακολουθήθηκε για την υλοποίηση της κατηγοριοποίησης των τυπικών προφίλ των καταναλωτών, δηλαδή η συλλογή και η διαμόρφωση των δεδομένων που τροφοδοτήθηκαν στον αλγόριθμο K-μέσων. Τα πρώτα αποτελέσματα που εξήχθησαν ανέδειξαν την ανάγκη εύρεσης βελτίωσης του αλγορίθμου λόγω της φύσης των δεδομένων, τα οποία επειδή συγκεντρώνονται σημαντικά σε μικρό τμήμα του συνολικού εύρους, ταξινομούνται σε ομάδες εξαιρετικά άνισα με μία εξ αυτών να περιέχει τη συντριπτική πλειοψηφία των δεδομένων. Έτσι, εφαρμόστηκαν οι εξής βελτιώσεις: η μείωση των διαστάσεων των δεδομένων, ο αλγόριθμος Mini Batch K-means, η υλοποίηση καθοδηγούμενης κατηγοριοποίησης με K-means (guided K-means) που έδωσε τη βέλτιστη απόκριση, καθώς και ο συνδυασμός τους.

Το επόμενο βήμα είναι η περιγραφή και η υλοποίηση μοντέλου πρόβλεψης το οποίο κατασκευάζεται για τη βραχυπρόθεσμη πρόβλεψη κάθε μίας από τις ομάδες που προκύπτουν από το βήμα της κατηγοριοποίησης. Έτσι, δημιουργούνται 15 μοντέλα που έχουν κατασκευαστεί με τεχνητά νευρωνικά δίκτυα και που το καθένα τροφοδοτείται από τη χρονοσειρά που περιγράφει το προφίλ της αντίστοιχης ομάδας για το έτος 2019. Ως αποτέλεσμα προκύπτει η πρόβλεψη φορτίου των επόμενων 24 ωρών για κάθε τυπικό προφίλ των Ελλήνων πελατών Μέσης Τάσης του διασυνδεδεμένου συστήματος, όπου παρατηρείται η αύξηση της ακρίβειάς της για ομάδες με μεγαλύτερη κατανάλωση.

Τέλος, αξιολογούνται τα αποτελέσματα που εξήχθησαν, παρουσιάζονται τα συμπεράσματα που προέκυψαν και προτείνονται μελλοντικά βήματα έρευνας με βάση την εργασία αυτή.

Λέξεις - Κλειδιά

Κατηγοριοποίηση – Βραχυπρόθεσμη πρόβλεψη φορτίου – Python – Αλγόριθμος K-μέσων – Μείωση Διαστάσεων – Mini Batch K-μέσων – Κατευθυνόμενη Κατηγοριοποίηση – Τεχνητό Νευρωνικό Δίκτυο

Post-graduate thesis:	**«Analysis, classification and load forecast of Medium Voltage consumers of the Greek electricity market»**
Student:	**Evangelos P. Voutsinas**
Supervisor:	**Nikolaos Hatziargyriou, Professor, E.C.E. N.T.U.A.**
Academic year:	**2020-2021**

Abstract

This post-graduate thesis focuses on the study and implementation of classification and short-term forecasting with Python in systems of Electric Power and particularly with regards to customers connected to the Greek Medium Voltage Power Grid. It was conducted for the Inter – Departmental Postgraduate Course: «Energy Production and Management» coordinated by the National Technical University of Athens and supervised by Professor Nikolaos Hatziargyriou.

The work solves the problem of forecast implementation when the load time series of a consumer are not known. The answer lies in implementing classification into consumer groups with known time series and classifying the unknown in them. Practically with the categorization that is implemented, a customer with known only his typical consumption profile can be registered in the respective group that is more representative for him and his load demand can be estimated.

Initially, the pattern recognition methods used are presented along with their importance for the electricity system. Then, the main categorization techniques are analyzed, as well as the evaluation indicators used to compare them. Furthermore, the most important methodologies the short-term forecasting are described and an analysis of artificial neural networks used to implement predictions in this thesis provided.

The procedure followed for the implementation of clustering the typical consumer profiles is described, starting with the collection and configuration of the data fed to the K-means algorithm. The first results obtained revealed the necessity to find an algorithm improvement due to the nature of the data that are significantly concentrated in a part of the total range. This element results in classification into extremely unequal groups, where only one of them is populated by the vast majority of data. Thus, the following improvements were implemented: the reduction of data dimensions, the Mini Batch K-means algorithm, the implementation of guided K-means categorization, as well as their combination. The best results were taken from guided K-means for 15 clusters in total.

Subsequently, a forecasting model is described and implemented for the short-term prediction of each of the groups resulting from the classification step. Thus, 15 models are created with artificial neural networks with the same architecture between them. Each one is fed by the time series that describes the typical profile of the corresponding cluster for the year 2019 which consists of the mean values of all consumers that are part of the same group. This results in a load forecast for the next 24 hours for each typical profile of every group that describes Greek Medium Voltage customers of the interconnected system. The outcome of forecasting shows that accuracy increases for clusters with higher energy consumption.

Finally, the results are evaluated and the conclusions drawn are presented, followed by the prospects and suggestions for future research on this topic.

Key words

Clustering – Short Term Load Forecast – Python – K-means – Dimensionality reduction – Mini Batch K-means – Guided Clustering – Artificial Neural Network

Πρόλογος

Η παρούσα διπλωματική εργασία έχει ως αντικείμενο τη μελέτη, την ανάλυση και την υλοποίηση κατηγοριοποίησης (clustering) ενός υποσυνόλου πελατών Ηλεκτρικής Ενέργειας στην Ελλάδα και αξιοποίησής της για τη βραχυπρόθεσμη πρόβλεψη φορτίου καταναλωτών Μέσης Τάσης του Ελληνικού Δικτύου Ηλεκτρικής Ενέργειας μέσω των τυπικών ημερήσιων καμπυλών τους για τα έτη 2018, 2019 και το πρώτο εξάμηνο του 2020.

Συγκεκριμένα, γίνεται μελέτη και παρουσίαση της διαδικασίας της κατηγοριοποίησης με την εφαρμογή μεθόδων αναγνώρισης προτύπων, ακολουθούμενη από την εύρεση τυπικών χρονολογικών ανά ώρα καμπυλών για κάθε πελάτη ξεχωριστά, οι οποίες μετά ταξινομούνται σε κατηγορίες με τη βοήθεια του αλγορίθμου K-μέσων. Η φύση των δεδομένων καθιστά το dataset που αυτά αποτελούν να χαρακτηρίζεται imbalanced, καθώς υπάρχει μεγάλη συγκέντρωσή τους σε μία περιοχή από το σύνολο της έκτασης που έχουν. Αυτό οδήγησε στην εφαρμογή παραλλαγών του αλγόριθμου για τη βελτιστοποίηση της απόκρισης της κατηγοριοποίησης. Τέλος, μελετάται και υλοποιείται με τεχνητό νευρωνικό δίκτυο η βραχυπρόθεσμη πρόβλεψη φορτίου για τις επόμενες 24 ώρες, η οποία αξιοποιεί τα αποτελέσματα που εξήχθησαν από την κατηγοριοποίηση.

Αναλυτικότερα στο **κεφάλαιο 1** αναφέρεται συνοπτικά η διαδικασία του σχεδιασμού της κατηγοριοποίησης προτύπων και γίνεται παρουσίαση των βασικότερων μεθόδων που χρησιμοποιούνται. Το κεφάλαιο ολοκληρώνεται με την παρουσίαση της Ελληνικής Αγοράς Ηλεκτρικής Ενέργειας και των τμημάτων που η χρήση της κατηγοριοποίησης θα ωφελούσε σημαντικά.

Στο **κεφάλαιο 2** αναλύονται οι κυριότερες τεχνικές κατηγοριοποίησης που εφαρμόζονται και οι δείκτες αξιολόγησης που χρησιμοποιούνται για τη σύγκριση μεταξύ των μεθόδων. Ακόμη, αναλύονται τα βήματα και ο μαθηματικός σκελετός του αλγορίθμου K-μέσων που θα υλοποιηθεί για την ομαδοποίηση των τυπικών προφίλ των καταναλωτών.

Το **κεφάλαιο 3** αναφέρεται στην πρόβλεψη φορτίου, δηλαδή στον προσδιορισμό της ζήτησης φορτίου για επόμενο χρονικό διάστημα το οποίο καθορίζει τον χαρακτηρισμό της ως βραχυπρόθεσμη, μεσοπρόθεσμη και μακροπρόθεσμη. Στη συνέχεια περιγράφονται οι σημαντικότερες μεθοδολογίες που χρησιμοποιούνται για την υλοποίησή της που διακρίνονται σε κλασικές και σε τεχνητής νοημοσύνης. Το κεφάλαιο ολοκληρώνεται με την ανάλυση των τεχνητών νευρωνικών δικτύων με τα οποία θα υλοποιηθεί η πρόβλεψη.

Στο **κεφάλαιο 4** παρουσιάζεται η υλοποίηση της κατηγοριοποίησης που έχει ως αρχικό στάδιο τη συλλογή και τη διαμόρφωση των δεδομένων. Έπειτα, αναλύεται η εφαρμογή του αλγόριθμου K-μέσων και παραλλαγών-βελτιώσεών του στην αναζήτηση λύσης της κατηγοριοποίησης του imbalanced dataset που αποτελεί το σύνολο των δεδομένων. Αυτό το στοιχείο επηρεάζει την αποτελεσματικότητα της και παρουσιάζονται τα αποτελέσματα ξεχωριστά για κάθε μία από αυτές τις τεχνικές. Συγκεκριμένα, αυτές είναι η μείωση των διαστάσεων των δεδομένων, ο αλγόριθμος Mini Batch K-means, η υλοποίηση καθοδηγούμενης κατηγοριοποίησης με K-means (guided K-means), καθώς και ο συνδυασμός τους.

Στο **κεφάλαιο 5** παρουσιάζεται η υλοποίηση της βραχυπρόθεσμης πρόβλεψης του μέσου τυπικού προφίλ για κάθε cluster με τη βοήθεια τεχνητών νευρωνικών δικτύων. Αρχικά, αναλύεται η επεξεργασία των δεδομένων που είναι απαραίτητη για τη συλλογή τους και τον συνδυασμό τους με την κατηγοριοποίηση που έχει προηγηθεί. Ακολουθεί η παρουσίαση της αρχιτεκτονικής του δικτύου που κατασκευάστηκε, αναλύονται τα στοιχεία του και παρουσιάζονται τα αποτελέσματα της πρόβλεψης των επόμενων 24 ωρών που εξήχθησαν για το σύνολο των ομάδων που ανήκουν οι καταναλωτές.

Στο **κεφάλαιο 6** καταγράφονται τα συμπεράσματα που προέκυψαν από τη διπλωματική εργασία και προτείνονται μελλοντικά βήματα έρευνας με βάση την εργασία που εκπονήθηκε.

Καταλήγοντας, καταγράφεται η βιβλιογραφία που χρησιμοποιήθηκε και στο παράρτημα παρουσιάζονται αναλυτικά διαγράμματα ανά σελίδα για τα αποτελέσματα που εξήχθησαν.

Στο σημείο αυτό θα ήθελα να ευχαριστήσω θερμά τον επιβλέποντα καθηγητή κ. Νικόλαο Χατζηαργυρίου για τη συνεργασία, την καθοδήγηση και τη στήριξη που μου παρείχε κατά την εκπόνηση της παρούσας εργασίας. Επίσης, θα ήθελα να ευχαριστήσω τον υποψήφιο διδάκτορα κ. Θεόδωρο Κωνσταντίνου για τη σημαντική βοήθεια που μου προσέφερε και τον Επίκουρο Καθηγητή ΠΑ.Δ.Α κ. Γ. Τσεκούρα για την παροχή του κειμένου του διδακτορικού του. Ακόμη, ευχαριστώ τον κ. Δημήτρη Βράνη Διευθυντή της Διεύθυνσης Χρηστών Δικτύου, τον κ. Δημήτρη Μερεντίτη Τομεάρχη του τομέα της Εκκαθάρισης Αγοράς, τον κ Νικόλαο Ιορδανίδη, αλλά και όλο το προσωπικό του ΔΕΔΔΗΕ για την παροχή των δεδομένων των Τηλεμετρούμενων πελατών Μέσης Τάσης που χρησιμοποιήθηκαν. Τέλος, ευχαριστώ την οικογένειά μου για την αμέριστη στήριξη που μου έδωσε σε όλα τα βήματα της ζωής μου.

Περιεχόμενα

Εισαγωγή

Η διπλωματική αυτή εργασία εκπονείται στα πλαίσια του Διατμηματικού Προγράμματος Μεταπτυχιακών Σπουδών «Παραγωγή και Διαχείριση Ενέργειας» του Εθνικού Μετσόβιου Πολυτεχνείου με επιβλέποντα καθηγητή τον κ. Νικόλαο Χατζηαργυρίου.

Ο στόχος της συγγραφής της είναι η υλοποίηση κατηγοριοποίησης των καταναλωτών Μέσης Τάσης του Ελληνικού Δικτύου Ηλεκτρικής Ενέργειας μέσω των τυπικών ημερήσιων καμπυλών τους τροφοδοτώντας μοντέλο βραχυπρόθεσμης πρόβλεψης των φορτίων τους.

Το θέμα της διπλωματικής εργασίας εμπνεύστηκε από την ανάγκη των σύγχρονων δικτύων για τη γνώση της ζήτησης φορτίου της επόμενης ημέρας και από την απελευθέρωση της αγοράς ηλεκτρικής ενέργειας. Η βραχυπρόθεσμη πρόβλεψη φορτίου που θα υλοποιηθεί με τεχνητό νευρωνικό δίκτυο καλείται να προσφέρει αυτήν την πληροφορία για τους καταναλωτές Μέσης Τάσης του διασυνδεδεμένου συστήματος, η οποία είναι εξαιρετικά σημαντική για τους προμηθευτές, τους παραγωγούς και τους διαχειριστές διανομής και μεταφοράς. Λαμβάνεται υπόψη στην εκτίμηση της αναγκαίας παραγόμενης ενέργειας, των ροών φορτίου και στη λήψη αποφάσεων που μπορούν να αποτρέψουν την υπερφόρτωση του δικτύου.

Υλοποιείται κατηγοριοποίηση με τον αλγόριθμο K-μέσων (K-means) και παραλλαγών - τεχνικών βελτιστοποίησής του με στόχο την εξαγωγή των καλύτερων δυνατών αποτελεσμάτων, τα οποία θα οδηγηθούν στο μοντέλο πρόβλεψης. Συγκεκριμένα, δοκιμάζεται η μείωση των διαστάσεων των δεδομένων, η εφαρμογή του Mini Batch K-means και η καθοδηγούμενη αρχικοποίησή του. Επίσης, η κατηγοριοποίηση είναι από μόνη της ένα σημαντικό εργαλείο στην αγορά ηλεκτρικής ενέργειας, καθώς μπορεί να οδηγήσει σε μια πιο προσωποποιημένη τιμολόγηση του πελάτη, βελτιστοποίηση των πόρων του προμηθευτή ηλεκτρικού ρεύματος και καλύτερη λήψη αποφάσεων σχετικά με την παραγωγή ηλεκτρικής ενέργειας. Η διπλωματική επικεντρώθηκε στην κατηγοριοποίηση, καθώς αυτή αντιμετώπισε το μεγάλο και ανισοκατανεμημένο σύνολο που αποτελούν τα δεδομένα. Έπειτα, για κάθε ομάδα που σχηματίστηκε, υλοποιείται νευρωνικό δίκτυο για τη βραχυπρόθεσμη πρόβλεψη φορτίου του μέσου τυπικού προφίλ της.

Η παρούσα εργασία δίνει λύση στο πρόβλημα της υλοποίησης πρόβλεψης όταν δεν είναι γνωστά τα δεδομένα της χρονοσειράς ζήτησης φορτίου ενός καταναλωτή. Η απάντηση βρίσκεται στην υλοποίηση κατηγοριοποίησης των καταναλωτών με γνωστές χρονοσειρές σε ομάδες και στη διεξαγωγή πρόβλεψης φορτίου για κάθε μία εξ αυτών. Το τυπικό προφίλ του άγνωστου πελάτη ταξινομείται σε μία από αυτές, όπου είναι γνωστό το σφάλμα πρόβλεψης που έχει υπολογιστεί για αυτή. Πρακτικά με τη μελέτη που υλοποιείται στην παρούσα εργασία, ένας πελάτης με γνωστό μόνο το τυπικό προφίλ κατανάλωσης φορτίου του μπορεί να καταχωρηθεί στην αντίστοιχη ομάδα που είναι πιο αντιπροσωπευτική για αυτόν και να γίνει εκτίμηση της ζήτησης φορτίου του.

Παραδείγματος χάρη, έστω ένας ζυγός στον οποίο δεν είναι γνωστή η πληροφορία των χρονοσειρών φορτίου όλων των x καταναλωτών που τροφοδοτεί, αλλά είναι μόνο για μερικούς εξ αυτών (y), ενώ για όλους είναι γνωστά τα τυπικά προφίλ φορτίου. Υλοποιείται κατηγοριοποίηση στα y προφίλ, τα οποία ταξινομούνται στα clusters που δημιουργούνται. Για την κάθε ομάδα που εξάγεται από το μοντέλο που έχει εκπαιδευτεί υπολογίζεται το μέσο τυπικό προφίλ. Η διαδικασία προϋποθέτει ότι η κατηγοριοποίηση υλοποιείται με τέτοιο τρόπο, ώστε οι ομάδες που προκύπτουν να είναι αντιπροσωπευτικές των καταναλωτών που κατατάσσονται σε αυτές, χωρίς να αλλοιώνουν τα χαρακτηριστικά τους. Έπειτα, η πρόβλεψη της κατανάλωσης της επόμενης ημέρας για την κάθε ομάδα έχει υπολογιστεί από το νευρωνικό δίκτυο υλοποιήθηκε. Με αυτόν τον τρόπο, είναι γνωστό το φορτίο των επόμενων 24 ωρών του τυχαίου χρήστη του δικτύου που θα τροφοδοτήσει ο ζυγός, με μόνη υπόθεση ως πληροφορία το τυπικό προφίλ του.

Κεφάλαιο 1
Αναγνώριση και κατηγοριοποίηση προτύπων

1.1 Συστήματα κατηγοριοποίησης

Η παρούσα διπλωματική πραγματεύεται την κατηγοριοποίηση καταναλωτών Μέσης Τάσης του Ελληνικού Δικτύου Ενέργειας. Η κατηγοριοποίηση (clustering) είναι η επιστημονική μέθοδος που ταξινομεί ένα σύνολο αντικειμένων σε ξεχωριστές ομάδες-κατηγορίες (clusters). Ο στόχος της είναι η ανάπτυξη μοντέλου που μπορεί να χρησιμοποιηθεί για την ταξινόμηση μελλοντικών δεδομένων. Τα αντικείμενα που κατηγοριοποιούνται ορίζονται ως πρότυπα, όρος που θα χρησιμοποιείται για αυτά κατά την ανάπτυξη της εργασίας. Η κατηγοριοποίηση, λόγω της ταχύτατης ανάπτυξης και διάδοσης των ηλεκτρονικών υπολογιστών αποτελεί τομέα αιχμής στην έρευνα, ενώ χρησιμοποιείται ήδη σε πληθώρα εφαρμογών. Τυπικές εφαρμογές αποτελούν όλα τα μηχανικά συστήματα με τεχνητή νοημοσύνη σε τομείς της βιομηχανίας, όπως στην επιθεώρηση ή αυτοματοποίηση της γραμμής παραγωγής, στην οικονομία, στην ιατρική, στη στρατιωτική βιομηχανία και φυσικά στην καθημερινότητα.

Η διαδικασία κατηγοριοποίησης προτύπων συνίσταται στην ταξινόμησή τους σε ομάδες των οποίων τα μέλη παρουσιάζουν κοινά χαρακτηριστικά, με τελικό αποτέλεσμα την οργάνωση του αρχικού συνόλου σε κατηγορίες με σαφώς μικρότερο πλήθος στοιχείων. Το γεγονός αυτό διευκολύνει την επεξεργασία των επιμέρους ομάδων, στοχεύοντας στην ανάλυσή τους που ποικίλει ανάλογα την εφαρμογή. Τέτοιο παράδειγμα αποτελεί η ταξινόμηση του πληθυσμού μιας χώρας σε ομάδες ανάλογα την περιοχή που διαμένουν και της ηλικίας τους ώστε να αναλυθεί η μέση ηλικία κάθε περιοχής της χώρας για αξιολόγηση του δημογραφικής ηλικίας της [1].

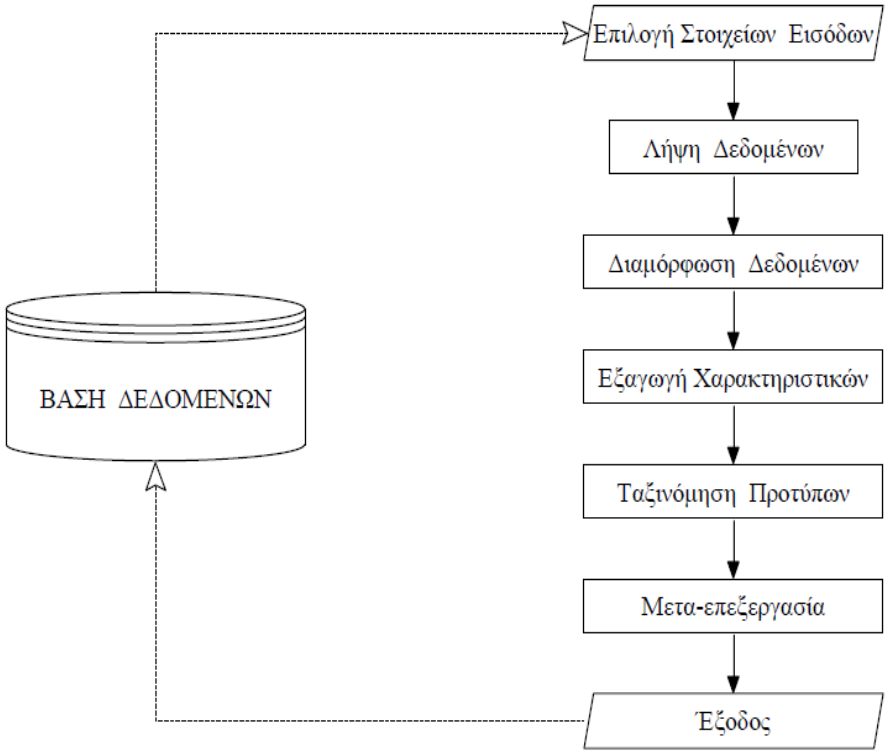
Ένα σύστημα αναγνώρισης και κατηγοριοποίησης προτύπων έχει ως είσοδο ένα σύνολο δεδομένων που η επεξεργασία τους παράγει ως έξοδο την ταξινόμησή τους σε ομάδες. Τα στάδια επεξεργασίας προτύπων που παρουσιάζονται στο διάγραμμα 1.1 είναι:

1. Λήψη (sensing) : τα πρότυπα εισάγονται στο σύστημα και μετασχηματίζονται μέσω μετατροπέα σε κατάλληλη απεικόνιση, ώστε να είναι εύχρηστα και να αποτελούν κατάλληλη βάση για τα επόμενα βήματα.
2. Διαμόρφωση(segmentation) : τα πρότυπα απομονώνονται από το σύνολο των δεδομένων της πληροφορίας και διαχωρίζονται από άλλα αντικείμενα.
3. Εξαγωγή χαρακτηριστικών (feature extraction) : οι ιδιότητες των προτύπων που κρίνονται σημαντικές και που διαχωρίζουν τα χαρακτηριστικά των προτύπων προσδιορίζονται και καταγράφονται.
4. Ταξινόμηση (classification) : οι ιδιότητες των προτύπων μελετώνται και αξιολογούνται, με τελικό προορισμό την καταχώρησή τους σε μία από τις ομάδες.
5. Μετά-επεξεργασία (post-processing) : λαμβάνονται υπόψη οι επιδράσεις του πλαισίου εφαρμογής (context) και το κόστος των ενδεχόμενων λαθών, ώστε να προσδιοριστεί η κατάλληλη δράση συνήθως από ανθρώπινο δυναμικό.

Η ροή των δραστηριοτήτων που παρουσιάστηκαν εκτός από ευθύγραμμη σε μερικά συστήματα αναγνώρισης και κατηγοριοποίησης προτύπων μπορεί να είναι και με ανατροφοδότηση, όπου τα δεδομένα

από την έξοδο οδηγούνται και πάλι σε ανώτερο στάδιο για επανάληψη των υπόλοιπων βημάτων μέχρι την επίτευξη του επιθυμητού αποτελέσματος [2].

Ο σχεδιασμός ενός συστήματος αναγνώρισης και κατηγοριοποίησης προτύπων αποτελείται από διαδικασίες που δεν είναι αυστηρά ορισμένες και περιλαμβάνουν συνεχείς δοκιμές και ελέγχους. Το στοιχείο αυτό βοηθά να αντιληφθούμε πως ένα τέλειο τέτοιο σύστημα είναι ανέφικτο, ωστόσο με τον διαρκή επανασχεδιασμό και την παραμετροποίησή του επιτυγχάνεται καλύτερη ανταπόκριση και αντίστοιχα αποτελέσματα. Τα προβλήματα που καλείται να λύσει το σύστημα είναι συνήθως σύνθετα προβλήματα με δεδομένα μεγάλης έκτασης, όπου για την ταξινόμησή τους απαιτείται μεγάλη υπολογιστική ισχύς. Έτσι, κρίνεται απαραίτητη η συνεισφορά προγραμματιστικών τεχνικών για την επίλυσή τους.



Διάγραμμα 1.1: Σύστημα κατηγοριοποίησης προτύπων

Ο σχεδιασμός του συστήματος αναγνώρισης και κατηγοριοποίησης αποτελείται από τα εξής βήματα:

1. Συλλογή δεδομένων (data collection) : το πρώτο αυτό βήμα τις περισσότερες φορές είναι το πιο χρονοβόρο και κοστοβόρο στον σχεδιασμό, όπου συγκεντρώνονται τα δεδομένα που θα χρησιμοποιηθούν στην εκπαίδευση και τον έλεγχο του συστήματος. Τα χαρακτηριστικά των δεδομένων επηρεάζουν την επιλογή κατάλληλων γνωρισμάτων-ιδιοτήτων αλλά και την ίδια την απόφαση για την υλοποίηση του βέλτιστου μοντέλου ταξινόμησης για την εφαρμογή.
2. Επιλογή γνωρισμάτων (features selection) : η επιλογή των ξεχωριστών γνωρισμάτων των δεδομένων εξαρτάται από τη φύση του προβλήματος.
3. Επιλογή μοντέλου ταξινόμησης (model selection) : η επιλογή του μοντέλου που θα χρησιμοποιηθεί εξαρτάται από τη φύση του προβλήματος..
4. Διαδικασία εκπαίδευσης (training process) : πραγματοποιείται η εκπαίδευση του μοντέλου της ταξινόμησης, όπου χρησιμοποιείται κομμάτι ή το σύνολο των δεδομένων που έχουν συλλεχθεί ώστε να προσδιοριστούν όλες οι παράμετροι του σχεδιαζόμενου συστήματος.
5. Αξιολόγηση (evaluation classifier) : τα αποτελέσματα αναλύονται και αξιολογούνται, η απόδοση του του συστήματος μετράται και προσδιορίζονται οι ανάγκες για βελτίωση των βασικών συνιστωσών.

Τα βήματα της διαδικασίας σχεδιασμού πολλές φορές επαναλαμβάνονται μέχρι να ληφθούν ικανοποιητικά αποτελέσματα, όπου τα δεδομένα που εξάγονται ανατροφοδοτούνται σε πρότερα στάδια οδηγώντας στη βελτίωση του σχεδιασμού [2].

1.2 Μέθοδοι αναγνώρισης προτύπων

Οι τεχνικές αναγνώρισης προτύπων χωρίζονται σε δύο βασικές κατηγορίες: σε επιβλεπόμενες (supervised pattern recognition) και σε μη επιβλεπόμενες (unsupervised pattern recognition). Η πρώτη κατηγορία αφορά μεθόδους όπου οι ομάδες διαχωρισμού ενός συνόλου δεδομένων είναι εξ αρχής γνωστές με ζητούμενο την ορθή τοποθέτηση των διανυσμάτων έπειτα από την εκπαίδευση του αντίστοιχου μοντέλου. Στις μη επιβλεπόμενες ζητείται η κατηγοριοποίηση χωρίς να είναι γνωστές οι κατηγορίες με σκοπό να ανακαλυφθούν οι ομοιότητες μεταξύ των διανυσμάτων και να σχηματιστούν κατάλληλες ομάδες, μέθοδοι που θα χρησιμοποιηθούν και στην υλοποίηση της παρούσας διπλωματικής.

Οι βασικότερες μέθοδοι επιβλεπόμενης αναγνώρισης προτύπων είναι:

- Ταξινόμηση κατά Bayes (Bayesian Classification), που περιλαμβάνει εκτίμηση άγνωστων συναρτήσεων πυκνότητας πιθανότητας και χρήσης πιθανοτικών μεθόδων ταξινόμησης.
- Γραμμικοί ταξινομητές (Linear Classifiers), οι οποίοι χρησιμοποιούνται σε γραμμικά διαχωρίσιμα προβλήματα χρησιμοποιώντας μεθόδους όπως του ελαχίστου τετραγωνικού σφάλματος, του απλούστερου νευρωνικού δικτύου τύπου perceptron κ.α..
- Μη γραμμικοί ταξινομητές (Nonlinear Classifiers), που αντιμετωπίζουν μη γραμμικά διαχωρίσιμα προβλήματα με μεθόδους όπως τα ακτινικά δίκτυα, τα τεχνητά νευρωνικά δίκτυα με πολλαπλά επίπεδα με μεθόδους ανάστροφης διάδοσης σφάλματος κ.α.
- Στοχαστικές μέθοδοι (Stochastic Methods), όπως η εκπαίδευση κατά Boltzmann, οι γενετικοί αλγόριθμοι κ.α.
- Μη μετρικές μέθοδοι (Nonmetric Methods), όπως τα δένδρα αποφάσεων, οι μέθοδοι συμβολοσειρών ή γραμματικών κανόνων.
- Μέθοδοι παραγωγής χαρακτηριστικών γνωρισμάτων (Feature Generation Methods), όπου χρησιμοποιούνται ο διακριτός μετασχηματισμός Fourier, ο διακριτός χρονικός μετασχηματισμός κυματομορφών, οι μετασχηματισμοί κατά Haar, κατά Hadamard κ.α.

Οι βασικότερες μέθοδοι μη επιβλεπόμενης αναγνώρισης προτύπων είναι οι ακόλουθες:

- Σειριακοί αλγόριθμοι (Sequential Algorithms), δηλαδή αλγόριθμοι που παράγουν απλή κατηγοριοποίηση και είναι ευθείες και γρήγορες μέθοδοι. Στις περισσότερες περιπτώσεις τα διανύσματα εισόδου παρουσιάζονται στον αλγόριθμο μία ή λίγες φορές και το τελικό αποτέλεσμα συνήθως εξαρτάται από τη σειρά με την οποία έγινε η παρουσίαση των προτύπων. Δίνουν συνήθως συμπαγείς ομάδες με σφαιρικό ή ελλειπτικό σχήμα, ανάλογα με το κριτήριο απόστασης που χρησιμοποιείται.
- Ιεραρχικοί αλγόριθμοι συγχώνευσης (Hierarchical Agglomerative Algorithms), στους οποίους παράγεται μία σειρά από κατηγοριοποιήσεις με τον αριθμό των δημιουργημένων ομάδων να μειώνεται. Η ομαδοποίηση που πραγματοποιείται σε κάθε βήμα προκύπτει από την ομαδοποίηση του προηγούμενου βήματος μέσω της συγχώνευσης δύο ομάδων. Κυριότεροι εκπρόσωποί τους θεωρούνται οι απλοί αλγόριθμοι και οι αλγόριθμοι ολοκληρωμένης σύνδεσης, οι οποίοι χωρίζονται σε αυτούς που βασίζονται στη θεωρία πινάκων και σε εκείνους που βασίζονται στη θεωρία γράφων. Οι συγκεκριμένοι αλγόριθμοι είναι κατάλληλοι για ανεύρεση επεκτεινόμενων, επιμηκυμένων (όπως η περίπτωση των απλών αλγορίθμων) και συμπαγών ομάδων (όπως η περίπτωση των αλγορίθμων ολοκληρωμένης σύνδεσης).
- Ιεραρχικοί διαιρούμενοι αλγόριθμοι (Hierarchical Divisive Algorithms), αλγόριθμοι που λειτουργούν με βάση την αντίστροφη διαδικασία από αυτή που περιεγράφηκε προηγουμένως, δηλαδή παράγουν μία σειρά από κατηγοριοποιήσεις με τον αριθμό των ομάδων να αυξάνει, αφού η ομαδοποίηση που πραγματοποιείται σε κάθε βήμα προκύπτει από την ομαδοποίηση του προηγούμενου βήματος μέσω της διαίρεσης μίας ομάδας σε δύο νέες.

- Αλγόριθμοι βασιζόμενοι στη βελτιστοποίηση της συνάρτησης κόστους (Algorithms based on Cost Function Optimization). Αυτή η κατηγορία περιλαμβάνει αλγόριθμους στους οποίους η «λογική» ποσοτικοποιείται από μία συνάρτηση κόστους *J*, βάσει της οποίας πραγματοποιείται η αξιολόγηση της κάθε ομαδοποίησης με τον αριθμό των κατηγοριών να διατηρείται συνήθως σταθερός με μεταβαλλόμενη τη σύνθεσή τους. Σε αυτούς τους αλγόριθμους χρησιμοποιούνται έννοιες διαφορικού λογισμού, ώστε να δημιουργούν επιτυχημένες κατηγοριοποιήσεις κατά την προσπάθεια βελτίωσης στης συνάρτησης J. Η διαδικασία της κατηγοριοποίησης ολοκληρώνεται όταν προσδιοριστεί ένα τοπικό ακρότατο (βέλτιστο) της συνάρτησης *J*. Οι αλγόριθμοι αυτής της κατηγορίας ονομάζονται και αλγόριθμοι βελτιστοποίησης της επαναληπτικής συνάρτησης (iterative function optimization schemes), με υποκατηγορίες τις ακόλουθες:
- Αλγόριθμοι σκληρής ομαδοποίησης (hard clustering algorithms), όπου το κάθε πρότυπο ανήκει αυστηρά και εξ ολοκλήρου σε μία συγκεκριμένη ομάδα. Η τοποθέτηση των προτύπων σε ξεχωριστές ομάδες διεκπεραιώνεται κατά τον βέλτιστο δυνατό τρόπο, σύμφωνα με το κριτήριο βελτιστοποίησης που έχει οριστεί. Ο πιο γνωστός αλγόριθμος αυτής της κατηγορίας είναι ο αλγόριθμος του Lloyd.
- Αλγόριθμοι πιθανοτικής ομαδοποίησης (probabilistic clustering algorithms), οι οποίοι ακολουθούν την κατανομή Bayes και το κάθε πρότυπο $\boldsymbol{x}$ καταχωρείται στην ομάδα *Ci* για την οποία η τιμή $P(Ci \mid \boldsymbol{x})$ μεγιστοποιείται. Αυτές οι πιθανότητες υπολογίζονται με χρήση μίας κατάλληλα ορισμένης διαδικασίας βελτιστοποίησης.
- Αλγόριθμοι ασαφής ομαδοποίησης (fuzzy clustering algorithms), όπου το πρότυπο ανήκει ξεχωριστά στην κάθε ομάδα με κάποιο βαθμό συμμετοχής.
- Αλγόριθμοι ενδεχόμενης ομαδοποίησης (possibilistic clustering algorithms), όπου μετράται το ενδεχόμενο ένα πρότυπο να ανήκει στην ομάδα *Ci*.
- Αλγόριθμοι ανίχνευσης συνόρων (boundary detection algorithms), οι οποίοι προσαρμόζουν επανειλημμένως τα σύνορα των περιοχών που εξαπλώνονται οι ομάδες αντί να προσδιορίζουν τις ομάδες με βάση τα χαρακτηριστικά των προτύπων. Αυτή η κατηγορία αλγόριθμων παρόλο που έχει αναπτυχθεί με βάση τη φιλοσοφία της βελτιστοποίησης της συνάρτησης κόστους, είναι ωστόσο τελείως διαφορετικής φύσεως.

Όλοι οι προηγούμενοι αλγόριθμοι δημιουργούν ομάδες οι οποίες αναπαρίστανται από κάποιον αντιπρόσωπο με στόχο να τις προσδιορίσουν κατά τον βέλτιστο τρόπο. Αντιθέτως, οι αλγόριθμοι ανίχνευσης συνόρων αναζητούν μεθόδους να οριοθετήσουν κατά βέλτιστο τρόπο τα σύνορα που υπάρχουν ανάμεσα στις ομάδες.

Ειδικές τεχνικές ομαδοποίησης που δεν εντάσσονται σε καμία από τις προηγούμενες κατηγορίες είναι οι εξής:

- Αλγόριθμοι ομαδοποίησης κλάδων και ορίων (branch and bound clustering algorithms) οι οποίοι δίνουν τη συνολικά βέλτιστη κατηγοριοποίηση για σταθερό αριθμό ομάδων με χρήση ενός εξ αρχής θεσπισμένου κριτηρίου, χωρίς όμως να εξετάζουν όλες τις δυνατές περιπτώσεις ομαδοποίησης. Ωστόσο, απαιτούν μεγάλη υπολογιστική ισχύ για να υλοποιηθούν.
- Γενετικοί αλγόριθμοι (genetic algorithms), οι οποίοι χρησιμοποιούν έναν αρχικό πληθυσμό πιθανών ομαδοποιήσεων τις οποίες διασταυρώνουν μεταξύ τους, δημιουργώντας νέους πληθυσμούς που γενικά περιέχουν καλύτερες ομαδοποιήσεις συγκριτικά με τις προηγούμενες γενιές κατηγοριοποιήσεων, πάντα σύμφωνα με κάποιο αυστηρά ορισμένο κριτήριο.
- Μέθοδοι στοχαστικής χαλάρωσης (stochastic relaxation methods), οι οποίες υπό ορισμένες συνθήκες εγγυώνται σύγκλιση στην πιθανότητα υλοποίησης μίας συνολικά βέλτιστης ομαδοποίησης, πάντα σύμφωνα με κάποιο αυστηρά ορισμένο κριτήριο. Έχουν όμως μειονέκτημα, τη μεγάλη υπολογιστική ισχύ που απαιτείται για την υλοποίησή τους.
- Αλγόριθμοι εύρεσης κοιλάδων (valley seeking clustering algorithms), οι οποίοι αντιμετωπίζουν τα πρότυπα ως ενδεχόμενα μίας (πολυδιάστατης) τυχαίας μεταβλητής $\boldsymbol{x}$. Η λογική τους βασίζεται στην υπόθεση πως για τις περιοχές της μεταβλητής $\boldsymbol{x}$, πολλά πρότυπα διαμένουν σύμφωνα με τις περιοχές των αυξανόμενων τιμών της σχετικής συνάρτησης πυκνότητας πιθανότητας της $\boldsymbol{x}$. Συνεπώς, ο προϋπολογισμός της συνάρτησης

πυκνότητας πιθανότητας μπορεί να δώσει πληροφορίες για τις περιοχές στις οποίες έχουν διαμορφωθεί οι ομάδες.

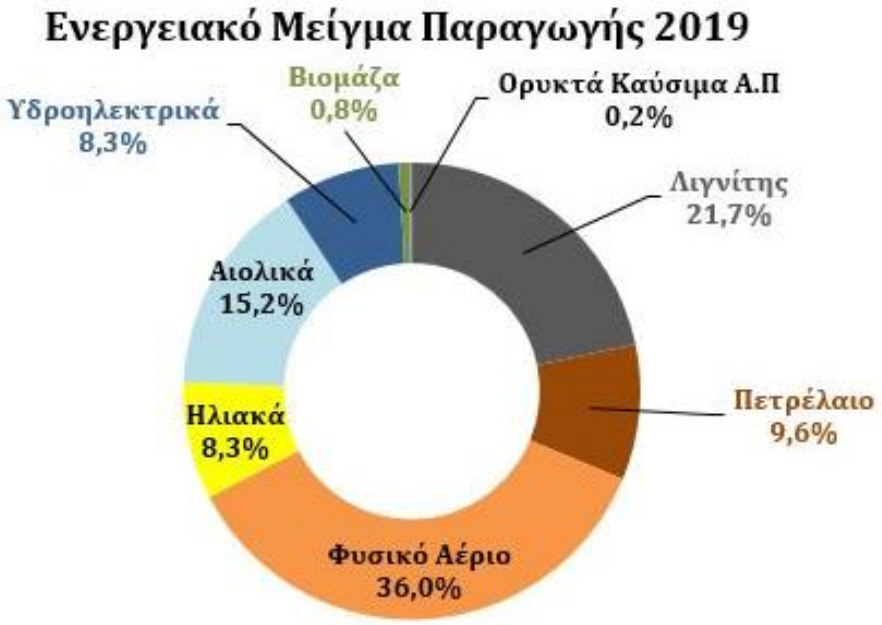
- Αλγόριθμοι ανταγωνιστικής μάθησης (competitive learning algorithms), οι οποίοι είναι επαναληπτικοί αλγόριθμοι και δεν εμπλέκουν τη συνάρτηση κόστους κατά την εφαρμογή τους. Οι αλγόριθμοι αυτοί δημιουργούν διάφορες ομαδοποιήσεις για να συγκλίνουν τελικά στην πιο λογική επιλογή, σύμφωνα με κάποιο κριτήριο απόστασης. Τυπικοί αντιπρόσωποι της κατηγορίας αυτής αποτελούν ο βασικός διανυσματικός αλγόριθμος ανταγωνιστικής μάθησης και ο αλγόριθμος διαρρέουσας μάθησης.
- Αλγόριθμοι βασισμένοι σε τεχνικές μορφολογικών μετασχηματισμών (algorithms based on morphological transformation techniques), οι οποίοι χρησιμοποιούν μορφολογικούς μετασχηματισμούς, ώστε να επιτύχουν καλύτερο διαχωρισμό ομάδων.

Εκτός του χαρακτηριστικού της επίβλεψης κριτήριο διαχωρισμού των μεθόδων κατηγοριοποίησης αποτελεί το εάν επιτρέπει η συμμετοχή του κάθε προτύπου σε περισσότερες από μία ομάδες, αποτυπώνοντας την ανάμειξη σε κάθε μία με ένα βαθμό συμμετοχής. Όταν επιτρέπεται να αντιστοιχεί σε μία μόνο ομάδα τότε πρόκειται για σκληρή ομαδοποίηση (hard clustering) και ασαφή ομαδοποίηση στην αντίθετη περίπτωση. Ακόμη, οι τεχνικές ομαδοποίησης διακρίνονται σε παραμετρικές και μη παραμετρικές. Στις πρώτες κάθε ομάδα περιγράφεται από κάποιο μοντέλο ή εκπρόσωπο με παραμέτρους που καθορίζονται κατά τη διαδικασία εύρεσης των ομάδων, ενώ στις δεύτερες το αποτέλεσμα της ομαδοποίησης είναι ο καθορισμός των ομάδων και των προτύπων που ανήκουν σε κάθε ομάδα.

Οι τεχνικές που παρουσιάστηκαν εμφανίζουν πλεονεκτήματα και μειονεκτήματα ως προς την αποτελεσματικότητα και την ταχύτητα ολοκλήρωσής τους, στοιχείο που αναδεικνύει την εξ αρχής επιλογή της τεχνικής που θα χρησιμοποιηθεί δυσδιάκριτη για κάθε πρόβλημα που καλείται να επιλύσει. Έτσι, πολλές φορές εκτελούνται περισσότερες από μία μεθόδους και με βάση την πιστότητα των αποτελεσμάτων προσδιορίζεται η καταλληλότερη [19].

1.3 Πρόβλεψη ηλεκτρικού φορτίου

Η ηλεκτρική ενέργεια είναι ανά πάσα στιγμή διαθέσιμη στο σημείο ζήτησης καλύπτοντας τις απαιτήσεις του καταναλωτή σε ποιότητα και ποσότητα με το ελάχιστο δυνατό κόστος [3]. Στη χώρα μας το σύστημα ηλεκτρικής ενέργειας αποτελείται σε επίπεδο μεταφοράς και διανομής από τρία επίπεδα τάσης, την υψηλή, τη μέση και τη χαμηλή τάση. Η παραγωγή της συνίσταται κυρίως στη μετατροπή κάποιας άλλης μορφής ενέργειας σε ηλεκτρική μέσω γεννητριών, με σύνηθες ενδιάμεσο στάδιο τη μηχανική ενέργεια μέσω κινητήρων και στροβίλων. Το σύνολο της ενέργειας που παρέχεται στον καταναλωτή προέρχεται από σταθμούς παραγωγής, όπως οι θερμοηλεκτρικοί, οι υδροηλεκτρικοί, οι πυρηνικοί και οι ανανεώσιμες πηγές ενέργειας (ηλιακή, αιολική, βιομάζα). Το ενεργειακό μίγμα από το οποίο παράχθηκε η ηλεκτρική ενέργεια που καταναλώθηκε για το έτος 2019 αναπαρίσταται στο Διάγραμμα 1.3 [4].



Διάγραμμα 1.3: Ενεργειακό μείγμα ελληνικής παραγωγής για το 2019.

Η παραγωγή και παροχή της ηλεκτρικής ενέργειας πραγματοποιείται μέσω του συστήματος ηλεκτρικής ενέργειας (Σ.Η.Ε.), το οποίο ορίζεται ως ένα σύνολο συσκευών, υπολογιστικών συστημάτων και ανθρώπινου δυναμικού το οποίο καλείται να τροφοδοτεί οπουδήποτε υπάρχει ζήτηση ηλεκτρικής ενέργειας, ακόμα και όταν αυτή μεταβάλλεται. Η παρεχόμενη ενέργεια πρέπει να διατηρεί σταθερή συχνότητα, να διατηρεί σταθερό το επίπεδο τάσης, να υπάρχει υψηλή αξιοπιστία τροφοδοτήσεως καθ' όλη τη διάρκεια του χρόνου, ώστε να εξασφαλίζεται η ποιότητά της [5].

Η ζήτηση ηλεκτρικής ενέργειας από τους καταναλωτές επηρεάζεται από πολλούς παράγοντες οι οποίοι μπορούν να ταξινομηθούν στις ακόλουθες κατηγορίες:

- Οικονομικοί παράγοντες: Η οικονομία και οι ρυθμοί ανάπτυξης κάθε περιοχής επηρεάζουν καθοριστικά τη ζήτηση ηλεκτρικής ενέργειας. Το βιοτικό επίπεδο των χρηστών επηρεάζει αναλογικά την κατανάλωση, έτσι το κατά κεφαλήν εισόδημα, δηλαδή το ΑΕΠ μιας χώρας διαιρεμένο δια τον πληθυσμό της, μπορεί να χρησιμοποιηθεί για μια μακροοικονομική ανάλυση ζήτηση της ηλεκτρικής ενέργειας. Σε

αντιδιαστολή τα τελευταία χρόνια μια αύξηση του βιοτικού επιπέδου συνεπάγεται και χρήση λιγότερο ενεργοβόρων συσκευών τόσο σε οικιακό όσο και σε βιομηχανικό επίπεδο [6].

- Μετεωρολογικοί παράγοντες: Ο καιρός αποτελεί προσδιοριστικό παράγοντα με πιο καθοριστική την επίδραση της θερμοκρασίας και δευτερογενείς παράγοντες την υγρασία, τις βροχοπτώσεις, τους άνεμους και την ηλιοφάνεια. Οι καιρικές συνθήκες προφανώς πέρα από την εποχή συνδυάζονται και από τη γεωγραφική περιοχή που ανήκει ο κάθε καταναλωτής, όπου ανάλογα τον κάθε τόπο η ζήτηση ηλεκτρικής ενέργειας έχει τα δικά της χαρακτηριστικά. Ακραίες καιρικές συνθήκες προκαλούν συνήθως αυξημένη ζήτηση φορτίου (π.χ. αύξηση χρήσης κλιματιστικών σε περιόδους καύσωνα)
- Εποχιακοί παράγοντες: Τα φορτία που καλείται να εξυπηρετήσει το σύστημα ηλεκτρικής ενέργειας παρουσιάζουν έντονες διακυμάνσεις τόσο κατά τη διάρκεια μιας μέρας όσο και του έτους. Αυτό μπορεί να οφείλεται στο εάν είναι μέρα ή νύχτα, στο είδος της ημέρας (π.χ. εργάσιμη, Σάββατο, Κυριακή, αργία), στην εποχή του χρόνου που συνοδεύεται από τις αντίστοιχες καιρικές συνθήκες, επηρεάζοντας την ανθρώπινη δραστηριότητα.
- Τυχαίοι παράγοντες: Αποτελούνται από τυχαίους παράγοντες όπως μεγάλες απεργίες, εκλογές, μετάδοση μεγάλων αθλητικών ή ψυχαγωγικών προγραμμάτων κ.α.
- Απροσδιόριστοι παράγοντες: Παράγοντες που είναι δύσκολο να προβλεφθούν και επηρεάζουν μακροπρόθεσμα την κατανάλωση, όπως οι προοπτικές ανάπτυξης μιας περιοχής και ο ρυθμός αύξησης του πληθυσμού της. Τέτοιο παράγοντα αποτελεί και η πανδημία του κορωνοϊού η οποία από το πρώτο τρίμηνο του 2020 εξαπλώθηκε και επηρέασε σε πολύ μεγάλο βαθμό την κατανάλωση ηλεκτρικής ενέργειας, ως αποτέλεσμα των μέτρων για την προσπάθεια περιορισμού της εξάπλωσης του ιού.

1.4 Κατηγοριοποίηση και πρόβλεψη στα συστήματα ηλεκτρικής ενέργειας

Οι μέθοδοι αναγνώρισης προτύπων έχουν εκτεταμένες εφαρμογές στα συστήματα ηλεκτρικής ενέργειας για την αντιμετώπιση προβλημάτων από τα αντίστοιχα κέντρα ελέγχου, αξιοπιστίας, βραχυπρόθεσμης πρόβλεψης και αξιοποίησης ταξινόμησης πελατών. Η οικονομική ανάπτυξη των χωρών σε συνδυασμό με την απελευθέρωση της αγοράς έχουν δημιουργήσει συνθήκες αυξημένου ανταγωνισμού μεταξύ των προμηθευτών παροχής ηλεκτρικής ενέργειας. Η ομαδοποίηση των πελατών δίνει τη δυνατότητα της μελέτης και αξιολόγησης των αναγκών της αγοράς και βάσει αυτών να υλοποιηθεί κατάλληλος σχεδιασμός των μονάδων παραγωγής, ενώ ταυτόχρονα συμβάλει και στον σχεδιασμό για εξοικονόμηση ενέργειας με κατάλληλη διαχείριση του φορτίου, βελτιώνοντας τον συντελεστή απόδοσης του ΣΗΕ. Οι προμηθευτές από την ταξινόμηση των πελατών και των ημερών μπορούν να διακρίνουν κατηγορίες πελατών με κοινά χαρακτηριστικά τιμολογώντας τους κατάλληλα. Ο καταναλωτής έχοντας την πληροφορία για την ομάδα στην οποία ανήκει δύναται να επιλέξει την αντίστοιχη αντιπροσωπευτική τιμολόγηση ή να προβεί σε δράση για τη διαχείριση ή την εξοικονόμηση της καταναλισκόμενης ενέργειας του.

Σημαντικός παράγοντας για την ομαλή και επικερδή λειτουργία του συστήματος ηλεκτρικής ενέργειας αποτελεί η δυνατότητα πρόβλεψης του μελλοντικού φορτίου των καταναλωτών, η οποία προϋποθέτει μια αποτελεσματική κατηγοριοποίηση καταναλωτών. Από 1η Νοεμβρίου 2020 η χονδρεμπορική αγορά ενέργειας λειτουργεί με το «target model», επιτρέποντας τον ημερήσιο ενεργειακό προγραμματισμό βάσει πρόβλεψης με στόχο τη μείωση του κόστους ενέργειας για τους τελικούς καταναλωτές και τη σύνδεση με τα ηλεκτρικά συστήματα της Ευρώπης [7], [8]. Η πρόβλεψη ζήτησης φορτίου χωρίζεται στις παρακάτω κατηγορίες:

- Την πολύ βραχυπρόθεσμη, όπου γίνεται η πρόβλεψη της ζήτησης του φορτίου για τα επόμενα 30 λεπτά έως μία ώρα, με βήματα μερικών δευτερολέπτων έως και λεπτού με στόχο την κάλυψη των αναγκών των συστημάτων αυτομάτου ελέγχου των μονάδων παραγωγής ηλεκτρικής ενέργειας (γεννητριών), τα οποία χρειάζονται όσο το δυνατό εγκυρότερη πληροφόρηση για τις μελλοντικές αλλαγές φόρτισής τους, ώστε να προχωρήσουν στις αναγκαίες μεταβολές της παροχής καυσίμου και ρύθμισης των επιπέδων τάσης εξόδου τους. Με αυτόν τον τρόπο είναι δυνατός ο έλεγχος της σχέσης συχνότητας-φορτίου και της ασφάλειας του συστήματος.
- Τη βραχυπρόθεσμη, όπου γίνεται η πρόβλεψη του φορτίου για το επόμενο 24ωρο έως μία εβδομάδα με χρονικό βήμα της μισής ή της μίας ώρας, που έχει ως στόχο τη ρύθμιση των βασικών επιπέδων λειτουργίας των μονάδων και παίζει σημαντικό ρόλο στη διαμόρφωση της σειράς ένταξης τους στην παραγωγική διαδικασία με βάση τα κριτήρια της οικονομικής κατανομής και λαμβάνοντας υπόψη τα αντίστοιχα προγράμματα συντήρησης. Αυτά την καθιστούν υπεύθυνη για την ενεργειακή διαχείριση του συστήματος.
- Τη μεσοπρόθεσμη, όπου η αντίστοιχη πρόβλεψη γίνεται για ένα έτος με χρονικό βήμα μίας εβδομάδας. Η μεσοπρόθεσμη πρόβλεψη χρησιμοποιείται για τη ρύθμιση των προγραμμάτων συντήρησης των μονάδων παραγωγής και αξιοποίησης των διαθέσιμων πόρων στις κατάλληλες χρονικές περιόδους του έτους.
- Τη μακροπρόθεσμη, όπου η αντίστοιχη πρόβλεψη γίνεται με χρονικό ορίζοντα 10-20 ετών με χρονικό βήμα ενός έτους. Η μακροπρόθεσμη πρόβλεψη φορτίου αφορά κυρίως την ετήσια αιχμή φορτίου και τη συνολική ετήσια ενέργεια για τα επόμενα 10 έως 20 χρόνια, ώστε να είναι δυνατός ο σχεδιασμός και η κατασκευή του συστήματος ηλεκτρικής ενέργειας, όπως η δημιουργία μονάδων παραγωγής, η κατασκευή γραμμών μεταφοράς ή η διαμόρφωση του δικτύου διανομής μίας μεγάλης περιοχής.

Η τιμολόγηση της ηλεκτρικής ενέργειας πρέπει να συνδυάζει τόσο το συμφέρον του πελάτη, όσο και αυτό του προμηθευτή. Η μέτρηση της καταναλισκόμενης ή αποδιδόμενης ενέργειας πραγματοποιείται από

κατάλληλους μετρητές που βρίσκονται στον χώρο λειτουργίας του συνδεδεμένου πελάτη στο σύστημα, πληροφορία που χρησιμοποιεί ο προμηθευτής για την κοστολόγηση της ενέργειας με συγκεκριμένη τιμή ανά kWh που ποικίλει ανάλογα την ιδιότητα του πελάτη (οικιακός, γεωργικός, εμπορικός, βιομηχανικός) και το επίπεδο τάσης που τροφοδοτείται. Στην Ελλάδα όλοι οι πελάτες που εξυπηρετούνται από τη μέση τάση (περίπου 13.000 παροχές) και ελάχιστοι από τη χαμηλή (περίπου 2.000 από τα 7 εκατομμύρια παροχές) μετρούνται ανά 15 λεπτά μέσω τηλεμέτρησης και τιμολογούνται με βάση τη μέγιστη ζήτησή τους για το διάστημα τιμολόγησης. Η κατηγοριοποίηση αυτών θα οδηγούσε σε σύνταξη κατάλληλων τιμολογίων για τον καθένα ξεχωριστά αναλόγως την ομάδα στην οποία ανήκει, ενώ μέχρι τώρα η τιμολόγησή του ήταν αναλόγως του επιχειρηματικού τομέα (χρήσης) που δραστηριοποιείται. Σε αυτή τη διπλωματική γίνεται επιλογή των πελατών μέσης τάσης που έχουν καταναλώσεις από 01-01-2018 έως και 01-06-2020 με δεδομένα ωριαίων μετρήσεων που παραχωρήθηκαν από τον ΔΕΔΔΗΕ για την κατάλληλη ταξινόμησή τους σε ομάδες, η οποία θα αξιοποιηθεί για την υλοποίηση πρόβλεψης για το φορτίο των επόμενων 24 ωρών. Η αναγνώριση προτύπων αποτελεί εναλλακτικό εργαλείο σχηματισμού τιμολογίων έναντι αυτής, με δειγματοληψία βασικών κατηγοριών πελατών [9].

Το αντικείμενο της κατηγοριοποίησης στα συστήματα ηλεκτρικής ενέργειας είναι συνδυασμός της εύρεσης δεδομένων, της αναγνώρισης προτύπων, της στατιστικής και της τεχνητής νοημοσύνης με όρια δυσδιάκριτα για την κάθε μία.

Κεφάλαιο 2
Τεχνικές Κατηγοριοποίησης

2.1 Εισαγωγή

Η κατηγοριοποίηση για ένα δίκτυο ηλεκτρικής ενέργειας διακρίνεται σε κατηγοριοποίηση βάσει του χρόνου και βάσει των πελατών. Στην πρώτη ως πρότυπα λαμβάνονται οι χρονικές μονάδες (κατά κανόνα οι ημέρες) που συνθέτουν ένα ευρύτερο χρονικό πλαίσιο (μήνας, έτος κλπ.). Αντίστοιχα, στην περίπτωση της κατηγοριοποίησης βάσει πελατών ως πρότυπα λαμβάνονται όλοι οι καταναλωτές της ηλεκτρικής ενέργειας, οποιαδήποτε χρήση και αν έχουν (εμπορική, βιομηχανική, γεωργική κλπ.), οι οποίοι εμφανίζουν εντελώς διαφορετική ηλεκτρική συμπεριφορά. Στις περιπτώσεις που αναφέρθηκαν τα πρότυπα ταξινομούνται στις ομάδες που σχηματίζονται με βάση τα κοινά χαρακτηριστικά τους με κριτήριο τον χρόνο ή τη ζήτηση φορτίου του κάθε καταναλωτή [19].

Στην κατηγοριοποίηση βάσει χρόνου όπου ως πρότυπα θεωρούνται οι ημέρες του έτους, η ομαδοποίηση τους γίνεται με την εύρεση και την ταξινόμηση σε ομάδες εκείνων των ημερών στις οποίες εμφανίζεται παραπλήσια ζήτηση φορτίου από τους καταναλωτές κατά τη διάρκεια ενός 24ώρου. Αρχική κατηγοριοποίηση θα μπορούσε να θεωρηθεί ο διαχωρισμός των ημερών σε δύο ομάδες, δηλαδή σε εργάσιμες και σε αργίες. Ωστόσο, πρέπει να ληφθούν υπόψη το είδος της αργίας, η εποχή του έτους, οι καιρικές συνθήκες, απεργίες, εκδηλώσεις και άλλοι παράγοντες, αυξάνοντας την πολυπλοκότητα. Έτσι, ακόμα και δύο διαφορετικές αργίες δε σημαίνει αυτόματα ότι θα έχουν παρόμοια ζήτηση φορτίου (πχ Πρωτοχρονιά και Δεκαπενταύγουστος). Οι αργίες, οι απεργίες και οι ημέρες των διακοπών χαρακτηρίζονται ως ανώμαλες λόγω της ιδιομορφίας που παρουσιάζουν στην κατανάλωση ενέργειας, με αποτέλεσμα να συνίσταται η δημιουργία διψήφιου αριθμού ομάδων για την κατηγοριοποίηση των ημερών.

Η κατηγοριοποίηση βάσει πελατών έχει ως πρότυπα όλους τους καταναλωτές της ηλεκτρικής ενέργειας οι οποίοι εμφανίζουν εντελώς διαφορετική ηλεκτρική συμπεριφορά μεταξύ τους. Η ομαδοποίηση πραγματοποιείται σε δύο σκέλη:

- Στο πρώτο στάδιο δημιουργούνται ομάδες βάσει διεθνών κριτηρίων, όπου οι πελάτες διαχωρίζονται με βάσει το επίπεδο τάσης που τροφοδοτούνται και ανάλογα με τη χρήση της ενέργειας που καταναλώνουν. Προκύπτουν οι ομάδες σε χαμηλής, μέσης και υψηλής τάσης, ο διαχωρισμός σε οικιακούς και μη καταναλωτές, καθώς και ταξινόμησή τους βάσει των ιδιαιτέρων γεωγραφικών και κλιματολογικών συνθηκών που είναι δυνατό να εμφανίζουν (π.χ. νησιώτικος/αστικός/ορεινός πληθυσμός).
- Έπειτα, εκτελείται νέα ειδικότερη κατηγοριοποίηση για την κάθε ομάδα που δημιουργήθηκε ξεχωριστά. Ο αριθμός των ομάδων που αυξάνει σημαντικά, καθώς πρέπει να εξασφαλίζεται η ομοιομορφία των προτύπων κάθε ομάδας ως προς την ηλεκτρική συμπεριφορά ώστε να συνυπάρχουν στην ίδια κατηγορία.

Η ηλεκτρική συμπεριφορά του καταναλωτή εντοπίζεται από το φορτίο που ζητάει κατά τη διάρκεια μίας χρονικής περιόδου. Η παρουσίαση της ζήτησης αυτής πραγματοποιείται είτε με διάγραμμα τιμών στο οποίο καταγράφονται τα ποσά και οι αντίστοιχες ώρες ζήτησής τους ή με κατασκευή χρονολογικών καμπυλών φορτίου, όπου αποτυπώνεται ποιοτικά η διακύμανση της καταναλισκόμενης ισχύος σε κάθε χρονικό διάστημα.

Το μειονέκτημα εντοπίζεται στον χρόνο που απαιτείται για την επεξεργασία μεγάλου πλήθους δεδομένων, οδηγώντας βέβαια σε αξιόπιστα αποτελέσματα. Στη βιβλιογραφία προτείνεται η χρήση δεικτών που προσομοιώνουν ικανοποιητικά την ηλεκτρική συμπεριφορά των πελατών και την αξιοποίηση αυτών για την κατηγοριοποίηση [10]. Οι δείκτες αυτοί χωρίζονται σε ημερήσιους και σε εβδομαδιαίους. Ως ημερήσιοι λαμβάνονται:

- ο λόγος της μέσης κατανάλωσης ισχύος προς τη μέγιστη
- ο λόγος της ελάχιστης κατανάλωσης προς τη μέγιστη
- δείκτης που εκφράζει την επίδραση της νύχτας για κάθε ημέρα
- δείκτης που εκφράζει την επίδραση του μεσημεριανού γεύματος της εργάσιμης ημέρας

Αντίστοιχα ορίζονται και οι εβδομαδιαίοι δείκτες. Οι ιδιαίτερες συνήθειες που χαρακτηρίζουν τους καταναλωτές διαφορετικών περιοχών καθιστούν την παραπάνω μεθοδολογία σχετικά αναποτελεσματική και επισημαίνεται πως χρειάζονται περαιτέρω έρευνες και μελέτες για την εύρεση δεικτών που να παρουσιάζουν τις χρήσιμες πληροφορίες για την ηλεκτρική συμπεριφορά των πελατών.

Ο κάθε καταναλωτής παρουσιάζει μοναδικά ηλεκτρικά χαρακτηριστικά με αποτέλεσμα ακόμα και η προσεγμένη και μελετημένη ταξινόμησή του σε μία συγκεκριμένη ομάδα βάσει αυστηρών κριτηρίων, είναι πιθανό να εμπεριέχει μια μικρή απόκλιση μεταξύ της πραγματικής ηλεκτρικής συμπεριφοράς του, που προσδιορίζεται από τη μετρούμενη κατανάλωσή του και τη θεωρητική, που προσδιορίζεται ως αυτή που του αποδίδεται από την ομάδα στην οποία ανήκει. Η επιλεγμένη ομαδοποίηση πρέπει να ικανοποιεί την ελαχιστοποίηση του σφάλματος που παρουσιάζεται και τον περιορισμό των ομάδων που δημιουργούνται, χαρακτηριστικά που είναι αλληλοαναιρούμενα αποτελώντας καίρια την εύρεση της χρυσής τομής μεταξύ τους.

Το άρθρο [11] υπαγορεύει ότι για να είναι επιτυχημένη η κατηγοριοποίηση πρέπει πρώτα να υλοποιηθούν συγκεκριμένες διεργασίες. Αρχικά, συλλέγονται όλες οι απαραίτητες πληροφορίες, όπως ιστορικά δεδομένα, στοιχεία καιρού, αριθμός πελατών κ.α., τα οποία καταχωρούνται σε βάση δεδομένων (database). Το σύνολο των διεργασιών για τη συλλογή και την αποθήκευση των δεδομένων ονομάζεται εύρεση γνώσης σε βάσεις δεδομένων, το οποίο αποτελείται από τα παρακάτω στάδια:

1. Προ-επεξεργασία (Preprocessing), η οποία περιλαμβάνει τη διόρθωση, την επιλογή και τον μετασχηματισμό σε επιθυμητή μορφή των δεδομένων. Στη φάση της διόρθωσης, γίνεται προσεκτικός έλεγχος των δεδομένων με ενδεχόμενη διόρθωση, συμπλήρωση και εξάλειψη κάποιων παραμέτρων π.χ. στοιχεία που λείπουν από το σύνολο των δεδομένων οποία πρέπει να συμπληρωθούν με γραμμική παρεμβολή ή άλλους τρόπους. Η επιλογή συνίσταται στην αξιολόγηση των δεδομένων αναλόγως την αντιπροσωπευτικότητα των πληροφοριών που περιέχουν, απορρίπτοντας όσες ερμηνεύονται ως θόρυβος (π.χ. μια λάθος μέτρηση).
2. Εύρεση δεδομένων (Data Mining), όπου μελετώνται τα δεδομένα που προκρίθηκαν κατά την προ-επεξεργασία, ώστε να βρεθούν χαρακτηριστικά που ίσως αγνοήθηκαν. Η τελική φάση της εύρεσης δεδομένων συνίσταται στην επιλογή της τεχνικής της κατηγοριοποίησης που θα χρησιμοποιηθεί με γνώμονα τα πρότυπα που έχουν επιλεγεί και το επιθυμητό αποτέλεσμα.
3. Μετά-επεξεργασία (Post-processing), που πραγματοποιείται όταν η τεχνική που έχει επιλεχθεί στο προηγούμενο στάδιο εκτελέσει την κατηγοριοποίηση των προτύπων και δώσει τον τελικό αριθμό των ομάδων. Τα αποτελέσματα που εξάγονται οδηγούν σε συμπεράσματα που αποσαφηνίζονται και προσδιορίζονται κάποιοι κανόνες γνώσης.

Οι αλγόριθμοι χρησιμοποιούν κάποιον κανόνα εκμάθησης όπως η οπισθοδιάδοση, η τυχαία μάθηση και η ανταγωνιστική μάθηση. Η ανταγωνιστική μάθηση (competitive learning) χρησιμοποιείται σε προβλήματα μάθησης χωρίς επίβλεψη, όπως η κατηγοριοποίηση, όπου η διαθέσιμη πληροφορία σχετικά με τις επιθυμητές εξόδους δεν παρέχεται και τα συνοπτικά βάρη ανανεώνονται μόνο με βάση τα δεδομένα εισόδου. Επίσης, υλοποιείται με τη χρήση ενός στρώματος, το οποίο αποτελείται από «ανταγωνιστικούς» νευρώνες και ονομάζεται ανταγωνιστικό επίπεδο (competitive layer). Όταν εισάγεται ένα διάνυσμα εισόδου στο ανταγωνιστικό επίπεδο οι νευρώνες αγωνίζονται για τη διεκδίκηση του συγκεκριμένου προτύπου. Ο νευρώνας

που προκρίνεται είναι συνήθως αυτός ο οποίος έχει μεγαλύτερη μεταβολή στις τιμές των βαρών του συγκριτικά με τους υπόλοιπους νευρώνες. Έτσι, επιτυγχάνεται αυτό-οργάνωση, όπου ορισμένοι νευρώνες μαθαίνουν να ανταποκρίνονται σε συγκεκριμένα ερεθίσματα (inputs). Ο νικητής εντοπίζεται όταν η απόσταση μεταξύ του διανύσματος εισόδου νευρώνα είναι η ελάχιστη συγκριτικά με τις αποστάσεις που αφορούν τους υπόλοιπους νευρώνες που ανταγωνίζεται. Η διαδικασία που περιεγράφηκε επαναλαμβάνεται για κάθε διάνυσμα εισόδου που εμφανίζεται στο ανταγωνιστικό επίπεδο [12].

Η ανταγωνιστική μάθηση εμπεριέχει ως μειονέκτημα το γεγονός ότι ορισμένα διανύσματα βάρους που επιλέγονται αρχικά τυχαία, μπορεί να βρίσκονται πολύ μακριά από οποιοδήποτε διάνυσμα εισόδου. Αυτό θα έχει αποτέλεσμα να έχουν μικρή πιθανότητα να ανανεωθούν κατά τον επανυπολογισμό των βαρών. Ωστόσο, αυτό μπορεί να επιλυθεί με αρχικοποίηση των βαρών σε δειγματικές τιμές που προκύπτουν από τα δεδομένα εισόδου, ώστε αφού παρουσιαστούν τα πρότυπα να λάβει χώρα ανανέωση όλων των βαρών. Εναλλακτικό τρόπο αντιμετώπισης αποτελεί η ανανέωση όλων των βαρών των νευρώνων, δηλαδή τόσο αυτών που κερδίζουν, όσο και αυτών που χάνουν, όπου στους δεύτερους θα χρησιμοποιείται πολύ μικρότερος συντελεστής (ρυθμού) μάθησης. Η συγκεκριμένη παραλλαγή της ανταγωνιστικής μάθησης ονομάζεται διαρρέουσα μάθηση.

Ως N ορίζεται ο συνολικός αριθμός των διανυσμάτων εισόδου προς ομαδοποίηση και το τυχαίο διάνυσμα εισόδου l συμβολίζεται ως $\overrightarrow{x_l}$, όπως δίνεται από τη σχέση 2.1 . Επίσης, ως d ορίζεται ο αριθμός των διαστάσεων του κάθε διανύσματος στην είσοδο.

$$\overrightarrow{x_l} = (x_{l1}, x_{l2}, \dots, x_{li}, \dots, x_{ld})^T \quad , \quad \text{όπου } 1 \le \text{i} \le \text{d} \text{ και } 1 \le \text{l} \le \text{N} \tag{2.1}$$

Έτσι, τα διανύσματα εισόδου, που είναι όσα οι ημέρες με το καθένα να έχει 24 διαστάσεις, μία για κάθε ώρα της ημέρας. Έτσι, για παράδειγμα το x_{310} αντιστοιχεί στις 10:00π.μ. της ημέρας 03/01/2018 με τελευταίο στοιχείο της ίδιας ημέρας το x_{324}.

Τα αναφερόμενα διανύσματα πρέπει να κατηγοριοποιηθούν σε M διαφορετικές ομάδες, όπου η κάθε ομάδα j έχει έναν αντιπρόσωπο – κέντρο ομάδας που την περιγράφει πλήρως και θα συμβολίζεται με $\overrightarrow{w_j}$. Για την κατηγοριοποίηση του πελάτη που αναφέρθηκε, κάθε τυχαία ομάδα αντιπροσωπεύεται από μια τυπική ημερήσια χρονολογική καμπύλη, η οποία παρουσιάζει σημαντικές ομοιότητες με τις ημερήσιες χρονολογικές καμπύλες των ημερών που έχουν ταξινομηθεί στη συγκεκριμένη ομάδα. Έτσι, διαπιστώνεται πως υπάρχει ποιοτική ομοιότητα μεταξύ των διανυσμάτων $\overrightarrow{x_l}$ και $\overrightarrow{w_j}$, καθώς αμφότερα έχουν τις ίδιες διαστάσεις d και ταυτόχρονα το κέντρο της τυχαίας ομάδας j δίνεται από τη σχέση 2.2:

$$\overrightarrow{w_j} = \left(w_{j1}, w_{j2}, \dots, w_{ji}, \dots, w_{jd}\right)^T \quad , \quad \text{όπου } 1 \le \text{i} \le \text{d} \text{ και } 1 \le \text{j} \le \text{N} \tag{2.2}$$

Αν ως θεώρηση είναι επιθυμητή η ταξινόμηση όλων των ημερών σε 10 ομάδες θα έχουμε M=10, με το κάθε κέντρο $\overrightarrow{w_j}$ να έχει 24 διαστάσεις. Έτσι, το κέντρο της δεύτερης ομάδας του συγκεκριμένου πελάτη θα δίνεται από τη σχέση 2.3:

$$\overrightarrow{w_2} = (w_{21}, w_{22}, \dots, w_{2i}, \dots, w_{224})^T \quad , \quad \text{όπου } 1 \le \text{i} \le 24 \tag{2.3}$$

Το υποσύνολο των διανυσμάτων που ανήκουν σε μία από τις ομάδες, έστω την k συμβολίζεται ως (Σχ. 2.4):

$$\Omega_k = \{ \overrightarrow{x_l}, l = 1 \dots N \} \text{ , όπου N είναι τα διανύσματα εισόδου} \tag{2.4}$$

Το σύνολο των κέντρων για κάθε ομάδα k ορίζονται ως W, τα οποία προκύπτουν από τις καμπύλες των προτύπων $\overrightarrow{x_l}$ που ανήκουν στη συγκεκριμένη ομάδα:

$$W = \{ \overrightarrow{w_k}, k = 1 \dots M \} \tag{2.5}$$

Στη συνέχεια του κεφαλαίου γίνεται αναλυτική παρουσίαση των μεθόδων κατηγοριοποίησης (είτε αλγοριθμικές, είτε με χρήση τεχνητής νοημοσύνης), καθώς και των τρόπων αξιολόγησης τους ακολουθούμενη από την ανάλυση της μεθοδολογίας που θα χρησιμοποιηθεί.

2.2 Αποστάσεις και Αξιολόγηση κατηγοριοποίησης

Στο πλαίσιο της μελέτης και αξιολόγησης των αλγορίθμων ταξινόμησης, ορίζονται τα ακόλουθα είδη αποστάσεων:

1. Ευκλείδεια απόσταση μεταξύ δύο τυχαίων διανυσμάτων εισόδου (έστω $\overrightarrow{x_{l1}}$ και $\overrightarrow{x_{l2}}$) του συνόλου *N*, η οποία δίνεται από τη σχέση:

$$d(\overrightarrow{x_{l1}}, \overrightarrow{x_{l2}}) = \sqrt{\frac{1}{d}\sum_{i=1}^{d}(x_{l1i} - x_{l2i})^2} \tag{2.6}$$

2. Η απόσταση μεταξύ της τυπικής καμπύλης $\overrightarrow{w_k}$ της ομάδας *k* και του υποσυνόλου Ω_k ορίζεται ως ο γεωμετρικός μέσος των ευκλείδειων αποστάσεων μεταξύ του κέντρου $\overrightarrow{w_k}$ και του κάθε προτύπου που εμπεριέχεται στο Ω_k.

$$d(\overrightarrow{w_k}, \Omega_\mathrm{k}) = \sqrt{\frac{1}{N_k}\sum_{\overrightarrow{x_l}\epsilon\Omega_\mathrm{k}} d^2(\overrightarrow{w_k}, \overrightarrow{x_l})} \tag{2.7}$$

3. Η μέση απόσταση των μελών της κάθε ομάδας μεταξύ τους (infra set mean distance) ορίζεται ως ο γεωμετρικός μέσος των ενδο-αποστάσεων (inter distances) μεταξύ των μελών του υποσυνόλου Ω_k ή *W*, σύμφωνα με τις σχέσεις:

$$\hat{d}(\Omega_\mathrm{k}) = \sqrt{\frac{1}{2 \cdot N_k}\sum_{\overrightarrow{x_l}\epsilon\Omega_\mathrm{k}} d^2(\overrightarrow{x_l}, \Omega_\mathrm{k})} \tag{2.8}$$

$$\hat{d}(\mathrm{W}) = \sqrt{\frac{1}{2 \cdot M}\sum_{k=1}^{M} d^2(\overrightarrow{w_k}, \mathrm{W})} \tag{2.9}$$

Η έγκυρη αξιολόγηση και σύγκριση των αποτελεσμάτων που εξάγονται από κάθε μέθοδο που υλοποιείται υπαγορεύει την επινόηση και χρήση μαθηματικών δεικτών που ονομάζονται δείκτες αξιολόγησης. Αυτοί είναι:

1. Δείκτης Αθροίσματος Τετραγωνικού Λάθους (Sum of Square Error – SSE) ή J, που εκφράζει την απόσταση του κάθε διανύσματος από το κέντρο της ομάδας, στο οποίο ανήκει για την ίδια τιμή βάρους [13]:

$$J = \frac{1}{N}\sum_{l=1}^{N} d^2\left(\overrightarrow{x_l}, \vec{w}_{k:\overrightarrow{x_l}\in\Omega_k}\right) \tag{2.10}$$

2. Δείκτης Mean Index Adequacy (M.I.A.) ή μέσος δείκτης καταλληλότητας, που δίνει το άθροισμα των τετραγώνων των αποστάσεων των διανυσμάτων εισόδου από τα αντίστοιχα κέντρα βάρους τους, προς το πλήθος των μελών της εκάστοτε ομάδας διαιρεμένο με τον συνολικό αριθμό πραγματικών ομάδων.

$$MIA = \sqrt{\frac{1}{M}\sum_{k=1}^{M} d^2(\overrightarrow{w_k}, \Omega_k)} \tag{2.11}$$

3. Δείκτης Clustering Dispersion Indicator (C.D.I) ή δείκτης διασποράς ομαδοποίησης, που δείχνει πόσο συμπαγείς είναι οι ομάδες λαμβάνοντας υπόψη τον λόγο του αθροίσματος των μέσων αποστάσεων των μελών της κάθε ομάδας μεταξύ τους προς τις ενδο-αποστάσεις των αντιπροσωπευτικών κέντρων τους.

$$CDI = \frac{1}{\hat{d}(\mathrm{W})}\cdot\sqrt{\frac{1}{M}\sum_{k=1}^{M}\hat{d}^2(\Omega_k)} \tag{2.12}$$

4. Δείκτης Similarity Matrix Indicator (S.M.I) ή δείκτης του πίνακα ομοιότητας, που ορίζει το μέγιστο μη διαγώνιο στοιχείο του συμμετρικού πίνακα ομοιότητας, του οποίου οι όροι υπολογίζονται από τη λογαριθμική συνάρτηση της ευκλείδειας απόστασης μεταξύ οποιουδήποτε ζεύγους καμπυλών φορτίου αντιπροσώπων.

$$SMI = max_{p>q}\left\{\left(1 - \frac{1}{\ln\left[d(\overrightarrow{w_p}, \overrightarrow{w_q})\right]}\right)^{-1}\right\} : p, q = 1, \dots, M \tag{2.13}$$

5. Δείκτης Davies-Bouldin (D.B.I.) με χρήση ευκλείδειων αποστάσεων, που αναπαριστά μία μέση ένδειξη των μέτρων ομοιότητας της κάθε αντιπροσωπευτικής ομάδας με την πιο όμοια ομάδα όλου του συστήματος.

$$DBI = \frac{1}{M}\cdot\sum_{k=1}^{M} max_{p\neq q}\left[\frac{\hat{d}(\Omega_p) + \hat{d}(\Omega_q)}{d(\overrightarrow{w_p}, \overrightarrow{w_q})}\right] : p, q = 1, \dots, M \tag{2.14}$$

6. Δείκτης Ratio of within cluster sum of squares to between cluster variation (WCBCR), δηλαδή ο λόγος του αθροίσματος των τετραγώνων των αποστάσεων κάθε διανύσματος εισόδου από το κέντρο της ομάδας που ανήκει προς τη διασπορά μεταξύ των κέντρων των ομάδων. Εκφράζει στον αριθμητή τις αποστάσεις των

διανυσμάτων εισόδου από τα αντίστοιχα κέντρα και στον παρονομαστή την ομοιότητα μεταξύ αντιπροσωπευτικών κέντρων [2].

$$\left[\sum_{k=1}^{M}\sum_{\overrightarrow{x_l}\in\Omega_k} d^2(\overrightarrow{w_k},\overrightarrow{x_l})\right] \Big/ \left[\sum_{l\le q<p}^{M} d^2(\overrightarrow{w_p},\overrightarrow{w_q})\right] \tag{2.15}$$

7. Δείκτης Silhouette Score, που αναπαρίσταται γραφικά, όπως φαίνεται στο Διάγραμμα 2.5 και χρησιμοποιείται για τη μελέτη της απόστασης διαχωρισμού μεταξύ των ομάδων που προκύπτουν, όπου αναδεικνύει πόσο κοντά βρίσκεται ένα σημείο συγκεκριμένου cluster από τα υπόλοιπα των γειτονικών του. Ο δείκτης αυτός έχει σύνολο τιμών [-1,1] και υπολογίζεται ως εξής: [14]

$$s(i) = \frac{b(i)-a(i)}{\max\{a(i),b(i)\}}\text{ , εάν ο αριθμός των σημείων του cluster είναι μεγαλύτερος από 1} \tag{2.16}$$

ή διαφορετικά

$$s(i) = \begin{cases} 1-\dfrac{a(i)}{b(i)}\text{ , εάν } a(i) < b(i) \\ 0 \qquad\quad \text{, εάν } a(i) = b(i) \\ \dfrac{a(i)}{b(i)} - 1\text{ , εάν } a(i) > b(i) \end{cases} \tag{2.17}$$

όπου:

- $a(i)$ είναι η μέση τιμή της απόστασης μεταξύ του στοιχείου i και των υπολοίπων που ανήκουν στο ίδιο cluster. Όσο μικρότερη η τιμή α(i), τόσο καλύτερη η τοποθέτηση του στοιχείου στο συγκεκριμένο cluster
- $b(i)$ είναι η ελάχιστη από όλες τις μέσες τιμές των αποστάσεων των σημείων i με τα σημεία ενός άλλου που δεν είναι μέλος, η οποία δείχνει τη γειτονική ομάδα του στοιχείου, δηλαδή την επόμενη καλύτερη από αυτή που τοποθετήθηκε

Διάγραμμα 2.5: Δείκτης Silhouette Score 6 clusters με μέση τιμή S= 0.450467 [15].

Στην παρούσα εργασία η εφαρμογή του κάθε αλγόριθμου θα αξιολογηθεί από τους δείκτες αθροίσματος τετραγωνικού λάθους (Sum of Square Error – SSE ή WSS), Silhouette Score και Davies Bouldin.

2.3 Ανάλυση τεχνικών κατηγοριοποίησης

Οι βασικές τεχνικές κατηγοριοποίησης προτύπων μη επιβλεπόμενης εκπαίδευσης και οι παραλλαγές τους αναλύονται στη συνέχεια του κεφαλαίου [16].

2.3.1 Κ-μέσων (k-means)

Ο αλγόριθμος Κ-μέσων (k-means) αποτελεί την απλούστερη μέθοδο παραμετρικής σημειακής ομαδοποίησης κατάλληλη για την ταξινόμηση των προτύπων σε συμπαγείς ομάδες και τη μέθοδο που θα χρησιμοποιηθεί στην υλοποίηση της κατηγοριοποίησης [17]. Κάθε ομάδα j αντιπροσωπεύεται από ένα διάνυσμα $\overrightarrow{w_j}$, το οποίο αποτελεί ανταγωνιστικό νευρώνα σε ανταγωνιστικό επίπεδο. Η εκπαίδευση του αλγόριθμου ξεκινά με την παρουσίαση των διανυσμάτων σε συγκεκριμένη σειρά, συνήθως χρονολογική, ίδια για κάθε εποχή – σειριακή παρουσίαση ανά εποχή. Η μορφή ανανέωσης των βαρών είναι ο λόγος που η τυχαία παρουσίαση ανά εποχή δε δίνει καλύτερα αποτελέσματα. Οι διεργασίες που εκτελεί ο αλγόριθμος είναι οι εξής:

1. Κανονικοποίηση τιμών εισόδου όλων των στοιχείων x_{li} των διανυσμάτων, αντιστοιχώντας 0 και 1 στην ελάχιστη και στη μέγιστη τιμή των στοιχείων. Έτσι, επιτυγχάνεται περιορισμένο εύρος τιμών για αποδοτικότερη επεξεργασία.

2. Αρχικοποίηση των M κέντρων $\overrightarrow{w_j}$, όπου j = 1,2,…,M. Αυτό το βήμα μπορεί να επιτευχθεί με διάφορους τρόπους, δύο εκ των οποίων είναι:
   i. Με επιλογή M διανυσμάτων από το αρχικό σύνολο των N διανυσμάτων εισόδου και σημαίνοντάς τα ως κέντρα για κάθε ομάδα.

ii. Με αρχικοποίηση μέσω του αλγόριθμου k-means++ που θα χρησιμοποιηθεί για την υλοποίηση της κατηγοριοποίησης, όπως αυτή αναλύεται στο Κεφ.4. Τα βήματα του αλγορίθμου είναι:
    1. Τυχαία επιλογή ενός διανύσματος εισόδου ως κέντρο.
    2. Υπολογισμός των αποστάσεων όλων των σημείων του dataset από το επιλεγμένο κέντρο ως εξής:

$$d_i = max_{(j:1\to m)} \left|\left|x_i - C_j\right|\right|^2 \text{, όπου m : κέντρα που έχουν ήδη επιλεχθεί} \quad (2.18)$$

    3. Ταξινόμηση όλων των αποστάσεων σε φθίνουσα σειρά και επιλογή ως κέντρου του σημείου με τη μεγαλύτερη απόσταση.
    4. Επανάληψη από το βήμα 2 έως ότου όλα τα κέντρα M να έχουν επιλεγεί.

3. Για κάθε πρότυπο εκπαίδευσης $\overrightarrow{x_l}$ υπολογίζονται οι αποστάσεις d($\overrightarrow{x_l}$, $\overrightarrow{w_k}$), όπου το διάνυσμα αυτό κατατάσσεται στην ομάδα Ω_k με τη μικρότερη ευκλείδεια απόσταση, δηλαδή βάσει του κριτηρίου:

$$d(\overrightarrow{x_l}, \overrightarrow{w_k}) = min_j(\overrightarrow{x_l}, \overrightarrow{w_j}) \quad (2.19)$$

4. Υπολογίζονται τα νέα κέντρα για κάθε ομάδα j με N_j τον αριθμό των προτύπων που έχουν ταξινομηθεί στην ομάδα j:

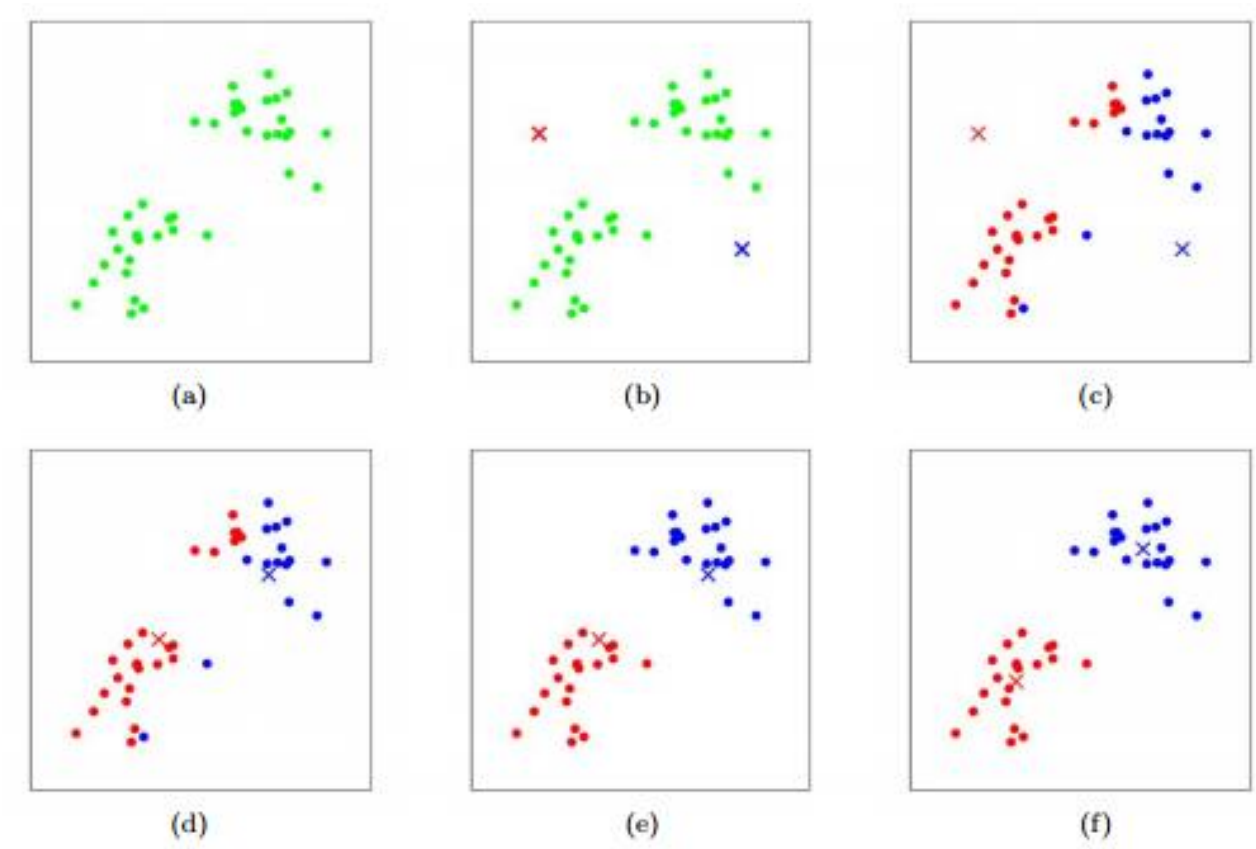
$$\overrightarrow{w_j}^{(t+1)} = \frac{1}{N_j^{(t)}} \sum_{\overrightarrow{x_l}\epsilon\Omega_j^{(t)}} \overrightarrow{x_l} \quad (2.20)$$

5. Έπειτα, πραγματοποιείται αύξηση του αριθμού των εποχών κατά ένα και ελέγχεται αν έχει συμπληρωθεί ο μέγιστος αριθμός των επιτρεπόμενων επαναλήψεων ή αν ισχύει $\left|\overrightarrow{w_j}^{(t)} - \overrightarrow{w_j}^{(t+1)}\right| < \varepsilon$ για κάθε j, δηλαδή να μην υπάρχει ουσιαστική αλλαγή στα βάρη. Αν κανένα από τα δύο δεν ισχύει, τότε επαναλαμβάνεται το βήμα 3.

Ταυτόχρονα, υπολογίζονται οι τιμές των δεικτών αξιολόγησης που αποτελούν το κριτήριο σύγκρισης μεταξύ των μεθόδων κατηγοριοποίησης. Ο αριθμός των ομάδων αρχικά είναι άγνωστος, με αποτέλεσμα την ανάγκη διενέργειας δοκιμών για τον προσδιορισμό του βέλτιστου πλήθους ομάδων ανάλογα τη χρονική περίοδο και το αντίστοιχο σύνολο ημερών.

Τα στάδια υλοποίησης του K-means οπτικά φαίνονται στο Διάγραμμα 2.6 που ακολουθεί. Τα βήματα που αποτυπώνονται είναι:

- a = Αρχικό dataset
- b = Επιλογή τυχαίου κέντρου του cluster
- c έως f = Υλοποίηση δύο επαναλήψεων του αλγορίθμου, όπου σε κάθε μία η είσοδος αντιστοιχίζεται στο πλησιέστερο κέντρο και έπειτα το κέντρο μετακινείται στον μέσο όρο των σημείων που του έχουν ανατεθεί.



Διάγραμμα 2.6: Στάδια υλοποίησης K-means [18].

2.3.2 Εκπαιδευόμενος Διανυσματικός Κβαντιστής - LVQ

Ο αλγόριθμος του εκπαιδευόμενου διανυσματικού κβαντιστή (Learning Vector Quantization) βασίζεται στην ανταγωνιστική μάθηση και αποτελεί παραλλαγή του αλγόριθμου K-μέσων, όπου σε κάθε βήμα επιλέγεται ένα διάνυσμα εισόδου και προσαρμόζεται σχετικά με το κέντρο της ομάδας j, στο οποίο τοποθετείται το πρότυπο. Η ομάδα που επιλέγεται μετακινεί το κέντρο της προς την κατεύθυνση του προτύπου, ενώ οι υπόλοιπες είτε μένουν ακίνητες, είτε κινούνται στην αντίθετη κατεύθυνση. Ο αλγόριθμος υλοποιείται με νευρωνικό δίκτυο, που αποτελείται από ένα ανταγωνιστικό επίπεδο, που οι νευρώνες είναι τόσοι όσες και οι ομάδες, με τον καθένα να αναπαρίσταται από το διάνυσμα που προκύπτει από τη σχέση 2.20.

Ο αλγόριθμος του εκπαιδευόμενου διανυσματικού κβαντιστή μετακινεί το κέντρο της νικήτριας ομάδας για κάποιο πρότυπο, εφόσον η ευκλείδεια απόσταση του κέντρου της από το πρότυπο είναι η ελάχιστη σε σχέση με των υπολοίπων ομάδων. Ειδικότερα, τα βήματα του αλγορίθμου είναι τα εξής:

α) Κανονικοποίηση των τιμών εισόδου των στοιχείων των διανυσμάτων, όπου αντιστοιχείται η ελάχιστη τιμή τους στο 0.1 και η μέγιστη στο 0.9, ώστε οι τιμές να κυμαίνονται σε περιορισμένο εύρος.

β) Καθορισμός του αριθμού *M* των κέντρων (ανταγωνιστικών νευρώνων) και αρχικοποίηση των βαρών στην τιμή $w_{ji} = 0.5$.

γ) Ο ρυθμός μάθησης η αρχικοποιείται, μέσω της σχέσης 2.21:

$$\eta(t) = \eta_o + \eta_o \cdot e^{-\frac{t}{T}} \tag{2.21}$$

Όπου:

- t : οι εποχές, δηλαδή η ακολουθία εμφανίσεων όλων των διανυσμάτων εισόδου στο νευρωνικό δίκτυο από μία φορά
- η_s : ο όρος που δεν επιτρέπει στον ρυθμό μάθησης να μειωθεί λιγότερο από μία συγκεκριμένη τιμή
- η_o: το μεταβλητό μέρος του ρυθμού μάθησης, ώστε όταν ξεκινά η μέθοδος ο ρυθμός να έχει μεγάλη τιμή, επιταχύνοντας τη σύγκλιση
- T: η χρονική παράμετρος

δ) Για κάθε εποχή πραγματοποιούνται τα ακόλουθα στάδια:

i. Σειριακή ή τυχαία παρουσίαση των προτύπων εκπαίδευσης $\overrightarrow{x_l}$

ii. Εύρεση του νικητή νευρώνα k υπολογίζοντας τις αποστάσεις $d\left(\overrightarrow{x_l}, \overrightarrow{w_j}\right)$ και επιλέγοντας την ομάδα που το κέντρο της έχει τη μικρότερη ευκλείδεια απόσταση από το διάνυσμα

iii. Υπολογισμός του νέου διανύσματος βαρών του νευρώνα k, δηλαδή της νέας θέσης $w_k^{(t')}$ για το κέντρο της νικήτριας ομάδας k μέσω της σχέσης 2.22, ενώ ταυτόχρονα τα βάρη των υπόλοιπων νευρώνων παραμένουν σταθερά:

$$w_k^{(t')} = w_{ki}^{(t)} + \eta \cdot (x_{li} - w_{ki}^{(t)}), \forall\, i = 1, \dots, d \tag{2.22}$$

ε) Με την ολοκλήρωση κάθε εποχής, το πλήθος των εποχών αυξάνεται κατά ένα με μεταβολή του ρυθμού μάθησης, σύμφωνα με τη σχέση 2.20 του βήματος γ. Έπειτα, ελέγχεται αν έχει ξεπεραστεί ο μέγιστος αριθμός επαναλήψεων ή ισχύει η σχέση $\left|\overrightarrow{w_j}^{(t)} - \overrightarrow{w_j}^{(t+1)}\right| < \varepsilon$ για κάθε j, δηλαδή δεν υπάρχει πλέον ουσιαστική αλλαγή των βαρών ή δεν υπάρχει ουσιαστική μεταβολή στη συνάρτηση σφάλματος. Σε οποιαδήποτε άλλη περίπτωση ο αλγόριθμος επαναλαμβάνεται από το βήμα δ [19].

2.3.3 Ασαφής ομαδοποίηση K-μέσων

Στους δύο προηγούμενους αλγόριθμους το κάθε πρότυπο ταξινομείται σε μία μοναδική ομάδα, καθώς αποτελούν τεχνικές σκληρής ομαδοποίησης, με αποτελέσματα να δημιουργούνται προβλήματα όταν χαρακτηριστικά ομάδων αλληλεπικαλύπτονται, όπου η ταξινόμηση ενός προτύπου σε μόνο μία από αυτές οδηγεί σε μη ποιοτικά αποτελέσματα. Στην ασαφή ομαδοποίηση υπάρχει επικάλυψη των ομάδων νεφών προσφέροντας μεγαλύτερη ευελιξία, καθώς το πρότυπο ανήκει σε περισσότερες από μία ομάδες με αντίστοιχο συντελεστή συμμετοχής για κάθε μία από αυτές [20]. Ο βαθμός συμμετοχής $u_{j,i}$, προβάλλει το ποσοστό που ανήκει το πρότυπο $\overrightarrow{x_l}$ σε κάθε ομάδα j και λαμβάνει τιμές 0 έως 1. Αν M ο συνολικός αριθμός των ομάδων, τότε για κάθε πρότυπο ισχύει:

$$\sum_{j=1}^{M} u_{j,i} = 1 \tag{2.23}$$

Η μετάβαση από την ασαφή στη σκληρή ομαδοποίηση πραγματοποιείται μέσω της αντιστοίχισης του κάθε προτύπου σε εκείνη την ομάδα που έχει τον μεγαλύτερο βαθμό συμμετοχής. Τα βήματα για την υλοποίηση του αλγορίθμου είναι:

α) Κανονικοποίηση των τιμών εισόδου των στοιχείων των διανυσμάτων σε τιμές με περιορισμένο εύρος από 0.1 έως 0.9.

β) Καθορισμός του αριθμού των *M* κέντρων – ανταγωνιστικών νευρώνων και αρχικοποίηση όλων των βαρών στην ίδια τιμή μέσω της γραμμικής παρεμβολής για τις τιμές a και a+b, όπως αυτές παρουσιάστηκαν στη σχέση 2.18.

γ) Για κάθε εποχή πραγματοποιούνται τα ακόλουθα:

i) Για τα *N* διανύσματα $\overrightarrow{x_l}$ προσδιορίζονται οι βαθμοί συμμετοχής τους για κάθε κέντρο M μέσω της σχέσης:

$$u_{l,j}^{(t+1)} = \frac{1}{\sum_{k=1}^{M} \frac{d\left(\overrightarrow{x_l}, \overrightarrow{w_j}^{(t)}\right)}{d\left(\overrightarrow{x_l}, \overrightarrow{w_k}^{(t)}\right)}} \tag{2.24}$$

ii) Για κάθε ένα κέντρο της ομάδας j προσδιορίζεται νέο κέντρο βάρους, που δίνεται από τη σχέση 2.24:

$$\overrightarrow{w_j}^{(t+1)} = \frac{\sum_{l=1}^{N} (u_{l,j}^{(t+1)})^q \cdot \overrightarrow{x_l}}{\sum_{l=1}^{N} (u_{l,j}^{(t+1)})^q} \tag{2.25}$$

όπου η παράμετρος *q* λαμβάνει οποιαδήποτε τιμή μεγαλύτερη της μονάδας. Εάν q=2, τότε χρησιμοποιείται η ευκλείδεια απόσταση.

δ) Όταν ολοκληρωθεί η εποχή, αυξάνεται ο αριθμός των εποχών κατά ένα και ελέγχεται αν έχει συμπληρωθεί ο μέγιστος αριθμός των επιτρεπόμενων επαναλήψεων ή εάν δεν υπάρχει ουσιαστική αλλαγή των βαρών, δηλαδή ότι ισχύει $\left|\vec{w}_j^{(t)} - \vec{w}_j^{(t+1)}\right| < \varepsilon$. Αν δεν ικανοποιούνται οποιαδήποτε από τις δύο αυτές συνθήκες τότε επαναλαμβάνεται η διαδικασία από το βήμα γ.

2.3.4 Αυτό-οργανωμένος χάρτης – S.O.M.

Το τεχνητό νευρωνικό δίκτυο του αυτο-οργανωμένου χάρτη (self-organized map – S.O.M.) βασίζεται στην έννοια της ανταγωνιστικής μάθησης, όπου οι νευρώνες του ανταγωνιστικού επιπέδου προσπαθούν να διεκδικήσουν το εκάστοτε πρότυπο εισόδου $\overrightarrow{x_l}$. Οι νευρώνες επιλέγουν στα πλαίσια της ανταγωνιστικής μάθησης ορισμένα πρότυπα εισόδου με αντίστοιχη ρύθμιση των βαρών τους ώστε να διαταχθούν με θέση σε κόμβους του πλέγματος σχετική με συγκεκριμένα χαρακτηριστικά των προτύπων δημιουργώντας τοπογραφικό τους χάρτη. Ο χάρτης αυτός αποτελείται από μία ή δύο διαστάσεις, γεγονός που καθιστά βασικό πλεονέκτημα της συγκεκριμένης μεθόδου έναντι άλλων που κάνουν χρήση περισσότερων διαστάσεων.

Ο αυτο-οργανωμένος χάρτης προήλθε από τον τρόπο λειτουργίας του φλοιού του εγκεφάλου. Συγκεκριμένα, στηρίζεται στην αρχή ότι η χωρική θέση ενός νευρώνα εξόδου σε έναν τοπογραφικό χάρτη

αντιστοιχεί σε έναν τομέα ή σε ένα χαρακτηριστικό των δεδομένων εισόδου. Ο βασικός σκοπός του δικτύου είναι η απεικόνιση ενός προτύπου εισόδου διάστασης d , σ' έναν διακριτό χάρτη μίας ή δύο διαστάσεων. Κάθε νευρώνας του πλέγματος συνδέεται πλήρως με όλους τους κόμβους του επιπέδου εισόδου. Η διαδικασία του αλγορίθμου περιγράφεται από τις παρακάτω φάσεις:

1. Αρχικοποίησης, στην οποία τα βάρη του δικτύου που συνδέουν τους νευρώνες εισόδου με τους νευρώνες του χάρτη αρχικοποιούνται είτε με χρήση τυχαίων αριθμών, είτε με συγκεκριμένη διάταξη.
2. Ανταγωνισμού, όπου οι νευρώνες για κάθε πρότυπο εισόδου υπολογίζουν την αντίστοιχη τιμή της συνάρτησης ανταγωνισμού των νευρώνων, με νικητή αυτόν που φέρει τη μεγαλύτερη τιμή.
3. Συνεργασίας, στην οποία ο νικητής νευρώνας καθορίζει τη χωρική θέση μίας γειτονιάς νευρώνων, παρέχοντας τη βάση για τη συνεργασία μεταξύ των γειτονικών νευρώνων.
4. Προσαρμογής των βαρών, όπου οι τιμές των βαρών των νευρώνων που ανήκουν στη νικητήρια γειτονιά να προσαρμόσουν τις τιμές των βαρών του νικητή, με στόχο την ενίσχυση της απόκρισης σε επόμενη εφαρμογή ενός παρομοίου προτύπου εκπαίδευσης.

Ο αλγόριθμος υλοποιείται στα εξής βήματα:

α) Κανονικοποίηση των τιμών εισόδου των στοιχείων των διανυσμάτων, αντιστοιχώντας τα σε πεδίο τιμών (0,1).

β) Καθορισμός του πλέγματος και του αριθμού των M κέντρων – ανταγωνιστικών νευρώνων και αρχικοποίησή τους. Κάθε νευρώνας τοποθετείται σε ένα πλέγμα μίας ή δύο διαστάσεων προσδιορίζοντας το πλήθος M των νευρώνων.

γ) Για κάθε πρότυπο εισόδου το οποίο παρουσιάζεται στο δίκτυο είτε σειριακά, είτε τυχαία εκτελούνται οι ακόλουθες διαδικασίες:

i) Εύρεση του νευρώνα νικητή με τη μικρότερη ευκλείδεια απόσταση από το πρότυπο εισόδου. Η θέση και το διάνυσμα βαρών του συγκεκριμένου νευρώνα αποτελούν την απόκριση του νευρωνικού δικτύου.

ii) Συνεργασία των νευρώνων, όπου ο νικητής προσδιορίζει το κέντρο της γειτονιάς, δηλαδή μιας περιοχής του πλέγματος που έχει συμμετρική απόσταση γύρω του, στην οποία οι νευρώνες θα συνεργαστούν επηρεασμένοι αντιστρόφως ανάλογα με την απόσταση από το κέντρο. Το πλάτος της γειτονιάς είναι ανεξάρτητο από τη θέση του νευρώνα νικητή και επηρεάζει τον βαθμό συμμετοχής των κοντινών νευρώνων .

iii) Αναπροσαρμογή των βαρών όλων των νευρώνων, όπου μετακινείται το διάνυσμα των μερών του νικητή και της γειτονιάς του προς την κατεύθυνση του διανύσματος εισόδου $\vec{x_l}$, λόγω της διαδικασίας προσαρμογής των γειτονικών νευρώνων, ενώ ταυτόχρονα μειώνεται ο χρόνος εκμάθησης.

iv) Έλεγχος εάν οι επαναλήψεις έχουν φτάσει τον μέγιστο αριθμό ή εάν δεν είναι ουσιώδης η περαιτέρω διαμόρφωση των βαρών. Σε οποιαδήποτε άλλη περίπτωση επαναλαμβάνεται το προηγούμενο βήμα.

Η διαδικασία προσαρμογής των βαρών χωρίζεται στη φάση αυτο-οργάνωσης και στη φάση σύγκλισης. Κατά την πρώτη διατάσσονται τοπολογικά τα διανύσματα βαρών έχοντας προσδιορίσει τις αρχικές τιμές των παραμέτρων του ρυθμού μάθησης και του πλάτους που θα έχει η "γειτονιά", δηλαδή ο αλγόριθμος εκτελείται για πληθώρα τιμών του ρυθμού μάθησης η_o. Στη φάση της σύγκλισης τελειοποιείται ο χάρτης χαρακτηριστικών, παρέχοντας μία ακριβή στατιστική ανάλυση των προτύπων με αριθμό επαναλήψεων πολλαπλάσιο του αριθμού νευρώνων [21].

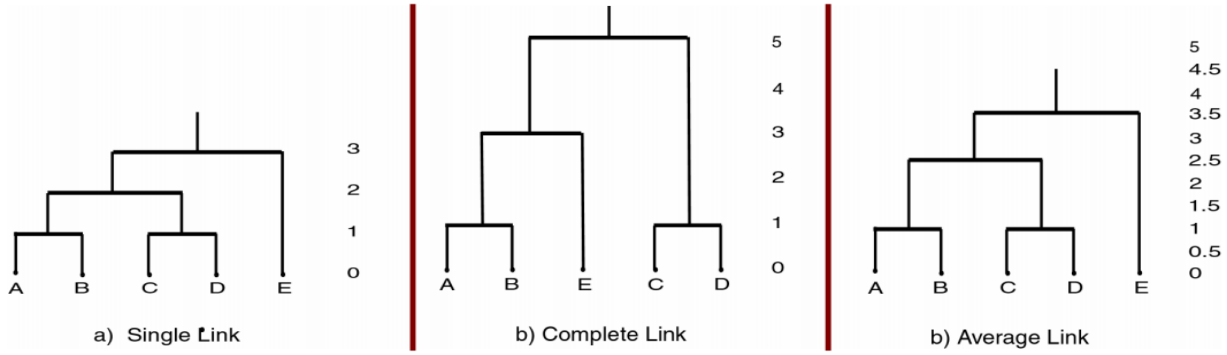
2.3.5 Ιεραρχικοί Αλγόριθμοι Συγχώνευσης.

Οι ιεραρχικοί αλγόριθμοι αναπτύσσουν μια ιεραρχία ομάδων μεταξύ τους, χωρίς να παράγουν μια συγκεκριμένη ομαδοποίηση. Οι συσσωρευτικοί αλγόριθμοι (agglomerative) είναι ιεραρχικοί αλγόριθμοι με προσέγγιση bottom – up. Ξεκινούν με κάθε στοιχείο να τοποθετείται σε μία ξεχωριστή ομάδα. Οι ομάδες που προκύπτουν συγχωνεύονται ανά επανάληψη έως ότου κριθεί ικανοποιητική η ομαδοποίηση. Οι ιεραρχικές τεχνικές συσταδοποίησης παρουσιάζονται με τη χρήση ενός δενδρογράμματος, όπου κάθε επίπεδο δείχνει τις συστάδες που υπάρχουν και κάθε συστάδα είναι ένωση αυτών του προηγούμενου επιπέδου. Τα βήματα του αλγορίθμου είναι:

1. Αρχικοποίηση, όπου το σύνολο των διανυσμάτων για το μηδενικό επίπεδο τίθεται ίσο με το σύνολο των διανυσμάτων εισόδου.
2. Αύξηση του επιπέδου κατά 1 και εύρεση και συγχώνευση των ομάδων που ικανοποιούν το κριτήριο της ελάχιστης μεταξύ τους απόστασης.
3. Σχηματισμός του νέου πίνακα-επιπέδου
4. Λήξη αλγορίθμου όταν όλα τα διανύσματα εισόδου είναι στο ίδιο σύνολο.

Ο αλγόριθμος εκτελεί επαναλήψεις ίσες με το πλήθος των προτύπων εισόδου μειωμένο κατά 1. Οι κυριότεροι ιεραρχικοί αλγόριθμοι συγχώνευσης είναι:

- Τεχνική απλού συνδέσμου (simple link), όπου ορίζει ως αποστάσεις μεταξύ των ομάδων την ελάχιστη μεταξύ των δύο πρώτων ομάδων που συγχωνεύονται.
- Τεχνική πλήρους συνδέσμου (complete link), όπου ορίζει ως αποστάσεις μεταξύ των ομάδων τη μέγιστη μεταξύ των δύο πρώτων ομάδων που συγχωνεύονται.
- Τεχνική μέσου συνδέσμου (average link), όπου ορίζει ως αποστάσεις μεταξύ των ομάδων τη μέση μεταξύ των δύο πρώτων ομάδων που συγχωνεύονται.
- Τεχνική σταθμισμένου μέσου όρου ομάδας ζευγαριού (weighted pair group method average), όπου ορίζει ως απόσταση της ομάδας C_s από τη νέα ομάδα C_q τον μέσο όρο των αποστάσεων της C_q από τις δύο αρχικές ομάδες που συγχωνεύονται (C_i, C_j).
- Τεχνική μη σταθμισμένου μέσου όρου ομάδας ζευγαριού (unweighted pair group method average), ο οποίος ορίζει ως απόσταση της ομάδας C_s από τη νέα ομάδα C_q το άθροισμα των αποστάσεων της C_q από τις δύο αρχικές ομάδες C_i, C_j που συγχωνεύονται με χρήση βαρών ανάλογων του πληθυσμού της κάθε ομάδας.
- Τεχνική Ward – ελάχιστης διασποράς, όπου η απόσταση είναι η σταθμισμένη τιμή των τετραγώνων των ευκλείδειων αποστάσεων των μέσων διανυσμάτων.
- Επίσης, υπάρχουν οι αλγόριθμοι μη σταθμισμένης κεντροειδούς μορφής ομάδας ζευγαριού (unweighted pair group method centroid) και σταθμισμένης κεντροειδούς μορφής ομάδας ζευγαριού (weighted pair group method centroid).

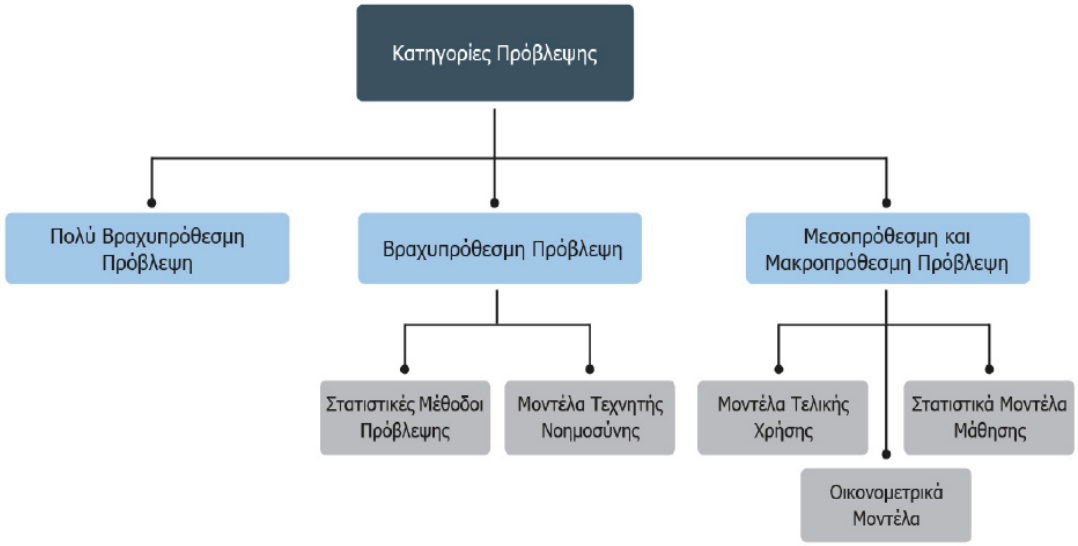


Διάγραμμα 2.7: Τεχνικές υλοποίησης ιεραρχικών αλγορίθμων [22].

Κεφάλαιο 3
Τεχνικές Πρόβλεψης φορτίου

3.1 Εισαγωγή

Η πρόβλεψη φορτίου συνίσταται στον προσδιορισμό της ζήτησης φορτίου για ένα επόμενο χρονικό διάστημα. Το χρονικό αυτό πλαίσιο μπορεί να είναι από 24 ώρες έως και 7 ημέρες, για ένα διάστημα 3 έως 10 ετών ή ακόμα μεγαλύτερο χαρακτηρίζοντας την πρόβλεψη ως βραχυπρόθεσμη, μεσοπρόθεσμη και μακροπρόθεσμη αντίστοιχα. Η πρώτη περίπτωση, η οποία θα αναλυθεί και θα υλοποιηθεί στην παρούσα διπλωματική, χρησιμοποιείται για την εύρυθμη και ασφαλή λειτουργία του συστήματος, ενώ η δεύτερη και η τρίτη αξιοποιούνται για τον σχεδιασμό, την κατασκευή και τη λήψη αποφάσεων που αφορούν το μέλλον του συστήματος ενέργειας. Η απελευθέρωση της αγοράς έχει αυξήσει και αυτή την ανάγκη για πιο ακριβείς προβλέψεις, καθώς η γνώση του ποσού ενέργειας που χρειάζεται να διαθέσει ο κάθε συμμετέχων σε αυτή είναι υψίστης σημασίας. Με αυτόν τον τρόπο αποφεύγονται υπερεκτιμήσεις που οδηγούν σε σπατάλη πόρων, αλλά και υποεκτιμήσεις που θα αυξάνουν το λειτουργικό κόστος του κάθε παραγωγού-προμηθευτή με ταυτόχρονο κίνδυνο της μείωσης της αξιοπιστίας του συστήματος. Οι κατηγορίες πρόβλεψης και οι βασικότερες μέθοδοι υλοποίησής τους παρουσιάζονται στο Διάγραμμα 3.1 που ακολουθεί.



Διάγραμμα 3.1: Βασικές κατηγορίες και μεθοδολογίες πρόβλεψης.

Η βραχυπρόθεσμη πρόβλεψη φορτίου αφορά στην προσέγγιση της ζήτησης ενέργειας για το επόμενο χρονικό διάστημα που αφορά από μία ώρα έως και μία εβδομάδα, βοηθώντας στην εκτίμηση των ροών φορτίου και στην αποφυγή υπερφορτίσεων. Η έγκαιρη λήψη αποφάσεων που διευκολύνονται με τη χρήση της συντελούν στη βελτίωση της αξιοπιστίας και στη μείωση βλαβών και διακοπών ρεύματος. Επίσης, η πρόβλεψη είναι σημαντική για την αξιολόγηση των συμβάσεων μεταξύ αγοράς και πελάτη κατά την τιμολόγηση της

ενέργειας που καταναλώνει. Τα βασικότερα στοιχεία εισόδου στη συγκεκριμένη πρόβλεψη είναι τα ωριαία φορτία των προηγούμενων ωρών ή ημερών, το είδος της ημέρας, καθώς το φορτίο διαφοροποιείται σημαντικά μεταξύ τους και των καιρικών συνθηκών λόγω κλίματος, αλλά και εποχής. Συγκεκριμένα, για την Ελλάδα θεωρείται πως η θερμοκρασία παρέχει όλες τις πληροφορίες που χρειάζονται για την επίδραση του καιρού στη ζήτηση φορτίου. [11]

Η υλοποίηση της πρόβλεψης βασίζεται στη συλλογή και επεξεργασία των δεδομένων σε μορφή χρονοσειρών, δηλαδή σειρές διαδοχικών παρατηρήσεων που περιγράφουν την εξέλιξη ενός μεγέθους στον χρόνο με μετέπειτα αναγνώριση των βασικών τους συνιστωσών (τάση, περιοδικότητα, εποχικότητα, ασυνέχειες). Οι παρατηρήσεις αυτές λαμβάνονται σε ορισμένες χρονικές στιγμές ή περιόδους που ισαπέχουν μεταξύ τους και για την ανάλυσή τους χρησιμοποιούνται δύο μοντέλα, τα ντετερμινιστικά και τα στοχαστικά. Τα πρώτα περιγράφουν την εξέλιξη ενός μεγέθους θεωρώντας πλήρη επίγνωση των παραγόντων που το επηρεάζουν, ενώ τα στοχαστικά λαμβάνουν υπόψη τους την επίδραση τυχαίου παράγοντα στην εξέλιξη του. Οι κατηγορίες μεθοδολογιών για τη βραχυπρόθεσμη πρόβλεψη είναι οι κλασικές - στατιστικές μέθοδοι και εκείνες που χρησιμοποιούν τεχνητή νοημοσύνη, οι οποίες θα μελετηθούν παρακάτω.

3.2 Μεθοδολογίες βραχυπρόθεσμης πρόβλεψης

Η βραχυπρόθεσμη πρόβλεψη φορτίου αποτελεί σημαντική διαδικασία για τα συστήματα ηλεκτρικής ενέργειας και για την υλοποίησή της έχουν αναπτυχθεί δύο βασικές κατηγορίες μεθοδολογιών, τα κλασικά-στατιστικά μοντέλα και τα μοντέλα τεχνητής νοημοσύνης που θα αναλυθούν παρακάτω.

3.2.1 Κλασικές Μεθοδολογίες

Οι βασικές κλασικές μεθοδολογίες που χρησιμοποιούνται για την πραγματοποίηση της βραχυπρόθεσμης πρόβλεψης φορτίου είναι:

- Απλοϊκή μέθοδος (Naive), που είναι η απλούστερη στατιστική μέθοδος και συνήθως παράγει μεγέθη που χρησιμοποιούνται ως αναφορά για σύγκριση με άλλα μοντέλα, διότι δεν εξάγει ακριβείς προβλέψεις. Με βάση αυτήν, η προβλεπόμενη τιμή του εξεταζόμενου μεγέθους είναι ίση με την παρατήρηση της ακριβώς προηγούμενης χρονικής περιόδου.
- Μέθοδοι Εξομάλυνσης (Smoothing), που στηρίζονται στις προηγούμενες παρατηρήσεις της χρονοσειράς ώστε να προσδιορίσουν μια εξομαλυμένη τιμή των δεδομένων, την οποία όταν την προεκτείνουν προκύπτει η πρόβλεψη. Τέτοιες είναι οι μέθοδοι κινητών μέσων όρων, των γενικών γραμμικών κινητών μέσω όρων και της εκθετικής εξομάλυνσης.
- Μέθοδοι αποσύνθεσης (Decomposition), που ανήκουν στο μοντέλο των χρονοσειρών και σκοπεύουν στον διαχωρισμό των βασικών συνιστωσών με τη μεγαλύτερη ορθότητα. Έτσι, απομονώνονται διαδοχικά τα χαρακτηριστικά εποχικότητας, τάσης και κυκλικότητας, ώστε να αναγνωριστεί η τυχαιότητα των παρατηρήσεων. Τέτοιες μέθοδοι είναι το προσθετικό και το πολλαπλασιαστικό μοντέλο, η σταθερή πολλαπλασιαστική και η μέθοδος Census.
- Παλινδρομικών μοντέλων (Regression Models), στα οποία γίνεται εύρεση συσχετίσεων μεταξύ του εξεταζόμενου μεγέθους και των μεταβλητών που το επηρεάζουν μέσω της εξίσωσης παλινδρόμησης. Ανάλογα με τον αριθμό αυτών των μεταβλητών, τα μοντέλα κατηγοριοποιούνται σε απλής ή πολλαπλής γραμμικής παλινδρόμησης. Στην πρόβλεψη ηλεκτρικής ενέργειας αυτά τα μοντέλα υποθέτουν πως το φορτίο διαχωρίζεται σε μια βασική τάση και μία γραμμική που εξαρτάται από τους παράγοντες που την επηρεάζουν (καιρός, τύπος μέρας, προφίλ πελάτη), βοηθώντας στην κατανόηση μεταξύ των σχέσεων εισόδου και εξόδου.
- Αυτοπαλινδρομικά μοντέλα (Autoregressive Models), στα οποία γίνεται η υπόθεση πως το φορτίο είναι γραμμικός συνδυασμός των προηγούμενων τιμών του, όπου για την εύρεση των άγνωστων συντελεστών της εξίσωσης χρησιμοποιείται η μέθοδος ελαχίστων τετραγώνων. Παραδείγματα αποτελούν τα ολοκληρωμένα και μη αυτοπαλινδρομικά μοντέλα κινητού μέσου όρου ARMA και ARIMA αντίστοιχα. Τα μοντέλα ARIMA είναι λόγω της πολύ καλής συμπεριφοράς τους, κύριο μέτρο σύγκρισης ως προς τα μοντέλα νευρωνικών δικτύων.
- Οικονομετρική μέθοδος, στην οποία εντάσσονται οι διαδικασίες που βασίζονται σε αιτιοκρατικές σχέσεις ανάμεσα στο μέγεθος προς πρόβλεψη και στις παραμέτρους που το επηρεάζουν. Τα οικονομετρικά μοντέλα αναπτύσσονται με τρόπο που περιγράφουν ένα συγκεκριμένο πρόβλημα, καθιστώντας τα μη ιδανικά για χρήση σε διαφορετικές περιπτώσεις.

Υπάρχουν και άλλες κλασικές μέθοδοι, όπως οι στοχαστικές, στις οποίες υπάρχει η παραδοχή ότι τα δεδομένα έχουν εσωτερική δομή, η προσέγγιση παρόμοιας ημέρας που βασίζεται στην αναζήτηση ιστορικών δεδομένων, η μέθοδος επαναληπτικών επανασταθμισμένων ελαχίστων τετραγώνων όπου χρησιμοποιούνται

συναρτήσεις αυτοσυσχέτισης στα ιστορικά δεδομένα. Επίσης, σε όλες τις προηγούμενες μεθόδους μπορούν να χρησιμοποιηθούν τεχνικές βελτιστοποίησης τύπου Hilbert και εκτιμητές τύπου Kernel [19].

3.2.2 Μεθοδολογίες Τεχνητής Νοημοσύνης

Τα μοντέλα τεχνητής νοημοσύνης είναι μαθηματικά – υπολογιστικά και βασισμένα στη δομή και τις αρχές λειτουργίας των βιολογικών νευρώνων. Η εφαρμογή τους δικαιολογείται από τη μορφή του προβλήματος της βραχυπρόθεσμης πρόβλεψης, το οποίο δεν καθορίζεται από αυστηρές μαθηματικές σχέσεις που προσδιορίζουν το φορτίο, καθώς οι παράμετροι δεν είναι όλες γνωστές. Οι κυριότερες μεθοδολογίες για την υλοποίησή τους είναι:

- Τα τεχνητά νευρωνικά δίκτυα (Artificial Neural Networks), που θα αναλυθούν παρακάτω, με τα οποία θα υλοποιηθεί το μοντέλο πρόβλεψης βασίζονται στη δομή και στις λειτουργίες των βιολογικών νευρώνων. Αυτά αποτελούνται από διασυνδεδεμένες ομάδες τεχνητών νευρώνων που επεξεργάζονται τις πληροφορίες μέσω συνδετικής προσέγγισης υπολογισμού.
- Οι μηχανές διανυσμάτων υποστήριξης (Support Vector Machines), που βασίζονται στη στατιστική θεωρία μάθησης χρησιμοποιώντας μη γραμμική χαρτογράφηση του χώρου εισόδου σε ένα πολυδιάστατο χώρο χαρακτηριστικών στον οποίο κατασκευάζεται ένα βέλτιστο υπερεπίπεδο. Στα διανύσματα υποστήριξης είναι σημαντική η επιλογή του κατάλληλου τύπου μηχανής μάθησης , όπως αντίστοιχα επιλογή αρχιτεκτονικής στα νευρωνικά δίκτυα. Έχουν ως στόχο την ελαχιστοποίηση του σφάλματος γενίκευσης και όχι την εκπαίδευση, καθιστώντας τα ισοδύναμα με την επίλυση ενός γραμμικά περιορισμένου προβλήματος δυαδικού προγραμματισμού [23].
- Η ασαφής λογική (Fuzzy Logic), που ξεφεύγει από τη Boolean λογική του ‘0’ ή ‘1’ και κατά την οποία μια μεταβλητή είναι συνδεδεμένη με ένα εύρος τιμών, όπως και ο ανθρώπινος τρόπος σκέψης δεν καθορίζεται από μόνο *ναι ή όχι*. Αποτελεί μέθοδο χαρτογράφησης ασαφών εισόδων σε ασαφείς εξόδους με βασικό πλεονέκτημα την απουσία ανάγκης προσδιορισμού ακριβών τιμών εισόδου ή μαθηματικού μοντέλου, γεγονός που αποτελεί και την αδυναμία τους για προβλήματα, όπου αυτό κρίνεται αναγκαίο.
- Τα δένδρα ταξινόμησης και παλινδρόμησης (Classification and Regression Trees), που βασίζονται σε μια ιεραρχική διαίρεση με τη μορφή δένδρου για τα διανύσματα εισόδου. Το δένδρο διευκρινίζει τη σχέση μεταξύ εισόδου και εξόδου σύμφωνα με τις συνθήκες διάσπασης, εντοπίζοντας κανόνες ενσωματωμένους στα δεδομένα και επιτρέποντας την απεικόνιση του προβλήματος. Η χρήση του αφορά συνεχή μεταβλητή εξόδου και η διαδικασία αφορά την κατάτμηση του χώρου εισόδου μέσα από εσωτερικούς κόμβους απόβασης σε τερματικά φύλλα, τα οποία ουσιαστικά αποτελούν την εξαγόμενη πρόβλεψη [25].

Άλλες μέθοδοι για την κατασκευή μοντέλων πρόβλεψης είναι τα έμπειρα συστήματα, που βασίζονται στη γνώση ενσωματώνοντας κανόνες και διαδικασίες από εμπειρογνώμονες, η παλινδρόμηση K κοντινότερων γειτόνων, όπου ως πρόβλεψη λαμβάνεται το αποτέλεσμα της διαδικασίας του μέσου όρου μεταξύ των γειτονικών σημείων και τα αυτοπαλινδρομικά μοντέλα με εξωγενή μεταβλητή (ARMAX) [26].

3.3 Τεχνητά Νευρωνικά Δίκτυα

Ένα τεχνητό νευρικό δίκτυο (Artificial Neural Networks - ANN) είναι το κομμάτι ενός υπολογιστικού συστήματος που έχει σχεδιαστεί για να προσομοιώνει τον τρόπο με τον οποίο ο ανθρώπινος εγκέφαλος μέσω των νευρώνων του αναλύει και επεξεργάζεται πληροφορίες. Αντίστοιχα, τα συγκεκριμένα δίκτυα είναι σε θέση να επικοινωνούν στέλνοντας σήματα μεταξύ ενός μεγάλου αριθμού συνδέσεων - τεχνητών νευρώνων που δέχονται πληροφορίες στους κόμβους εισόδου και συγκεντρώνει δεδομένα. Η συνάρτηση μεταφοράς του περιγράφει τον τρόπο με τον οποίο το σταθμισμένο άθροισμα των δεδομένων εισόδου καταλήγει ως μία έξοδος.

Τα νευρωνικά δίκτυα είναι το θεμέλιο της τεχνητής νοημοσύνης και επιλύει προβλήματα που αποδεικνύονται αδύνατα ή δύσκολα από τα ανθρώπινα ή στατιστικά πρότυπα. Τα ANN έχουν αναπτυχθεί ιδιαίτερα τα τελευταία χρόνια με εφαρμογές σε όλους τους κλάδους τόσο των επιχειρήσεων και της βιομηχανίας, όσο και στην καθημερινότητα μέσω της ανάπτυξης της τεχνολογίας. Η χρήση τους προσφέρει πολλά οφέλη στην κατηγοριοποίηση και στην πρόβλεψη. Συνεχώς, κερδίζουν έδαφος ως κύρια εργαλεία για αυτές. Χαρακτηριστικό τους είναι ότι δεν προγραμματίζονται, αλλά μαθαίνουν από κάποιον επιλεγμένο αλγόριθμο εκπαίδευσης και με βελτίωση των αποτελεσμάτων ανάλογη με τον αριθμό δεδομένων που τους παρέχονται. Στην αντιμετώπιση ενός προβλήματος, ένα σημαντικό πλεονέκτημα των νευρωνικών δικτύων έγκειται στη δυνατότητά τους να δουν και να επεξεργαστούν πολυδιάστατα πρότυπα εισόδου, σε αντίθεση με τον άνθρωπο ο οποίος έχει τρισδιάστατη άποψη του χώρου.

Η ικανότητα που παρουσιάζουν τα νευρωνικά δίκτυα στον εντοπισμό των ομοιοτήτων μεταξύ δεδομένων προσφέρει αυξημένες δυνατότητες στην επίλυση προβλημάτων ταξινόμησης. Ωστόσο, αδυνατούν να δώσουν ολοκληρωμένη ερμηνεία στα συμπεράσματα που εξάγουν και να εξηγήσουν την πορεία που τα οδήγησε προς αυτά. Τα πολυεπίπεδα νευρωνικά δίκτυα είναι σε θέση να δώσουν τη βέλτιστη λύση ακόμα και σε ένα αυθαίρετο πρόβλημα κατηγοριοποίησης, υλοποιώντας γραμμικούς διαχωρισμούς, με τα πρότυπα να είναι διατεταγμένα στο χώρο κατά μη γραμμικό τρόπο. Η ικανότητα αυτή των δικτύων, οφείλεται στο γεγονός πως χρησιμοποιούν αρκετά απλούς αλγόριθμους, εκεί που η μορφή της μη γραμμικότητας μπορεί να εξαχθεί από τα σύνολα της εκπαίδευσης.

Η εποικοδομητική χρήση ενός νευρωνικού δικτύου συνίσταται στη δημιουργία κατάλληλων αλγόριθμων με διαμορφωμένα πλαίσια εκπαίδευσης, ώστε να το δίκτυο αντιμετωπίζει αποτελεσματικά την τρέχουσα μορφή του εκάστοτε προβλήματος. Πρέπει να ληφθεί υπόψη ότι τα νευρωνικά δίκτυα είναι κατάλληλα για παρεμβολή (interpolation) και όχι για εξαγωγή συμπερασμάτων (extrapolation). Αυτό σημαίνει ότι τα σύνολα εκπαίδευσης πρέπει να αποτελούνται από τον πλήρη χώρο προτύπων, τα οποία χρειάζεται να ταξινομούνται (αναγνωρίζονται) σωστά κατά τη διάρκεια της εκμετάλλευσης του δικτύου που χρησιμοποιείται για την εξαγωγή του αποτελέσματος. Ακόμη, υπάρχει το ζήτημα του χρόνου εκπαίδευσης, ιδιαίτερα σε περιπτώσεις εφαρμογών πραγματικού χρόνου, όπως είναι η πρόβλεψη φορτίου. Ο χρόνος εκπαίδευσης είναι ανάλογος με την πολυπλοκότητα της δομής του δικτύου και του μεγάλου αριθμού των προτύπων του συνόλου μάθησης.

Η χρήση των νευρωνικών δικτύων εμπεριέχει το μειονέκτημα της κανονικοποίησης που επιλέγει ή ρυθμίζει την πολυπλοκότητα του δικτύου, καθώς ο αριθμός των εισόδων και εξόδων είναι γνωστός από τον χώρο των προτύπων και του αριθμού των ομάδων, ο συνολικός αριθμός των βαρών και των παραμέτρων είναι άγνωστος. Η χρήση πολλών ελευθέρων παραμέτρων οδηγεί σε απώλεια γενίκευσης, ενώ στην αντίθετη περίπτωση προκαλείται ανεπαρκής μάθηση από τα σύνολα εκπαίδευσης. Συνεπώς, η επιλογή της τοπολογίας του δικτύου εξαρτάται άμεσα από η φύση του προβλήματος που αντιμετωπίζεται.

Η μάθηση των νευρωνικών δικτύων μπορεί να διακριθεί σε επιβλεπόμενη και μη επιβλεπόμενη μάθηση, όπου στη δεύτερη περίπτωση δεν υπάρχει διαθέσιμος ένας εξωτερικός «δάσκαλος» ή οδηγίες (ενισχυτικό σήμα) από το περιβάλλον, καθώς για την εκπαίδευση των νευρωνικών δικτύων μπορούν να χρησιμοποιηθούν

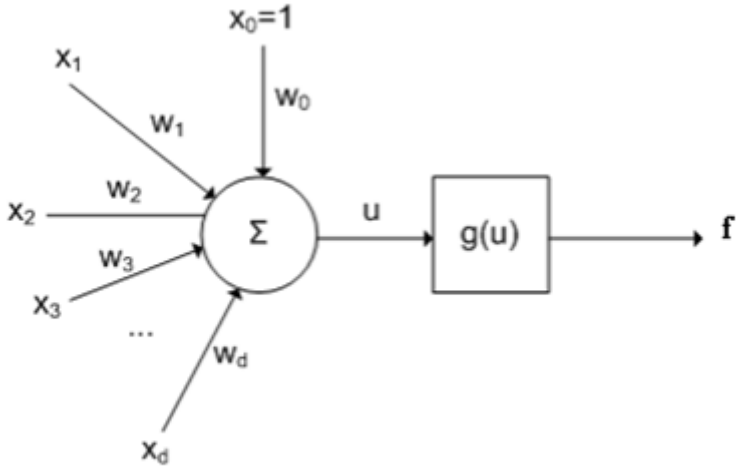
μόνο τα διανύσματα (πρότυπα) εισόδου. Ένα σύστημα μη επιβλεπόμενης μάθησης λειτουργεί εξάγοντας ιδιότητες ή χαρακτηριστικά των προτύπων τα οποία του παρουσιάζονται μη γνωρίζοντας τις κατηγορίες και τον αριθμό τους. Έτσι, σκοπός της μάθησης είναι η αυτο-οργάνωση και η ανακάλυψη διαφόρων χαρακτηριστικών ιδιοτήτων των δεδομένων εισόδου [17].

Ο τεχνητός νευρώνας, όπως αυτός παρουσιάζεται στο Διάγραμμα 3.2 είναι η βασική μονάδα επεξεργασίας του ΤΝΔ ο οποίος αποτελείται από d συνδέσεις εισόδου x_i και χαρακτηρίζεται από μία τιμή βάρους w_i [27]. Ο υπολογισμός που συντελείται σε αυτόν διακρίνεται σε:

1. Υπολογισμός συνολικής εισόδου:

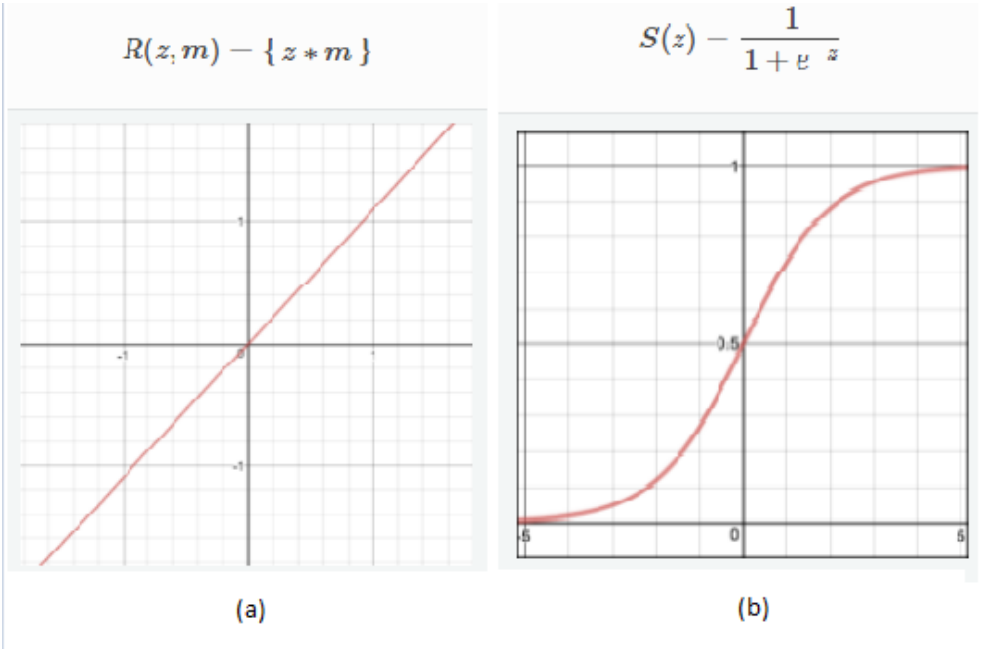
$$u(x) = \sum_{i=1}^{d} w_i x_i + w_o \tag{3.1}$$

2. Υπολογισμός εξόδου του νευρώνα έπειτα από την αλληλεπίδραση της συνολικής εισόδου u(x) με μία συνάρτηση ενεργοποίησης g: $f = g(u)$

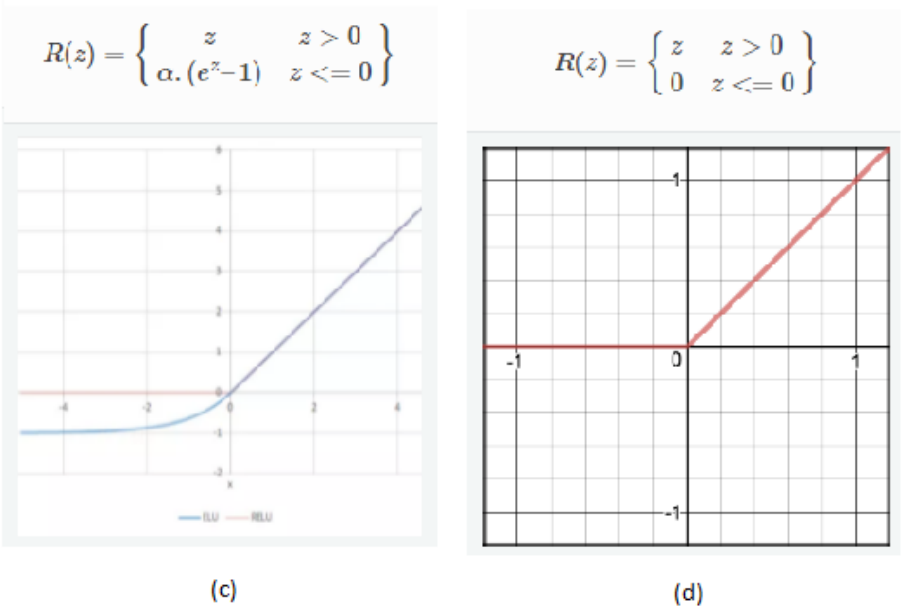


Διάγραμμα 3.2: Μοντέλο τεχνητού νευρώνα [27].

Η βηματική συνάρτηση θεωρείται ότι αποτελεί τη συνάρτηση ενεργοποίησης του βιολογικού νευρώνα, ενώ στα τεχνητά νευρωνικά δίκτυα συνήθως οι συναρτήσεις ενεργοποίησης έχουν μορφή παρόμοια αυτής. Οι συνηθέστεροι τύποι συναρτήσεων ενεργοποίησης είναι η γραμμική (Linear), η σιγμοειδής (Sigmoid), η εκθετική γραμμική (ELU) και η ανορθωμένη γραμμική (ReLU), όπου τα γραφήματά τους παρατίθενται στα Διαγράμματα 3.3 και 3.4 [28].



Διάγραμμα 3.3: Συναρτήσεις ενεργοποίησης ΤΝΔ: α) Γραμμική και β) Σιγμοειδής



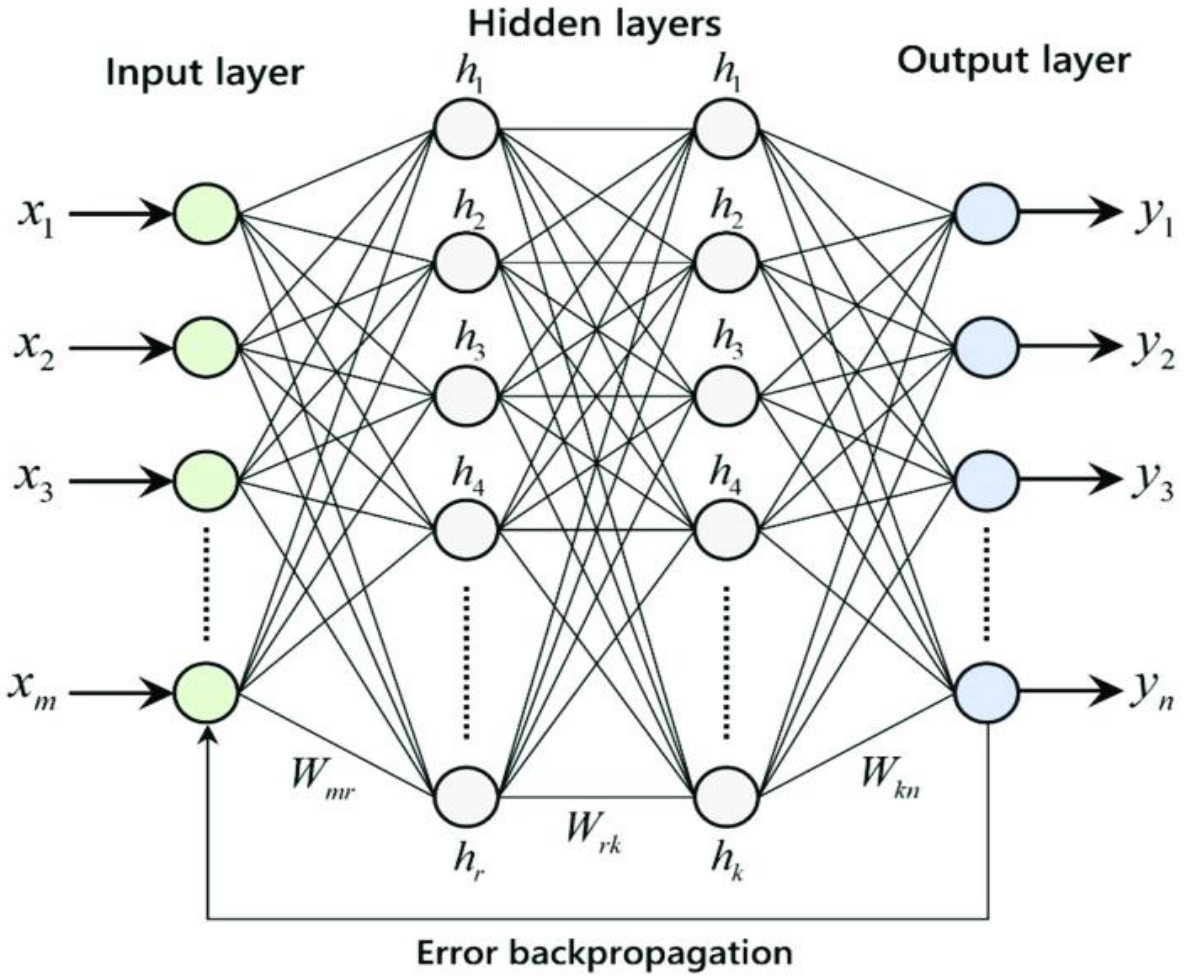
Διάγραμμα 3.4: Συναρτήσεις ενεργοποίησης ΤΝΔ: α) ELU και β) ReLU

Τα νευρωνικά δίκτυα χαρακτηρίζονται από την αρχιτεκτονική τους, τη λειτουργία που επιτελούν και τη μέθοδο εκπαίδευσής τους. Η αρχιτεκτονική του δικτύου καθορίζει τη διάταξη των συνδέσεων, τον αριθμό και τον τύπο των νευρώνων και ανάλογα με αυτή τα δίκτυα κατηγοριοποιούνται σε [29]:

- Δίκτυα πρόσθιας τροφοδότησης ενός επιπέδου (Feed Forward), που το επίπεδο εισόδου δείχνει το επόμενο επίπεδο που αποτελεί το επίπεδο εξόδου.
- Πολυεπίπεδα δίκτυα πρόσθιας τροφοδότησης (Multi-Layer Perceptron-M.L.P), που αποτελούν την πιο συνηθισμένη μορφή ΤΝΔ που εφαρμόζεται. Η δομή τους συνίσταται από ένα στρώμα εισόδου (input), ένα ή περισσότερα κρυμμένα στρώματα (hidden) και από ένα στρώμα εξόδου (output). Το κάθε στρώμα αποτελείται από έναν αριθμό νευρώνων (neurons) με τον καθένα να συνδέεται μόνο με τους αντίστοιχους του επόμενου στρώματος μέσω διαφορετικών ελεύθερων παραμέτρων που καλούνται βάρη (weights).

- Αναδρομικά δίκτυα (Recurrent) στα οποία υπάρχει τουλάχιστον ένας βρόχος ανάδρασης, δηλαδή η έξοδος του νευρώνα ανατροφοδοτεί την είσοδο των άλλων νευρώνων του ίδιου επιπέδου.

Ο αριθμός των επιπέδων και των νευρώνων ανά επίπεδο, δηλαδή η δομή του νευρωνικού δικτύου είναι εξέχουσας σημασίας για την άρτια λειτουργία του και ποικίλει ανάλογα τη φύση του προβλήματος για το οποίο εφαρμόζεται. Η χρήση υπερβολικού αριθμού νευρώνων στα κρυφά στρώματα ενός δικτύου εμπρόσθιας τροφοδότησης πολλών επιπέδων, ενδέχεται να οδηγήσει σε απομνημόνευση των συνόλων εκπαίδευσης με ορατό τον κίνδυνο απώλειας της ικανότητας γενίκευσης, ενώ αντίθετα για ένα πολύπλοκο και σύνθετο πρόβλημα, η έλλειψη ικανού αριθμού νευρώνων είναι πιθανό να εμποδίσει την ικανοποιητική ταξινόμηση των προτύπων. Οι νευρώνες λειτουργούν παράλληλα, ταυτόχρονα και συνήθως συναντώνται σε μεγάλους αριθμούς, καθιστώντας τα νευρωνικά δίκτυα χαρακτηριστικό παράδειγμα μαζικά παράλληλου υπολογισμού. Στο διάγραμμα 3.5 παρουσιάζεται μια τυπική δομή ενός ανατροφοδοτούμενου πολυεπίπεδου νευρωνικού [30], στο οποίο φαίνεται ότι η έξοδος του κάθε κρυφού νευρώνα μεταδίδεται εκ νέου σε όλους τους νευρώνες του επόμενου στρώματος, οι οποίοι τις επεξεργάζονται, παράγοντας νέες εξόδους.



Διάγραμμα 3.5: Δομή ανατροφοδοτούμενου πολυεπίπεδου νευρωνικού δικτύου.

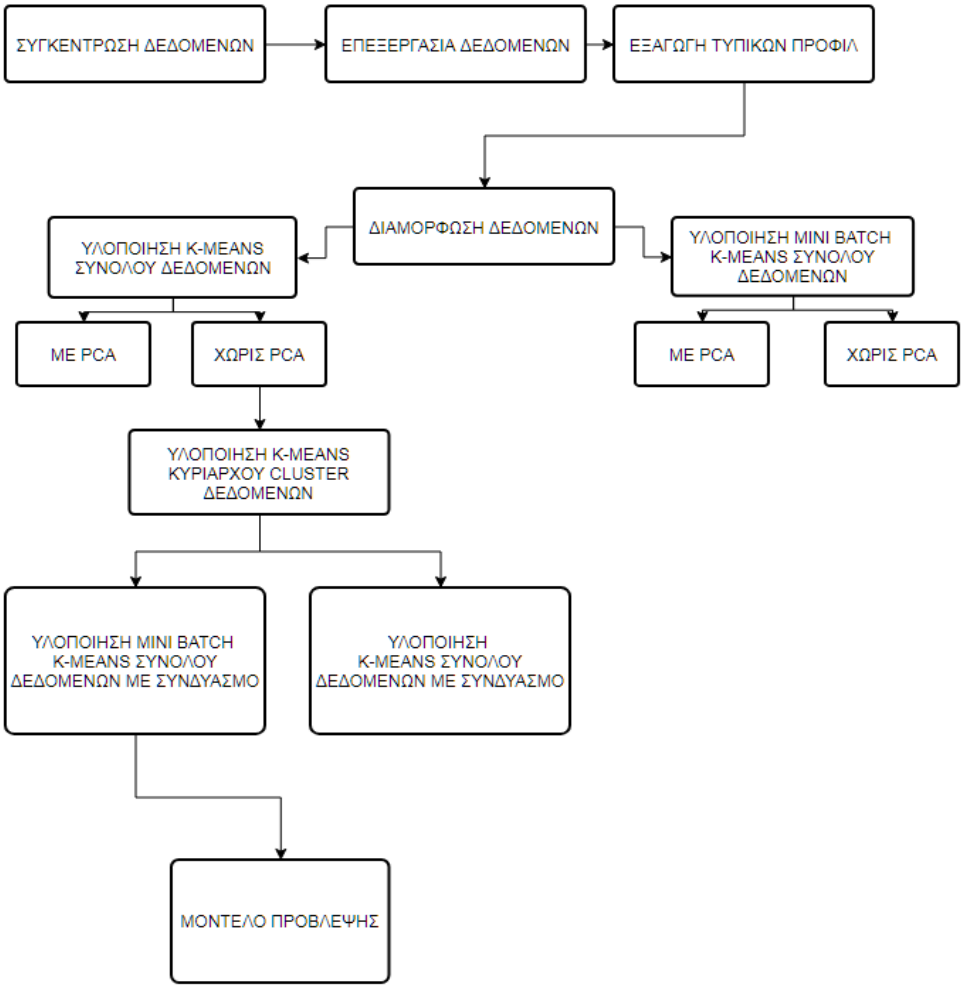
Οι παράμετροι του δικτύου, δηλαδή οι τιμές των βαρών που θα ικανοποιούν αυτές τις προδιαγραφές επιτυγχάνεται μέσω της διαδικασίας της μάθησης, η οποία αποτυπώνεται στις διασυνδέσεις των μονάδων και τις τιμές των βαρών. Η εκπαίδευση επιτυγχάνεται με τη συνεχή τροποποίηση των τιμών των βαρών. Στην

πλειοψηφία τους τα τεχνητά νευρωνικά δίκτυα εκπαιδεύονται με στόχο να αναπτύξουν κατάλληλη εσωτερική δομή, ώστε να εξάγουν ικανοποιητικά αποτελέσματα, όταν τους ζητηθεί να αναγνωρίσουν ή να κατηγοριοποιήσουν καινούρια άγνωστα δεδομένα - πρότυπα που θα έχουν ομοιότητες με τα δεδομένα στα οποία εκπαιδεύτηκαν. Το δίκτυο που παρουσιάζεται έχει ανατροφοδότηση (back propagation) στέλνοντας την έξοδο στην είσοδο για βελτίωση των βαρών. Η διαδικασία επαναλαμβάνεται έως ότου οι τιμές φτάσουν σε ικανοποιητικά ακρίβεια αποτελεσμάτων στους νευρώνες του στρώματος εξόδου.

Κεφάλαιο 4 :
Υλοποίηση κατηγοριοποίησης

4.1 Εισαγωγή

Στο κεφάλαιο αυτό αναλύεται η κατηγοριοποίηση που υλοποιείται σε Python στο Google Colab και εφαρμόζεται με σκοπό τη μετέπειτα τροφοδότηση των αποτελεσμάτων σε μοντέλο πρόβλεψης . Ο αρχικός σχεδιασμός αφορούσε την εφαρμογή του αλγόριθμου K-means σε όλα τα δεδομένα, ωστόσο η φύση αυτών όπως θα εξηγηθεί παρακάτω υπαγόρευσε την ανάγκη αναζήτησης παραλλαγών του αλγορίθμου για βελτίωση των αποτελεσμάτων που εξάγει. Στόχο αποτελεί η ταξινόμηση των καταναλωτών Μέσης Τάσης που τροφοδοτούνται από το Ελληνικό διασυνδεδεμένο σύστημα σε ομάδες που θα εκφράζουν τα χαρακτηριστικά των τυπικών προφίλ που τις αποτελούν. Το διάγραμμα ροής που εξηγεί τη διαδικασία που ακολουθείται, παρουσιάζεται στο Διάγραμμα 4.1.



Διάγραμμα 4.1: Διάγραμμα ροής για την υλοποίηση της κατηγοριοποίησης.

4.2 Συλλογή και διαμόρφωση των δεδομένων

Τα δεδομένα αποτελούνται από τις ωριαίες μετρήσεις Μέσης Τάσης του Ελληνικού Δικτύου Ηλεκτρικής Ενέργειας για το διασυνδεδεμένο σύστημα που δόθηκαν από τον ΔΕΔΔΗΕ για το διάστημα από 01-01-2018 έως και 30-06-2020. Οι ωριαίες μετρήσεις είναι αποθηκευμένες με τα πεδία που δόθηκαν μέσω κατάλληλων εντολών SQL (queries) από το server της Διεύθυνσης Χρηστών Δικτύου του Διαχειριστή να είναι:

- READING_DATE: Ημερομηνία μέτρησης
- READING_HOUR: Ώρα μέτρησης
- CUST_NUM: Αριθμός – Κλειδί ανά παροχή Μέσης Τάσης (διαφορετικός από τον Αριθμό Παροχής)
- ACTIVE_CONSUMPTION: Μέτρηση πραγματικής ισχύος σε kW

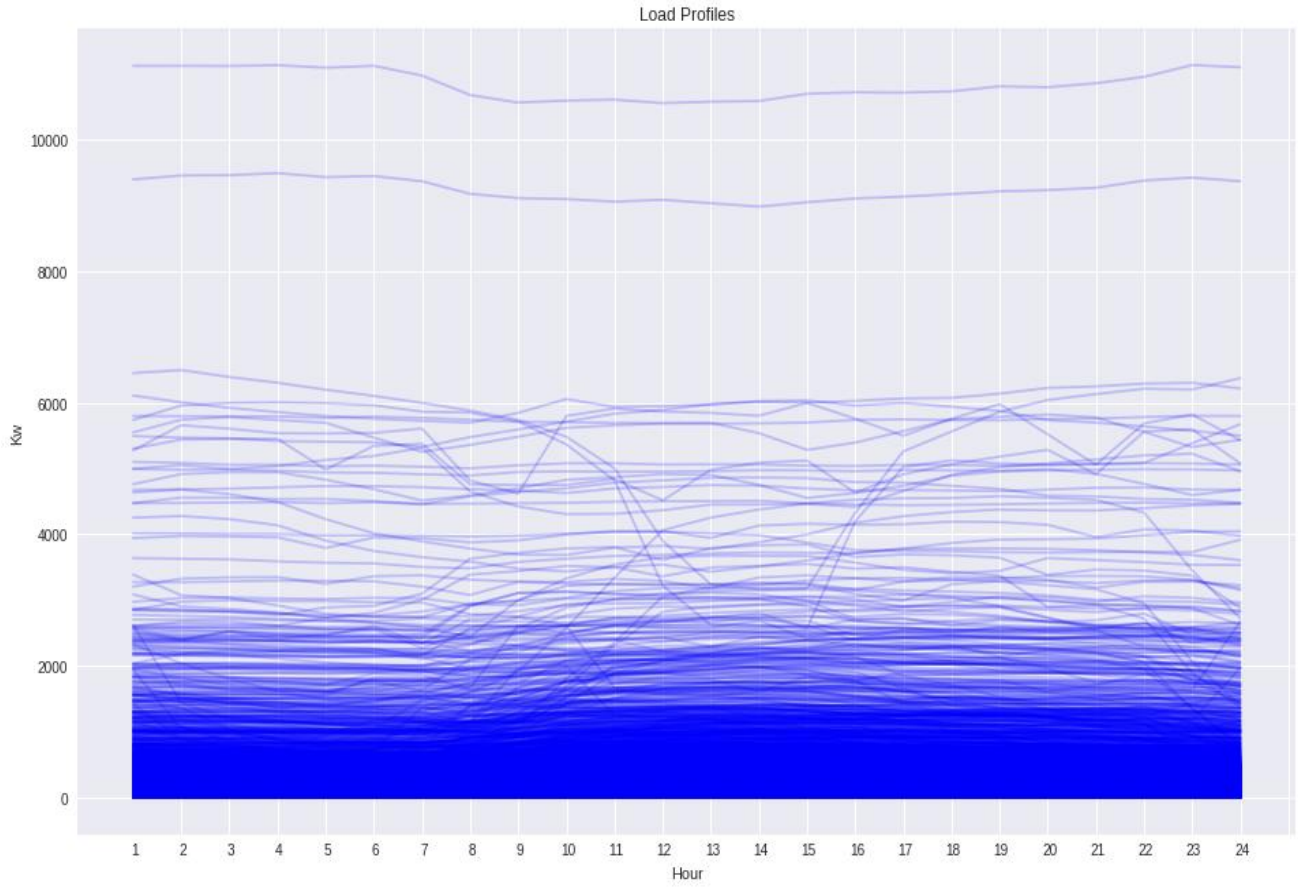
Τα δεδομένα που συλλέχθηκαν αφορούν 13.240 καταναλωτές με τον καθένα να έχει ωριαίες μετρήσεις στο χρονικό διάστημα που αναφέρθηκε, ενώ υπήρξε αναπόφευκτο να υποστούν σειρά από τροποποιήσεις ώστε να είναι σε κατάλληλη μορφή ώστε να οδηγηθούν στο εκάστοτε μοντέλο αναγνώρισης προτύπων. Σημειώνεται πως τα παραπάνω δεδομένα παραχωρήθηκαν με τη συγκεκριμένη μορφή που είναι σύμφωνη με την αρχή προστασίας προσωπικών δεδομένων των καταναλωτών. Αρχική μεταβολή είναι η διαγραφή της 25^{ης} ώρας που χρησιμοποιεί σε όλη τη διάρκεια του έτους στα συστήματά του ο ΔΕΔΔΗΕ, ώστε να προνοεί για τις δύο ημέρες του χρόνου που αλλάζει η ώρα.

Το dataset που δημιουργήθηκε είναι γιγαντιαίων διαστάσεων αν αναλογιστεί κανείς μια μέση διάρκεια τροφοδότησης της κάθε παροχής τα 2 έτη, δηλαδή έχοντας ως εισόδους 13240*24*365*2 = 231.964.800 στοιχεία χρονοσειρών, με ταυτόχρονη ύπαρξη πολλών παροχών με μηδενική κατανάλωση, διακοπή εντός του χρονικού διαστήματος, πλασματικές ακραίες τιμές κ.α. Το γεγονός αυτό οδήγησε στην επεξεργασία των δεδομένων με αρχική παράλειψη όσων παροχών έχουν μέση ημερήσια κατανάλωση 0 και όσων έχουν κατανάλωση εντός του 2020, καθώς η επίδραση της πανδημίας του κορωνοϊού SARS-CoV-2 και των μέτρων αντιμετώπισής της άλλαξαν τις καταναλώσεις ηλεκτρισμού τόσο στην Ελλάδα, όσο και παγκόσμια. Οι παροχές πλέον ανέρχονται σε 12.995, όπου για την κάθε μία από αυτές εξάχθηκε το τυπικό ημερήσιο προφίλ (μέση ωριαία κατανάλωση κατά τη διάρκεια μίας ημέρας), το οποίο θα αποτελέσει και το αντικείμενο της κατηγοριοποίησης με σύνολο εισόδων 12.995 διανύσματα 24 διαστάσεων.

Τα δεδομένα διαμορφωθήκαν για την εισαγωγή τους στο μοντέλο ως ένα αρχείο txt, όπου κάθε αριθμός παροχής ταυτίστηκε με μία μοναδική ημερομηνία έχοντας δηλαδή ωριαίες μετρήσεις για συνολικά 12995 διαδοχικές ημέρες. Η μετατροπή τους έγινε με τη βοήθεια συναρτήσεων του Microsoft Excel ώστε να έχουν τελική μορφή:

```
Date;Time;Global_active_power
01/01/1950;01:00:00;28.30
01/01/1950;02:00:00;25.41
01/01/1950;03:00:00;23.75
01/01/1950;04:00:00;22.72
01/01/1950;05:00:00;22.14
01/01/1950;06:00:00;21.54
01/01/1950;07:00:00;21.39
…
```

Τα μοντέλα κατηγοριοποίησης και πρόβλεψης υλοποιήθηκαν στο Google Colab με χρήση της γλώσσας προγραμματισμού Python, καθώς και όλα τα διαγράμματα συμπεριλαμβάνονται στην παρούσα εργασία. Σε αυτό το στάδιο χρησιμοποιήθηκαν οι βιβλιοθήκες *google.colab, pandas, numpy* και *matplotlib* για τη φόρτωση του αρχείου, τη μετατροπή του σε dataset, πίνακα και τη σχεδίασή του αντίστοιχα. Το σύνολο των τυπικών ημερήσιων προφίλ παρουσιάζονται στο Διάγραμμα 4.2, στο οποίο ξεχωρίζει η μεγάλη συγκέντρωση παροχών με χαμηλή σχετικά κατανάλωση (μικρότερη από 1.000 kW ανά ώρα). Το χαρακτηριστικό αυτό οδήγησε στη διαγραφή παροχών με ελάχιστη μέση ημερήσια κατανάλωση, δηλαδή μικρότερη των 5 kWh ανά ημέρα, καθιστώντας τον αριθμό παροχών – ημερών του τελικού dataset σε 12.104.



Διάγραμμα 4.2: Τυπικά προφίλ 12995 πελατών Μέσης Τάσης διασυνδεδεμένου συστήματος.

Η μορφή του dataset όπως τελικά αυτό μετατρέπεται για εισαγωγή στο αλγοριθμικό μοντέλο είναι της μορφής που ακολουθεί, όπου προβάλλονται οι 5 πρώτες εγγραφές του με κάθε μία να αφορά έναν πελάτη.

hour	0	1	2	3	4	5	6	7	8	9	10	11
1950-01-01	34.20	27.81	25.05	23.66	22.63	22.03	21.47	21.35	23.47	30.69	39.74	49.13
1950-01-02	17.02	17.02	17.04	17.05	17.07	17.08	16.98	15.91	17.76	46.71	65.12	68.53
1950-01-03	373.72	362.08	352.32	344.65	339.16	337.12	341.51	385.39	479.50	562.56	596.69	612.07
1950-01-04	3.38	3.38	3.41	3.44	3.47	3.49	3.47	3.41	3.30	3.09	3.01	2.99
1950-01-05	72.59	71.44	70.73	70.18	69.65	69.63	69.29	69.66	100.51	146.72	163.94	168.77

…

Διάγραμμα 4.3: Μορφή dataset που εισάγεται στο μοντέλο.

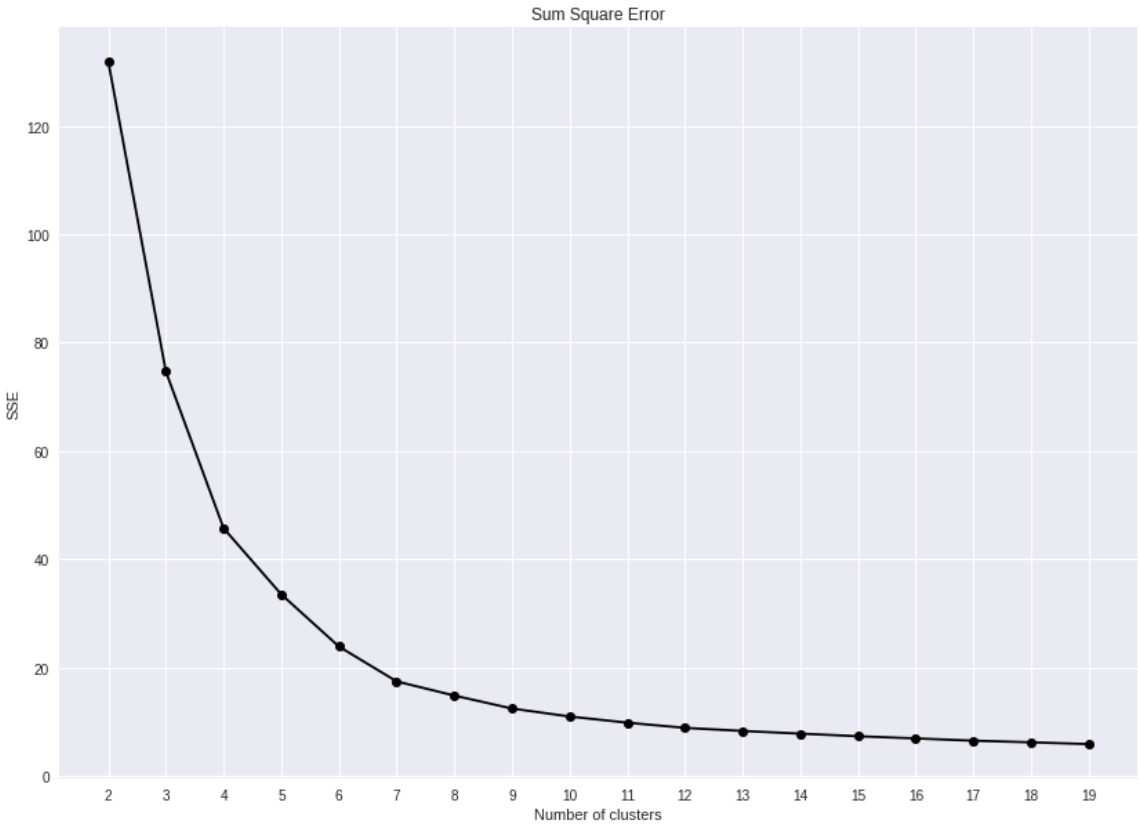
4.3 Υλοποίηση αλγορίθμου K-means

Η κατηγοριοποίηση – clustering υλοποιείται στη διπλωματική εργασία με εφαρμογή του αλγορίθμου K-means και παραλλαγών του που κρίθηκαν απαραίτητες για τα συγκεκριμένα δεδομένα ώστε να βελτιστοποιηθεί η απόκριση του. Η κατασκευή του αλγόριθμου έγινε με τη βοήθεια της open-source βιβλιοθήκης για την Python: Sklearn (Scikit - learn) και των υποβιβλιοθηκών που περιέχει [31]. Αρχικά, εφαρμόζεται ο αλγόριθμος για όλες τις διαστάσεις των δεδομένων, δηλαδή και τις 24 και έπειτα δοκιμάζεται η εφαρμογή μείωσης διαστάσεων μέσω της μεθόδου Ανάλυσης Κύριων Συνιστωσών (PCA) ώστε να διαπιστωθεί εάν αυτή βοηθά στη βελτιστοποίηση της κατηγοριοποίησης.

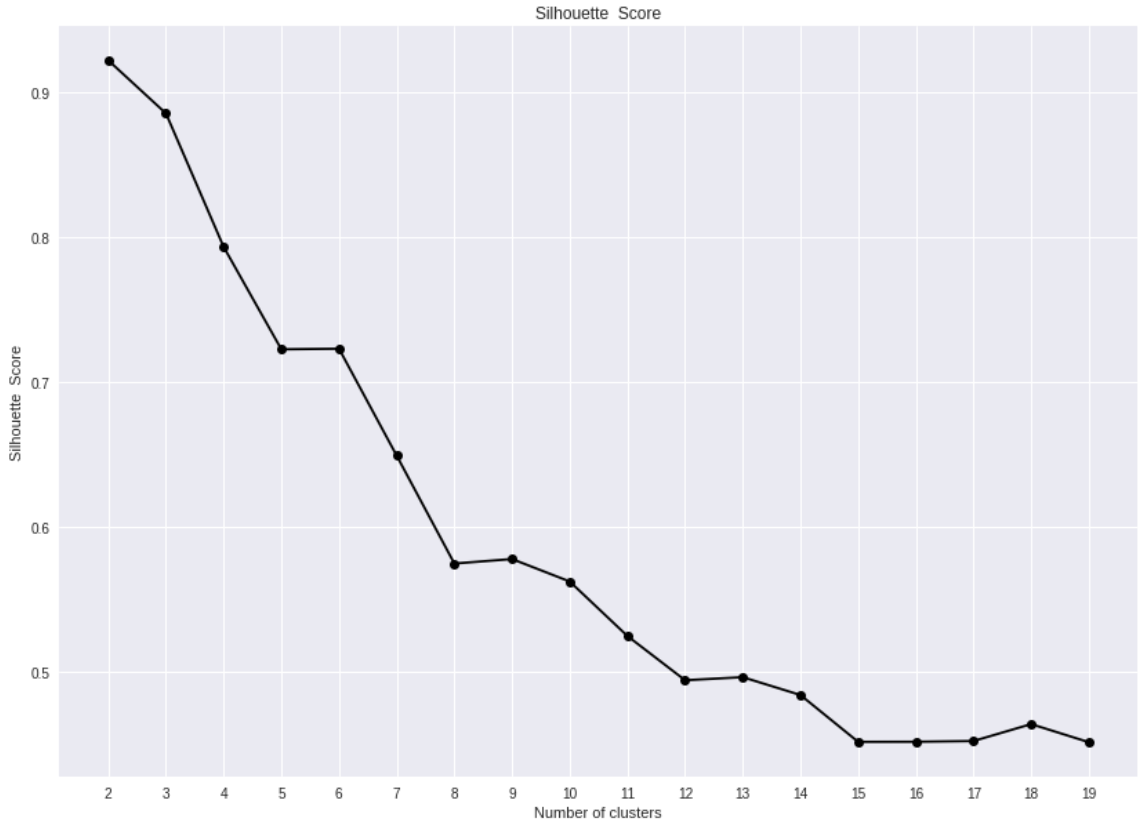
Πρώτο βήμα έπειτα από τη διαμόρφωση των δεδομένων είναι η κανονικοποίησή τους σε εύρος τιμών [0,1], χρησιμοποιώντας την κλάση MinMaxScaler, ενώ για αρχικοποίηση των αλγορίθμων χρησιμοποιήθηκε ο αλγόριθμος `'k-means++'` που βελτιστοποιεί τη διαδικασία επιλογής των αρχικών τιμών – κέντρων. Έπειτα, υπολογίζονται οι δείκτες αξιολόγησης αθροίσματος τετραγωνικού λάθους (Sum of Square Error – SSE ή WSS) [32], Silhouette Score [33], Davies Bouldin [34] από 2 έως 20 clusters και για τα δύο σενάρια για την εύρεση του ιδανικού αριθμού τους. Ακολουθεί η διαδικασία που ακολουθήθηκε και για τις δύο περιπτώσεις.

4.3.1 Κατηγοριοποίηση με K-means

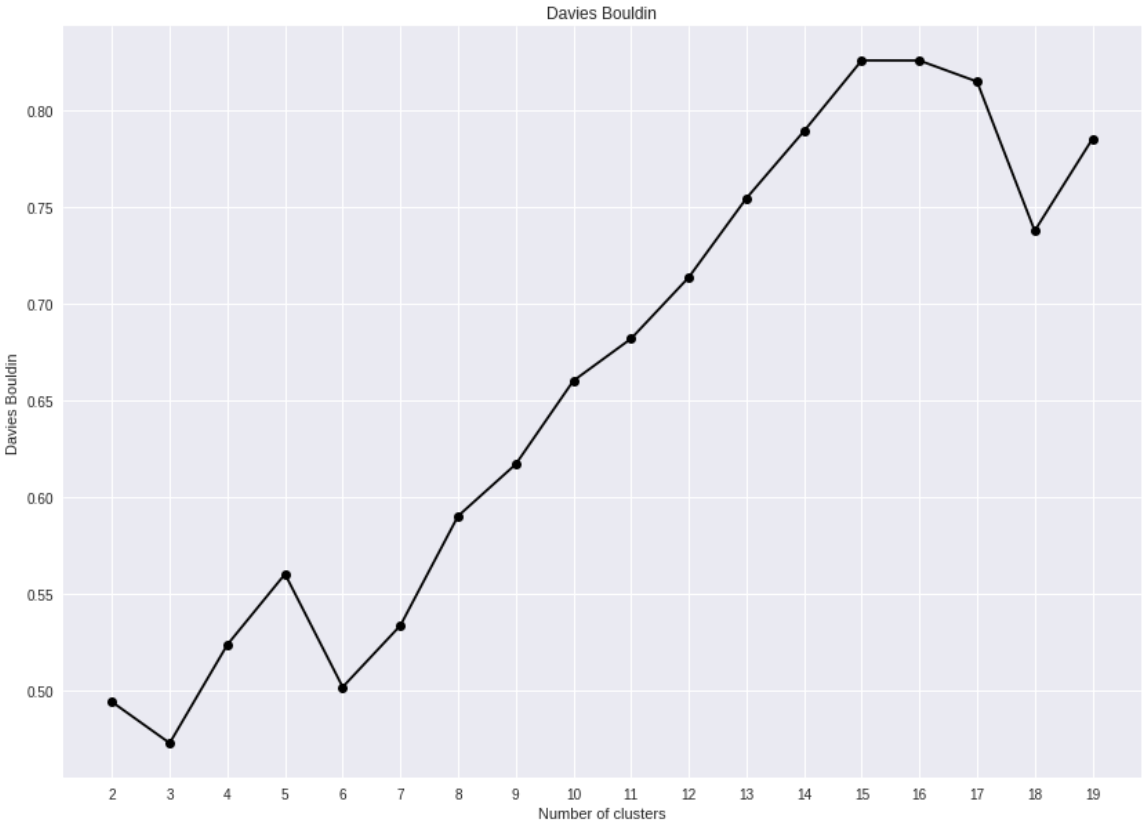
Τα αποτελέσματα των δεικτών αξιολόγησης φαίνονται στα Διαγράμματα 4.4 έως και 4.6, από τα οποία εξάγεται το συμπέρασμα ότι ο ιδανικός αριθμός των ομάδων είναι n = 4 ή n = 5.



Διάγραμμα 4.4: Δείκτης αξιολόγησης Sum of Square Error – SSE για αριθμό cluster n = 2 έως 20.

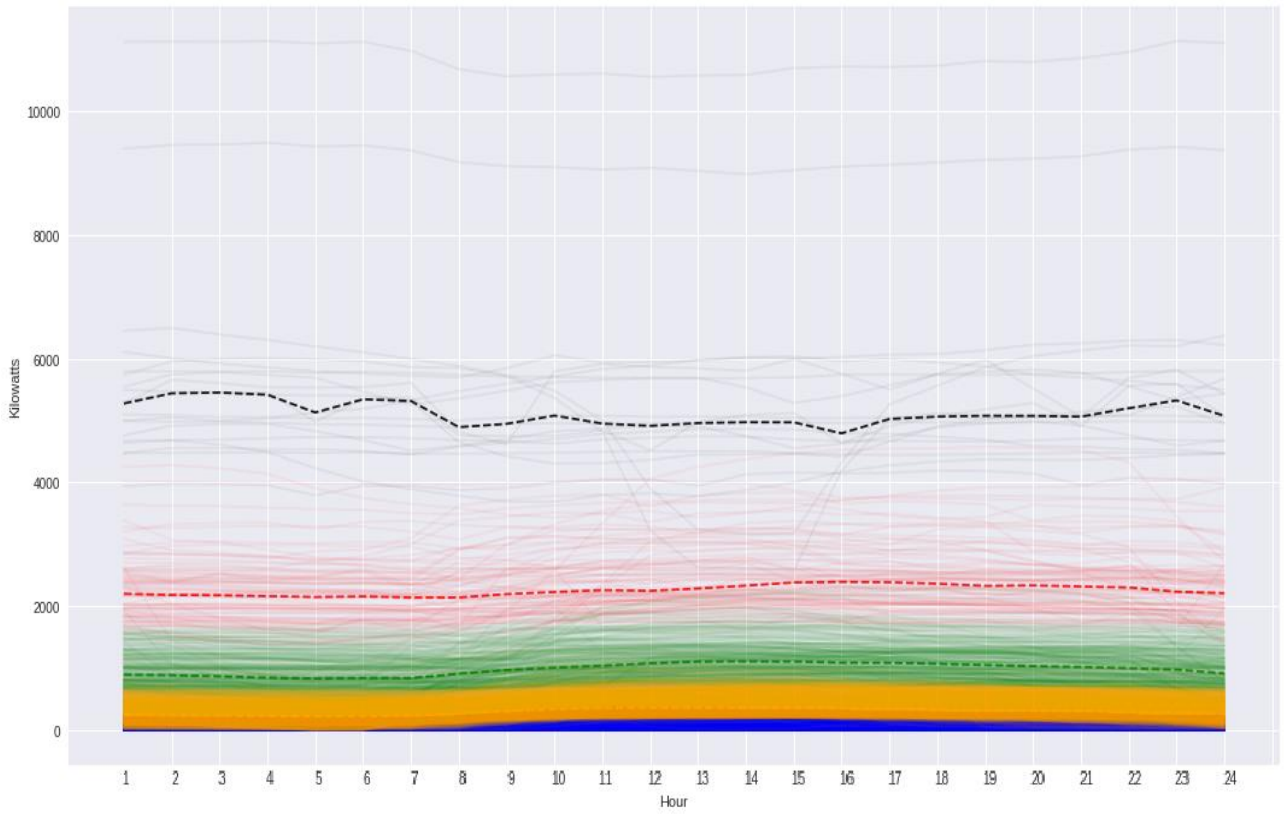


Διάγραμμα 4.5: Δείκτης αξιολόγησης Silhouette Score για αριθμό cluster n = 2 έως 20.



Διάγραμμα 4.6: Δείκτης αξιολόγησης Davies Bouldin για αριθμό cluster n = 2 έως 20.

Η επιλογή για τον προτεινόμενο αριθμό ομάδων έγινε βάσει του συνδυασμού του κανόνα του αγκώνα για τον δείκτη SSE, της βέλτιστης μεγαλύτερης τιμής του δείκτη Silhouette Score και της βέλτιστης ελάχιστης τιμής του δείκτη Davies Bouldin. Επιλέχθηκε n = 5 με τα αντίστοιχα αποτελέσματα να φαίνονται στο Διάγραμμα 4.7, όπου με κάθε χρώμα αντιπροσωπεύει και ένα cluster και οι διακεκομμένες γραμμές το κέντρο (centroid) του.



Διάγραμμα 4.7: Αποτελέσματα K-means για n=5.

Η κατανομή των δεδομένων στις ομάδες, καθώς και το ποσοστό του πλήθους της κάθε ομάδας προς το σύνολο, δηλαδή τα 12.104 τυπικά προφίλ όπως φαίνεται στον Πίνακα 4.1.

Cluster	**Αριθμός τυπικών προφίλ**	**Ποσοστό επί του συνόλου προφίλ**
1	10353	85.53371 %
2	84	0.69399 %
3	277	2.2885 %
4	19	0.15697 %
5	1371	11.32683 %

Πίνακας 4.1: Αποτελέσματα K-means για n=5.

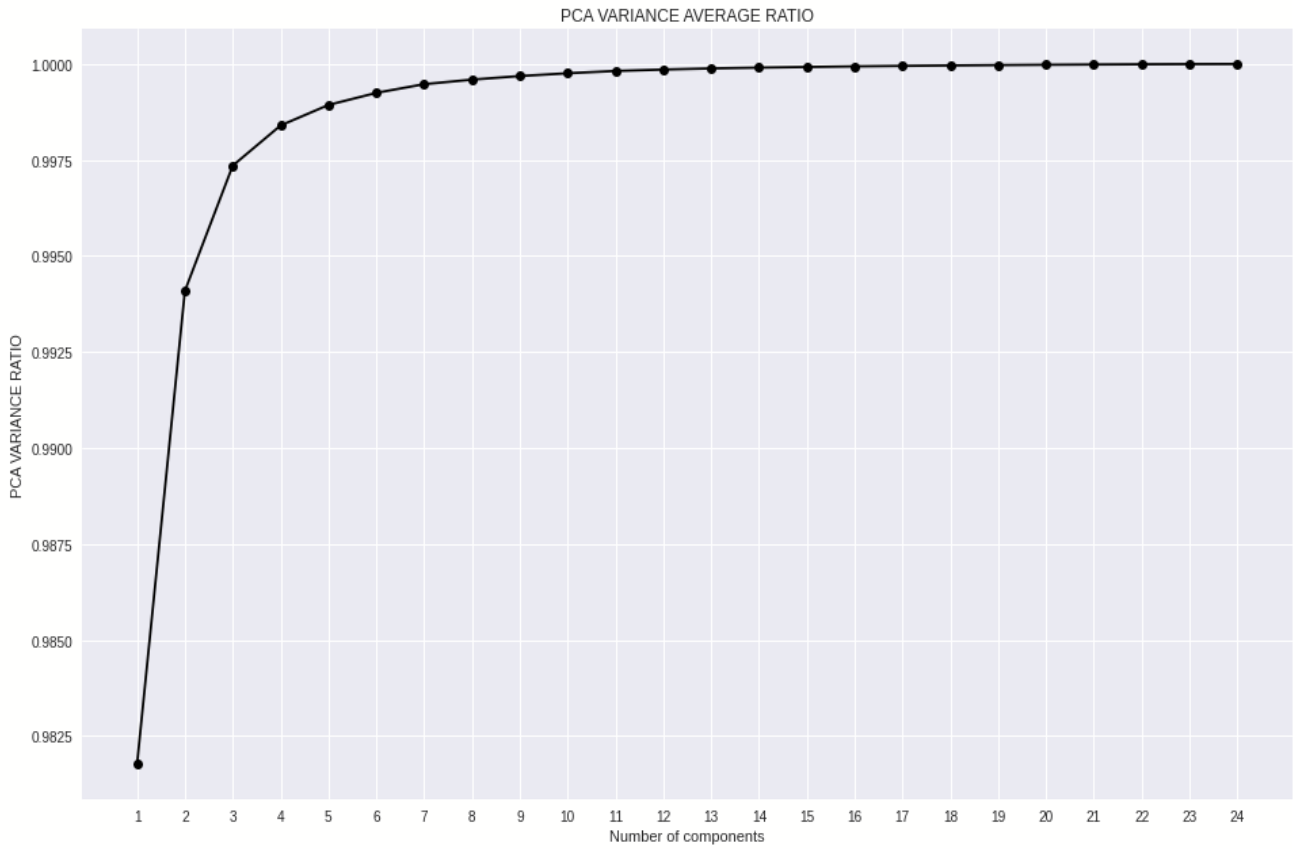
Η άνιση ταξινόμηση των δεδομένων σε ένα μόνο cluster εξηγείται από τη φύση των δεδομένων που όπως παρουσιάστηκε έχει μεγάλη συγκέντρωση στα χαμηλά φορτία, ωστόσο αποτελεί πρόβλημα στην αποδοτικότητα της ομαδοποίησης και της μετέπειτα αξιοποίησής της για την πρόβλεψη των επόμενων 24

ωρών. Στη συνέχεια γίνεται προσπάθεια επίλυσης αυτού με τη μείωση των διαστάσεων των δεδομένων ώστε να χρησιμοποιηθεί μόνο η απαραίτητη πληροφορία για την κατηγοριοποίησή τους.

4.3.2 Κατηγοριοποίηση με K-means και μείωση των διαστάσεων

Η μείωση των διαστάσεων επιτυγχάνεται με την εφαρμογή της μεθόδου Ανάλυσης Κύριων Συνιστωσών (Principal Component Analysis - PCA), η οποία χρησιμοποιείται για μεγάλα datasets μειώνοντας το πλήθος των μεταβλητών, διατηρώντας όμως την απαραίτητη πληροφορία. Η ανάλυση κυρίων συνιστωσών είναι μια στατιστική διαδικασία με την οποία αναπαριστούμε ένα πίνακα συνδιακύμανσης ενός συνόλου μεταβλητών μέσα από ένα διαφορετικό νέο σύνολο μεταβλητών που προκύπτουν από το γραμμικό συνδυασμό των πρώτων. Η μέθοδος μέσα από τους γραμμικούς συνδυασμούς των αρχικών μεταβλητών δημιουργεί ένα νέο σύστημα συντεταγμένων μετατοπίζοντας και περιστρέφοντας το παλαιό σύστημα. Οι νέοι άξονες καθορίζουν τις κατευθύνσεις που παρουσιάζουν τις μέγιστες μεταβολές των δεδομένων. Οι κύριες συνιστώσες της PCA βρίσκονται από τον υπολογισμό των ιδιοδιανυσμάτων και των αντιστοίχων ιδιοτιμών του πίνακα συνδιακύμανσης, όπως αυτός υπολογίζεται από τα δεδομένα που υπάρχουν. Η μείωση των διαστάσεων συνοδεύεται από τη μείωση της ακρίβειας της πληροφορίας, οπότε και η εύρεση της χρυσής τομής μεταξύ ακρίβειας και διαστάσεων είναι καθοριστική [35], [36].

Στο διάγραμμα που ακολουθεί προβάλλεται το ποσοστό της πληροφορίας που διατηρείται (PCA Variance Ratio) από τα τυπικά προφίλ ανάλογα με τον αριθμό διαστάσεων που επιλέγεται για διαστάσεις από 1 έως και 24. Ο υπολογισμός έγινε με τη βοήθεια της κλάσης PCA της βιβλιοθήκης `sklearn.decomposition` [37].



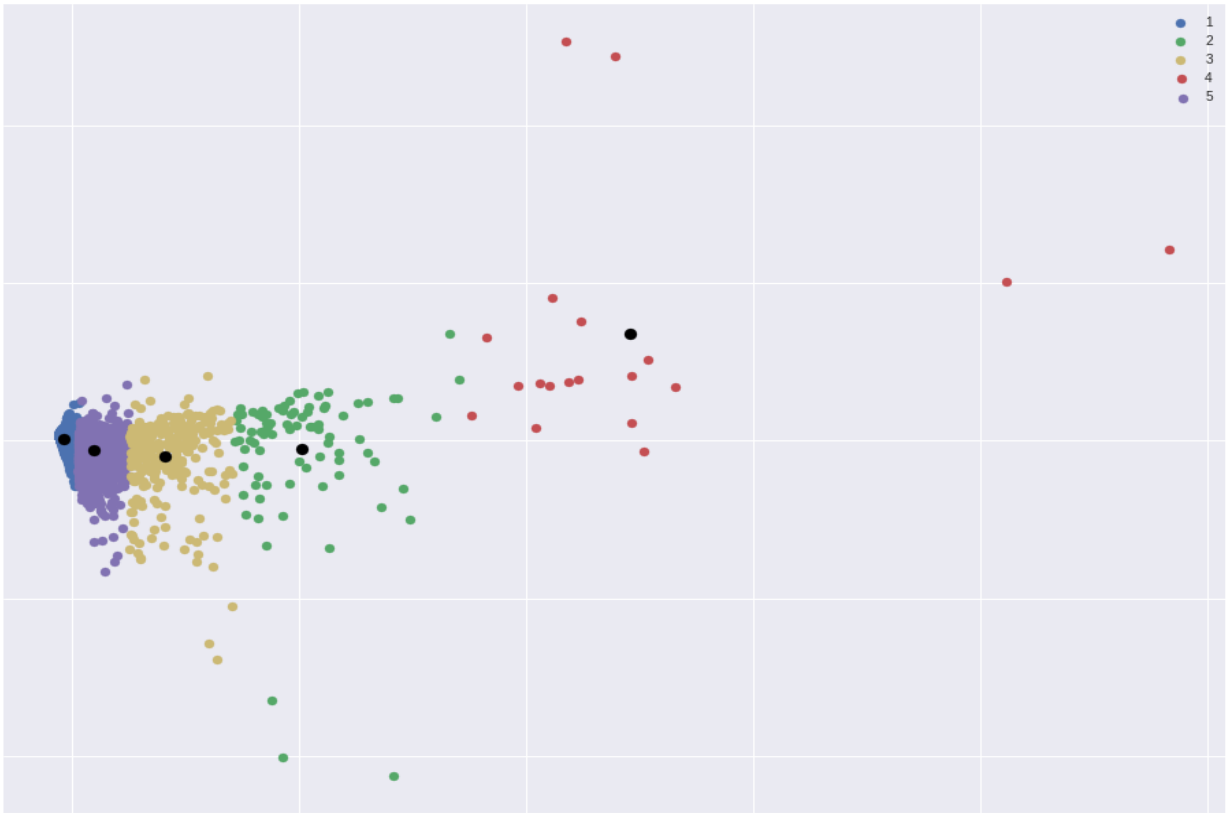
Διάγραμμα 4.8: Ποσοστό διατήρησης πληροφορίας από εφαρμογή του PCA για 1 έως και 24 διαστάσεις.

Επιλέγονται d= 12 διαστάσεις, καθώς είναι οι ελάχιστες που διατηρούν ολόκληρη την πληροφορία και έπειτα γίνεται μετασχηματισμός των δεδομένων σε αυτές με σκοπό την εφαρμογή του K-means στο dataset που προκύπτει. Στη συνέχεια, διερευνήθηκε εκ νέου ο ιδανικός αριθμός cluster χωρίς ωστόσο να προκύπτει σημαντική διαφορά τους δείκτες αξιολόγησης που εξήχθησαν για όλες τις διαστάσεις, τα διαγράμματα των οποίων παρουσιάζονται στο Παράρτημα Π.1 - Π.3. Έτσι, επιλέχθηκαν και πάλι 5 clusters με τα αποτελέσματα της κατηγοριοποίησης να παρουσιάζονται στον παρακάτω πίνακα.

Cluster	**Αριθμός τυπικών προφίλ**	**Ποσοστό επί του συνόλου προφίλ**
1	10354	85.54197 %
2	84	0.69399 %
3	277	2.2885 %
4	19	0.15697 %
5	1370	11.31857 %

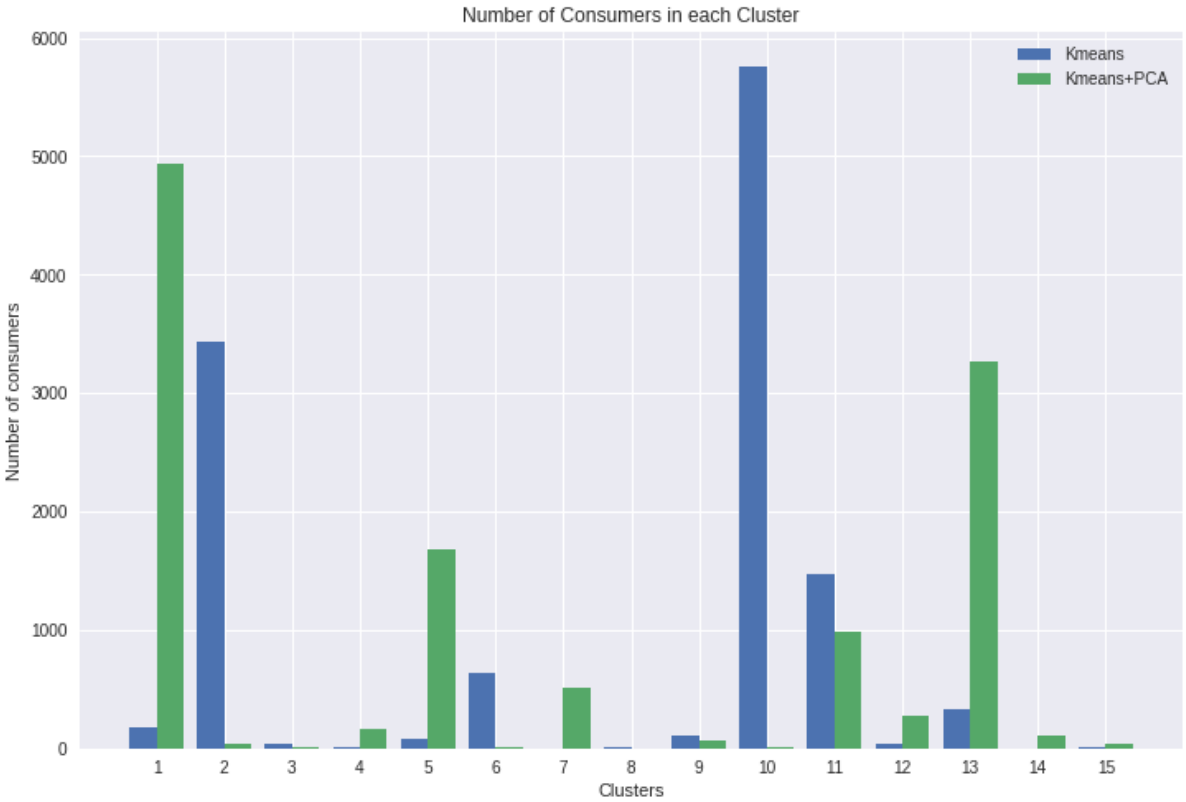
Πίνακας 4.2: Αποτελέσματα K-means με μείωση διαστάσεων (d=12) για n=5.

Η μείωση των διαστάσεων σε 2 από 24 επιτρέπει την αποτύπωση της εξόδου σε ένα δισδιάστατο διάγραμμα όπου κάθε cluster έχει συγκεκριμένο χρώμα με τα κέντρα τους να είναι τα μαύρα σημεία (Διάγρ. 4.9). Σημειώνεται ότι το cluster 1 προβάλλεται έτσι λόγω της μεγάλης πυκνότητάς του.



Διάγραμμα 4.9: Δισδιάστατη αποτύπωση K-means με μείωση διαστάσεων (d=12) για n=5 clusters.

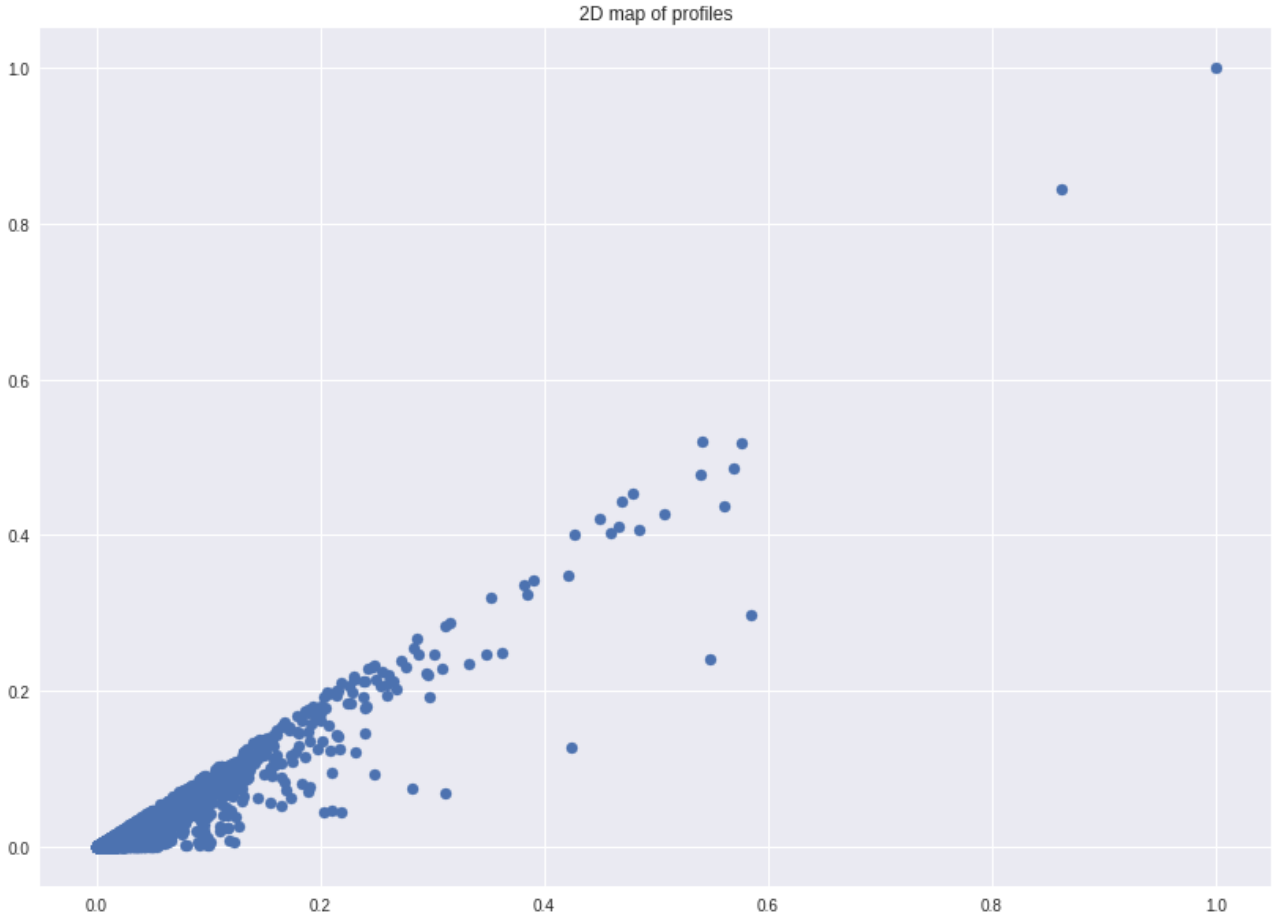
Τα αποτελέσματα είναι κοινά με αυτά που ελήφθησαν στην περίπτωση 4.2.1, οπότε κρίνεται αναγκαία η διεξαγωγή περαιτέρω διερεύνησης καλύτερου τρόπου για την κατηγοριοποίηση, η οποία παρουσιάζεται στο κεφάλαιο που ακολουθεί και αφορά την εφαρμογή μιας παραλλαγής του K-means τον αλγόριθμο Mini Batch K-means. Τέλος, δοκιμάστηκε επίσης η αύξηση του αριθμού των cluster σε 15 που δεν προσέφερε ωστόσο κάποια βελτίωση, όπου η ταξινόμηση σε ομάδες ήταν η εξής:



Διάγραμμα 4.10: Αποτελέσματα K-means (μπλε) και K-means με PCA(πράσινο) για n=15.

4.4 Υλοποίηση αλγορίθμου Mini Batch K-means

Ο αλγόριθμος Mini batch K-means αποτελεί παραλλαγή του κλασικού με τη διαφορά πως οι αποστάσεις που καθορίζουν την τοποθέτηση των κέντρων δεν υπολογίζονται από το σύνολο των δεδομένων σε κάθε επανάληψη, αλλά από τυχαία πακέτα δεδομένων σταθερού μήκους. Με αυτό τον τρόπο αποφεύγεται η κυριαρχία των σημείων με υψηλή συγκέντρωση στο dataset, όπως αυτή παρουσιάστηκε στις προηγούμενες περιπτώσεις επηρεάζοντας τα αποτελέσματα της κατηγοριοποίησης. Ταυτόχρονα, εξοικονομείται χρόνος υπολογισμού στη διαδικασία με τη θεωρητική απώλεια ακρίβειας της κατηγοριοποίησης [38], [39]. Ωστόσο, στην παρούσα κατάσταση που το clustering καλείται να δώσει λύση σε σημαντικά ασύμμετρο dataset, όπως αυτό παρουσιάζεται δισδιάστατα στο Διάγραμμα 4.11 όπου υπάρχει τεράστια συγκέντρωση στοιχείων σε συγκεκριμένο χώρο, η ακρίβεια και αντίστοιχα η αξιοποίηση για τη μετέπειτα πρόβλεψη ήδη διακυβεύεται.

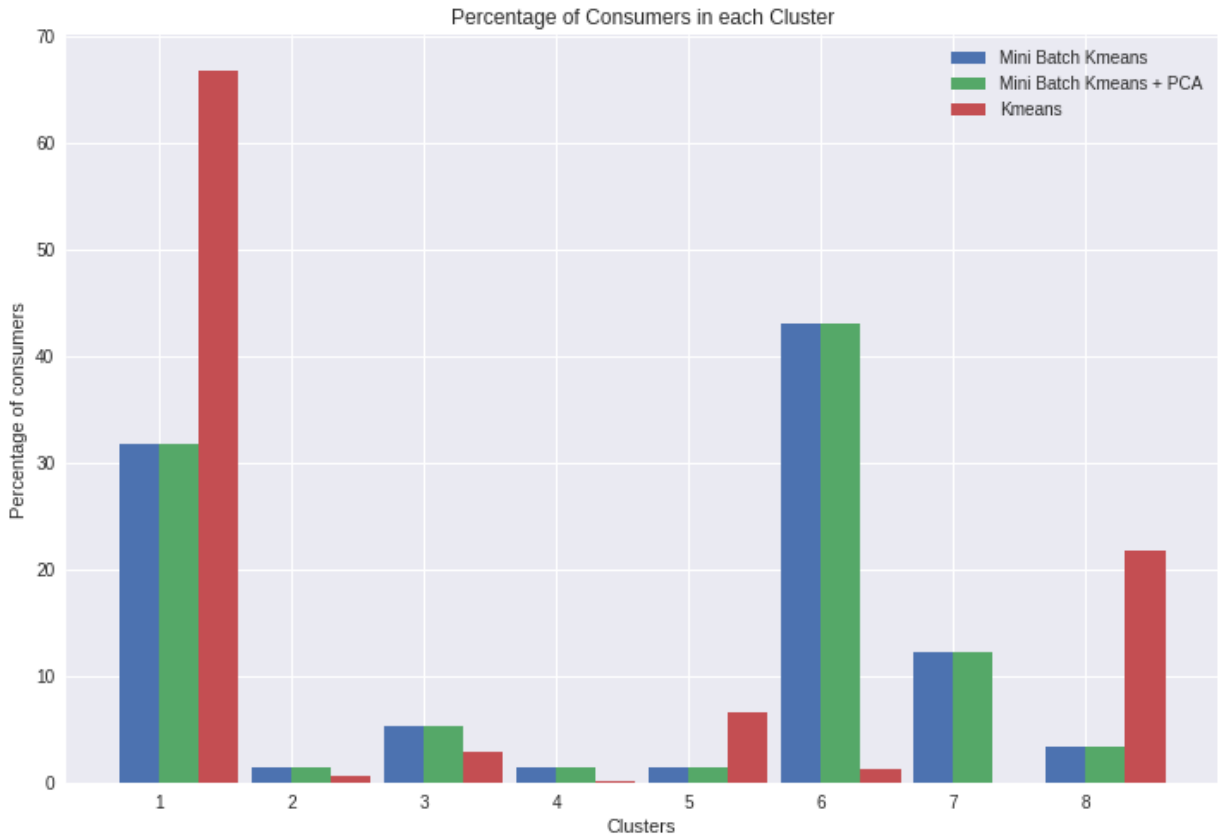


Διάγραμμα 4.11: Δισδιάστατη αποτύπωση όλων των τυπικών προφίλ καταναλώσεων.

Ο Mini Batch K-means αρχικοποιείται παρόμοια με τον κλασικό, με μόνη διαφορά τον προσδιορισμό της παραμέτρου Batch Size, δηλαδή του πλήθους των διανυσμάτων από τα οποία υπολογίζονται οι αποστάσεις. Επιλέχθηκε μετά από δοκιμές Batch_size =24 και εξήχθησαν αποτελέσματα για όλες και για μειωμένες μέσω PCA διαστάσεις για n = 8 clusters.

	Όλες οι διαστάσεις (d=24)		Μειωμένες διαστάσεις (d=12)	
Cluster	**Αριθμός τυπικών προφίλ**	**Ποσοστό επί του συνόλου προφίλ**	**Αριθμός τυπικών προφίλ**	**Ποσοστό επί του συνόλου προφίλ**
1	3835	31.68374 %	3835	31.68374 %
2	178	1.47059 %	178	1.47059 %
3	639	5.27925 %	639	5.27925 %
4	179	1.47885 %	180	1.48711 %
5	179	1.47885 %	179	1.47885 %
6	5199	42.95274 %	5199	42.95274 %
7	1487	12.28519 %	1487	12.28519 %
8	408	3.37079 %	407	3.36252 %

Πίνακας 4.3: Αποτελέσματα Mini Batch K-means με όλες και με μειωμένες διαστάσεις(d=12) για n=8.



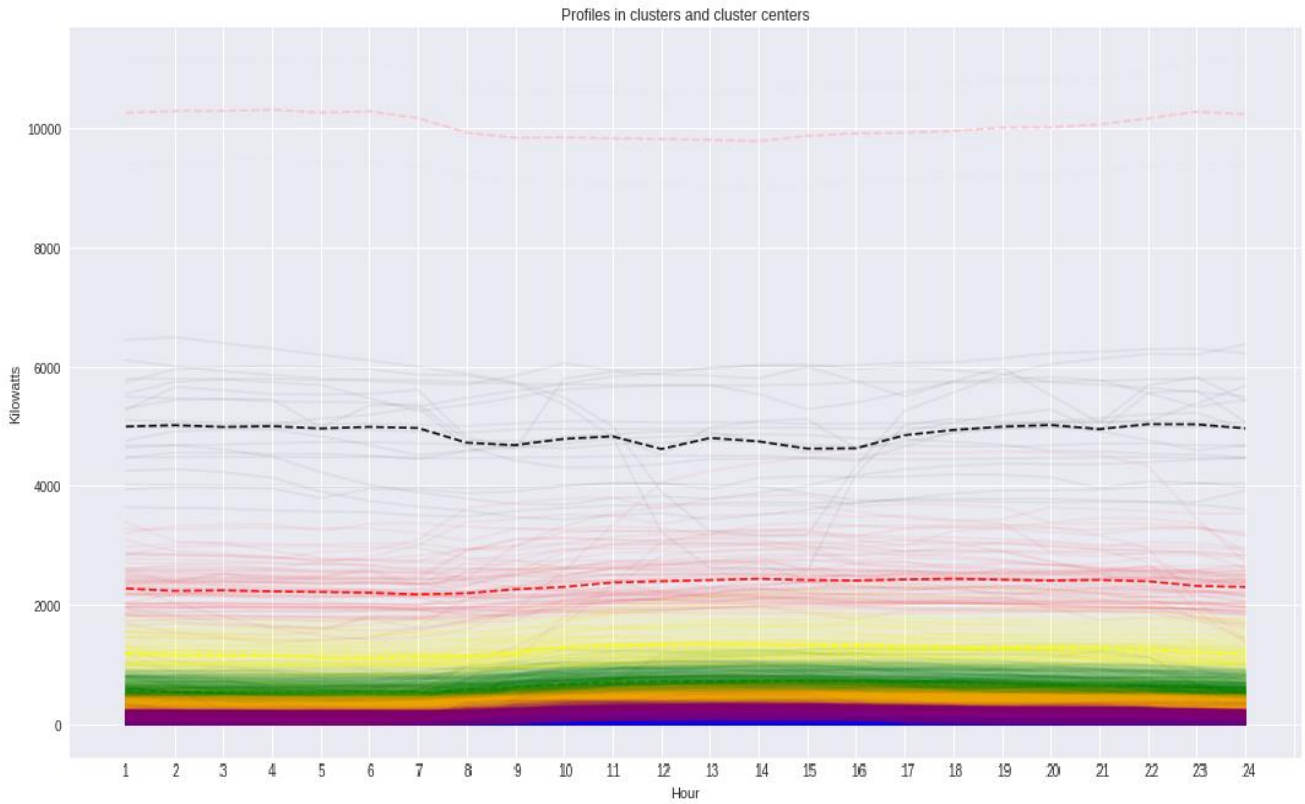
Διάγραμμα 4.12: Ποσοστά επί του συνόλου του πλήθους τυπικών προφίλ ανά ομάδα χρησιμοποιώντας: Mini Batch K-Means (μπλε), Mini Batch K-means με PCA (πράσινο) και K-means (κόκκινο) για n=8.

Το διάγραμμα 4.12 παρουσιάζει τη σύγκριση μεταξύ των τριών παραλλαγών του K-means η οποία αναδεικνύει τη βελτίωση της κατανομής των δεδομένων σε cluster, χωρίς ωστόσο να έχει φτάσει σε ικανοποιητικά επίπεδα χρησιμοποιώντας τον Mini Batch, ενώ για άλλη μια φορά η μείωση διαστάσεων δε φαίνεται να προσφέρει κάποιο όφελος. Τα δύο αυτά στοιχεία οδηγούν σε περαιτέρω διερεύνηση με χρήση Guided K-means χωρίς μείωση διαστάσεων που θα αναλυθεί στο επόμενο κεφάλαιο.

4.5 Υλοποίηση αλγορίθμου Guided K-means

Η επόμενη παραλλαγή του αλγόριθμου K-means αφορά την καθοδηγούμενη (guided) υλοποίηση του για την τελική κατηγοριοποίηση των δεδομένων. Σε αυτή τα κέντρα των ομάδων δεν αρχικοποιούνται από τον αλγόριθμο `'k-means++'` όπως σε όλες τις προηγούμενες περιπτώσεις, αλλά εισάγονται χειροκίνητα, ενώ ο αλγόριθμος εκτελείται συνολικά 3 φορές. Τα βήματα που θα ακολουθηθούν είναι:

1. Εφαρμογή του K-means για όλο το dataset για n = 8 clusters (ίδια με αυτή που παρουσιάστηκε στο προηγούμενο κεφάλαιο) και αποθήκευση των κέντρων του.



Διάγραμμα 4.13: Αποτελέσματα K-means για n=8 για όλο το dataset.

2. Εξαγωγή του πυκνότερου και πολυπληθέστερου cluster με n = 1 που περιέχει 8083 στοιχεία (66,78 % του συνόλου) ως ξεχωριστό dataset
3. Εφαρμογή K-means και Mini Batch K-means στο δεύτερο dataset για n = 8.
4. Εισαγωγή των 7 κέντρων που προέκυψαν από το βήμα 1 (παραλείπεται το κέντρο του cluster που χρησιμοποιήθηκε ως dataset) και των 8 από το βήμα 3 ως αρχικοποίηση για εφαρμογή εκ νέου των αλγορίθμων K-means και Mini Batch K-means στο σύνολο των δεδομένων.

Ακολουθεί η ανάλυση των βημάτων 2 έως και 4.

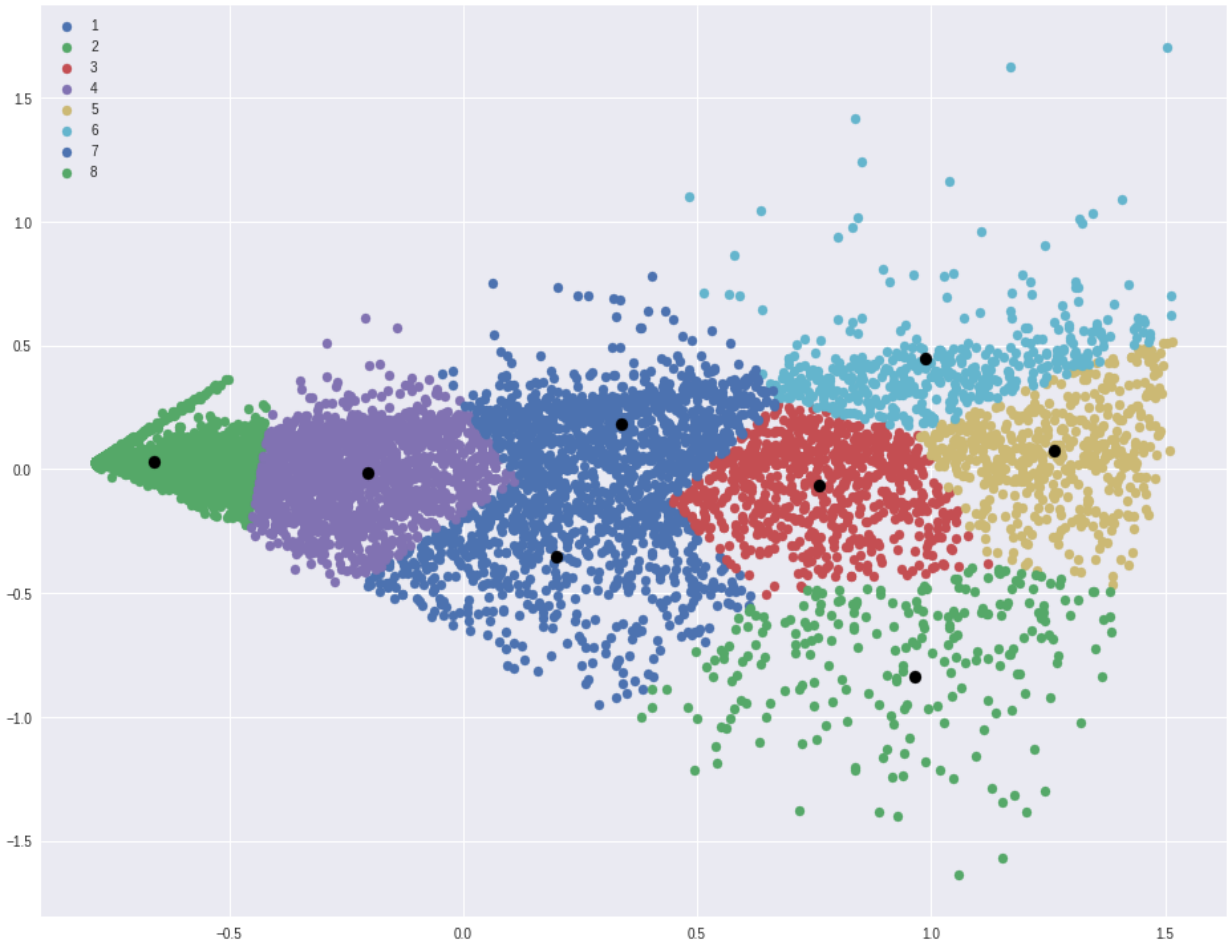
4.5.1 Κατηγοριοποίηση πολυπληθέστερου cluster

Στο dataset με τα δεδομένα του πολυπληθέστερου cluster εφαρμόζεται ο K-means με τα αποτελέσματα για την πλήρωση των cluster να παρουσιάζονται στον Πίνακα 4.4.

	K-means		**Mini Batch K-means**	
Cluster	**Αριθμός τυπικών προφίλ**	**Ποσοστό επί του συνόλου προφίλ**	**Αριθμός τυπικών προφίλ**	**Ποσοστό επί του συνόλου προφίλ**
1	584	7.22504 %	583	7.21267 %
2	2584	31.96833 %	3117	38.56241 %
3	674	8.33849 %	1076	13.31189 %
4	935	11.56749 %	1886	23.33292 %
5	486	6.01262 %	44	0.54435 %
6	331	4.09501 %	892	11.03551 %
7	871	10.7757 %	182	2.25164 %
8	1618	20.01732 %	303	3.74861 %

Πίνακας 4.4: Αποτελέσματα K-means και Mini Batch K-means στο πολυπληθέστερο cluster για n=8.

Επιλέγεται η χρησιμοποίηση του K-means για το συγκεκριμένο στάδιο καθώς εξάγει τα πιο ισορροπημένα αποτελέσματα. Η δισδιάστατη αποτύπωση της κατηγοριοποίησής τους στο επίπεδο προβάλλεται στο Διάγραμμα 4.14, όπου φαίνεται η άμβλυνση της ανισοκατανομής μεταξύ τους. Το μαύρο σημείο που σχεδιάζεται σε κάθε χρωματική ομάδα αποτελεί το κέντρο του εκάστοτε cluster.



Διάγραμμα 4.14:: Δισδιάστατη αποτύπωση K-means του πολυπληθέστερου cluster για n=8.

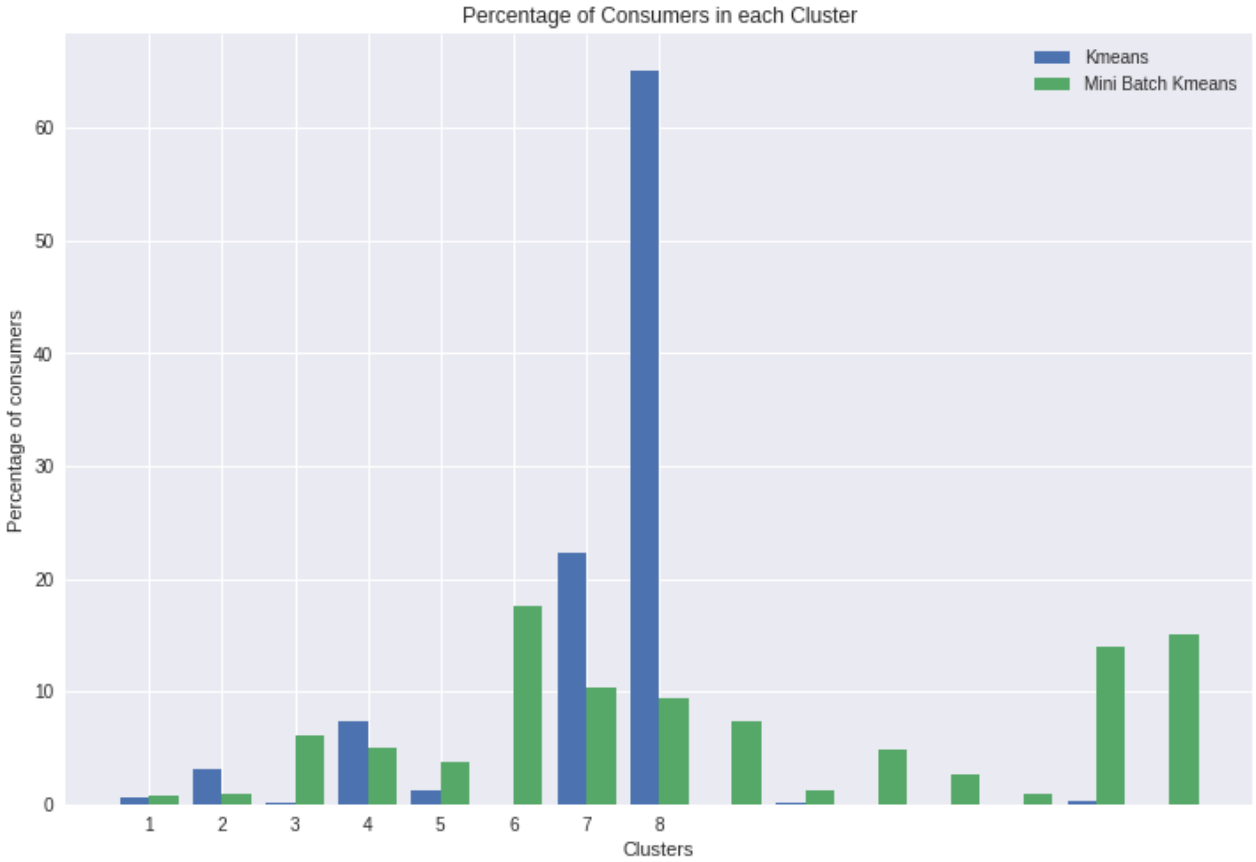
4.5.2 Συνολική κατηγοριοποίηση

Το σύνολο των δεδομένων ταξινομείται εκ νέου με τους αλγόριθμος K-means και Mini Batch K-means για την εύρεση της τελικής μορφής που θα κατηγοριοποιηθούν τα τυπικά προφίλ. Η αρχικοποίηση του αλγορίθμου γίνεται από τα διανύσματα που αποτελούνται από τα 7 κέντρα της αρχικής ολικής κατηγοριοποίησης και τα 8 που προέκυψαν από αυτή που εφαρμόστηκε στο πολυπληθέστερο cluster. Τελικό αποτέλεσμα είναι η κατηγοριοποίηση σε 15 cluster, όπως αυτή παρουσιάζεται παρακάτω.

	K-means		**Mini Batch K-means**	
Cluster	**Αριθμός τυπικών προφίλ**	**Ποσοστό επί του συνόλου προφίλ**	**Αριθμός τυπικών προφίλ**	**Ποσοστό επί του συνόλου προφίλ**
1	68	0.5618 %	91	0.75182 %
2	370	3.05684 %	115	0.9501 %
3	8	0.06609 %	744	6.14673 %
4	881	7.27859 %	605	4.99835 %
5	149	1.231 %	454	3.75083 %
6	2	0.01652 %	2132	17.61401 %
7	2690	22.22406 %	1249	10.3189 %
8	7878	65.08592 %	1143	9.44316 %
9	1	0.00826 %	889	7.34468 %
10	10	0.08262 %	139	1.14838 %
11	5	0.04131 %	582	4.80833 %
12	1	0.00826 %	325	2.68506 %
13	5	0.04131 %	110	0.90879 %
14	35	0.28916 %	1698	14.02842 %
15	1	0.00826 %	1828	15.10245 %

Πίνακας 4.5: Αποτελέσματα K-means και Mini Batch K-means στο σύνολο των δεδομένων cluster για n=15.

Στο Διάγραμμα 4.15 συγκρίνονται τα ποσοστά πλήρωσης των ομάδων από την εφαρμογή του K-means με αυτά που εξάγονται από τον Mini Batch K-means για τη συνολική κατηγοριοποίηση όλων των καταναλωτών.



Διάγραμμα 4.15: Ποσοστά επί του συνόλου του πλήθους τυπικών προφίλ ανά ομάδα χρησιμοποιώντας: K-means (μπλε) και Mini Batch K-means (πράσινο) για n=15.

Το συμπέρασμα που προκύπτει είναι ότι η καθοδηγούμενη (guided) παραλλαγή του αλγόριθμου Mini Batch K-means δίνει τα καλύτερα αποτελέσματα συγκριτικά με όλες τις δοκιμές που διεξήχθησαν με στόχο την πιο ομοιογενή κατανομή του πλήθους των συνολικών τυπικών προφίλ σε clusters, ώστε να αξιοποιηθούν κατάλληλα για την υλοποίηση της βραχυχρόνιας πρόβλεψης.

Κεφάλαιο 5 :
Υλοποίηση βραχυπρόθεσμης πρόβλεψης

5.1 Συλλογή και επεξεργασία των δεδομένων

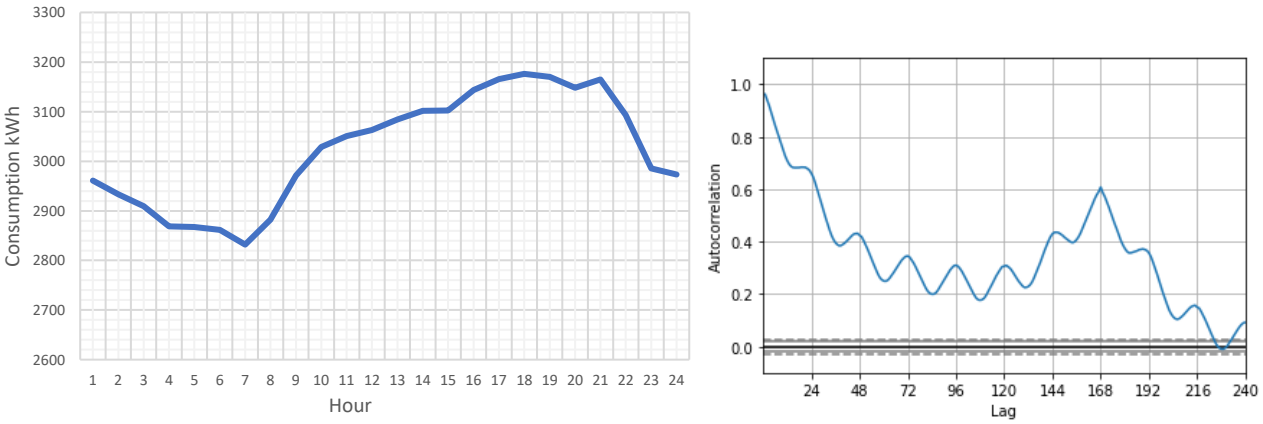
Η κατηγοριοποίηση που υλοποιήθηκε και περιεγράφηκε στο προηγούμενο κεφάλαιο ταξινόμησε τα τυπικά προφίλ των καταναλωτών σε 15 ξεχωριστές ομάδες. Για κάθε πελάτη του κάθε cluster λήφθηκε η χρονοσειρά των ωριαίων μετρήσεων κατανάλωσής του για τη χρονική περίοδο από 01/01/2019 έως και 31/12/2019. Η αρχική επεξεργασία συνίσταται στη διαγραφή της 25ης ώρας που χρησιμοποιείται στις μετρήσεις για την ημέρα που αλλάζει η ώρα σε χειμερινή και στην αντίστοιχη εισαγωγή του μέσου όρου μεταξύ των ωρών που συμβαίνει η θερινή αλλαγή, η οποία για το έτος 2019 ήταν στις 31/03. Έτσι, προέκυψαν 8.760 μετρήσεις για καθένα από τους 12.104 καταναλωτές, από τα οποία για κάθε cluster υπολογίστηκε ο μέσος όρος των μετρήσεων για κάθε ώρα του 2019. Έτσι, η μέση χρονοσειρά θεωρείται ότι εκφράζει την ομάδα από την οποία προκύπτει και καθίσταται ως το dataset, που θα χρησιμοποιηθεί στο κάθε μοντέλο που κατασκευάζεται ανά cluster.

Η αυτοσυσχέτιση είναι η συσχέτιση μεταξύ των τιμών ενός σήματος με τις τιμές του ίδιου σήματος σε διαδοχικές χρονικές περιόδους. Η διαδικασία της αυτοσυσχέτισης παρέχει ένα μέτρο της ομοιότητας ή συμφωνίας μεταξύ ενός δοσμένου σήματος και ενός αντίγραφου του σήματος καθυστερημένου κατά μία μεταβλητή χρονική ποσότητα. Η συνάρτηση αυτοσυσχέτισης ενός σήματος υπολογίζεται ως εξής:

$$R_g(\tau) = \int_{-\infty}^{\infty} g(t)\, g^*(t-\tau)dt \text{ , όπου τ είναι η καθυστέρηση χρόνου} \qquad (5.1)$$

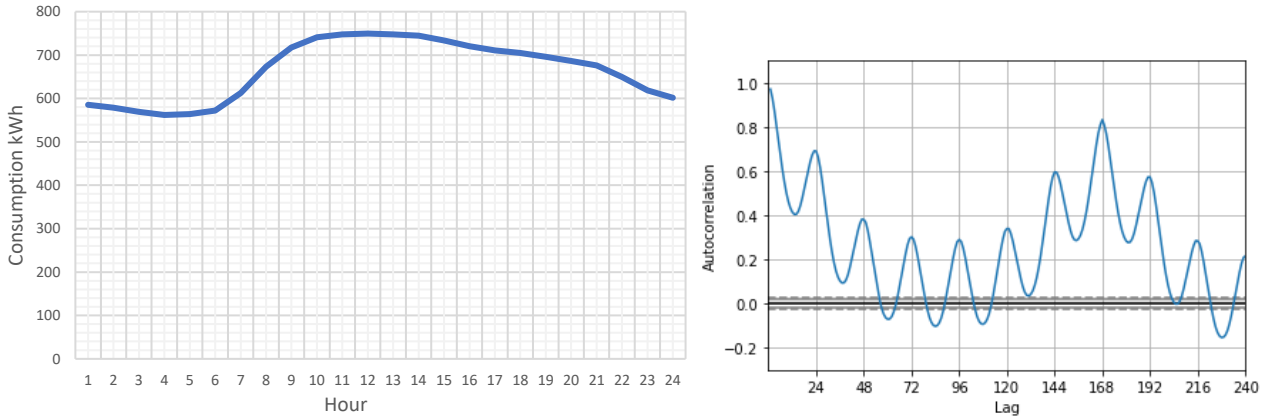
Στα διαγράμματα 5.1 έως 5.15 παρουσιάζονται τα τυπικά προφίλ και η αυτοσυσχέτιση της χρονοσειράς των ωριαίων μετρήσεων ανά cluster για το 2019.

Cluster 1



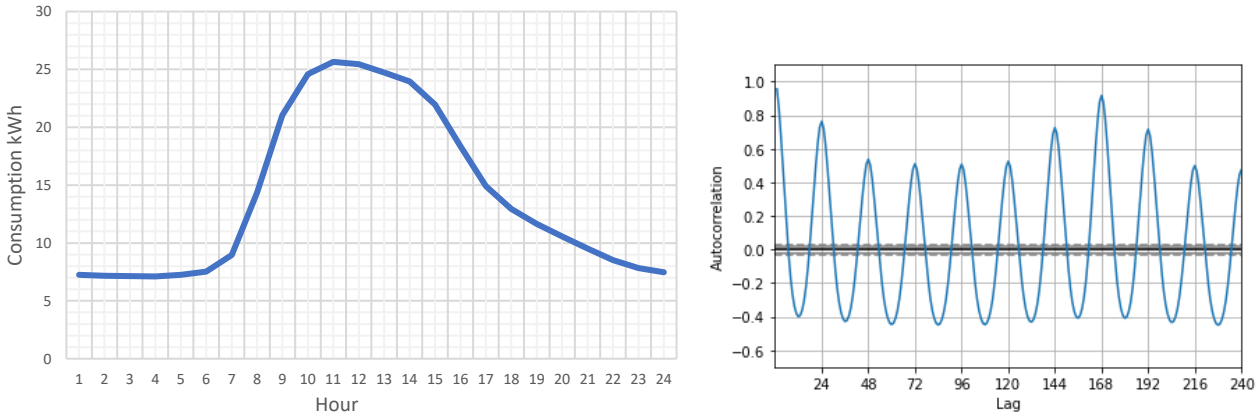
Διάγραμμα 5.1: a) Τυπικό προφίλ κατανάλωσης και **b)** αυτοσυσχέτιση της χρονοσειράς του για το cluster 1.

Cluster 2



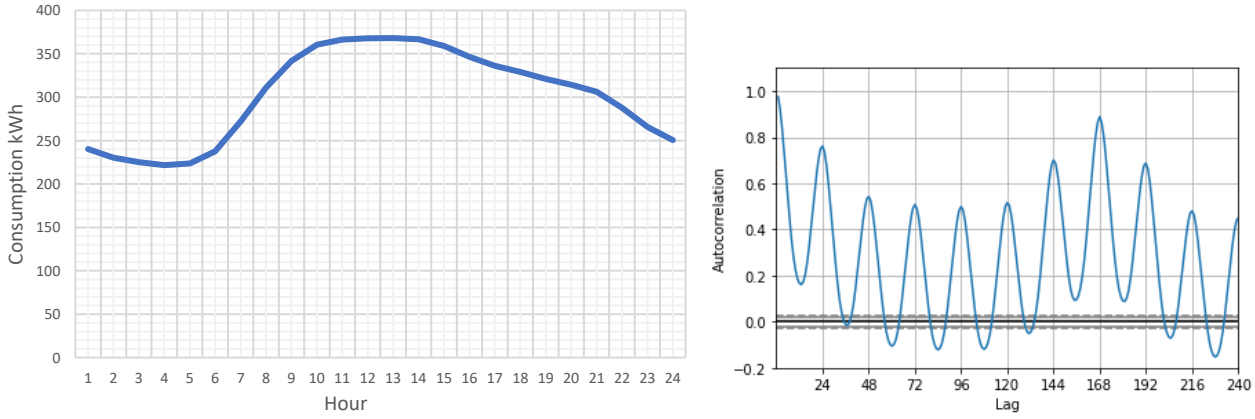
Διάγραμμα 5.2: a) Τυπικό προφίλ κατανάλωσης και **b)** αυτοσυσχέτιση της χρονοσειράς του για το cluster 2.

Cluster 3



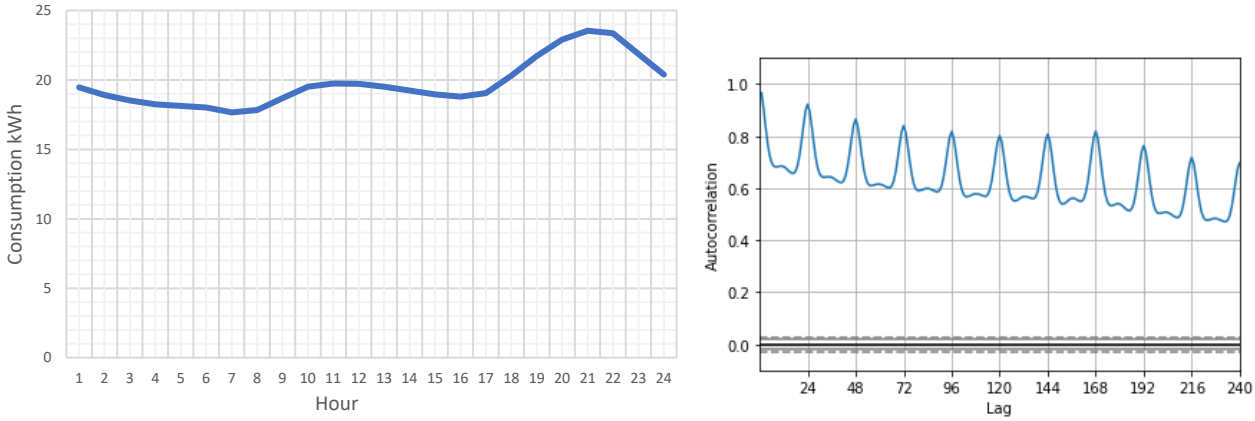
Διάγραμμα 5.3: a) Τυπικό προφίλ κατανάλωσης και **b)** αυτοσυσχέτιση της χρονοσειράς του για το cluster 3.

Cluster 4



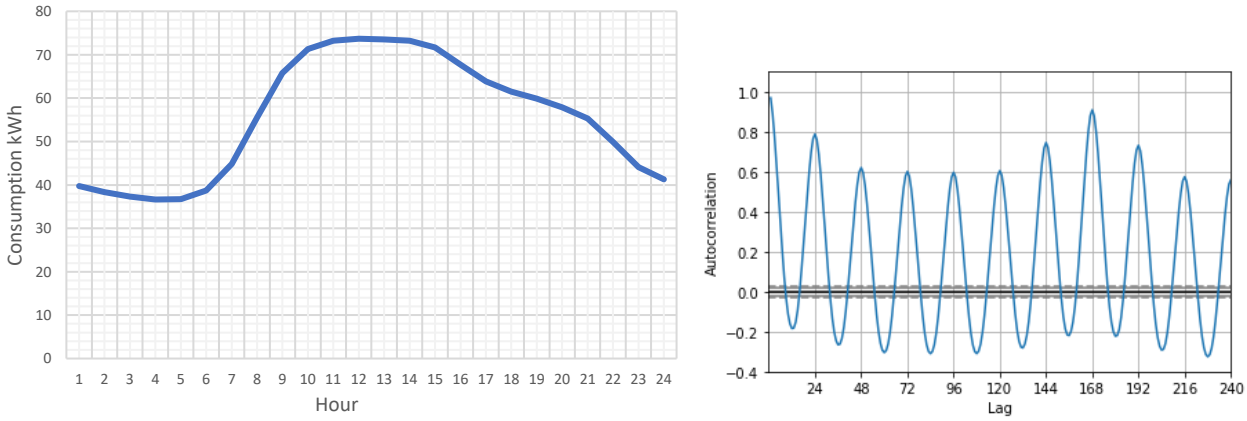
Διάγραμμα 5.4: a) Τυπικό προφίλ κατανάλωσης και **b)** αυτοσυσχέτιση της χρονοσειράς του για το cluster 4.

Cluster 5



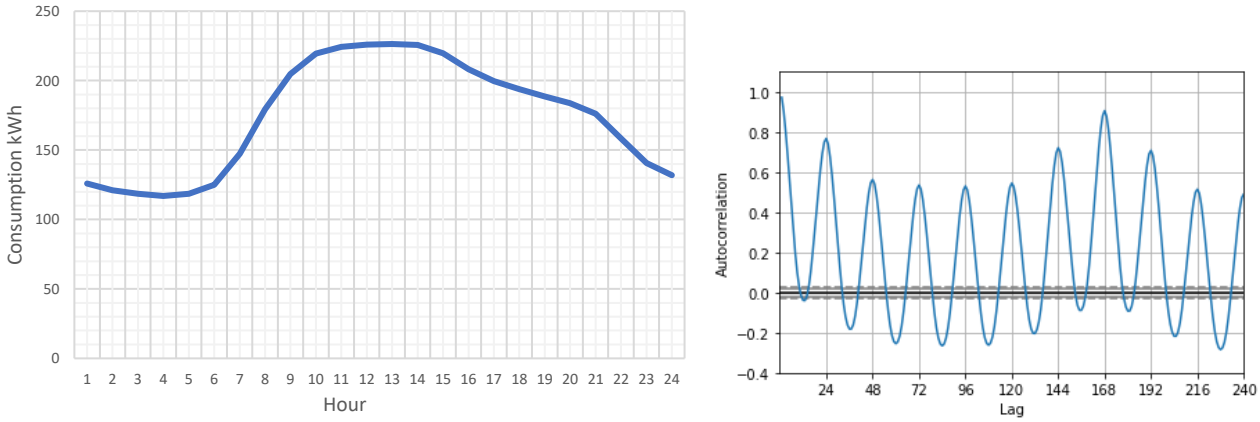
Διάγραμμα 5.5: a) Τυπικό προφίλ κατανάλωσης και **b)** αυτοσυσχέτιση της χρονοσειράς του για το cluster 5.

Cluster 6



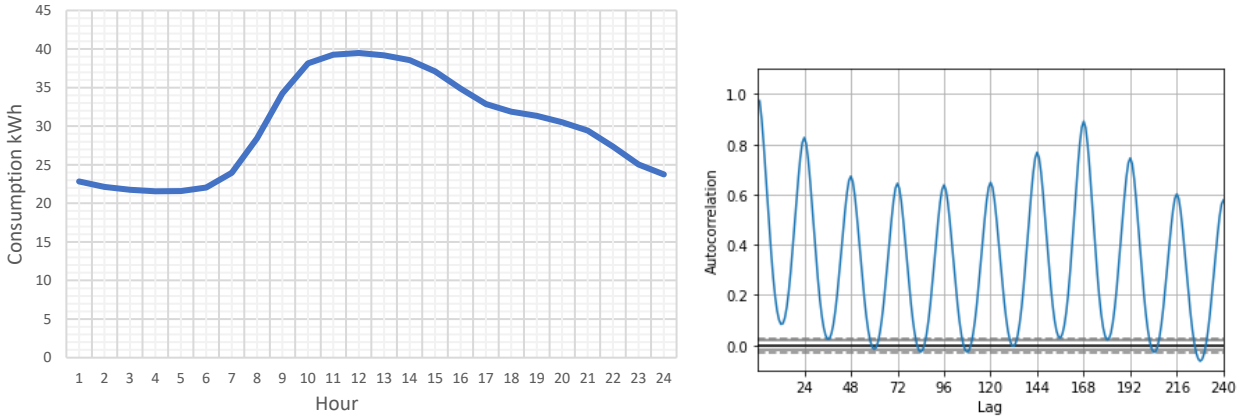
Διάγραμμα 5.6: a) Τυπικό προφίλ κατανάλωσης και **b)** αυτοσυσχέτιση της χρονοσειράς του για το cluster 6.

Cluster 7



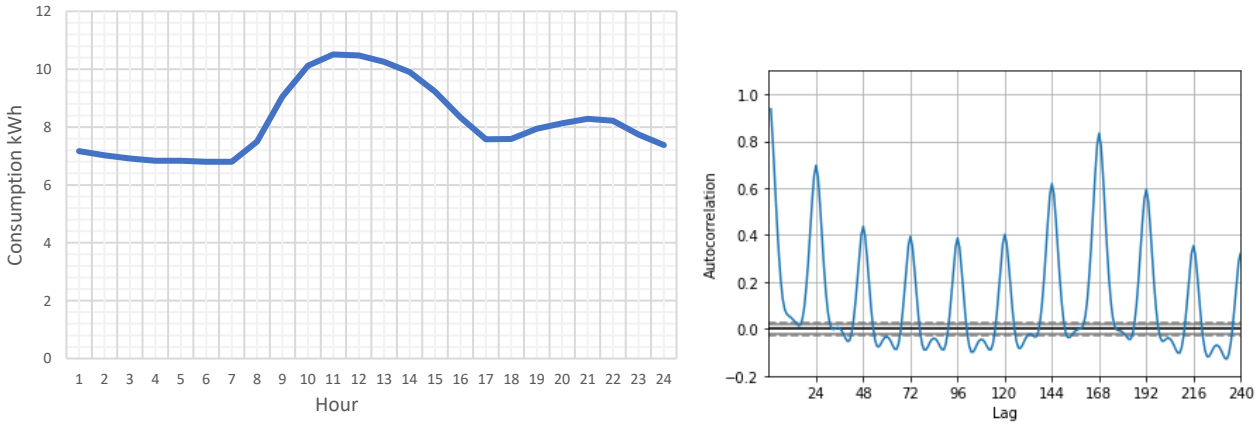
Διάγραμμα 5.7: a) Τυπικό προφίλ κατανάλωσης και **b)** αυτοσυσχέτιση της χρονοσειράς του για το cluster 7.

Cluster 8



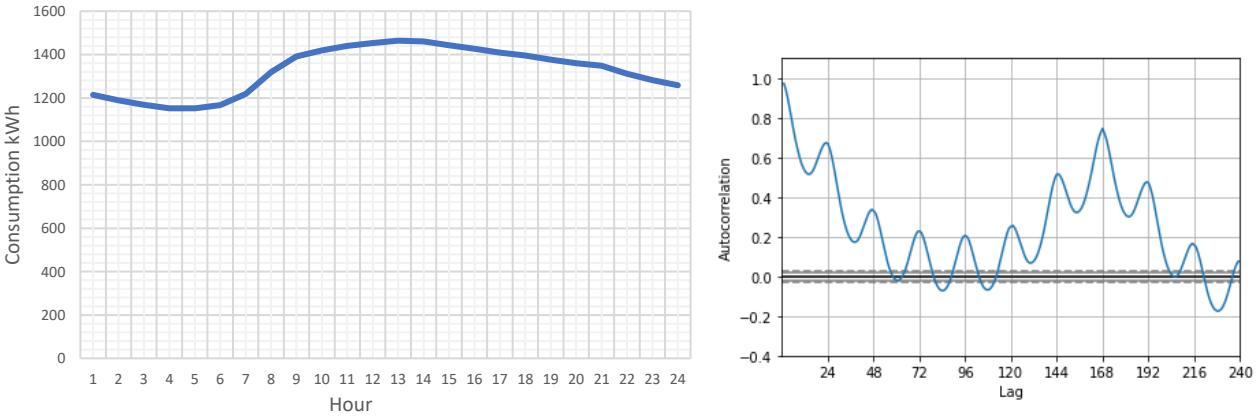
Διάγραμμα 5.8: a) Τυπικό προφίλ κατανάλωσης και **b)** αυτοσυσχέτιση της χρονοσειράς του για το cluster 8.

Cluster 9



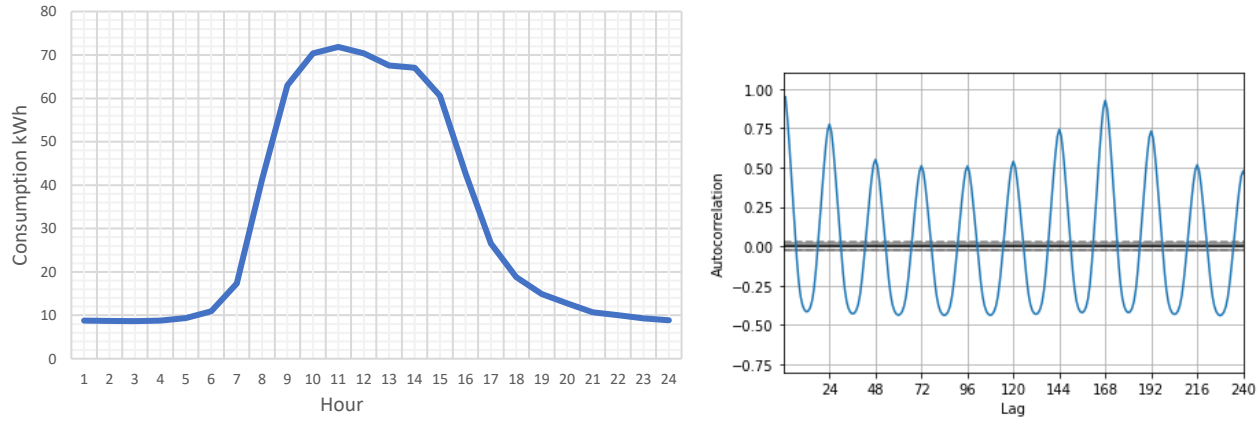
Διάγραμμα 5.9: a) Τυπικό προφίλ κατανάλωσης και **b)** αυτοσυσχέτιση της χρονοσειράς του για το cluster 9.

Cluster 10



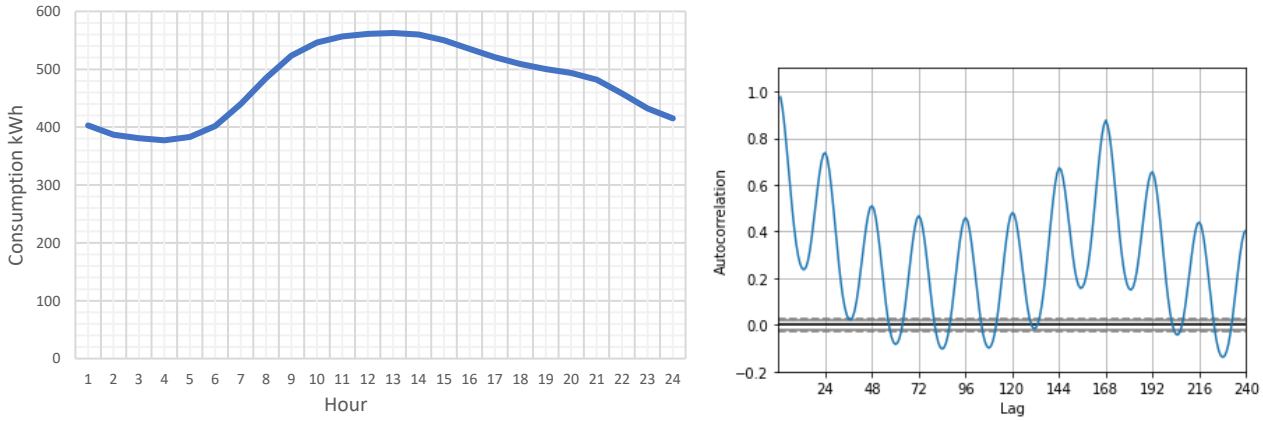
Διάγραμμα 5.10: a) Τυπικό προφίλ κατανάλωσης και **b)** αυτοσυσχέτιση της χρονοσειράς του για το cluster 10.

Cluster 11



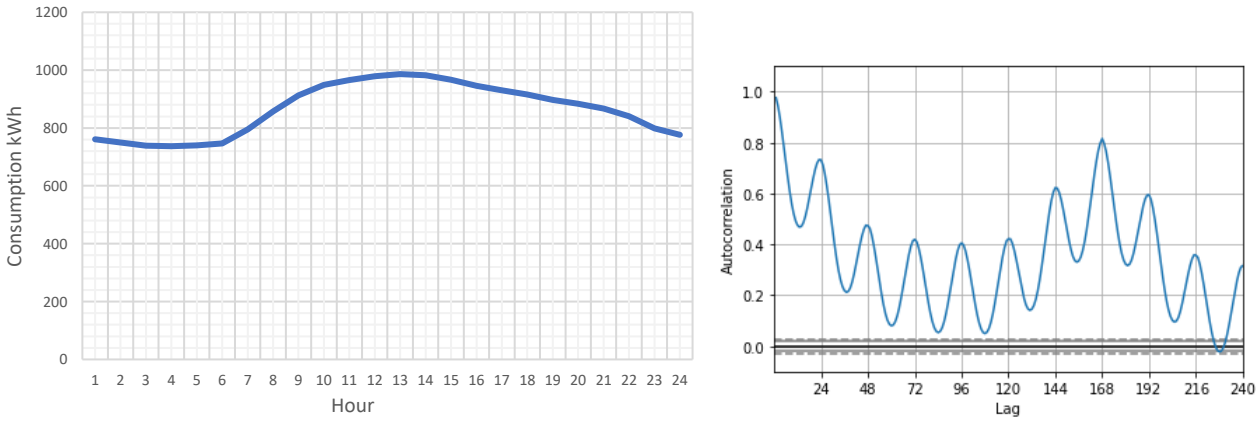
Διάγραμμα 5.11: a) Τυπικό προφίλ κατανάλωσης και **b)** αυτοσυσχέτιση της χρονοσειράς του για το cluster 11.

Cluster 12



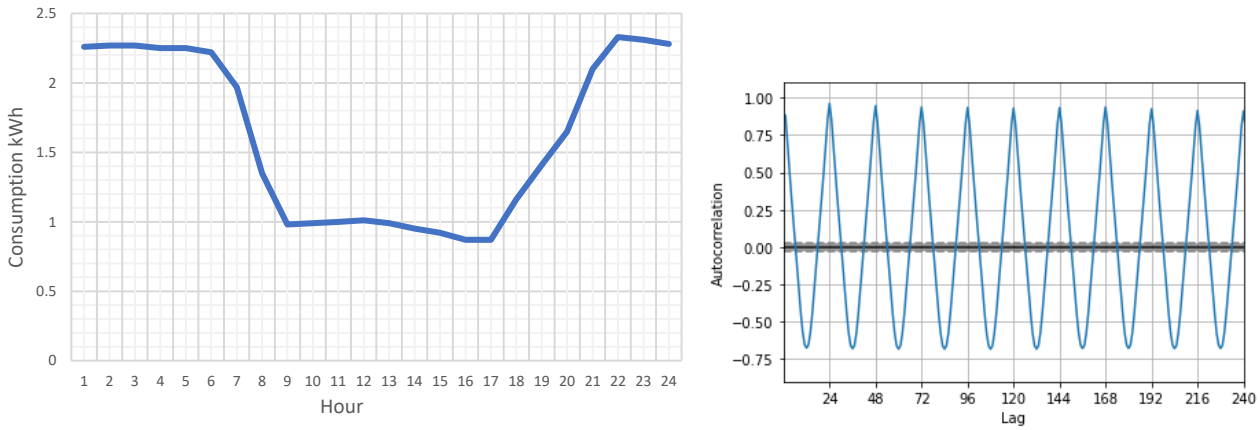
Διάγραμμα 5.12: a) Τυπικό προφίλ κατανάλωσης και **b)** αυτοσυσχέτιση της χρονοσειράς του για το cluster 12.

Cluster 13



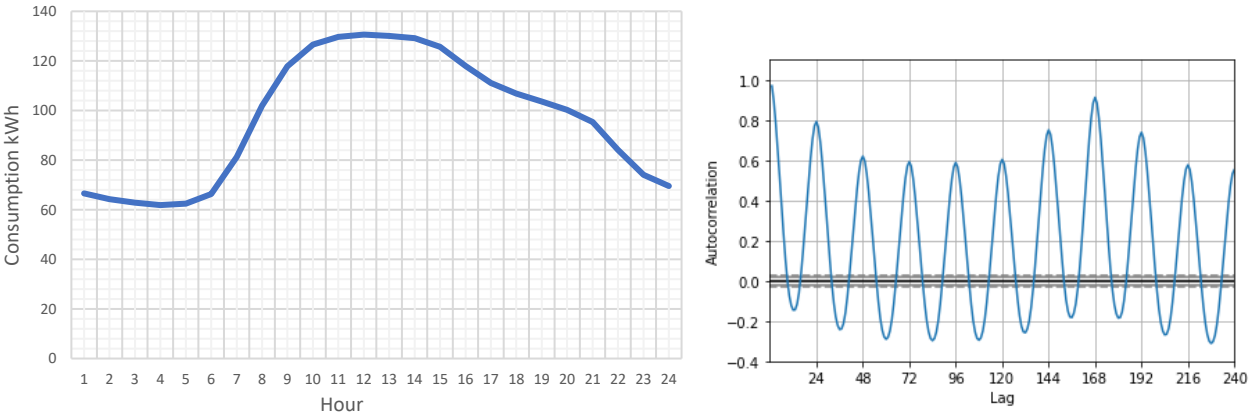
Διάγραμμα 5.13: a) Τυπικό προφίλ κατανάλωσης και **b)** αυτοσυσχέτιση της χρονοσειράς του για το cluster 13.

Cluster 14



Διάγραμμα 5.14: a) Τυπικό προφίλ κατανάλωσης και **b)** αυτοσυσχέτιση της χρονοσειράς του για το cluster 14.

Cluster 15



Διάγραμμα 5.15: a) Τυπικό προφίλ κατανάλωσης και **b)** αυτοσυσχέτιση της χρονοσειράς του για το cluster 15.

Τελευταίο βήμα σε αυτό το στάδιο αποτελεί η κανονικοποίηση των δεδομένων σε τιμές από 0 έως 1 για την εισαγωγή τους στο νευρωνικό δίκτυο.

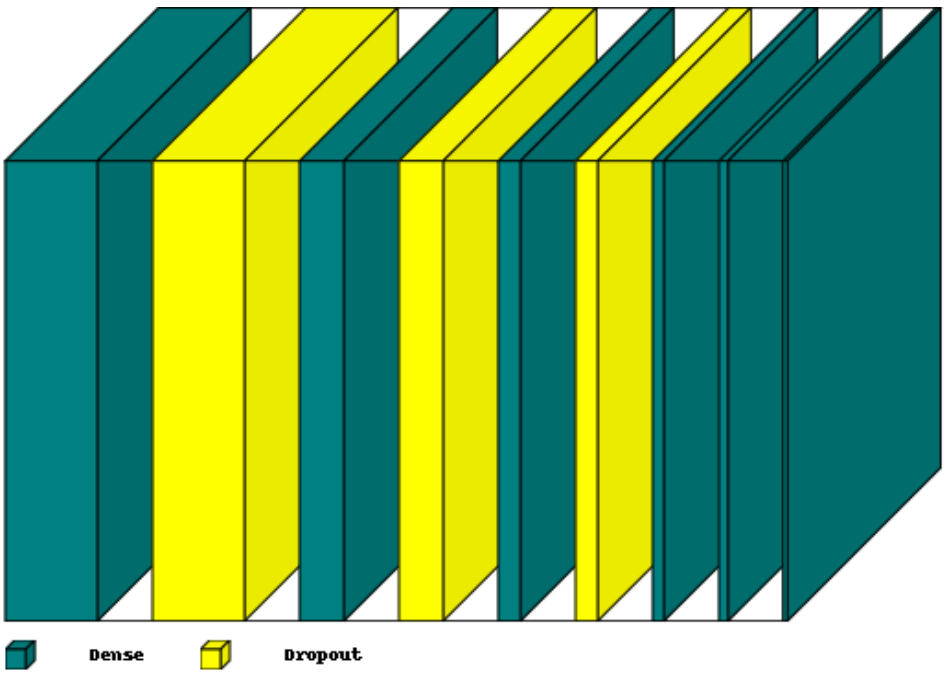
5.2 Περιγραφή μοντέλου

Η υλοποίηση της βραχυπρόθεσμης πρόβλεψης των πελατών υλοποιείται μέσω μοντέλου που προγραμματίστηκε με Python στο Google Colab. Το σύνολο των δεδομένων χωρίζεται σε δύο τμήματα, το τμήμα εκπαίδευσης (Train Set) και το τμήμα επαλήθευσης (Test Set). Το νευρωνικό δίκτυο που υλοποιήθηκε αποτελείται από διαδοχικά Dense επίπεδα (layers) νευρώνων που το καθένα αποτελείται από 500,240,120,60,40 και το εξόδου από 24 νευρώνες. Σε κάθε layer οι νευρώνες λαμβάνουν ως είσοδο τις εξόδους όλων όσων ανήκουν στο προηγούμενο. Στο παρασκήνιο ο καθένας εκτελεί πολλαπλασιασμό με μία μήτρα, οι τιμές της οποίας αποτελούν τις παραμέτρους που εκπαιδεύονται σε κάθε εποχή [40]. Η συνάρτηση ενεργοποίησης που χρησιμοποιείται είναι η Exponential Linear Unit (ELU) για όλα τα επίπεδα εκτός του εξόδου που είναι γραμμική και η απόκριση που αξιολογείται από το μέσο απόλυτο ποσοστιαίο σφάλμα της επαλήθευσης.

Επίσης, μεταξύ των στρωμάτων χρησιμοποιούνται Dropout layers, τα οποία έχουν μία πιθανότητα που ορίζεται ως παράμετρος να αγνοήσουν εισερχόμενες μονάδες κατά τη διάρκεια της εκπαίδευσης. Η τεχνική αυτή χρησιμοποιείται για την αποφυγή του overfitting, δηλαδή το νευρωνικό δίκτυο δεν γενικεύει για όλα τα δεδομένα, αλλά απομνημονεύει τμήματα του Train Set και αντιμετωπίζει το Test Set ως αυτά. Αυτή η κατάσταση φαίνεται όταν το σφάλμα εκπαίδευσης (Train Loss) είναι μικρότερο του σφάλματος πρόβλεψης (Validation Loss). Επιπλέον, επιταχύνει τη διαδικασία της εκπαίδευσης, καθώς όταν ενεργοποιείται το Dropout απαιτούνται λιγότερα βάρη σε κάθε πέρασμα [42], [43].

Μία ακόμα παράμετρος που ρυθμίζεται είναι του παραθύρου που "σκανάρει" τα δεδομένα στο μοντέλο. Επιλέγεται look_back = 24, καθώς η αυτοσυσχέτιση των περισσότερων από τις χρονοσειρές εκεί έχει τιμή μεγαλύτερη του 0,5, που θεωρείται ικανοποιητική. Ακόμη, ορίζεται ο ορίζοντας της πρόβλεψης στις 24 ώρες και άρα η αντίστοιχη μεταβλητή είναι horizon=24 και οι εποχές που εκτελείται το μοντέλο στις 4.000.

Τέλος, χρησιμοποιείται η μεταβλητή call_back, η οποία σταματά τη διεξαγωγή περαιτέρω επαναλήψεων του μοντέλου και αποθηκεύει τους συντελεστές βαρών του όταν αυτό κριθεί απαραίτητο, δηλαδή όταν το σφάλμα επαλήθευσης αρχίζει να αυξάνεται [43].



Διάγραμμα 5.16: Αρχιτεκτονική του μοντέλου πρόβλεψης.

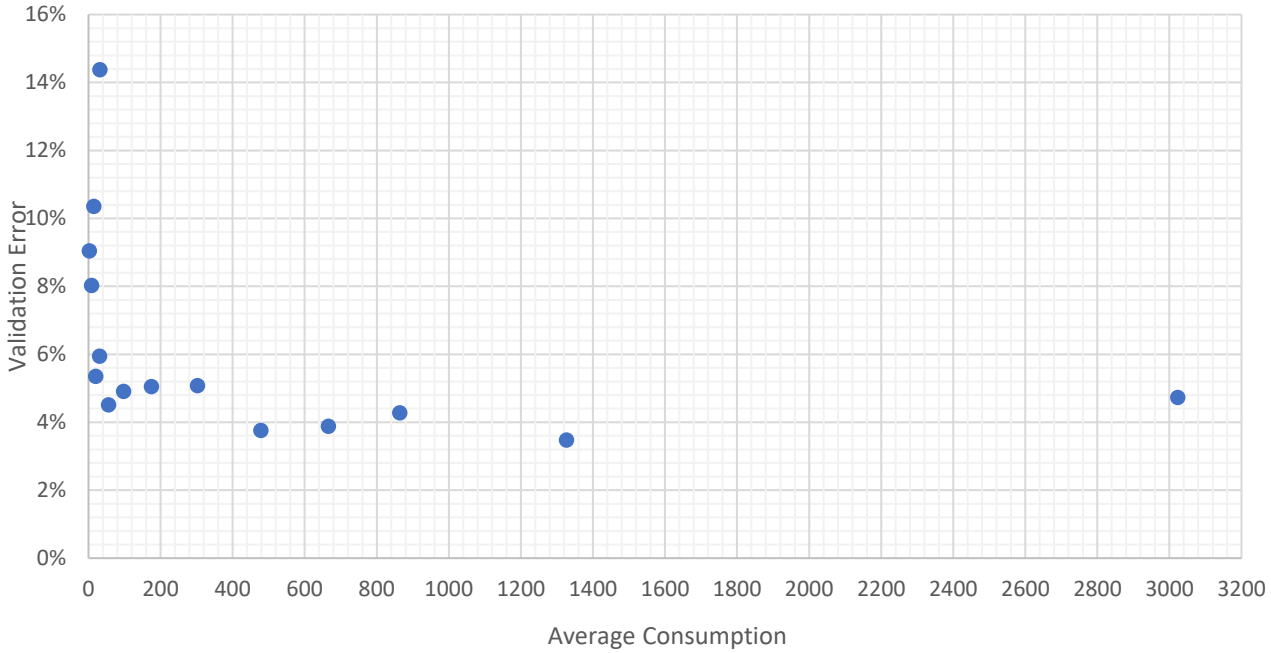
5.3 Αποτελέσματα πρόβλεψης

Τα αποτελέσματα που πάρθηκαν για την πρόβλεψη, δηλαδή το ποσοστό σφάλματος επαλήθευσης για κάθε cluster παρουσιάζονται στον Πίνακα 5.1.

	Αποτελέσματα Clustering		**Στοιχεία Χρονοσειράς**		**Αποτελέσματα Forecasting**
Cluster	**Αριθμός τυπικών προφίλ**	**Ποσοστό επί του συνόλου προφίλ**	**Min (kW)**	**Max (kW)**	**Σφάλμα πρόβλεψης**
1	91	0.75182 %	1482.39	4096.85	4.73352 %
2	115	0.9501 %	341.97	1079.65	3.88699 %
3	744	6.14673 %	5.24	44.25	10.36061 %
4	605	4.99835 %	139.25	560.95	5.08289 %
5	454	3.75083 %	11.36	37.18	5.35566 %
6	2132	17.61401 %	28.50	118.83	4.51396 %
7	1249	10.3189 %	77.40	352.87	5.05187 %
8	1143	9.44316 %	14.56	70.11	5.95071 %
9	889	7.34468 %	4.31	16.1	8.02943 %
10	139	1.14838 %	572.74	1997.56	3.47878 %
11	582	4.80833 %	6.27	130.13	14.38416 %
12	325	2.68506 %	239.69	819.05	3.76750 %
13	110	0.90879 %	383.30	1420.64	4.28375 %
14	1698	14.02842 %	0.55	4.26	9.04613 %
15	1828	15.10245 %	44.77	208.19	4.90974 %

Πίνακας 5.1: Αποτελέσματα βραχυπρόθεσμης πρόβλεψης για κάθε cluster.

Στο διάγραμμα 5.16 σχεδιάζεται το σφάλμα σε σχέση με τη μέση τιμή της ενέργειας της κατανάλωσης, όπου παρατηρείται ότι όσο χαμηλότερη είναι η κατανάλωση, τόσο αυξάνεται το σφάλμα πρόβλεψης και αντίστροφα.



Διάγραμμα 5.17: Ποσοστιαίο λάθος πρόβλεψης σχετικά με τη μέση κατανάλωση του τυπικού προφίλ κάθε cluster.

Τα αναλυτικά αποτελέσματα των μοντέλων παρουσιάζονται στο Παράρτημα Π.5 – Π. 19, στα οποία σχεδιάζονται τα σφάλματα για το Train Set (Loss) και για το Test Set (Val_Loss) που υπολογίζονται σε κάθε εποχή.

Κεφάλαιο 6:
Συμπεράσματα και Μελλοντική Έρευνα

6.1 Συμπεράσματα

Στη διπλωματική εργασία που παρουσιάστηκε υλοποιήθηκε η κατηγοριοποίηση και η βραχυπρόθεσμη πρόβλεψη φορτίου των καταναλωτών Μέσης Τάσης της Ελληνικής Αγοράς Ηλεκτρικής Ενέργειας του διασυνδεδεμένου δικτύου. Επίσης, προσέφερε λύση στην κατηγοριοποίηση σε clusters τεράστιου όγκου και μεγάλης ανομοιογένειας δεδομένων που αποτελούν οι ωριαίες μετρήσεις των πελατών. Αυτήν ακολούθησε και η βραχυπρόθεσμη πρόβλεψη φορτίου τους με νευρωνικό δίκτυο.

Τα δεδομένα δόθηκαν από τον Διαχειριστή Ελληνικού Δικτύου Διανομής Ηλεκτρικής Ενέργειας μέσω της Διεύθυνσης Χρηστών Δικτύου και διαμορφώθηκαν κατάλληλα, όπως παρουσιάζεται στο Κεφάλαιο 4.1. Το imbalanced dataset που αποτελούν τα ημερήσια τυπικά προφίλ των πελατών, αντιμετωπίστηκε με την εξέταση πολλών σεναρίων εφαρμογής του αλγόριθμου K-means και των παραλλαγών του. Τα συμπεράσματα που προέκυψαν για κάθε σενάριο είναι:

1. Η εφαρμογή του K-means για τον προτεινόμενο αριθμό cluster (n=5) που υπολογίστηκε από τους δείκτες αξιολόγησης αθροίσματος τετραγωνικού λάθους (Sum of Square Error – SSE), Silhouette Score, Davies Bouldin δεν ήταν ικανή να ταξινομήσει τα τυπικά προφίλ που αφορούν χαμηλές καταναλώσεις σε ομάδες, αλλά τα κατέτασσε σε μία. Έτσι, το συγκεκριμένο cluster είχε το 85,5 % των δεδομένων, γεγονός που δεν ήταν αποδεκτό ως προς τη διάκριση των καταναλωτών βάσει των χαρακτηριστικών τους. (Κεφ. 4.3.1)
2. Η μείωση διαστάσεων από 24 σε 12 με την εφαρμογή την Ανάλυσης σε Κύριες Συνιστώσες (PCA) δεν προσέφερε όφελος στο παραπάνω πρόβλημα, καθώς τα αποτελέσματα ήταν σχεδόν ίσα με το 1.(Κεφ. 4.3.2)
3. Δοκιμάστηκε η αύξηση του αριθμού των cluster σε n= 8 και n=15, χωρίς όμως να δώσει λύση, καθώς σχηματίστηκε και πάλι ομάδα που περιείχε το 66,8 % και 47.6 % των πελατών αντίστοιχα. Παρόμοια εικόνα δίνει και η εφαρμογή του PCA για το ίδιο n. (Κεφ. 4.3.2)
4. Εφαρμόστηκε ο αλγόριθμος Mini Batch K-means για 8 clusters δίνοντας βελτιωμένα αποτελέσματα, αφού σε 24 και σε 12 διαστάσεις το πολυπληθέστερο περιείχε το 42,95 % των δεδομένων.(Κεφ. 4.4)
5. Υλοποιήθηκε καθοδηγούμενη κατηγοριοποίηση (guided clustering) σε στάδια. Στο πρώτο εκτελείται ο αλγόριθμος του K-means για n=8 και απομονώνεται το πολυπληθέστερο cluster που προκύπτει, το οποίο περιέχει 8.083 καταναλωτές. Στο δεύτερο στάδιο εφαρμόζεται εκ νέου ο αλγόριθμος για τις τυπικές καταναλώσεις αυτών των πελατών της συγκεκριμένης ομάδας. Τέλος, το σύνολο των δεδομένων ταξινομείται, όπου τα κέντρα του πρώτου σταδίου, εκτός από αυτό που αφορά την πολυπληθέστερη ομάδα και του δεύτερου χρησιμοποιούνται ως αρχικοποίηση, δηλαδή ως ορισμένα εξ αρχής κέντρα για τον αλγόριθμο. Δοκιμάζονται οι αλγόριθμοι K-means και Mini Batch K-means και επιλέγεται ο δεύτερος που δίνει αποτελέσματα που κρίνονται ικανοποιητικά για την κατηγοριοποίηση, καθώς το πλήθος ανά ομάδα δεν είναι τόσο άνισο, όσο όλων των προηγούμενων μεθόδων. (Κεφ. 4.5)

Έτσι, η καθοδηγούμενη κατηγοριοποίηση αντιμετώπισε το imbalanced dataset ταξινομώντας τους 12.104 πελάτες σε clusters που παρουσιάζουν κοινά χαρακτηριστικά και μπορούν πλέον να αξιοποιηθούν για το μοντέλο πρόβλεψης, αυξάνοντας την ακρίβεια της εκτίμησης λόγω των πιο αντιπροσωπευτικών ομάδων.

Η διαμόρφωση των δεδομένων για την τροφοδοσία του μοντέλου πρόβλεψης συνίσταται στη συλλογή των χρονοσειρών που περιέχουν τις ωριαίες μετρήσεις των πελατών ανά cluster για τις 365 ημέρες του 2019. Για κάθε ομάδα υπολογίστηκε η μέση χρονοσειρά, δηλαδή ο μέσος όρος των μετρήσεων ανά ώρα των καταναλωτών που ανήκουν σε αυτή. Τελικό αποτέλεσμα είναι 15 χρονοσειρές που η κάθε μία περιέχει 8.760 ενδείξεις, όπου για την κάθε μία κατασκευάζεται ένα νευρωνικό δίκτυο. Για κάθε χρονοσειρά κατασκευάζεται το ημερήσιο τυπικό προφίλ που εκφράζει το συγκεκριμένο cluster και η αυτοσυσχέτισή του. Τα μοντέλα που κατασκευάστηκαν έχουν ίδια αρχιτεκτονική νευρωνικού δικτύου, όπως αυτή παρουσιάστηκε στο Κεφ. 5.2, για την πρόβλεψη των επόμενων 24 ωρών. Τα αποτελέσματα που εξήχθησαν για κάθε cluster υποδεικνύουν την αύξηση της ακρίβειας της πρόβλεψης για χρονοσειρές με μεγαλύτερη μέση κατανάλωση.

Συμπερασματικά, η εργασία δίνει λύση στο πρόβλημα της υλοποίησης πρόβλεψης όταν δεν είναι γνωστά τα δεδομένα της χρονοσειράς ζήτησης φορτίου ενός καταναλωτή. Η πρόταση που υλοποιήθηκε αφορά την υλοποίηση κατηγοριοποίησης των καταναλωτών με γνωστές χρονοσειρές σε ομάδες και πρόβλεψη φορτίου για κάθε μία εξ αυτών. Το τυπικό προφίλ του άγνωστου πελάτη ταξινομείται σε μία από αυτές, όπου είναι γνωστό το σφάλμα πρόβλεψης που έχει υπολογιστεί για αυτή. Έτσι, ένας πελάτης με γνωστό μόνο το τυπικό ημερήσιο προφίλ φορτίου του μπορεί να καταχωρηθεί στην αντίστοιχη ομάδα που είναι πιο αντιπροσωπευτική για αυτόν και να γίνει εκτίμηση της ζήτησης φορτίου του για τις επόμενες ώρες.

6.2 Μελλοντική Έρευνα

Επόμενα ή εναλλακτικά βήματα που προτείνονται για την αξιοποίηση της υπάρχουσας διπλωματικής αφορούν τα εξής πεδία:

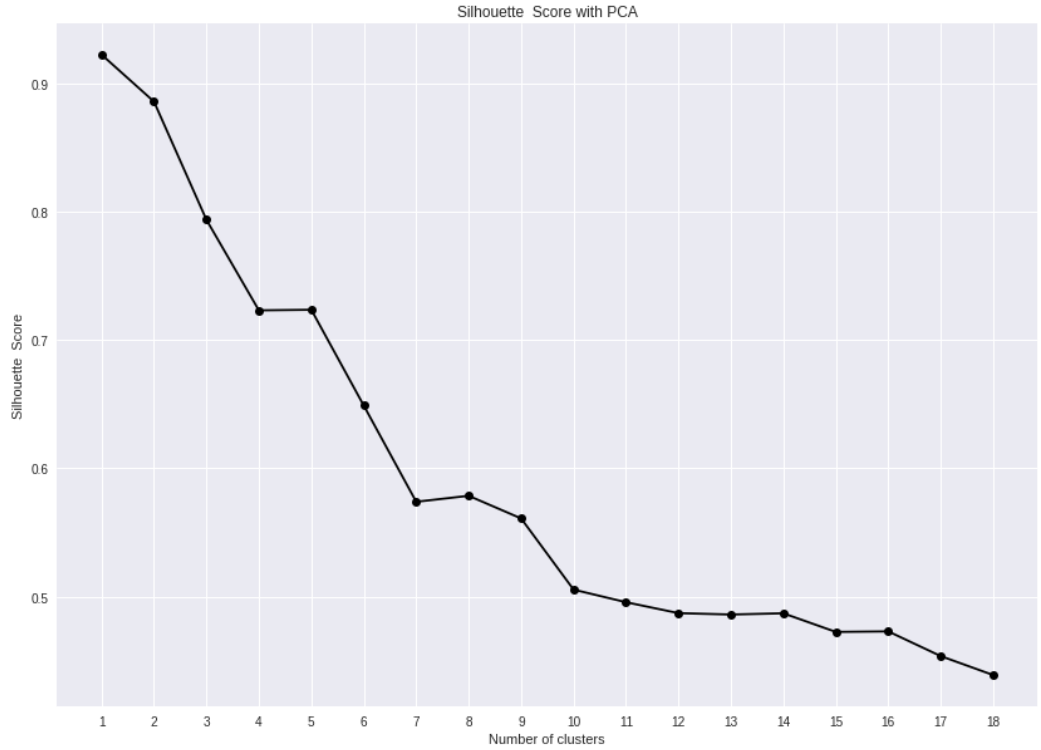
Α) Βελτιστοποίηση της διαδικασίας που υλοποιήθηκε:

- Η κατηγοριοποίηση να υλοποιηθεί με είσοδο τις χρονοσειρές των πελατών και όχι τα τυπικά προφίλ τους.
- Η κατηγοριοποίηση και η πρόβλεψη να παίρνουν υπόψη και άλλες παραμέτρους, όπως η χρήση του καταναλωτή, η εποχικότητα, η επίδραση των καιρικών συνθηκών με τις αντίστοιχες πληροφορίες να παρέχονται ως δεδομένα.
- Να διερευνηθούν και άλλοι αλγόριθμοι κατηγοριοποίησης, όπως αυτοί περιεγράφηκαν στο Κεφ. 2.
- Τα δεδομένα που τροφοδοτούνται στο μοντέλο πρόβλεψης να μην είναι το μέσο τυπικό προφίλ αλλά ένα σταθμισμένο μέσο προφίλ που εξάγεται ανά cluster.
- Να διερευνηθούν και άλλοι τρόποι βραχυπρόθεσμης πρόβλεψης, ιδιαίτερα τα νευρωνικά δίκτυα με διαφορετική αρχιτεκτονική, πχ μονοδιάστατα συνελικτικά.

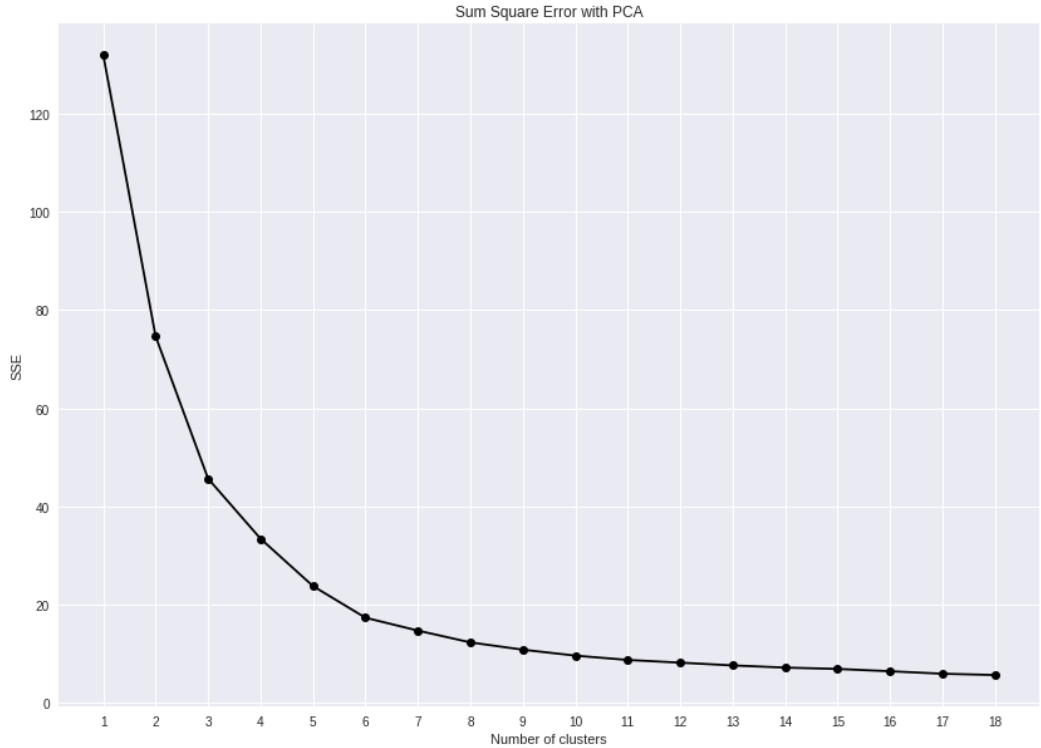
Β) Διερεύνηση της προσφορά της κατηγοριοποίησης όταν προηγείται της πρόβλεψης με τους εξής τρόπους:

- Ποσοτικοποίηση των σφαλμάτων για πελάτες ανά cluster και σύγκρισή τους με τα αντίστοιχα αποτελέσματα που θα εξήγαγε η πρόβλεψη στο σύνολο των δεδομένων χωρίς clustering.
- Τα δεδομένα που τροφοδοτούνται στο κάθε μοντέλο πρόβλεψης να αποτελούνται από την αθροιστική χρονοσειρά των πελατών του cluster και να συγκριθούν τα αποτελέσματα με την αθροιστική χρονοσειρά όλων των πελατών.

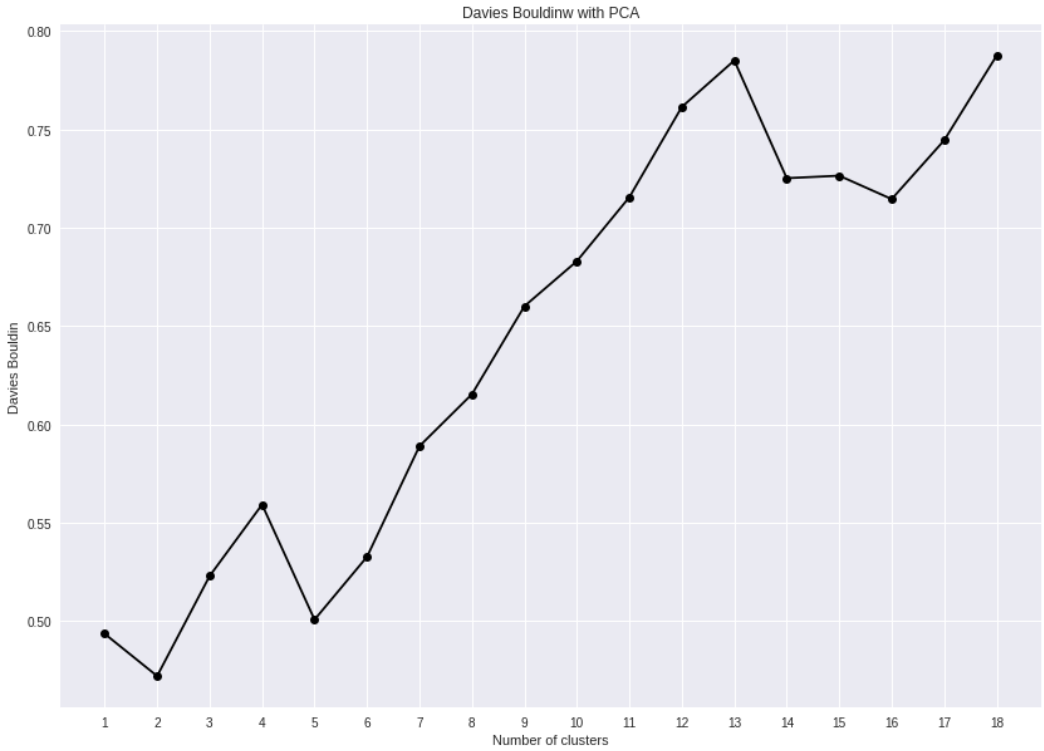
Παράρτημα



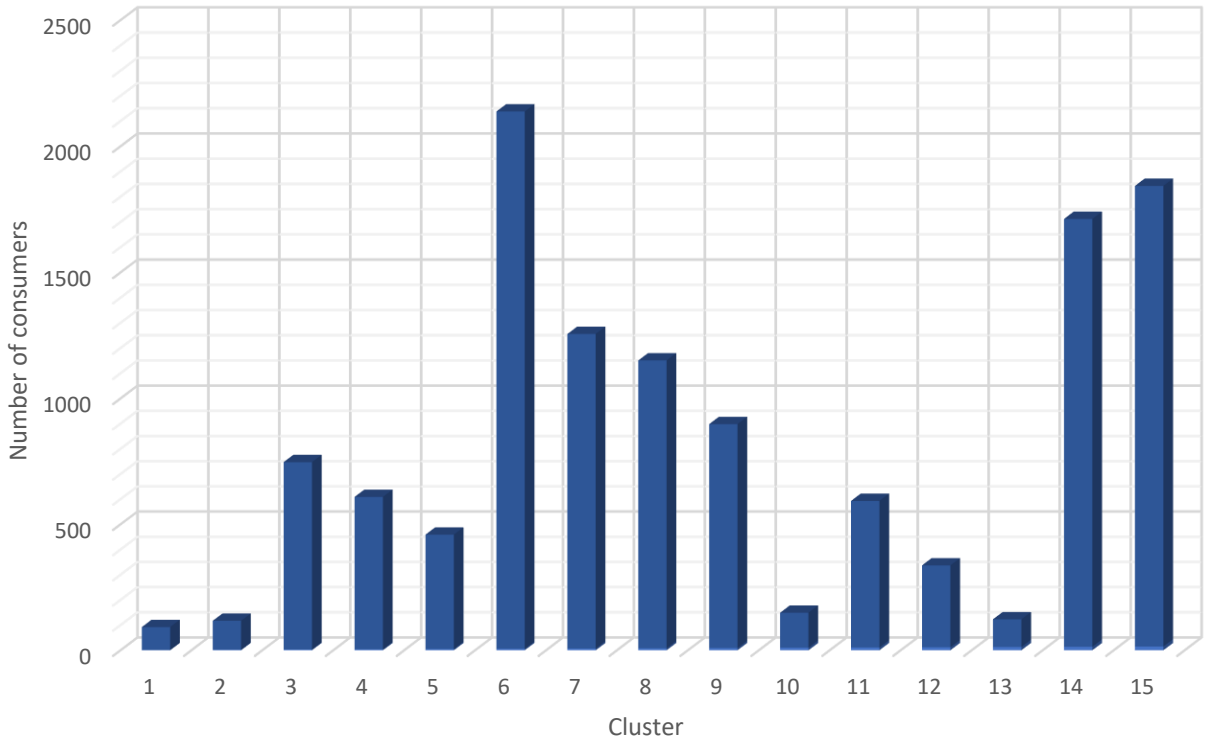
Π.1: Δείκτης αξιολόγησης Silhouette Score για αριθμό cluster n = 2 έως 20 και 12 διαστάσεις μετά την εφαρμογή του PCA. **(**Κεφάλαιο 4.3)



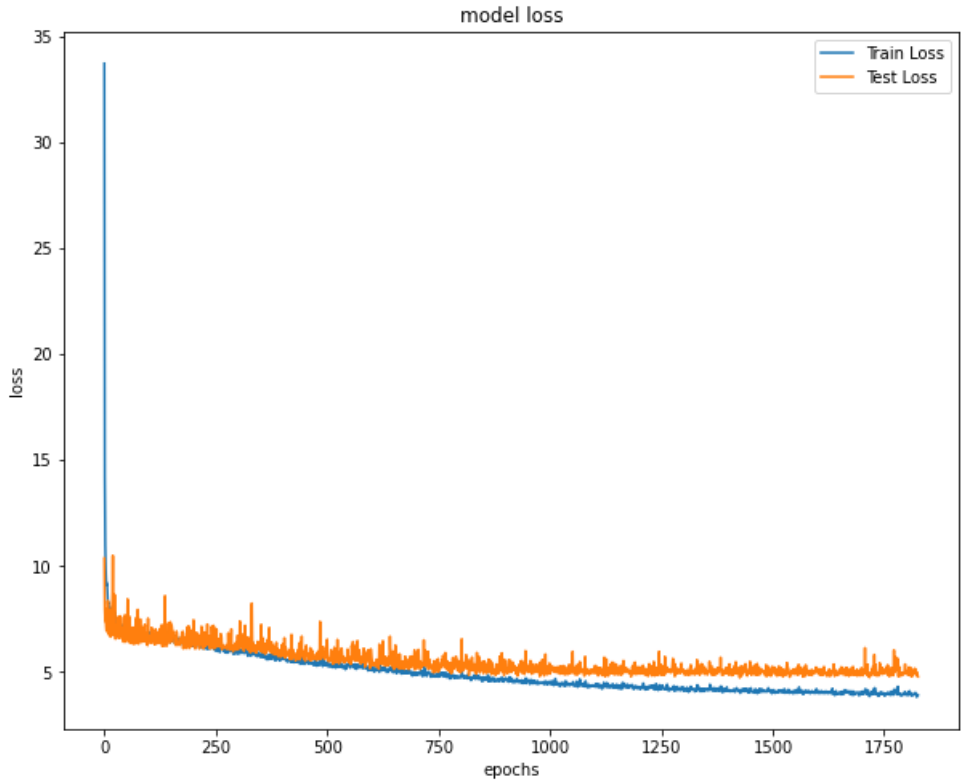
Π.2: Δείκτης αξιολόγησης Sum of Square Error – SSE για αριθμό cluster n = 2 έως 20 και 12 διαστάσεις μετά την εφαρμογή του PCA. **(**Κεφάλαιο 4.3)



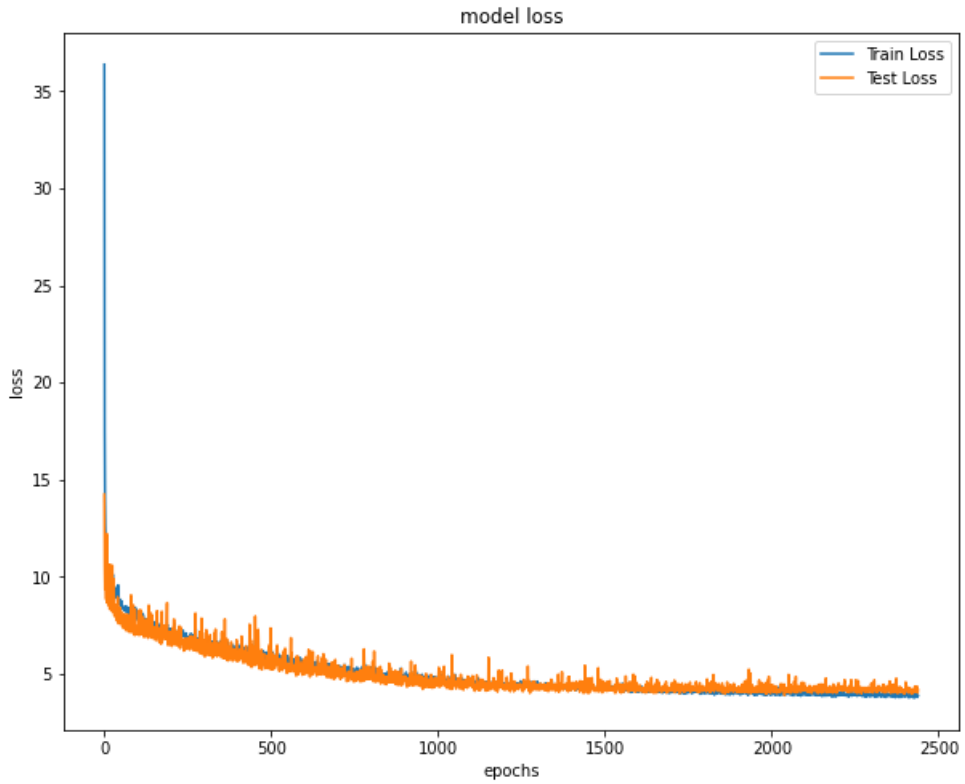
Π.3: Δείκτης αξιολόγησης Davies Bouldin για αριθμό cluster n = 2 έως 20 και 12 διαστάσεις μετά την εφαρμογή του PCA. **(**Κεφάλαιο 4.3)



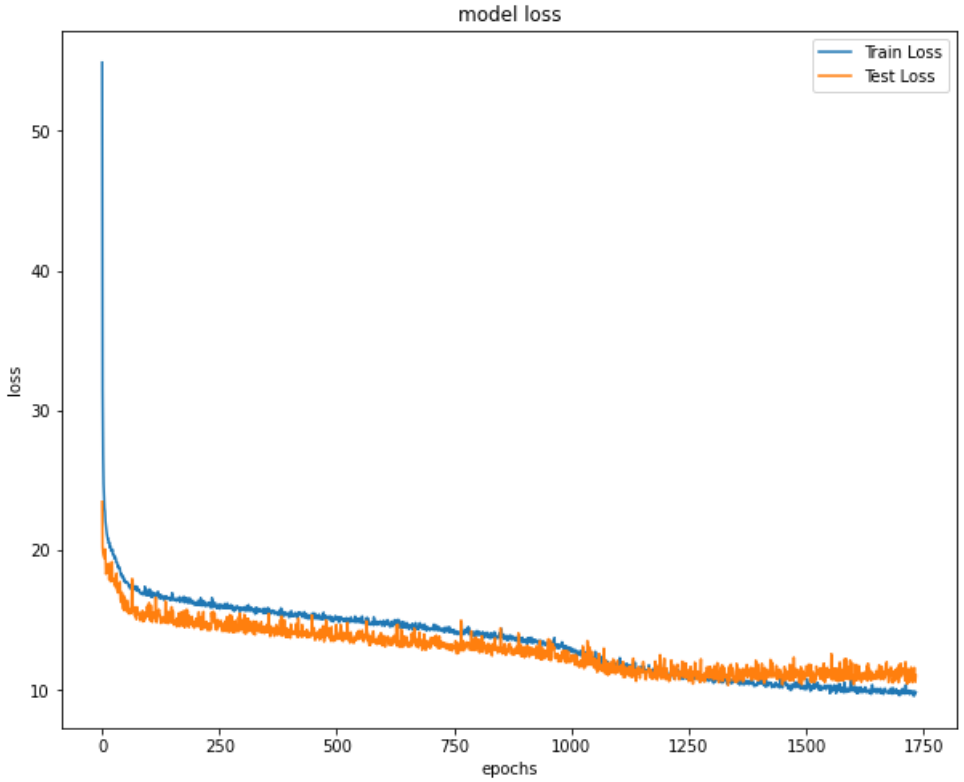
Π.4: Ποσοστά επί του συνόλου του πλήθους τυπικών προφίλ ανά ομάδα χρησιμοποιώντας: Mini Batch K-means για n=15. **(**Κεφάλαιο 4.5.2)



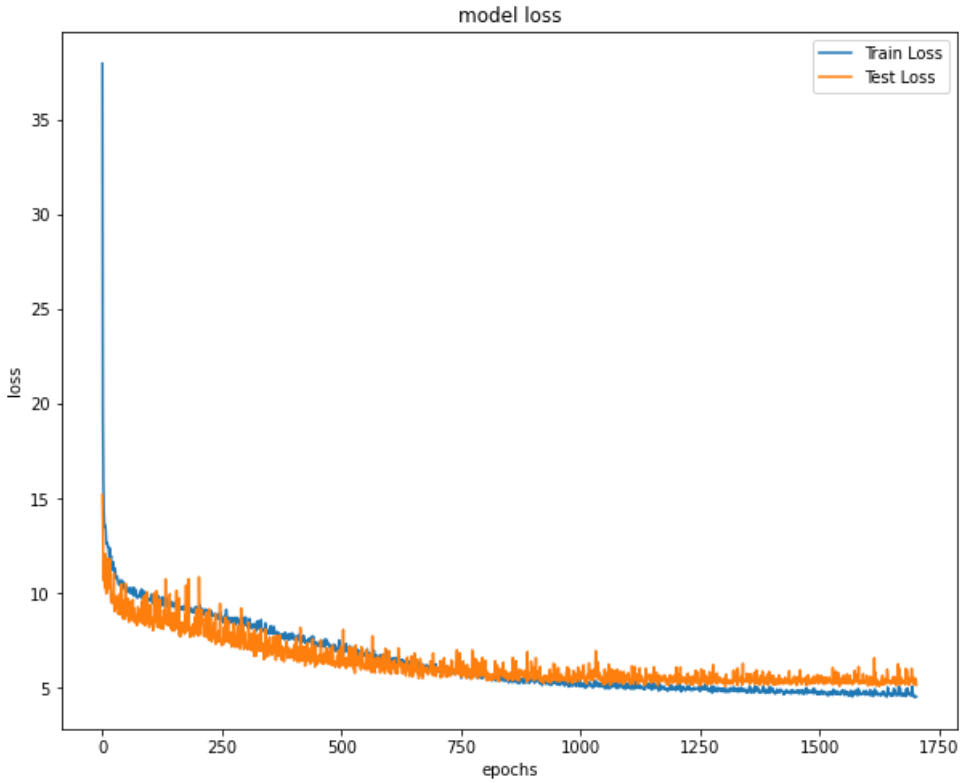
Π.5: Σφάλμα εκπαίδευσης (Loss) και σφάλμα πρόβλεψης (Val_Loss) για το cluster 1. (Κεφάλαιο 5.3)



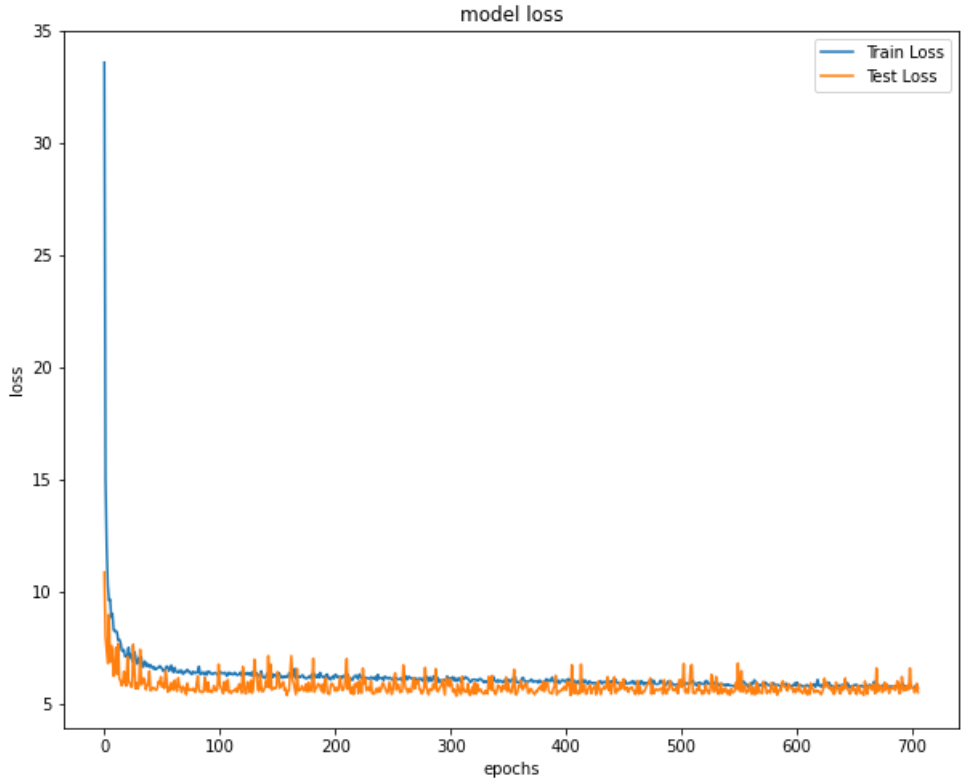
Π.6: Σφάλμα εκπαίδευσης (Loss) και σφάλμα πρόβλεψης (Val_Loss) για το cluster 2. (Κεφάλαιο 5.3)



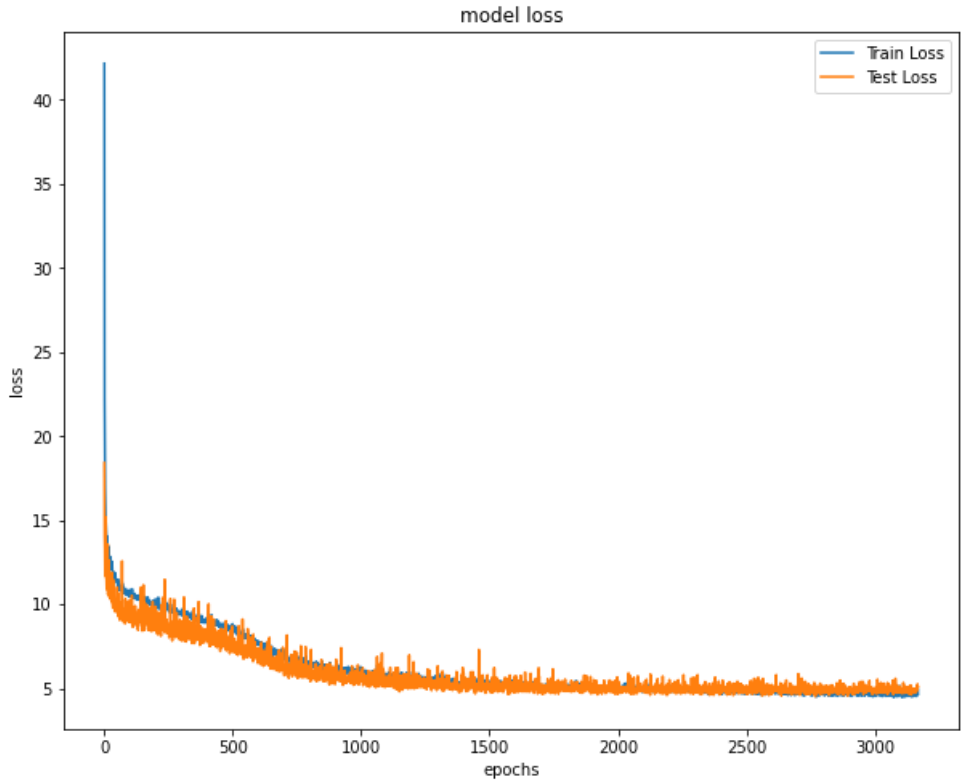
Π.7: Σφάλμα εκπαίδευσης (Loss) και σφάλμα πρόβλεψης (Val_Loss) για το cluster 3. (Κεφάλαιο 5.3)



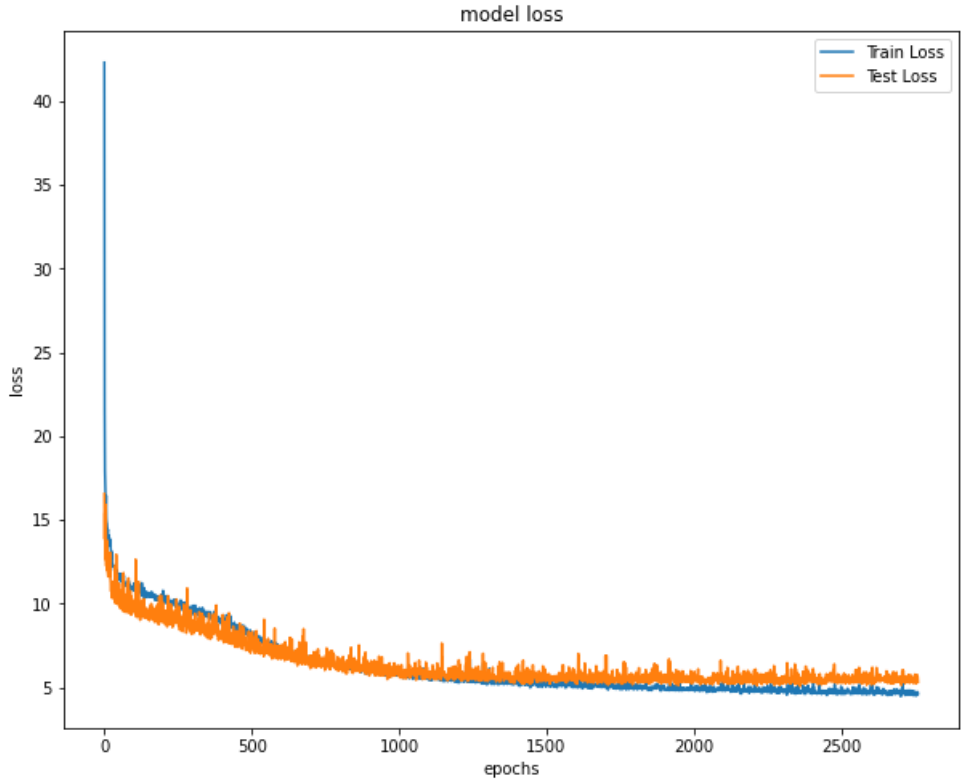
Π.8: Σφάλμα εκπαίδευσης (Loss) και σφάλμα πρόβλεψης (Val_Loss) για το cluster 4. (Κεφάλαιο 5.3)



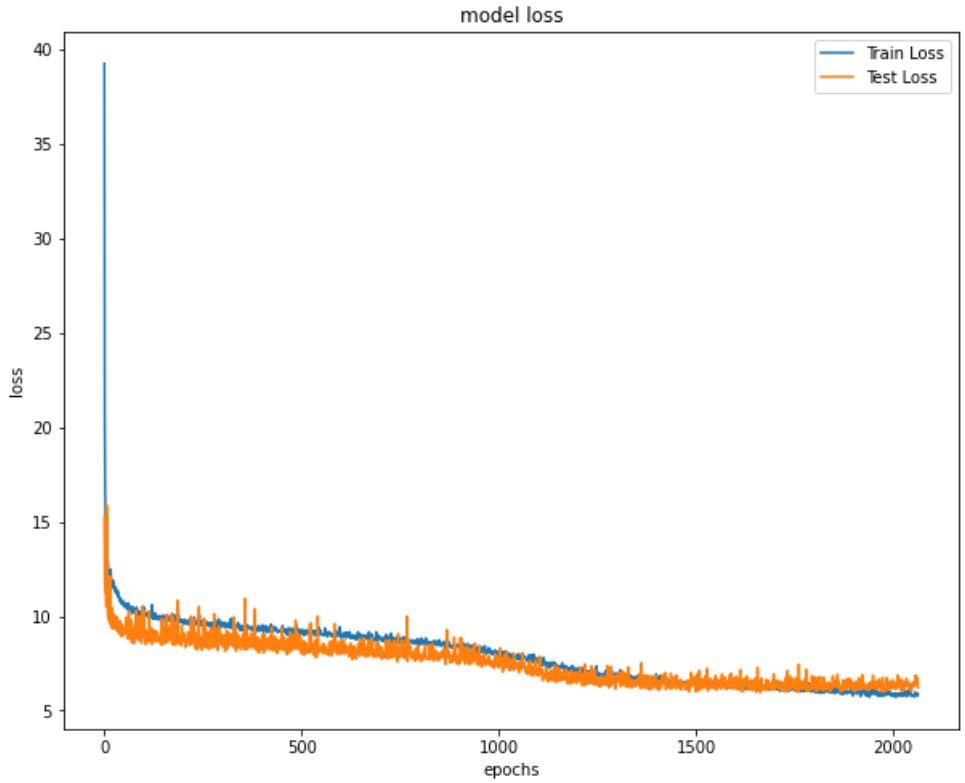
Π.9: Σφάλμα εκπαίδευσης (Loss) και σφάλμα πρόβλεψης (Val_Loss) για το cluster 5. (Κεφάλαιο 5.3)



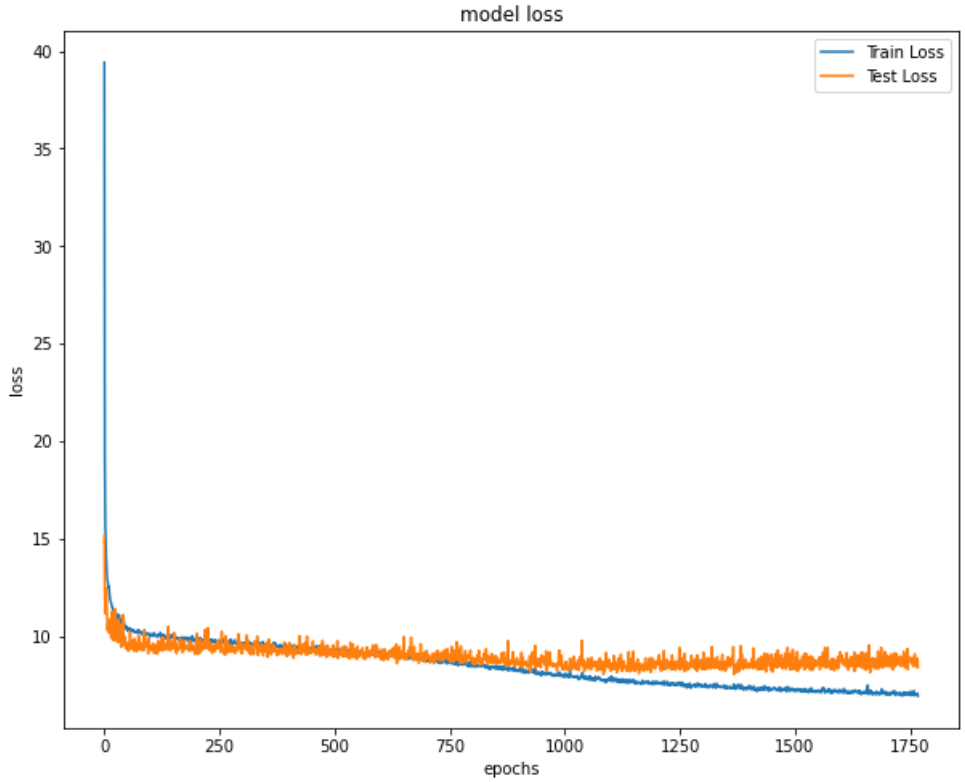
Π.10: Σφάλμα εκπαίδευσης (Loss) και σφάλμα πρόβλεψης (Val_Loss) για το cluster 6. (Κεφάλαιο 5.3)



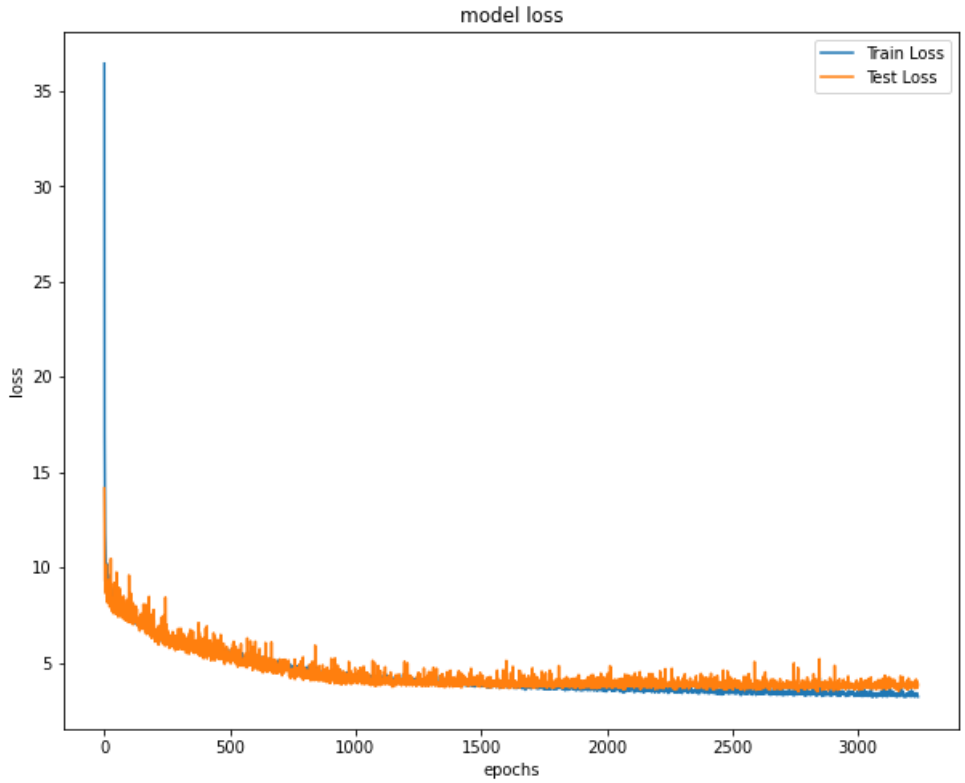
Π.11: Σφάλμα εκπαίδευσης (Loss) και σφάλμα πρόβλεψης (Val_Loss) για το cluster 7. (Κεφάλαιο 5.3)



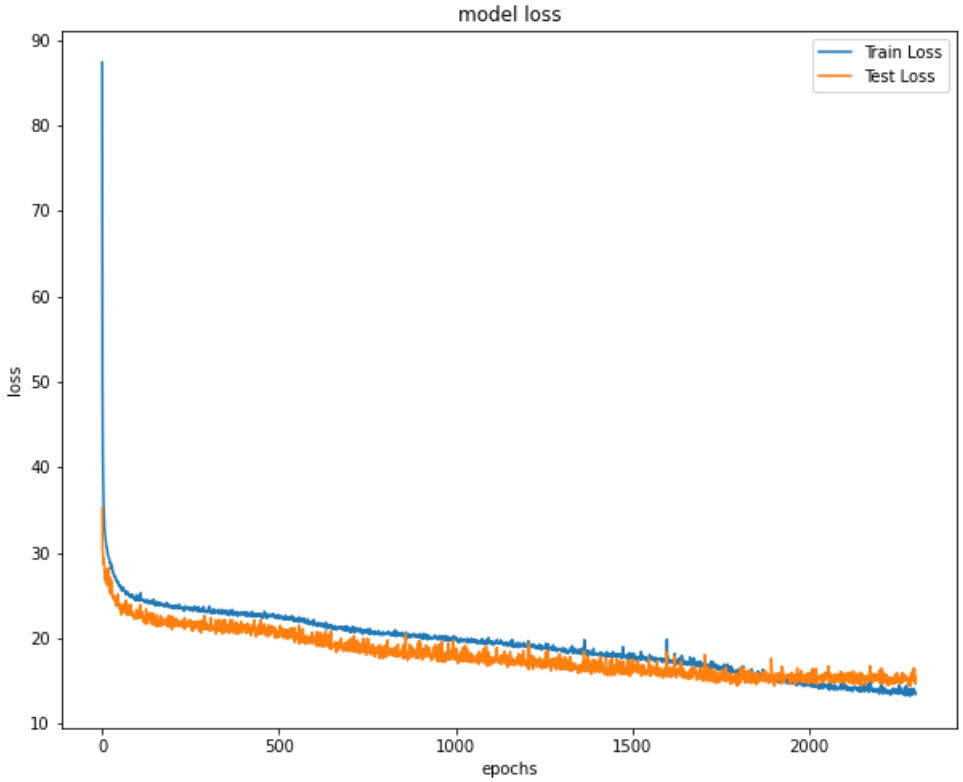
Π.12: Σφάλμα εκπαίδευσης (Loss) και σφάλμα πρόβλεψης (Val_Loss) για το cluster 8. (Κεφάλαιο 5.3)



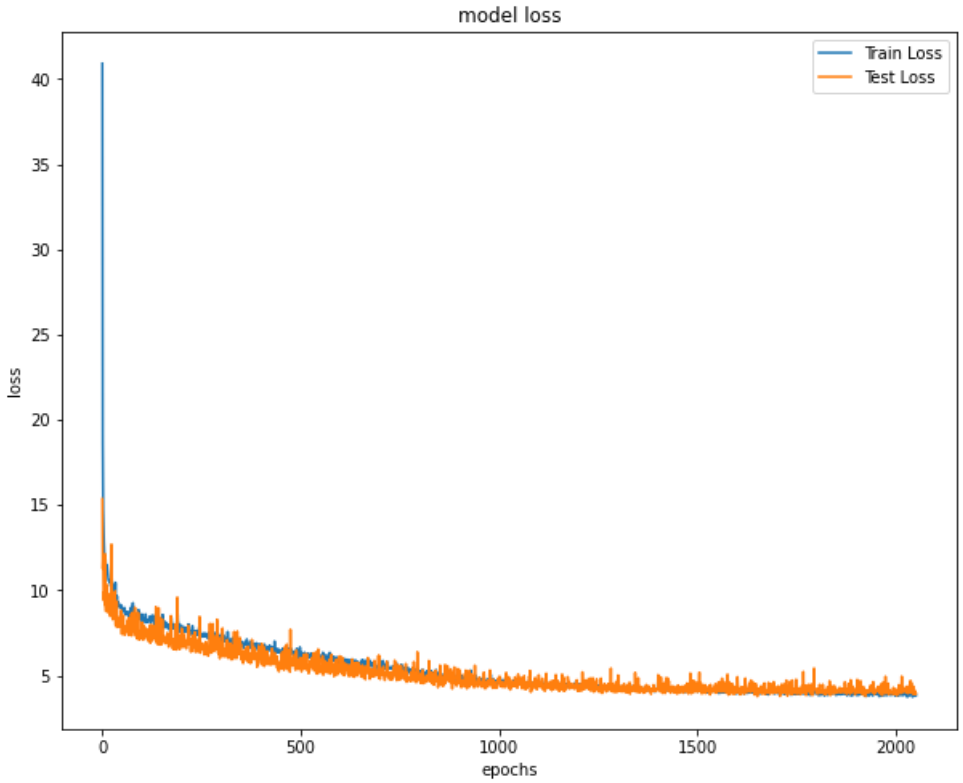
Π.13: Σφάλμα εκπαίδευσης (Loss) και σφάλμα πρόβλεψης (Val_Loss) για το cluster 9. (Κεφάλαιο 5.3)



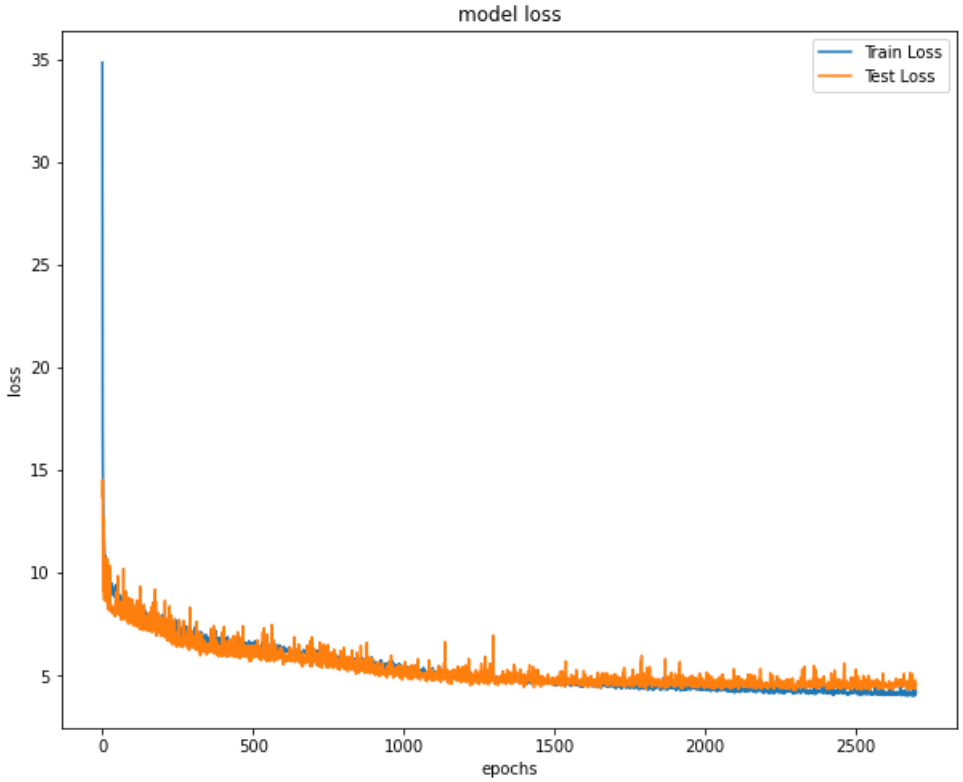
Π.14: Σφάλμα εκπαίδευσης (Loss) και σφάλμα πρόβλεψης (Val_Loss) για το cluster 10. (Κεφάλαιο 5.3)



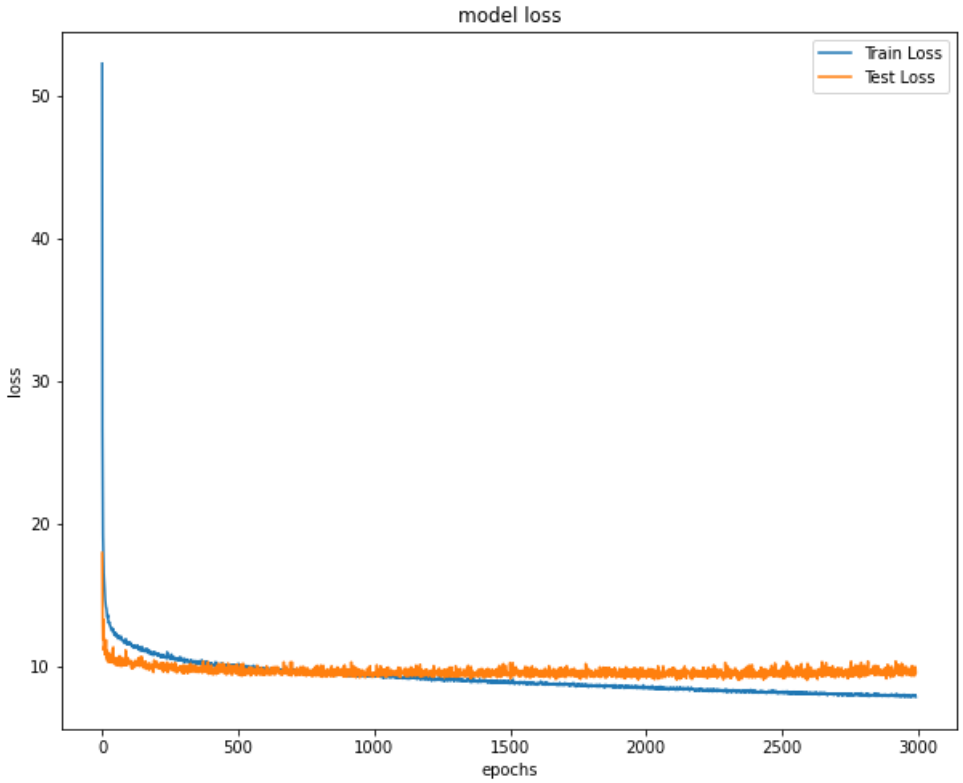
Π.15: Σφάλμα εκπαίδευσης (Loss) και σφάλμα πρόβλεψης (Val_Loss) για το cluster 11. (Κεφάλαιο 5.3)



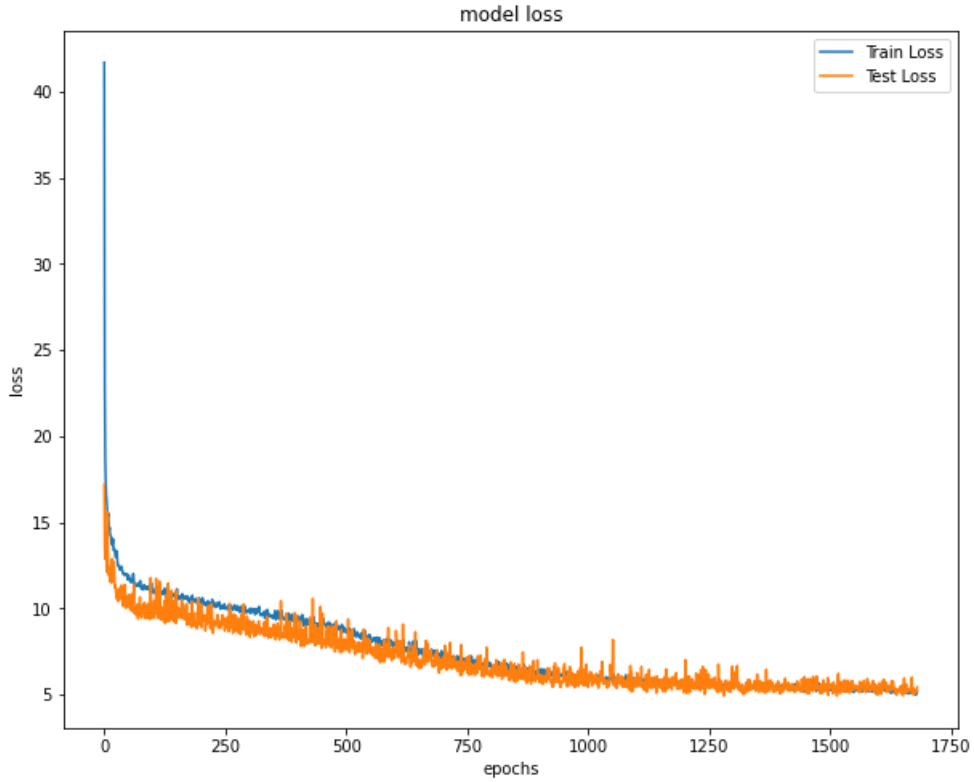
Π.16: Σφάλμα εκπαίδευσης (Loss) και σφάλμα πρόβλεψης (Val_Loss) για το cluster 12.



Π.17: Σφάλμα εκπαίδευσης (Loss) και σφάλμα πρόβλεψης (Val_Loss) για το cluster 13. (Κεφάλαιο 5.3)



Π.18: Σφάλμα εκπαίδευσης (Loss) και σφάλμα πρόβλεψης (Val_Loss) για το cluster 14. (Κεφάλαιο 5.3)



Π.19: Σφάλμα εκπαίδευσης (Loss) και σφάλμα πρόβλεψης (Val_Loss) για το cluster 15. (Κεφάλαιο 5.3)

Βιβλιογραφία

[1] S. Theodoridis S. and K. Koutroumbas, *Pattern Recognition*. 4th Ed. Academic Press, 2009.

[2] R. Duda, P. Hart and D. Stork, *Pattern classification*. 2nd Ed. USA: Wiley - Interscience, 2000.

[3] Γ. Κονταξής and Β.Κ. Παπαδιάς Β.Κ., *Ηλεκτρική Οικονομία.* 1η Έκδοση Αθήνα: Εκδόσεις Ε.Μ.Π, 1996.

[4] "Ενεργειακό Μείγμα", *Διαχειριστής ΑΠΕ & Εγγυήσεων Προέλευσης Α.Ε.* - *ΔΑΠΕΕΠ Α.Ε.* 2021. [Online]. Available at: https://www.dapeep.gr/viosimi-anaptixi/energeiako-meigma/ [Accessed 1 April 2021].

[5] Ν. Βοβός and Γ. Γιαννακόπουλος, *Ανάλυση Συστημάτων Ηλεκτρικής Ενέργειας.* 1η Έκδοση Αθήνα: Εκδόσεις Ζήτη, 2008.

[6] H. Amarawickrama and L. Hunt, "Electricity demand for Sri Lanka: A time series analysis", *Energy*, 33(5), 2008, pp.724-739.

[7] "Θεσμικό Πλαίσιο Ηλεκτρισμού", *Ρυθμιστική Αρχή Ενέργειας* - *PAE*. 2021 . [Online] Available at: https://www.rae.gr/θεσμικό-πλαίσιο-ηλεκτρισμού [Accessed 5 April 2021].

[8] "Directive (EU) 2019/944 of the European Parliament and of the Council of 5 June 2019 on common rules for the internal market for electricity and amending" Directive 2012/27/EU, PE/10/2019/REV/1, *OJ L 158, 14.6.2019, p. 125–199*

[9] C. Chen, J. Hwang and C. Huang, 1997. "Application of load survey systems to proper tariff design." *IEEE Transactions on Power Systems*, 12(4), 1997, pp.1746-1751.

[10] G. Chicco, R. Napoli, P. Postolache, M. Scutariu and T. Cornel, "Electric energy customer characterisation for developing dedicated market strategies." *2001 IEEE Porto Power Tech Conference 10th-13th September*, 2001, Porto, Portugal, 6 pp.

[11] V. Figueiredo, F.J. Duarte, F. Rodrigues, Z. Vale, C. Ramos and B. Gouveia, "Electric Customer Characterization by Clustering", *ISAP 2003*, August-September 2003, Lemnos, Greece.

[12] "Competitive Learning", *Stanford University, Web.stanford.edu*, 2021. *6* [online] Available at: https://web.stanford.edu/group/pdplab/pdphandbook/handbookch7.html [Accessed 15 May 2021].

[13] "Error Sum of Squares", *Stanford University, Hlab.stanford.edu*, 2021. [Online] Available at: https://hlab.stanford.edu/brian/error_sum_of_squares.html [Accessed 10 May 2021].

[14] P. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis", *Journal of Computational and Applied Mathematics*, 1987 20, pp.53-65.

[15] "Selecting the number of clusters with silhouette analysis on KMeans clustering", *Scikit-learn*, 2021. [Online] Available at: https://scikit-learn.org/stable/auto_examples/cluster/plot_kmeans_silhouette_analysis.html [Accessed 11 May 2021].

[16] G. Chicco, R. Napoli, F. Piglione, P. Postolache, M. Scutariu and C. Toader, "Load Pattern-Based Classification of Electricity Customers", *IEEE Transactions on Power Systems*, 2004, 19(2), pp.1232-1239.

[17] Α. Λύκας, *Υπολογιστική Νοημοσύνη.* 1η Έκδοση Ιωάννινα: Πανεπιστημιακές Σημειώσεις Ιωαννίνων, 1999.

[18] "CS221", *Stanford University, Stanford*.edu, 2021. . [Online] Available at: https://stanford.edu/~cpiech/cs221/handouts/kmeans.html [Accessed 9 June 2021].

[19] Γ. Τσεκούρας, "Συμβολή στη βραχυπρόθεσμη και μεσοπρόθεσμη πρόβλεψη ζήτησης φορτίου και ενέργειας συστημάτων ηλεκτρικής ενέργειας με χρήση μεθόδων αναγνώρισης προτύπων." Διδακτορική Διατριβή στο τμήμα Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών, Εθνικό Μετσόβιο Πολυτεχνείο, Αθήνα, 2006.

[20] "Fuzzy k means", *Priceton University, Cs.princeton.edu*, 2021. [Online] Available at: https://www.cs.princeton.edu/courses/archive/fall08/cos436/Duda/C/fk_means.htm [Accessed 10 May 2021].

[21] T. Kohonen, "The self-organizing map", in *Proceedings of the IEEE*, vol. 78, no. 9, pp. 1464-1480, Sept. 1990, doi: 10.1109/5.58325.

[22] "Hierarchical Clustering", *MATLAB & Simulink, Mathworks.com*, 2021. [Online] Available at: https://www.mathworks.com/help/stats/hierarchical-clustering.html [Accessed 12 May 2021].

[23] N. Cristianini and J. Shawe-Taylor, *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*. Cambridge: Cambridge University Press, 2000.

[24] N. Ahmed, A. Atiya, N. Gayar and H. El-Shishiny, "An Empirical Comparison of Machine Learning Models for Time Series Forecasting", *Econometric Reviews*, vol. 29, no. 5-6, pp. 594-621, 2010.

[25] L. Breiman, J.H. Friedman, R.A. Olshen and C.J. Stone, *Classification and regression trees.* Belmont, CA: Wadsworth, Inc, 1984.

[26] Π Λαδάς, "Βραχυπρόθεσμη πρόβλεψη ενεργειακής ζήτησης Προσεγγίσεις βασισμένες στη Μηχανική Μάθηση.", Διπλωματική Εργ. στο τμήμα Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών, Εθνικό Μετσόβιο Πολυτεχνείο, Αθήνα, 2014.

[27] Ι. Σαλάτας, "Υλοποίηση και εφαρμογή Τεχνητών Νευρωνικών Δικτύων για την πρόβλεψη χρονοσειρών συναλλαγματικών ισοτιμιών.", Διπλωματική Εργ. στο τμήμα Θετικών Σπουδών και Τεχνολογίας, Ελληνικό Ανοιχτό Πανεπιστήμιο, Πάτρα, 2013.

[28] ML Glossary documentation. "*Activation Functions*", [Online]. Available: https://ml-cheatsheet.readthedocs.io/en/latest/activation_functions.html#sigmoid [Accessed: 15-May-2021].

[29] Λ. Θεοδόση – Κοκκίνου, "Τεχνητά Νευρωνικά Δίκτυα και εφαρμογές στα Συστήματα Αυτόματου Ελέγχου.", Διπλωματική Εργ. στο τμήμα Ηλεκτρολόγων Μηχανικών και Τεχνολογίας Υπολογιστών, Πανεπιστήμιο Πατρών, Πάτρα, 2013.

[30] P. Fernández-Cabán, F. Masters and B. Phillips, "Predicting Roof Pressures on a Low-Rise Structure From Freestream Turbulence Using Artificial Neural Networks", *Frontiers in Built Environment*, 2018.

[31] "Kmeans", *Scikit-learn*. 2021. [Online] Available at: https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html [Accessed 4 March 2021].

[32] "Mean_Squared_Error", *Scikit-learn*, 2021. [Online]. Available at: https://scikit-learn.org/stable/modules/generated/sklearn.metrics.mean_squared_error.html. [Accessed: 05- Mar- 2021].

[33] "Silhouette_Score", *Scikit-learn*, 2021. [Online]. Available at: https://scikit-learn.org/stable/modules/generated/sklearn.metrics.silhouette_score.html. [Accessed: 05- May- 2021].

[34] "Davies_Bouldin_score", *Scikit-learn*, 2021. [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.metrics.davies_bouldin_score.html. [Accessed: 05- May- 2021].

[35] I. Jolliffe, *Principal component analysis*. New York: Springer, 2011.

[36] "A Step-by-Step Explanation of Principal Component Analysis (PCA)", *Built In - Data Science*, 2021. [Online]. Available: https://builtin.com/data-science/step-step-explanation-principal-component-analysis. [Accessed: 06- May- 2021].

[37] "PCA", *Scikit-learn*, 2021. [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.PCA.html. [Accessed: 06- May- 2021].

[38] "ML | Mini Batch K-means clustering algorithm", *GeeksforGeeks*, 2021. [Online]. Available: https://www.geeksforgeeks.org/ml-mini-batch-k-means-clustering-algorithm/. [Accessed: 07- May- 20211].

[39] "MiniBatchKMeans", *Scikit-learn*, 2021. [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.cluster.MiniBatchKMeans.html. [Accessed: 07- May- 2021].

[40] Sharma, P., 2021. *Keras Dense Layer Explained*. [Online] MLK - Machine Learning Knowledge. Available at: https://machinelearningknowledge.ai/keras-dense-layer-explained-for-beginners [Accessed 1 June 2021].

[41] Grosser, S. and Michel, J., 2021. *ENNUI ~ Elegant Neural Network User Interface ~*. [Online] MIT - Math.mit.edu. Available at: https://math.mit.edu/ennui [Accessed 1 June 2021].

[42] Brownlee, J., 2021. *Introduction to Dropout for Regularizing Deep Neural Networks*. [Online] Machine Learning Mastery. Available at: https://machinelearningmastery.com/dropout-for-regularizing-deep-neural-networks [Accessed 1 June 2021].

[43] Hinno, R., 2021. *Callbacks in neural networks*. [Online] Medium. Available at: https://towardsdatascience.com/callbacks-in-neural-networks-b0b006df7626 [Accessed 1 June 2021].