Σχολή Θετικών Επιστημών και Τεχνολογίας

Διπλωματική Εργασία

Εφαρμογή Μεθόδων Μηχανικής Μάθησης για την Υποστήριξη Λήψης Αποφάσεων στον Δημόσιο Τομέα: Περίπτωση Οργανισμού Τοπικής Αυτοδιοίκησης

Δημήτριος Ράπτης

Επιβλέπων καθηγητής: Ανδρέας Καναβός

Πάτρα, Μάιος 2026

Εφαρμογή Μεθόδων Μηχανικής Μάθησης για την Υποστήριξη Λήψης Αποφάσεων στον Δημόσιο Τομέα: Περίπτωση Οργανισμού Τοπικής Αυτοδιοίκησης

Δημήτριος Ράπτης

Επιτροπή Επίβλεψης Διπλωματικής Εργασίας

Επιβλέπων Καθηγητής	Συν-Επιβλέπουσα Καθηγήτρια
Ανδρέας Καναβός	Ελένη Χριστοπούλου
Αναπληρωτής Καθηγητής	ΕΔΙΠ
Ιόνιο Πανεπιστήμιο	Ιόνιο Πανεπιστήμιο

Πάτρα, Μάιος 2026

Ευχαριστίες

Ευχαριστώ θερμά τον επιβλέποντά μου κ. Ανδρέα Καναβό για την πολύτιμη και ουσιαστική υποστήριξή του για την ολοκλήρωση της παρούσας εργασίας. Αφιερώνεται στα παιδιά μου Θωμά και Κωνσταντίνο, με την προσδοκία να αποτελέσει αφορμή για την επίτευξη των δικών τους εκπαιδευτικών και ερευνητικών στόχων και στη σύζυγό μου Κατερίνα για την συμβολή της ως φιλόλογος στην κειμενική επιμέλεια. Ιδιαίτερες ευχαριστίες συνολικά στον οργανισμό του ΕΑΠ για την ευκαιρία που μου έδωσε να συνεχίσω να μαθαίνω...

Περίληψη

Ο δημόσιος τομέας αποτελεί βασικό δομικό στοιχείο της κοινωνικής οργάνωσης, παρέχοντας ζωτικές υπηρεσίες προς τους πολίτες, μέσω της αποκλειστικής – σε αρκετές περιπτώσεις – διαχείρισης κρίσιμων πόρων. Τόσο ο τρόπος όσο και η διαδικασία με την οποία λαμβάνονται οι αποφάσεις από τη Διοίκηση σχετικά με την ορθολογική διαχείριση των διαθέσιμων πόρων, αποτελούν βασική παράμετρο για την επιτυχή άσκηση της πολιτικής και διοικητικής λειτουργίας. Η εν λόγω διαδικασία συχνά επηρεάζεται από παράγοντες όπως πολιτικές και κοινωνικές παρεμβάσεις, διοικητική πολυπλοκότητα, ελλιπή ή διάσπαρτα δεδομένα, καθώς και από περιορισμένους οικονομικούς και ανθρώπινους πόρους.

Ιδιαίτερα η Τοπική Αυτοδιοίκηση (ΟΤΑ Α' βαθμού – Δήμοι) καλείται να διαχειριστεί πολύπλοκα ζητήματα που σχετίζονται με τη βέλτιστη αξιοποίηση των διαθέσιμων πόρων προς όφελος της τοπικής κοινωνίας. Προκλήσεις όπως η μείωση του ενεργειακού κόστους και του κόστους συλλογής απορριμμάτων, απαιτούν σύγχρονα εργαλεία υποστήριξης για την ορθολογική κατανομή των διαθέσιμων οικονομικών και ανθρώπινων πόρων. Η αυξανόμενη διαθεσιμότητα δεδομένων (οικονομικών, ενεργειακών, γεωχωρικών, κοινωνικών, περιβαλλοντικών) καθιστά ελκυστική την ενσωμάτωση μεθόδων Μηχανικής Μάθησης (Machine Learning - ML) στην υποστήριξη της λήψης αποφάσεων.

Η εφαρμογή τεχνικών Machine Learning - ML στον δημόσιο τομέα εξελίσσεται ραγδαία, προσφέροντας σημαντικές δυνατότητες για την ενίσχυση της αποτελεσματικότητας και της διαφάνειας των διοικητικών λειτουργιών. Ειδικότερα, η αξιοποίηση Συστημάτων Υποστήριξης Αποφάσεων (Decision Support Systems - DSS) σε επίπεδο Οργανισμών Τοπικής Αυτοδιοίκησης (ΟΤΑ) πρώτου βαθμού, δύναται να λειτουργήσει ως μοχλός μετασχηματισμού της δημόσιας διοίκησης προς όφελος των τοπικών κοινωνιών, ενσωματώνοντας χαρακτηριστικά όπως η ευελιξία, η απόδοση και η λογοδοσία.

Η παρούσα εργασία επικεντρώνεται στην ανάπτυξη και εφαρμογή αλγοριθμικών μοντέλων Μηχανικής Μάθησης, αξιοποιώντας επιχειρησιακά δεδομένα ενός Ο.Τ.Α. που εντάσσεται στους Μεγάλους Ηπειρωτικούς Δήμους, με στόχο τη βελτίωση της διοικητικής αποδοτικότητας και την υποστήριξη στρατηγικών αποφάσεων στον κρίσιμο τομέα της ενεργειακής διαχείρισης. Εξετάζονται τεχνικές επιβλεπόμενης μάθησης (π.χ. Random Forests, Gradient Boosting) για την πρόβλεψη της κατανάλωσης ηλεκτρικής ενέργειας σε

δημοτικά κτίρια και υποδομές, με σκοπό τον σχεδιασμό παρεμβάσεων ενεργειακής αναβάθμισης. Τα αποτελέσματα αναδεικνύουν τη δυναμική των μεθόδων Machine Learning - ML στην παροχή τεκμηριωμένων υποδείξεων, ενισχύοντας τόσο την επιχειρησιακή λειτουργία όσο και τον στρατηγικό σχεδιασμό των Ο.Τ.Α. Παράλληλα, αναλύονται οι προκλήσεις που ανακύπτουν σχετικά με τη διαχείριση και ποιότητα των δημόσιων δεδομένων, την ανάγκη για διαφάνεια και ερμηνευσιμότητα (Explainable AI) των αλγορίθμων, καθώς και τα ηθικά διλήμματα που συνοδεύουν την αυτοματοποιημένη λήψη αποφάσεων στον δημόσιο τομέα.

Λέξεις – Κλειδιά

Μηχανική Μάθηση, Συστήματα Υποστήριξης Αποφάσεων, Οργανισμός Τοπικής Αυτοδιοίκησης.

Application of Machine Learning Methods for Decision Support in the Public Sector: Case of a Local Government Organization

Dimitrios Raptis

Abstract

The public sector constitutes a fundamental structural component of social organization, providing vital services to citizens through the exclusive — in several cases — management of critical resources. Both the manner and the process by which decisions are made by the Administration regarding the rational management of available resources constitute a key parameter for the successful exercise of political and administrative functions. This process is often influenced by factors such as political interventions, administrative complexity, incomplete or fragmented data, as well as limited financial and human resources.

In particular, Local Government Authorities (first-level local authorities — Municipalities) are called upon to manage complex issues related to the optimal utilization of available resources for the benefit of the local community. Challenges such as reducing energy costs and waste collection costs require modern support tools for the rational allocation of available financial and human resources. The increasing availability of data (financial, energy-related, geospatial, social, and environmental) makes the integration of Machine Learning (ML) methods into decision-making support especially attractive.

The application of Machine Learning (ML) techniques in the public sector is developing rapidly, offering significant potential for enhancing the efficiency and transparency of administrative functions. More specifically, the utilization of Decision Support Systems (DSS) at the level of first-degree Local Government Organizations (LGOs) can serve as a lever for transforming public administration for the benefit of local communities, by incorporating features such as flexibility, performance, and accountability.

This study focuses on the development and application of algorithmic Machine Learning models, utilizing operational data from a Local Government Organization classified among the Large Mainland Municipalities, with the aim of improving administrative efficiency and supporting strategic decision-making. Supervised learning techniques (e.g., Random Forests, Gradient Boosting) are examined for forecasting electricity consumption in municipal buildings and infrastructure, with the goal of planning energy-upgrade interventions. The results highlight the potential of Machine Learning (ML) methods to provide evidence-based recommendations, strengthening both the operational function and the strategic planning of Local Government Organizations. At the same time, the study analyzes the challenges arising in relation to the management and quality of public data, the need for transparency and interpretability (Explainable AI) of algorithms, as well as the ethical dilemmas accompanying automated decision-making in the public sector.

Keywords

Machine Learning, Decision Support Systems, Local Government Organization.

Περιεχόμενα

Κατάλογος Εικόνων / Σχημάτων

Κατάλογος Πινάκων

Συντομογραφίες & Ακρωνύμια

ΔΕ	Διπλωματική Εργασία
ΕΑΠ	Ελληνικό Ανοικτό Πανεπιστήμιο
ΤΑ	Τοπική Αυτοδιοίκηση
ΟΤΑ	Οργανισμοί Τοπικής Αυτοδιοίκησης
ML	Machine Learning
DSS	Decision Support Systems
ΟΟΣΑ	Οργανισμός Οικονομικής Συνεργασίας και Ανάπτυξης

1. Εισαγωγή

1.1 Δημόσιος Τομέας και Τοπική Αυτοδιοίκηση

Ο Δημόσιος Τομέας αποτελεί τον βασικό θεσμό μέσω του οποίου ασκείται η διοίκηση στο κράτος, παρέχονται υπηρεσίες στους πολίτες και εφαρμόζονται πολιτικές που υπηρετούν το δημόσιο συμφέρον. Ο Δημόσιος Τομέας περιλαμβάνει τη Γενική Κυβέρνηση (Κεντρική Κυβέρνηση, Οργανισμούς Τοπικής Αυτοδιοίκησης και Οργανισμούς Κοινωνικής Ασφάλισης) καθώς και τις Δημόσιες Επιχειρήσεις και Οργανισμούς (ΔΕΚΟ) που δεν εντάσσονται στη Γενική Κυβέρνηση (Σχήμα 1.1).

Σχήμα 1.1 Ιεραρχική δομή Δημόσιου Τομέα

Η Τοπική Αυτοδιοίκηση, ως αυτοτελές επίπεδο διακυβέρνησης (υποτομέας S1313 ΕΛΣΤΑΤ, 2024· Μητρώο Φορέων Γενικής Κυβέρνησης, κατά ESA 2010), περιλαμβάνει τους Οργανισμούς Τοπικής Αυτοδιοίκησης (ΟΤΑ) α’ και β’ βαθμού, δηλαδή τους δήμους και τις περιφέρειες. Συνιστά κρίσιμο θεσμικό πυλώνα του κράτους, λειτουργώντας ως συνδετικός κρίκος μεταξύ της κεντρικής διοίκησης και της τοπικής κοινωνίας. Οι ΟΤΑ συμβάλλουν καθοριστικά στην εξυπηρέτηση της καθημερινότητας των πολιτών και στη διαμόρφωση στρατηγικών αποφάσεων προς όφελος της τοπικής κοινωνίας (Χλέπας, 2020).

Οι βασικές αρμοδιότητες των ΟΤΑ α' βαθμού – δηλαδή των δήμων – εκτείνονται σε ένα ευρύ φάσμα λειτουργικών πεδίων, όπως:

- Υπηρεσίες καθαριότητας, δημοτικός φωτισμός, ύδρευση και αποχέτευση
- Διαχείριση δημόσιων και κοινόχρηστων υποδομών
- Παροχή κοινωνικών, προνοιακών και πολιτιστικών υπηρεσιών
- Περιβαλλοντική προστασία και ενεργειακή διαχείριση
- Προώθηση ψηφιακής διακυβέρνησης και διαφάνειας
- Διαχείριση τεχνικών έργων, έργων οδοποιίας και αστικών υποδομών
- Έκδοση πιστοποιητικών, αδειών και άλλων διοικητικών εγγράφων
- Εποπτεία σχολικών μονάδων και πολιτιστικών φορέων
- Σχεδιασμός και υλοποίηση τοπικών αναπτυξιακών πολιτικών

Η λειτουργία των ΟΤΑ διέπεται από τις αρχές της επικουρικότητας και της εγγύτητας (Ν. 3463/2006, άρθρο 75), ενώ η διοικητική και οικονομική τους αυτοτέλεια κατοχυρώνεται συνταγματικά (Σύνταγμα της Ελλάδας, άρθρο 102). Ωστόσο, στην πράξη η αποτελεσματικότητα τους περιορίζεται από δομικά και λειτουργικά εμπόδια. Μεταξύ αυτών συγκαταλέγονται η υπερβολική γραφειοκρατία, η σοβαρή υποστελέχωση – ιδιαίτερα σε εξειδικευμένο προσωπικό – οι περιορισμένοι χρηματοδοτικοί πόροι (OECD, 2020) και η καθυστέρηση στην υιοθέτηση ψηφιακών εργαλείων και υποδομών (OECD, 2022).

Η αντιμετώπιση αυτών των προκλήσεων απαιτεί την αναβάθμιση της ικανότητας των ΟΤΑ, η οποία προϋποθέτει τη μετάβαση σε ένα μοντέλο τοπικής διακυβέρνησης που θα βασίζεται στην υιοθέτηση σύγχρονων τεχνολογικών εργαλείων, ώστε να ανταποκρίνονται στις αυξανόμενες απαιτήσεις των τοπικών κοινωνιών.

1.2 Προκλήσεις στη Λήψη Αποφάσεων

Η διαδικασία λήψης αποφάσεων από τα αρμόδια όργανα του δημόσιου τομέα - είτε μονομελή είτε συλλογικά - και ιδίως στους ΟΤΑ, είναι συχνά περίπλοκη, καθώς επηρεάζεται από ένα πλήθος αλληλοσυγκρουόμενων κοινωνικών, οικονομικών, πολιτικών και διοικητικών παραμέτρων. Η επιδίωξη της εξυπηρέτησης του δημόσιου συμφέροντος,

υπό συνθήκες αβεβαιότητας και περιορισμένων πόρων, προϋποθέτει τη συνεκτίμηση πολλών και συχνά σύνθετων μεταβλητών ενώ απαιτεί επαρκή γνώση και τεχνική εξειδίκευση για την αξιοποίηση δεδομένων και αναλύσεων (Janssen et al., 2017).
Η διαφάνεια και η αποτελεσματικότητα της εκάστοτε διοίκησης, με βάση τον στρατηγικό της σχεδιασμό, αντιμετωπίζει πολυάριθμα προβλήματα κατά τη διαδικασία λήψης αποφάσεων στους ΟΤΑ, όπως:

• Πληθώρα, πολυπλοκότητα και ασάφεια δεδομένων

Οι ΟΤΑ παράγουν και διαχειρίζονται έναν τεράστιο όγκο δεδομένων που σχετίζονται με οικονομικές λειτουργίες, τεχνικές υπηρεσίες, κοινωνικές παροχές, δημοτική περιουσία, δαπάνες ενέργειας και αιτήματα πολιτών. Ωστόσο, αυτά τα δεδομένα είναι συχνά ασύνδετα, ελλιπώς καταγεγραμμένα ή μη επικαιροποιημένα, ενώ απουσιάζει η ύπαρξη ενιαίων και διαλειτουργικών πληροφοριακών συστημάτων. Αυτή η κατάσταση δυσχεραίνει την τεκμηρίωση για την αποτύπωση της πραγματικότητας και περιορίζει τη δυνατότητα ορθολογικής ανάλυσης των δεδομένων κατά τη λήψη των σχετικών αποφάσεων.

• Περιορισμένοι ανθρώπινοι και τεχνολογικοί πόροι

Η λειτουργία των ΟΤΑ συχνά καθορίζεται από αυστηρούς δημοσιονομικούς περιορισμούς και ελλείψεις σε εξειδικευμένο προσωπικό, όπως αναλυτές δεδομένων, μηχανικούς στον τομέα των ΤΠΕ ή οικονομολόγους. Ταυτόχρονα, οι τεχνολογικές υποδομές σε πολλές περιπτώσεις είναι παρωχημένες, γεγονός που καθιστά δύσκολη την αξιοποίηση σύγχρονων τεχνολογικών εργαλείων όπως τα Συστήματα Υποστήριξης Αποφάσεων (Irani et al., 2023).

• Πολυτομεακός και σύνθετος χαρακτήρας των ζητημάτων

Θέματα όπως η ενεργειακή διαχείριση, η αστική κινητικότητα ή η κοινωνική προστασία απαιτούν τη συνεργασία διαφορετικών διοικητικών μονάδων και ιεραρχικών επιπέδων, καθώς και το συντονισμό ετερογενών επιστημονικών και τεχνικών ειδικοτήτων.

• Πολιτικές πιέσεις και ασυνέχεια πολιτικών

Η διαδικασία λήψης αποφάσεων ενδέχεται να επηρεάζεται από πολιτικές σκοπιμότητες ή κοινωνικές πιέσεις, διοικητική πολυπλοκότητα και περιορισμένη ενσωμάτωση εργαλείων στρατηγικού σχεδιασμού και αξιολόγησης, γεγονός που υπονομεύει τη μακροπρόθεσμη στρατηγική σχεδίαση (Plimakis, 2018). Η εναλλαγή διοικήσεων συχνά συνοδεύεται από αλλαγή προτεραιοτήτων και διακοπή ή επαναπροσδιορισμό υφιστάμενων πολιτικών.

• Έλλειψη εργαλείων πρόβλεψης και αναλυτικής τεκμηρίωσης

Η λήψη αποφάσεων συνήθως βασίζεται σε εμπειρική γνώση ή σε αποσπασματικά ιστορικά και διάσπαρτα δεδομένα. Η μη αξιοποίηση σύγχρονων αναλυτικών μεθόδων και τεχνικών μηχανικής μάθησης μπορεί να περιορίσει την ακρίβεια και τη δυνατότητα τεκμηριωμένης στρατηγικής στοχοθεσίας, καθώς δεν αξιοποιούνται πλήρως τα διαθέσιμα δεδομένα για την πρόβλεψη και υποστήριξη των αποφάσεων (Jordan & Mitchell, 2015).

Σχήμα 1.2 Προκλήσεις στη Λήψη Αποφάσεων

Η σημερινή πρόκληση για την Τοπική Αυτοδιοίκηση είναι η μετάβαση σε ένα μοντέλο διοίκησης βασισμένο στην αξιοποίηση και ανάλυση έγκυρων λειτουργικών δεδομένων που προκύπτουν από την καθημερινή δραστηριότητα των δήμων στο πεδίο. Η υιοθέτηση Συστημάτων Υποστήριξης Αποφάσεων (DSS) με τεχνικές Μηχανικής Μάθησης (ML), μπορεί να βελτιώσει την τεκμηρίωση και την αποτελεσματικότητα των αποφάσεων της διοίκησης (Janssen et al., 2017; Wirtz et al., 2019). Αυτό απαιτεί κατάλληλες δομές και διαδικασίες, ενδυναμώνοντας και επικοινωνώντας στην τοπική κοινωνία τον ρόλο των ΟΤΑ ως υπεύθυνων και προσαρμοστικών φορέων διοίκησης, ενισχύοντας τη διαφάνεια.

1.3 Εφαρμογές Μηχανικής Μάθησης και Τοπική Αυτοδιοίκηση

Η Μηχανική Μάθηση (Machine Learning – ML) αποτελεί ένα σύγχρονο και συνεχώς εξελισσόμενο εργαλείο για την υποστήριξη λήψης αποφάσεων με ιδιαίτερη αξία σε διοικητικά και επιχειρησιακά περιβάλλοντα που χαρακτηρίζονται από πολυπλοκότητα και ιδιαίτερα μεγάλο όγκο δεδομένων. Μέσω της ικανότητάς της να επεξεργάζεται μεγάλα και ετερογενή σύνολα πληροφοριών συμβάλλει στην ανάδειξη προτύπων, την εξαγωγή χρήσιμων γνώσεων και την παραγωγή ασφαλών προβλέψεων που βασίζονται σε έγκυρα δεδομένα (Witten et al., 2016). Σε αντίθεση με τις παραδοσιακές στατιστικές προσεγγίσεις, οι αλγόριθμοι μηχανικής μάθησης διαθέτουν ικανότητα προσαρμογής και μπορούν, σε πολλές περιπτώσεις, να βελτιώνουν την ακρίβειά των προβλέψεών τους όταν εκπαιδεύονται με περισσότερα δεδομένα (Jordan & Mitchell, 2015). Η υιοθέτηση τέτοιων προηγμένων τεχνολογιών από τους Οργανισμούς Τοπικής Αυτοδιοίκησης (ΟΤΑ) στην Ελλάδα βρίσκεται ακόμη σε αρχικό στάδιο σε σχέση με το μέσο όρο των χωρών του ΟΟΣΑ (OECD, 2022). Καθίσταται αναγκαία η προώθηση πρωτοβουλιών που ενθαρρύνουν τη μετάβαση σε μηχανισμούς ψηφιακής διακυβέρνησης, με στόχο την αξιοποίηση των διάσπαρτων και συχνά υποχρησιμοποιούμενων δεδομένων για την ενίσχυση της αποτελεσματικότητας της διοίκησης και της ποιότητας των παρεχόμενων υπηρεσιών (OECD, 2020). Ενδεικτικά σε επίπεδο ΟΤΑ μπορούν να εφαρμοστούν τεχνικές μηχανικής μάθησης για την υποστήριξη λήψης επιχειρησιακών αποφάσεων στους τομείς:

Πρόβλεψη αιτημάτων πολιτών και διαχείριση ανθρώπινων πόρων

Μέσω αλγορίθμων επιβλεπόμενης μάθησης (supervised learning), οι ΟΤΑ μπορούν να προβλέψουν τον αναμενόμενο φόρτο αιτημάτων πολιτών ανά γεωγραφική περιοχή, τύπο υπηρεσίας ή χρονική περίοδο. Με αυτόν τον τρόπο, μπορεί να προγραμματιστεί η εξατομικευμένη διάθεση του ανθρώπινου δυναμικού σε επιμέρους υπηρεσίες, με σκοπό την ταχύτερη απόκριση και τη βελτίωση της εξυπηρέτησης των πολιτών.

Βελτιστοποίηση αποκομιδής απορριμμάτων

Η χρήση αισθητήρων πληρότητας σε κάδους απορριμμάτων και δεδομένων κίνησης οχημάτων (GPS), μπορούν να τροφοδοτήσουν με δεδομένα, μοντέλα μηχανικής μάθησης, με σκοπό την πρόβλεψη βέλτιστων διαδρομών αποκομιδής. Οι τεχνικές αυτές οδηγούν στη μείωση του κόστους κατανάλωσης των καυσίμων, στην ορθολογική διαχείριση του προσωπικού και στην μείωση του περιβαλλοντικού αποτυπώματος.

Ενεργειακή διαχείριση δημοτικών κτιρίων & υποδομών

Οι αλγόριθμοι μηχανικής μάθησης μπορούν να μοντελοποιήσουν την ενεργειακή συμπεριφορά των δημοτικών εγκαταστάσεων. Με βάση ιστορικά δεδομένα και περιβαλλοντικές παραμέτρους, μπορούν να προβλέπουν την κατανάλωση ενέργειας και να εντοπίζουν αποκλίσεις που υποδεικνύουν σπατάλη ή βλάβες, όπως π.χ. στα συστήματα θέρμανσης και ψύξης (HVAC - Heating, Ventilation, and Air Conditioning). Η εφαρμογή τέτοιων συστημάτων επιτρέπει την έγκαιρη παρέμβαση με στόχο την εξοικονόμηση ενέργειας και την παράταση της διάρκειας ζωής των ηλεκτρομηχανολογικών εγκαταστάσεων (Ahmad et al., 2017).

Προληπτική συντήρηση τεχνικών έργων

Δεδομένα από αισθητήρες, τεχνικές αναφορές μηχανικών και εξωτερικές συνθήκες μπορούν να χρησιμοποιηθούν για την πρόβλεψη βλαβών σε οδοποιία, φωτισμό ή υποδομές άρδευσης και ύδρευσης, μειώνοντας το κόστος που προκαλούν οι έκτακτες και εκτός του ετήσιου προγραμματισμού παρεμβάσεις από τη διακοπή λειτουργίας των δημοτικών υποδομών. Ανάλογες τεχνικές εφαρμόζονται σε βιομηχανικά περιβάλλοντα όπου σχεδιάζονται συντηρήσεις βάσει πρόβλεψης (predictive maintenance) (Lee et al., 2014).

Υποστήριξη στρατηγικού σχεδιασμού

Η μηχανική μάθηση μπορεί να υποστηρίξει μακροπρόθεσμες αποφάσεις με βάση σενάρια και προβλέψεις που προκύπτουν από ιστορικά δεδομένα. Παραδείγματα αποτελούν η καλύτερη κατανομή κοινωνικών πόρων και η βελτίωση της ποιότητας των παρεχόμενων υπηρεσιών μέσω τεκμηριωμένης ανάλυσης των δεδομένων (Wirtz et al., 2019).

Η ενσωμάτωση των τεχνικών μηχανικής μάθησης σε Συστήματα Υποστήριξης Αποφάσεων (DSS) δεν αποσκοπεί στην αντικατάσταση της ανθρώπινης κρίσης αλλά στην συστηματική υποστήριξή της (Shim et al., 2002). Στους ΟΤΑ σε επίπεδο διοίκησης δίνει τη δυνατότητα λήψης τεκμηριωμένων αποφάσεων, μεταβαίνοντας από την εμπειρική κρίση στην κρίση που βασίζεται στην ανάλυση των πραγματικών λειτουργικών δεδομένων (data-driven policy). Απαραίτητη προϋπόθεση φυσικά, αποτελεί η εφαρμογή κατάλληλων μηχανισμών διακυβέρνησης και ελέγχου της ποιότητας των δεδομένων ώστε να διασφαλίζεται η αξιοπιστία τους στη διαδικασία λήψης αποφάσεων (Janssen et al., 2017). Με τον τρόπο αυτό, μπορεί να μειωθεί η αβεβαιότητα που δημιουργεί η εμπειρική κρίση ενισχύοντας τη διαφάνεια και τη λογοδοσία των διοικητικών διαδικασιών.

2. Θεωρητικό Υπόβαθρο

2.1 Εισαγωγή στη Μηχανική Μάθηση

Η Μηχανική Μάθηση (Machine Learning – ML) αποτελεί έναν θεμελιώδη και ραγδαία αναπτυσσόμενο επιστημονικό κλάδο της Τεχνητής Νοημοσύνης (Artificial Intelligence – AI), ο οποίος εξειδικεύεται στη μελέτη και ανάπτυξη αλγορίθμων που επιτρέπουν στα υπολογιστικά συστήματα να μαθαίνουν από δεδομένα, να εντοπίζουν πρότυπα και να βελτιώνουν την απόδοσή τους σε συγκεκριμένες εργασίες χωρίς ρητό προγραμματισμό (Jordan & Mitchell, 2015; Witten et al., 2016). Ιστορικά συνδέεται στενά με την πρόοδο των μαθηματικών και της στατιστικής, την αύξηση της διατιθέμενης υπολογιστικής ισχύος σε συνδυασμό με τις προσιτές οικονομικές απαιτήσεις απόκτησής της, καθώς και με την αυξανόμενη ανάγκη για την αυτοματοποιημένη λήψη αποφάσεων σε δυναμικά περιβάλλοντα που μεταβάλλονται διαρκώς.

Ένας από τους πιο καθιερωμένους ορισμούς της Μηχανικής Μάθησης διατυπώθηκε από τον Tom Mitchell ως εξής:

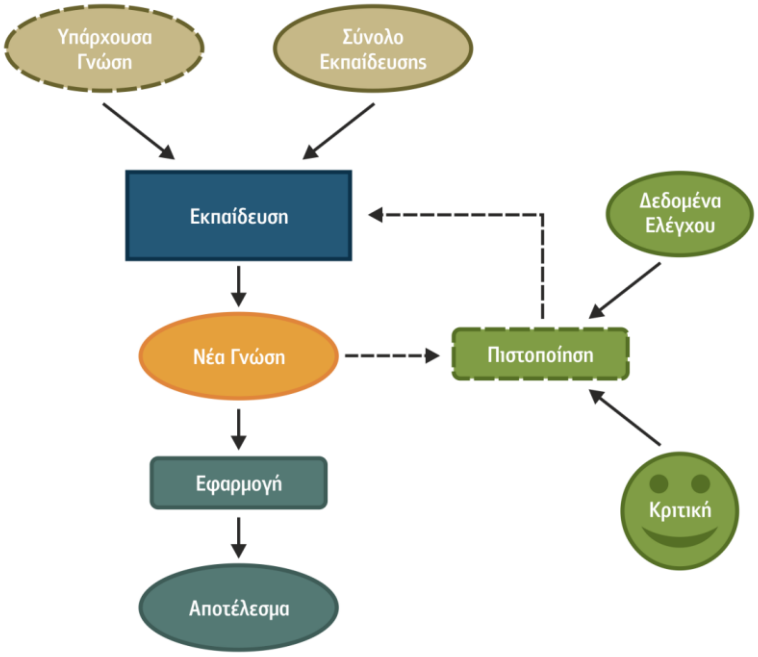
«*Ένα πρόγραμμα υπολογιστή λέμε ότι μαθαίνει από την εμπειρία E ως προς κάποια κλάση εργασιών T και μέτρο απόδοσης P, αν η απόδοσή του σε εργασίες από το T, όπως μετριέται από το P, βελτιώνεται μέσω της εμπειρίας E*» (Mitchell, 1997).

Ο ορισμός αυτός αναδεικνύει τη δυναμική φύση της μάθησης μέσω εμπειρίας (E), καθώς και τη μέτρηση της απόδοσης (P) ως βασικό κριτήριο αξιολόγησης. Σε αντίθεση με τις παραδοσιακές μεθόδους προγραμματισμού όπου ο αλγοριθμικός μηχανισμός επίλυσης προβλημάτων είναι προκαθορισμένος, στη Μηχανική Μάθηση το σύστημα «ανακαλύπτει» τη γνώση μέσα από τα ίδια τα δεδομένα με σκοπό να **γενικεύει** και να προσαρμόζεται σε νέα, άγνωστα δεδομένα με βάση τα μοτίβα που έχει εντοπίσει από την εμπειρία του, βελτιώνοντας την απόδοσή του.

Σημειώνεται ότι παρόλο που οι όροι Μηχανική Μάθηση και Εξόρυξη Γνώσης από Δεδομένα (Knowledge Discovery in Databases - KDD) συχνά χρησιμοποιούνται αλληλοκαλυπτικά, ωστόσο δεν είναι ταυτόσημοι. Η Μηχανική Μάθηση αποτελεί επιμέρους στάδιο της ευρύτερης διαδικασίας της Εξόρυξης Γνώσης από Δεδομένα (data mining). Η Μηχανική Μάθηση επικεντρώνεται στην ανάπτυξη μοντέλων χρησιμοποιώντας

αλγόριθμους που μπορούν να γενικεύσουν από παραδείγματα (στιγμιότυπα-παρατηρήσεις) και να πραγματοποιούν προβλέψεις ή ταξινομήσεις (Witten et al., 2016).

Η Εξόρυξη Γνώσης από Δεδομένα (KDD) είναι μια διαδικασία που περιλαμβάνει διακριτά στάδια όπως η προεπεξεργασία, η ανάλυση, η εξόρυξη, η αξιολόγηση και ερμηνεία των δεδομένων αξιοποιώντας τεχνικές όπως η μηχανική μάθηση, η οποία αποτελεί το εργαλείο που αφορά μέρος της διαδικασίας του «data mining», παρέχοντας μοντέλα πρόβλεψης ή αναγνώρισης προτύπων που αξιοποιούνται στην ερμηνεία των δεδομένων.



Σχήμα 2.1 Λειτουργία αλγορίθμων Μηχανικής Μάθησης (πηγή https://repository.kallipos.gr/handle/11419/3382?locale=el)

2.2 Κατηγορίες Μηχανικής Μάθησης

Τα μοντέλα Μηχανικής Μάθησης (Machine Learning – ML) κατατάσσονται σε τρεις βασικές κατηγορίες ανάλογα με τη φύση της μάθησης και τη διαθεσιμότητα των εποπτευόμενων δεδομένων: **επιβλεπόμενη μάθηση (supervised learning), μη επιβλεπόμενη μάθηση (unsupervised learning)** και **ενισχυτική μάθηση (reinforcement learning)**. Η διαφορά ανάμεσα στις τρεις κατηγορίες έγκειται στη δομή των δεδομένων εκπαίδευσης, στον τρόπο με τον οποίο ο αλγόριθμος μαθαίνει από αυτά και στους στόχους της μάθησης (Goodfellow et al., 2016).

2.2.1 Επιβλεπόμενη Μάθηση (Supervised Learning)

Η επιβλεπόμενη μάθηση είναι μία από τις πλέον θεμελιώδεις μεθόδους της Μηχανικής Μάθησης και χρησιμοποιείται κυρίως για την ανάπτυξη προβλεπτικών μοντέλων. Η βασική της αρχή συνίσταται στην εκπαίδευση ενός αλγορίθμου πάνω σε ένα σύνολο παραδειγμάτων (training set), το οποίο περιλαμβάνει γνωστές εισόδους (predictor variables ή attributes) και τις αντίστοιχες επιθυμητές εξόδους (target values ή labels).

Κατά τη διαδικασία εκπαίδευσης το μοντέλο επιχειρεί να μάθει τη συσχέτιση μεταξύ των μεταβλητών εισόδου και της εξαρτημένης μεταβλητής εξόδου, δημιουργώντας μια συνάρτηση απεικόνισης ή πρόγνωσης (predictor function) της μορφής:

$$f: X \rightarrow Y$$

όπου X είναι το σύνολο των εισόδων και Y το σύνολο των αντιστοίχων εξόδων. Ο στόχος είναι να αποκτήσει το μοντέλο μέσω της μάθησης την ικανότητα της γενίκευσης ώστε να προβλέπει σωστά την έξοδο για νέα, άγνωστα δεδομένα.

Αφού ολοκληρωθεί η εκπαίδευση, η αξιολόγηση της απόδοσης του μοντέλου πραγματοποιείται με την εφαρμογή του σε ξεχωριστό σύνολο δοκιμής (test set) το οποίο δεν έχει χρησιμοποιηθεί κατά τη φάση της εκμάθησης. Οι προβλέψεις συγκρίνονται με τις πραγματικές τιμές, ενώ μετρικές απόδοσης όπως η ακρίβεια (accuracy), το μέσο τετραγωνικό σφάλμα (MSE) ή η F1-score χρησιμοποιούνται για να εκτιμηθεί η αποτελεσματικότητα του μοντέλου (Witten et al., 2016). Αλγόριθμοι επιβλεπόμενης μηχανικής μάθησης:

- **Δέντρα Απόφασης (Decision Trees):** Χρησιμοποιούν μια ιεραρχική, δενδροειδή δομή για να καταλήξουν σε προβλέψεις ή αποφάσεις με βάση διαδοχικές ερωτήσεις (διαχωρισμούς) στα χαρακτηριστικά των δεδομένων. Παρέχουν καλή ερμηνευσιμότητα, δεν απαιτούν κανονικοποίηση, αλλά μπορεί να οδηγήσουν σε υπερεκπαίδευση (overfitting) και σε θορυβώδη δεδομένα (λανθασμένες αποφάσεις). Εφαρμογές: Διάγνωση, χρηματοπιστωτικές αποφάσεις, ανάλυση κινδύνου.
- **Κ-Κοντινότεροι Γείτονες (K-Nearest Neighbors – KNN):** Πρόκειται για έναν μη παραμετρικό αλγόριθμο που βασίζεται στην ομοιότητα μεταξύ παραδειγμάτων για να προβλέψει την κατηγορία ή την τιμή ενός νέου δείγματος. Η πρόβλεψη γίνεται βάσει των

«k» πλησιέστερων γειτόνων ενός νέου δείγματος σύμφωνα με κάποιο μέτρο απόστασης (π.χ. Ευκλείδεια απόσταση). Εφαρμογές: Αναγνώριση εικόνων, βιομετρικά συστήματα.

- **Λογιστική Παλινδρόμηση (Logistic Regression):** Κατάλληλη για δυαδικά προβλήματα ταξινόμησης, προβλέπει την πιθανότητα ένταξης μιας παρατήρησης (ενός δείγματος) σε μια κατηγορία χρησιμοποιώντας μια λογιστική συνάρτηση. Εφαρμογές: Ιατρική διάγνωση, προβλέψεις binary.
- **Random Forests:** Αποτελούνται από πολλά ανεξάρτητα δέντρα απόφασης (ensemble learning) που εκπαιδεύονται τυχαία σε διαφορετικά υποσύνολα δεδομένων και χαρακτηριστικών. Παρέχουν υψηλή ακρίβεια, καλή γενίκευση και είναι ανθεκτικά στο overfitting. Η τελική πρόβλεψη προκύπτει από το μέσο όρο των προβλέψεων όλων των δέντρων. Εφαρμογές: Ανάλυση ρίσκου, βιολογία, προβλέψεις κειμένου.
- **Μηχανές Διανυσμάτων Υποστήριξης (Support Vector Machines – SVM):** Δημιουργούν υπερεπίπεδα (hyperplanes) που διαχωρίζουν τις κατηγορίες των δεδομένων με τον καλύτερο δυνατό τρόπο εστιάζοντας στα όρια μεταξύ τους (margins). Είναι αποτελεσματικοί σε υψηλής διαστασιμότητας δεδομένα. Εφαρμογές: Αναγνώριση κειμένου/εικόνας.
- **Νευρωνικά Δίκτυα (Neural Networks):** Εμπνευσμένα από τη βιολογία, συνίστανται από στρώματα τεχνητών νευρώνων (αλγοριθμικών μονάδων επεξεργασίας δεδομένων). Μπορούν να αποδώσουν καλά σε πολύπλοκα προβλήματα πρόβλεψης και ταξινόμησης, ιδιαίτερα όταν ενσωματώνονται σε βαθιές αρχιτεκτονικές μάθησης (deep learning). Εφαρμογές: Όραση υπολογιστών, φυσική γλώσσα, φωνητική αναγνώριση.
- **Γραμμική Παλινδρόμηση (Linear Regression):** Χρησιμοποιείται για πρόβλεψη συνεχών τιμών (regression tasks) επιχειρώντας να μοντελοποιήσει τη σχέση μεταξύ μιας ή περισσότερων ανεξάρτητων μεταβλητών (X) και μιας εξαρτημένης μεταβλητής (y), χρησιμοποιώντας μια ευθεία γραμμή. Εφαρμογές: Πρόβλεψη κατανάλωσης ενέργειας ανά περιοχή ή εποχή.
- **Gradient Boosting:** Δημιουργεί ένα σύνολο (ensemble) αδύναμων μοντέλων, συνήθως δέντρων απόφασης (decision trees), όπου μέσω της διαδοχικής εκπαίδευσης επιχειρείται η διόρθωση των σφαλμάτων πρόβλεψης με σκοπό τη σταδιακή βελτιστοποίηση (gradient optimization). Εφαρμογές: Πρόβλεψη κόστους έργων, κατανάλωσης νερού, αναγκών πρόνοιας.

2.2.2 Μη Επιβλεπόμενη Μάθηση (Unsupervised Learning)

Τα δεδομένα εισόδου δεν συνοδεύονται από ετικέτες-στόχους ή αναμενόμενες εξόδους. Ο αλγόριθμος προσπαθεί να εντοπίσει και να μοντελοποιήσει πρότυπα, ομάδες ή συσχετίσεις. Η προσέγγιση αυτή είναι ιδιαίτερα χρήσιμη για ανάλυση συσχετισμών, μείωση διαστάσεων και κυρίως για **ομαδοποίηση (clustering)**, όπου τα δεδομένα κατανέμονται σε ομάδες με βάση την εγγύτητα ή την ομοιότητά τους π.χ. ομαδοποίηση γεωγραφικών περιοχών με βάση την κοινωνικοοικονομική συμπεριφορά (Witten et al., 2016). Αλγόριθμοι μη επιβλεπόμενης μηχανικής μάθησης:

• **K-Means Clustering:** Ο K-Means προσπαθεί να χωρίσει ένα σύνολο δεδομένων σε *K* ομάδες (clusters), έτσι ώστε κάθε παρατήρηση να ανήκει στην ομάδα (cluster) του πλησιέστερου κέντρου. Αρχικά επιλέγονται K τυχαία σημεία ως αρχικά κέντρα και κάθε σημείο δεδομένων ενσωματώνεται στο πλησιέστερο κέντρο με βάση την Ευκλείδεια απόσταση. Εν συνεχεία υπολογίζονται τα νέα κέντρα με βάση το μέσο όρο των σημείων κάθε ομάδας. Εφαρμογές: Ομαδοποίηση περιοχών βάσει δραστηριότητας (οικιστική, εμπορική, βιομηχανική)

• **DBSCAN:** Αλγόριθμος συσταδοποίησης που ανακαλύπτει ομάδες (clusters) με βάση την πυκνότητα σημείων στο χώρο. Διακρίνει συνεκτικές περιοχές υψηλής πυκνότητας από αραιές περιοχές που χαρακτηρίζονται ως θόρυβος. Δεν απαιτεί ορισμό αρχικών κέντρων ως ομάδες (K-Means) ενώ αντιμετωπίζει την συσταδοποίηση οποιουδήποτε σχήματος βασιζόμενος σε δύο παραμέτρους: eps (ακτίνα γειτονιάς) και min_samples (ελάχιστος αριθμός σημείων πυκνής περιοχής). Εφαρμογές: Χαρτογράφηση κοινωνικής συμπεριφοράς, ανάλυση κυκλοφορίας (GPS διαδρομές).

2.2.3 Ενισχυτική Μάθηση (Reinforcement Learning)

Ο αλγόριθμος μαθαίνει μέσω δοκιμής και σφάλματος, αλληλεπιδρώντας με ένα δυναμικό περιβάλλον. Σε αντίθεση με άλλες μεθόδους που βασίζονται σε προκαθορισμένα ζεύγη εισόδου–εξόδου, εδώ ο αλγόριθμος ή πράκτορας (agent) λαμβάνει αποφάσεις, εκτελεί ενέργειες και λαμβάνει θετική ή αρνητική ενίσχυση ανάλογα με την απόδοσή του μέσω ενός μηχανισμού ανταμοιβής (reward signal). Με την πάροδο του χρόνου και την επανάληψη αλληλεπιδράσεων, ο πράκτορας αναπτύσσει μια πολιτική (policy), δηλαδή μια στρατηγική που καθορίζει ποιες ενέργειες είναι βέλτιστες σε κάθε κατάσταση, με σκοπό τη μεγιστοποίηση της θετικά σωρευτικής ανταμοιβής (Goodfellow et al., 2016). Η ενισχυτική

μάθηση εφαρμόζεται κυρίως σε δυναμικά περιβάλλοντα με χαρακτηριστικά αβεβαιότητας όπου απαιτείται συνεχής προσαρμογή και λήψη αποφάσεων σε πραγματικό χρόνο.

2.3 Μεθοδολογία εφαρμογής Μοντέλου Μηχανικής Μάθησης

Η διαδικασία εφαρμογής της Μηχανικής Μάθησης περιλαμβάνει μια σειρά βημάτων που στόχο έχουν την **κατασκευή, εκπαίδευση, αξιολόγηση και αξιοποίηση** ενός μοντέλου πρόβλεψης, ταξινόμησης ή ομαδοποίησης με βάση τα δεδομένα που έχουμε στη διάθεσή μας. Η διαδικασία αυτή συνοψίζεται στα εξής βασικά βήματα:

- **Κατανόηση και Καθορισμός του Προβλήματος**

Πρέπει να οριστεί με απόλυτη σαφήνεια τι επιχειρούμε να λύσουμε και ποιος είναι ο επιχειρησιακός στόχος της επίλυσης του προβλήματος (Janssen et al., 2017). Ποιες είναι οι μεταβλητές-δεδομένα **εισόδου** που καθορίζουν την πρόβλεψη-έξοδο και άρα την απόφαση για την επίτευξη του στόχου; Είναι πρόβλημα **παλινδρόμησης**, **ταξινόμησης**, ή **ομαδοποίησης**;

- **Συλλογή και Επιλογή Δεδομένων**

Η ποιότητα, η πληρότητα και η εγκυρότητα των δεδομένων αποτελούν βασικούς παράγοντες για την επιτυχία κάθε συστήματος Μηχανικής Μάθησης. Η συγκέντρωση όλων των απαραίτητων δεδομένων που θα αξιοποιηθούν για την εκπαίδευση, αξιολόγηση και πρόβλεψη του μοντέλου, μπορούν να προέρχονται από αρχεία Excel/CSV, Βάσεις δεδομένων, Open Data portals, Αισθητήρες IoT και Συμμετοχικές Πλατφόρμες Αλληλεπίδρασης με τους πολίτες (CRM) (Misuraca & van Noordt, 2020). Για την εκπαίδευση του μοντέλου, σε προβλήματα παλινδρόμησης, επιλέγονται εκείνα τα χαρακτηριστικά (features – attributes) των δεδομένων, τα οποία προσδιορίζουν με το καλύτερο τρόπο τη σχέση μεταξύ εισόδου και εξόδου - δηλαδή της τιμής στόχου.

- **Καθαρισμός και Προεπεξεργασία Δεδομένων**

Τα δεδομένα μπορεί να παρουσιάζουν ελλιπείς τιμές λόγω σφαλμάτων καταγραφής ή ανθρώπινου λάθους ανάλογα με την πηγή από την οποία αντλούνται. Μπορεί να περιέχουν μη λογικές τιμές που δεν σχετίζονται με την πραγματικότητα (θόρυβος) ή να βρίσκονται εκτός ορίων (outliers-ανωμαλίες). Συνεπώς είναι σημαντικό να βελτιωθεί η αξιοπιστία των αρχικών δεδομένων επεμβαίνοντας στις τιμές των χαρακτηριστικών (features) των

επιμέρους εγγραφών με αντικατάσταση (μέσος όρος, συχνότερη τιμή), κωδικοποίηση (0,1) ή κανονικοποίηση (εύρος τιμών από 0 έως 1) των προβληματικών τιμών ή και διαγραφή ολόκληρων εγγραφών (Witten et al., 2016).

- **Εξερεύνηση και Ανάλυση Δεδομένων (Exploratory Data Analysis - EDA)**

Είναι σημαντικό να γίνει κατανοητή η δομή των δεδομένων καθώς επηρεάζει άμεσα την ανάλυση τους με τον έγκαιρο εντοπισμό ανωμαλιών (outliers), ελλιπών ή προβληματικών τιμών, καθώς και για τον προσδιορισμό των συσχετίσεων μεταξύ των μεταβλητών εισόδου και της μεταβλητής-στόχου σε προβλήματα παλινδρόμησης. Η διαδικασία αυτή είναι απαραίτητη για την προληπτική διόρθωση των σφαλμάτων που ενδέχεται να επηρεάσουν την απόδοση, τη γενίκευση ή την αξιοπιστία του μοντέλου μηχανικής μάθησης. Η διερεύνηση πραγματοποιείται κυρίως μέσω τεχνικών οπτικοποίησης, όπως τα ιστογράμματα, τα διαγράμματα διασποράς και οι πίνακες συσχέτισης, που προσφέρουν μια εποπτική κατανόηση της κατανομής των δεδομένων και των μεταξύ τους σχέσεων. Η οπτική αναπαράσταση διευκολύνει την ερμηνεία της συμπεριφοράς των χαρακτηριστικών για την επιλογή ή τον μετασχηματισμό τους συμβάλλοντας στη διαμόρφωση τεκμηριωμένων αποφάσεων (Witten et al., 2016).

- **Διαχωρισμός Εκπαιδευτικού και Ελεγκτικού Συνόλου Δεδομένων**

Ο διαχωρισμός των δεδομένων σε εκπαιδευτικό (training set) και ελεγκτικό (test set) σύνολο επιτρέπει την εκπαίδευση του μοντέλου σε ένα υποσύνολο των δεδομένων και στη συνέχεια την αξιολόγηση του με βάση την ικανότητα να γενικεύει σε νέα, άγνωστα δεδομένα που δεν χρησιμοποιήθηκαν κατά την εκπαίδευση (Witten et al., 2016). Η κατανομή ανάμεσα στα δύο σύνολα σχετίζεται με το μέγεθος αλλά και τα χαρακτηριστικά των δεδομένων. Συνήθως υιοθετείται μια αναλογία 80:20

- **Επιλογή και Εκπαίδευση Αλγορίθμου**

Η επιλογή του κατάλληλου αλγόριθμου μηχανικής μάθησης εξαρτάται από το είδος του προβλήματος (π.χ. ταξινόμηση, παλινδρόμηση, ομαδοποίηση), τη δομή και τον όγκο της διατιθέμενης πληροφορίας που μας παρέχουν τα δεδομένα καθώς και από τις δυνατότητες των υπολογιστικών συστημάτων που έχουμε στη διάθεσή μας. Κατά τη διαδικασία της εκπαίδευσης στην επιβλεπόμενη μηχανική μάθηση, ο αλγόριθμος προσπαθεί να μάθει τη συσχέτιση μεταξύ των χαρακτηριστικών εισόδου (features) και της εξαρτημένης μεταβλητής-στόχου (target label). Αυτό επιτυγχάνεται με την κατασκευή μιας συνάρτησης με σκοπό την ελαχιστοποίηση του κόστους (loss function), όπως για παράδειγμα η μέση

τετραγωνική απόκλιση (Mean Squared Error - MSE) για προβλήματα παλινδρόμησης ή η διασταυρούμενη εντροπία (Cross-Entropy Loss) για προβλήματα ταξινόμησης.

- **Αξιολόγηση Απόδοσης**

Η αξιολόγηση της απόδοσης ενός μοντέλου μηχανικής μάθησης βασίζεται στη χρήση κατάλληλων μετρικών απόδοσης που μας παρέχουν τη δυνατότητα να προσδιορίσουμε την ικανότητα του να γενικεύει σε νέα, άγνωστα δεδομένα. Σε προβλήματα ταξινόμησης χρησιμοποιούνται μετρικές όπως η ακρίβεια (accuracy), η ευαισθησία (recall), η ακρίβεια πρόβλεψης (precision) και ο συντελεστής F1 (F1-score) ενώ σε προβλήματα παλινδρόμησης συνήθως χρησιμοποιείται η μέση τετραγωνική απόκλιση (MSE) (Witten et al., 2016). Η επιλογή της κατάλληλης μετρικής μας παρέχει την απαιτούμενη πληροφόρηση που αφορά την ποιότητα των προβλέψεων του εφαρμοζόμενου μοντέλου για την αποφυγή σφαλμάτων, όπως αυτό της υπερπροσαρμογής (overfitting) ή της υποπροσαρμογής (underfitting).

- **Βελτιστοποίηση Μοντέλου (Model Tuning)**

Σε έναν αλγόριθμο μηχανικής μάθησης επιβάλλονται ρυθμίσεις εκ των προτέρων (υπέρπαράμετροι) οι οποίες δεν μπορούν να προσδιοριστούν κατά το στάδιο της εκπαίδευσης από το διατιθέμενο σύνολο δεδομένων (dataset).

Τέτοιες ρυθμίσεις έχουν να κάνουν με το ρυθμό μάθησης (learning rate), τον αριθμό των κόμβων σε δέντρα απόφασης ή το πλήθος των επιπέδων σε νευρωνικό δίκτυο αλλά και την πολυπλοκότητα του μηχανισμού ανταμοιβών σε αλγορίθμους SVM (Witten et al., 2016).

Η βελτίωση της απόδοσης του αλγόριθμου που εφαρμόζει το μοντέλο της μηχανικής μάθησης, μπορεί να επιτευχθεί επίσης, με την επιλογή συγκεκριμένων χαρακτηριστικών (feature selection) του συνόλου των δεδομένων ή με την προσθήκη νέων.

- **Αποθήκευση και Ενσωμάτωση σε Σύστημα ή Dashboard**

Μετά την εκπαίδευση, αξιολόγηση και αποθήκευση του μοντέλου ακολουθεί η ενσωμάτωσή του σε ένα λειτουργικό πληροφοριακό σύστημα ή διαδραστικό περιβάλλον (dashboard), προκειμένου να μπορεί να χρησιμοποιείται σε πραγματικό χρόνο ή κατά απαίτηση από τελικούς χρήστες, οργανισμούς ή άλλες εφαρμογές. Η διαδικασία αυτή αποτελεί το τελικό στάδιο ενός κύκλου μηχανικής μάθησης και συνδέεται άμεσα με την παραγωγική αξιοποίηση της τεχνητής νοημοσύνης σε επιχειρησιακό ή δημόσιο περιβάλλον (Janssen et al., 2017). Η παρουσίαση των αποτελεσμάτων μέσω dashboard επιτρέπει την οπτική ερμηνεία των προβλέψεων από χρήστες χωρίς τεχνικό υπόβαθρο, ενώ παράλληλα

προσφέρει δυναμικό έλεγχο και αλληλεπίδραση με τα δεδομένα. Η σωστή ενσωμάτωση εξασφαλίζει την αξία της επέκτασης του μοντέλου πέρα από το ερευνητικό στάδιο και σε πραγματικό επιχειρησιακό περιβάλλον.

2.4 Συστήματα Υποστήριξης Αποφάσεων (DSS)

Τα Συστήματα Υποστήριξης Αποφάσεων (Decision Support Systems – DSS) αποτελούν πληροφοριακά υπολογιστικά συστήματα που έχουν ως κύριο σκοπό την υποστήριξη της ανθρώπινης κρίσης κατά τη λήψη σύνθετων, πολύπλοκων ή στρατηγικής φύσης αποφάσεων. Συνδυάζουν δεδομένα, τεχνικές ανάλυσης και υπολογιστικά μοντέλα παρέχοντας στους χρήστες τεκμηριωμένες, ποσοτικοποιημένες και διαφανείς εναλλακτικές λύσεις. Σε αντίθεση με τα αυτοματοποιημένα συστήματα που εκτελούν ρητά καθορισμένες εντολές, ο ρόλος των DSS είναι αποκλειστικά συμβουλευτικός και δεν αντικαθιστούν σε καμία περίπτωση τον ανθρώπινο παράγοντα αλλά λειτουργούν συμπληρωματικά προς αυτόν ενισχύοντας τη διαδικασία ανάλυσης, πρόβλεψης και αξιολόγησης των δεδομένων, ιδίως σε σύνθετα περιβάλλοντα αβεβαιότητας (Shim et al., 2002).

Σύμφωνα με την κλασική θεώρηση του Sprague (1980) ένα DSS αποτελείται από τρία βασικά δομικά υποσυστήματα:

Βάση Δεδομένων (Data Management): Αποθηκεύει και διαχειρίζεται τα δεδομένα που απαιτούνται για τη λήψη αποφάσεων, τόσο από εσωτερικές πηγές του οργανισμού όσο και από εξωτερικές πηγές.

Βάση Μοντέλων Ανάλυσης (Model Management): Περιλαμβάνει το σύνολο των αναλυτικών εργαλείων, μαθηματικών μοντέλων και αλγορίθμων που χρησιμοποιούνται για την επεξεργασία και ερμηνεία των δεδομένων.

Διεπαφή Χρήστη (User Interface): Επιτρέπει την αλληλεπίδραση του τελικού χρήστη με το σύστημα προσφέροντας φιλικά και διαδραστικά εργαλεία για την εισαγωγή δεδομένων, την προβολή αποτελεσμάτων και την πλοήγηση στις λειτουργίες του DSS.

3. Ερευνητική Μεθοδολογία

Η έρευνα για την εφαρμογή και το ρόλο των τεχνικών Μηχανικής Μάθησης με στόχο την εξόρυξη γνώσης για την υποστήριξη λήψης τεκμηριωμένων αποφάσεων από την Διοίκηση και τις αρμόδιες υπηρεσίες σε έναν ΟΤΑ, θα γίνει στον κρίσιμο λειτουργικό τομέα της ενεργειακής κατανάλωσης των δημοτικών υποδομών. Για τις ανάγκες της παρούσας εργασίας χρησιμοποιήθηκε η γλώσσα Python 3.7 με τις απαραίτητες βιβλιοθήκες μέσω του εργαλείου PyCharm. Η εφαρμογή έγινε σε Η/Υ LENOVO ideacentre 510-15IKL με μνήμη RAM 12 GB, SSD, CPU Intel® Core™ i7-7700 3.60 GHz, WINDOWS 10 64bit.

3.1. Ενεργειακή Κατανάλωση Δημοτικών Υποδομών

Στο κεφάλαιο αυτό παρουσιάζεται η εφαρμογή μοντέλων Μηχανικής Μάθησης (Machine Learning – ML) με στόχο την πρόβλεψη της μηνιαίας κατανάλωσης ηλεκτρικής ενέργειας σε δημοτικά κτίρια και υποδομές ενός Δήμου (ΟΤΑ). Η μελέτη εφαρμογής έχει ως στόχο την ανάδειξη του τρόπου με τον οποίο ένα προγνωστικό μοντέλο μπορεί να συμβάλλει στον στρατηγικό σχεδιασμό της ενεργειακής πολιτικής ενός Δήμου ενισχύοντας τις αποφάσεις της Διοίκησης για την ορθολογική διαχείριση των οικονομικών πόρων, την πρόληψη σπατάλης και τη βελτίωση της ενεργειακής απόδοσης. Μέσα από την ανάλυση ιστορικών δεδομένων κατανάλωσης το μοντέλο μπορεί να αποτελέσει σημαντικό εργαλείο υποστήριξης αποφάσεων, παρέχοντας στη Διοίκηση τεκμηριωμένες προβλέψεις για την καλύτερη κατανομή των ενεργειακών δαπανών και την έγκαιρη ανίχνευση αδικαιολόγητων αποκλίσεων (Ahmad et al., 2017).

3.2. Περιγραφή προβλήματος

Οι Δήμοι διαχειρίζονται πλήθος δημοτικών κτιρίων και εγκαταστάσεων όπως κτίρια στέγασης δημοτικών υπηρεσιών, σχολικά κτίρια, εγκαταστάσεις αθλητισμού και πολιτισμού, εγκαταστάσεις οδοφωτισμού, ύδρευσης και άρδευσης. Οι δαπάνες που σχετίζονται με την κατανάλωση της ενέργειας αποτελούν κρίσιμο παράγοντα τόσο για την

ομαλή εκτέλεση του ετήσιου δημοτικού προϋπολογισμού όσο και για την επίτευξη στόχων οικονομικής αλλά και περιβαλλοντικής βιωσιμότητας σε έναν Δήμο. Η έλλειψη της δυνατότητας για δυναμική πρόβλεψη της μηνιαίας κατανάλωσης ενέργειας (kWh) έχει ως αποτέλεσμα την μη έγκαιρη αναγνώριση φαινομένων ασυνήθιστης ενεργειακής συμπεριφοράς των δημοτικών υποδομών που θα οδηγούσε σε στοχευμένες διορθωτικές παρεμβάσεις. Η δυνατότητα αξιοποίησης των δεδομένων ενεργειακής κατανάλωσης βοηθά τις αρμόδιες τεχνικές υπηρεσίες στην παρακολούθηση και ανάλυση αυτών, σε πραγματικό χρόνο, ενισχύοντας την τεκμηρίωση των αποφάσεων που θα πρέπει να ληφθούν σε συνεργασία με τη Διοίκηση.

3.3. Συλλογή Δεδομένων

Τα δεδομένα που θα χρησιμοποιηθούν για τους σκοπούς της παρούσας εργασίας αφορούν συγκεκριμένο Δήμο, για τα οποία έχει γίνει η κατάλληλη ανωνυμοποίηση. Η άντληση τους έχει προκύψει από τους μηνιαίους λογαριασμούς του παρόχου ηλεκτρικής ενέργειας των ετών 2020 έως και το 2025, οι οποίοι ενσωματώνονται σε ειδικό Σύστημα Διαχείρισης Λογαριασμών Ηλεκτρικού Ρεύματος που έχει προμηθευθεί και χρησιμοποιεί ο συγκεκριμένος Δήμος.

3.4. Προεπεξεργασία αρχικών δεδομένων

Το αρχείο με τα ιστορικά δεδομένα της ενεργειακής κατανάλωσης δημοτικών κτιρίων και υποδομών (energ_2020-2025.csv) που διατέθηκε από το συγκεκριμένο Δήμο, περιέχει 41.572 εγγραφές (η πρώτη γραμμή είναι επικεφαλίδα) με πληροφορίες οι οποίες πρέπει να ελεγχθούν για την ποιότητά τους πριν την εφαρμογή σε αυτά μεθόδων μηχανικής μάθησης με σκοπό την δημιουργία προβλεπτικών μοντέλων. Ο έλεγχος αυτός έχει να κάνει με την αντιμετώπιση της παρουσίας θορύβου, δηλαδή τιμών που δεν αντιπροσωπεύουν πραγματικά γεγονότα, ακραίων τιμών, δηλαδή τιμών που αποκλίνουν σε πολύ μεγάλο βαθμό, ελλιπών ή διπλότυπων εγγραφών στο σύνολο των παρατηρήσεων (εγγραφών) του διατιθέμενου αρχείου (data set).

3.4.1 Πληροφορίες αρχείου ενεργειακών δεδομένων

Πριν από οποιαδήποτε ενέργεια ελέγχου των δεδομένων του αρχείου, πρέπει να εμφανίσουμε τις πληροφορίες που εμπεριέχονται σε αυτό. Όπως ήδη αναφέρθηκε σε όλα τα στάδια που θα περιγράψουμε χρησιμοποιείται η γλώσσα Python με τις αντίστοιχες απαραίτητες βιβλιοθήκες. Με δεδομένο ότι ένα αρχείο CSV (Comma Separated Values) υποδηλώνει ένα αρχείο κειμένου στο οποίο τα δεδομένα που εμφανίζονται σε διακριτές εγγραφές διαχωρίζονται με κόμμα (,) εντούτοις σε πολλές περιπτώσεις όταν εξάγονται από κάποια εφαρμογή (π.χ. Microsoft Excel, Βάσεις Δεδομένων, κτλ.) ο διαχωρισμός αυτός μπορεί να γίνεται με διαφορετικό οριοθέτη (delimiter). Για την άντληση και εμφάνιση των πληροφοριών του αρχείου εκτελούμε τον κώδικα της Εικόνας 3.1 πραγματοποιώντας την ανάγνωση ικανού δείγματος εγγραφών (8000) για την ανίχνευση και εντοπισμό του οριοθέτη (delimiter).

```
#Ορισμός τοπικής διαδρομής αποθηκευμένου αρχείου
filepath="C:/dataset/energ_2020-2025.csv"

#Άνοιγμα αρχείου και δημιουργία αντικειμένου ροής
with open(filepath, "r", encoding="utf-8-sig") as file:
    sample = file.read(8000) #διαβαζουμε δείγμα του αρχειου
    dialect = csv.Sniffer().sniff(sample) #μέθοδος εντοπισμού διαχωριστικού

print ("Οριοθέτης πεδίων αρχείου csv:", "(",dialect.delimiter,")",end='\n\n')
#Δημιουργία data frame
df = pd.read_csv(filepath, sep=dialect.delimiter, encoding="utf-8-sig")
#Πληροφορίες data set
print("Πληροφορίες dataset", end='\n\n')
print(df.info(), end='\n\n')
```

Εικόνα 3.1 Ανάγνωση αρχείου ενεργειακών δεδομένων

Διαπιστώνουμε ότι ο delimiter στο csv αρχείο που μας διατέθηκε είναι το ερωτηματικό (;) ενώ τα χαρακτηριστικά (μεταβλητές) φαίνονται στην Εικόνα 3.2

Το dataset αποτελείται από 41.571 εγγραφές και 19 στήλες χαρακτηριστικών (μεταβλητών) εκ των οποίων κάποιες εμφανίζουν διαφορετικό πλήθος εγγραφών γεγονός που υποδηλώνει την ύπαρξη ελλιπών τιμών (NaN). Αξιολογώντας τις ετικέτες (labels) των επιμέρους στηλών και την ποιοτική αξία των πληροφοριών που μας παρέχουν αυτές για την επίτευξη των ερευνητικών στόχων της παρούσας εργασίας, θα γίνει επιλογή συγκεκριμένων χαρακτηριστικών μετά τον έλεγχο που θα λάβει χώρα σε επόμενο στάδιο της προεπεξεργασίας των πρωτογενών δεδομένων .

```
Οριοθέτης πεδίων αρχείου csv: ( ; )

Πληροφορίες dataset

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 41571 entries, 0 to 41570
Data columns (total 19 columns):
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   Ετος                  41571 non-null  int64
 1   Μήνας                 41571 non-null  int64
 2   Ημ/νία Λογ/σμου       41571 non-null  object
 3   Αρ. Πολλαπλου         41571 non-null  object
 4   Ονομα Πολλαπλού       41571 non-null  object
 5   Αρ.Παροχής            41571 non-null  int64
 6   Τιμολόγιο             41571 non-null  object
 7   Ον.Πελάτη             41571 non-null  object
 8   Οδός                  41412 non-null  object
 9   Αριθ.                 3833 non-null   object
 10  Αρ.Μετρητή            41571 non-null  object
 11  Ημέρες Κατανάλωσης    41571 non-null  int64
 12  Συντ.Kwh              41571 non-null  int64
 13  Κατανάλωση KWh        41571 non-null  int64
 14  Αξία Ενέργειας        41571 non-null  object
 15  Συν Προμηθ Ρευματος   41571 non-null  object
 16  Σύνολο Λογαριασμού    41571 non-null  object
 17  Σύν.Τρεχ.Μήνα         41571 non-null  object
 18  Τύπος Λογ/σμου        40339 non-null  object
dtypes: int64(6), object(13)
```

Εικόνα 3.2 Πληροφορίες αρχείου ενεργειακών δεδομένων

3.4.2 Κατηγορίες χαρακτηριστικών ενεργειακών δεδομένων

Από τις πληροφορίες που έχουμε λάβει για το αρχείο παρατηρούμε ότι τα χαρακτηριστικά χωρίζονται σε:

- **Αριθμητικά (int/float):** «Έτος», «Μήνας», «Αρ. Παροχής», «Ημέρες Κατανάλωσης», «Συντ.Kwh», «Κατανάλωση Kwh» και
- **Κατηγορικά (object):** «Ημ/νία Λογ/σμου», «Αρ. Πολλαπλου», «Ονομα Πολλαπλού», «Τιμολόγιο|, «Ον.Πελάτη», «Οδός», «Αριθ.», «Αρ.Μετρητή», «Αξία Ενέργειας», «Συν Προμηθ Ρευματος», «Σύνολο Λογαριασμού», «Σύν.Τρεχ.Μήνα», «Τύπος Λογ/σμου».

Τα κατηγορικά χαρακτηριστικά: «**Αξία Ενέργειας**», «**Συν Προμηθ Ρευματος**», «**Σύνολο Λογαριασμού**», «**Σύν.Τρεχ.Μήνα**» αντιπροσωπεύουν οικονομικά ποσά άρα απαιτείται η μετατροπή τους σε αριθμητικά ώστε να έχουν εφαρμογή σε αυτά οι ιδιότητες (λειτουργίες) των αριθμών (=, ≠, <, ≤, >, ≥, +, -, *, /) σε περίπτωση επιλογής τους. Συνήθως στα πεδία αυτά μπορεί να εμφανίζονται μη αριθμητικοί χαρακτήρες όπως το σύμβολο του ευρώ (€), κόμματα (,) αντί για τελείες ή παύλες (-) αντί για μηδενικά (0). Οι χαρακτήρες αυτοί θα

πρέπει να αντικατασταθούν. Το χαρακτηριστικό «Αρ.Πολλαπλού» αποτελεί προφανώς την κωδικοποίηση του χαρακτηριστικού «Όνομα Πολλαπλού» άρα δεν έχει νόημα η εφαρμογή των ιδιοτήτων (λειτουργιών) των αριθμών. Συνεπώς αυτό το χαρακτηριστικό θα παραμείνει κατηγορικό. Το ίδιο ισχύει και για το κατηγορικό χαρακτηριστικό «Αρ.Μετρητή» το οποίο αποτελεί την κωδικοποιημένη αναφορά του μετρητή της παροχής του ρεύματος σε κάθε εγκατάσταση, οπότε δεν απαιτούνται οι ιδιότητες (λειτουργίες) των αριθμών. Με απλά λόγια δεν έχει νόημα να συγκρίνουμε αριθμητικά δύο αριθμούς μετρητών παροχής ρεύματος. Αντιθέτως, το χαρακτηριστικό «Ημ/νία Λογ/σμου» εμφανίζεται ως κατηγορικό με αποτέλεσμα αλγοριθμικά να αντιμετωπίζεται ως κείμενο-συμβολοσειρά. Τα ιστορικά δεδομένα του αρχείου της ενεργειακής κατανάλωσης των δημοτικών κτιρίων/υποδομών αποτελούνται από εγγραφές (παρατηρήσεις) που έχουν καταγραφεί διαδοχικά στο χρόνο (time series – χρονοσειρές) όπου η κατανάλωση του ρεύματος συνδέεται άμεσα με την ημερομηνία έκδοσης του λογαριασμού και τις ημέρες κατανάλωσης. Παρότι στο αρχείο μας υπάρχουν τα αριθμητικά χαρακτηριστικά «Έτος», «Μήνας» και «Ημέρες Κατανάλωσης», εντούτοις το χαρακτηριστικό «Ημ/νία Λογ/σμου» επειδή είναι βασικό για την έρευνά μας, θα πρέπει να μετατραπεί σε τύπο δεδομένων που αναπαριστά αποκλειστικά χρονική πληροφορία (datetime) και όχι κείμενο για να μπορέσουμε να αναλύσουμε και να προβλέψουμε χρονικά την ενεργειακή συμπεριφορά των δημοτικών κτιρίων/υποδομών (Εικόνα 3.3).

```
Ελεγχος πληροφοριών μετά τις μετατροπές

Ετος                          int64
Μήνας                         int64
Ημ/νία Λογ/σμου       datetime64[ns]
Αρ. Πολλαπλου                object
Ονομα Πολλαπλού              object
Αρ.Παροχής                    int64
Τιμολόγιο                    object
Ον.Πελάτη                    object
Οδός                         object
Αριθ.                        object
Αρ.Μετρητή                   object
Ημέρες Κατανάλωσης            int64
Συντ.Kwh                      int64
Κατανάλωση KWh                int64
Αξία Ενέργειας              float64
Συν Προμηθ Ρευματος         float64
Σύνολο Λογαριασμού          float64
Σύν.Τρεχ.Μήνα               float64
Τύπος Λογ/σμου               object
dtype: object
```

Εικόνα 3.3 Μετατροπή και εμφάνιση τύπων δεδομένων

3.4.3 Ανάλυση μοναδικότητας, εύρεση ελλιπών τιμών & διπλοεγγραφών

Με δεδομένο ότι στις πληροφορίες του αρχείου περιλαμβάνεται ο αριθμός παροχής («Αρ.Παροχής») και ο αριθμός μετρητή («Αρ.Μετρητή»), ελέγχουμε τον αριθμό των μοναδικών τιμών όλων των χαρακτηριστικών για να διαπιστώσουμε το πλήθος των διαφορετικών δημοτικών κτιρίων/υποδομών (Εικόνα 3.4).

```
#Πλήθος μοναδικών τιμών ανα χαρακτηριστικό
print("Πλήθος μοναδικών τιμών ανά χακτηριστικο", end='\n\n')
print(df.nunique(), end='\n\n')
```

Εικόνα 3.4 Κώδικας εμφάνισης μοναδικών τιμών

Από την ανάλυση των μοναδικών τιμών του αρχείου των ιστορικών δεδομένων όπως εμφανιζονται στην Εικόνα 3.5, παρατηρούμε ότι τα πρωτογενή δεδομένα (row data) αφορούν λογαριασμούς που έχουν εκδοθεί για έξι (6) έτη ενώ καλύπτονται και οι δώδεκα (12) μήνες του χρόνου. Έχουμε 868 διαφορετικές παροχές ρεύματος («Αρ.Παροχής») και 1008 μετρητές κατανάλωσης ρεύματος («Αρ.Μετρητή») οι οποίοι αφορούν με βάση το χαρακτηριστικό «Ον.Πελάτη» 754 δημοτικές υποδομές.

```
Πλήθος μοναδικών τιμών ανά χακτηριστικο

Ετος                      6
Μήνας                    12
Ημ/νία Λογ/σμου        1372
Αρ. Πολλαπλου            25
Ονομα Πολλαπλού          25
Αρ.Παροχής              868
Τιμολόγιο                 8
Ον.Πελάτη               754
Οδός                    169
Αριθ.                    65
Αρ.Μετρητή              1008
Ημέρες Κατανάλωσης       223
Συντ.Kwh                   4
Κατανάλωση KWh          6748
Αξία Ενέργειας         21432
Συν Προμηθ Ρευματος    23390
Σύνολο Λογαριασμού      3022
Σύν.Τρεχ.Μήνα           2938
Τύπος Λογ/σμου             4
dtype: int64
```

Εικόνα 3.5 Πλήθος μοναδικών τιμών ανά χαρακτηριστικό

Αυτό μπορεί να συμβαίνει διότι σε ένα κτίριο/υποδομή μπορεί να υφίσταται μια παροχή ρεύματος αλλά να έχουν εγκατασταθεί δύο μετρητές π.χ. για μέτρηση της κατανάλωσης την ημέρα και μέτρηση την νύχτα, ή στο ίδιο κτίριο/υποδομή να στεγάζονται δύο διαφορετικές οντότητες π.χ. συγκρότημα σχολείων. Μπορεί επίσης στην ίδια υποδομή να υφίστανται περισσότερες από μια παροχές ρεύματος που εξυπηρετούν ξεχωριστά σημεία κατανάλωσης

με διαφορετικούς μετρητές. Από τις πληροφορίες της Εικόνας 3.5 και συγκεκριμένα από τα χαρακτηριστικά «Αρ. Πολλαπλού» και «Όνομα Πολλαπλού» συμπεραίνουμε ότι ο συγκεκριμένος Δήμος έχει 25 διαφορετικές διοικητικές ενότητες (π.χ. Δημοτικές Κοινότητες), ενώ τα διαφορετικά τιμολόγια των συνδέσεων των παροχών στα κτίρια/υποδομές εμπίπτουν σε 8 κατηγορίες με βάση το χαρακτηριστικό «Τιμολόγιο».

Όπως ήδη αναφέρθηκε υπάρχουν ελλιπείς τιμές στις επιμέρους στήλες των εγγραφών του αρχείου, οπότε θα προβούμε σε έλεγχο και εύρεση αυτών για το σύνολο των χαρακτηριστικών (Εικόνα 3.6 & Εικόνα 3.7).

```
#Έλεγχος ελλιπών τιμών
print("Ελλιπείς τιμές ανα χαρακτηριστικό", end='\n\n')
print(df.isnull().sum(), end='\n\n')
```

Εικόνα 3.6 Κώδικας εμφάνισης ελλιπών τιμών

```
Ελλιπείς τιμές ανα χαρακτηριστικό

Ετος                        0
Μήνας                       0
Ημ/νία Λογ/σμου             0
Αρ. Πολλαπλου               0
Ονομα Πολλαπλού             0
Αρ.Παροχής                  0
Τιμολόγιο                   0
Ον.Πελάτη                   0
Οδός                      159
Αριθ.                   37738
Αρ.Μετρητή                  0
Ημέρες Κατανάλωσης          0
Συντ.Kwh                    0
Κατανάλωση KWh              0
Αξία Ενέργειας              0
Συν Προμηθ Ρευματος         0
Σύνολο Λογαριασμού          0
Σύν.Τρεχ.Μήνα               0
Τύπος Λογ/σμου           1232
dtype: int64
```

Εικόνα 3.7 Ελλιπείς τιμές ανά χαρακτηριστικό

Με δεδομένο ότι έχουμε 754 μοναδικές τιμές στο χαρακτηριστικό «Ον.Πελάτη», τα ιστορικά δεδομένα της ενεργειακής κατανάλωσης αφορούν σε ισάριθμες δημοτικές υποδομές. Οι συνολικές παροχές ρεύματος («Αρ.Παροχής») σύμφωνα με τις πληροφορίες που έχουμε λάβει κατά το στάδιο εύρεσης των μοναδικών τιμών όλων των χαρακτηριστικών (Εικόνα 3.5) ανέρχονται σε 868. Αυτό σημαίνει ότι εφόσον τα κτίρια/υποδομές είναι 754 σε κάποια από αυτά, όπως έχει ήδη λεχθεί, υφίστανται περισσότερες από μια παροχές. Ελέγχοντας την παραπάνω προσέγγιση πράγματι

διαπιστώνουμε ότι 56 δημοτικές υποδομές, όπως αυτές μοναδικά διαχωρίζονται στους λογαριασμούς κατανάλωσης του παρόχου με βάση το χαρακτηριστικό «Ον.Πελάτη», έχουν περισσότερες από μια παροχή ρεύματος (Εικόνα 3.8 & Πίνακας 3.1).

Στο σημείο αυτό πρέπει να διευκρινιστεί ότι στις περιπτώσεις αυτές μιλάμε για την ίδια υποδομή με βάση το χαρακτηριστικό «Ον.Πελάτη» π.χ. ΦΟΠ Δ.Δ. Σ/ΟΥ αλλά με περισσότερες από μια συνδέσεις παροχών ρεύματος προφανώς λόγω των τεχνικών απαιτήσεων αυτής της υποδομής για την ενεργειακή εξυπηρέτηση διαφορετικών σημείων.

```
# Ομαδοποίηση ανά πελάτη και μέτρηση μοναδικών παροχών
multiple_supply = (
    df.groupby("Ον.Πελάτη")["Αρ.Παροχής"]
      .nunique()       # διαφορετικές παροχές ανά πελάτη
      .reset_index()
)

# Πελάτες που έχουν πάνω από 1 παροχή - ταξινόμηση κατά πλήθος παροχών
multiple_supply = multiple_supply[multiple_supply["Αρ.Παροχής"] > 1]
multiple_supply = multiple_supply.sort_values(by="Αρ.Παροχής", ascending=False)
```

Εικόνα 3.8 Κώδικας ομαδοποίησης πελατών με περισσότερες από μια παροχές

Α/Α	Ον.Πελάτη	Πλήθος Αρ.Παροχής
1	ΔΗΜΟΣ Α/ΑΣ	21
2	ΦΟΠ ΔΗΜΟΥ Μ/Σ	12
3	ΚΟΙΝ. Κ/ΗΣ	5
4	ΕΡΓΑΤΙΚΕΣ ΚΑΤΟΙΚΙΕΣ Π/ΟΣ	5
5	ΦΟΠ Δ.Δ.Σ/ΟΥ	5
6	ΦΟΠ Σ/ΟΥ	5
7	ΠΛΑΤΕΙΑ Π/Υ	4
8	Δ.Δ. Σ/ΟΥ	4
9	ΚΟΙΝ. ΚΑ/ΩΝ	3
10	ΚΟΙΝ. Κ/ΟΥ	3
11	ΚΟΙΝ. Μ/Σ	3
12	ΠΛΑΤΕΙΑ Π/ΟΣ	3
13	ΔΗΜΟΣ Μ/Σ	3
14	ΚΟΙΝΟΤΙΚΟ ΚΑΤΑΣΤΗΜΑ ΝΕ/Υ	3
15	ΦΟΠ ΝΗΣ/Υ	3
16	ΦΟΠ Μ/Η	3
17	ΠΛΑΤΕΙΑ ΚΑ/ΡΙΟΥ	3
18	ΠΛΑΤΕΙΑ ΝΗ/ΟΥ	3
19	ΠΕΖΟΔΡΟΜΙΟ ΒΡ/Ι	3
20	ΚΟΙΝ. ΛΟ/ΟΥ	3
21	2ος ΒΡΕΦΟΝΗΠΙΑΚΟΣ ΣΤΑΘΜΟΣ Α/ΑΣ	3

22	ΑΝΤΛΙΟΣΤΑΣΙΟ ΑΡΔΕΥΣΗΣ Μ/Η	2
23	1ο ΝΗΠΙΑΓΩΓΕΙΟ Α/ΑΣ	2
24	ΔΗΜΑΡΧΕΙΟ Μ/Σ	2
25	ΚΟΙΝΟΤΙΚΟ ΚΑΤΑΣΤΗΜΑ Ν. ΠΡ/Υ	2
26	ΚΟΙΝΟΤΙΚΟ ΚΑΤΑΣΤΗΜΑ ΚΥ/Σ	2
27	ΔΗΜΟΤΙΚΟ ΘΕΑΤΡΟ Π/ΟΣ	2
28	ΔΗΜΟΤΙΚΗ ΑΓΟΡΑ Α/ΑΣ	2
29	ΔΗΜΟΤΙΚΟ ΣΧΟΛΕΙΟ Π/ΟΣ	2
30	ΔΗΜΟΤΙΚΟ ΣΧΟΛΕΙΟ Τ/ΩΝ	2
31	ΚΟΙΝ. ΒΡ/ΙΟΥ	2
32	ΚΟΙΝ. ΚΑ/ΡΙΟΥ	2
33	5ο ΝΗΠΙΑΓΩΓΕΙΟ Α/ΑΣ	2
34	3ο ΝΗΠΙΑΓΩΓΕΙΟ Α/ΑΣ	2
35	ΠΛΑΤΕΙΑ ΕΠ/ΗΣ	2
36	ΠΛΑΤΕΙΑ Κ/ΟΥ	2
37	ΠΛΑΤΕΙΑ ΛΙ/Ι	2
38	ΠΛΑΤΕΙΑ ΕΝΑΝΤΙ Κ.Υ Α/ΑΣ	2
39	ΠΕΖΟΔΡΟΜΙΟ ΤΡ/Α	2
40	ΠΕΖΟΔΡΟΜΙΟ ΒΕ/ΛΟΥ Δ.	2
41	ΦΟΠ Α/ΑΣ Υ/Σ Γ2-6	2
42	ΠΛΑΤΕΙΑ ΤΡ/Α	2
43	ΠΛΑΤΕΙΑ ΟΣΕ ΑΛ/Α	2
44	ΠΛΑΤΕΙΑ ΠΑ/ΡΙ	2
45	ΠΛΑΤΕΙΑ ΝΗ/Ι	2
46	ΦΟΠ Κ/ΗΣ	2
47	ΦΟΠ Α/ΑΣ Υ/Σ Γ5-4	2
48	ΦΟΠ Δ.Δ.Π.ΠΡ/Υ	2
49	ΦΟΠ ΚΟΙΝΟΤ Π/ΟΥ	2
50	ΦΟΠ ΚΟΙΝ.Μ/Σ	2
51	ΦΟΠ ΝΕ/Ο	2
52	ΦΟΠ Μ/Σ	2
53	ΦΟΠ ΝΕ/Υ	2
54	ΦΟΠ Π/ΟΣ	2
55	ΦΟΠ ΠΡ/Υ	2
56	ΦΩΤΙΣΜΟΣ Π/Υ	2
		170

Πίνακας 3.1 Πελάτες με περισσότερες από μια παροχές

Οι επιπλέον παροχές που προκύπτουν για τους 56 πελάτες (εξαιρουμένων των 56 ισάριθμων παροχών που αντιστοιχούν σε αυτούς), ανέρχονται σε 170-56 = 114 παροχές. Εάν από το σύνολο των 868 παροχών αφαιρεθούν οι 114 επιπλέον παροχές (868-114=754), προκύπτει ο αριθμός των 754 μοναδικών πελατών. Το ίδιο ακριβώς συμβαίνει και με το πλήθος των 1008 μετρητών («Αρ.Μετρητή») όπου εντοπίζονται περισσότεροι του ενός μετρητή σε 171

δημοτικές υποδομές με βάση το χαρακτηριστικό «Ον.Πελάτη» ενώ 4 δημοτικές υποδομές έχουν 2 κοινούς μετρητές (Εικόνα 3.9 & Πίνακας 3.2).

```
# Ομαδοποίηση ανά πελάτη και μέτρηση μοναδικών μετρητών
multiple_counter = (
    df.groupby("Ον.Πελάτη")["Αρ.Μετρητή"]
      .nunique()       # διαφορετικοί μετρητές ανά πελάτη
      .reset_index()
)

# Πελάτες που έχουν πάνω από 1 μετρητή - ταξινόμηση κατά πλήθος μετρητων
multiple_counter = multiple_counter[multiple_counter["Αρ.Μετρητή"] > 1]
multiple_counter = multiple_counter.sort_values(by="Αρ.Μετρητή", ascending=False)
```

Εικόνα 3.9 Κώδικας ομαδοποίησης πελατών με περισσότερους από έναν μετρητή

Α/Α	Ον.Πελάτη	Πλήθος Αρ.Μετρητή
1	ΔΗΜΟΣ Α/ΑΣ	**15**
2	ΦΟΠ ΔΗΜΟΥ Μ/Σ	13
3	ΕΡΓΑΤΙΚΕΣ ΚΑΤΟΙΚΙΕΣ Π/ΟΣ	9
4	ΦΟΠ Δ.Δ.Σ/ΟΥ	6
5	ΚΟΙΝ. Κ/ΟΥ	5
6	ΦΟΠ Σ/ΟΥ	5
7	ΚΟΙΝ. Κ/ΗΣ	5
8	ΔΗΜΟΤΙΚΟ ΣΧΟΛΕΙΟ Π/ΟΣ	4
9	ΑΝΤΛΙΟΣΤΑΣΙΟ ΑΡΔΕΥΣΗΣ Μ/Η	4
10	Δ.Δ. Σ/ΟΥ	4
11	ΠΛΑΤΕΙΑ ΚΑ/Υ	4
12	ΠΛΑΤΕΙΑ Π/Υ	4
13	ΠΕΖΟΔΡΟΜΙΟ ΕΥ/ΛΑ Ν.	4
14	ΔΗΜΟΣ Μ/Σ	4
15	ΚΟΙΝΟΤΙΚΟ ΚΑΤΑΣΤΗΜΑ ΝΕ/Υ	4
16	ΦΟΠ Μ/Η	4
17	ΠΛΑΤΕΙΑ ΝΗ/ΟΥ	4
18	ΠΕΖΟΔΡΟΜΙΟ ΒΡ/Ι	3
19	ΠΛΑΤΕΙΑ ΟΣΕ Α/Α	3
20	ΦΟΠ Δ.Δ.ΝΗΣ/Υ	3
21	ΦΟΠ Δ.Δ.Π.ΠΡ/Υ	3
22	ΦΟΠ Π.ΣΚ/Ι Υ/Σ 1	3
23	ΦΟΠ ΝΗΣ/Υ	3
24	ΦΟΠ ΚΟΙΝΟΤΗΤΟΣ ΠΡ/ΟΣ	3
25	ΦΟΠ ΝΕ/Ο	3
26	ΠΛΑΤΕΙΑ Π/ΟΣ	3
27	ΚΟΙΝ. Μ/Σ	3
28	ΚΟΙΝ. ΚΑ/ΩΝ	3
29	ΝΗΠΙΑΓΩΓΕΙΟ ΚΕ/Ι	3

30	ΔΗΜΟΤΙΚΟ ΣΧΟΛΕΙΟ Τ/ΩΝ	3
31	ΠΑΙΔΙΚΟΣ ΣΤΑΘΜΟΣ Μ/Σ	3
32	ΔΗΜΑΡΧΕΙΟ Α/ΑΣ	3
33	ΓΕΩΤΡΗΣΗ ΤΡ/ΑΣ	3
34	ΚΟΙΝ. ΛΟ/ΟΥ	3
35	2ος ΒΡΕΦΟΝΗΠΙΑΚΟΣ ΣΤΑΘΜΟΣ Α/ΑΣ	3
36	ΓΥΜΝΑΣΤΗΡΙΟ ΚΛΕΙΣΤΟ Μ/Σ	3
37	ΓΕΩΤΡΗΣΗ ΕΠ/ΗΣ	3
38	ΠΕΖΟΔΡΟΜΙΟ ΤΡ/Α	3
39	1ος ΒΡΕΦΟΝΗΠΙΑΚΟΣ ΣΤΑΘΜΟΣ Α/ΑΣ	2
40	5ο ΝΗΠΙΑΓΩΓΕΙΟ Α/ΑΣ	2
41	1ο ΝΗΠΙΑΓΩΓΕΙΟ Α/ΑΣ	2
42	3ο ΝΗΠΙΑΓΩΓΕΙΟ Α/ΑΣ	2
43	2ο ΝΗΠΙΑΓΩΓΕΙΟ Π/ΟΣ	2
44	ΓΕΩΤΡΗΣΗ ΚΑ/Α	2
45	ΓΕΩΤΡΗΣΗ ΒΡ/ΙΟΥ	2
46	ΓΕΩΤΡΗΣΗ ΑΡ/Υ No1	2
47	Δ.Δ. Π.ΠΡ/Υ	2
48	ΓΥΜΝΑΣΤΗΡΙΟ ΠΡ/Α	2
49	ΓΥΜΝΑΣΤΗΡΙΟ ΤΡ/Α	2
50	ΓΥΜΝΑΣΤΗΡΙΟ Κ/ΡΙ	2
51	ΓΥΜΝΑΣΤΗΡΙΟ Π/Υ	2
52	ΓΥΜΝΑΣΙΟ Κ/ΗΣ	2
53	ΓΥΜΝΑΣΙΟ Μ/Σ (+ΛΥΚΕΙΟ)	2
54	ΓΥΜΝΑΣΙΟ Τ/ΩΝ	2
55	ΓΥΜΝΑΣΙΟ Π/ΟΣ	2
56	ΓΥΜΝΑΣΙΟ ΚΑ/ΑΣ	2
57	ΓΗΠΕΔΟ ΚΑ/Υ	2
58	ΓΗΠΕΔΟ Π.Σ/ΟΥ	2
59	ΓΗΠΕΔΟ ΣΧ/Α	2
60	ΑΠΟΔΥΤΗΡΙΑ ΓΥΜΝΑΣΤΗΡΙΟΥ Π/ΟΣ	2
61	ΑΠΟΔΥΤΗΡΙΑ ΓΗΠΕΔΟΥ ΛΙ/ΙΟΥ	2
62	ΑΝΤΛΙΟΣΤΑΣΙΟ ΑΡΔΕΥΣΗΣ ΠΡ/Υ No2	2
63	ΓΕΩΤΡΣΗΣΗ ΑΡΔΕΥΣΗΣ ΝΕ/Ο (ΤΣ/ΡΙ)	2
64	ΓΕΩΤΡΗΣΗ ΑΡΔΕΥΣΗΣ ΝΗΣ/Υ (ΓΕ/ΚΟ)	2
65	ΓΕΩΤΡΗΣΗ ΑΡΔΕΥΣΗΣ ΝΕ/Υ (ΓΕ/ΚΟ)	2
66	ΓΕΩΤΡΗΣΗ ΑΡΔΕΥΣΗΣ ΑΓ/ΑΣ (ΣΥ/ΟΣ)	2
67	ΓΕΝΙΚΟ ΛΥΚΕΙΟ Π/ΟΣ Κ/ΗΣ	2
68	ΑΝΤΛΙΟΣΤΑΣΙΟ ΑΡΔΕΥΣΗΣ Μ/Σ (ΤΣ/ΕΣ ΠΑΛΙΟ)	2
69	ΑΝΤΛΙΟΣΤΑΣΙΟ ΑΡΔΕΥΣΗΣ Μ/Σ (ΝΟ/ΚΟ)	2
70	ΑΝΤΛΙΟΣΤΑΣΙΟ ΑΡΔΕΥΣΗΣ Μ/Σ (ΜΠ/ΕΣ)	2
71	ΑΝΤΛΙΟΣΤΑΣΙΟ ΑΡΔΕΥΣΗΣ ΝΕ/Ο (ΠΕ/ΡΑ)	2
72	ΑΝΤΛΙΟΣΤΑΣΙΟ ΑΡΔΕΥΣΗΣ Μ/Η (ΠΕ/ΡΑ)	2
73	7ο ΝΗΠΙΑΓΩΓΕΙΟ Α/ΑΣ	2

74	ΑΜΦΙΘΕΑΤΡΟ ΔΗΜΟΥ Α/ΑΣ	2
75	ΑΝΤΛΙΟΣΤΑΣΙΟ ΑΡΔΕΥΣΗΣ Α.Τ/ΑΣ	2
76	1ο ΛΥΚΕΙΟ Α/ΑΣ	2
77	ΔΗΜΟΤΙΚΟ ΣΧΟΛΕΙΟ ΚΛ/Ι	2
78	ΔΗΜΟΤΙΚΟ ΣΧΟΛΕΙΟ Κ/ΡΙ	2
79	ΔΗΜΟΤΙΚΟ ΣΧΟΛΕΙΟ ΚΑ/Α	2
80	ΔΗΜΟΤΙΚΟ ΘΕΑΤΡΟ Π/ΟΣ	2
81	ΔΗΜΟΤΙΚΟ ΑΘΛΗΤΙΚΟ ΚΕΝΤΡΟ Α/ΑΣ	2
82	ΔΗΜΟΤΙΚΗ ΑΓΟΡΑ Α/ΑΣ	2
83	ΔΗΜΟΣ Α/ΑΣ Δ/Δ Λ	2
84	ΓΕΩΤΡΗΣΗ ΣΧ/Α	2
85	ΔΗΜΟΣ Α/ΑΣ Δ/Δ Κ	2
86	ΔΗΜΟΣ Α/ΑΣ (ΠΕΡΙΒ ΑΛΛΩ	2
87	ΔΗΜΑΡΧΕΙΟ Μ/Σ	2
88	ΓΕΩΤΡΗΣΗ Π/ΟΥ No1	2
89	ΚΕΠ Π/ΟΣ	2
90	ΔΗΜΟΤΙΚΟ ΣΧΟΛΕΙΟ ΝΗ/Ι	2
91	ΚΑΠΗ Α/ΑΣ	2
92	ΔΗΜΟΤΙΚΟ ΣΧΟΛΕΙΟ ΛΟ/ΟΣ	2
93	ΓΕΩΤΡΗΣΗ ΠΡ/Υ	2
94	ΓΕΩΤΡΗΣΗ Π/Υ No1	2
95	ΓΕΩΤΡΗΣΗ Π/Υ No3	2
96	ΓΕΩΤΡΗΣΗ ΚΑ/Υ	2
97	ΓΕΩΤΡΗΣΗ Κ/ΟΥ	2
98	ΓΕΩΤΡΗΣΗ ΚΥ/Σ No1	2
99	ΓΕΩΤΡΗΣΗ ΚΥ/Σ No2	2
100	ΓΕΩΤΡΗΣΗ ΝΗ/ΟΥ No2	2
101	ΓΕΩΤΡΗΣΗ ΞΕ/Η No1	2
102	ΓΕΩΤΡΗΣΗ Π. ΠΡ/Υ No1	2
103	ΓΕΩΤΡΗΣΗ Π.Σ/ΟΥ	2
104	ΓΕΩΤΡΗΣΗ ΚΥ/Σ No3	2
105	ΓΕΩΤΡΗΣΗ ΛΟ/ΟΥ	2
106	ΓΕΩΤΡΗΣΗ Μ/Σ No1	2
107	ΓΕΩΤΡΗΣΗ ΝΕΟΥ ΠΡ/Υ No1	2
108	ΠΑΙΔΙΚΟΣ ΣΤΑΘΜΟΣ Σ/ΟΥ	2
109	ΠΕΖΟΔΡΟΜΙΟ ΒΕ/ΛΟΥ Δ.	2
110	ΝΗΠΙΑΓΩΓΕΙΟ ΠΑ/ΡΑΣ	2
111	ΝΗΠΙΑΓΩΓΕΙΟ Τ/ΩΝ	2
112	ΚΟΙΝΟΤΙΚΟ ΚΑΤΑΣΤΗΜΑ ΚΛ/ΟΥ	2
113	ΚΟΙΝΟΤΙΚΟ ΚΑΤΑΣΤΗΜΑ ΚΥ/Σ	2
114	ΚΟΙΝΟΤΙΚΟ ΚΑΤΑΣΤΗΜΑ Ν. ΠΡ/Υ	2
115	ΚΟΙΝΟΤΙΚΟ ΚΑΤΑΣΤΗΜΑ Σ/ΟΥ 3	2
116	ΚΟΙΝ. Π/ΟΣ	2
117	ΚΟΙΝΟΤΙΚΟ ΚΑΤΑΣΤΗΜΑ ΑΓ/ΑΣ 2	2

118	ΚΟΙΝ. ΝΗ/ΟΥ	2
119	ΔΗΜΟΤΙΚΟ ΣΧΟΛΕΙΟ Σ/ΑΣ	2
120	ΔΗΜΟΤΙΚΟ ΣΧΟΛΕΙΟ ΝΕ/Ο	2
121	ΚΟΙΝ. ΚΑ/Υ	2
122	ΚΟΙΝ. ΒΡ/ΙΟΥ	2
123	ΚΟΙΝ. ΕΠ/ΗΣ	2
124	ΠΛΑΤΕΙΑ ΠΑ/ΡΙ	2
125	ΠΛΑΤΕΙΑ ΛΙ/Ι	2
126	ΠΛΑΤΕΙΑ ΜΑΚΕΔΟΝΙΑΣ Δ - ΜΠ/ΣΗ	2
127	ΠΛΑΤΕΙΑ ΝΗ/Ι	2
128	ΠΛΑΤΕΙΑ ΤΡ/Α	2
129	ΣΤΑΔΙΟ ΠΑ/ΡΙ	2
130	ΣΥΝΟΙΚΙΑ ΘΡΑΚΗΣ Μ/Σ	2
131	ΤΕΧΝΙΚΗ ΥΠΗΡΕΣΙΑ Δ.Ε. Α/ΩΝ	2
132	ΦΟΠ Α/ΑΣ Υ/Σ Γ2-6	2
133	ΦΟΠ Α/ΑΣ Υ/Σ Β4-7 ΑΝΤΥΠΑ Μ. 38	2
134	ΠΛΑΤΕΙΑ ΕΠ/ΗΣ	2
135	ΠΛΑΤΕΙΑ ΕΝΑΝΤΙ Κ.Υ Α/ΑΣ	2
136	ΠΛΑΤΕΙΑ Κ/ΟΥ	2
137	ΠΛΑΤΕΙΑ Κ/ΡΙ	2
138	ΠΕΖΟΔΡΟΜΟΣ Π/Υ	2
139	ΠΕΖΟΔΡΟΜΙΟ ΠΡ/ΑΣ	2
140	ΦΟΠ ΚΑ/ΑΣ Υ/Σ 4	2
141	ΦΟΠ Α/ΑΣ Υ/Σ Γ5-4	2
142	ΦΟΠ Α/ΑΣ Υ/ΣΑ1-10	2
143	ΦΟΠ ΒΡ/Ι Υ/Σ 11	2
144	ΦΟΠ ΑΓ.ΤΡ/Α	2
145	ΦΟΠ Α/ΑΣ Υ/Σ Β3-2 28ΗΣ ΟΚΤΩΒΡΙΟΥ 101	2
146	ΦΟΠ ΚΛ/ΟΥ Υ/Σ 6	2
147	ΦΟΠ Κ/ΡΙ Υ/Σ 149	2
148	ΦΟΠ Κ/ΗΣ Υ/Σ 4	2
149	ΦΟΠ Κ/ΗΣ Υ/Σ 3	2
150	ΦΟΠ Κ/ΗΣ	2
151	ΦΟΠ ΚΟΙΝΟΤ Π/ΟΥ	2
152	ΦΟΠ ΚΟΙΝ.Μ/Σ	2
153	ΦΟΠ ΚΛ/Ι Υ/Σ22	2
154	ΦΟΠ ΚΛ/ΟΥ Υ/Σ 16	2
155	ΦΟΠ ΚΛ/ΟΥ Υ/Σ 17	2
156	ΦΟΠ Ν.ΠΡ/Υ ΠΡΟΣ ΑΓ.Τ	2
157	ΦΟΠ Ν. ΠΡ/ΟΣ	2
158	ΦΟΠ Μ/Σ	2
159	ΦΟΠ ΝΗΣ/Υ Υ/Σ 89	2
160	ΦΟΠ ΞΕ/Η Υ/Σ 2	2
161	ΦΟΠ Π.ΣΚ/Ι Υ/Σ 3	2

162	ΦΟΠ ΠΑ/ΡΙ Υ/Σ 15	2
163	ΦΟΠ ΝΕ/Υ	2
164	ΦΟΠ ΠΑ/ΑΣ Υ/Σ 5	2
165	ΦΟΠ Π/ΟΣ	2
166	ΦΟΠ ΠΡ/Υ	2
167	ΦΟΠ Π/ΟΣ Υ/Σ 28	2
168	ΦΟΠ ΡΑ/Η Υ/Σ 3	2
169	ΦΟΠ Σ/ΟΥ Υ/Σ43	2
170	ΦΟΠ Τ/ΩΝ Υ/Σ 1	2
171	ΦΩΤΙΣΜΟΣ Π/Υ	2
172	ΑΝΤΛΙΟΣΤΑΣΙΟ ΑΡΔΕΥΣΗΣ Μ/Η (Γ/Ι 1)	
173	ΔΗΜΟΣ Μ/Σ (ΑΡΔΕΥΣ Η)	1
174	ΔΗΜΟΣ Μ/Σ (ΑΡΔΕΥΤ ΙΚΟ)	
175	ΑΝΤΛΙΟΣΤΑΣΙΟ ΑΡΔΕΥΣΗΣ Μ/Η (Γ/Ι 2)	1
		429

Πίνακας 3.2 Πελάτες με περισσότερους από έναν μετρητές ή με κοινούς μετρητές

Συνεχίζουμε την έρευνα μας με τον εντοπισμό διπλότυπων εγγραφών εκτελώντας τον κώδικα της Εικόνας 3.10.

```
#Εντοπισμός Διπλοεγγραφών
print("Εντοπισμός Διπλοεγγραφών")
duplicates = df[df.duplicated()]
print(duplicates, end='\n\n')
```

Εικόνα 3.10 Κώδικας εντοπισμού διπλοεγγραφών

Εντοπίζονται τρείς διπλότυπες εγγραφές όπως αυτές φαίνονται στην Εικόνα 3.11, που αντιστοιχούν στις γραμμές 11712, 37899 και 38892 και οι οποίες θα πρέπει να διαγραφούν (Εικόνα 3.12).

```
Εντοπισμός Διπλοεγγραφών
       Ετος  Μήνας  ... Σύν.Τρεχ.Μήνα Τύπος Λογ/σμου
11712  2024     2   ...           0.0          NaN
37899  2025    10   ...           5.0       ΕΚΚΑΘ
38892  2025     6   ...           5.0       ΕΚΚΑΘ

[3 rows x 19 columns]
```

Εικόνα 3.11 Διπλότυπες εγγραφές

```
#Καθαρισμός διπλοεγγραφών
print("Αρχικό μέγεθος:", df.shape,end='\n\n')
df = df.drop_duplicates()
print("Μέγεθος μετά τον καθαρισμό:", df.shape, end='\n\n')

#Εξαγωγή σε νέο CSV
df.to_csv("clean_energ.csv", index=False, encoding="utf-8-sig")

print("Εκκαθαρισμένο αρχείο: clean_energ.csv")
```

Εικόνα 3.12 Κώδικας καθαρισμού διπλότυπων εγγραφών

Συγκρίνοντας το αρχικό μέγεθος του αρχείου (energy_2020-2025.csv) και το μέγεθος του νέου αρχείου (clean_energ.csv) μετά τον καθαρισμό (Εικόνα 3.13), παρατηρούμε ότι το πλήθος των εγγραφών από 41.571 έχει διαμορφωθεί στις 41.568 εγγραφές ενώ το πλήθος των στηλών έχει παραμείνει το ίδιο. Το αποτέλεσμα αυτό είναι σωστό διότι από τις 41.571 αρχικές εγγραφές έχουμε διαγράψει τις 3 διπλότυπες εγγραφές (41.571-3=41.568).

```
Αρχικό μέγεθος: (41571, 19)

Μέγεθος μετά τον καθαρισμό: (41568, 19)

Εκκαθαρισμένο αρχείο: clean_energ.csv
```

Εικόνα 3.13 Μέγεθος νέου αρχείου μετά τον καθαρισμό

Με την ανάλυση που επιχειρήθηκε μέχρι στιγμής στα δεδομένα του αρχείου έγινε προσπάθεια να προσδιοριστεί το πλήθος των διαφορετικών δημοτικών υποδομών λαμβάνοντας υπόψη τον αριθμό των μοναδικών τιμών στα χαρακτηριστικά «Αρ.Παροχής», «Ον.Πελάτη» και «Αρ.Μετρητή».

Από τον έλεγχο που προηγήθηκε διαπιστώνουμε ότι **η μοναδικότητα μιας δημοτικής υποδομής** εξαρτάται αποκλειστικά από το χαρακτηριστικό «Ον.Πελάτη». Η προσέγγιση αυτή είναι σημαντική για την περαιτέρω συνέχεια της ερευνητικής μας μεθοδολογίας.

3.4.4 Κατηγοριοποίηση ενεργειακών δεδομένων

Όπως έχει ήδη αναφερθεί, στόχος της παρούσας εργασίας είναι η πρόβλεψη της μηνιαίας ενεργειακής κατανάλωσης των δημοτικών υποδομών. Επομένως η εξαρτημένη μεταβλητή-στόχος της έρευνάς μας είναι η πρόβλεψη της τιμής του χαρακτηριστικού **«Κατανάλωση KWh» (**Εικόνα 3.2), αξιοποιώντας και αναλύοντας τα ιστορικά δεδομένα του αρχείου που μας διατέθηκε. Παράμετροι όπως η εποχικότητα με βάση το μήνα κατανάλωσης σε συνδυασμό με την κατηγορία της δημοτικής εγκατάστασης, συντελούν καθοριστικά στην ενεργειακή συμπεριφορά των δημοτικών υποδομών.

Στο εκκαθαρισμένο αρχείο **clean_energ.csv** το κατηγορικό χαρακτηριστικό **«Ον.Πελάτη»** εμφανίζεται με **754 μοναδικές τιμές** (Εικόνα 3.5). Εάν αντιμετωπίσουμε το πλήθος των 754 διαφορετικών δημοτικών υποδομών ως ισάριθμες διαφορετικές κατηγορίες-κλάσεις, τότε το χαρακτηριστικό «Ον.Πελάτη» αποκτά υπερβολική διάσταση (high cardinality). Η

υπερβολική διάσταση δημιουργεί προβλήματα στην ανάλυση και οπτικοποίηση των δεδομένων, ενώ κατά την εφαρμογή μοντέλων μηχανικής μάθησης αυξάνεται ο κίνδυνος υπερπροσαρμογής (overfitting) δυσχεραίνοντας την ερμηνευσιμότητα των αποτελεσμάτων (Goodfellow et al., 2016). Η προτεινόμενη λύση στο πρόβλημα αυτό είναι η μείωση των πολλών διαφορετικών κατηγοριών μέσω της **εννοιολογικής κατηγοριοποίησης** των 754 δημοτικών υποδομών, αξιοποιώντας την πληροφορία που περιέχεται στο πεδίο «Ον.Πελάτη» (Witten et al., 2016).

Συγκεκριμένα η προσέγγιση αυτή στοχεύει στον εντοπισμό της συχνότητας εμφάνισης συγκεκριμένων λέξεων-κλειδιών μέσα στο πεδίο αυτό ώστε κάθε δημοτική υποδομή να ενταχθεί σε μια **σαφώς καθορισμένη θεματική κατηγορία (κλάση)** (π.χ. *Αθλητισμός*, *Πολιτισμός*, *Δημοτικός Φωτισμός, κτλ.*).

```
# ανάγνωση εκκαθαρισμένου αρχείου CSV
df = pd.read_csv("clean_energ.csv", encoding="utf-8-sig")
# Δημιουργία ενιαίας συμβολοσειράς με τις τιμές της στήλης 'Ον.Πελάτη'
names = df["Ον.Πελάτη"].astype(str)
# Μετατροπή σε κεφαλαία για σωστή μέτρηση με το ίδιο όνομα
names = names.str.upper()
# Αφαίρεση σημείων στίξης και ειδικών χαρακτήρων
names = names.apply(lambda x: re.sub(r"[^A-ΩA-Z0-9 ]", " ", x))
# Διαχωρισμός όλων των λέξεων
all_words = " ".join(names).split()
# θέλουμε μόνο λέξεις όχι αριθμούς
clean_words = [w for w in all_words if not w.isdigit()]
# Μέτρηση συχνότητας εμφάνισης λέξεων-κλειδιών
word_freq = Counter(clean_words)
# Μετατροπή σε DataFrame για εύκολη προβολή
word_freq_df = pd.DataFrame(word_freq.items(), columns=["Λέξη", "Συχνότητα"])
# Φιλτράρουμε λέξεις με μικρή σημασία ή πολύ σπάνιες
word_freq_df = word_freq_df[word_freq_df["Συχνότητα"] > 5]
# Ταξινόμηση κατά συχνότητα
word_freq_df = word_freq_df.sort_values(by="Συχνότητα", ascending=False)
```

Εικόνα 3.14 Κώδικας εμφάνισης συχνότητας λέξεων

Από την ανάλυση της συχνότητας εμφάνισης των λέξεων στο πεδίο «Ον.Πελάτη», επιλέχθηκαν τμήματα ή ολόκληρες λέξεις–κλειδιά για τις οποίες υπάρχει ασφαλής γνώση σχετικά με την κατηγορία στην οποία πρέπει να ενταχθούν.

```python
#====== Αντιστοίχιση λέξεων-κλειδιά σε κατηγορίες ==========
mapping = {
    "ΦΩΠ": "ΦΩΤΙΣΜΟΣ ΟΔΩΝ ΠΛΑΤΕΙΩΝ",
    "ΑΡΔΕΥΣΗΣ": "ΑΝΤΛΙΟΣΤΑΣΙΑ ΑΡΔΕΥΣΗΣ",
    "ΔΕΥΑ": "ΑΝΤΛΙΟΣΤΑΣΙΑ ΥΔΡΕΥΣΗΣ",
    "ΠΛΑΤΕΙΑ": "ΦΩΤΙΣΜΟΣ ΠΛΑΤΕΙΩΝ",
    "ΣΧΟΛΕΙΟ": "ΔΗΜΟΤΙΚΑ ΣΧΟΛΕΙΑ",
    "ΠΕΖΟΔΡΟΜΙΟ": "ΦΩΤΙΣΜΟΣ ΠΕΖΟΔΡΟΜΙΩΝ",
    "ΝΗΠΙΑΓΩΓΕΙΟ": "ΝΗΠΙΑΓΩΓΕΙΑ",
    "ΚΑΤΑΣΤΗΜΑ": "ΚΟΙΝΟΤΙΚΑ ΚΑΤΑΣΤΗΜΑΤΑ",
    "ΓΥΜΝΑΣΤΗΡΙΟ": "ΑΘΛΗΤΙΚΕΣ ΕΓΚΑΤΑΣΤΑΣΕΙΣ",
    "ΓΗΠΕΔΟ": "ΑΘΛΗΤΙΚΕΣ ΕΓΚΑΤΑΣΤΑΣΕΙΣ",
    "ΓΥΜΝΑΣΙΟ": "ΣΧΟΛΕΙΑ Β'ΒΑΘΜΙΑΣ ΕΚΠΑΙΔΕΥΣΗΣ",
    "ΑΠΟΔΥΤΗΡΙΑ": "ΑΘΛΗΤΙΚΕΣ ΕΓΚΑΤΑΣΤΑΣΕΙΣ",
    "ΕΡΓΑΤΙΚΕΣ": "ΕΡΓΑΤΙΚΕΣ ΚΑΤΟΙΚΙΕΣ",
    "ΙΑΤΡΕΙΟ": "ΔΗΜΟΤΙΚΑ ΙΑΤΡΕΙΑ",
    "ΛΥΚΕΙΟ": "ΣΧΟΛΕΙΑ Β'ΒΑΘΜΙΑΣ ΕΚΠΑΙΔΕΥΣΗΣ",
    "ΓΕΦΥΡΑ": "ΦΩΤΙΣΜΟΣ ΟΔΩΝ ΠΛΑΤΕΙΩΝ",
    "ΔΗΜΑΡΧΕΙΟ": "ΔΗΜΟΤΙΚΑ ΚΤΙΡΙΑ",
    "ΚΕΠ": "ΚΕΝΤΡΟ ΕΞΥΠΗΡΕΤΗΣΗΣ ΠΟΛΙΤΩΝ",
    "ΝΕΚΡΟΤΑΦ": "ΔΗΜΟΤΙΚΑ ΚΟΙΜΗΤΗΡΙΑ",
    "ΒΡΕΦ": "ΥΠΟΔΟΜΕΣ ΠΡΟΣΧΟΛΙΚΗΣ ΑΓΩΓΗΣ",
    "ΕΕΕΕΚ": "ΣΧΟΛΕΙΑ Β'ΒΑΘΜΙΑΣ ΕΚΠΑΙΔΕΥΣΗΣ",
    "ΠΟΔΗΛΑΤΟΔΡΟ": "ΦΩΤΙΣΜΟΣ ΠΕΖΟΔΡΟΜΙΩΝ",
    "ΠΑΙΔΙΚΗ": "ΠΑΙΔΙΚΕΣ ΧΑΡΕΣ",
    "ΣΗΜΑΤΟΔΟΤ": "ΟΔΟΣΗΜΑΝΣΗ",
    "ΘΕΑΤΡΟ": "ΠΟΛΙΤΙΣΜΟΣ",
    "ΠΑΙΔΙΚΟΣ": "ΥΠΟΔΟΜΕΣ ΠΡΟΣΧΟΛΙΚΗΣ ΑΓΩΓΗΣ",
    "ΑΓΟΡΑ": "ΔΗΜΟΤΙΚΗ ΑΓΟΡΑ",
    "ΚΑΠΗ": "ΚΑΠΗ",
    "ΣΥΝΕΔΡΙΑΚΟ": "ΠΟΛΙΤΙΣΜΟΣ",
    "ΦΙΛΑΡΜΟΝΙΚΗ": "ΠΟΛΙΤΙΣΜΟΣ",
    "ΕΠΑΛ": "ΣΧΟΛΕΙΑ Β'ΒΑΘΜΙΑΣ ΕΚΠΑΙΔΕΥΣΗΣ",
    "ΣΤΑΔΙΟ": "ΑΘΛΗΤΙΚΕΣ ΕΓΚΑΤΑΣΤΑΣΕΙΣ",
    "ΑΘΛΗΤΙΚΟ": "ΑΘΛΗΤΙΚΕΣ ΕΓΚΑΤΑΣΤΑΣΕΙΣ",
    "ΑΜΦΙΘΕΑΤΡΟ": "ΠΟΛΙΤΙΣΜΟΣ",
    "ΚΟΛΥΜΒΗΤΗΡΙΟ": "ΑΘΛΗΤΙΚΕΣ ΕΓΚΑΤΑΣΤΑΣΕΙΣ",
    "ΣΥΝΟΙΚΙΑ": "ΦΩΤΙΣΜΟΣ ΠΕΖΟΔΡΟΜΙΩΝ",
    "ΠΝΕΥΜΑΤΙΚΟ": "ΠΟΛΙΤΙΣΜΟΣ",
    "ΠΟΛΙΤΙΣΤΙΚΟ": "ΠΟΛΙΤΙΣΜΟΣ",
    "ΜΟΡΦ": "ΠΟΛΙΤΙΣΜΟΣ",
    "ΚΔΑΠ": "ΚΔΑΠ",
    "ΠΕΖΟΔΡΟΜΟΣ": "ΦΩΤΙΣΜΟΣ ΠΕΖΟΔΡΟΜΙΩΝ",
    "ΚΟΙΜΗΤΗΡΙΑ": "ΔΗΜΟΤΙΚΑ ΚΟΙΜΗΤΗΡΙΑ",
    "ΦΩΤΙΣΜ": "ΦΩΤΙΣΜΟΣ ΟΔΩΝ ΠΛΑΤΕΙΩΝ",
    "ΒΙΟΛΟΓ": "ΕΓΚΑΤΑΣΤΑΣΗ ΕΠΕΞΕΡΓΑΣΙΑΣ ΛΥΜΑΤΩΝ"

}
#Συνάρτηση κατηγοριοποίησης
def castomer_category(name):
    nane = str(name).upper()

    for keyword, category in mapping.items():
        if keyword in name:
            return category
    return "Λοιπά Κτίρια/Υποδομές"

# Δημιουργία νέας στήλης "Κατηγορία Πελάτη"
df["Κατηγορία Υποδομής"] = df["Ον.Πελάτη"].apply(castomer_category)

# Αποθήκευση σε νέο CSV
df.to_csv("data_categorized.csv", index=False, encoding="utf-8-sig")
print("Νέο CSV με κατηγορίες :","data_categorized.csv", end='\n\n')
```

Εικόνα 3.15 Κώδικας κατηγοριοποίησης

Με τη χρήση ενός λεξικού mapping στην Python, όπως φαίνεται στην Εικόνα 3.15, πραγματοποιείται η απαραίτητη αντιστοίχιση των συχνών λέξεων (ή τμημάτων τους) στις αντίστοιχες θεματικές κατηγορίες για την κατηγοριοποίηση των 754 μοναδικών πελατών-καταναλωτών ρεύματος. Το αποτέλεσμα αυτής της διαδικασίας αποθηκεύεται στο αρχείο **data_categorized.csv** με την προσθήκη νέας στήλης (χαρακτηριστικού) με ετικέτα **«Κατηγορία Υποδομής»**. Στη συνέχεια απεικονίζονται οι πληροφορίες του νέου αρχείου και οι μοναδικές τιμές των χαρακτηριστικών του ώστε να επαληθευτεί η ορθότητα της κατηγοριοποίησης.

```
Πληροφορίες αρχείου κατηγοριοποίησης καταναλωτών

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 41568 entries, 0 to 41567
Data columns (total 20 columns):
 #   Column               Non-Null Count  Dtype
---  ------               --------------  -----
 0   Ετος                 41568 non-null  int64
 1   Μήνας                41568 non-null  int64
 2   Ημ/νία Λογ/σμου      41568 non-null  object
 3   Αρ. Πολλαπλου        41568 non-null  object
 4   Ονομα Πολλαπλού      41568 non-null  object
 5   Αρ.Παροχής           41568 non-null  int64
 6   Τιμολόγιο            41568 non-null  object
 7   Ον.Πελάτη            41568 non-null  object
 8   Οδός                 41450 non-null  object
 9   Αριθ.                13012 non-null  object
 10  Αρ.Μετρητή           41506 non-null  object
 11  Ημέρες Κατανάλωσης   41568 non-null  int64
 12  Συντ.Kwh             41568 non-null  int64
 13  Κατανάλωση KWh       41568 non-null  int64
 14  Αξία Ενέργειας       41568 non-null  float64
 15  Συν Προμηθ Ρευματος  41568 non-null  float64
 16  Σύνολο Λογαριασμού   41568 non-null  float64
 17  Σύν.Τρεχ.Μήνα        41568 non-null  float64
 18  Τύπος Λογ/σμου       40445 non-null  object
 19  Κατηγορία Υποδομής   41568 non-null  object
dtypes: float64(4), int64(6), object(10)
```

```
Πλήθος μοναδικών τιμών ανά χακτηριστικο

Ετος                    6
Μήνας                  12
Ημ/νία Λογ/σμου      1372
Αρ. Πολλαπλου          25
Ονομα Πολλαπλού        25
Αρ.Παροχής            868
Τιμολόγιο               8
Ον.Πελάτη             754
Οδός                  169
Αριθ.                  65
Αρ.Μετρητή           1008
Ημέρες Κατανάλωσης    223
Συντ.Kwh                4
Κατανάλωση KWh       6748
Αξία Ενέργειας      21432
Συν Προμηθ Ρευματος 23390
Σύνολο Λογαριασμού   3022
Σύν.Τρεχ.Μήνα        2938
Τύπος Λογ/σμου          4
Κατηγορία Υποδομής     24
dtype: int64
```

Εικόνα 3.16 Πληροφορίες αρχείου κατηγοριοποίησης

Μετά την εφαρμογή του κώδικα κατηγοριοποίησης των δημοτικών υποδομών και την αποθήκευση των δεδομένων στο αρχείο data_categorized.csv, παρατηρείται ότι το χαρακτηριστικό «Κατηγορία Υποδομής» μειώνει τις **754** αρχικές ξεχωριστές κατηγορίες (βάσει των μοναδικών τιμών του πεδίου «Ον.Πελάτη») σε **24 θεματικές κατηγορίες (κλάσεις)**.

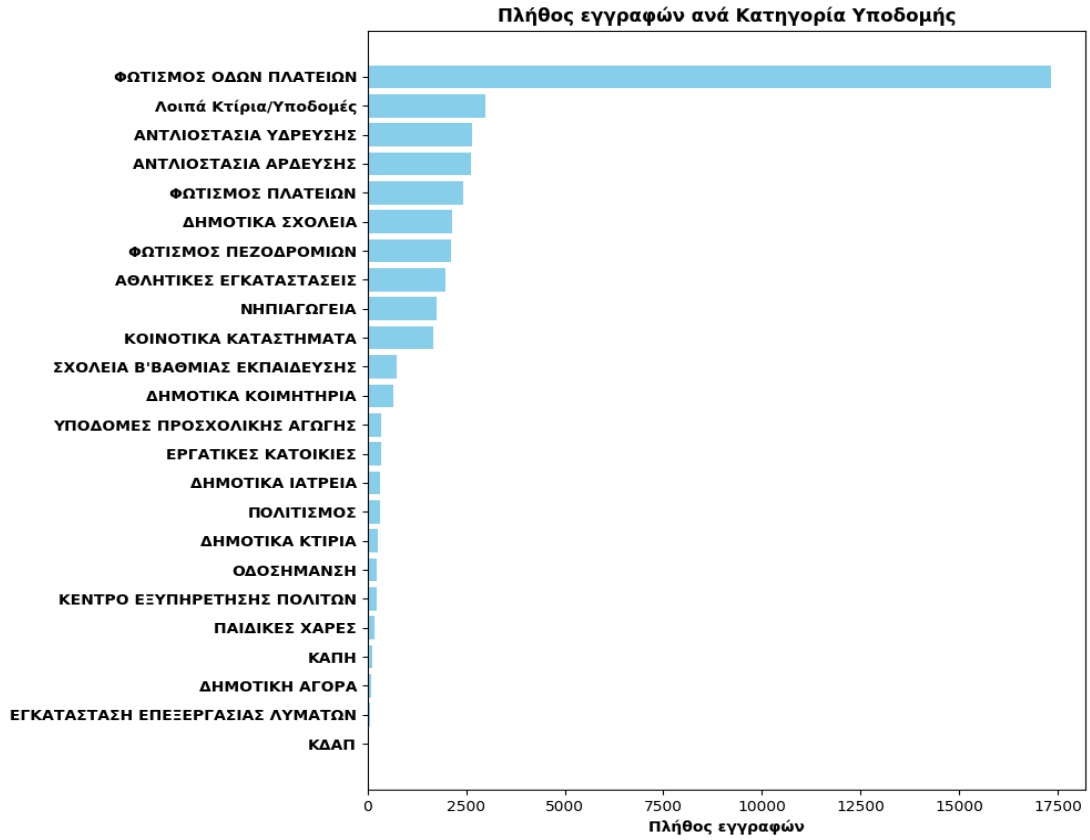
Ο έλεγχος του πλήθους των παρατηρήσεων, των πελατών και των παροχών ανά κατηγορία δείχνει ότι το συνολικό πλήθος των εγγραφών όλων των κατηγοριών ανέρχεται σε **41.568**,

το σύνολο των πελατών–καταναλωτών σε **754** και το σύνολο των παροχών σε **868** (Πίνακας 3.3). Τα αποτελέσματα αυτά επιβεβαιώνουν ότι όλοι οι μοναδικοί πελάτες–καταναλωτές έχουν αντιστοιχιστεί σε κάποια από τις 24 νέες θεματικές κατηγορίες του χαρακτηριστικού «Κατηγορία Υποδομής».

Τέλος, διευκρινίζεται ότι μεταξύ των νέων θεματικών κατηγοριών περιλαμβάνεται και η κατηγορία «**Λοιπά κτίρια/υποδομές**» στην οποία έχουν ενταχθεί πελάτες για τους οποίους δεν είναι εφικτή κάποια πιο συγκεκριμένη ή ασφαλής κατάταξη, διασφαλίζοντας έτσι ότι κανένας πελάτης δεν παραμένει χωρίς κατηγορία.

Α/Α	Κατηγορία Υποδομής	Πλήθος παρατηρήσεων	Πλήθος πελατών	Πλήθος παροχών
1	ΑΘΛΗΤΙΚΕΣ ΕΓΚΑΤΑΣΤΑΣΕΙΣ	1972	32	32
2	ΑΝΤΛΙΟΣΤΑΣΙΑ ΑΡΔΕΥΣΗΣ	2630	42	43
3	ΑΝΤΛΙΟΣΤΑΣΙΑ ΥΔΡΕΥΣΗΣ	2658	40	40
4	ΔΗΜΟΤΙΚΑ ΙΑΤΡΕΙΑ	319	7	7
5	ΔΗΜΟΤΙΚΑ ΚΟΙΜΗΤΗΡΙΑ	642	14	14
6	ΔΗΜΟΤΙΚΑ ΚΤΙΡΙΑ	256	4	5
7	ΔΗΜΟΤΙΚΑ ΣΧΟΛΕΙΑ	2148	39	41
8	ΔΗΜΟΤΙΚΗ ΑΓΟΡΑ	90	1	2
9	ΕΓΚΑΤΑΣΤΑΣΗ ΕΠΕΞΕΡΓΑΣΙΑΣ ΛΥΜΑΤΩΝ	71	1	1
10	ΕΡΓΑΤΙΚΕΣ ΚΑΤΟΙΚΙΕΣ	339	1	5
11	ΚΑΠΗ	118	2	2
12	ΚΔΑΠ	45	1	1
13	ΚΕΝΤΡΟ ΕΞΥΠΗΡΕΤΗΣΗΣ ΠΟΛΙΤΩΝ	222	3	3
14	ΚΟΙΝΟΤΙΚΑ ΚΑΤΑΣΤΗΜΑΤΑ	1653	30	34
15	Λοιπά Κτίρια/Υποδομές	3000	42	81
16	ΝΗΠΙΑΓΩΓΕΙΑ	1754	34	37
17	ΟΔΟΣΗΜΑΝΣΗ	224	5	5
18	ΠΑΙΔΙΚΕΣ ΧΑΡΕΣ	180	4	4
19	ΠΟΛΙΤΙΣΜΟΣ	306	4	5
20	ΣΧΟΛΕΙΑ Β'ΒΑΘΜΙΑΣ ΕΚΠΑΙΔΕΥΣΗΣ	738	11	11
21	ΥΠΟΔΟΜΕΣ ΠΡΟΣΧΟΛΙΚΗΣ ΑΓΩΓΗΣ	341	7	9
22	ΦΩΤΙΣΜΟΣ ΟΔΩΝ ΠΛΑΤΕΙΩΝ	17331	351	386
23	ΦΩΤΙΣΜΟΣ ΠΕΖΟΔΡΟΜΙΩΝ	2113	43	47
24	ΦΩΤΙΣΜΟΣ ΠΛΑΤΕΙΩΝ	2418	36	53
		41568	**754**	**868**

Πίνακας 3.3 Πλήθος εγγραφών - πελατών – παροχών ανά Κατηγορία Υποδομής



Εικόνα 3.17 Οπτικοποίηση αποτελεσμάτων κατηγοριοποίησης

3.5 Μετατροπή αρχικών δεδομένων

Τα δεδομένα ενεργειακής κατανάλωσης των δημοτικών υποδομών αποτελούνται από στιγμιότυπα (παρατηρήσεις) που έχουν καταγραφεί διαδοχικά στο χρόνο. Αυτό σημαίνει ότι συνιστούν χρονοσειρές (time series) όπου η κατανάλωση του ρεύματος - η τιμή δηλαδή του χαρακτηριστικού «Κατανάλωση KWh» - συνδέεται άμεσα με την πάροδο του χρόνου ως αποτέλεσμα τόσο της ημερομηνίας έκδοσης του λογαριασμού όσο και του πλήθους των ημερών κατανάλωσης. Παράγοντες όπως η εποχικότητα αλλά και η ιδιαιτερότητα κάθε πελάτη (π.χ. σχολεία, αντλιοστάσια, κτλ.) διαμορφώνουν καθοριστικά την ενεργειακή συμπεριφορά των δημοτικών υποδομών. Δεδομένου ότι πρόκειται για διακριτές χρονοσειρές (ανά μήνα), η χρονική αυτοσυσχέτιση μεταξύ των παρατηρήσεων - δηλαδή η χρονική απόσταση μεταξύ των διαδοχικών μετρήσεων για την ίδια δημοτική υποδομή ή κατηγορία υποδομών - αποτελεί βασικό κριτήριο τόσο για την επιλογή όσο και για τη σύγκριση μοντέλων μηχανικής μάθησης.

Τα πρωτογενή δεδομένα (row data) αφορούν τη μηνιαία κατανάλωση των δημοτικών υποδομών. Στην Εικόνα 3.16 παρατηρούμε ότι το χαρακτηριστικό **«Ημέρες Κατανάλωσης»** παρουσιάζει **223** διαφορετικές μοναδικές τιμές. Ο αριθμός αυτός καταδεικνύει ότι στο αρχείο περιλαμβάνονται λογαριασμοί κατανάλωσης οι οποίοι καλύπτουν χρονικές περιόδους σημαντικά μεγαλύτερες ή μικρότερες από μια τυπική μηνιαία περίοδο (31, 30, 28 ή 29 ημέρες, ανάλογα με τον μήνα και το έτος). Ως συνέπεια ο ίδιος πελάτης μπορεί σε κάποιο έτος να εμφανίζεται με έναν μόνο λογαριασμό, ενώ σε άλλο έτος να εμφανίζεται με περίπου δώδεκα λογαριασμούς (έναν ανά μήνα). Επιπλέον για τις εγγραφές που αφορούν λογαριασμούς με ημερομηνία έκδοσης εντός του 2020, μέρος της καταγεγραμμένης κατανάλωσης (και των αντίστοιχων «Ημερών Κατανάλωσης») ενδέχεται να αφορά περίοδο που εκκινεί εντός του 2019.

Εάν η ανάλυση βασιστεί αποκλειστικά στην ημερομηνία έκδοσης των λογαριασμών, υπάρχει κίνδυνος να εισαχθεί μεροληψία (bias) στην εκτίμηση της μηνιαίας ενεργειακής συμπεριφοράς μιας δημοτικής υποδομής. Ένας λογαριασμός μιας παροχής μπορεί να αποδίδεται σε συγκεκριμένο μήνα με βάση την ημερομηνία έκδοσης του, ενώ στην πραγματικότητα να καλύπτει διαφορετικό εύρος ημερών κατανάλωσης (άρα και μήνες) σε σχέση με τον λογαριασμό της ίδιας παροχής σε άλλο έτος για τον ίδιο μήνα, όπου το πλήθος των ημερών κατανάλωσης ταυτίζεται ή προσεγγίζει τις πραγματικές ημέρες του μήνα.

3.5.1 Δημιουργία νέων χαρακτηριστικών (feature creation)

Με βάση την ανωτέρω ανάλυση κρίνεται απαραίτητη η μετατροπή των αθροιστικών τιμών του χαρακτηριστικού «Κατανάλωση KWh», οι οποίες συνδέονται με συγκεκριμένο αριθμό ημερών κατανάλωσης (χαρακτηριστικό «Ημέρες Κατανάλωσης»), σε τιμές που απεικονίζουν την κατανάλωση ανά ημέρα.

Για τον σκοπό αυτό δημιουργούνται νέα χαρακτηριστικά: «**Ημερομηνία**», «**Έτος κατανάλωσης**», «**Μήνας κατανάλωσης**», «**Ημερήσια αξία ενέργειας**» και «**Ημερήσια κατανάλωση KWh**» ανά αριθμό παροχής ρεύματος και τα αποτελέσματα αποθηκεύονται σε νέο αρχείο με όνομα **data_per_day.csv**.

Η διαδικασία αυτή υλοποιείται τεχνητά μέσω μιας συνάρτησης (Εικόνα 3.18), η οποία επεκτείνει κάθε εγγραφή λογαριασμού που αφορά συγκεκριμένο αριθμό παροχής («Αρ.Παροχής») ο οποίος ανήκει σε συγκεκριμένο πελάτη («Ον.Πελάτη») που εντάσσεται σε μια συγκεκριμένη κατηγορία υποδομής («Κατηγορία Υποδομής»), σε τόσες νέες εγγραφές όσες είναι οι «**Ημέρες Κατανάλωσης**» που προηγούνται (συμπεριλαμβανομένης και) της ημερομηνίας έκδοσής του.

```
# Συνάρτηση ημερήσιας κατανάλωσης
def daily_consume(df,
                  year_col='Ετος',
                  month_col='Μήνας',
                  bill_date_col='Ημ/νία Λογ/σμου',
                  multiple_num_col='Αρ. Πολλαπλου',
                  multiple_name_col='Ονομα Πολλαπλού',
                  power_supply_num_col='Αρ.Παροχής',
                  invoice_col='Τιμολόγιο',
                  customer_col='Ον.Πελάτη',
                  address_col='Οδός',
                  number_col='Αριθ.',
                  counter_num_col='Αρ.Μετρητή',
                  days_consumption_col='Ημέρες Κατανάλωσης',
                  coefficient_KWh_col='Συντ.Kwh',
                  kwh_col='Κατανάλωση KWh',
                  price_energy_col='Αξία Ενέργειας',
                  total_supply_col='Συν Προμηθ Ρευματος',
                  total_account_col='Σύνολο Λογαριασμού',
                  total_current_month_col='Σύν.Τρεχ.Μήνα',
                  type_account_col='Τύπος Λογ/σμου',
                  category_col='Κατηγορία Υποδομής',
                  # νέα γνωρίσματα-χαρακτηριστικά
                  date_first_col='Ημερομηνία',
                  year_consumtion_col='Ετος κατανάλωσης',
                  month_consumtion_col='Μήνας κατανάλωσης',
                  price_energy_per_day_col='Ημερήσια αξία ενέργειας',
                  Kwh_per_day_col='Ημερήσια κατανάλωση KWh'
                  ):
```

Εικόνα 3.18 Συνάρτηση ημερήσιας τεχνητής κατανομής κατανάλωσης

```
# Βοηθητική λίστα δημιουργίας τελικού data frame
records = []
# Ανάγνωση γραμμών data frame <-- data_categorized.csv
for _, r in df.iterrows():
    d = int(r[days_consumption_col])
    if d <= 0:
        continue
    end = r[bill_date_col].normalize()       # ημερομηνία τέλους
    start = end - pd.Timedelta(days=d - 1)   # ημερομηνία αρχής
    # Υπολογισμός ημερήσιας κατανάλωσης - αξίας
    kwh_per_day = round(float(r[kwh_col]) / d, 2)
    price_energy_per_day = round(float(r[price_energy_col]) / d, 2)
```

```
    # Δημιουργία ημερήσιου εύρους ημερών
    idx = pd.date_range(start, end, freq='D')
    # αντιστοίχιση ημέρας με όλα τα γνωρίσματα
    for day in idx:
        records.append({
            year_col: r[year_col],
            month_col: r[month_col],
            bill_date_col: r[bill_date_col],
            multiple_num_col: r[multiple_num_col],
            multiple_name_col: r[multiple_name_col],
            power_supply_num_col: r[power_supply_num_col],
            invoice_col: r[invoice_col],
            customer_col: r[customer_col],
            address_col: r[address_col],
            number_col: r[number_col],
            counter_num_col: r[counter_num_col],
            days_consumption_col: r[days_consumption_col],
            coefficient_KWh_col: r[coefficient_KWh_col],
            kwh_col: r[kwh_col],
            price_energy_col: r[price_energy_col],
            total_supply_col: r[total_supply_col],
            total_account_col: r[total_account_col],
            total_current_month_col: r[total_current_month_col],
            type_account_col: r[type_account_col],
            category_col: r[category_col],
            # νέα γνωρίσματα-χαρακτηριστικά
            date_first_col: day,
            year_consumtion_col: day.year,
            month_consumtion_col: day.month,
            price_energy_per_day_col: price_energy_per_day,
            Kwh_per_day_col: kwh_per_day,
        })
# δημιουργία τελικού data frame
daily = pd.DataFrame.from_records(records)
return daily
```

Εικόνα 3.19 Σώμα συνάρτησης ημερήσιας κατανομής ενέργειας

Για κάθε λογαριασμό ρεύματος δημιουργείται ένα εύρος διαδοχικών ημερομηνιών που αντιστοιχούν στις «Ημέρες Κατανάλωσης» που προηγούνται μέχρι και την ημερομηνία έκδοσής του (Εικόνα 3.19). Σε κάθε μία από αυτές τις ημερήσιες εγγραφές αποδίδεται ένα

κλάσμα της συνολικής κατανάλωσης του λογαριασμού, με ομοιόμορφη κατανομή της τιμής του χαρακτηριστικού «Κατανάλωση KWh» στις αντίστοιχες ημέρες κατανάλωσης του χαρακτηριστικού «Ημερομηνία».

Από τις πληροφορίες του νέου αρχείου data_per_day.csv (Εικόνα 3.20) προκύπτει αύξηση των εγγραφών από 41.571 σε 2.466.260 και η προσθήκη των νέων χαρακτηριστικών: «**Ημερομηνία**», «**Έτος κατανάλωσης**», «**Μήνας κατανάλωσης**», «**Ημερήσια αξία ενέργειας**» και «**Ημερήσια κατανάλωση KWh**» ανά αριθμό παροχής.

```
Πληροφορίες αρχείου ημερήσιας κατανομής KWh
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2466260 entries, 0 to 2466259
Data columns (total 25 columns):
 #   Column                  Dtype
---  ------                  -----
 0   Ετος                    int64
 1   Μήνας                   int64
 2   Ημ/νία Λογ/σμου         datetime64[ns]
 3   Αρ. Πολλαπλου           object
 4   Ονομα Πολλαπλού         object
 5   Αρ.Παροχής              int64
 6   Τιμολόγιο               object
 7   Ον.Πελάτη               object
 8   Οδός                    object
 9   Αριθ.                   object
 10  Αρ.Μετρητή              object
 11  Ημέρες Κατανάλωσης      int64
 12  Συντ.Kwh                int64
 13  Κατανάλωση KWh          int64
 14  Αξία Ενέργειας          float64
 15  Συν Προμηθ Ρευματος     float64
 16  Σύνολο Λογαριασμού      float64
 17  Σύν.Τρεχ.Μήνα           float64
 18  Τύπος Λογ/σμου          object
 19  Κατηγορία Υποδομής      object
 20  Ημερομηνία              datetime64[ns]
 21  Ετος κατανάλωσης        int64
 22  Μήνας κατανάλωσης       int64
 23  Ημερήσια αξία ενέργειας float64
 24  Ημερήσια κατανάλωση KWh float64
dtypes: datetime64[ns](2), float64(6), int64(8), object(9)
```

```
Πλήθος μοναδικών τιμών ανά χακτηριστικο

Ετος                        6
Μήνας                      12
Ημ/νία Λογ/σμου          1372
Αρ. Πολλαπλου              25
Ονομα Πολλαπλού            25
Αρ.Παροχής                868
Τιμολόγιο                   8
Ον.Πελάτη                 754
Οδός                      169
Αριθ.                      65
Αρ.Μετρητή                1008
Ημέρες Κατανάλωσης         223
Συντ.Kwh                     4
Κατανάλωση KWh            6748
Αξία Ενέργειας           21432
Συν Προμηθ Ρευματος      23390
Σύνολο Λογαριασμού        3022
Σύν.Τρεχ.Μήνα             2938
Τύπος Λογ/σμου               4
Κατηγορία Υποδομής          24
Ημερομηνία                2537
Ετος κατανάλωσης             7
Μήνας κατανάλωσης           12
Ημερήσια αξία ενέργειας   3992
Ημερήσια κατανάλωση KWh    8793
dtype: int64
```

Εικόνα 3.20 Πληροφορίες αρχείου ημερήσιας κατανομής ενέργειας

Κάθε εγγραφή (από τις 2.466.260 γραμμές) του νέου αρχείου data_per_day.csv εμφανίζει πλέον την ημερήσια κατανάλωση KWh («Ημερήσια κατανάλωση KWh») που αντιστοιχεί σε συγκεκριμένη ημέρα του έτους («Ημερομηνία»). Αυτό έχει ως αποτέλεσμα μια πιο ομαλή ημερήσια ροή των δεδομένων στο χρόνο που αποτελεί βασική ιδιότητα των χρονικών σειρών.

Η προσέγγιση που υλοποιήθηκε, αν και τεχνητή, μας δίνει τη δυνατότητα να ανακατασκευάσουμε με ικανοποιητική ακρίβεια τους πραγματικούς μήνες στους οποίους αντιστοιχεί η κατανάλωση ηλεκτρικής ενέργειας ανά έτος (Εικόνα 3.21), διαμορφώνοντας μια πιο συνεπή χρονική βάση για την ανάλυση και τη μοντελοποίηση των δεδομένων.

Η ημερήσια τεχνητή κατανομή παρότι εφαρμόζεται στο σύνολο των λογαριασμών (εγγραφών), ουσιαστικά αφορά μόνο εκείνους τους λογαριασμούς για τους οποίους οι ημέρες κατανάλωσης αποκλίνουν σημαντικά από τις ημέρες ενός τυπικού μήνα.

Για τους μήνες αυτούς που πλέον εμφανίζονται ενδεχομένως με μια πιο ομοιόμορφη κατανομή ενέργειας στο εσωτερικό τους, η παρέμβαση αυτή δεν αλλοιώνει το συνολικό ύψος της κατανάλωσης αλλά βελτιώνει τη συγκρισιμότητα των μετρήσεων μεταξύ διαφορετικών ετών και υποδομών, μειώνοντας τη μεροληψία (bias) που εισάγεται από τις άνισες περιόδους κατανάλωσης.

```
# Ανακατασκευή πραγματικών μηνών κατανάλωσης
df_month = (
    df
     .groupby(["Ημ/νία Λογ/σμου","Αρ.Παροχής","Αρ.Μετρητή",
               "Μήνας κατανάλωσης","Ετος κατανάλωσης",],as_index=False)
     .agg({
        "Ημερήσια κατανάλωση KWh": "sum",     # κατανάλωσης KWh μήνα
        "Ημερήσια αξία ενέργειας": "sum",     # αξία ενέργειας μήνα

        "Ημερομηνία": "first",
        "Ετος": "first",
        "Μήνας": "first",
        "Αρ. Πολλαπλου": "first",
        "Ονομα Πολλαπλού": "first",
        "Τιμολόγιο": "first",
        "Ον.Πελάτη": "first",
        "Οδός": "first",
        "Αριθ.": "first",
        "Ημέρες Κατανάλωσης": "first",
        "Συντ.Kwh": "first",
        "Κατανάλωση KWh": "first",
        "Αξία Ενέργειας": "first",
        "Συν Προμηθ Ρευματος": "first",
        "Σύνολο Λογαριασμού": "first",
        "Σύν.Τρεχ.Μήνα": "first",
        "Τύπος Λογ/σμου": "first",
        "Κατηγορία Υποδομής": "first",
    }).rename(columns={
        "Ημερομηνία": "1η Ημερομηνία Μήνα",
        "Ημερήσια κατανάλωση KWh": "Μηνιαία κατανάλωση KWh",
        "Ημερήσια αξία ενέργειας": "Μηνιαία αξία ενέργειας",
    })
```

```python
# Συγκεντρωτική μηνιαία κατανάλωση
df_monthly = (
    df_month
     .groupby(["Αρ.Παροχής","Ετος κατανάλωσης","Μήνας κατανάλωσης"],as_index=False)
     .agg({
        "Μηνιαία κατανάλωση KWh": "sum",     # αθροιστική μηνιαία κατανάλωση KWh
        "Μηνιαία αξία ενέργειας": "sum",     # αθροιστική μηνιαία αξία ενέργειας

        "Ημ/νία Λογ/σμου": "first",
        "1η Ημερομηνία Μήνα": "first",
        "Ετος": "first",
        "Μήνας": "first",
        "Αρ. Πολλαπλου": "first",
        "Ονομα Πολλαπλού": "first",
        "Τιμολόγιο": "first",
        "Ον.Πελάτη": "first",
        "Οδός": "first",
        "Αριθ.": "first",
        "Αρ.Μετρητή": "first",
        "Ημέρες Κατανάλωσης": "first",
        "Συντ.Kwh": "first",
        "Κατανάλωση KWh": "first",
        "Αξία Ενέργειας": "first",
        "Συν Προμηθ Ρευματος": "first",
        "Σύνολο Λογαριασμού": "first",
        "Σύν.Τρεχ.Μήνα": "first",
        "Τύπος Λογ/σμου": "first",
        "Κατηγορία Υποδομής": "first",
    }))
```

Εικόνα 3.21 Κώδικας ανακατασκευής μηνιαίας ενεργειακής κατανάλωσης

Πληροφορίες μηνιαίου αρχείου (data_per_month.csv)

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 61935 entries, 0 to 61934
Data columns (total 25 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   Αρ.Παροχής             61935 non-null  int64
 1   Ετος κατανάλωσης       61935 non-null  int64
 2   Μήνας κατανάλωσης      61935 non-null  int64
 3   Μηνιαία κατανάλωση KWh  61935 non-null  float64
 4   Μηνιαία αξία ενέργειας  61935 non-null  float64
 5   Ημ/νία Λογ/σμου         61935 non-null  object
 6   1η Ημερομηνία Μήνα      61935 non-null  object
 7   Ετος                    61935 non-null  int64
 8   Μήνας                   61935 non-null  int64
 9   Αρ. Πολλαπλου           61935 non-null  object
 10  Ονομα Πολλαπλού         61935 non-null  object
 11  Τιμολόγιο               61935 non-null  object
 12  Ον.Πελάτη               61935 non-null  object
 13  Οδός                    61694 non-null  object
 14  Αριθ.                   6045 non-null   object
 15  Αρ.Μετρητή              61935 non-null  string
 16  Ημέρες Κατανάλωσης      61935 non-null  int64
 17  Συντ.Kwh                61935 non-null  int64
 18  Κατανάλωση KWh          61935 non-null  int64
 19  Αξία Ενέργειας          61935 non-null  float64
 20  Συν Προμηθ Ρευματος     61935 non-null  float64
 21  Σύνολο Λογαριασμού      61935 non-null  float64
 22  Σύν.Τρεχ.Μήνα           61935 non-null  float64
 23  Τύπος Λογ/σμου          61913 non-null  object
 24  Κατηγορία Υποδομής      61935 non-null  object
dtypes: float64(6), int64(8), object(10), string(1)
```

Πλήθος μοναδικών τιμών ανά χαρτηριστικο

Αρ.Παροχής	868
Ετος κατανάλωσης	7
Μήνας κατανάλωσης	12
Μηνιαία κατανάλωση KWh	31952
Μηνιαία αξία ενέργειας	23568
Ημ/νία Λογ/σμου	1367
1η Ημερομηνία Μήνα	341
Ετος	6
Μήνας	12
Αρ. Πολλαπλου	25
Ονομα Πολλαπλού	25
Τιμολόγιο	8
Ον.Πελάτη	754
Οδός	169
Αριθ.	65
Αρ.Μετρητή	1006
Ημέρες Κατανάλωσης	213
Συντ.Kwh	4
Κατανάλωση KWh	6631
Αξία Ενέργειας	21177
Συν Προμηθ Ρευματος	23100
Σύνολο Λογαριασμού	2906
Σύν.Τρεχ.Μήνα	2848
Τύπος Λογ/σμου	4
Κατηγορία Υποδομής	24

dtype: int64

Εικόνα 3.22 Πληροφορίες νέου αρχείου μηνιαίας κατανάλωσης

Στην Εικόνα 3.22 παρουσιάζονται οι πληροφορίες του αρχείου **data_per_month.csv**, όπου αξιοποιώντας τα χαρακτηριστικά «Ημ/νία Λογ/σμου», «Αρ.Παροχής», «Μήνας κατανάλωσης», «Ετος κατανάλωσης» από το αρχείο της ημερήσιας τεχνητής κατανομής (data_per_day.csv**)** έχει ανακατασκευαστεί η συγκεντρωτική μηνιαία αναφορά κατανάλωσης **με βάση τον αριθμό παροχής** ο οποίος προσδιορίζει μοναδικά κάθε λογαριασμό ηλεκτρικού ρεύματος. Τονίζεται, όπως έχει ήδη αναφερθεί στην ενότητα 3.4.3, ο αριθμός παροχής δεν προσδιορίζει μοναδικά κάθε πελάτη («Ον.Πελάτη») διότι ο ίδιος πελάτης μπορεί να έχει περισσότερες από μία παροχές ρεύματος.

Ελέγχοντας την ακεραιότητα της ανωτέρω διαδικασίας σχετικά με τον κίνδυνο απώλειας δεδομένων μετά από τις αναγκαίες μετατροπές διαπιστώνουμε ότι ο αριθμός των μοναδικών τιμών των χαρακτηριστικών **«Αρ.Παροχής» = 868**, **«Ον.Πελάτη» = 754** και **«Κατηγορία Υποδομής»=24** των αρχείων **energy_2020-2025.csv** (Εικόνα 3.5), **data_categorized.csv** (Εικόνα 3.16) και **data_per_month.csv** (Εικόνα 3.22) παραμένει ο ίδιος. Διευκρινίζεται ότι το χαρακτηριστικό «Αρ.Μετρητή» εμφανίζεται με μικρότερο αριθμό μοναδικών τιμών, διότι στην ανακατασκευή της μηνιαίας κατανάλωσης χρησιμοποιήθηκε αποκλειστικά το χαρακτηριστικό «Αρ.Παροχής». Τα στιγμιότυπα (εγγραφές) σε σχέση με το αρχείο clean_energ.csv και το αρχείο κατηγοριοποίησης data_categorized.csv έχουν αυξηθεί από 41.568 σε 61.935.

3.6. Ερμηνευσιμότητα χαρακτηριστικών

Η διαδικασία που ακολουθήθηκε στις ενότητες 3.4.4 και 3.5.1, οδήγησε στη διαμόρφωση ενός πιο συμπαγούς και δομημένου συνόλου δεδομένων όπου:

- οι πολλές και ετερογενείς ονομασίες δημοτικών υποδομών έχουν αντικατασταθεί από λίγες, συνεκτικές θεματικές κατηγορίες (χαρακτηριστικό **«Κατηγορία Υποδομής»**),
- η χρονική πληροφορία έχει αποτυπωθεί με ακρίβεια τόσο σε επίπεδο έτους (**«Έτος κατανάλωσης»**) όσο και σε επίπεδο μήνα πραγματικής κατανάλωσης **(«Μήνας κατανάλωσης»)**

Τα αρχικά χαρακτηριστικά **«Έτος»** και **«Μήνας»** του αρχείου energ_2020-2025.csv ενσωματώνονται ουσιαστικά στα νέα γνωρίσματα **«Έτος κατανάλωσης»** και **«Μήνας**

κατανάλωσης» ώστε να αποτυπώνεται ο πραγματικός χρόνος κατανάλωσης και όχι απλώς ο χρόνος έκδοσης του λογαριασμού.

Με βάση τις πληροφορίες του νέου αρχείο data_per_month.csv (Εικόνα 3.22) ως μεταβλητή-στόχος της ανάλυσης ορίζεται πλέον το χαρακτηριστικό:

y = «Μηνιαία κατανάλωση KWh».

Τα χαρακτηριστικά: **«Ημ/νία Λογ/σμου»**, **«Ετος»**, **«Μήνας»**, **«Ημέρες Κατανάλωσης»**, **«Κατανάλωση KWh»**, **«Αξία Ενέργειας»**, **«Συν Προμηθ Ρευματος»**, **«Σύνολο Λογαριασμού»** και **«Σύν.Τρεχ.Μήνα»** συνδέονται αποκλειστικά με τις αρχικές εγγραφές των λογαριασμών του αρχείου των ενεργειακών δεδομένων energ_2020-2025.csv και όχι με τις επιμέρους εγγραφές (στιγμιότυπα) που προέκυψαν από την εφαρμογή της ημερήσιας τεχνητής κατανομής της ενέργειας και τον υπολογισμό της τεχνητής (εν μέρει) πραγματικής μηνιαίας κατανάλωσης και αξίας στο αρχείο data_per_month.csv. Σε αρκετές περιπτώσεις η πληροφορία τους είτε έχει ήδη αναδιατυπωθεί (π.χ. σε «Έτος κατανάλωσης», «Μήνας κατανάλωσης») είτε δεν είναι πλήρως συνεπής με τη νέα μηνιαία χρονοσειρά.

Συνεπώς, η διατήρησή των ανωτέρω χαρακτηριστικών στο τελικό σύνολο δεδομένων, εισάγει πλεονάζουσα και εν μέρει ασύμβατη πληροφορία σε σχέση με τη νέα, ανακατασκευασμένη μηνιαία χρονοσειρά και για τον λόγο αυτό δεν συμπεριλήφθηκαν στα επόμενα στάδια της ερευνητικής διαδικασίας. Επίσης το χαρακτηριστικό **«1η Ημερομηνία Μήνα»** (rename «Ημερομηνία»), προσδιορίζει την ημερομηνία έναρξης της κατανάλωσης σε κάθε μήνα του χαρακτηριστικού «Μήνας κατανάλωσης» και θα χρησιμοποιηθεί **μόνο** για τη χρονική ταξινόμηση των παρατηρήσεων του αρχείου data_per_month.csv.

Στη συνέχεια εξετάζεται η ερμηνευσιμότητα (Explainable AI) των υπόλοιπων χαρακτηριστικών (ανεξάρτητες μεταβλητές X) του αρχείου data_per_month.csv και εκτιμάται η συμβολή τους στη διαμόρφωση της μεταβλητής-στόχου **y = «Μηνιαία κατανάλωση KWh»** (Molnar et al., 2024)**.**

Ο σκοπός της ενέργειας αυτής είναι διττός:

1. **Ανίχνευση και επιλογή χαρακτηριστικών (feature subset selection)**, δηλαδή ο εντοπισμός και η διατήρηση μόνο εκείνων των χαρακτηριστικών που συνδέονται άμεσα με τη διαμόρφωση της τιμής του χαρακτηριστικού **«Μηνιαία κατανάλωση KWh»** ανά πελάτη, και κατ' επέκταση ανά κατηγορία δημοτικής υποδομής.

2. **Περαιτέρω μείωση της διάστασης του προβλήματος** με στόχο τη βελτίωση της ερμηνευσιμότητας και της ανάλυσης των ενεργειακών δεδομένων καθώς και τον περιορισμό των απαιτήσεων της ερευνητικής μεθοδολογίας σε αποθηκευτικό χώρο και υπολογιστική ισχύ.

Το χαρακτηριστικό «Μηνιαία κατανάλωση KWh» ορίζεται ως εξαρτημένη μεταβλητή-στόχος y ενώ τα υπόλοιπα αριθμητικά και κατηγορικά χαρακτηριστικά, ορίζονται ως ανεξάρτητες μεταβλητές *X*. Δεδομένου ότι η μεταβλητή-στόχος y είναι συνεχής αριθμητική τιμή, η πρόβλεψή της συνιστά πρόβλημα παλινδρόμησης και υλοποιείται με τον αλγόριθμο μηχανικής μάθησης Random Forest στο αρχείο data_per_month.csv. Στο πλαίσιο αυτό πραγματοποιείται ανάλυση και εκτίμηση της σημαντικότητας κάθε χαρακτηριστικού του συνόλου *X,* στην πρόβλεψης της τιμής στόχου y (Goodfellow et al., 2016).

Ο Random Forest είναι ένας αλγόριθμος που βασίζεται σε ένα σύνολο (ensemble) από δέντρα απόφασης (forest of decision trees). Κάθε δέντρο εκπαιδεύεται σε διαφορετικό, τυχαίο υποσύνολο των παρατηρήσεων (bootstrap δείγμα), καθώς και σε τυχαίο υποσύνολο χαρακτηριστικών σε κάθε κόμβο διάσπασης. Στόχος του μοντέλου είναι να «μάθει» τη σχέση μεταξύ των χαρακτηριστικών του συνόλου X (π.χ. «Αρ.Παροχής», «Ετος κατανάλωσης», «Μήνας κατανάλωσης», κτλ.) και της μεταβλητής-στόχου y = «Μηνιαία κατανάλωση KWh» (Hastie et al., 2009).

Η τελική πρόβλεψη προκύπτει ως ο μέσος όρος των προβλέψεων όλων των δέντρων, γεγονός που οδηγεί σε αυξημένη σταθερότητα και μειωμένο κίνδυνο υπερπροσαρμογής (overfitting) σε σχέση με ένα μεμονωμένο δέντρο απόφασης.

Η εκπαίδευση του Random Forest γίνεται σε ένα σύνολο εκπαίδευσης (training set) στο οποίο κατασκευάζονται πολλά δέντρα απόφασης. Κάθε δέντρο επιλέγει με επαναδειγματοληψία (bootstrap) το δικό του τυχαίο δείγμα παρατηρήσεων/εγγραφών. Σε κάθε δέντρο η διαδικασία εκκινεί από τη ρίζα (ριζικός κόμβος απόφασης – root node) όπου εφαρμόζεται ένας αρχικός κανόνας διάσπασης πάνω σε κάποιο χαρακτηριστικό του συνόλου X (π.χ. «Έτος κατανάλωσης» = 2020).

Σε κάθε εσωτερικό κόμβο του δέντρου ο αλγόριθμος εξετάζει υποψήφιους κανόνες διάσπασης πάνω σε ένα τυχαίο υποσύνολο των χαρακτηριστικών X και επιλέγει εκείνον που επιτυγχάνει τη μεγαλύτερη μείωση μιας συνάρτησης κόστους, η οποία για προβλήματα παλινδρόμησης είναι συνήθως το Μέσο Τετραγωνικό Σφάλμα (Mean Squared Error -

MSE) (Hastie et al., 2009). Η συνάρτηση κόστους σχετίζεται με την απόκλιση μεταξύ της πραγματικής τιμής της μεταβλητής στόχου y_{true} («Μηνιαία κατανάλωση KWh») και της αντίστοιχης τιμής πρόβλεψης y_{pred}.

Έτσι σε κάθε κόμβο οι παρατηρήσεις κατευθύνονται στο αριστερό ή στο δεξί υποδέντρο ανάλογα με το αν ικανοποιούν ή όχι τον κανόνα διάσπασης. Η διαδικασία συνεχίζεται με διαδοχικές αποφάσεις σε εσωτερικούς κόμβους μέχρι να ικανοποιηθούν κάποια κριτήρια τερματισμού (π.χ. μέγιστο βάθος δέντρου, ελάχιστος αριθμός παρατηρήσεων ανά κόμβο κτλ.). Με αυτόν τον τρόπο επιτυγχάνεται σταδιακή βελτίωση στην ακρίβεια των προβλέψεων της εξαρτημένης μεταβλητής y = «Μηνιαία κατανάλωση KWh». Κατά την εκπαίδευση του αλγορίθμου κάθε χαρακτηριστικό του συνόλου X συμβάλει μέσω των κόμβων στους οποίους έχει χρησιμοποιηθεί ως κριτήριο διάσπασης στη δυνατότητα του μοντέλου να αποδώσει ακριβείς προβλέψεις για την μεταβλητή-στόχο y.

Οι τελικοί κόμβοι (φύλλα – leaves) δεν ταυτίζονται με τις πραγματικές τιμές της μεταβλητής-στόχου y_{true}, αλλά περιέχουν την εκτίμηση (πρόβλεψη) της τιμής y_{pred} , η οποία σε δέντρα παλινδρόμησης αντιστοιχεί συνήθως στη μέση τιμή του χαρακτηριστικού «Μηνιαία κατανάλωση KWh» των παρατηρήσεων που καταλήγουν στο συγκεκριμένο φύλλο.

Για παράδειγμα ένα φύλλο (τελικός κόμβος – «Μηνιαία κατανάλωση KWh») μπορεί να αντιστοιχεί σε ένα μονοπάτι αποφάσεων:

- **node1** → Έτος κατανάλωσης = 2020
- **node2** → Μήνας κατανάλωσης = 2
- **node3** → Κατηγορία Υποδομής = ΦΩΤΙΣΜΟΣ ΟΔΩΝ ΠΛΑΤΕΙΩΝ

το οποίο δίνει ως πρόβλεψη μια μέση τιμή π.χ. **leaf** → y_{pred} = 2.500 KWh. Μέσω αυτής της διαδικασίας ο αλγόριθμος μαθαίνει τη σχέση των ανεξάρτητων μεταβλητών X με την εξαρτημένη μεταβλητή y (Hastie et al., 2009).

Όσο περισσότερο συμβάλλει ένα χαρακτηριστικό του συνόλου X στην βελτίωση της ακρίβειας των προβλέψεων της μεταβλητής-στόχου τόσο πιο σημαντική είναι η σχέση της εξάρτησης του με την εξαρτημένη μεταβλητή y = «Μηνιαία κατανάλωση KWh».

Η εκτίμηση του μέτρου «σημαντικότητας» των χαρακτηριστικών (feature importance) που ανήκουν στο σύνολο των ανεξάρτητων μεταβλητών του συνόλου X στην παρούσα εργασία πραγματοποιείται με δύο τεχνικές ερμηνείας: με την ενσωματωμένη ιδιότητα (attribute)

του Random Forest **feature_importances_** και τη μέθοδο **Permutation Importance** που είναι ανεξάρτητη από το μοντέλο (Hastie et al., 2009· Kuhn & Johnson, 2013).

Η ενσωματωμένη ιδιότητα feature_importances_ του Random Forest θεωρεί ένα χαρακτηριστικό σημαντικό όταν αυτό χρησιμοποιείται ως μεταβλητή στους κόμβους πολλών δέντρων για την εύρεση του βέλτιστου κριτηρίου διάσπασης που μειώνει το σφάλμα πρόβλεψης της μεταβλητής-στόχου, όπως το Μέσο Τετραγωνικό Σφάλμα – MSE σε προβλήματα παλινδρόμησης.

Η μέθοδος Permutation Importance διαταράσσει σκόπιμα τη σχέση εξάρτησης κάποιου χαρακτηριστικού με την εξαρτημένη μεταβλητή στόχο y, όπως αυτή έχει ήδη ενσωματωθεί στο μοντέλο κατά το στάδιο της εκπαίδευσης, μέσω τυχαίας αναδιάταξης (permutation) των τιμών του στο σύνολο ελέγχου. Το μοντέλο δεν εκπαιδεύεται ξανά απλώς αξιολογείται εκ νέου η απόδοσή του πάνω σε δεδομένα όπου οι τιμές του συγκεκριμένου χαρακτηριστικού έχουν «ανακατευτεί» (permute) τεχνητά. Εάν το αποτέλεσμα αυτής της διαταραχής επιδεινώνει την ακρίβεια των προβλέψεων και κατ' επέκταση μειώνει την απόδοση του μοντέλου, τότε το συγκεκριμένο χαρακτηριστικό κρίνεται σημαντικό για την εφαρμογή μεθόδων μηχανικής μάθησης (υψηλό feature importance). Αντίθετα, αν η απόδοση δεν μεταβάλλεται ουσιαστικά το χαρακτηριστικό θεωρείται λιγότερο σημαντικό.

Ο έλεγχος της ορθότητας των προβλέψεων του Random Forest γίνεται σε ένα ξεχωριστό υποσύνολο ελέγχου (test set) που δεν χρησιμοποιείται για την εκπαίδευση αλλά μόνο για την αξιολόγηση της απόδοσης του μοντέλου με τη χρήση κατάλληλων μετρικών για προβλήματα παλινδρόμησης, όπως ο συντελεστής προσδιορισμού R^2 (coefficient of determination), το Μέσο Απόλυτο Σφάλμα (Mean Absolute Error-MAE), το Μέσο Τετραγωνικό Σφάλμα (Mean Squared Error – MSE), και η Τετραγωνική Ρίζα (Root Mean Squared Error – RMSE) του MSE.

Συντελεστής προσδιορισμού (coefficient of determination) R^2

Για κάθε παρατήρηση $i = 1, \dots, n$*:*

- *πραγματική τιμή:* $\boldsymbol{y_{true_i}}$ *(πραγματική κατανάλωση KWh)*
- *πρόβλεψη μοντέλου:* $\boldsymbol{y_{pred_i}}$
- *μέσος όρος όλων των πραγματικών τιμών:*

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_{true_i}$$

- άθροισμα των τετραγωνικών υπολοίπων (σφάλματα του μοντέλου)

$$SS_{\text{res}} = \sum_{i=1}^{n} \left(y_{true_i} - y_{pred_i}\right)^2$$

- συνολική διασπορά των παρατηρήσεων γύρω από τον μέσο όρο $\bar{y}$

$$SS_{\text{tot}} = \sum_{i=1}^{n} (\boldsymbol{y_{true_i}} - \bar{\boldsymbol{y}})^2$$

Ο δείκτης R^2 ορίζεται ως:

$$R^2 = 1 - \frac{SS_{\text{res}}}{SS_{\text{tot}}}$$

και εκφράζει το ποσοστό της συνολικής διακύμανσης της εξαρτημένης μεταβλητής-στόχου y = «Μηνιαία κατανάλωση KWh» που εξηγείται από το μοντέλο. Τιμές κοντά στη μονάδα υποδηλώνουν καλή προσαρμογή και τιμές κοντά στο μηδέν ή αρνητικές υποδεικνύουν ανεπαρκή προσαρμογή.

Μέσο Απόλυτο Σφάλμα - Mean Absolute Error MAE

Για κάθε παρατήρηση $i = 1, \dots, n$*:*

- *πραγματική τιμή:* $\boldsymbol{y_{true_i}}$ *(πραγματική κατανάλωση KWh)*
- *πρόβλεψη μοντέλου:* $\boldsymbol{y_{pred_i}}$
- Το μέσο απόλυτο σφάλμα είναι:

$$\boldsymbol{MAE} = \frac{1}{n}\sum_{i=1}^{n} |\boldsymbol{y_{true_i}} - \boldsymbol{y_{pred_i}}|$$

και υπολογίζει το μέσο όρο των λαθών πρόβλεψης του μοντέλου σε KWh, για την μεταβλητή-στόχο y = «Μηνιαία κατανάλωση KWh» ανά μήνα (kWh/μήνα)

Μέσο Τετραγωνικό Σφάλμα - Mean Squared Error MSE

Για κάθε παρατήρηση $i = 1, \dots, n$*:*

- *πραγματική τιμή:* $\boldsymbol{y_{true_i}}$ *(πραγματική κατανάλωση KWh)*
- *πρόβλεψη μοντέλου:* $\boldsymbol{y_{pred_i}}$
- ορίζουμε το σφάλμα ως:

$$e_i = \boldsymbol{y_{true_i}} - \boldsymbol{y_{pred_i}}$$

- ορίζουμε το τετραγωνικό σφάλμα ως:

$$e_i^2 = (\boldsymbol{y_{true_i}} - \boldsymbol{y_{pred_i}})^2$$

- Το Μέσο Τετραγωνικό Σφάλμα είναι ο μέσος όρος:

$$\boldsymbol{MSE} = \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{y_{true_i}} - \boldsymbol{y_{pred_i}})^2$$

και εντοπίζει τα μεγαλύτερα σφάλματα στην πρόβλεψη της μεταβλητής-στόχου y λόγω ύψωσης στο τετράγωνο.

Τετραγωνική Ρίζα MSE – Root Mean Squared Error RMSE

$$\textbf{RMSE} = \sqrt{\boldsymbol{MSE}} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{y_{true_i}} - \boldsymbol{y_{pred_i}})^2}$$

Υπολογίζει το λάθος πρόβλεψης της μεταβλητής-στόχου y σε KWh ανά μήνα, ενσωματώνοντας την πληροφορία με τα μεγαλύτερα σφάλματα πρόβλεψης που έχουν εντοπιστεί με βάση τη μετρική MSE.

Με αυτόν τον τρόπο αξιολογείται ο βαθμός γενίκευσης του μοντέλου Random Forest σε νέα άγνωστα ενεργειακά δεδομένα και παράλληλα επιβεβαιώνεται η αξιοπιστία της εκτίμησης του μέτρου της σημαντικότητας των χαρακτηριστικών του συνόλου X στη διαμόρφωση της τιμής της μεταβλητής-στόχου y.

3.6.1 Feature Selection με Random Forest σε Python

Η υλοποίηση του αλγορίθμου Random Forest στα ενεργειακά δεδομένα των δημοτικών υποδομών στη γλώσσα Python πραγματοποιείται με την κλάση **RandomForestRegressor** της βιβλιοθήκης scikit-learn**,** η οποία αντιστοιχεί σε μοντέλο παλινδρόμησης πολλών δέντρων απόφασης (ensemble).

Ως μεταβλητή–στόχος y ορίστηκε το χαρακτηριστικό «Μηνιαία κατανάλωση KWh», ενώ ως ανεξάρτητες μεταβλητές εισόδου X χρησιμοποιήθηκαν τα αριθμητικά χαρακτηριστικά «Αρ.Παροχής», «Ετος κατανάλωσης», «Μήνας κατανάλωσης», «Μηνιαία αξία ενέργειας»,

«Συντ.Kwh» και τα κατηγορικά χαρακτηριστικά «Αρ. Πολλαπλου», «Ονομα Πολλαπλού», «Τιμολόγιο», «Ον.Πελάτη», «Οδός», «Αριθ.», «Αρ.Μετρητή», «Τύπος Λογ/σμου», «Κατηγορία Υποδομής» **(**Εικόνα 3.23).

```
# Ταξινόμηση με βάση την 1η ημερομηνία έναρξης της κατανάλωσης κάθε μήνα
df = df.sort_values(['1η Ημερομηνία Μήνα']).reset_index(drop=True)
df.to_csv("data_per_month_sorted.csv", index=False, encoding="utf-8-sig")
# Ορισμός στόχου y & ανεξάρτητων μεταβλητών X
target_col = 'Μηνιαία κατανάλωση KWh'
numeric_features = ['Αρ.Παροχής',
                    'Ετος κατανάλωσης',
                    'Μήνας κατανάλωσης',
                    'Μηνιαία αξία ενέργειας',
                    'Συντ.Kwh']
categorical_features = ['Αρ. Πολλαπλου',
                        'Ονομα Πολλαπλού',
                        'Τιμολόγιο',
                        'Ον.Πελάτη',
                        'Οδός',
                        'Αριθ.',
                        'Αρ.Μετρητή',
                        'Τύπος Λογ/σμου',
                        'Κατηγορία Υποδομής']

# αντίγραφο data frame και μετατροπη στόχου σε float
X = df[numeric_features + categorical_features].copy()
y = df[target_col].astype(float).round(3)
```

Εικόνα 3.23 Ορισμός μεταβλητών εισόδου X και στόχου y

Αρχικά, για τα κατηγορικά χαρακτηριστικά του συνόλου X («Αρ. Πολλαπλου», «Ονομα Πολλαπλού», «Τιμολόγιο», «Ον.Πελάτη», «Οδός», «Αριθ.», «Αρ.Μετρητή», «Τύπος Λογ/σμου», «Κατηγορία Υποδομής»), απαιτείται κατάλληλη προεπεξεργασία - μετασχηματισμός ώστε να είναι εφικτή η λειτουργία του μοντέλου RandomForestRegressor, το οποίο δέχεται ως είσοδο μόνο αριθμητικούς πίνακες.

Για τα αριθμητικά χαρακτηριστικά («Αρ.Παροχής», «Ετος κατανάλωσης», «Μήνας κατανάλωσης», «Μηνιαία αξία ενέργειας», «Συντ.Kwh»), τυπικά δεν απαιτείται μετασχηματισμός για δέντρα απόφασης καθώς οι αποφάσεις που λαμβάνονται στους κόμβους διάσπασης βασίζονται σε συγκρίσεις (thresholds) και όχι σε αποστάσεις. Παρ' όλα αυτά, στο πλαίσιο της παρούσας εργασίας εφαρμόζεται η κλάση **StandardScaler** για κλιμάκωση (standardization) των αριθμητικών χαρακτηριστικών, λόγω του διαφορετικού αριθμητικού εύρους των τιμών σε αυτά με σκοπό την υιοθέτηση μιας συνεκτικής ροής

προεπεξεργασίας η οποία μπορεί να χρησιμοποιηθεί μελλοντικά και σε άλλα μοντέλα (π.χ. γραμμικά μοντέλα ή SVM) (Hastie et al., 2009).

Η κλάση **ColumnTransformer** όπως φαίνεται στην Εικόνα 3.24, επιτρέπει την εφαρμογή διαφορετικών μεθόδων προεπεξεργασίας σε διαφορετικά υποσύνολα χαρακτηριστικών (αριθμητικά – κατηγορικά) και τον συνδυασμό των αποτελεσμάτων σε έναν ενιαίο πίνακα χαρακτηριστικών.

Στα κατηγορικά χαρακτηριστικά εφαρμόζεται κωδικοποίηση μέσω της κλάσης **OneHotEncoder**, η οποία μετατρέπει όλες τις διακριτές τιμές ενός χαρακτηριστικού (π.χ. «Κατηγορία Υποδομής») σε ισάριθμες διαφορετικές στήλες οι οποίες λαμβάνουν αριθμητικές τιμές 0/1 (one-hot encoded features)

```
# Προεπεξεργασία μετασχηματισμένων χαρακτηριστικών
preprocess = ColumnTransformer(
    transformers=[
        ('num', StandardScaler(), numeric_features),
        ('cat', OneHotEncoder(handle_unknown='ignore'), categorical_features),
    ]
)
```

Εικόνα 3.24 Μετασχηματισμός χαρακτηριστικών

Στη συνέχεια, δημιουργείται το αντικείμενο **model** του μοντέλου **RandomForestRegressor** με την κατάλληλη παραμετροποίηση συγκεκριμένων μόνο ορισμάτων (υπερπαράμετροι). Το αντικείμενο της προεπεξεργασίας (**preprocess**) που δημιουργείται από την κλάση ColumnTransformer και το αντικείμενο (**model**) του μοντέλου RandomForestRegressor ενσωματώνονται σε ένα ενιαίο αντικείμενο ροής (**pipe**) μέσω της κλάσης **Pipeline**.

Με αυτόν τον τρόπο κάθε κλήση της μεθόδου **fit()** στο εκπαιδευτικό σύνολο ή της μεθόδου **predict()** στο σύνολο ελέγχου εφαρμόζει αυτόματα τα ίδια βήματα μετασχηματισμού στα δεδομένα εισόδου, διασφαλίζοντας μια συνεκτική διαδικασία με δυνατότητα επανάληψης τόσο κατά το στάδιο της εκπαίδευσης όσο και της πρόβλεψης.

Επιπλέον, η χρήση Pipeline όπως φαίνεται στην Εικόνα 3.25, διευκολύνει την αντικατάσταση του τελικού μοντέλου με κάποιον άλλο αλγόριθμο μηχανικής μάθησης με την ίδια ή με τροποποιημένη προεπεξεργασία (π.χ. σε πειράματα σύγκρισης μοντέλων ή σε διαδικασίες *grid search*).

```
# Αντικείμενο Random Forest
model = RandomForestRegressor(
    random_state=42,
   )

# Ενιαίο αντικείμενο ροής (προεπεξεργασία-RF)
pipe = Pipeline([
    ('preprocess', preprocess),
    ('model', model)
])
```

Εικόνα 3.25 Ορισμός μοντέλου και ενιαίου αντικειμένου ροής

Για τη βελτιστοποίηση των υπερπαραμέτρων του Random Forest χρησιμοποιήθηκε η διαδικασία **grid search** σε συνδυασμό με την τεχνική της επαλήθευσης μέσω διασταύρωσης **cross-validation** που υλοποιείται από την κλάση **GridSearchCV** της βιβλιοθήκης scikit-learn. Στο πλαίσιο αυτό ορίστηκε ένα λεξικό υποψήφιων τιμών (param grid – Εικόνα 3.26) με τις βασικές υπερπαραμέτρους του μοντέλου όπως:

- αριθμός δέντρων στο δάσος (n_estimators),
- μέγιστο βάθος των δέντρων (max_depth),
- ελάχιστος αριθμός δειγμάτων για διάσπαση κόμβου (min_samples_split),
- ελάχιστος αριθμός δειγμάτων σε φύλλο (min_samples_leaf),
- μέγιστος αριθμός χαρακτηριστικών που εξετάζονται σε κάθε διάσπαση (max_features).

```
# ορισμός βασικών υπερπαραμέτρων
param_grid = {
    "model__n_estimators": [200, 300, 500],
    "model__max_depth": [None, 10, 20],
    "model__min_samples_split": [2, 5],
    "model__min_samples_leaf": [1, 2],
    "model__max_features": ["sqrt", "log2"],
}
```

Εικόνα 3.26 Ορισμός λεξικού βασικών υπερπαραμέτρων

Τα ενεργειακά δεδομένα πρώτα ταξινομούνται με βάση το χαρακτηριστικό «1η Ημερομηνία Μήνα» και στη συνέχεια διαχωρίζονται σε σύνολο εκπαίδευσης (training set) σε ποσοστό 83,785% και σε σύνολο ελέγχου (test set) σε ποσοστό 16,215% (test_size=0.16215**)** με τη συνάρτηση **train_test_split**(), διατηρώντας τη χρονική τους σειρά (shuffle=False**)** όπως φαίνεται στην Εικόνα 3.27. Η επιλογή της συγκεκριμένης αναλογίας διαχωρισμού του αρχείου έγινε ώστε τα δύο σύνολα δεδομένων να περιέχουν πλήρη ετήσια συνεχή κάλυψη.

Έτσι από τις 61.935 εγγραφές, οι 51.892 (83,785%) καλύπτουν την χρονική περίοδο 2019-2024, ενώ οι υπόλοιπες 10.043 (16,215%) καλύπτουν το 2025.

```
# Διαχωρισμός δεδομένων σε train/test
X_train, X_test, y_train, y_test = train_test_split(
    X, y,
    test_size=0.16215,
    shuffle=False   # διατήρηση χρονικής σειράς
)
```

Εικόνα 3.27 Συνάρτηση διαχωρισμού δεδομένων σε train/test

Η διαδικασία grid search της Εικόνας 3.28 εφαρμόζεται μόνο στο σύνολο εκπαίδευσης (training set), το οποίο στη συνέχεια με την τεχνική της επαλήθευσης με διασταύρωση cross-validation χωρίζεται σε 5 τμήματα (*folds)* μέσω της συνάρτησης **TimeSeriesSplit()**. Με αυτόν τον τρόπο κάθε επόμενο τμήμα (fold) χρησιμοποιεί ως σύνολο εκπαίδευσης όλα τα προηγούμενα χρονικά δεδομένα και ως σύνολο ελέγχου ένα μεταγενέστερο χρονικό τμήμα.

Έτσι προσομοιώνεται πιο ρεαλιστικά το σενάριο πρόβλεψης των μελλοντικών τιμών βάσει ιστορικών δεδομένων. Η διαδικασία του αντικειμένου grid_search (κλάση GridSearchCV) εκπαιδεύει διαδοχικά το αντικείμενο pipe για κάθε συνδυασμό υπερπαραμέτρων (param_grid) και για κάθε τμήμα (*fold*) του cross-validation (αντικείμενο *tscv*) υπολογίζει το μέσο όρο του συντελεστή R^2 στα σύνολα ελέγχου και τελικά επιλέγει τον συνδυασμό που μεγιστοποιεί την απόδοση.

```
# διαδικασία grid search
grid_search = GridSearchCV(
    estimator=pipe,
    param_grid=param_grid,
    cv=tscv,
    scoring="r2",
    n_jobs=-1,
    verbose=2,
)

print("---Εκτέλεση GridSearchCV στο training set--- ")
# εκπαίδευση
grid_search.fit(X_train, y_train)
```

Εικόνα 3.28 Διαδικασία εύρεσης βέλτιστων υπερπαραμέτρων

Το τελικό εκπαιδευμένο μοντέλο (best_estimator_) με τις βέλτιστες υπερπαραμέτρους (Εικόνα 3.29) χρησιμοποιήθηκε για τα επόμενα βήματα της τελικής αξιολόγησης και ανάλυσης της σημαντικότητας των χαρακτηριστικών του συνόλου *X* στη διαμόρφωση της μεταβλητής-στόχου y.

```
# ορισμός βέλτιστων υπέρ-παραμέτρων μοντέλου (pipe)
pipe = grid_search.best_estimator_

# Πρόβλεψη μεταβλητής στόχου στο σύνολο ελέγχου
y_pred = pipe.predict(X_test)
```

Εικόνα 3.29 Ορισμός βέλτιστων υπερπαραμέτρων και πρόβλεψη στόχου

Η απόδοση του μοντέλου αξιολογείται στο σύνολο ελέγχου (test set) με τον προσδιορισμό:

- του συντελεστή R2 (μέθοδος score())
- του Μέσου Απόλυτου Σφάλματος – MAE (συνάρτηση mean_absolute_error())
- του Μέσου Τετραγωνικού Σφάλματος – MSE (συνάρτηση mean_squared_error())
- της Τετραγωνικής Ρίζας (Root Mean Error – RMSE) του Μέσου Τετραγωνικού Σφάλματος MSE

όπως φαίνεται στην Εικόνα 3.30.

```
# Μετρικές απόδοσης στο σύνολο ελέγχου
r2 = pipe.score(X_test, y_test)
mae = mean_absolute_error(y_test, y_pred)
mse = mean_squared_error(y_test, y_pred)
rmse = np.sqrt(mse)
```

Εικόνα 3.30 Μετρικές απόδοσης

Για την αξιολόγηση της αξιοπιστίας της απόδοσης και της σταθερότητας του μοντέλου μετά τον προσδιορισμό των βέλτιστων υπερπαραμέτρων χρησιμοποιείται εκ νέου η τεχνική της επαλήθευσης με διασταύρωση (*cross-validation*) μόνο στο αρχικό σύνολο εκπαίδευσης (training set), με δύο διαφορετικά αντικείμενα που προκύπτουν από τις κλάσεις: **TimeSeriesSplit()** και **K-Fold()** (Εικόνα 3.31)**.**

```
# χρονικός διαχωρισμός δεδομένων
tscv = TimeSeriesSplit(n_splits=5)

# τυχαίος διαχωρισμός δεδομένων
cv = KFold(n_splits=5,shuffle=True,random_state=42)
```

Εικόνα 3.31 Χρονικός & Τυχαίος διαχωρισμός δεδομένων

Η τεχνική αυτή χωρίζει το αρχικό σύνολο εκπαίδευσης του αρχείου των ενεργειακών δεδομένων σε k τμήματα (*folds*) τα οποία προκύπτουν:

- είτε τυχαία μέσω της κλάσης KFold()
- είτε χρονικά, διατηρώντας αυστηρά τη χρονική σειρά των παρατηρήσεων μέσω της κλάσης TimeSeriesSplit().

Η συνάρτηση **cross_validate** της βιβλιοθήκης scikit-learn (Εικόνα 3.32) εκτελείται *k* φορές και σε κάθε επανάληψη επιλέγει ένα τμήμα ως σύνολο ελέγχου (test set), ενώ τα υπόλοιπα *k-1* τμήματα επιλέγονται ως σύνολο εκπαίδευσης (training set). Στο τέλος προκύπτουν για κάθε μετρική αξιολόγησης (π.χ. R^2, MSE) *k* διαφορετικές τιμές από τις οποίες υπολογίζεται ο μέσος όρος και η τυπική απόκλιση (Hastie et al., 2009).

Με την μέθοδο αυτή όλα τα στιγμιότυπα των παρατηρήσεων (του αρχικού συνόλου εκπαίδευσης) των ενεργειακών δεδομένων χρησιμοποιούνται - σε διαφορετικά τμήματα (*folds*) - τόσο ως σύνολο εκπαίδευσης όσο και ως σύνολο ελέγχου (Witten et al., 2016).

```
# Μετρικές απόδοσης για KFold και TimeSeriesSplit
scoring = {
    "r2": "r2",
    "mae": "neg_mean_absolute_error",
    "mse": "neg_mean_squared_error",
}

# cross-validation με TimeSeriesSplit στο training set
cv_results = cross_validate(
    pipe,X_train,y_train,
    cv=tscv,
    scoring=scoring,
    return_train_score=False
)
# Υπολογισμός μέσων τιμών & τυπικών αποκλίσεων TimeSeriesSplit
r2_scores   = cv_results["test_r2"]
mae_scores  = -cv_results["test_mae"]
mse_scores  = -cv_results["test_mse"]
rmse_scores = np.sqrt(mse_scores)
# Αποθήκευση μέσων τιμών TimeSeriesSplit
r2_tscv_mean   = r2_scores.mean()
mae_tscv_mean  = mae_scores.mean()
rmse_tscv_mean = rmse_scores.mean()
```

```
# cross-validation με K-Fold στο training set
cv_results = cross_validate(
    pipe,X_train,y_train,
    cv=cv,
    scoring=scoring,
    return_train_score=False
)
# Υπολογισμός μέσων τιμών & τυπικών αποκλίσεων KFold
r2_scores  = cv_results["test_r2"]
mae_scores = -cv_results["test_mae"]
mse_scores = -cv_results["test_mse"]
rmse_scores = np.sqrt(mse_scores)
# Αποθήκευση μέσων τιμών KFold
r2_kfold_mean   = r2_scores.mean()
mae_kfold_mean  = mae_scores.mean()
rmse_kfold_mean = rmse_scores.mean()
```

Εικόνα 3.32 Μετρικές απόδοσης με cross-validation στο training set

Αυτό επιτρέπει να εκτιμηθεί και να συγκριθεί η απόδοση του μοντέλου με πιο αξιόπιστο και σταθερό τρόπο σε σχέση με την εκτίμηση που προκύπτει από έναν μόνο διαχωρισμό των δεδομένων σε σύνολο εκπαίδευσης και ελέγχου.

Προκειμένου να εντοπιστούν και να επιλεγούν μόνο εκείνα τα χαρακτηριστικά που συμβάλλουν περισσότερο στην πρόβλεψη της μεταβλητής στόχου «Μηνιαία κατανάλωση KWh», εφαρμόζεται η ενσωματωμένη ιδιότητα του Random Forest feature_importances_ και η συνάρτηση permutation importance.

Η ιδιότητα **feature_importances_** αποτελεί μια εσωτερική μετρική για την εκτίμηση της σημαντικότητας κάθε μετασχηματισμένου χαρακτηριστικού του προβλεπτικού μοντέλου Random Forest. Η μετρική αυτή βασίζεται στην μείωση του σφάλματος (MSE) που προκαλεί ένα χαρακτηριστικό κάθε φορά που χρησιμοποιείται ως κριτήριο διάσπασης στους εσωτερικούς κόμβους των δέντρων. Η συνολική εκτίμηση της σημαντικότητας ενός χαρακτηριστικού προκύπτει από το άθροισμα των μειώσεων των σφαλμάτων σε όλα τα δέντρα του δάσους και την κανονικοποίηση των τιμών ώστε το συνολικό άθροισμα να ισούται με τη μονάδα.

```python
# Σημαντικότητα χαρακτηριστικών με RF feature_importances_
model_fitted = pipe.named_steps["model"]
preprocess_fitted = pipe.named_steps["preprocess"]
feature_names_transformed = preprocess_fitted.get_feature_names_out()
# βοηθητική λίστα αναδόμησης αρχικών feature
feature_names = []
# αφαίρεση 'num' και 'cat' απο τα μετασχηματισμένα δεδομένα
for name in feature_names_transformed:
    if name.startswith("num__"):
        # "num__Ετος κατανάλωσης" -> "Ετος κατανάλωσης"
        orig = name.split("num__")[1]
    elif name.startswith("cat__"):
        # παράδειγμα: "cat__Κατηγορία Υποδομής_ΠΟΛΙΤΙΣΜΟΣ" -> "Κατηγορία Υποδομής"
        after_prefix = name.split("cat__")[1]
        orig = after_prefix.split("_", 1)[0]
    else:
        # αν δεν υπάρχει prefix (num,cat)
        orig = name
    feature_names.append(orig)
# Δημιουργία Series με importance's και ομαδοποίηση ανά αρχικό feature
fi_orig = (
    pd.Series(model_fitted.feature_importances_, index=feature_names)
      .groupby(level=0).sum()      # άθροισμα όλων των dummies του ίδιου feature
      .sort_values(ascending=False)
)
```

Εικόνα 3.33 Random Forest feature_importances_

Η συνάρτηση **permutation importance** χρησιμοποιεί το ήδη εκπαιδευμένο μοντέλο (pipe) και, ως σύνολο αναφοράς το σύνολο ελέγχου (test set). Για κάθε χαρακτηριστικό X_j προκαλεί τυχαία αναδιάταξη (*permutation*) των τιμών του μεταξύ των εγγραφών, διατηρώντας όμως την αρχική διάταξη των υπολοίπων χαρακτηριστικών. Η αναδιάταξη αυτή διαταράσσει την αρχική σχέση εξάρτησης που έχει «μάθει» το μοντέλο μεταξύ του συγκεκριμένου χαρακτηριστικού και της εξαρτημένης μεταβλητής y κατά το στάδιο της εκπαίδευσης.

```
# Συνάρτηση permutation importance με R²
result = permutation_importance(
    pipe,X_test,y_test,
    scoring="r2",
    n_repeats=10,
    random_state=42,
    n_jobs=-1,
)
# Σημαντικότητα χαρακτηριστικών με permutation importance R²
pi = (pd.Series(result.importances_mean,index=X_test.columns)
                .sort_values(ascending=False))
```

Εικόνα 3.34 Permutation Importance

Για κάθε χαρακτηριστικό υπολογίζεται η διαφορά ανάμεσα στην αρχική (baseline) τιμή ενός δείκτη απόδοσης (στην περίπτωσή μας του R^2 στο test set) και στη μέση τιμή του ίδιου δείκτη μετά από επαναλαμβανόμενες τυχαίες αναδιατάξεις των τιμών του χαρακτηριστικού. Η διαφορά αυτή (baseline R^2 – permuted R^2), κατά μέσο όρο, αποτελεί το μέτρο σημαντικότητας του χαρακτηριστικού, η οποία όσο μεγαλύτερη είναι τόσο μεγαλύτερη είναι η συμβολή του συγκεκριμένου χαρακτηριστικού στην πρόβλεψη της μεταβλητής y = «Μηνιαία κατανάλωση KWh» (Molnar, 2020).

3.6.2 Αποτελέσματα Random Forest για την επιλογή χαρακτηριστικών

Το μοντέλο εκπαιδεύτηκε συνολικά 360 φορές (72 συνδυασμοί υπερπαραμέτρων × 5 folds). Η συνολική χρονική διάρκεια της διαδικασίας GridSearchCV ανήλθε σε 8.189 seconds ≈ 2,27 ώρες όπως φαίνεται στην Εικόνα 3.35.

Βέλτιστες υπέρ-παράμετροι μοντέλου Random Forest στο training set

```
---Εκτέλεση GridSearchCV στο training set---
Fitting 5 folds for each of 72 candidates, totalling 360 fits

Καλύτερες υπερπαράμετροι (GridSearchCV):
{'model__max_depth': None,
'model__max_features': 'sqrt',
'model__min_samples_leaf': 1,
'model__min_samples_split': 2,
'model__n_estimators': 200}

Καλύτερη μέση τιμή συντελεστή R² από GridSearchCV: 0.7504

Χρόνος εκτέλεσης grid search: 8189.675 seconds
```

Εικόνα 3.35 Βέλτιστες υπερπαράμετροι Random Forest

Η διαδικασία βελτιστοποίησης υπερπαραμέτρων (GridSearchCV) σε συνδυασμό με τη χρήση της κλάσης TimeSeriesSplit(n_splits=5) ώστε να διατηρείται η χρονική σειρά κατά τον διαχωρισμό των παρατηρήσεων του εκπαιδευτικού συνόλου, ανέδειξε ως βέλτιστο σύνολο υπερπαραμέτρων για το προβλεπτικό μοντέλο Random Forest τις τιμές: n_estimators = 200, max_depth = None, max_features = 'sqrt', min_samples_split = 2 και min_samples_leaf = 1.

Ο συνδυασμός αυτός οδήγησε σε μέση επίδοση του συντελεστή προσδιορισμού **R^2 = 0.7504**, γεγονός που υποδηλώνει ότι το μοντέλο Random Forest εξηγεί περίπου το 75% της διακύμανσης της μηνιαίας κατανάλωσης ηλεκτρικής ενέργειας.

Το βέλτιστο πλήθος των δέντρων του δάσους προσδιορίστηκε σε n_estimators = 200, επιλογή που ενισχύει τη σταθερότητα των προβλέψεων, διατηρώντας τις απαιτήσεις σε υπολογιστικούς πόρους σχετικά χαμηλά σε περίπτωση επιλογής περισσότερων δέντρων. Παράλληλα, το μέγιστο βάθος των δέντρων δεν περιορίζεται (max_depth = None), επιτρέποντας στο μοντέλο να αποτυπώνει πολύπλοκες και μη γραμμικές σχέσεις μεταξύ των ανεξάρτητων μεταβλητών X και της εξαρτημένης μεταβλητής-στόχου y. Αυτό μπορεί να αυξήσει τον κίνδυνο υπερπροσαρμογής σε επίπεδο μεμονωμένων δέντρων, ωστόσο μετριάζεται από τον μηχανισμό συνάθροισης του Random Forest που προκύπτει από τον μέσο όρο των προβλέψεων όλων των δέντρων.

Η ρύθμιση max_features = 'sqrt' σημαίνει ότι σε κάθε κόμβο διάσπασης επιλέγεται ένα τυχαίο υποσύνολο χαρακτηριστικών μεγέθους $\sqrt{p}$ (όπου p το πλήθος χαρακτηριστικών), μειώνοντας τη συσχέτιση μεταξύ των δέντρων και βελτιώνοντας τη γενίκευση του μοντέλου. Επιπλέον, απαιτούνται τουλάχιστον δύο παρατηρήσεις σε κάθε κόμβο για να πραγματοποιηθεί διάσπαση (min_samples_split = 2), ενώ κάθε φύλλο του δέντρου μπορεί να αντιστοιχεί ακόμη και σε μία μόνο παρατήρηση (min_samples_leaf = 1). Στο επίπεδο

της πρόβλεψης κάθε φύλλο αποδίδει μια τιμή για τη μεταβλητή-στόχο y=«Μηνιαία κατανάλωση kWh».

Αξιολόγηση μοντέλου στο σύνολο ελέγχου (test set)

Η απόδοση του μοντέλου αξιολογήθηκε στο σύνολο ελέγχου (test set) που δεν χρησιμοποιήθηκε σε κανένα στάδιο της επιλογής των υπερπαραμέτρων (Εικόνα 3.36).

```
Αξιολόγηση απόδοσης μοντέλου στο σύνολο ελέγχου (test set)
============================================================
R²   : 0.7397
MAE  : 327.9529
MSE  : 9102051.7184
RMSE : 3016.9607
============================================================
```

Εικόνα 3.36 Αξιολόγηση μοντέλου στο test set

Η μέση τιμή για τον συντελεστή προσδιορισμού ανήλθε σε $R^2 = 0.7397$. Η τιμή αυτή βρίσκεται πολύ κοντά στην τιμή που προέκυψε από την διαδικασία grid search για την εύρεση των βέλτιστων υπερπαραμέτρων με την τεχνική της επαλήθευσης μέσω διασταύρωσης (cross-validation) εφαρμόζοντας την μέθοδο TimeSeriesSplit. Το Μέσο Απόλυτο Σφάλμα – MAE δείχνει ότι το μέσο σφάλμα στην πρόβλεψη της μεταβλητής-στόχου y= «Μηνιαία κατανάλωση KWh» είναι περίπου 328 KWh το μήνα. Αντίστοιχα, η ρίζα του μέσου τετραγωνικού σφάλματος (RMSE), η οποία δίνει μεγαλύτερη βαρύτητα στα μεγάλα σφάλματα πρόβλεψης που μπορεί να οφείλονται ενδεχομένως σε ακραίες τιμές (outliers) ή στην εποχικότητα των δεδομένων, δείχνει ότι το τυπικό μέγεθος του σφάλματος πρόβλεψης ανέρχεται περίπου σε 3017 KWh ανά μήνα.

Απόδοση μοντέλου με TimeSeriesSplit – cross validation στο training set

Η εφαρμογή TimeSeriesSplit cross validation στο training set, αποδίδει μέση τιμή για τον συντελεστή προσδιορισμού $R^2 = 0.7504$ με απόκλιση ± 0.0823. Αυτή είναι η καλύτερη μέση τιμή που προέκυψε κατά τη διαδικασία εύρεσης των βέλτιστων υπερπαραμέτρων μέσω GridSearchCV λόγω της ρύθμισης cv=tscv, γεγονός που ενισχύει την αξιοπιστία της απόδοσης.

Η τήρηση της χρονικής σειράς των παρατηρήσεων σημαίνει ότι το μοντέλο για την εύρεση των βέλτιστων υπερπαραμέτρων μέσω GridSearchCV εκπαιδεύεται σε δεδομένα που προηγούνται χρονικά από τα δεδομένα στα οποία ελέγχεται η απόδοσή του.

```
Αναλυτικά αποτελέσματα Cross-validation στο training set με TimeSeriesSplit (5-splits)
================================================================================
Τμήμα 1: R² = 0.8454,  MAE = 342.2916,  RMSE = 1794.0152
Τμήμα 2: R² = 0.8385,  MAE = 361.7032,  RMSE = 2142.5624
Τμήμα 3: R² = 0.6386,  MAE = 792.0486,  RMSE = 2829.9256
Τμήμα 4: R² = 0.7466,  MAE = 350.3948,  RMSE = 2352.7872
Τμήμα 5: R² = 0.6829,  MAE = 492.6163,  RMSE = 3019.9023

Μέσες τιμές & τυπικές αποκλίσεις:
R² = 0.7504, std = 0.0823
MAE = 467.81, std = 171.20
RMSE = 2427.84, std = 447.43
================================================================================
```

Εικόνα 3.37 Cross-validation με TimeSeriesSplit

Παρατηρώντας τις αποδόσεις στα επιμέρους τμήματα (Εικόνα 3.37) διαπιστώνουμε ότι οι τιμές του συντελεστή R^2 κυμαίνονται από περίπου 0.64 έως 0.85 γεγονός που υποδηλώνει σαφή μεταβλητότητα στη διακύμανση της μηνιαίας κατανάλωσης ενέργειας σε KWh. Το Μέσο Απόλυτο Σφάλμα – MAE δείχνει ότι το μέσο λάθος στην πρόβλεψη της μεταβλητής-στόχου y= «Μηνιαία κατανάλωση KWh» είναι περίπου 468 KWh το μήνα με απόκλιση περίπου ± 171 KWh. Αντίστοιχα το RMSE που ενσωματώνει μεγαλύτερα σφάλματα πρόβλεψης που μπορεί να οφείλονται είτε σε ακραίες τιμές (outliers) ή στην εποχικότητα των δεδομένων η οποία ήδη διαπιστώνεται στα επιμέρους τμήματα (folds), δείχνει ότι το λάθος πρόβλεψης είναι περίπου 2428 KWh ανά μήνα με απόκλιση περίπου ± 447 KWh.

Απόδοση μοντέλου με Kfold – cross validation στο training set

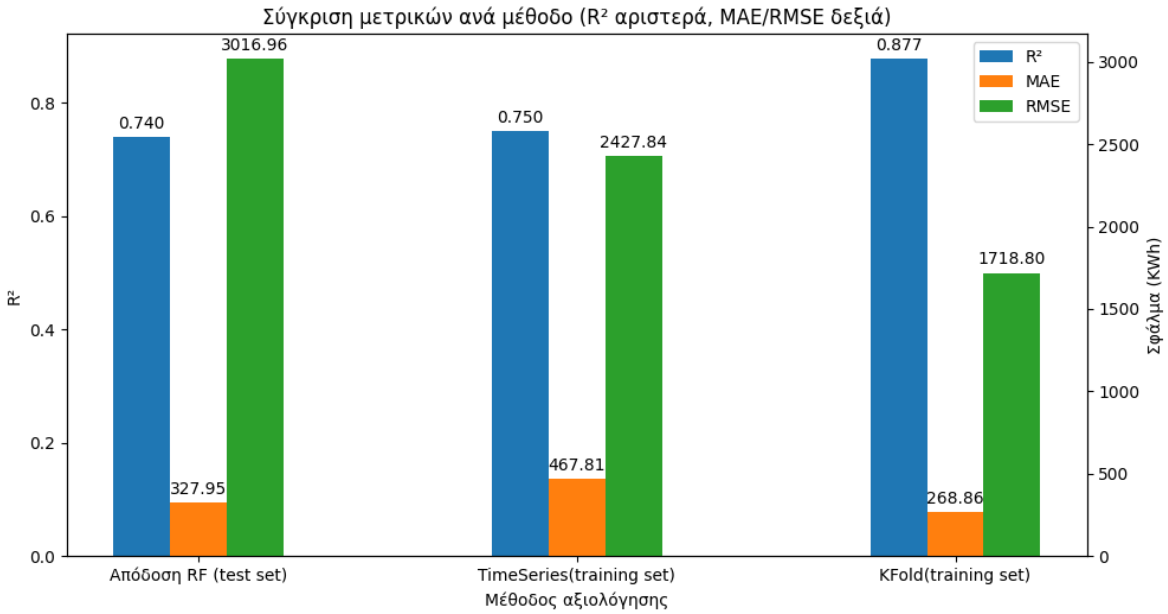
Η εφαρμογή της μεθόδου Kfold – cross validation οδηγεί σε καλύτερη απόδοση με μέση τιμή συντελεστή προσδιορισμού $R^2 = 0.8771$, γεγονός που υποδηλώνει ότι το μοντέλο εξηγεί περίπου το 88,% της διακύμανσης της μεταβλητής-στόχου. Η βελτίωση αυτή δεν θεωρείται πλήρως αξιόπιστη καθώς η μέθοδος K-Fold δεν λαμβάνει υπόψη τη χρονική σειρά των παρατηρήσεων. Κατά το στάδιο της εκπαίδευσης είναι πιθανό να χρησιμοποιεί μεταγενέστερες χρονικά παρατηρήσεις οι οποίες «αποκαλύπτουν» έμμεσα πληροφορία που στην πράξη δεν θα ήταν διαθέσιμη κατά τον χρόνο της πρόβλεψης (μορφή διαρροής πληροφορίας – data leakage). Το Μέσο Απόλυτο Σφάλμα – MAE δείχνει ότι το μέσο σφάλμα στην πρόβλεψη της μεταβλητής-στόχου y= «Μηνιαία κατανάλωση KWh» ανέρχεται περίπου σε 269 KWh ανά μήνα με τυπική απόκλιση περίπου ± 4.5 KWh μεταξύ των διαφορετικών τμημάτων διαχωρισμού (folds). Αντίστοιχα η ρίζα του μέσου τετραγωνικού σφάλματος (RMSE) που ενσωματώνει μεγαλύτερα σφάλματα πρόβλεψης που μπορεί να οφείλονται είτε σε ακραίες τιμές (outliers) ή στην εποχικότητα των

δεδομένων, εκτιμάται σε περίπου 1719 KWh ανά μήνα με απόκλιση περίπου ± 99 KWh (Εικόνα 3.38)

```
Αποτελέσματα Cross-validation στο training set με K-Fold (5-fold)
========================================================================
R² = 0.8771, std = 0.0155
MAE = 268.8645, std = 4.5443
RMSE = 1718.8041, std = 98.7560
========================================================================
```

Εικόνα 3.38 Cross-validation με K-Fold

Τα αποτελέσματα αυτά (Εικόνα 3.39) υποδεικνύουν πολύ καλή προσαρμογή του μοντέλου στο συγκεκριμένο σχήμα επικύρωσης, αλλά ταυτόχρονα ενισχύουν την ανάγκη χρήσης μεθόδων που τηρούν τη χρονική δομή των δεδομένων (π.χ. TimeSeriesSplit) για μια πιο ρεαλιστική εκτίμηση της ικανότητας γενίκευσης του μοντέλου σε νέα άγνωστα δεδομένα.



Εικόνα 3.39 Οπτικοποίηση αποτελεσμάτων GridSearchCV, TimeSeriesSplit, KFold

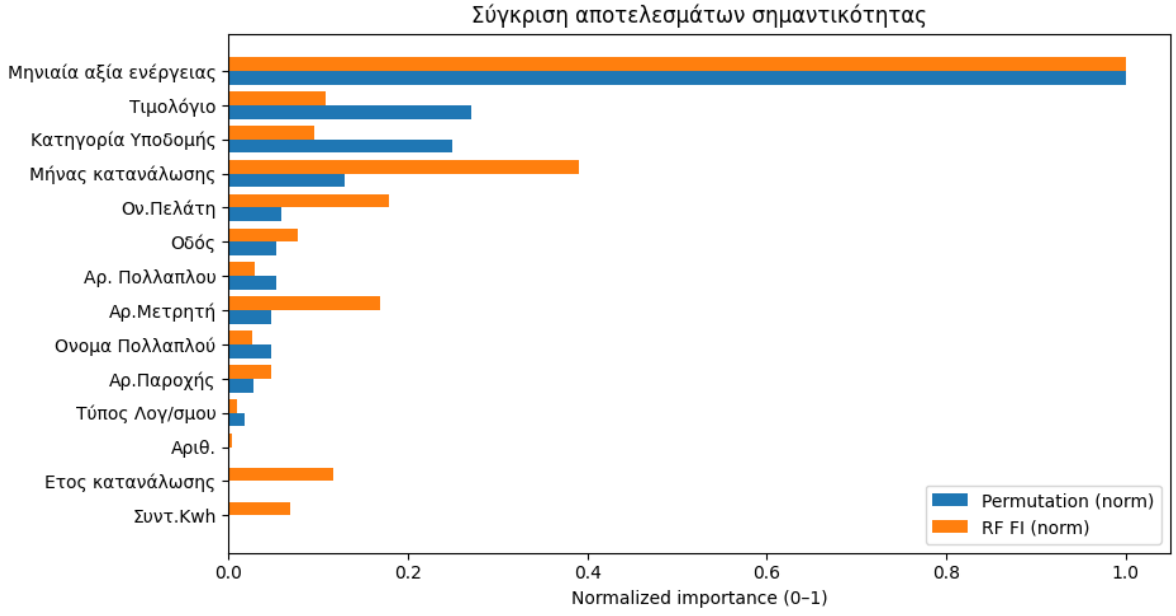
Σημαντικότητα χαρακτηριστικών

Η συνάρτηση permutation_importance εφαρμόστηκε στο εκπαιδευμένο μοντέλο pipe, το οποίο είχε ρυθμιστεί με τις βέλτιστες υπερπαραμέτρους του αλγορίθμου Random Forest. Η εκπαίδευση πραγματοποιήθηκε στο training set και η αξιολόγηση της απόδοσης στο test set, τα οποία προέκυψαν από τον διαχωρισμό των ενεργειακών δεδομένων σε ποσοστό 83,785% ως σύνολο εκπαίδευσης (training set) και σε ποσοστό 16,215% ως σύνολο ελέγχου (test set) με τη χρήση της συνάρτησης train_test_split. Επισημαίνεται ότι πριν τον διαχωρισμό τα δεδομένα ταξινομήθηκαν χρονικά με βάση το χαρακτηριστικό «1η

Ημερομηνία Μήνα». Η ρύθμιση της παραμέτρου shuffle=False αποκλείει τον τυχαίο διαχωρισμό των παρατηρήσεων με αποτέλεσμα τα δεδομένα του εκπαιδευτικού συνόλου (training set) να περιέχουν πληροφορίες χρονικά προγενέστερες από τα δεδομένα του συνόλου ελέγχου (test set) όπου αξιολογείται η ικανότητα γενίκευσης (πρόβλεψης) του μοντέλου.

Σύγκριση σημαντικότητας (Permutation_Importance - RF feature_importances_)		
Μεταβλητή Εισόδου X	**Permutation_Importance**	**RF_feature_importances**
Μηνιαία αξία ενέργειας	0.565346	0.430553
Τιμολόγιο	0.152662	0.046691
Κατηγορία Υποδομής	0.140788	0.041453
Μήνας κατανάλωσης	0.073509	0.167936
Ον.Πελάτη	0.033198	0.076997
Οδός	0.030396	0.033131
Αρ. Πολλαπλου	0.030218	0.012396
Αρ.Μετρητή	0.027051	0.072865
Ονομα Πολλαπλού	0.026652	0.011671
Αρ.Παροχής	0.015391	0.020342
Τύπος Λογ/σμου	0.009946	0.004258
Αριθ.	0.000897	0.001547
Ετος κατανάλωσης	0.000000	0.050326
Συντ.Kwh	0.000000	0.029832

Πίνακας 3.4 Βαθμός σημαντικότητας χαρακτηριστικών



Εικόνα 3.40 Οπτικοποίηση σημαντικότητας χαρακτηριστικών

Τόσο η εσωτερική μετρική feature_importances_ του Random Forest, όσο και η μέθοδος Permutation Importance ανέδειξαν ως πιο σημαντικό χαρακτηριστικό την **«Μηνιαία αξία ενέργειας»** (Πίνακας 3.4 & Εικόνα 3.40)**,** γεγονός που υποδηλώνει ότι η συγκεκριμένη μεταβλητή συμβάλλει καθοριστικά στη μείωση του σφάλματος πρόβλεψης.

Παράλληλα, παρατηρείται ότι ορισμένα χαρακτηριστικά εμφανίζουν σχετικά υψηλή σημαντικότητα με τη feature_importances_ αλλά μηδενική μέσω Permutation Importance. Αυτό συμβαίνει διότι η εσωτερική μετρική feature_importances_ του Random Forest σε δένδρα απόφασης ευνοεί χαρακτηριστικά με πολλές μοναδικές τιμές, τα οποία επιτρέπουν την εφαρμογή πολλών κανόνων διάσπασης στους κόμβους των δέντρων με αποτέλεσμα τη μείωση του σφάλματος και την αύξηση της σημαντικότητα τους. Αυτό όμως δεν συνεπάγεται και αντίστοιχη βελτίωση στη γενίκευση της πρόβλεψης. Αυτό ακριβώς αποτυπώνει η μέθοδος Permutation Importance όπου όταν οι τιμές ενός χαρακτηριστικού αναδιατάσσονται τυχαία στο σύνολο ελέγχου και δεν μεταβάλλεται η απόδοση του μοντέλου, τότε το χαρακτηριστικό εμφανίζει πολύ χαμηλή ή μηδενική σημαντικότητα.

Συμπέρασμα

Η μεταβλητή **«Μηνιαία αξία ενέργειας»** εξαρτάται άμεσα από την μεταβλητή **«Μηνιαία κατανάλωση KWh»** καθώς υπολογίζεται με βάση αυτήν. Το συγκεκριμένο χαρακτηριστικό όταν συμπεριλαμβάνεται στο σύνολο των ανεξάρτητων μεταβλητών X εμφανίζει σχεδόν γραμμική σχέση με την μεταβλητή-στόχο. Το γεγονός αυτό οδηγεί σε τεχνητή βελτίωση της εκτιμώμενης απόδοσης του μοντέλου και σε λάθος εκτίμηση της σημαντικότητας του χαρακτηριστικού καθώς αποκτά πρόσβαση σε πληροφορία η οποία στην πράξη δεν θα είναι διαθέσιμη τη στιγμή της πρόβλεψης. Το φαινόμενο αυτό είναι γνωστό ως *target leakage* (διαρροή πληροφορίας) (Kapoor & Narayanan, 2023).

Η πρόβλεψη της μεταβλητής-στόχου «Μηνιαία κατανάλωση KWh» πρέπει να βασίζεται αποκλειστικά σε ανεξάρτητες μεταβλητές των οποίων οι τιμές είναι ήδη γνωστές κατά τη χρονική στιγμή της πρόβλεψης, χωρίς να ενσωματώνουν άμεσα ή έμμεσα πληροφορία που διαρρέει από τον στόχο.

Για την εξάλειψη του φαινομένου target leakage η ανάλυση της σημαντικότητας των χαρακτηριστικών επαναλήφθηκε εξαιρώντας το χαρακτηριστικό «Μηνιαία αξία ενέργειας», ώστε να προκύψει μια πιο ρεαλιστική εκτίμηση της συμβολής των υπόλοιπων χαρακτηριστικών στην πρόβλεψη της μεταβλητής-στόχου «Μηνιαία κατανάλωση KWh».

Το μοντέλο εκπαιδεύτηκε ξανά 360 φορές (72 συνδυασμοί υπερπαραμέτρων × 5 folds). Η συνολική χρονική διάρκεια της διαδικασίας GridSearchCV ανήλθε σε 9.308 seconds ≈ 2,58 ώρες. Οι βέλτιστες υπερπαράμετροι που προέκυψαν για το μοντέλο καθώς και η αξιολόγηση του στο σύνολο ελέγχου παρουσιάζονται στην Εικόνα 3.41.

```
Καλύτερες υπερπαράμετροι (GridSearchCV):
{'model__max_depth': None,
'model__max_features': 'log2',
'model__min_samples_leaf': 2,
'model__min_samples_split': 2,
'model__n_estimators': 300}

Καλύτερη μέση τιμή συντελεστή R² από GridSearchCV: 0.4957

Χρόνος εκτέλεσης grid search: 9308.458 seconds

Αξιολόγηση απόδοσης μοντέλου στο σύνολο ελέγχου (test set)
============================================================
R²   : 0.4014
MAE  : 696.5205
MSE  : 20928261.0320
RMSE : 4574.7416
============================================================


Αναλυτικά αποτελέσματα Cross-validation στο training set με TimeSeriesSplit (5-splits)
================================================================================
Τμήμα 1: R² = 0.4262,  MAE = 811.9103,  RMSE = 3456.5708
Τμήμα 2: R² = 0.5049,  MAE = 752.8299,  RMSE = 3750.9184
Τμήμα 3: R² = 0.6960,  MAE = 552.5489,  RMSE = 2595.7588
Τμήμα 4: R² = 0.4980,  MAE = 600.6415,  RMSE = 3311.1909
Τμήμα 5: R² = 0.3537,  MAE = 856.7367,  RMSE = 4311.2492

Μέσες τιμές & τυπικές αποκλίσεις:
R² = 0.4957, std = 0.1142
MAE = 714.93, std = 118.64
RMSE = 3485.14, std = 561.28
================================================================================


Αποτελέσματα Cross-validation στο training set με K-Fold (5-fold)
==================================================================
R² = 0.6245, std = 0.0143
MAE = 589.9591, std = 19.4529
RMSE = 3013.6942, std = 148.3335
==================================================================
```

Εικόνα 3.41 Αποτελέσματα GridSearchCV, TimeSeriesSplit, KFold without leakage

Ανάλυση αποτελεσμάτων:

Η διαδικασία βελτιστοποίησης υπερπαραμέτρων (GridSearchCV) ανέδειξε ως βέλτιστο σύνολο υπερπαραμέτρων μετά την αφαίρεση του χαρακτηριστικού «Μηνιαία αξία

ενέργειας» που προκαλούσε τη διαρροή πληροφορίας (leakage) τις τιμές: n_estimators = 300, max_depth = None, max_features = 'log2', min_samples_split = 2 και min_samples_leaf = 2. Ο συνδυασμός αυτός οδήγησε σε μέση επίδοση του συντελεστή προσδιορισμού $\mathbf{R^2 = 0.4957}$, γεγονός που υποδηλώνει ότι το μοντέλο Random Forest εξηγεί στο συγκεκριμένο σύνολο δεδομένων περίπου το 50% της διακύμανσης της μηνιαίας κατανάλωσης ηλεκτρικής ενέργειας χωρίς το φαινόμενο leakage.

Το βέλτιστο πλήθος των δέντρων του δάσους αυξήθηκε σε n_estimators = 300, παρέχοντας σταθερότητα προβλέψεων με αύξηση του υπολογιστικού κόστους σε σχέση με την προηγούμενη λύση (200 δέντρα). Παράλληλα, το μέγιστο βάθος των δέντρων συνεχίζει να μην περιορίζεται (max_depth = None) γεγονός που δείχνει ότι το μοντέλο εξακολουθεί να αξιοποιεί τη δυνατότητα αποτύπωσης μη γραμμικών και σύνθετων σχέσεων μεταξύ των ανεξάρτητων μεταβλητών X και της εξαρτημένης μεταβλητής-στόχου y.

Ωστόσο, η αλλαγή στις υπόλοιπες υπερπαραμέτρους υποδηλώνει τον περιορισμό της υπερπροσαρμογής. Ειδικότερα, η ρύθμιση max_features = 'log2' σημαίνει ότι σε κάθε κόμβο διάσπασης εξετάζεται ένα τυχαίο υποσύνολο χαρακτηριστικών μεγέθους $\log_2(p)$ το οποίο είναι μικρότερο από το $\sqrt{p}$, μειώνοντας τη συσχέτιση μεταξύ των δέντρων και βελτιώνοντας τη γενίκευση του μοντέλου.

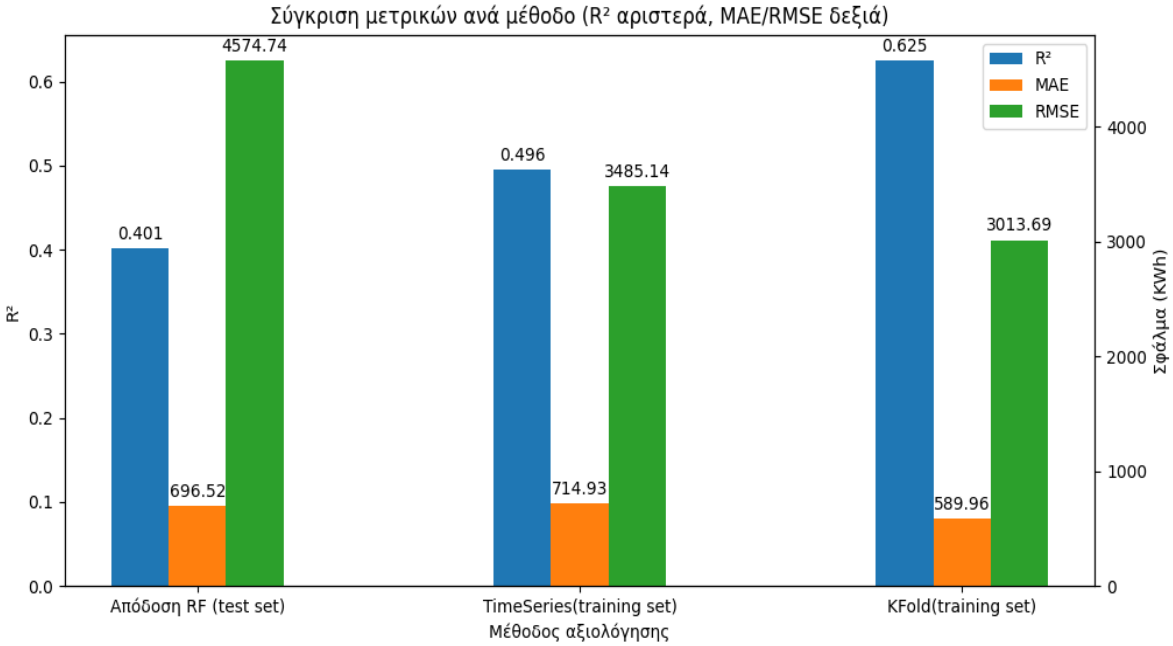
Παράλληλα, η αύξηση του min_samples_leaf από 1 σε 2 λειτουργεί ως μορφή κανονικοποίησης καθώς περιορίζει την ανάπτυξη φύλλων που βασίζονται σε πολύ λίγες παρατηρήσεις και συνεπώς μειώνει την πολυπλοκότητα του μοντέλου. Όταν min_samples_leaf = 1 επιτρέπονται φύλλα που αντιστοιχούν σε μία μόνο παρατήρηση, κάτι που αυξάνει την ευελιξία του μοντέλου αλλά και τον κίνδυνο υπερπροσαρμογής. Αντίθετα, με min_samples_leaf = 2 επιβάλλεται κάθε τελικό φύλλο να στηρίζεται σε τουλάχιστον δύο παρατηρήσεις, περιορίζοντας την υπερβολική κατάτμηση του χώρου των χαρακτηριστικών και βελτιώνοντας τη γενίκευση σε νέα δεδομένα.

Η παράμετρος min_samples_split = 2 παραμένει αμετάβλητη, υποδεικνύοντας ότι το μοντέλο εξακολουθεί να επιτρέπει διασπάσεις με τον ελάχιστο δυνατό αριθμό παρατηρήσεων σε έναν κόμβο του δέντρου απόφασης.

Παράλληλα, καταγράφεται μείωση της απόδοσης (R^2/MAE/RMSE), τόσο κατά τη διαδικασία αναζήτησης των βέλτιστων υπερπαραμέτρων GridSearchCV με την τεχνική cross-validation (TimeSeriesSplit) στο training set όσο και κατά την τελική αξιολόγηση του

μοντέλου στο σύνολο ελέγχου (test set) το οποίο δεν χρησιμοποιήθηκε σε κανένα στάδιο (Εικόνα 3.42). Η μέση τιμή για τον συντελεστή προσδιορισμού, ανήλθε σε $R^2 = 0.4014$. Το Μέσο Απόλυτο Σφάλμα – MAE δείχνει ότι το μέσο σφάλμα στην πρόβλεψη της μεταβλητής-στόχου y= «Μηνιαία κατανάλωση KWh» είναι περίπου 696 KWh το μήνα. Αντίστοιχα, η ρίζα του μέσου τετραγωνικού σφάλματος (RMSE), η οποία δίνει μεγαλύτερη βαρύτητα στα μεγάλα σφάλματα πρόβλεψης, που μπορεί να οφείλονται ενδεχομένως σε ακραίες τιμές (outliers) ή στην εποχικότητα των δεδομένων, δείχνει ότι το τυπικό μέγεθος του σφάλματος πρόβλεψης ανέρχεται περίπου σε 4575 KWh ανά μήνα.

Η πτώση του συντελεστή R^2 είναι αναμενόμενη καθώς συνδέεται με τον περιορισμό της υπερπροσαρμογής (overfitting) και την αποτίμηση της πραγματικής ικανότητας του μοντέλου για την πρόβλεψη της μεταβλητής-στόχου, **με βάση τα συγκεκριμένα χαρακτηριστικά εισόδου**. Το μοντέλο αξιοποιούσε πληροφορία που δεν θα ήταν διαθέσιμη σε πραγματικές συνθήκες πρόβλεψης. Αντιθέτως, το νέο αποτέλεσμα παρουσιάζει μεγαλύτερη αξιοπιστία και μια πιο ρεαλιστική εκτίμηση της ικανότητας γενίκευσης σε νέα, άγνωστα δεδομένα.



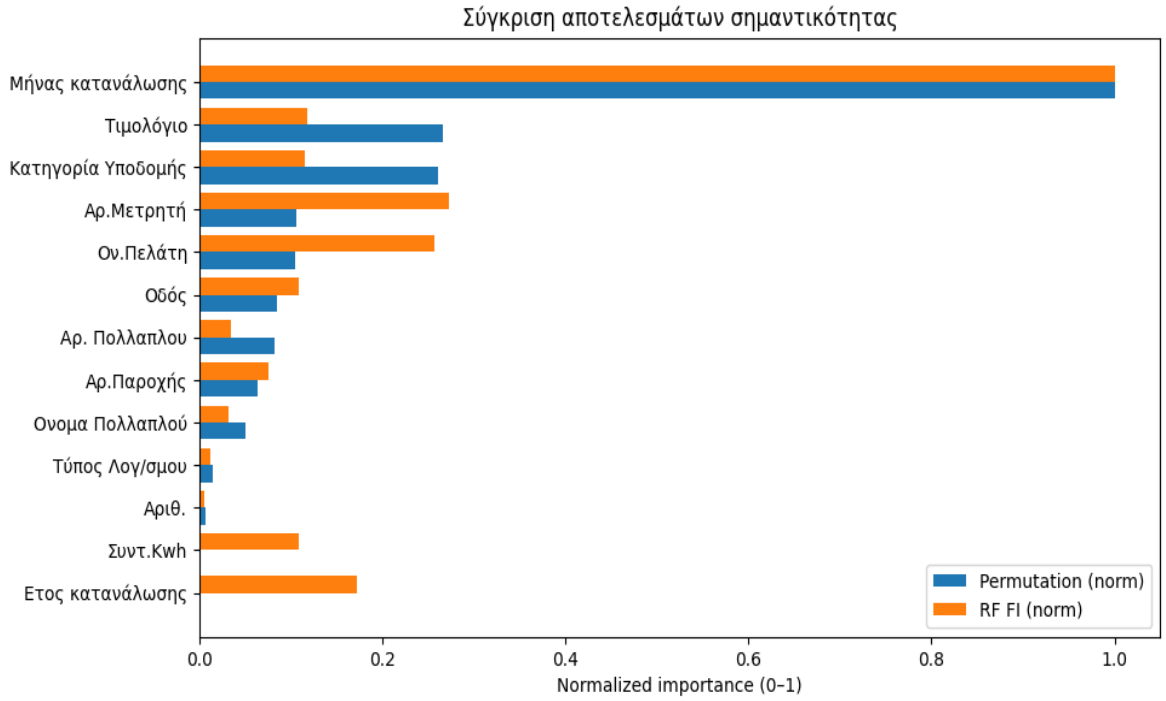
Εικόνα 3.42 Οπτικοποίηση αποτελεσμάτων χωρίς leakage

Το χαρακτηριστικό «**Μήνας κατανάλωσης**» και στις δύο τεχνικές εκτίμησης (Permutation Importance και RF_feature_importances_) συγκεντρώνει το μεγαλύτερο βαθμό (Πίνακας 3.5 & Εικόνα 3.43). Οι παρατηρήσεις του χαρακτηριστικού «Ον.Πελάτη», όπως αναλύθηκε

και στην ενότητα 3.4.4, έχουν αντιστοιχιστεί στις 24 θεματικές κατηγορίες του χαρακτηριστικού «Κατηγορία Υποδομής» το οποίο κατατάσσεται ως το τρίτο πιο σημαντικό γνώρισμα με βάση την μέθοδο Permutation Importance.

Σύγκριση σημαντικότητας (Permutation_Importance - RF feature_importances_)		
Μεταβλητές Εισόδου X	**Permutation_Importance**	**RF_feature_importances**
Μήνας κατανάλωσης	0.265670	0.432658
Τιμολόγιο	0.070637	0.051071
Κατηγορία Υποδομής	0.069387	0.049858
Αρ.Μετρητή	0.028074	0.118105
Ον.Πελάτη	0.027776	0.110816
Οδός	0.022390	0.046826
Αρ. Πολλαπλου	0.021857	0.014802
Αρ.Παροχής	0.017023	0.032803
Ονομα Πολλαπλού	0.013531	0.013845
Τύπος Λογ/σμου	0.003899	0.005225
Αριθ.	0.001615	0.002553
Συντ.Kwh	0.000000	0.047107
Ετος κατανάλωσης	0.000000	0.074331

Πίνακας 3.5 Βαθμός σημαντικότητας χαρακτηριστικών χωρίς leakage



Εικόνα 3.43 Οπτικοποίηση σημαντικότητας χαρακτηριστικών χωρίς leakage

Τα αποτελέσματα των δύο τεχνικών εκτίμησης του βαθμού σημαντικότητας χωρίς το φαινόμενο της διαρροής πληροφορίας από το στόχο (target leakage), αναδεικνύουν το χαρακτηριστικό **«Μήνας κατανάλωσης»** ως το πιο σημαντικό. Αυτό δείχνει ότι η εποχικότητα σε συνδυασμό με την κατηγορία της δημοτικής εγκατάστασης (χαρακτηριστικό **«Κατηγορία Υποδομής»)**, συντελούν καθοριστικά στην ενεργειακή συμπεριφορά των δημοτικών κτιρίων/υποδομών στο συγκεκριμένο σχήμα δεδομένων.

Αντίθετα, χαρακτηριστικά όπως «Οδός», «Αριθ.» σχετίζονται αποκλειστικά με τα στοιχεία της διεύθυνσης ενός καταναλωτή, ενώ τα χαρακτηριστικά «Αρ.Μετρητή» και «Αρ.Πολλαπλού» λειτουργούν κυρίως ως αναγνωριστικά (IDs). Τα χαρακτηριστικά αυτά διαφοροποιούν τα επιμέρους σημεία κατανάλωσης χωρίς να διαμορφώνουν την ενεργειακή συμπεριφορά των δημοτικών υποδομών. Το χαρακτηριστικό «Τύπος Λογ/σμου» αποτυπώνει το είδος του λογαριασμού (π.χ. Έναντι ή Εκκαθαριστικός), δηλαδή ένα γνώρισμα που σχετίζεται με τον τρόπο έκδοσης και εκκαθάρισης του λογαριασμού και όχι με την πραγματική ενεργειακή κατανάλωση. Αντίστοιχα, το χαρακτηριστικό «Συντ.Kwh» αποτελεί τον συντελεστή μετασχηματισμού της ένδειξης του μετρητή - ιδιαίτερα σε περιπτώσεις δημοτικών υποδομών με μεγάλες καταναλώσεις - ώστε να υπολογίζεται η πραγματική κατανάλωση σε KWh. Παρότι είναι κρίσιμος για τον υπολογισμό της κατανάλωσης από την ένδειξη του μετρητή, δεν αποτελεί βασική παράμετρο που να εξηγεί τις διακυμάνσεις της κατανάλωσης στον χρόνο.

Τα παραπάνω χαρακτηριστικά έχουν περιορισμένη ή μηδενική συνδρομή στη διαμόρφωση της συμπεριφοράς της μεταβλητής-στόχου y = «Μηνιαία κατανάλωση kWh» για τις δημοτικές υποδομές/εγκαταστάσεις. Δεν αποτυπώνουν άμεσα τους πραγματικούς παράγοντες που καθορίζουν την ενεργειακή κατανάλωση (π.χ. χρήση εγκατάστασης, εποχικότητα), αλλά κυρίως πληροφορίες ταυτοποίησης ή διαδικαστικά στοιχεία τιμολόγησης/μέτρησης. Το χαρακτηριστικό «Αρ.Παροχής» το οποίο λειτουργεί επίσης ως αναγνωριστικό (ID) και κατατάσσεται αρκετά χαμηλά και στις δύο μεθόδους, εντούτοις θεωρείται σημαντικό διότι συνδέεται μονοσήμαντα με κάθε λογαριασμό ρεύματος.

Για τους σκοπούς της παρούσας εργασίας επιλέγονται, με βάση τον συνδυασμό της ερμηνείας των αποτελεσμάτων και των δυο μεθόδων, ως μεταβλητές εισόδου X , μόνο τα χαρακτηριστικά: **«Έτος κατανάλωσης», «Μήνας κατανάλωσης», «Αρ.Παροχής», «Τιμολόγιο», «Όνομα Πολλαπλού»** και **«Κατηγορία Υποδομής»** (αρχείο data.csv)

3.6.3 Ανάλυση δεδομένων μετά την επιλογή χαρακτηριστικών

Το τελικό αρχείο (data.csv) το οποίο περιλαμβάνει τα σημαντικότερα χαρακτηριστικά για την πρόβλεψη της μηνιαίας ενεργειακής κατανάλωσης στο συγκεκριμένο σχήμα δεδομένων, αποτελεί ένα συνεκτικό και κατάλληλα διαμορφωμένο σύνολο δεδομένων. Τα δεδομένα είναι οργανωμένα ως πολλαπλές χρονοσειρές (panel time series), καθώς η κατανάλωση της ηλεκτρικής ενέργειας των δημοτικών εγκαταστάσεων, καταγράφεται σε μηνιαία βάση ανά αριθμό παροχής, ενώ παράλληλα υπάρχει ομαδοποίηση και ανάλυση ανά κατηγορία δημοτικής υποδομής.

Μια χρονοσειρά μηνιαίας κατανάλωσης y_t αποτελείται από τέσσερα βασικά συστατικά (Cleveland et al., 1990):

- **Τάση** (trend $\boldsymbol{T_t}$): αποτυπώνει την μακροπρόθεσμη πορεία της ενεργειακής κατανάλωσης, (αύξηση ή μείωση) σε βάθος χρόνου.
- **Εποχικότητα** (seasonality $\boldsymbol{S_t}$): αποτυπώνει τις διακυμάνσεις της ενεργειακής κατανάλωσης μέσα από σταθερά μοτίβα (σταθερά χρονικά διαστήματα) χειμώνας/καλοκαίρι
- **Κυκλικότητα** (cyclicity $\boldsymbol{C_t}$): αποτυπώνει τις διακυμάνσεις της ενεργειακής κατανάλωσης για μεγαλύτερα χρονικά διαστήματα από την εποχικότητα οι οποίες εμφανίζονται ακαθόριστα χωρίς σταθερή περίοδο.
- **Τυχαία δεδομένα – Θόρυβος** (irregular/noise $\boldsymbol{R_t}$): αποτυπώνουν μεταβολές στη διακύμανση της ενεργειακής κατανάλωσης που δεν εξηγούνται από την τάση την εποχικότητα και την κυκλικότητα και οφείλονται σε τυχαία γεγονότα.

Η εποχικότητα στην ενεργειακή κατανάλωση των δημοτικών υποδομών αποτελεί κρίσιμη παράμετρο, καθώς οι ανάγκες της κατανάλωσης επηρεάζονται άμεσα από την εποχή και κυρίως, το μήνα κατά τον οποίο συντελείται η κατανάλωση του ηλεκτρικού ρεύματος. Κατά τους θερινούς μήνες οι απαιτήσεις κατανάλωσης ηλεκτρικού ρεύματος για το φωτισμό των κτιρίων και των κοινόχρηστων χώρων (πεζοδρόμια, πλατείες, κτλ.) είναι μικρότερες σε σχέση με τους χειμερινούς μήνες όπου η διάρκεια της νύχτας είναι πολύ μεγαλύτερη. Αντιθέτως οι απαιτήσεις κατανάλωσης για κλιματισμό των κτιρίων αυξάνονται ραγδαία το καλοκαίρι σε σχέση με το χειμώνα λόγω υψηλών θερμοκρασιών.

Η αξιοπιστία και η ποιότητα των παρατηρήσεων σε ενεργειακά δεδομένα, προϋποθέτει τον έλεγχο της σχέσης τους με την εποχή που αυτά καταγράφονται και την μακροπρόθεσμη

τάση που διαμορφώνεται ως αποτέλεσμα αυτής της καταγραφής στο χρόνο. Παρατηρήσεις που ακολουθούν επαναλαμβανόμενα εποχικά μοτίβα και υιοθετούν συγκεκριμένη μακροπρόθεσμη τάση (ανοδική ή καθοδική), δεν χαρακτηρίζονται ως εσφαλμένες παρατηρήσεις. Παρατηρήσεις που δεν ερμηνεύονται με την εποχικότητα αλλά και με την μακροπρόθεσμη τάση, πρέπει να ελεγχθούν και να αξιολογηθούν αν αποτελούν θόρυβο ή ακραίες τιμές (outliers).

Για τη διερεύνηση της εποχικής ανάλυσης και αποτύπωσης της μακροπρόθεσμης τάσης των αριθμητικών τιμών της χρονοσειράς «Μηνιαία κατανάλωση KWh» ανά κατηγορία υποδομής, θα εφαρμοστεί η μέθοδος **STL – Seasonal - Trend decomposition using Loess**. Η μέθοδος αυτή διασπά την τιμή της χρονοσειράς σε τρείς επιμέρους τιμές. Στην τιμή που σχετίζεται με την εποχικότητα (Seasonal), στην τιμή που σχετίζεται με την μακροπρόθεσμη τάση (Trend) και στην τιμή που δεν μπορεί να ερμηνευτεί βάσει εποχικότητας ή/και τάσης, η οποία χαρακτηρίζεται ως υπόλοιπο (Residuals).

```
# STL decomposition (period=12 για μηνιαία)
stl = STL(y, period=12, robust=True).fit()
trend = stl.trend
seasonal = stl.seasonal
resid = stl.resid
```

Εικόνα 3.44 STL αποσύνθεση χρονοσειράς

Αυτό επιτυγχάνεται μέσω της στατιστικής μεθόδου LOESS - **Lo**cally **E**stimated **S**catterplot **S**moothing στον αλγόριθμο STL. Η μέθοδος LOESS για κάθε χρονική στιγμή t εφαρμόζει για την παρατήρηση y_t τοπική πολυωνυμική παλινδρόμηση σε ένα παράθυρο γειτονικών παρατηρήσεων y_{ti} - οι οποίες ενδεχομένως παρουσιάζουν μη γραμμική συμπεριφορά – αποδίδοντας βάρη με τη συνάρτηση tricube. Τα βάρη υπολογίζονται με βάση την κανονικοποιημένη χρονική απόσταση των παρατηρήσεων y_{ti} από την παρατήρηση y_t , $u_i = \frac{|t_i - t|}{d(t)}$ όπου $d(t)$ η χρονική απόσταση μεταξύ της πιο απομακρυσμένης γειτονικής παρατήρησης y_k από την y_t , όπου k το πλήθος των γειτόνων. Εάν η απόλυτη τιμή της χρονικής απόστασης $|u_i|$ είναι μικρότερη από τη μονάδα $|u_i| < 1$ τότε $W(u_i) = (1 - |u|^3)^3$, ενώ εάν είναι μεγαλύτερη ή ίση $|u_i| \geq 1$ τότε $W(u_i) = 0$. Τα βάρη φθίνουν όσο η χρονική απόσταση από το t αυξάνεται.

Ο αλγόριθμος STL για κάθε παρατήρηση y_t της χρονοσειράς «Μηνιαία κατανάλωση KWh» χρησιμοποιεί την πολυωνυμική παλινδρόμηση LOESS σε κυλιόμενα παράθυρα για να

εκτιμήσει την τιμή της εποχικότητας *S(t)* και την τιμή της μακροπρόθεσμης τάσης *T(t)* και τελικά να προσδιορίσει το υπόλοιπο *R(t)* για το οποίο δεν υπάρχει ερμηνεία σημαίνοντάς το για περαιτέρω έλεγχο. Η κάθε παρατήρηση y_t της χρονοσειράς αποτελείται από το άθροισμα των τριών στοιχείων: $y_t = S(t) + T(t) + R(t)$. Βασικές παράμετροι λειτουργίας του αλγορίθμου STL είναι η περίοδος (period) της εποχικότητας, το πλάτος του παραθύρου LOESS για την εποχικότητα (seasonal), το πλάτος του παραθύρου LOESS για την τάση (trend) και η παράμετρος της ανθεκτικότητας (robust) για την μείωση της επίδρασης των ακραίων τιμών.

Στο πλαίσιο των ενεργειακών δεδομένων των δημοτικών υποδομών η κυκλικότητα C_t μπορεί να ερμηνευτεί ως οι μεσοπρόθεσμες αποκλίσεις της κατανάλωσης γύρω από τη μακροπρόθεσμη πορεία-τάση, οι οποίες δεν σχετίζονται με την εποχικότητα και δεν εμφανίζονται με σταθερή περιοδικότητα. Για την διερεύνηση αυτών των αποκλίσεων και την περαιτέρω αποτύπωση της μακροχρόνιας πορείας αξιοποιείται συμπληρωματικά το φίλτρο **Hodrick–Prescott (HP)** (Hodrick & Prescott, 1997).

```
# Cyclicity: αφαιρούμε εποχικότητα και παίρνουμε "κύκλο" με HP filter
# Για μηνιαία δεδομένα συνήθως λ=129600
deseasonal = y - seasonal
cycle, hp_trend = hpfilter(deseasonal, lamb=129600)
```

Εικόνα 3.45 Φίλτρο Hodrick–Prescott κυκλικότητα χρονοσειράς

Το φίλτρο διαχωρίζει μια χρονοσειρά μηνιαίας κατανάλωσης y_t σε μακροπρόθεσμη τάση T_t και σε κυκλική συνιστώσα C_t, ώστε $y_t = T_t + C_t$. Η τάση T_t εκτιμάται ως μια ομαλή καμπύλη που παραμένει κοντά στις πραγματικές τιμές ενώ ταυτόχρονα αποφεύγει τις απότομες μεταβολές. Αυτό επιτυγχάνεται μέσω ενός προβλήματος βελτιστοποίησης που εξισορροπεί την προσαρμογή στα δεδομένα με έναν όρο εξομάλυνσης που περιορίζει τις απότομες μεταβολές στην καμπυλότητα της τάσης.

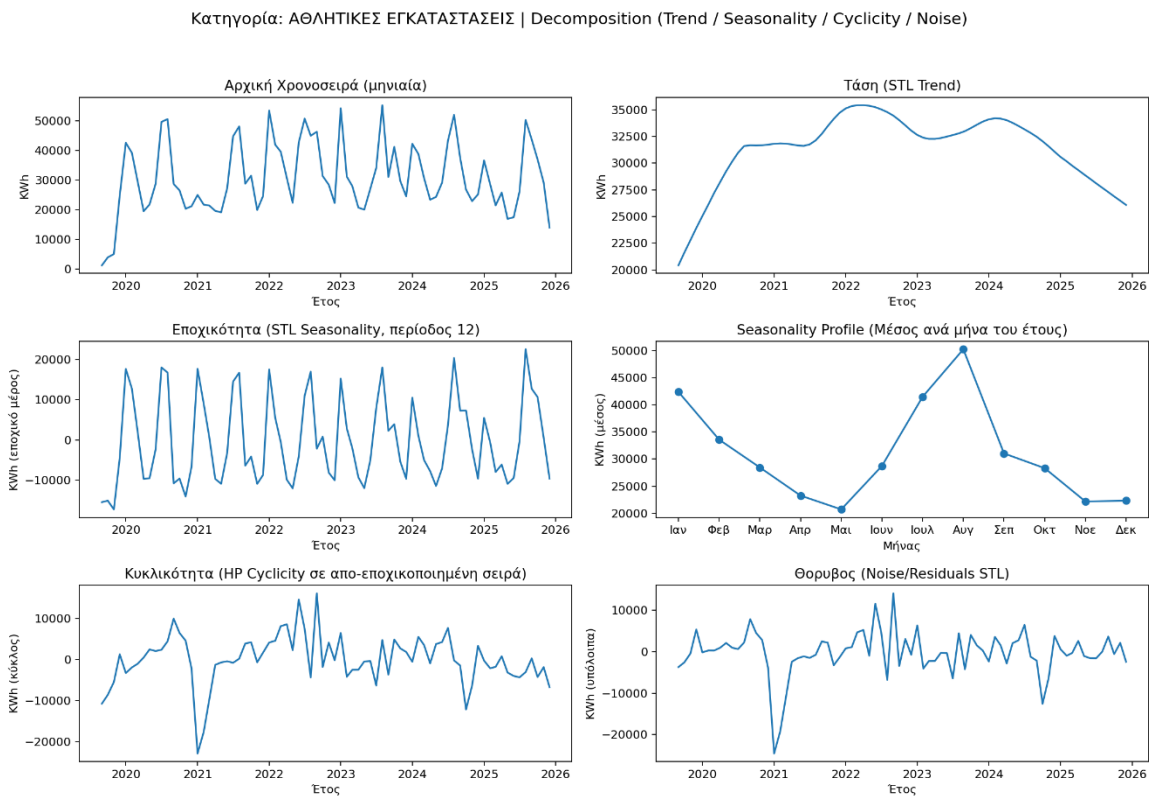
Ο βαθμός εξομάλυνσης στο φίλτρο Hodrick–Prescott καθορίζεται από την παράμετρο λ, η οποία ελέγχει την αυστηρότητα της ομαλοποίησης. Όσο μεγαλύτερη είναι η τιμή της, τόσο πιο αργά μεταβαλλόμενη γίνεται η εκτιμώμενη τάση T_t ενώ αντίστοιχα μεγαλύτερο μέρος της μεσοπρόθεσμης μεταβλητότητας αποδίδεται στην κυκλική συνιστώσα C_t. Για δεδομένα διαφορετικής συχνότητας η επιλογή της παραμέτρου αυτής πρέπει να ρυθμίζεται αναλόγως

με τη σχετική βιβλιογραφία να προτείνει κατάλληλη προσαρμογή της παραμέτρου λ (Ravn & Uhlig, 2002).

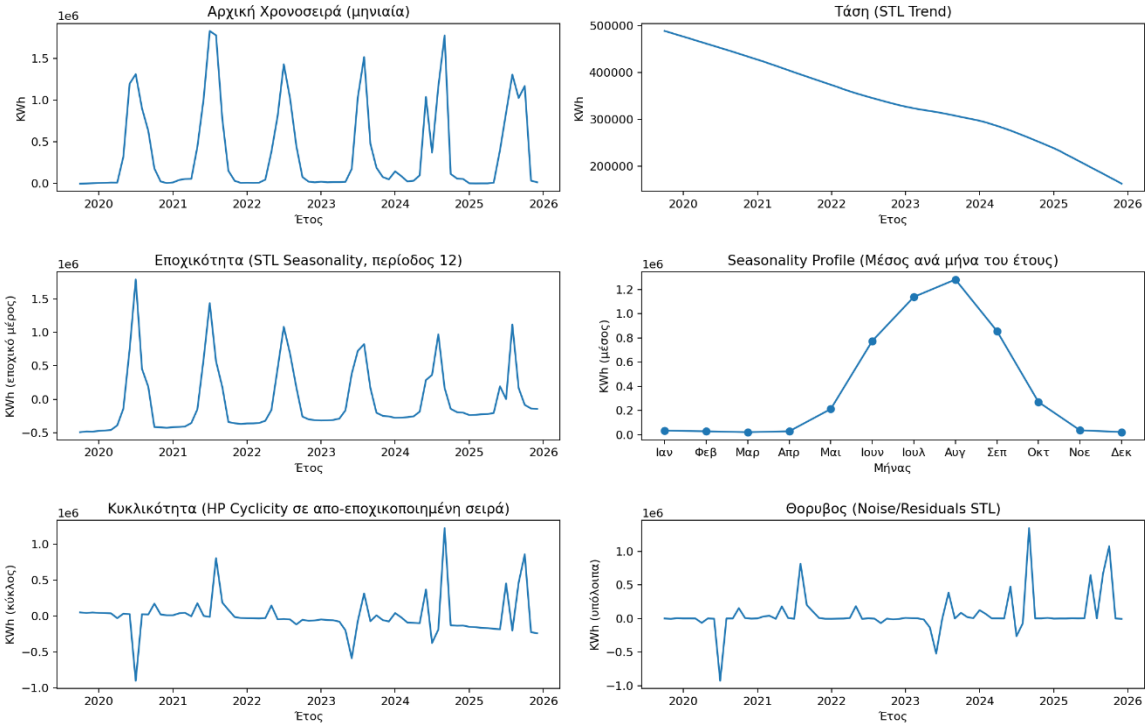
Η χρήση του HP filter είναι ιδιαίτερα χρήσιμη, καθώς μπορεί να εφαρμοστεί σε χρονοσειρά αφού έχει αφαιρεθεί η εποχικότητα ($y_t - S_t$) για την εκτίμηση μιας πιο καθαρής μακροπρόθεσμης τάσης και την ανάδειξη των κυκλικών αποκλίσεων. Με αυτόν τον τρόπο, το φίλτρο συμβάλλει στην καλύτερη κατανόηση της δομής της χρονοσειράς και στην αξιολόγηση παρατηρήσεων που δεν εξηγούνται επαρκώς από την εποχικότητα και την τάση, πριν αυτές χαρακτηριστούν ως θόρυβος ή ακραίες τιμές.

Η ανωτέρω σύντομη ανάλυση παρατίθεται απλά για τις ανάγκες της οπτικής αναπαράστασης και κατανόησης της συμπεριφοράς των ενεργειακών δεδομένων του τελικού αρχείου data.csv μέσω κατάλληλης υλοποίησης σε Python.

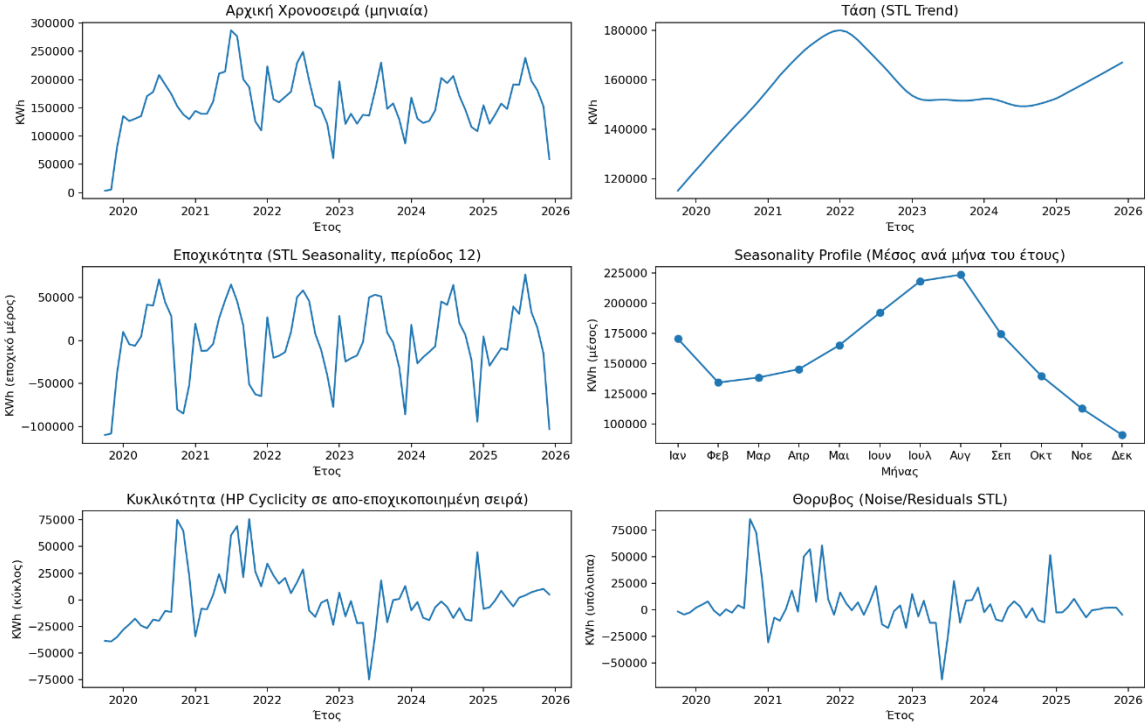
Στη συνέχεια παρατίθενται συγκεντρωτικά διαγράμματα που αποτυπώνουν τα βασικά συστατικά της χρονοσειράς «Μηνιαία κατανάλωση KWh», με σκοπό την οπτική απεικόνιση της ενεργειακής συμπεριφοράς των κατηγοριοποιημένων δημοτικών υποδομών. Η ανάλυση βασίζεται αποκλειστικά στα πρωτογενή δεδομένα (raw data) όπως αυτά διαμορφώθηκαν μέχρι το σημείο αυτό.

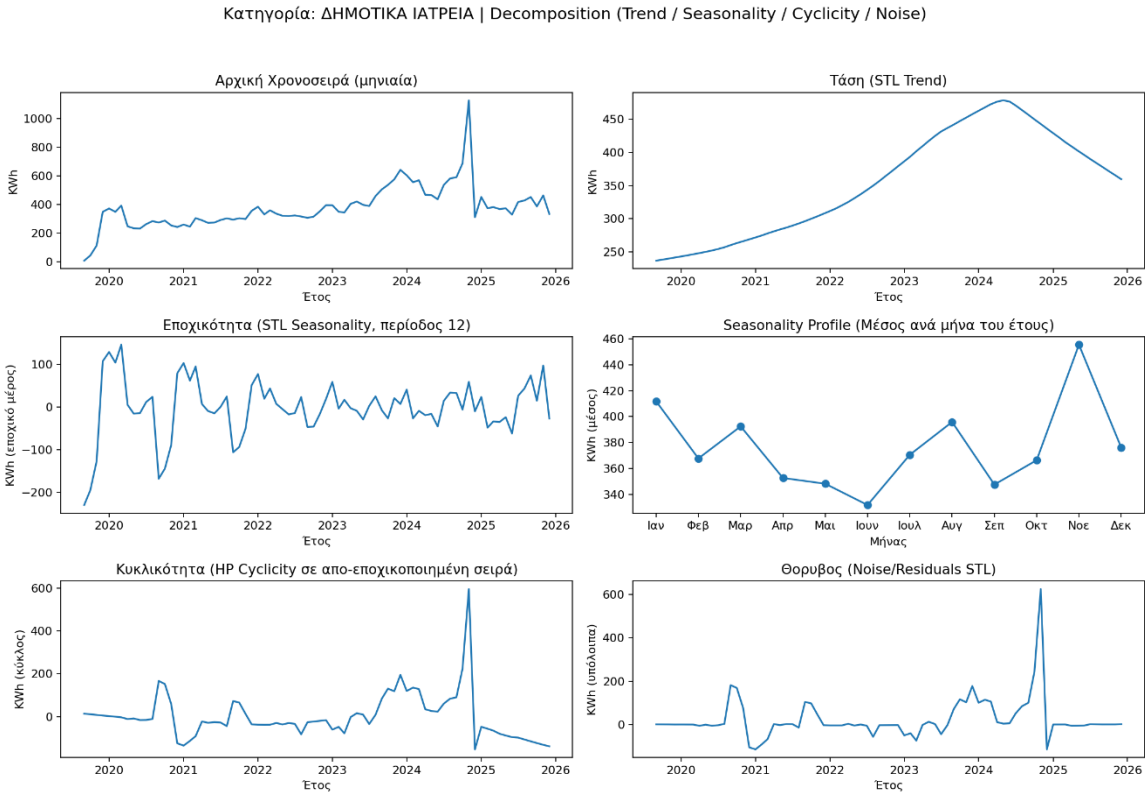


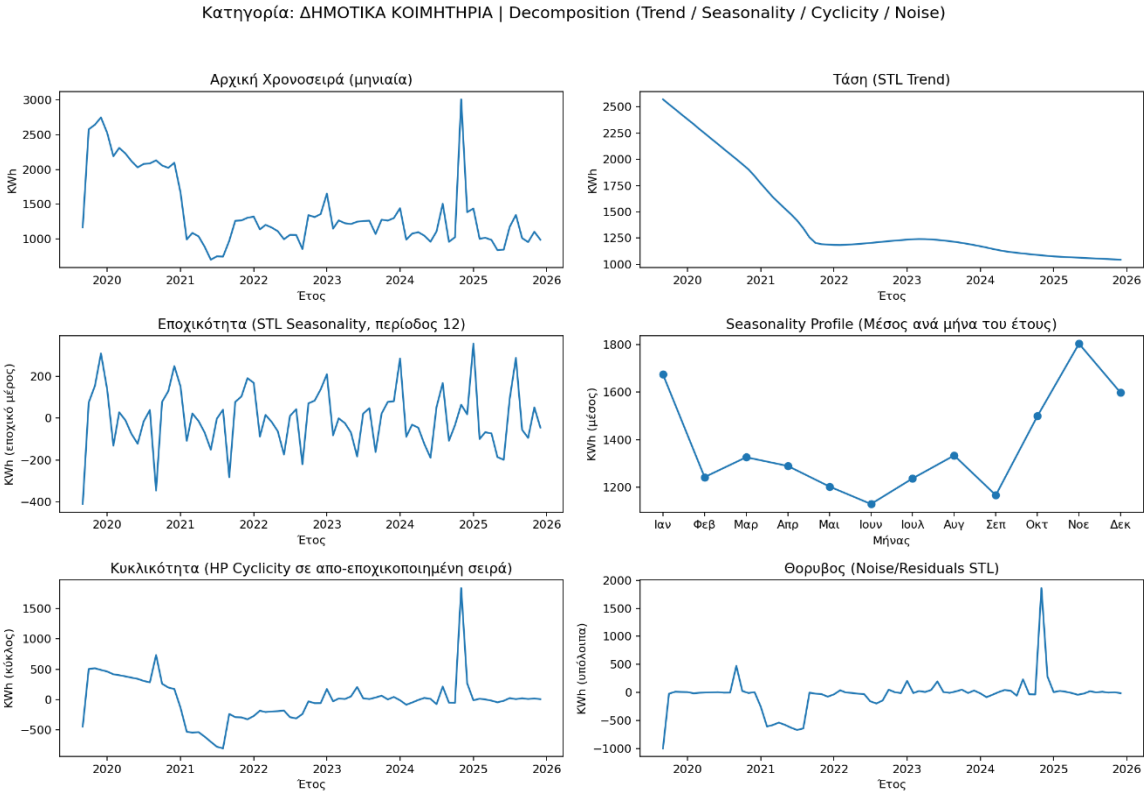
Κατηγορία: ΑΝΤΛΙΟΣΤΑΣΙΑ ΑΡΔΕΥΣΗΣ | Decomposition (Trend / Seasonality / Cyclicity / Noise)



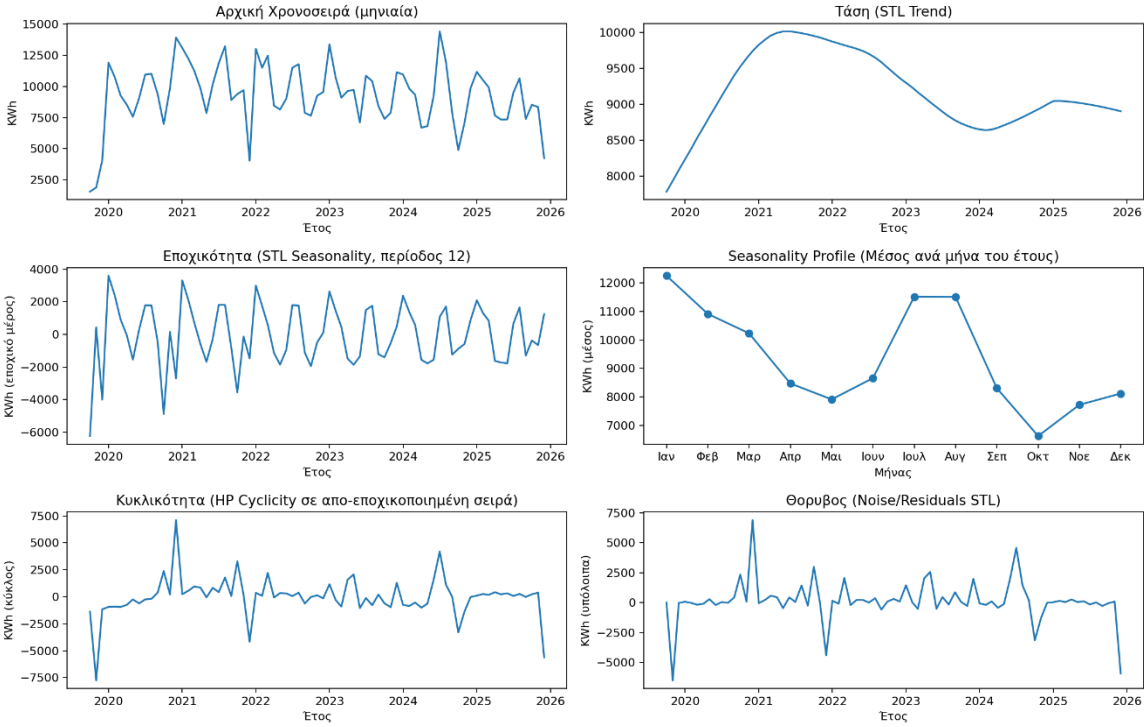
Κατηγορία: ΑΝΤΛΙΟΣΤΑΣΙΑ ΥΔΡΕΥΣΗΣ | Decomposition (Trend / Seasonality / Cyclicity / Noise)



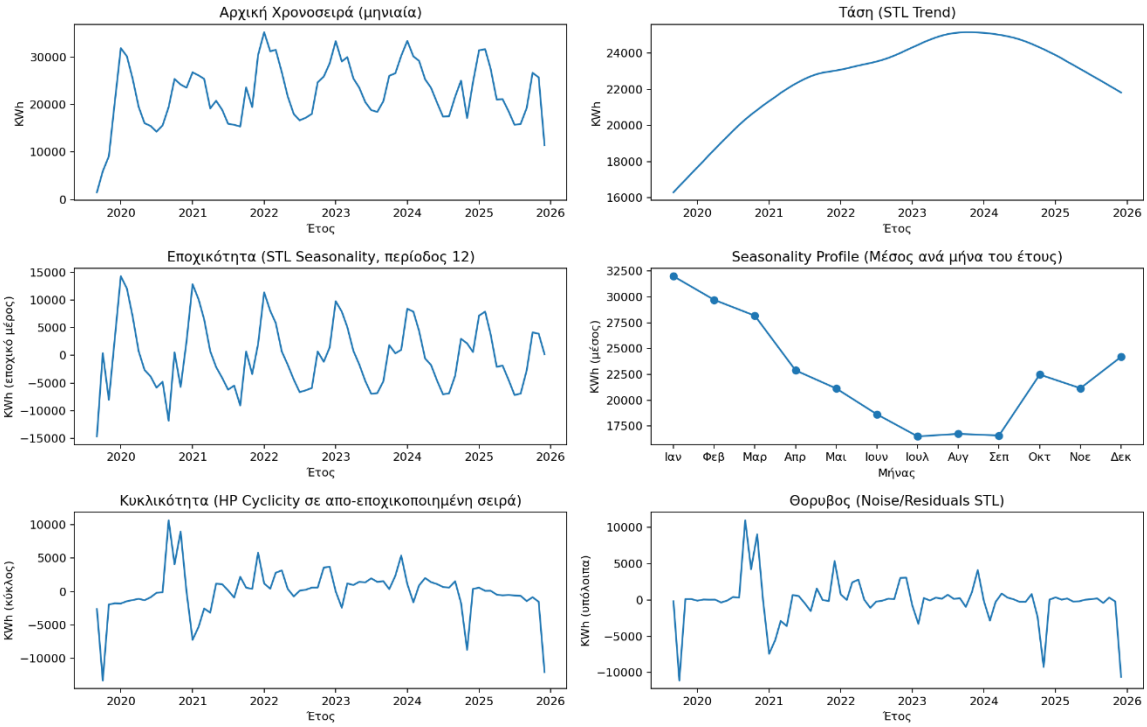

Κατηγορία: ΔΗΜΟΤΙΚΑ ΙΑΤΡΕΙΑ | Decomposition (Trend / Seasonality / Cyclicity / Noise)
Αρχική Χρονοσειρά (μηνιαία)
Τάση (STL Trend)
Εποχικότητα (STL Seasonality, περίοδος 12)
Seasonality Profile (Μέσος ανά μήνα του έτους)
Κυκλικότητα (HP Cyclicity σε απο-εποχικοποιημένη σειρά)
Θορυβος (Noise/Residuals STL)
KWh
Έτος
Μήνας
Ιαν
Φεβ
Μαρ
Απρ
Μαι
Ιουν
Ιουλ
Αυγ
Σεπ
Οκτ
Νοε
Δεκ


Κατηγορία: ΔΗΜΟΤΙΚΑ ΚΟΙΜΗΤΗΡΙΑ | Decomposition (Trend / Seasonality / Cyclicity / Noise)
Αρχική Χρονοσειρά (μηνιαία)
Τάση (STL Trend)
Εποχικότητα (STL Seasonality, περίοδος 12)
Seasonality Profile (Μέσος ανά μήνα του έτους)
Κυκλικότητα (HP Cyclicity σε απο-εποχικοποιημένη σειρά)
Θορυβος (Noise/Residuals STL)
KWh
Έτος
Μήνας
Ιαν
Φεβ
Μαρ
Απρ
Μαι
Ιουν
Ιουλ
Αυγ
Σεπ
Οκτ
Νοε
Δεκ

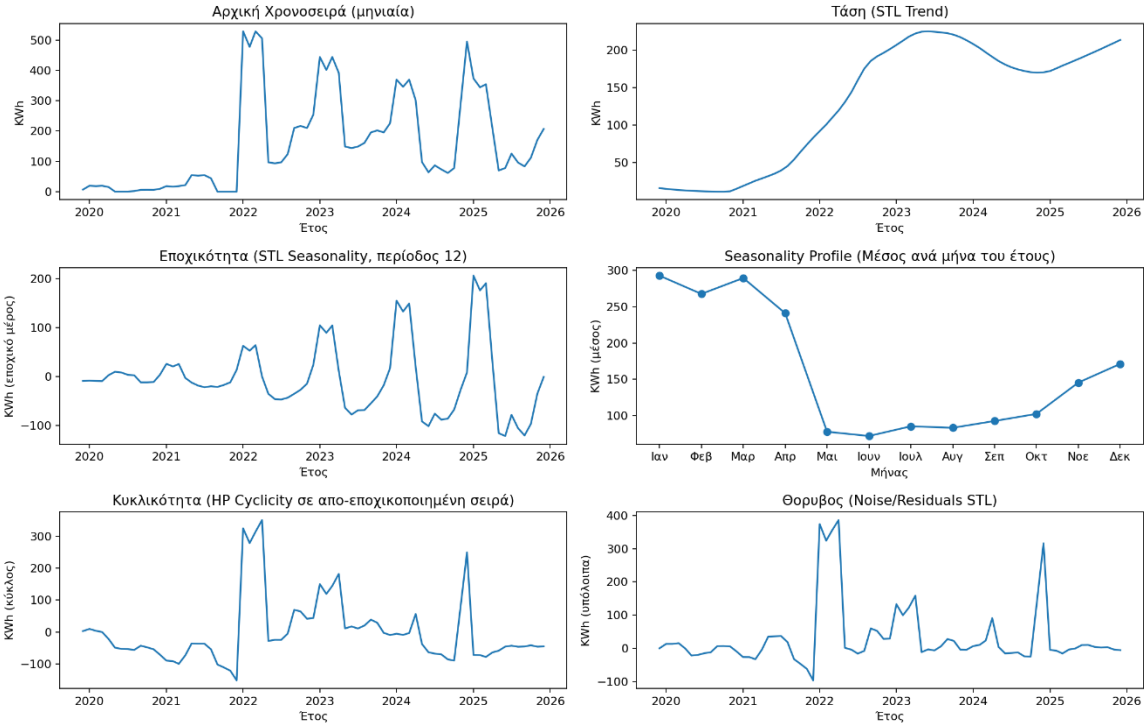
Κατηγορία: ΔΗΜΟΤΙΚΑ ΚΤΙΡΙΑ | Decomposition (Trend / Seasonality / Cyclicity / Noise)



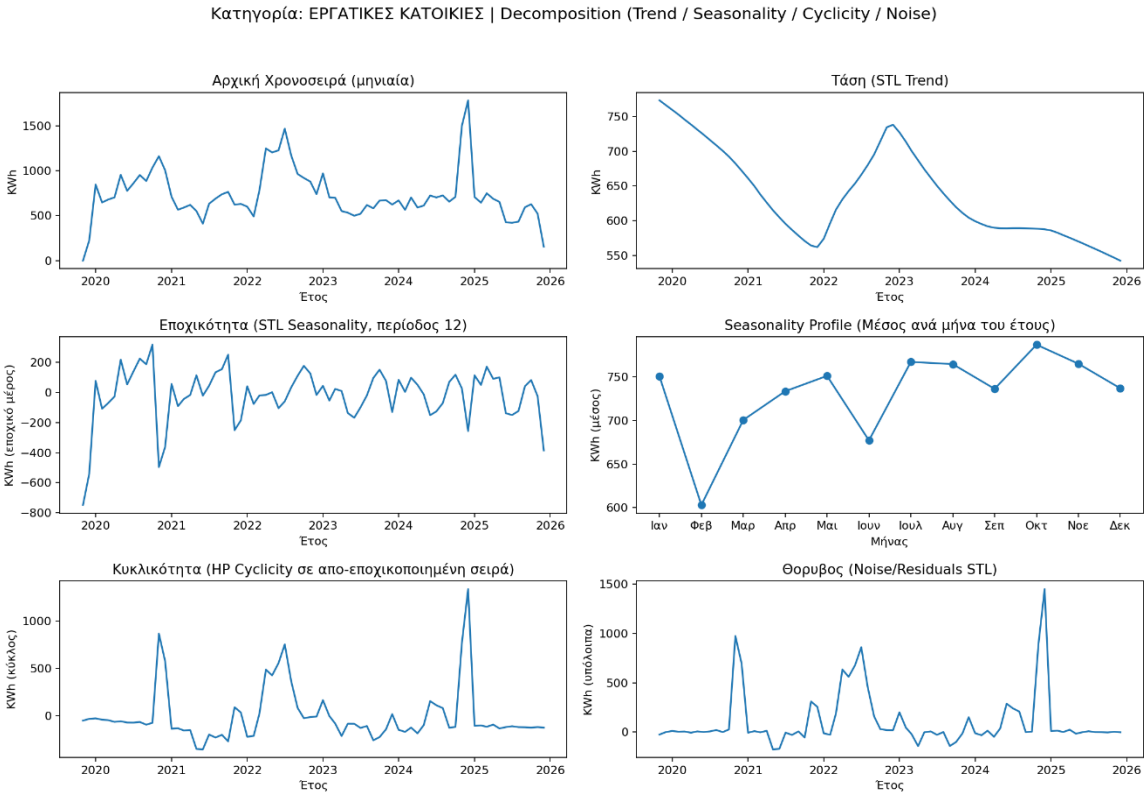
Κατηγορία: ΔΗΜΟΤΙΚΑ ΣΧΟΛΕΙΑ | Decomposition (Trend / Seasonality / Cyclicity / Noise)

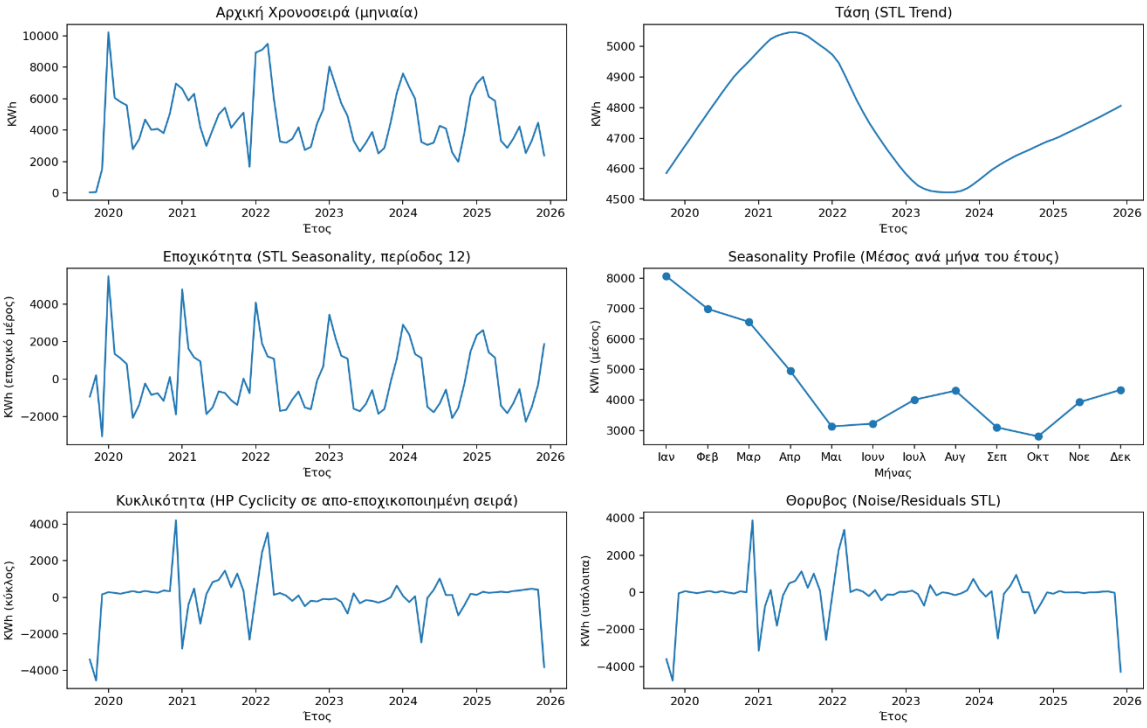


Κατηγορία: ΔΗΜΟΤΙΚΗ ΑΓΟΡΑ | Decomposition (Trend / Seasonality / Cyclicity / Noise)

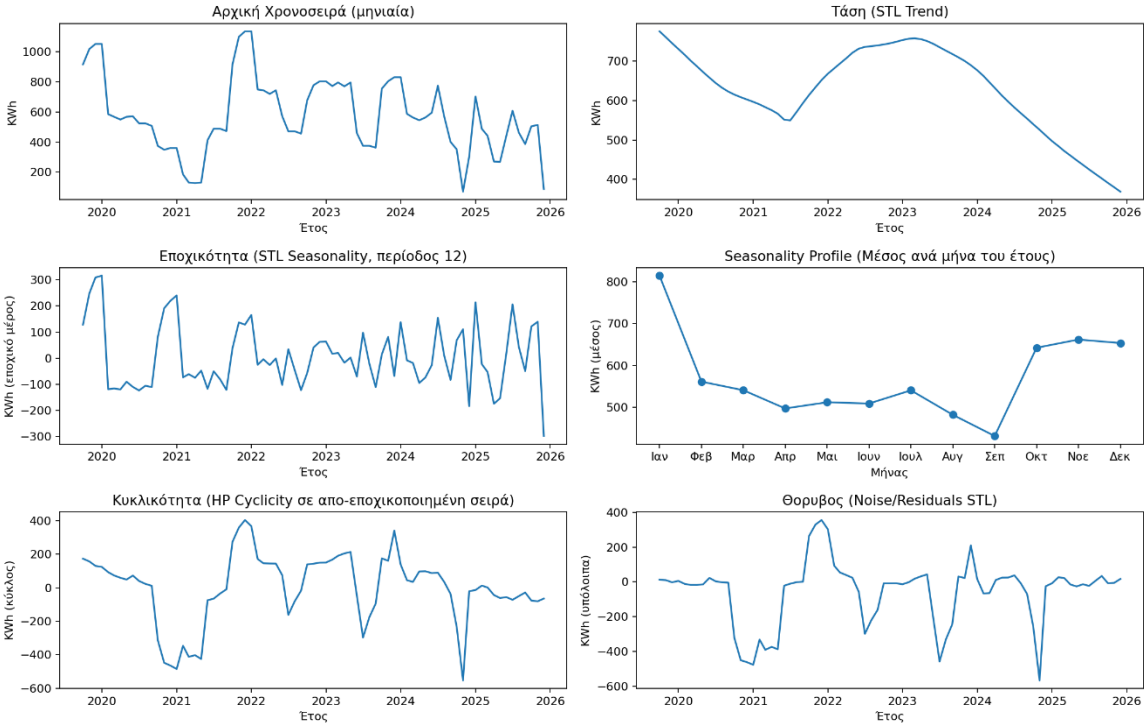


Κατηγορία: ΕΓΚΑΤΑΣΤΑΣΗ ΕΠΕΞΕΡΓΑΣΙΑΣ ΛΥΜΑΤΩΝ | Decomposition (Trend / Seasonality / Cyclicity / Noise)

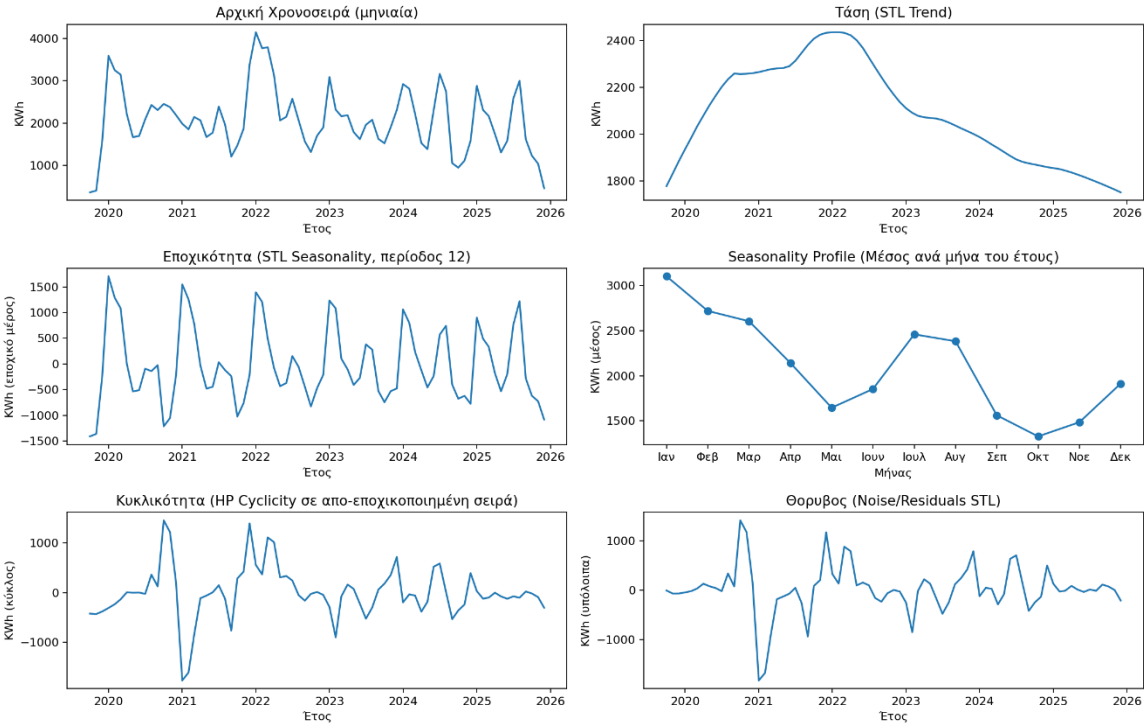

Κατηγορία: ΕΡΓΑΤΙΚΕΣ ΚΑΤΟΙΚΙΕΣ | Decomposition (Trend / Seasonality / Cyclicity / Noise)
Αρχική Χρονοσειρά (μηνιαία)
Τάση (STL Trend)
Εποχικότητα (STL Seasonality, περίοδος 12)
Seasonality Profile (Μέσος ανά μήνα του έτους)
Κυκλικότητα (HP Cyclicity σε απο-εποχικοποιημένη σειρά)
Θορυβος (Noise/Residuals STL)
KWh
Έτος
Μήνας
Ιαν
Φεβ
Μαρ
Απρ
Μαι
Ιουν
Ιουλ
Αυγ
Σεπ
Οκτ
Νοε
Δεκ

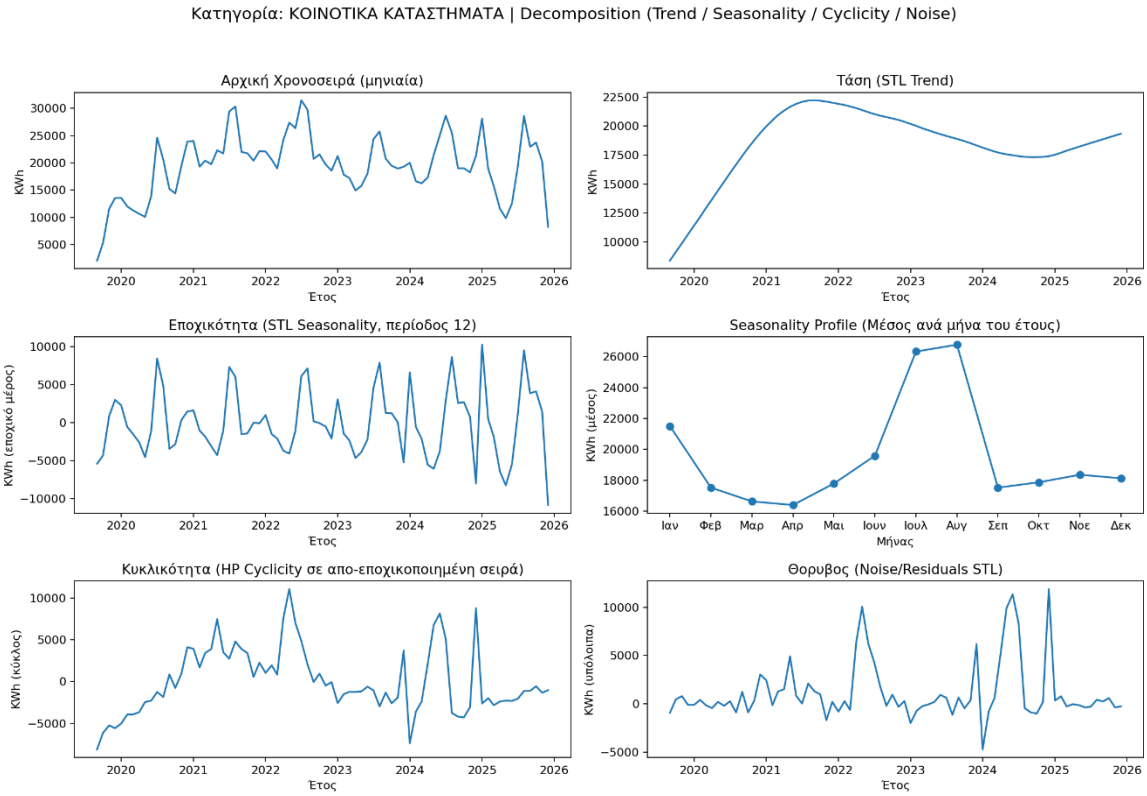

Κατηγορία: ΚΑΠΗ | Decomposition (Trend / Seasonality / Cyclicity / Noise)
Αρχική Χρονοσειρά (μηνιαία)
Τάση (STL Trend)
Εποχικότητα (STL Seasonality, περίοδος 12)
Seasonality Profile (Μέσος ανά μήνα του έτους)
Κυκλικότητα (HP Cyclicity σε απο-εποχικοποιημένη σειρά)
Θορυβος (Noise/Residuals STL)
KWh
Έτος
Μήνας
Ιαν
Φεβ
Μαρ
Απρ
Μαι
Ιουν
Ιουλ
Αυγ
Σεπ
Οκτ
Νοε
Δεκ

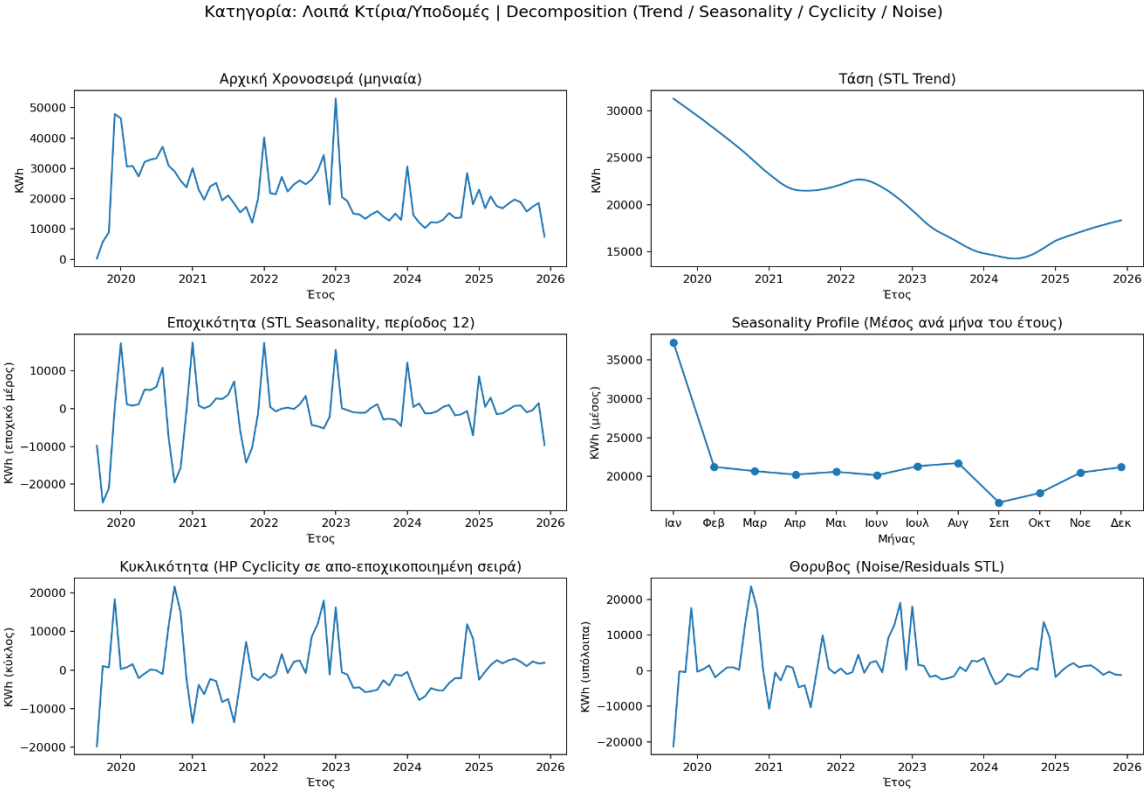
Κατηγορία: ΚΔΑΠ | Decomposition (Trend / Seasonality / Cyclicity / Noise)



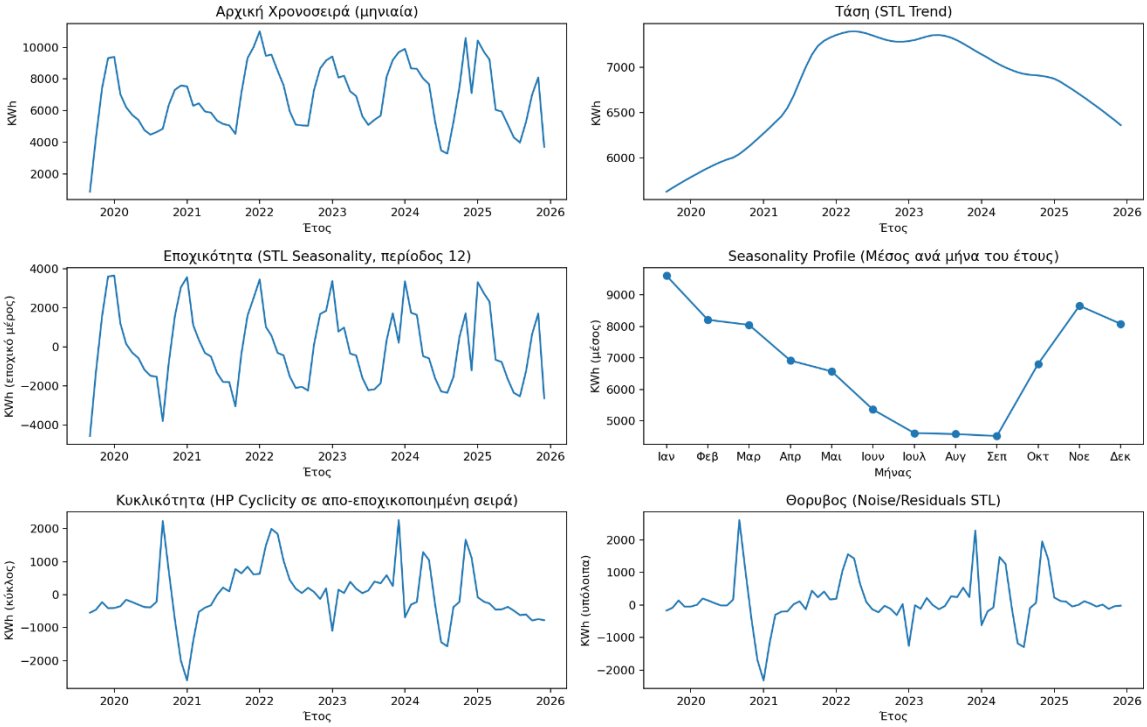
Κατηγορία: ΚΕΝΤΡΟ ΕΞΥΠΗΡΕΤΗΣΗΣ ΠΟΛΙΤΩΝ | Decomposition (Trend / Seasonality / Cyclicity / Noise)



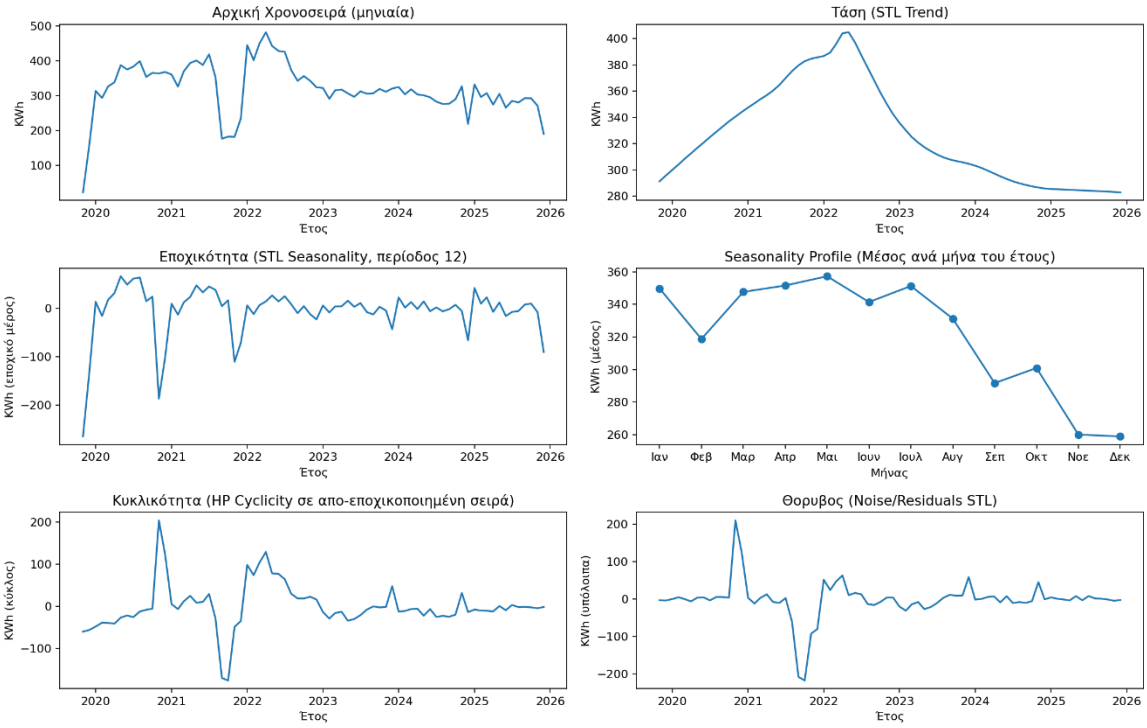

Κατηγορία: ΚΟΙΝΟΤΙΚΑ ΚΑΤΑΣΤΗΜΑΤΑ | Decomposition (Trend / Seasonality / Cyclicity / Noise)
Αρχική Χρονοσειρά (μηνιαία)
Τάση (STL Trend)
Εποχικότητα (STL Seasonality, περίοδος 12)
Seasonality Profile (Μέσος ανά μήνα του έτους)
Κυκλικότητα (HP Cyclicity σε απο-εποχικοποιημένη σειρά)
Θορυβος (Noise/Residuals STL)
KWh
Έτος
Μήνας
Ιαν
Φεβ
Μαρ
Απρ
Μαι
Ιουν
Ιουλ
Αυγ
Σεπ
Οκτ
Νοε
Δεκ


Κατηγορία: Λοιπά Κτίρια/Υποδομές | Decomposition (Trend / Seasonality / Cyclicity / Noise)
Αρχική Χρονοσειρά (μηνιαία)
Τάση (STL Trend)
Εποχικότητα (STL Seasonality, περίοδος 12)
Seasonality Profile (Μέσος ανά μήνα του έτους)
Κυκλικότητα (HP Cyclicity σε απο-εποχικοποιημένη σειρά)
Θορυβος (Noise/Residuals STL)
KWh
Έτος
Μήνας
Ιαν
Φεβ
Μαρ
Απρ
Μαι
Ιουν
Ιουλ
Αυγ
Σεπ
Οκτ
Νοε
Δεκ

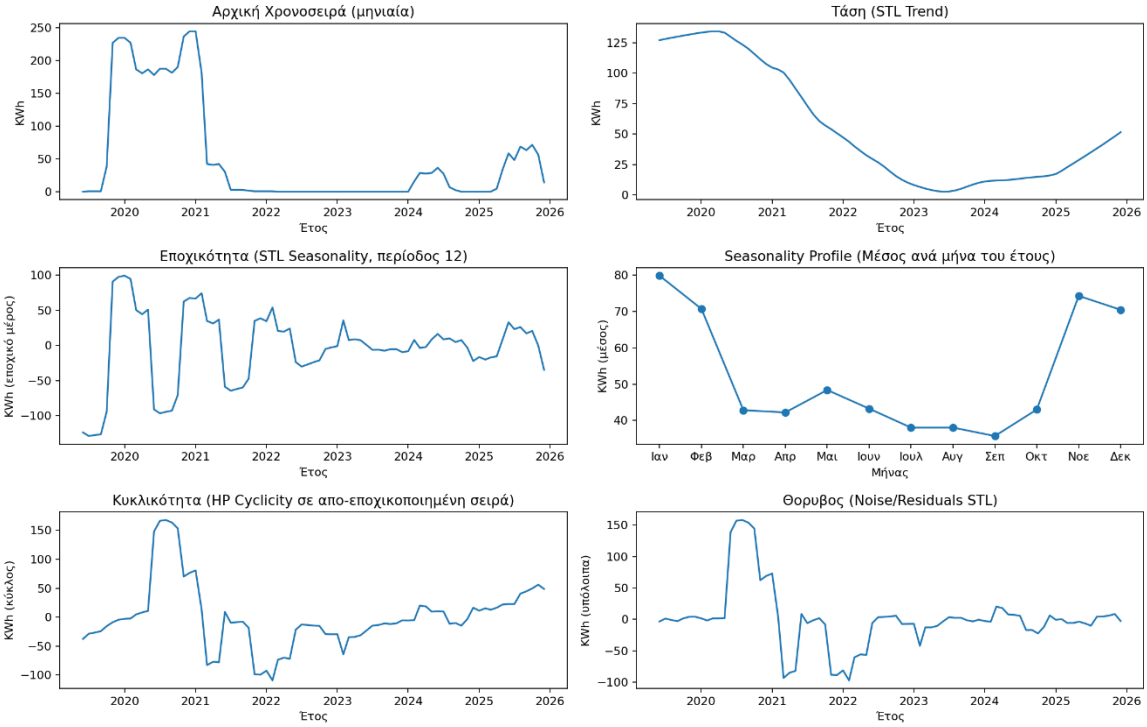
Κατηγορία: ΝΗΠΙΑΓΩΓΕΙΑ | Decomposition (Trend / Seasonality / Cyclicity / Noise)



Κατηγορία: ΟΔΟΣΗΜΑΝΣΗ | Decomposition (Trend / Seasonality / Cyclicity / Noise)

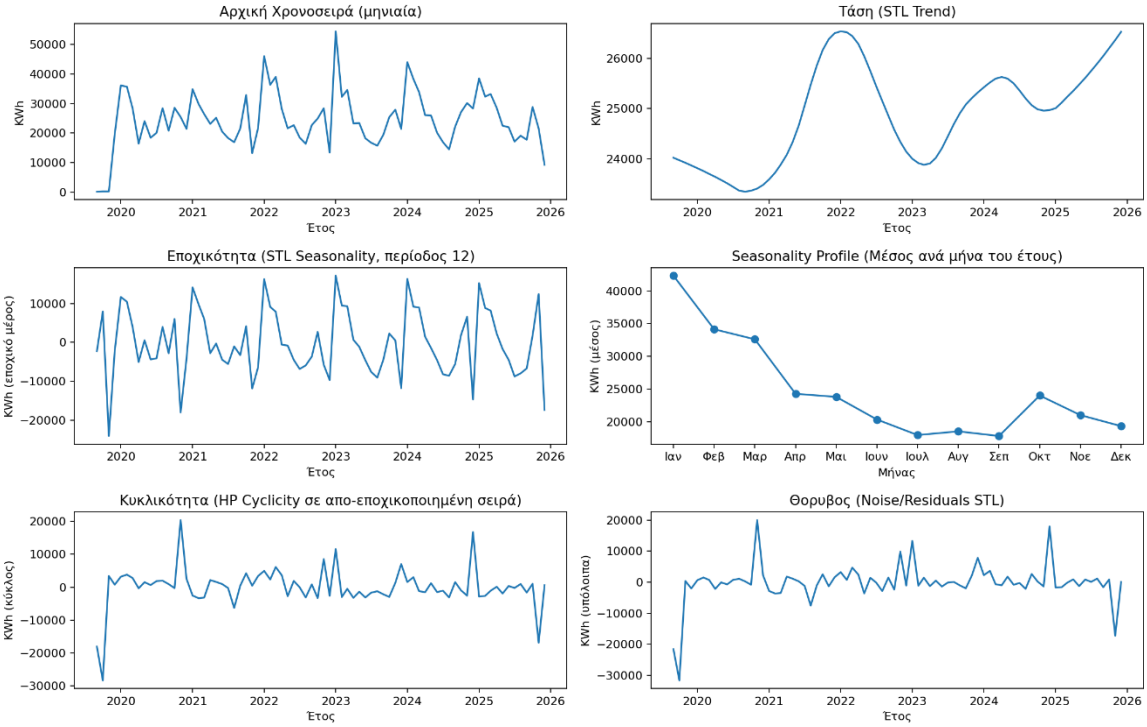


Κατηγορία: ΠΑΙΔΙΚΕΣ ΧΑΡΕΣ | Decomposition (Trend / Seasonality / Cyclicity / Noise)

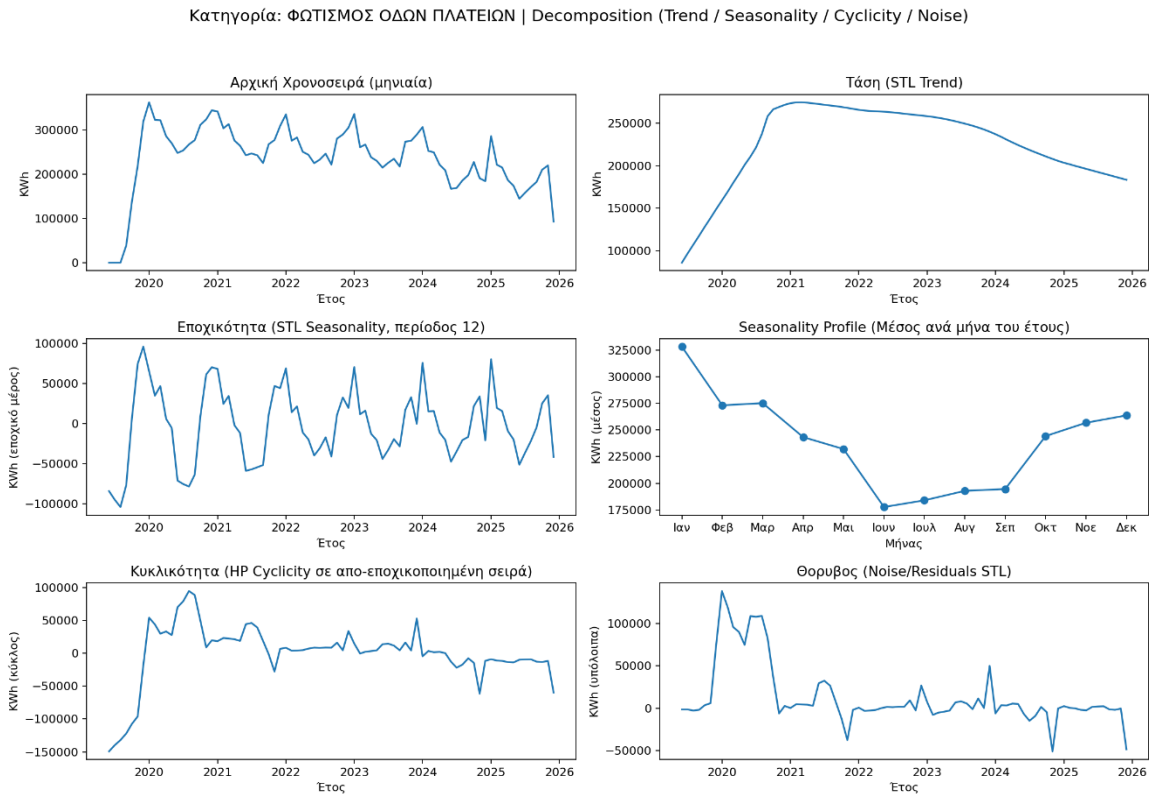


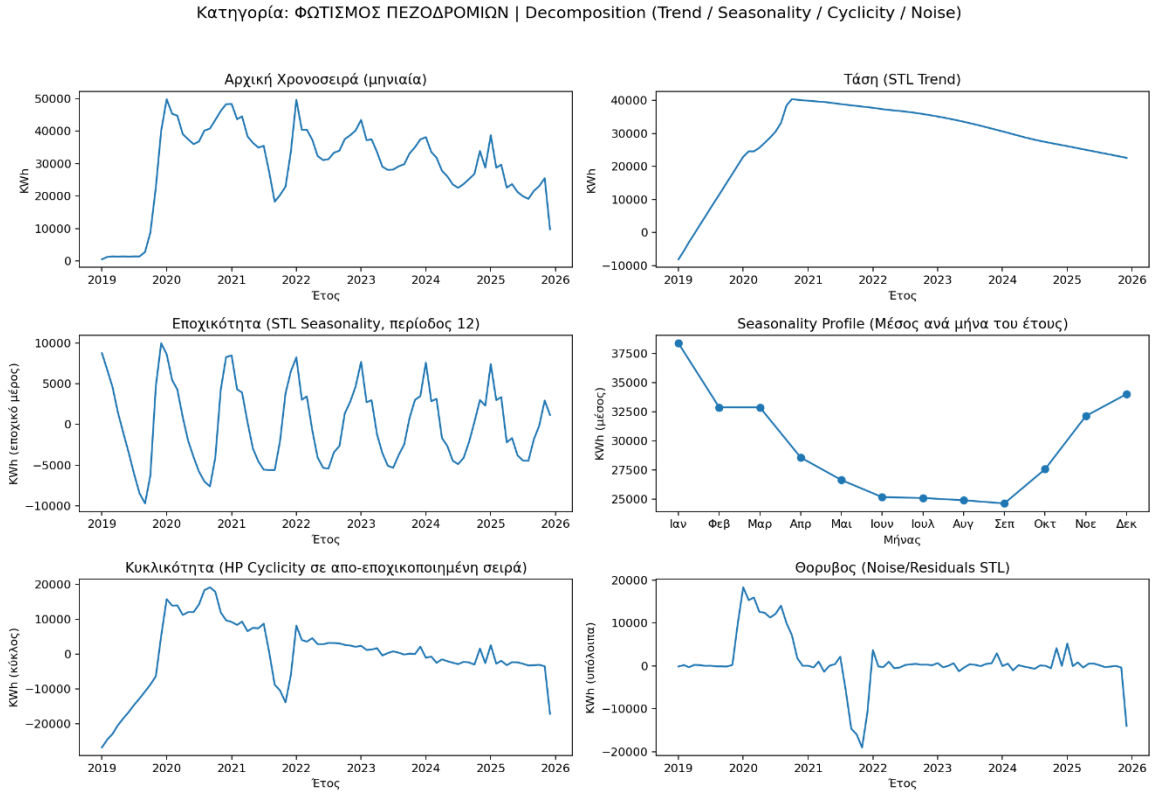
Κατηγορία: ΠΟΛΙΤΙΣΜΟΣ | Decomposition (Trend / Seasonality / Cyclicity / Noise)

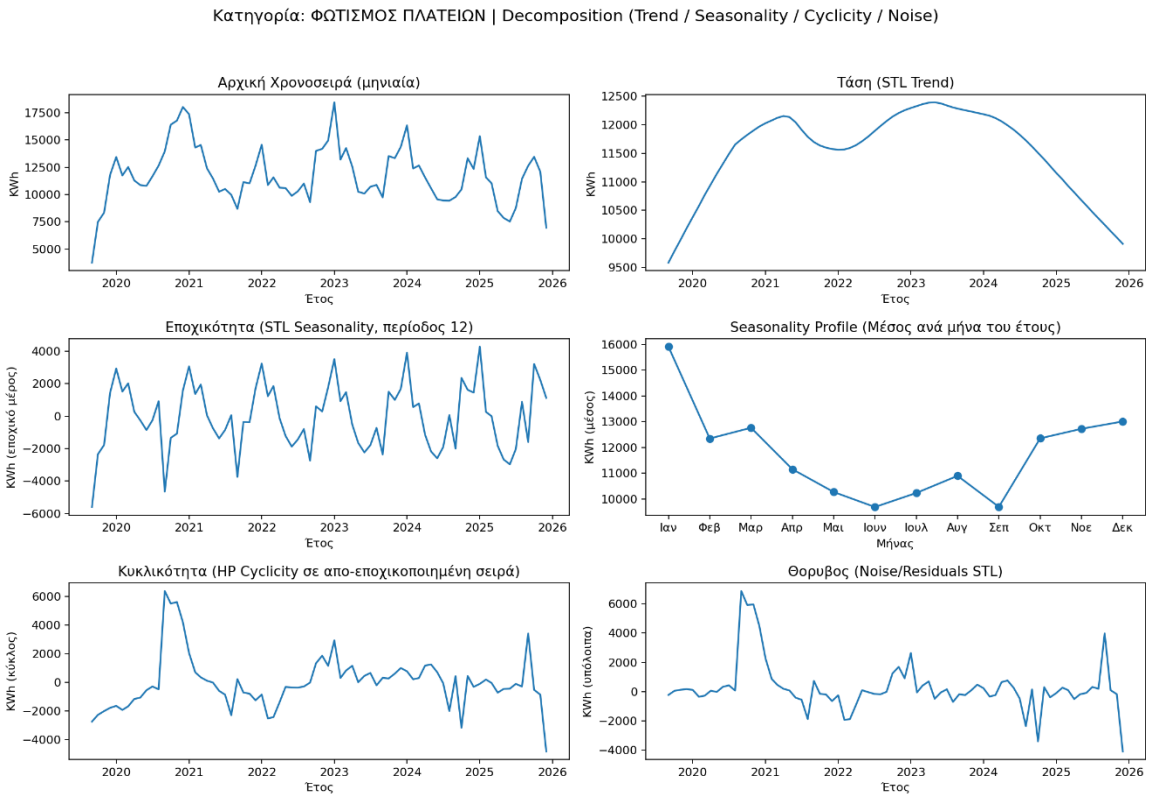
Κατηγορία: ΣΧΟΛΕΙΑ Β'ΒΑΘΜΙΑΣ ΕΚΠΑΙΔΕΥΣΗΣ | Decomposition (Trend / Seasonality / Cyclicity / Noise)



Κατηγορία: ΥΠΟΔΟΜΕΣ ΠΡΟΣΧΟΛΙΚΗΣ ΑΓΩΓΗΣ | Decomposition (Trend / Seasonality / Cyclicity / Noise)


Κατηγορία: ΦΩΤΙΣΜΟΣ ΟΔΩΝ ΠΛΑΤΕΙΩΝ | Decomposition (Trend / Seasonality / Cyclicity / Noise)
Αρχική Χρονοσειρά (μηνιαία)
Τάση (STL Trend)
Εποχικότητα (STL Seasonality, περίοδος 12)
Seasonality Profile (Μέσος ανά μήνα του έτους)
Κυκλικότητα (HP Cyclicity σε απο-εποχικοποιημένη σειρά)
Θορυβος (Noise/Residuals STL)
KWh
Έτος
Μήνας
Ιαν
Φεβ
Μαρ
Απρ
Μαι
Ιουν
Ιουλ
Αυγ
Σεπ
Οκτ
Νοε
Δεκ


Κατηγορία: ΦΩΤΙΣΜΟΣ ΠΕΖΟΔΡΟΜΙΩΝ | Decomposition (Trend / Seasonality / Cyclicity / Noise)
Αρχική Χρονοσειρά (μηνιαία)
Τάση (STL Trend)
Εποχικότητα (STL Seasonality, περίοδος 12)
Seasonality Profile (Μέσος ανά μήνα του έτους)
Κυκλικότητα (HP Cyclicity σε απο-εποχικοποιημένη σειρά)
Θορυβος (Noise/Residuals STL)
KWh
Έτος
Μήνας
Ιαν
Φεβ
Μαρ
Απρ
Μαι
Ιουν
Ιουλ
Αυγ
Σεπ
Οκτ
Νοε
Δεκ



Εικόνα 3.46 Trend/Seasonality/Cyclicity/Noise ανά κατηγορία υποδομής

3.6.4 Προετοιμασία τελικού αρχείου δεδομένων για μηχανική μάθηση

Οι ενέργειες που υλοποιήθηκαν στις ενότητες 3.4 έως και 3.6.2, είχαν ως σκοπό την δημιουργία ενός τελικού αρχείου δεδομένων (data.csv) το οποίο περιλαμβάνει εγγραφές που αντιστοιχούν στους πραγματικούς μηνιαίους λογαριασμούς κατανάλωσης ηλεκτρικής ενέργειας των δημοτικών εγκαταστάσεων. Κάθε εγγραφή περιγράφεται από τον αριθμό παροχής, το έτος και τον μήνα κατανάλωσης, το είδος του τιμολογίου, το όνομα πολλαπλού καθώς και την αντίστοιχη κατηγορία δημοτικής υποδομής, ενώ ως μεταβλητή-στόχος ορίζεται η μηνιαία κατανάλωση σε KWh.

Οι παροχές ηλεκτρικού ρεύματος ανέρχονται σε 868. Μετά την τεχνητή ημερήσια κατανομή της κατανάλωσης (με βάση την ημερομηνία έκδοσης του αρχικού λογαριασμού σε συνάρτηση με τις ημέρες κατανάλωσης που αντιστοιχούν σε αυτόν) πραγματοποιήθηκε εκ νέου ανακατασκευή της μηνιαίας κατανάλωσης με άθροιση της ημερήσιας κατανάλωσης ανά παροχή και μήνα για κάθε έτος. Η διαδικασία αυτή συνοψίζει την λεπτομερή ημερήσια κατανάλωση σε υψηλότερο επίπεδο (Han, Kamber, & Pei, 2011), με αποτέλεσμα τη

δημιουργία ενός αρχείου με συνεπή μηνιαία αναφορά (data_per_month.csv) με 61.935 εγγραφές (Εικόνα 3.22).

Στο τελικό αρχείο data.csv (Ενότητα 3.6.2) διατηρούνται μόνο τα πλέον σημαντικά χαρακτηριστικά (στήλες) ανά εγγραφή (παρατήρηση), όπως αυτά προέκυψαν από τον έλεγχο σημαντικότητας που προηγήθηκε στο αρχείο των πρωτογενών δεδομένων και τα οποία καθορίζουν την συμπεριφορά της ενεργειακής κατανάλωσης.

Το πλήθος των 61.935 εγγραφών που αντιστοιχούν στη μεταβλητή-στόχο «Μηνιαία κατανάλωση KWh» υποδηλώνει ότι υπάρχουν παροχές που δεν εμφανίζονται σε όλους τους μήνες κάθε έτους. Το τελικό αρχείο data.csv καλύπτει, με βάση τις μοναδικές τιμές των χαρακτηριστικών «Έτος κατανάλωσης» και «Μήνας κατανάλωσης» (Εικόνα 3.20 & 3.22), 84 μήνες (7*12) δηλαδή το χρονικό διάστημα από 01-01-2019 έως και 31-12-2025 (Ενότητα 3.5). Εάν όλες οι παροχές εμφάνιζαν κατανάλωση για το σύνολο των 84 μηνών τότε το πλήθος των εγγραφών του αρχείου θα ανέρχονταν σε 868*84= 72.912. Συνεπώς λείπουν 10.977 μηνιαίες παρατηρήσεις (72.912 – 61.935), με αποτέλεσμα να δημιουργείται ασυνέχεια στη χρονική ροή των δεδομένων ανά παροχή.

Η πρόβλεψη της μηνιαίας ενεργειακής κατανάλωσης, πρώτα ανά αριθμό παροχής και εν συνεχεία αθροιστικά ανά κατηγορία δημοτικής υποδομής, μπορεί να υλοποιηθεί με αλγόριθμους επιβλεπόμενης μάθησης (supervised learning) που επιλύουν προβλήματα παλινδρόμησης σε χρονοσειρές. Επιβλεπόμενη, γιατί είναι γνωστή η τιμή της εξαρτημένης μεταβλητής-στόχου y = «Μηνιαία κατανάλωση KWh» ανά αριθμό παροχής από τα ιστορικά δεδομένα και αθροιστικά ανά κατηγορία υποδομής, που είναι και το ζητούμενο για επιχειρησιακούς λόγους. Στην επιβλεπόμενη μηχανική μάθηση κατασκευάζεται μια συνάρτηση f (x) συσχέτισης των ανεξάρτητων μεταβλητών εισόδου X για την πρόβλεψη της εξαρτημένης μεταβλητής-στόχου y.

Βασικοί παράγοντες στις χρονοσειρές για την εύρεση κατάλληλων εισόδων X σε μοντέλα μηχανικής μάθησης, είναι η αποτύπωση της χρονικής αυτοσυσχέτισης (εξάρτησης) μεταξύ των διαδοχικών παρατηρήσεων, ο εντοπισμός της εποχικότητας, η τοπική συμπεριφορά της κατανάλωσης σε οριοθετημένα μικρά χρονικά διαστήματα και η διαφοροποίηση της κατανάλωσης ανά κατηγορία υποδομής. Η προσέγγιση αυτή επιβάλει την δημιουργία νέων χαρακτηριστικών (feature engineering) που θα αποτυπώνουν επιπλέον χρονικές πληροφορίες στο τελικό αρχείο data.csv. Η προσθήκη νέων κατάλληλων χαρακτηριστικών

αποτελεί βασική παράμετρο για τη βελτίωση της απόδοσης των μοντέλων μηχανικής μάθησης. Οι τιμές τους υπολογίζονται αποκλειστικά από τις τιμές συγκεκριμένων αρχικών χαρακτηριστικών που επιλέχθηκαν ως τα πιο σημαντικά (Ενότητα 3.6.2).

Σκοπός λοιπόν είναι η δημιουργία κατάλληλων εισόδων X ως ανεξάρτητων μεταβλητών που θα τροφοδοτήσουν τους αλγόριθμους μηχανικής μάθησης για την πρόβλεψη της μεταβλητής-στόχου «Μηνιαία κατανάλωση KWh» επιτυγχάνοντας την μέγιστη δυνατή απόδοση.

Τονίζεται ότι η συγκεκριμένη ενέργεια δεν αποτελεί επανάληψη της αρχικής προεπεξεργασίας-μετατροπής των αρχικών δεδομένων που περιγράφεται στις ενότητες 3.4 και 3.5. Εκεί ο στόχος ήταν η μετατροπή των λογαριασμών με βάση την ημερομηνία έκδοσης και τις ημέρες κατανάλωσης σε συνεπή μηνιαία αποτύπωση. Εδώ ο στόχος είναι η δημιουργία κατάλληλων εισόδων X για την εισαγωγή τους σε αλγόριθμους επιβλεπόμενης μάθησης (Kuhn & Johnson, 2013).

Στη συνέχεια παρουσιάζονται συνοπτικά οι βασικές ενέργειες που υλοποιούνται μέσω κατάλληλων συναρτήσεων σε Python για την δημιουργία ενός πλήρους συνόλου δεδομένων που οργανώνονται ως πολλαπλές χρονοσειρές (panel time series):

1. Από τα χαρακτηριστικά «Ετος κατανάλωσης» και «Μήνας κατανάλωσης» δημιουργείται η πρώτη ημέρα κάθε μήνα (date index) για τη σωστή χρονολογική ταξινόμηση των δεδομένων και για τον ασφαλή διαχωρισμό σε σύνολο εκπαίδευσης και ελέγχου.
2. Κατασκευή πλήρους πλέγματος μηνών ανά έτος (month/year) για όλες τις παροχές ρεύματος με σκοπό την διόρθωση της ασυνέχειας στις χρονοσειρές των ενεργειακών δεδομένων και με συμπλήρωση ως NaN της μεταβλητής-στόχου «Μηνιαία κατανάλωση KWh» για τους μήνες όπου δεν υπάρχει εγγραφή (έλλειψη παρατήρησης).
3. Προσθήκη χαρακτηριστικών εποχικότητας μέσω κυκλικής κωδικοποίησης των μηνών: month_sin=sin(2*π*month/12) και month_cos=(2*π*month/12), για την αποφυγή της ασυνέχειας που προκαλείται στο τέλος και την αρχή ενός έτους για κάθε παροχή. Η μετατροπή των ακέραιων τιμών (1,..,12) του χαρακτηριστικού «Μήνας κατανάλωσης» σε μορφή γωνίας φ = (2*π*month/12) διατηρεί στον ετήσιο κύκλο τη χρονική συνέχεια των μηνών. Ο 12^{ος} μήνας (Δεκέμβριος) είναι χρονικά κοντά στον 1^{ο} μήνα (Ιανουάριος).
4. Προσθήκη χαρακτηριστικών που αποτυπώνουν προηγούμενες χρονικά τιμές της κατανάλωσης (lags features) μιας παροχής (lag_1, lag_2, lag_3, lag_6) και την κατανάλωση

του ίδιου μήνα του προηγούμενου έτους (lag_12). Η αποτύπωση της εξάρτησης της τρέχουσας τιμής της κατανάλωσης από προηγούμενες χρονικά τιμές της (χρονική αυτοσυσχέτιση) ενισχύει την ικανότητα μάθησης των αλγορίθμων περιορίζοντας το φαινόμενο της διαρροής πληροφορίας (data leakage).

5. Προσθήκη στατιστικών χαρακτηριστικών που περιγράφουν την τοπική συμπεριφορά της ενεργειακής κατανάλωσης ανά παροχή, με δείκτες όπως ο κυλιόμενος μέσος όρος (rolling_mean) σε παράθυρα (rolling windows) που περιλαμβάνουν την μέση κατανάλωση των προηγούμενων 3, 6 και 12 μηνών (rolling_mean_3, rolling_mean_6, rolling_mean_12) και βελτιώνουν τη σταθερότητα των προβλέψεων.

Μετά την προσθήκη της απαραίτητης χρονικής πληροφορίας (χαρακτηριστικών) **ανά παροχή**, παράγεται ένα ενδιάμεσο αρχείο (add_lag_rolling_features.csv) με το σύνολο των αρχικών και των νέων χαρακτηριστικών (Εικόνα 3.47).

Το αρχείο αυτό έχει την δομή που απαιτείται για την εκπαίδευση και εφαρμογή αλγορίθμων επιβλεπόμενης μηχανικής μάθησης σε χρονοσειρές ενεργειακών δεδομένων.

Πληροφορίες ενδιάμεσου αρχείου δεδομένων

```
<class 'pandas.core.frame.DataFrame'>
Index: 72912 entries, 15540 to 64931
Data columns (total 23 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   Αρ.Παροχής             72912 non-null  object
 1   Ετος κατανάλωσης       61935 non-null  float64
 2   Μήνας κατανάλωσης      61935 non-null  float64
 3   Μηνιαία κατανάλωση KWh 61935 non-null  float64
 4   Ονομα Πολλαπλού        62999 non-null  object
 5   Τιμολόγιο              62999 non-null  object
 6   Κατηγορία Υποδομής     62999 non-null  object
 7   date                    72912 non-null  datetime64[ns]
 8   year                    72912 non-null  int64
 9   month                   72912 non-null  int64
 10  month_sin               72912 non-null  float64
 11  month_cos               72912 non-null  float64
 12  lag_1                   61103 non-null  float64
 13  lag_2                   60264 non-null  float64
 14  lag_3                   59425 non-null  float64
 15  lag_6                   56926 non-null  float64
 16  lag_12                  51892 non-null  float64
 17  roll_mean_3             59163 non-null  float64
 18  roll_std_3              59163 non-null  float64
 19  roll_mean_6             56312 non-null  float64
 20  roll_std_6              56312 non-null  float64
 21  roll_mean_12            50645 non-null  float64
 22  roll_std_12             50645 non-null  float64
dtypes: datetime64[ns](1), float64(16), int64(2), object(4)
```

Πλήθος μοναδικών τιμών

```
Αρ.Παροχής              868
Ετος κατανάλωσης          7
Μήνας κατανάλωσης        12
Μηνιαία κατανάλωση KWh 31952
Ονομα Πολλαπλού           25
Τιμολόγιο                  8
Κατηγορία Υποδομής        24
date                      84
year                       7
month                     12
month_sin                 11
month_cos                 11
lag_1                  31579
lag_2                  31091
lag_3                  30554
lag_6                  29006
lag_12                 25820
roll_mean_3            47587
roll_std_3             52077
roll_mean_6            49274
roll_std_6             50754
roll_mean_12           44915
roll_std_12            45534
dtype: int64
```

Εικόνα 3.47 Πληροφορίες και μοναδικές τιμές τελικού αρχείου δεδομένων

Από τα δεδομένα του παραγόμενου ενδιάμεσου αρχείου διατηρούνται μόνο εκείνες οι παρατηρήσεις (61.935 εγγραφές) για τις οποίες υπάρχει τιμή στη μεταβλητή-στόχο y = «Μηνιαία κατανάλωση KWh». Οι παρατηρήσεις ταξινομούνται χρονολογικά και διαχωρίζονται με βάση την πρώτη ημέρα κάθε μήνα (date index) σε σύνολο εκπαίδευσης και ελέγχου ως εξής:

- Σύνολο εκπαίδευσης (training set) από 01-01-2019 έως 31-12-2024, παρέχοντας ένα επαρκές σύνολο ιστορικού βάθους 72 μηνών που αποτελείται από 51.892 εγγραφές.
- Σύνολο ελέγχου (test set) από 01-1-2025 έως 31-12-2025, πλήρως άγνωστο στους αλγόριθμους που θα χρησιμοποιηθούν για την πρόβλεψη της μηνιαίας κατανάλωσης σε KWh που αποτελείται από 10.043 εγγραφές.

3.7 Εφαρμογή Αλγορίθμων Μηχανικής Μάθησης (ML)

Στόχος της εφαρμογής των αλγορίθμων μηχανικής μάθησης είναι η πρόβλεψη της μηνιαίας ενεργειακής κατανάλωσης **πρώτα ανά αριθμό παροχής και εν συνεχεία αθροιστικά για κάθε κατηγορία δημοτικής υποδομής**. Οι 868 παροχές που αντιστοιχούν σε 754 μοναδικούς πελάτες-καταναλωτές έχουν αντιστοιχιστεί σε 24 κατηγορίες δημοτικών υποδομών. Το χαρακτηριστικό «Αρ.Παροχής» έχει διατηρηθεί σε όλους τους μετασχηματισμούς των δεδομένων που προηγήθηκαν, διότι μας παρέχει άμεσα την πληροφορία της κατανάλωσης σε επίπεδο δημοτικής εγκατάστασης και όχι σε επίπεδο πελάτη ο οποίος μπορεί να έχει περισσότερες από μια παροχές. Αυτό έχει ως αποτέλεσμα οι παρατηρήσεις της κατανάλωσης να είναι περισσότερες στο τελικό αρχείο συνεπώς αυξάνεται η συνεισφορά τους στο τελικό αποτέλεσμα της κατηγορίας στην οποία ανήκουν.

Τονίζεται ότι η προσέγγιση της εφαρμογής μοντέλων μηχανικής μάθησης ανά παροχή θα οδηγούσε σε μέγιστη εξατομικευμένη προσαρμογή αλλά με πολύ υψηλό υπολογιστικό κόστος εκπαίδευσης και παρουσίασης των αποτελεσμάτων (ενότητα 3.4.4).

Αυτό που έχει αξία για την υποστήριξη στην λήψη των αποφάσεων της διοίκησης ενός οργανισμού τοπικής αυτοδιοίκησης είναι η συνολική εικόνα της ενεργειακής συμπεριφοράς κάθε λειτουργικού τομέα των δημοτικών υποδομών (σχολεία, αθλητικές εγκαταστάσεις, κτλ) και ενδεχομένως η περαιτέρω εξειδίκευση ανά συγκεκριμένη δημοτική υποδομή/εγκατάσταση. Με βάση την ανωτέρω προσέγγιση τα μοντέλα που

χρησιμοποιούνται εκπαιδεύονται στο σύνολο των παροχών (global model) ενώ η αποτύπωση των αποτελεσμάτων γίνεται ανά μήνα και κατηγορία υποδομής.

Η πρόβλεψη της μηνιαίας κατανάλωσης πρώτα σε επίπεδο παροχής και εν συνεχεία αθροιστικά ανά κατηγορία δημοτικής υποδομής (bottom-up) αποτελεί ιεραρχική ομαδοποιημένη πρόβλεψη (Athanasopoulos & Kourentzes, 2023).

3.7.1 Προεπεξεργασία μετασχηματισμού δεδομένων εισόδου

Η απαραίτητη προεπεξεργασία μετασχηματισμού των δεδομένων σε κατάλληλη μορφή για την εισαγωγή τους σε μοντέλα μηχανικής μάθησης υλοποιείται μέσω της βιβλιοθήκης scikit-learn της Python. Τα χαρακτηριστικά εισόδου X διαχωρίζονται σε:

- **αριθμητικά** (year, month, month_sin, month_cos, lag_1, lag_2, lag_3, lag_6, lag_12, roll_mean_3, roll_std_3, roll_mean_6, roll_std_6, roll_mean_12, roll_std_12) και
- **κατηγορικά** (Αρ.Παροχής, Τιμολόγιο, Όνομα Πολλαπλού, Κατηγορία Υποδομής)

Η απαραίτητη προεπεξεργασία μετασχηματισμού γίνεται μέσω της κλάσης PipeLine ξεχωριστά για τα αριθμητικά και τα κατηγορικά χαρακτηριστικά και στο τέλος ενιαία μέσω της κλάσης ColumnTransformer.

Στα αριθμητικά χαρακτηριστικά μέσω της PipeLine οι ελλιπείς τιμές λόγω μη ύπαρξης μηνών κατανάλωσης αντικαθίστανται από τη διάμεσο (median) του κάθε χαρακτηριστικού μέσω της κλάσης SimpleImputer(strategy="median"), ενώ ταυτόχρονα υλοποιείται κλιμάκωση (scaling) στις τιμές τους μέσω της κλάσης StandardScaler που είναι απαραίτητη σε γραμμικά μοντέλα αλγορίθμων μηχανικής μάθησης για λόγους σύγκρισης.

Στα κατηγορικά χαρακτηριστικά μέσω της PipeLine, οι ελλιπείς τιμές αντικαθίστανται από την τιμή που εμφανίζεται πιο συχνά (most_frequent) σε κάθε χαρακτηριστικό, μέσω της κλάσης SimpleImputer(strategy="most_frequent"), ενώ ταυτόχρονα υλοποιείται κωδικοποίηση αυτών μέσω της κλάσης OneHotEncoder ώστε να προκύψουν οι απαιτούμενοι αριθμητικοί πίνακες για τους αλγόριθμους μηχανικής μάθησης.

Το τελικό αποτέλεσμα, όπως ήδη αναφέρθηκε, παράγεται μέσω της κλάσης ColumnTransformer που επιτρέπει την ταυτόχρονη προεπεξεργασία αριθμητικών και κατηγορικών χαρακτηριστικών επιστρέφοντας έναν ενιαίο πίνακα μετασχηματισμένων δεδομένων. Στη συνέχεια μέσω της κλάσης PipeLine δημιουργείται ενιαίο αντικείμενο ροής

για την ταυτόχρονη προεπεξεργασία και εφαρμογή μοντέλων μηχανικής μάθησης, ώστε να υλοποιούνται ακριβώς τα ίδια βήματα τόσο κατά την εκπαίδευση όσο και κατά την πρόβλεψη.

```python
# Συνάρτηση προεπεξεργασίας και μετασχηματισμού των δεδομένων
def build_preprocessor(numeric_cols: List[str], categorical_cols: List[str]) -> ColumnTransformer:
    num_pipe = Pipeline(steps=[
        ("imputer", SimpleImputer(strategy="median")),
        ("scaler", StandardScaler())
    ])
    cat_pipe = Pipeline(steps=[
        ("imputer", SimpleImputer(strategy="most_frequent")),
        ("ohe", OneHotEncoder(handle_unknown="ignore"))
    ])
    return ColumnTransformer(
        transformers=[
            ("num", num_pipe, numeric_cols),
            ("cat", cat_pipe, categorical_cols),
        ],
        remainder="drop"
    )

# Συνάρτηση δημιουργίας αντικειμένου ροής (προεπεξεργασία+μοντέλο)
def make_pipeline(model, numeric_cols, categorical_cols, needs_dense: bool) -> Pipeline:
    pre = build_preprocessor(numeric_cols, categorical_cols)
    if needs_dense:
        return Pipeline(steps=[("prep", pre), ("dense", ToDense()), ("model", model)])
    return Pipeline(steps=[("prep", pre), ("model", model)])
```

Εικόνα 3.48 Προεπεξεργασία μετασχηματισμού των δεδομένων

3.7.2 Επιλογή Αλγορίθμων μηχανικής μάθησης

Η επιλογή έγινε με γνώμονα τη χρήση αλγορίθμων από όλες τις βασικές οικογένειες με σκοπό να εξαχθούν συγκριτικά συμπεράσματα για την ανάδειξη του καταλληλότερου προβλεπτικού μοντέλου. Με δεδομένο ότι τα μετασχηματισμένα πλέον χρονικά δεδομένα εμφανίζουν έντονη εποχικότητα, μακροπρόθεσμη τάση και μη γραμμικές σχέσεις (ενότητα 3.6.3) επιλέχθηκαν 12 αλγόριθμοι από διαφορετικές οικογένειες μοντέλων, οι οποίοι μαθαίνουν τις σχέσεις και την αλληλεπίδραση των μεταβλητών εισόδου X με διαφορετική μεθοδολογία.

Οι αλγόριθμοι υλοποιούνται μέσω κατάλληλων κλάσεων της βιβλιοθήκης scikit-learn στην Python. Παρακάτω δίνεται μια συνοπτική περιγραφή της λειτουργίας τους.

Γραμμικά μοντέλα με κανονικοποίηση (Regularized Linear Models):

Ridge()

Το Ridge Regression είναι ένα γραμμικό μοντέλο παλινδρόμησης στο οποίο η πρόβλεψη της κατανάλωσης y_{pred} για κάθε παροχή και μήνα υπολογίζεται ως γραμμικός συνδυασμός των χαρακτηριστικών εισόδου X. Το σύνολο των εισόδων X περιλαμβάνει αριθμητικά χαρακτηριστικά όπως lags, rolling στατιστικά, εποχικότητα sin/cos και κατηγορικά τα οποία μετατρέπονται σε αριθμητική μορφή μέσω one-hot κωδικοποίηση.

Η εκπαίδευση του μοντέλου βασίζεται στην ελαχιστοποίηση της συνάρτηση κόστους:

$$\| y_{true} - Xw \|_2^2 + \alpha \| w \|_2^2$$

Ο πρώτος όρος $\| y_{true} - Xw \|_2^2$ εκφράζει το άθροισμα των τετραγωνικών σφάλματων (τετράγωνο ευκλείδειας απόστασης) μεταξύ των πραγματικών τιμών κατανάλωσης y_{true} και των προβλέψεων $y_{pred} = Xw$, αποτυπώνοντας πόσο κοντά βρίσκονται συνολικά οι προβλέψεις του μοντέλου στις πραγματικές τιμές της κατανάλωσης.

Ο δεύτερος όρος $\alpha \| w \|_2^2$ αποτελεί την L2 κανονικοποίηση του διανύσματος βαρών w. Η κανονικοποίηση λειτουργεί ως ποινή στη συνάρτηση κόστους. Τα βάρη w_j για κάθε χαρακτηριστικό εισόδου X_j προκύπτουν από την επίλυση ενός κανονικοποιημένου προβλήματος ελαχίστων τετραγώνων και λειτουργούν ως συντελεστές βαρύτητας που καθορίζουν την συμβολή του κάθε χαρακτηριστικού στην τελική πρόβλεψη της κατανάλωσης. Με αυτόν τον τρόπο το μοντέλο περιορίζει την υπερβολική αύξηση των συντελεστών, κάτι που είναι ιδιαίτερα χρήσιμο όταν τα χαρακτηριστικά εισόδου X_j (όπως στην δική μας περίπτωση) είναι πολλά και συσχετισμένα (π.χ. lags/rolling) και όταν το πλήθος τους αυξάνεται λόγω One-Hot κωδικοποίησης (π.χ. χαρακτηριστικό «Αρ. Παροχής»).

Η υπερπαράμετρος α ρυθμίζει τον συμβιβασμό μεταξύ ακρίβειας προσαρμογής (εκπαίδευση σε ιστορικά δεδομένα) και σταθερότητας των προβλέψεων της κατανάλωσης (σε νέα άγνωστα δεδομένα). Μικρές τιμές α δίνουν πιο ευέλικτο μοντέλο ενώ μεγαλύτερες τιμές α δίνουν πιο συντηρητικό και σταθερό μοντέλο το οποίο τείνει να γενικεύει καλύτερα σε νέα άγνωστα δεδομένα (νέοι μήνες) (Kuhn & Johnson, 2013).

Elastic Net()

Η βασική διαφορά του **Elastic Net** σε σχέση με το **Ridge** είναι ότι η εκπαίδευσή του βασίζεται στην ελαχιστοποίηση μιας συνάρτησης κόστους που συνδυάζει σφάλμα προσαρμογής και διπλή κανονικοποίηση (L1 & L2):

$$\| y_{true} - Xw \|_2^2 + \alpha \left((1-\rho) \frac{\| w \|_2^2}{2} + \rho \| w \|_1 \right)$$

Ο πρώτος όρος έχει αναλυθεί προηγουμένως και εκφράζει το άθροισμα τετραγωνικών σφαλμάτων μεταξύ των πραγματικών τιμών και των προβλέψεων. Ο δεύτερος όρος αποτελεί τον όρο κανονικοποίησης και ενσωματώνει ταυτόχρονα:

- L2 (Ridge) κανονικοποίηση $\| w \|_2^2 = \sum_j w_j^2$ υπολογίζει το άθροισμα των τετραγώνων των βαρών του μοντέλου, περιορίζοντας την αύξηση των συντελεστών βάρους w μειώνοντας την πιθανότητα υπερπροσαρμογής και βελτιώνοντας τη σταθερότητα όταν υπάρχουν συσχετισμένα χαρακτηριστικά εισόδου.
- L1 (Lasso) κανονικοποίηση $\| w \|_1 = \sum_j | w_j |$ υπολογίζει το άθροισμα των απόλυτων τιμών των συντελεστών (δηλαδή επιβάλλει γραμμική ποινή, όχι τετραγωνική). Εξαιτίας αυτής της μορφής, τείνει να οδηγεί μερικούς συντελεστές ακριβώς στο μηδέν λειτουργώντας έτσι και ως μηχανισμός επιλογής χαρακτηριστικών (feature selection).

Οι υπερπαράμετροι α και ρ (συνήθως l1_ratio) ρυθμίζουν αντίστοιχα το βαθμό της κανονικοποίησης και την αναλογία μεταξύ L2 και L1. Όταν $\rho \rightarrow 0$ το μοντέλο προσεγγίζει την κανονικοποίηση τύπου L2 (Ridge) ενώ όταν $\rho \rightarrow 1$ προσεγγίζει την L1(Lasso).

Δέντρα Αποφάσεων (decision trees)

Random Forest

Η λειτουργία του αλγόριθμου Random Forest έχει αναλυθεί διεξοδικά στην ενότητα 3.6

ExtraTrees

Ο αλγόριθμος ExtraTrees, όπως και ο Random Forest, βασίζεται στη δημιουργία πολλών δέντρων απόφασης, ενώ η τελική πρόβλεψη της μηνιαίας κατανάλωσης προκύπτει ως ο μέσος όρος των εκτιμήσεων όλων των δέντρων. Μια βασική διαφορά σε σχέση με τον Random Forest είναι ότι εξ ορισμού στο scikit-learn ο Random Forest εκπαιδεύει κάθε δέντρο με επαναδειγματοληψία (bootstrap) στο δικό του τυχαίο δείγμα παρατηρήσεων του εκπαιδευτικού συνόλου (bootstrap=True), ενώ στο ExtraTrees από προεπιλογή χρησιμοποιείται ολόκληρο το εκπαιδευτικό σύνολο για κάθε δέντρο (bootstrap=False). Και τα δυο μοντέλα εξετάζουν σε κάθε κόμβο ένα τυχαίο δείγμα χαρακτηριστικών όμως

διαφέρουν στον τρόπο επιλογής των κανόνων διάσπασης (split). Ο Random Forest επιλέγει επαναληπτικά το βέλτιστο κατώφλι (threshold) διάσπασης για τα υποψήφια χαρακτηριστικά που οδηγεί στη μέγιστη μείωση της συνάρτησης κόστους (MSE). Αντίθετα στο ExtraTrees το κατώφλι (κανόνας) διάσπασης στα υποψήφια χαρακτηριστικά παράγεται τυχαία και επιλέγεται εκείνο που μειώνει στο μέγιστο την ίδια συνάρτηση κόστους.

HistGB

Το μοντέλο κατασκευάζει διαδοχικά δέντρα απόφασης με σκοπό την ενίσχυση (boosting) της ικανότητας πρόβλεψης. Σε κάθε επανάληψη το νέο δέντρο εκπαιδεύεται ώστε να μάθει και να διορθώσει τα σφάλματα πρόβλεψης που προκύπτουν από το άθροισμα των εκτιμήσεων όλων των προηγούμενων δέντρων. Πριν από την κατασκευή των δέντρων τα χαρακτηριστικά με συνεχείς τιμές διακριτοποιούνται (binning) και αντιστοιχίζονται σε ακέραιους κωδικούς (bins). Για κάθε χαρακτηριστικό σχηματίζεται ένα ιστόγραμμα συχνοτήτων ανά κωδικό (bin) και η αναζήτηση των κανόνων διάσπασης στους κόμβους γίνεται πλέον με βάση τα όρια των κωδικών (bins) και όχι σε όλες τις μοναδικές συνεχείς τιμές. Η προσέγγιση αυτή μειώνει την υπολογιστική πολυπλοκότητα της διαδικασίας εύρεσης κανόνων διάσπασης και οδηγεί σε μείωση του χρόνου εκπαίδευσης.

GBDT

Η λειτουργία του Gradient Boosting Decision Trees είναι παρόμοια με αυτή του HistGB με τη διαφορά ότι δεν γίνεται διακριτοποίηση των χαρακτηριστικών με συνεχείς τιμές. Αυτό σημαίνει ότι η αναζήτηση των κανόνων διάσπασης στους κόμβους των δέντρων, γίνεται με βάση τις αρχικές τιμές των χαρακτηριστικών αυξάνοντας την υπολογιστική πολυπλοκότητα και συνεπώς τον χρόνο εκπαίδευσης του αλγορίθμου.

Οι αλγόριθμοι που βασίζονται σε δέντρα απόφασης μπορούν να εντοπίζουν μη γραμμικές σχέσεις και εμφανίζουν καλύτερη προσαρμογή πρόβλεψης.

Μηχανές Διανυσμάτων Υποστήριξης - SVM

LinearSVR

Ο αλγόριθμος Linear Support Vector Regression είναι ένας γραμμικός αλγόριθμος παλινδρόμησης της οικογένειας SVM για την πρόβλεψη συνεχών τιμών όπως η «Μηνιαία κατανάλωση KWh». Μέσω μιας συνάρτησης πρόβλεψης ($y_{pred} = w \cdot X + b$) προσπαθεί να προσδιορίσει μια **βέλτιστη ευθεία** (υπερεπίπεδο) που περιγράφει τη συσχέτιση των

χαρακτηριστικών εισόδου X - αριθμητικά lags, rolling στατιστικά, δείκτες εποχικότητας sin/cos και κατηγορικά μέσω one-hot - με τη μεταβλητή-στόχο «Μηνιαία κατανάλωση KWh». Γύρω από την ευθεία ορίζεται μια ζώνη πλάτους ε (epsilon) η οποία εκφράζει το αποδεκτό περιθώριο σφάλματος. Η βέλτιστη ευθεία είναι εκείνη που ενσωματώνει όσο το δυνατόν περισσότερες παρατηρήσεις εντός της ζώνης ανοχής ε μέσα στην οποία τα σφάλματα αγνοούνται. Οι παρατηρήσεις που βρίσκονται στα όρια ή εκτός της ζώνης ονομάζονται διανύσματα υποστήριξης - **support vectors.** Τα διανύσματα υποστήριξης είναι οι παρατηρήσεις που επηρεάζουν περισσότερο τον προσδιορισμό της ευθείας δηλαδή της πρόβλεψης. Η παράμετρος C καθορίζει τη μείωση μεγάλων σφαλμάτων πρόβλεψης (μεγάλο C) και τη σταθερότητα γενίκευσης (μικρό C) του μοντέλου. Στην παρούσα εφαρμογή, όπου έχει πραγματοποιηθεί κλιμάκωση στα χαρακτηριστικά, το LinearSVR λειτουργεί ως αποδοτικό γραμμικό μοντέλο που αξιοποιεί την πληροφορία των feature engineered χαρακτηριστικών για την πρόβλεψη σε επίπεδο παροχής.

SVR_RBF

Ο αλγόριθμος Support Vector Regression με πυρήνα RBF (Radial Basis Function) σε αντίθεση με τον LinearSVR εντοπίζει μη-γραμμικές σχέσεις μεταξύ των εισόδων X με την μεταβλητή-στόχο «Μηνιαία κατανάλωση KWh». Ο αλγόριθμος κατά την εκπαίδευση ορίζει μια **καμπύλη πρόβλεψης** που ενσωματώνει εντός μιας ζώνης πλάτους ε (epsilon) γύρω από την καμπύλη παρατηρήσεις με αποδεκτό περιθώριο σφάλματος. Η παράμετρος C λειτουργεί με τον ίδιο ακριβώς τρόπο όπως και στον LinearSVR. Ο πυρήνας RBF χρησιμοποιεί μια συνάρτηση βάσης (basis function) που υπολογίζει την ομοιότητα μεταξύ μιας παρατήρησης και ενός διανύσματος υποστήριξης (support vector) με βάση την απόσταση τους στον χώρο των χαρακτηριστικών. Έτσι το μοντέλο μπορεί να μάθει καμπύλες/μη γραμμικά μοτίβα π.χ. αλληλεπιδράσεις χαρακτηριστικών με εποχικότητα.

Κ-Κοντινότεροι Γείτονες (k-Nearest Neighbours)

KNN_uniform

Ο αλγόριθμος κατά την εκπαίδευση δεν προσαρμόζει κάποια συνάρτηση πρόβλεψης με βάση το εκπαιδευτικό σύνολο, απλώς αποθηκεύει τις παρατηρήσεις. Για κάθε νέο διάνυσμα χαρακτηριστικών X_{new} (παρατήρηση) υπολογίζει την **Ευκλείδεια απόσταση** από όλες τις παρατηρήσεις του εκπαιδευτικού συνόλου X_i.

$$d(x_{\text{new}}, x_i) = \| x_{\text{new}} - x_i \|_2 = \sqrt{\sum_{j=1}^{p} (x_{\text{new},j} - x_{i,j})^2}$$

και επιλέγει το σύνολο $N_k(X_{new})$ με τους k πιο πλησιέστερους γείτονες. Σε όλους του *k* κοντινούς γείτονες αποδίδεται ίσο βάρος (weights="uniform"). Η πρόβλεψη της «Μηνιαίας κατανάλωσης KWh» είναι ο απλός μέσος όρος των αντίστοιχων τιμών της μηνιαίας κατανάλωσης των k γειτόνων.

$$y_{pred}(x_{new}) = \frac{1}{k} \sum_{i \in \mathcal{N}_k(x_{new})} y_{true\_i}$$

Ο συντελεστής *k* ρυθμίζει τον βαθμό εξομάλυνσης της πρόβλεψης. Μικρό *k (λίγοι γείτονες)* οδηγεί σε προβλέψεις που επηρεάζονται περισσότερο από θόρυβο και ακραίες τιμές ενώ μεγάλο *k* (*περισσότεροι γείτονες*) οδηγεί σε πιο σταθερές προβλέψεις.

KNN_distance

Η διαφορά από τον KNN_uniform είναι ότι ο αλγόριθμος αποδίδει διαφορετικό σταθμισμένο βάρος στους *k* κοντινούς γείτονες ανάλογα με το πόσο κοντά βρίσκονται στο νέο διάνυσμα χαρακτηριστικών X_{new}. Η τελική πρόβλεψη της μηνιαίας κατανάλωσης προκύπτει από τον σταθμισμένο μέσο όρο των τιμών της κατανάλωσης των *k* κοντινών γειτόνων.

Ο αλγόριθμος KNN βασίζεται σε αποστάσεις συνεπώς η κλιμάκωση/τυποποίηση των χαρακτηριστικών είναι κρίσιμη ώστε να μην υπερέχουν οι μεταβλητές με μεγαλύτερη κλίμακα στον υπολογισμό της απόστασης.

Νευρωνικά Δίκτυα (*Neural Networks*)

MLP_Adam / MLP_LBFGS

Οι αλγόριθμοι των νευρωνικών δικτύων βασίζουν τη λειτουργία τους σε επίπεδα (layers) όπου κάθε επίπεδο αποτελείται από πολλούς νευρώνες που λειτουργούν παράλληλα. Ο νευρώνας είναι μια υπολογιστική μονάδα που δέχεται ως είσοδο αριθμητικές τιμές και υπολογίζει ενδιάμεσες τιμές τις οποίες προωθεί ως είσοδο στο επόμενο επίπεδο.

Στον αλγόριθμο Multi-Layer Perceptron, είτε με βελτιστοποιητή Adam είτε με LBFGS, κάθε επίπεδο του δικτύου αποτελεί και ένα στάδιο μετασχηματισμού του διανύσματος των χαρακτηριστικών εισόδου X (παρατηρήσεων).

Ένα νευρωνικό δίκτυο αποτελείται από το επίπεδο εισόδου (input layer) ένα ή περισσότερα κρυφά επίπεδα (hidden layers) και το επίπεδο εξόδου (output layer).

Το επίπεδο εισόδου αναλαμβάνει μόνο την εισαγωγή των χαρακτηριστικών εισόδου X, δηλαδή τα κανονικοποιημένα (scaled) αριθμητικά και τα κωδικοποιημένα (one-hot) κατηγορικά χαρακτηριστικά, στο δίκτυο χωρίς κάποιο υπολογισμό.

Στα κρυφά ενδιάμεσα επίπεδα κάθε νευρώνας υπολογίζει ένα γραμμικό άθροισμα της μορφής $z=w \cdot x+b$, όπου w είναι τα βάρη που καθορίζουν τη συμβολή κάθε χαρακτηριστικού εισόδου και b ένας σταθερός όρος (bias) που μετατοπίζει το αποτέλεσμα. Στο αποτέλεσμα αυτό εφαρμόζεται ένας τελικός κανόνας μέσω μιας συνάρτησης ενεργοποίησης (activation function, ReLU ή Tanh) διαμορφώνοντάς το κατάλληλα ώστε να αποτελέσει είσοδο στους νευρώνες του επόμενου επιπέδου.

Οι έξοδοι των hidden layers δεν είναι οι τελικές προβλέψεις κατανάλωσης. Αποτελούν ενδιάμεσες αναπαραστάσεις νέων ενδιάμεσων χαρακτηριστικών που κατασκευάζονται αυτόματα και διαδίδονται από επίπεδο σε επίπεδο (forward pass) και καταλήγουν στο επίπεδο εξόδου όπου παράγεται η τελική πρόβλεψη της μηνιαίας ενεργειακής κατανάλωσης.

Το σφάλμα μεταξύ πραγματικής τιμής και τιμής πρόβλεψης χρησιμοποιείται για την προς τα πίσω ενημέρωση (backpropagation) και διόρθωση των βαρών w και των σταθερών όρων b (bias) με σκοπό τη σταδιακή μείωσή του. Αυτό επιτυγχάνεται μέσω του βελτιστοποιητή (optimizer) που επιλέγεται (Adam ή LBFGS).

Τέλος, η πολυπλοκότητα του μοντέλου καθορίζεται από την αρχιτεκτονική του δικτύου. Ο αριθμός των κρυφών επιπέδων (hidden layers) καθορίζει το βάθος (depth) ενώ ο αριθμός των νευρώνων σε κάθε κρυφό επίπεδο καθορίζει το πλάτος (width) και τη χωρητικότητα του μοντέλου να μάθει πιο σύνθετες σχέσεις από τα χαρακτηριστικά εισόδου X.

3.7.3 Ρύθμιση υπερπαραμέτρων στο training set

Η ρύθμιση των βέλτιστων υπερπαραμέτρων στους αλγόριθμους μηχανικής μάθησης που επιλέχθηκαν πραγματοποιήθηκε με τη διαδικασία grid search σε συνδυασμό με την τεχνική της επαλήθευσης μέσω διασταύρωσης cross-validation (Hastie et al., 2009) που υλοποιείται από την κλάση GridSearchCV της βιβλιοθήκης scikit-learn στην Python. Η διαδικασία υλοποιήθηκε αποκλειστικά στο εκπαιδευτικό σύνολο (ενότητα 3.6.4).

Με τη χρήση της κλάσης ModelSpec (Εικόνα 3.49) η οποία υλοποιείται ως κλάση τύπου @dataclass ορίζεται για κάθε μοντέλο μηχανικής μάθησης ένα ενιαίο αντικείμενο στο οποίο ενσωματώνονται ως ιδιότητες οι πληροφορίες:

- της κατηγορίας στην οποία ανήκει το μοντέλο (group)
- το όνομα του μοντέλου (name)
- ο αλγόριθμος της βιβλιοθήκης scikit learn που υλοποιεί το συγκεκριμένο μοντέλο (estimator)
- η ανάγκη ή μη μετατροπής των χαρακτηριστικών εισόδου X σε πυκνή μορφή (needs_dense) και
- ένα λεξικό υποψήφιων τιμών (param grid) με τις βασικές υπερπαραμέτρους προς διερεύνηση για κάθε μοντέλο.

```python
# Κλάση decorater για τη δημιουργία αντικειμένων ανα μοντέλο
@dataclass
class ModelSpec:
    group: str
    name: str
    estimator: Any
    needs_dense: bool
    param_grid: Dict[str, List[Any]]
```

Εικόνα 3.49 Δημιουργία αντικειμένων Machine Learning

Με την προσέγγιση αυτή επιτυγχάνεται αυτοματοποίηση της διαδικασίας ρύθμισης των υπερπαραμέτρων για κάθε μοντέλο με ενιαίο και συνεπή τρόπο. Στη συνέχεια δημιουργείται μια λίστα με τα αντικείμενα των μοντέλων μηχανικής μάθησης (Εικόνα 3.50) όπως αυτά κατασκευάζονται μέσω της κλάσης ModelSpec.

```python
# Λίστα αντικειμένων μοντέλων μηχανικής μάθησης
specs: List[ModelSpec] = [
    # Linear
    ModelSpec(
        group="Linear", name="Ridge", needs_dense=False,
        estimator=Ridge(),
        param_grid={"model__alpha": [0.3, 1, 3, 5, 10, 30, 100]}),
    ModelSpec(
        group="Linear", name="ElasticNet", needs_dense=False,
        estimator=ElasticNet(random_state=0, max_iter=CD_MAX_ITER, tol=CD_TOL, selection="random"),
        param_grid={
            "model__alpha": [3e-4, 1e-3, 3e-3, 1e-2, 3e-2, 1e-1, 3e-1],
            "model__l1_ratio": [0.2, 0.5, 0.8]}),
```

```
# Forests
ModelSpec(
    group="Forests", name="RandomForest", needs_dense=False,
    estimator=RandomForestRegressor(random_state=0, n_jobs=-1),
    param_grid={
        "model__n_estimators": [200, 300, 500],
        "model__max_depth": [None, 10, 20],
        "model__min_samples_leaf": [1, 2, 3],
        "model__max_features": ["sqrt", "log2"], }),
ModelSpec(
    group="Forests", name="ExtraTrees", needs_dense=False,
    estimator=ExtraTreesRegressor(random_state=0, n_jobs=-1),
    param_grid={
        "model__n_estimators": [200, 300, 500],
        "model__max_depth": [None, 10, 20],
        "model__min_samples_leaf": [1, 2, 3],
        "model__max_features": ["sqrt", "log2"], }),
# Boosting
ModelSpec(
    group="Boosting", name="HistGB", needs_dense=True,
    estimator=HistGradientBoostingRegressor(random_state=0),
    param_grid={
        "model__learning_rate": [0.05, 0.1],
        "model__max_depth": [None, 10], }),
ModelSpec(
    group="Boosting", name="GBDT", needs_dense=True,
    estimator=GradientBoostingRegressor(random_state=0),
    param_grid={
        "model__learning_rate": [0.05, 0.1],
        "model__max_depth": [3, 5], }),
# SVR
ModelSpec(
    group="SVR", name="LinearSVR", needs_dense=False,
    estimator=LinearSVR(random_state=0, max_iter=10000),
    param_grid={"model__C": [0.5, 1.0, 2.0, 5.0]}),
ModelSpec(
    group="SVR", name="SVR_RBF", needs_dense=True,
    estimator=SVR(kernel="rbf"),
    param_grid={"model__C": [1.0, 3.0], "model__gamma": ["scale", 0.01]}),
```

```
# kNN
ModelSpec(
    group="kNN", name="KNN_uniform", needs_dense=True,
    estimator=KNeighborsRegressor(),
    param_grid={"model__n_neighbors": [5, 10, 20, 30], "model__weights": ["uniform"]}),
ModelSpec(
    group="kNN", name="KNN_distance", needs_dense=True,
    estimator=KNeighborsRegressor(),
    param_grid={"model__n_neighbors": [5, 10, 20, 30], "model__weights": ["distance"]}),
```

```
    # Neural Networks
    ModelSpec(
        group="Neural", name="MLP_Adam", needs_dense=True,
        estimator=MLPRegressor(
            random_state=0,
            solver="adam",
            activation="relu",
            max_iter=5000,
            early_stopping=True,
            validation_fraction=0.1,
            n_iter_no_change=30,
            learning_rate="adaptive"
        ),
        param_grid={
            "model__hidden_layer_sizes": [(64,), (128,), (64, 32)],
            "model__alpha": [1e-4, 1e-3],
            "model__learning_rate_init": [3e-4, 1e-3], }),
    ModelSpec(
        group="Neural", name="MLP_LBFGS", needs_dense=True,
        estimator=MLPRegressor(
            random_state=0,
            solver="lbfgs",
            activation="tanh",
            max_iter=5000,
            tol=1e-4
        ),
        param_grid={
            "model__hidden_layer_sizes": [(64,), (128,), (128, 64)],
            "model__alpha": [1e-5, 1e-4, 1e-3], }),
]
```

Εικόνα 3.50 Λίστα μοντέλων Machine Learning

Το εκπαιδευτικό σύνολο δεδομένων (training set: 1/1/2019 – 31/12/2024 ενότητα 3.6.4) χωρίζεται σε διαδοχικά χρονικά τμήματα τύπου forward rolling folds μέσω μιας ειδικής συνάρτησης κυλιόμενου διαχωρισμού (Εικόνα 3.51). Στο πρώτο τμήμα (fold) οι πρώτοι 36 μήνες (0-35) χρησιμοποιούνται ως σύνολο εκπαίδευσης και οι επόμενοι 12 (36-47) ως σύνολο επικύρωσης (validation set). Σε κάθε επόμενο τμήμα, το σύνολο εκπαίδευσης επεκτείνεται κατά 6 μήνες ενσωματώνοντας νέες χρονικά παρατηρήσεις, ενώ το σύνολο επικύρωσης μετατοπίζεται επίσης κατά 6 μήνες προς τα εμπρός διατηρώντας σταθερή διάρκεια 12 μηνών (Πίνακας 3.6). Με αυτόν τον τρόπο διασφαλίζεται ότι η εκπαίδευση προηγείται χρονικά της επικύρωσης αποφεύγοντας διαρροή πληροφορίας από μελλοντικές παρατηρήσεις (Hyndman & Athanasopoulos, 2018).

```python
# Κυλιόμενος διαχωρσμός δεδομένων (forward rolling folds)
cv_splits, folds_calendar = make_cv_splits_and_calendar(
    df_tr,
    start=TRAIN_START,
    end=TRAIN_END,
    val_size_months=12,
    step_months=6,
    min_train_months=36
)
folds_calendar.to_csv(
    os.path.join(OUT_DIR, "tscv_folds_calendar.csv"),
    index=False, encoding="utf-8-sig"
)
```

Εικόνα 3.51 Συνάρτηση κυλιόμενου διαχωρισμού training set

split	train_end	val_start	val_end	n_train_rows	n_val_rows
0	1/12/2021	1/1/2022	1/12/2022	21680	10017
1	1/6/2022	1/7/2022	1/6/2023	26686	10046
2	1/12/2022	1/1/2023	1/12/2023	31697	10086
3	1/6/2023	1/7/2023	1/6/2024	36732	10104
4	1/12/2023	1/1/2024	1/12/2024	41783	10109

Πίνακας 3.6 Τμήματα (folds) training – validation εκπαιδευτικού συνόλου

Σημειώνεται ότι για τον διαχωρισμό των δεδομένων σε τμήματα (folds) δεν χρησιμοποιείται η συνάρτηση TimeSeriesSplit της βιβλιοθήκης scikit-learn της Python διότι δεν διασφαλίζεται η πληρότητα του συνόλου των παρατηρήσεων (εγγραφών) που αντιστοιχούν σε κάθε μήνα. Αυτό θα είχε ως συνέπεια στα τμήματα διαχωρισμού (folds) κάποιες παρατηρήσεις του ίδιου μήνα να βρίσκονται στο εκπαιδευτικό σύνολο και κάποιες άλλες στο σύνολο επικύρωσης, γεγονός που θα προκαλούσε διαρροή πληροφορίας (leakage).

Η κλάση GridSearchCV (Εικόνα 3.52) εκπαιδεύει τον αλγόριθμο (estimator) του μοντέλου μηχανικής μάθησης (name) κάθε οικογένειας (group) για όλους τους συνδυασμούς των υπερπαραμέτρων (n_candidates, Πίνακας 3.7) του αντίστοιχου λεξικού param_grid στο εκπαιδευτικό σύνολο και αξιολογεί την απόδοση στο αντίστοιχο σύνολο επικύρωσης κάθε τμήματος (n_folds_used, Πίνακας 3.7) υπολογίζοντας τη μετρική RMSE και τον συντελεστή προσδιορισμού R^2 ανά κατηγορία δημοτικής υποδομής.

```
# Βρόχος επανάληψης εκπαιδευτικής διαδικασίας grid search ανά μοντέλο
for spec in specs:
    cat_cols = categorical_cols_mlp if spec.group == "Neural" else categorical_cols_default
    pipe = make_pipeline(spec.estimator, numeric_cols, cat_cols, spec.needs_dense)

    gs = GridSearchCV(
        estimator=pipe,
        param_grid=spec.param_grid,
        cv= cv_splits,
        scoring=scoring,
        refit=best_rmse_or_r2,
        n_jobs=-1,
        verbose=0,
        return_train_score=False
    )
    t_gs0 = time.perf_counter()
    gs.fit(X_tr, y_tr)
```

Εικόνα 3.52 GridSearch cross validation

Ειδικότερα πριν τον υπολογισμό των μετρικών οι προβλέψεις και οι πραγματικές τιμές κατανάλωσης όλων των παροχών αθροίζονται (aggregation) σε μηνιαία βάση για κάθε κατηγορία δημοτικής υποδομής (Εικόνα 3.53), ώστε η αξιολόγηση να πραγματοποιείται στο επίπεδο των συνολικών μηνιαίων καταναλώσεων για κάθε λειτουργικό τομέα που αποτελεί και το τελικό επίπεδο ενδιαφέροντος.

```
# Λεξικό μετρικών με τιμές απο τις αντίστοιχες συναρτήσεις υπολογισμού
scoring = {"neg_RMSE": neg_rmse_agg_scorer, "R2": r2_agg_scorer}
```

```
# Συνάρτηση υπολογισμού RMSE ανα μοντελο
def neg_rmse_agg_scorer(estimator, X, y_true):
    # Πρόβλεψη σε επίπεδο παροχή/μήνα απο το X
    y_pred = estimator.predict(X)
    y_pred = np.maximum(y_pred, 0.0)
    # Βοηθητικό data frame
    tmp = pd.DataFrame({
        "date": X["date"].values,
        "cat": X[COL_CAT].values,
        "y_true": np.asarray(y_true),
        "y_pred": np.asarray(y_pred),
    })
    # Ομαδοποίηση κατηγορία/μήνα με άθροιση τιμών true-pred
    agg = tmp.groupby(["date", "cat"], as_index=False).agg(
        y_true=("y_true", "sum"),
        y_pred=("y_pred", "sum")
    )
    return -rmse(agg["y_true"].values, agg["y_pred"].values)
```

```python
# Συνάρτηση υπολογισμού R2 ανα μοντελο
def r2_agg_scorer(estimator, X, y_true):
    # Πρόβλεψη σε επίπεδο παροχή/μήνα απο το X
    y_pred = estimator.predict(X)
    y_pred = np.maximum(y_pred, 0.0)
    # Βοηθητικό data frame
    tmp = pd.DataFrame({
        "date": X["date"].values,
        "cat": X[COL_CAT].values,
        "y_true": np.asarray(y_true),
        "y_pred": np.asarray(y_pred),
    })
    # Ομαδοποίηση κατηγορία/μήνα με άθροιση τιμών true-pred
    agg = tmp.groupby(["date", "cat"], as_index=False).agg(
        y_true=("y_true", "sum"),
        y_pred=("y_pred", "sum")
    )
    return float(r2_score(agg["y_true"].values, agg["y_pred"].values))
```

Εικόνα 3.53 Υπολογισμός μετρικών απόδοσης GridSearchCV - RMSE / R^2

Η επιλογή των βέλτιστων υπερπαραμέτρων για κάθε αλγόριθμο προκύπτει από τη σύγκριση της μέσης απόδοσης στα τμήματα (folds) πρώτα με βάση το μέσο μικρότερο RMSE (CV_mean_RMSE) και σε περίπτωση ισοβαθμίας με βάση το μέσο μεγαλύτερο συντελεστή προσδιορισμού R^2 (CV_mean_R2) όπως φαίνεται στην Εικόνα 3.54.

```python
# Συνάρτηση εύρεσης καλύτερου RMSE ή R2
def best_rmse_or_r2(cv_results: Dict[str, Any]) -> int:
    primary = np.asarray(cv_results["mean_test_neg_RMSE"])
    secondary = np.asarray(cv_results["mean_test_R2"])

    best_primary = np.nanmax(primary)
    cand = np.flatnonzero(np.isclose(primary, best_primary, atol=1e-12, rtol=0))

    if len(cand) == 1:
        return int(cand[0])

    best_secondary = np.nanmax(secondary[cand])
    cand2 = cand[np.isclose(secondary[cand], best_secondary, atol=1e-12, rtol=0)]
    return int(cand2[0])
```

Εικόνα 3.54 Συνάρτηση εύρεσης καλύτερου RMSE / R^2 - GridSearchCV

Στον Πίνακα 3.7 παρουσιάζονται συνολικά τα αποτελέσματα της απόδοσης όλων των μοντέλων όπως αυτά εκπαιδεύτηκαν μέσω της διαδικασίας grid search cross validation στο

εκπαιδευτικό σύνολο (training set) για όλους τους υποψήφιους συνδυασμούς υπερπαραμέτρων (n_candidates) που ορίστηκαν στο λεξικό param_grid.

group	model	CV_mean_RMSE	CV_mean_R2	n_folds_used	n_candidates	time (sec)
Linear	ElasticNet	70611.639262	0.706598	5	21	768.76
Linear	Ridge	71505.678948	0.698732	5	7	8.54
Forests	RandomForest	58961.326510	0.785378	5	54	4633.80
Forests	ExtraTrees	60425.759662	0.780719	5	54	5394.24
Boosting	HistGB	63945.567109	0.748334	5	4	92.01
Boosting	GBDT	65508.934833	0.737536	5	4	1463.97
SVR	LinearSVR	70295.537738	0.704958	5	4	5.91
SVR	SVR_RBF	118137.064769	0.195808	5	4	11340.84
kNN	KNN_uniform	63425.962308	0.750712	5	4	198
kNN	KNN_distance	64757.404523	0.742975	5	4	177.35
Neural	MLP_LBFGS	58389.836579	0.789881	5	9	11334.80
Neural	MLP_Adam	65413.147906	0.724416	5	12	1128.08

Πίνακας 3.7 Αποτελέσματα RMSE / R^2 επιλογής υπερπαραμέτρων GridSearchCV

Τα αποτελέσματα της διαδικασίας grid search cross-validation ανέδειξαν για κάθε μοντέλο τις βέλτιστες υπερπαραμέτρους που φαίνονται στον Πίνακα 3.8

group	model	model__alpha	model__l1_ratio
Linear	ElasticNet	0.3	0.2
Linear	Ridge	0.3	

group	model	model__max_depth	model__max_features	model__min_samples_leaf	model__n_estimators
Forests	RandomForest	None	sqrt	1	200
Forests	ExtraTrees	None	log2	1	300

group	model	model__learning_rate	model__max_depth
Boosting	HistGB	0.10	None
Boosting	GBDT	0.05	3

group	model	model__C	model__gamma
SVR	LinearSVR	2.0	
SVR	SVR_RBF	3.0	0.01

group	model	model__n_neighbors	model__weights
kNN	KNN_uniform	30	uniform
kNN	KNN_distance	30	distance

group	model	model__alpha	model__hidden_layer_sizes	model__learning_rate_init
Neural	MLP_LBFGS	0.001	(64,)	
Neural	MLP_Adam	0.001	(64, 32)	0.001

Πίνακας 3.8 Βέλτιστες υπερπαράμετροι

3.7.4 Συγκεντρωτική Αξιολόγηση στο test set

Μετά τον προσδιορισμό των βέλτιστων υπερπαραμέτρων για κάθε μοντέλο με βάση τη μέση επίδοση στα διαδοχικά τμήματα (folds) του εκπαιδευτικού συνόλου με κριτήριο το RMSE ή /και τον συντελεστή R^2 πραγματοποιήθηκε εκ νέου εκπαίδευση του αντίστοιχου βέλτιστου μοντέλου (**refit**=best_rmse_or_r2) σε ολόκληρο το αρχικό εκπαιδευτικό σύνολο.

Στη συνέχεια πραγματοποιήθηκε πρόβλεψη της μηνιαίας ενεργειακής κατανάλωσης όλων των παροχών στο σύνολο ελέγχου σε τρείς χρονικούς ορίζοντες 3, 6 και 12 μηνών με στόχο την αξιολόγηση της βραχυπρόθεσμης, μεσοπρόθεσμης και μακροπρόθεσμης απόδοσης αντίστοιχα των μοντέλων μηχανικής μάθησης.

Το σύνολο ελέγχου το οποίο καλύπτει τη χρονική περίοδο από 1/1/2025 έως 31/12/2025 αποτελείται από 10.043 παρατηρήσεις (υποενότητα 3.6.4) και δεν έχει χρησιμοποιηθεί σε κανένα στάδιο της διαδικασίας επιλογής υπερπαραμέτρων (grid search). Αποτελεί ένα πλήρως άγνωστο και αθέατο σύνολο δεδομένων για τα ήδη εκπαιδευμένα μοντέλα.

Η αξιολόγηση της απόδοσης κάθε μοντέλου (name) ανά οικογένεια αλγορίθμων (group) πραγματοποιείται με τις μετρικές R^2, MAE και RMSE οι οποίες χρησιμοποιούνται σε προβλήματα παλινδρόμησης (Εικόνα 3.58).

Όπως και στη διαδικασία grid search πριν τον υπολογισμό των μετρικών εφαρμόζεται άθροιση (aggregation) των πραγματικών και των προβλεπόμενων τιμών κατανάλωσης όλων των παροχών σε μηνιαία βάση για κάθε κατηγορία δημοτικής υποδομής. Η επιλογή

αυτή ευθυγραμμίζει την αξιολόγηση με βάση το λειτουργικό και διοικητικό επίπεδο ενδιαφέροντος (Εικόνα 3.55) περιορίζοντας την επίδραση τυχόν διακυμάνσεων σε επίπεδο μεμονωμένης παροχής.

```python
# Συνάρτηση μηνιαίας άθροισης κατανάλωσης παροχών (y_true-y_pred) ανα μήνα και κατηγορία
def aggregate_by_category(df_period: pd.DataFrame, y_pred: np.ndarray) -> pd.DataFrame:
    tmp = df_period[["date", COL_CAT, COL_Y]].copy()
    tmp["y_pred"] = y_pred
    agg = tmp.groupby(["date", COL_CAT], as_index=False).agg(
        y_true=(COL_Y, "sum"),
        y_pred=("y_pred", "sum")
    )
    agg.to_csv(os.path.join(OUT_DIR, "aggregate_by_category.csv"),
               index=False, encoding="utf-8-sig")
    return agg
```

Εικόνα 3.55 Μηνιαία άθροιση πραγματικών & προβλεπόμενων τιμών ανά κατηγορία

Οι μετρικές R^2, MAE και RMSE υπολογίζονται για κάθε χρονικό ορίζοντα τόσο στο σύνολο όλων των κατηγοριών των δημοτικών υποδομών (Εικόνα 3.56) όσο και ανά κατηγορία δημοτικής υποδομής (Εικόνα 3.57).

```python
# Συνάρτηση υπολογισμού μετρικών απόδοσης συγκεντρωτικά για όλες τις κατηγορίες
def metrics_from_agg(agg_cat: pd.DataFrame) -> Dict[str, float]:
    y_true = agg_cat["y_true"].values
    y_pred = agg_cat["y_pred"].values
    return {
        "R2": float(r2_score(y_true, y_pred)),
        "MAE": float(mean_absolute_error(y_true, y_pred)),
        "RMSE": rmse(y_true, y_pred),
        "n_points": int(len(agg_cat))
    }
```

Εικόνα 3.56 Υπολογισμός μετρικών απόδοσης συνολικά στο test set

```python
# Συνάρτηση υπολογισμού μετρικών απόδοσης ανα κατηγορία
def metrics_by_category(agg_cat: pd.DataFrame) -> pd.DataFrame:
    rows = []
    for cat, g in agg_cat.groupby(COL_CAT):
        rows.append({
            "category": cat,
            "y_true_overall": g["y_true"].sum(),
            "y_pred_overall": g["y_pred"].sum(),
            "R2": float(r2_score(g["y_true"].values, g["y_pred"].values)),
            "MAE": float(mean_absolute_error(g["y_true"].values, g["y_pred"].values)),
            "RMSE": rmse(g["y_true"].values, g["y_pred"].values),
            "n_months": int(len(g))
        })
    return pd.DataFrame(rows).sort_values("RMSE")
```

Εικόνα 3.57 Υπολογισμός μετρικών απόδοσης ανά κατηγορία υποδομής

```
overall_rows = []
by_cat_rows = []
agg_test_store = {}
test_predict_time_map = {}
for spec in specs:
    best_pipe = best_estimator_map[spec.name]
    # Πρόβλεψη κατανάλωσης παροχή/μήνα στο X_ts (test set)
    t_pred0 = time.perf_counter()
    y_pred = best_pipe.predict(X_ts)
    pred_sec = time.perf_counter() - t_pred0
    test_predict_time_map[spec.name] = float(pred_sec)
    print(f"[TIMER] TEST predict ({spec.name}): {pred_sec:.2f} sec")
    y_pred = np.maximum(y_pred, 0.0)
    # Ομαδοποίηση κατηγορία/μήνα και άθροιση true-pred
    agg_test = aggregate_by_category(df_ts, y_pred)
    agg_test_store[spec.name] = agg_test

    # Μετρικές συγκεντρωτικά για όλες τις κατηγορίες/μήνα
    m = metrics_from_agg(agg_test)
    overall_rows.append({
        "group": spec.group,
        "model": spec.name,
        "y_true_overall": agg_test["y_true"].sum(),
        "y_pred_overall": agg_test["y_pred"].sum(),
        "R2": m["R2"],
        "MAE": m["MAE"],
        "RMSE": m["RMSE"],
        "n_points": m["n_points"],
        "test_predict_seconds": float(test_predict_time_map.get(spec.name, np.nan))})
    # Μετρικές ανά κατηγορία/μήνα
    by_cat = metrics_by_category(agg_test)
    by_cat["group"] = spec.group
    by_cat["model"] = spec.name
    by_cat_rows.append(by_cat)

df_overall = pd.DataFrame(overall_rows).sort_values("RMSE")
df_by_cat = pd.concat(by_cat_rows, ignore_index=True)
df_overall.to_csv(os.path.join(OUT_DIR, "metrics_overall_models_TEST.csv"),
                  index=False, encoding="utf-8-sig")
df_by_cat.to_csv(os.path.join(OUT_DIR, "metrics_by_category_and_model_TEST.csv"),
                 index=False, encoding="utf-8-sig")

best_model_name = df_overall.iloc[0]["model"]
best_by_cat = (df_by_cat
               .sort_values(["category", "R2", "RMSE", "MAE"],ascending=[True, False, True, True])
               .groupby("category", as_index=False)
               .head(1)
               .reset_index(drop=True))
```

Εικόνα 3.58 Πρόβλεψη-αξιολόγηση απόδοσης στο test set

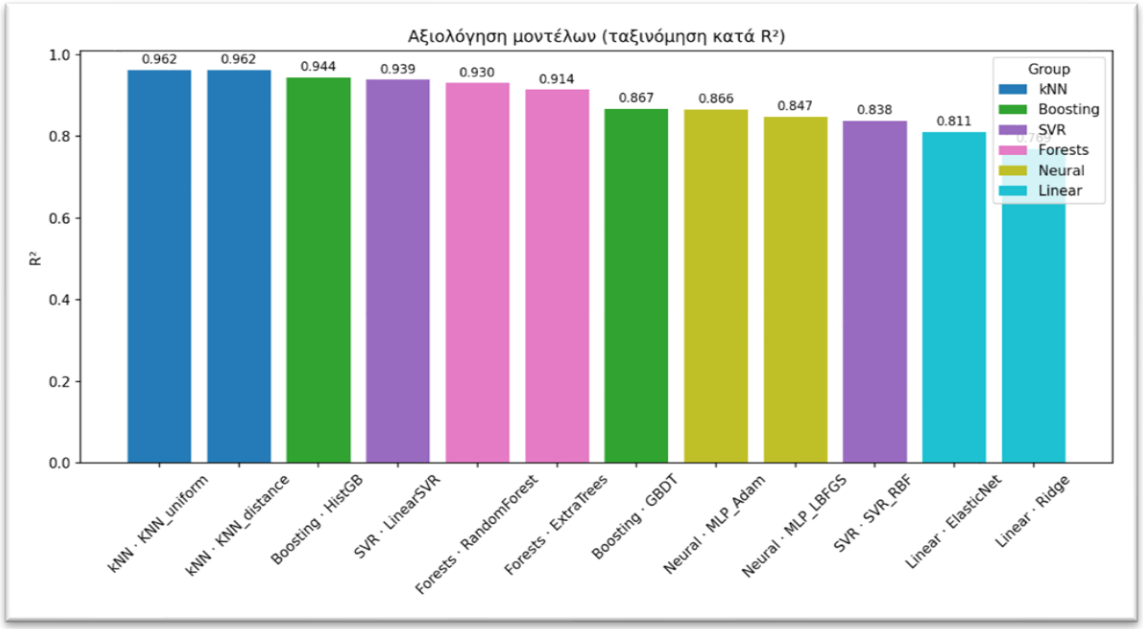
Οι παρατηρήσεις των ενεργειακών δεδομένων έχουν αντιστοιχιστεί σε 24 κατηγορίες (κλάσεις) δημοτικών υποδομών (χαρακτηριστικό «Κατηγορία Υποδομής»). Μετά την άθροιση των πραγματικών τιμών και των τιμών πρόβλεψης της ενεργειακής κατανάλωσης όλων των παροχών ανά μήνα και κατηγορία δημοτικής υποδομής για κάθε χρονικό ορίζοντα

προκύπτει συγκεκριμένος αριθμός μηνιαίων αθροιστικών παρατηρήσεων (n_points) στο σύνολο ελέγχου για την αξιολόγηση των μοντέλων μηχανικής μάθησης.

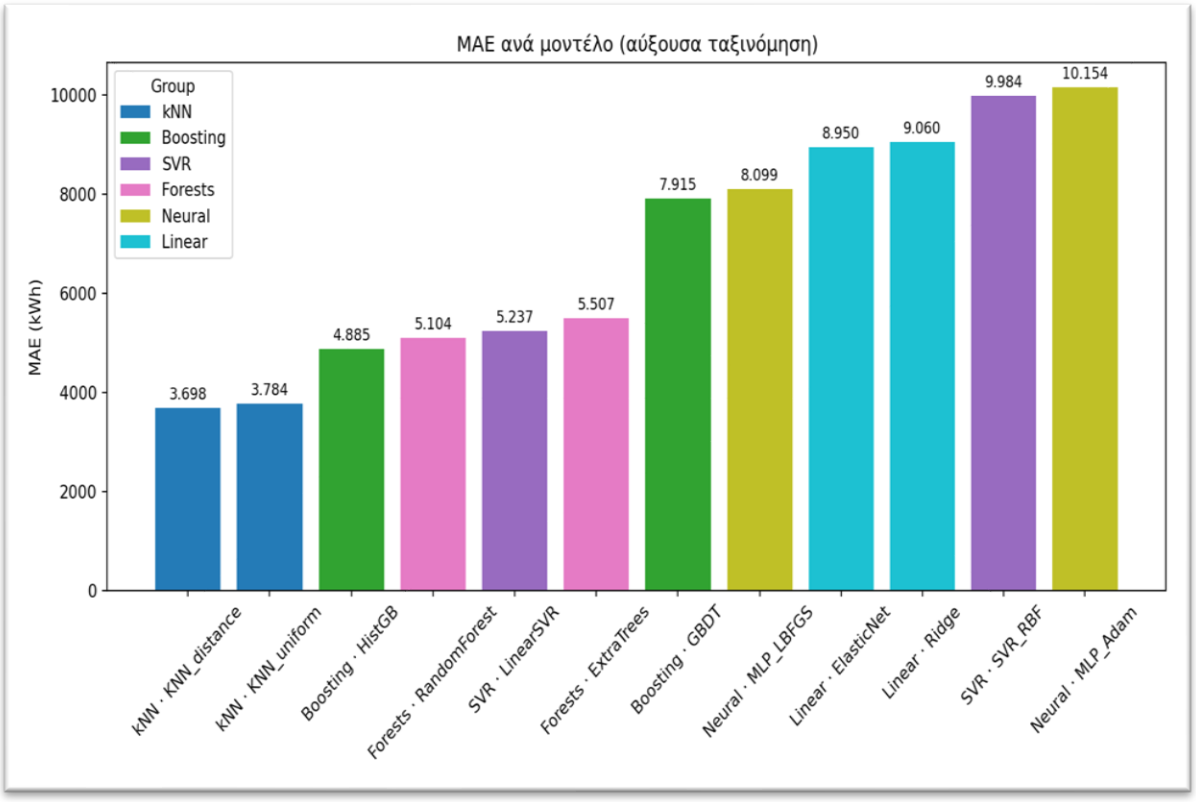
Αποτελέσματα βραχυπρόθεσμης πρόβλεψης 3 μηνών που αντιστοιχούν σε 2.523 μηνιαίες παρατηρήσεις ανά παροχή ρεύματος οι οποίες αθροίζονται σε 3 μήνες * 24 κατηγορίες = 72 μηνιαίες παρατηρήσεις ανά κατηγορία (Πίνακας 3.9, Εικόνες 3.59 – 3.62), με συνολική κατανάλωση **y_true_overall** = 1.832.193,02 kWh :

group	model	y_pred_overall (kWh)	R^2	MAE (kWh)	RMSE (kWh)	n_points	time (sec)
kNN	KNN_uniform	1.824.954,64	0.961858	3783.727491	10424.203466	72	2.08
kNN	KNN_distance	1.843.027,32	0.961702	3697.860570	10445.519852	72	1.98
Boosting	HistGB	1.907.964,76	0.944361	4885.316304	12590.208940	72	0.03
SVR	LinearSVR	1.826.736,22	0.938563	5236.631620	13229.921985	72	0.01
Forests	RandomForest	2.038.956,86	0.930405	5103.734905	14080.923744	72	0.11
Forests	ExtraTrees	2.311.188,62	0.913989	5506.974451	15653.792453	72	0.12
Boosting	GBDT	2.076.606,22	0.867314	7915.156578	19442.651469	72	0.03
Neural	MLP_Adam	2.103.903,89	0.865893	10153.916100	19546.435972	72	0.01
Neural	MLP_LBFGS	1.715.365,22	0.847463	8099.082250	20846.332929	72	0.01
SVR	SVR_RBF	1.467.590,49	0.837810	9983.562975	21495.823325	72	97.51
Linear	ElasticNet	1.671.248,53	0.810760	8949.870288	23219.305508	72	0.01
Linear	Ridge	1.951.328,84	0.769315	9059.809637	25636.104464	72	0.01

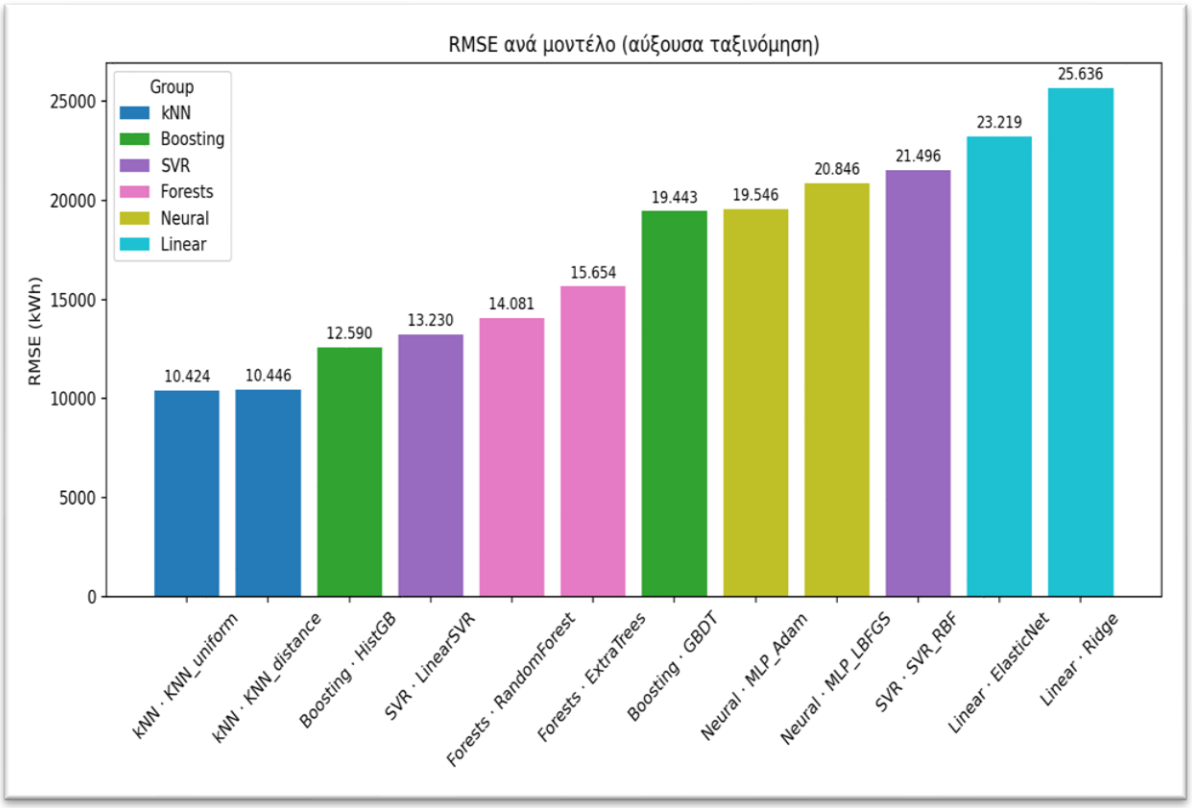
Πίνακας 3.9 Βέλτιστο μοντέλο βραχυπρόθεσμης πρόβλεψης συνολικά



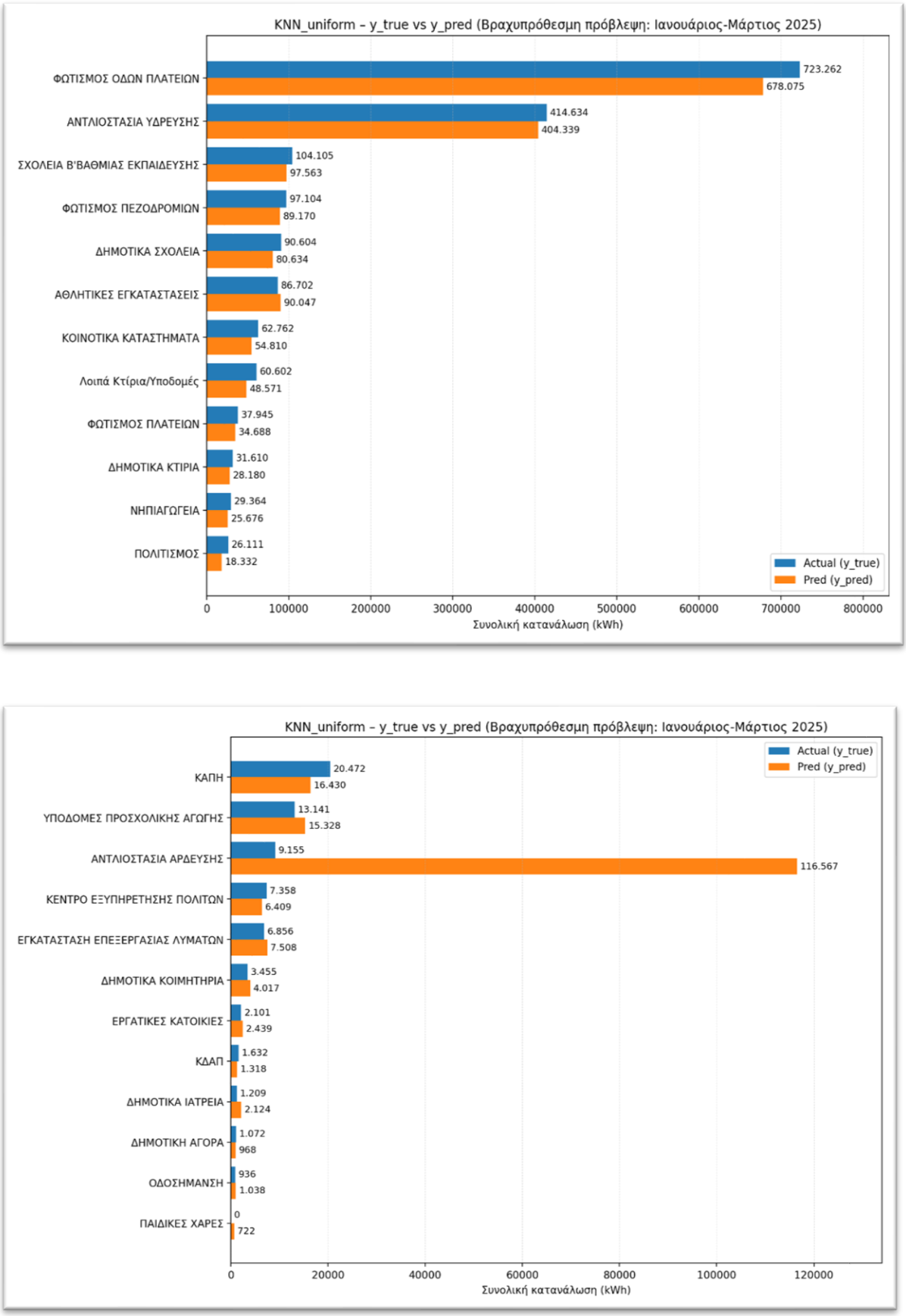
Εικόνα 3.59 Ιστόγραμμα βραχυπρόθεσμης διακύμανση R^2



Εικόνα 3.60 Ιστόγραμμα MAE βραχυπρόθεσμης πρόβλεψης



Εικόνα 3.61 Ιστόγραμμα RMSE βραχυπρόθεσμης πρόβλεψης

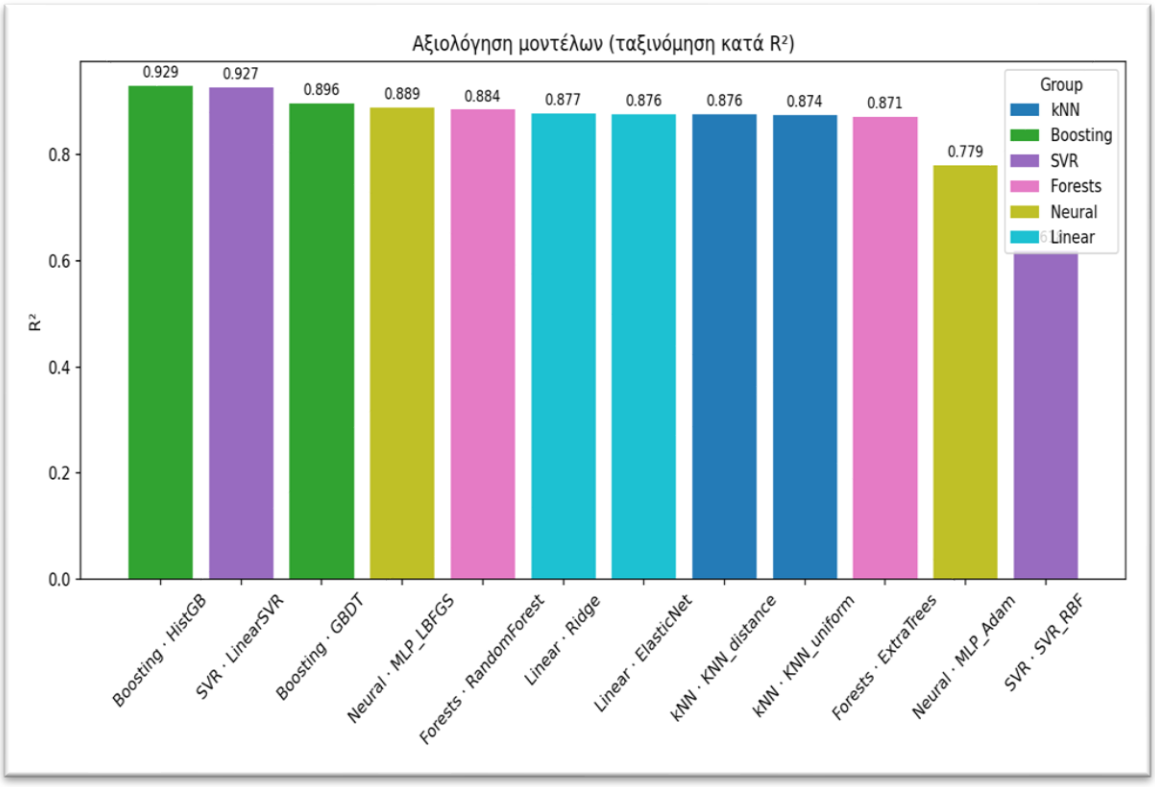


Εικόνα 3.62 Ιστόγραμμα βέλτιστου μοντέλου βραχυπρόθεσμης πρόβλεψης

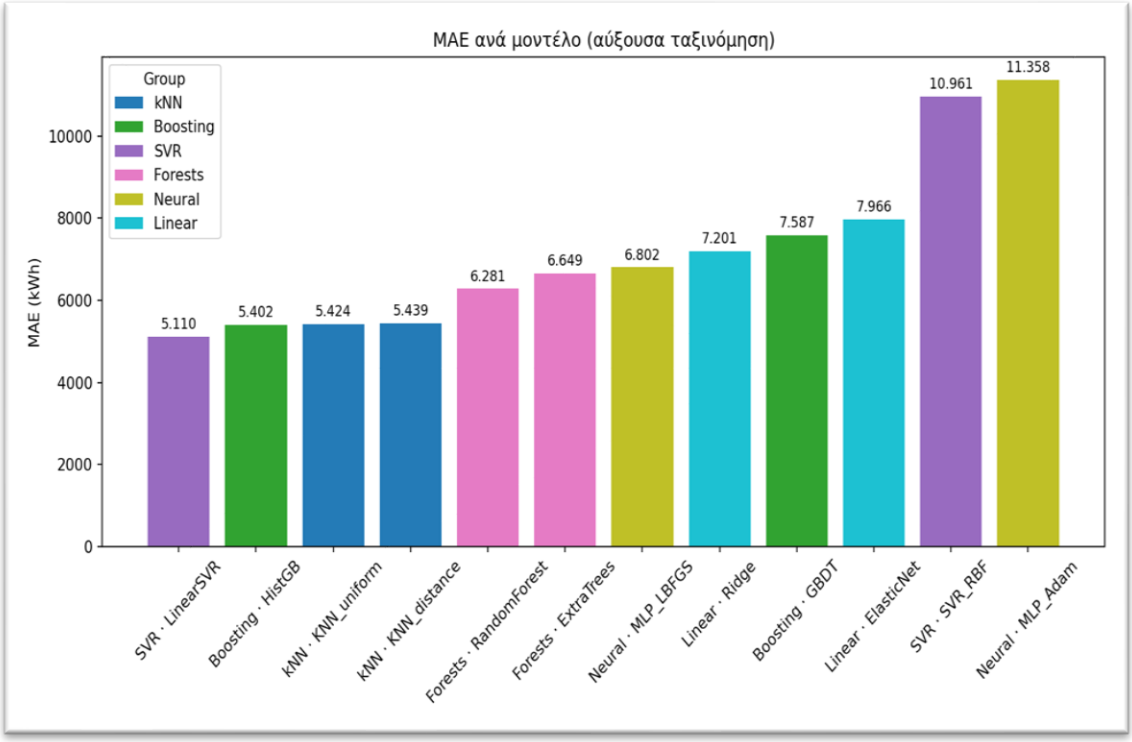
Αποτελέσματα μεσοπρόθεσμης πρόβλεψης 6 μηνών που αντιστοιχούν σε 5.034 μηνιαίες παρατηρήσεις ανά παροχή ρεύματος οι οποίες αθροίζονται σε 6 μήνες * 24 κατηγορίες = 144 μηνιαίες παρατηρήσεις ανά κατηγορία (Πίνακας 3.10, Εικόνες 3.63 – 3.66), με συνολική κατανάλωση **y_true_overall** = 3.716.800,03 kWh :

group	model	y_pred_overall (kWh)	R^2	MAE (kWh)	RMSE (kWh)	n_points	time (sec)
Boosting	HistGB	3.992.265,95	0.929015	5401.877422	15620.790022	144	0.06
SVR	LinearSVR	3.604.488,03	0.926642	5110.167220	15879.785111	144	0.02
Boosting	GBDT	4.191.870,44	0.896205	7587.062350	18889.051926	144	0.05
Neural	MLP_LBFGS	3.471.787,88	0.888956	6801.954313	19537.510501	144	0.02
Forests	RandomForest	4.167.840,01	0.884419	6281.041831	19932.663142	144	0.13
Linear	Ridge	4.010.290,56	0.877408	7201.473086	20528.242950	144	0.02
Linear	ElasticNet	3.769.347,35	0.875537	7965.705046	20684.344702	144	0.02
kNN	KNN_distance	3.910.382,06	0.875510	5438.901083	20686.554631	144	4.21
kNN	KNN_uniform	3.900.290,37	0.874203	5424.469738	20794.851231	144	4.39
Forests	ExtraTrees	4.338.634,65	0.870576	6648.818484	21092.540633	144	0.17
Neural	MLP_Adam	4.376.534,91	0.779476	11358.199023	27532.727136	144	0.01
SVR	SVR_RBF	2.785.572,39	0.618346	10960.521642	36220.658947	144	180.53

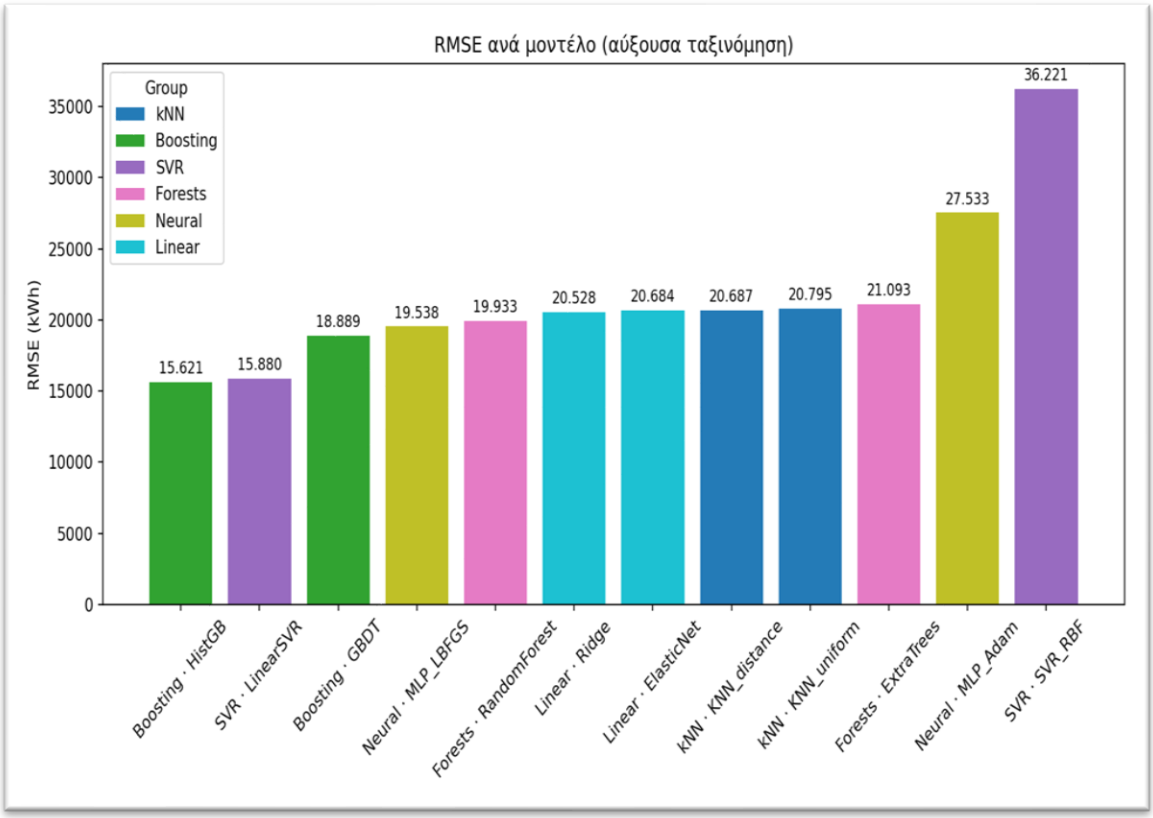
Πίνακας 3.10 Βέλτιστο μοντέλο μεσοπρόθεσμης πρόβλεψης συνολικά



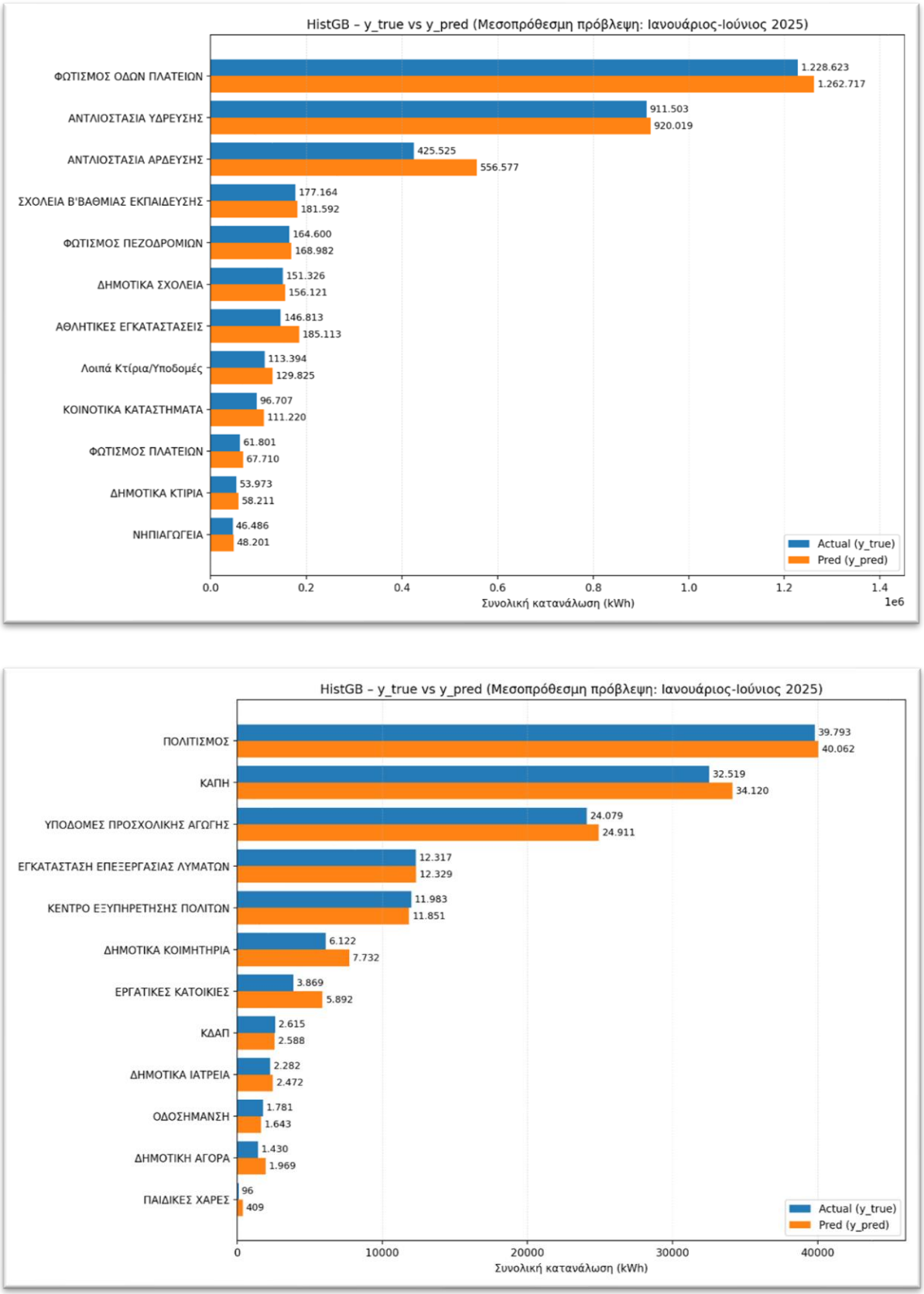
Εικόνα 3.63 Ιστόγραμμα μεσοπρόθεσμης διακύμανσης R^2



Εικόνα 3.64 Ιστόγραμμα MAE μεσοπρόθεσμης πρόβλεψης



Εικόνα 3.65 Ιστόγραμμα RMSE μεσοπρόθεσμης πρόβλεψης

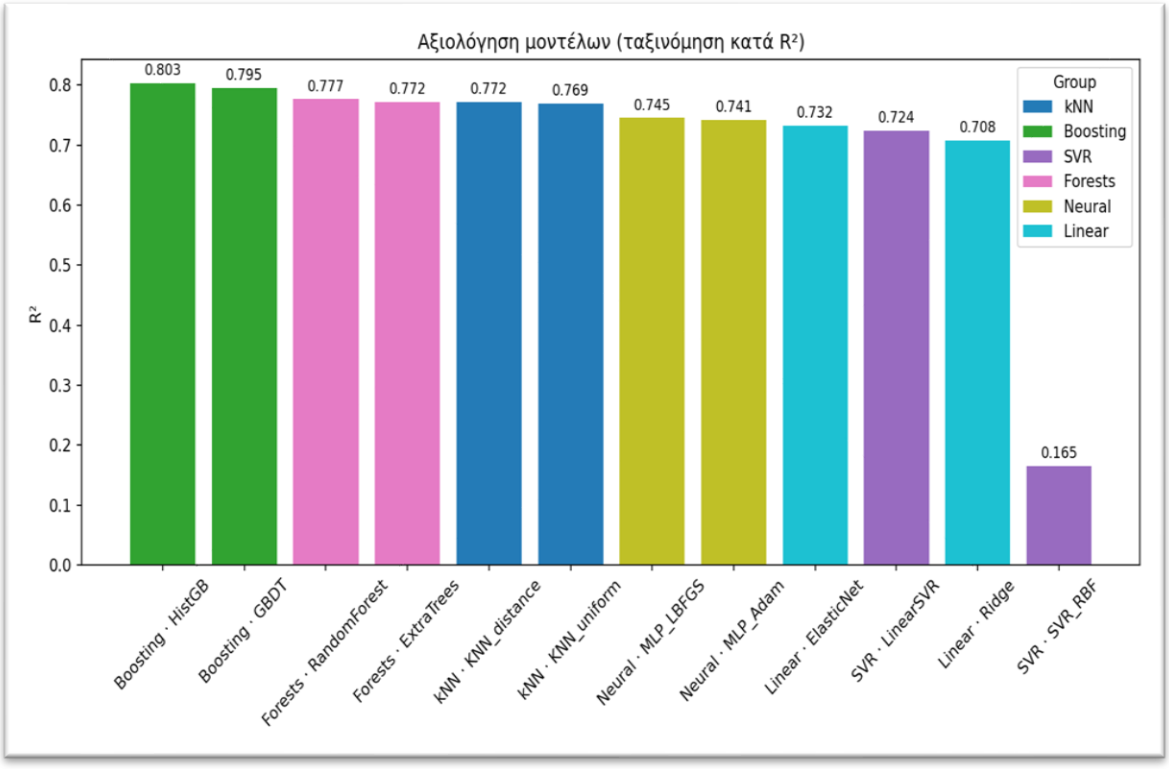


Εικόνα 3.66 Ιστόγραμμα βέλτιστου μοντέλου μεσοπρόθεσμης πρόβλεψης

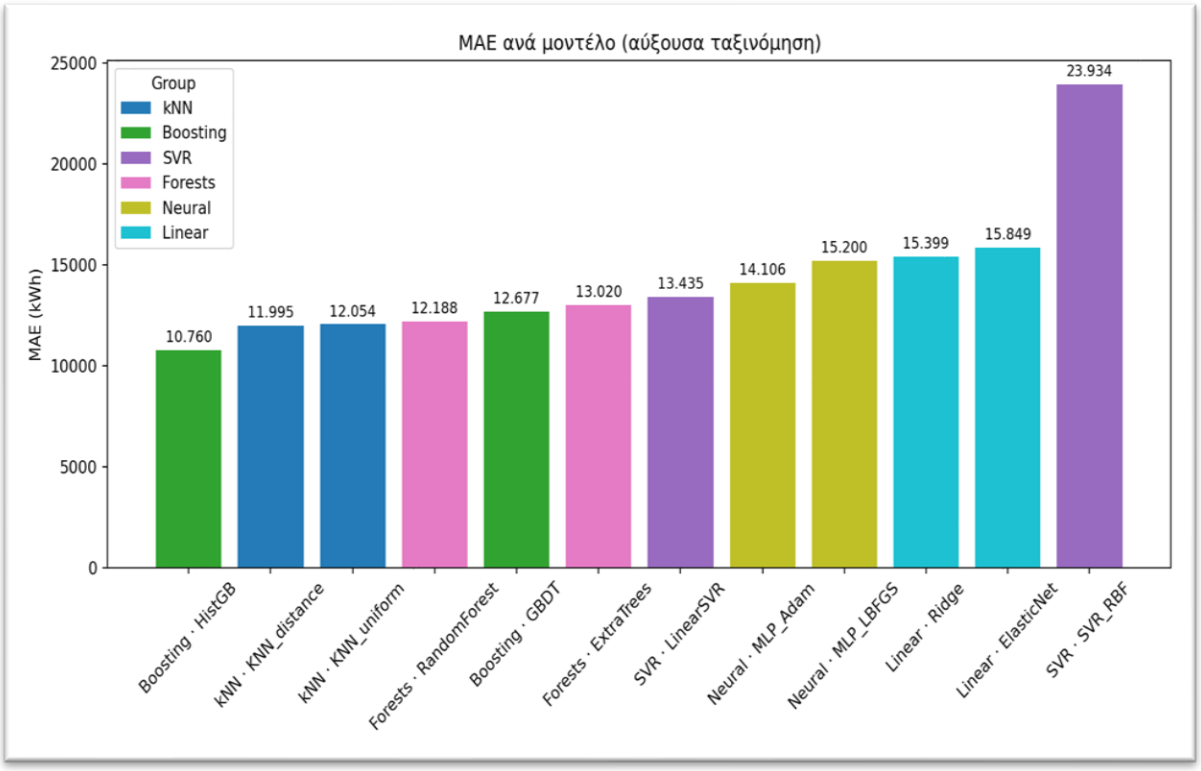
Αποτελέσματα μακροπρόθεσμης πρόβλεψης 12 μηνών που αντιστοιχούν σε 10.043 μηνιαίες παρατηρήσεις ανά παροχή ρεύματος (ολόκληρο το test set) οι οποίες αθροίζονται σε 12 μήνες * 24 κατηγορίες = 288 μηνιαίες παρατηρήσεις ανά κατηγορία (Πίνακας 3.11, Εικόνες 3.67 - 3.70), με συνολική κατανάλωση **y_true_overall** = 11.204.938,5 kWh :

group	model	y_pred_overall (kWh)	R^2	MAE (kWh)	RMSE (kWh)	n_points	time (sec)
Boosting	HistGB	10.188.861,68	0.802913	10760.464445	60911.417839	288	0.17
Boosting	GBDT	10.702.164,81	0.794626	12677.475813	62178.771656	288	0.12
Forests	RandomForest	10.450.105,99	0.776712	12187.537045	64833.945316	288	0.43
Forests	ExtraTrees	10.650.953,59	0.771892	13019.870867	65530.010706	288	0.80
kNN	KNN_distance	10.175.962,92	0.771596	11994.945233	65572.495720	288	7.28
kNN	KNN_uniform	10.137.960,12	0.768731	12053.931345	65982.371818	288	7.26
Neural	MLP_LBFGS	12.419.011,39	0.745311	15199.989460	69242.751308	288	0.03
Neural	MLP_Adam	10.372.814,27	0.741499	14105.984882	69759.138240	288	0.03
Linear	ElasticNet	9.936.205,00	0.731793	15849.264938	71056.673331	288	0.04
SVR	LinearSVR	10.115.848,02	0.724444	13434.854766	72023.552996	288	0.04
Linear	Ridge	11.002.510,52	0.708075	15399.033977	74131.873223	288	0.04
SVR	SVR_RBF	5.522.156.66	0.165477	23933.962564	125339.713635	288	366.55

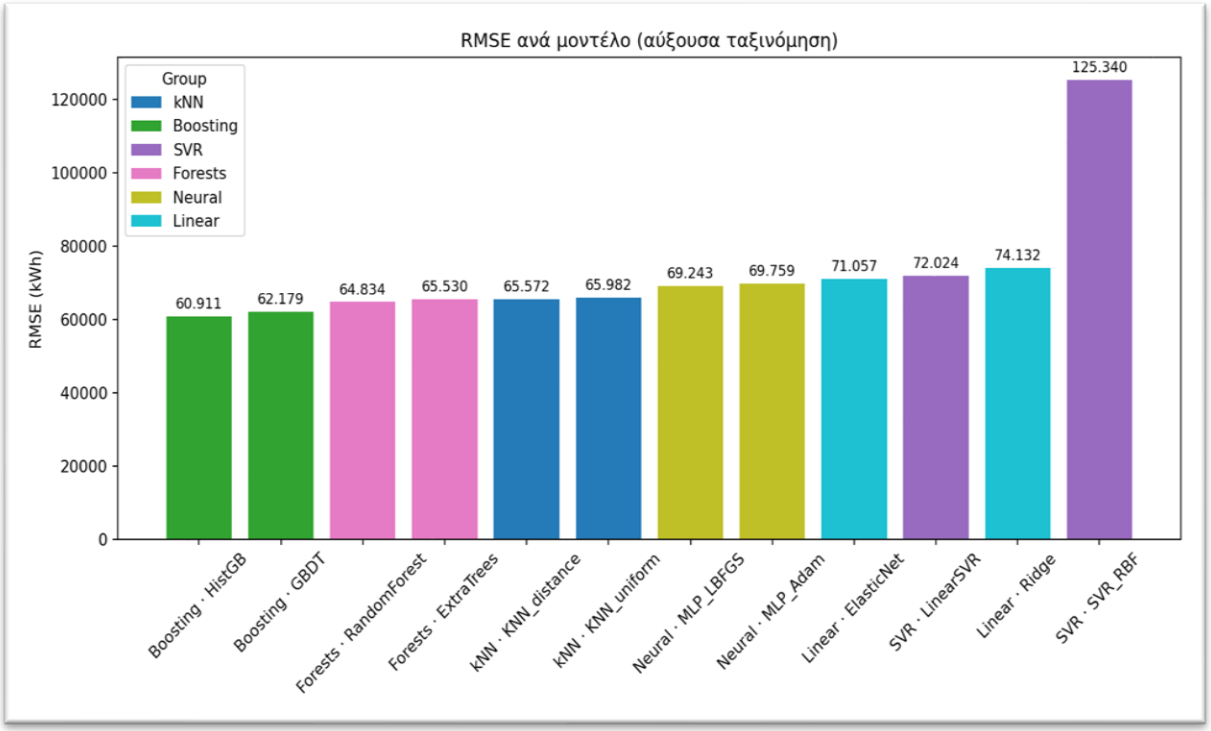
Πίνακας 3.11 Βέλτιστο μοντέλο μακροπρόθεσμης πρόβλεψης συνολικά



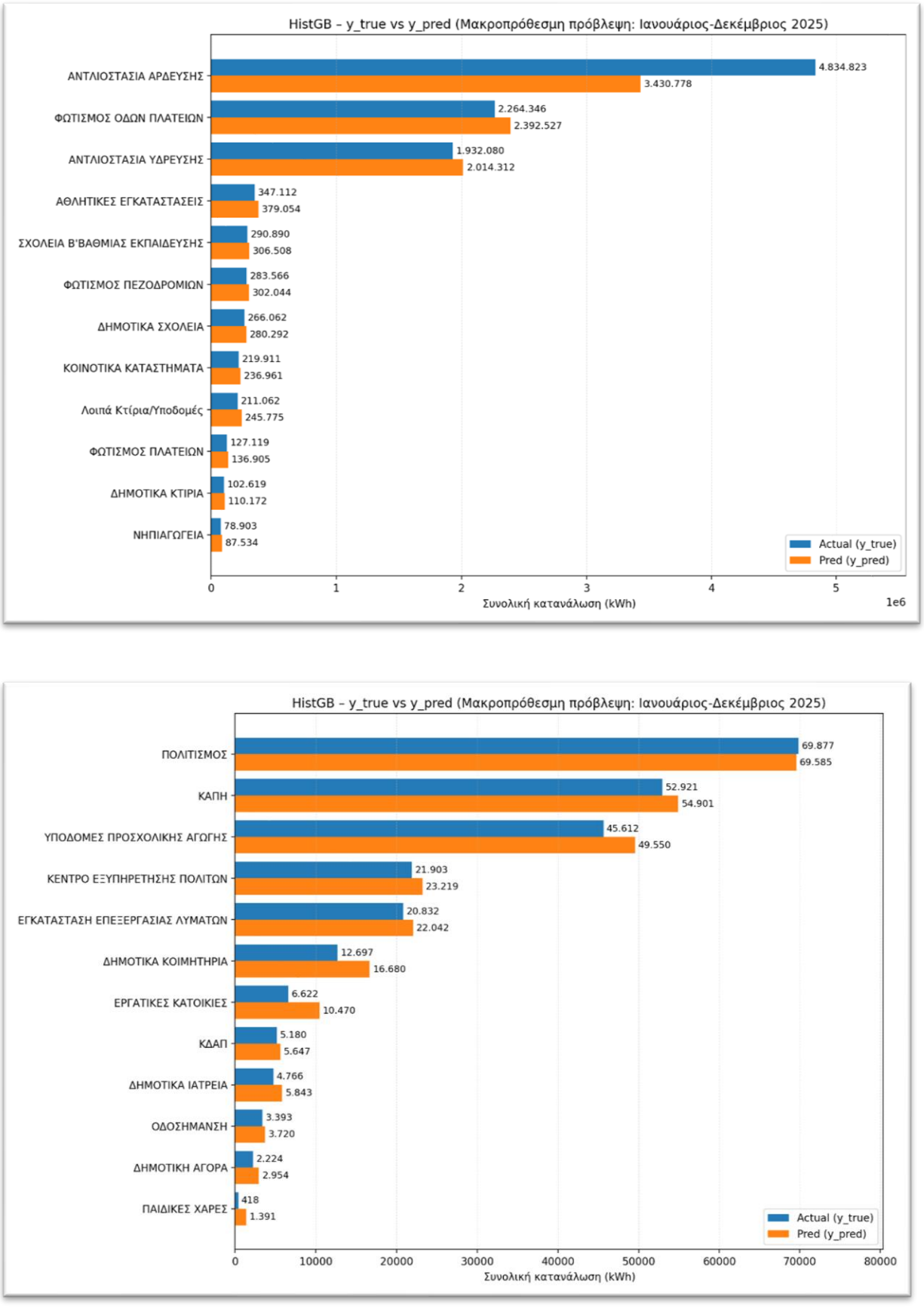
Εικόνα 3.67 Ιστόγραμμα μακροπρόθεσμης διακύμανσης R^2



Εικόνα 3.68 Ιστόγραμμα MAE μακροπρόθεσμης πρόβλεψης



Εικόνα 3.69 Ιστόγραμμα RMSE μακροπρόθεσμης πρόβλεψης



Εικόνα 3.70 Ιστόγραμμα βέλτιστου μοντέλου μακροπρόθεσμης πρόβλεψης

3.7.5 Ανάλυση Συγκεντρωτικής Αξιολόγησης

Η συγκεντρωτική και ενιαία αξιολόγηση των αλγόριθμων μηχανικής μάθησης με κριτήριο την ικανότητα τους να προβλέπουν την μηνιαία ενεργειακή κατανάλωση αρχικά σε επίπεδο μεμονωμένων παροχών και στη συνέχεια σε αθροιστικό επίπεδο ανά μήνα και κατηγορία δημοτικής υποδομής, είναι ιδιαίτερα σημαντική τόσο για την ορθή ερμηνεία των αποτελεσμάτων όσο και για την ανάδειξη του καταλληλότερου προβλεπτικού μοντέλου σε κάθε χρονικό ορίζοντα.

Η κατάταξη της απόδοσης των μοντέλων πραγματοποιείται με βάση τη μικρότερη τιμή της μετρικής RMSE (kWh/μήνα) η οποία αποτυπώνει την ευαισθησία σε μεγάλες αποκλίσεις μεταξύ πραγματικών και προβλεπόμενων τιμών της μηνιαίας ενεργειακής κατανάλωσης και είναι ιδιαίτερα κατάλληλη για την ανάδειξη των σφαλμάτων που εμφανίζονται σε ορισμένους μήνες. Παράλληλα για την πληρέστερη αποτίμηση της προβλεπτικής ικανότητας των μοντέλων υπολογίζονται η μετρική MAE (kWh/μήνα) που εκφράζει το μέσο απόλυτο σφάλμα πρόβλεψης καθώς και ο συντελεστής προσδιορισμού R^2 ο οποίος αποτυπώνει το ποσοστό της διακύμανσης της μεταβλητής-στόχου που εξηγείται από το μοντέλο. Όσο υψηλότερη είναι η τιμή του συντελεστή R^2 τόσο καλύτερη είναι η προσαρμογή και η ικανότητα γενίκευσης του μοντέλου σε νέα άγνωστα δεδομένα.

Συγκρίνοντας τα αποτελέσματα της βραχυπρόθεσμης, μεσοπρόθεσμης και μακροπρόθεσμης πρόβλεψης διαπιστώνεται ότι η απόδοση όλων των μοντέλων μειώνεται όσο αυξάνεται ο χρονικός ορίζοντας πρόβλεψης. Στη βραχυπρόθεσμη πρόβλεψη τις καλύτερες επιδόσεις παρουσιάζουν τα μοντέλα kNN με μέγιστη τιμή $\boldsymbol{R^2=0.961858}$, γεγονός που δείχνει πολύ καλή προσαρμογή στα συγκεντρωτικά δεδομένα των 3 μηνών. Στη μεσοπρόθεσμη πρόβλεψη των 6 μηνών το μοντέλο HistGradientBoosting αναδεικνύεται ως το πιο αποδοτικό επιτυγχάνοντας $\boldsymbol{R^2=0.929015,}$ ενώ στη μακροπρόθεσμη πρόβλεψη των 12 μηνών διατηρεί και πάλι την πρώτη θέση με $\boldsymbol{R^2=0.802913}$.

Η συνολική εικόνα υποδηλώνει ότι όσο διευρύνεται το βάθος του χρονικού ορίζοντα πρόβλεψης τόσο μειώνεται η ικανότητα γενίκευσης των μοντέλων. Η τάση αυτή μπορεί να αποδοθεί στο γεγονός ότι, σε μεγαλύτερους χρονικούς ορίζοντες η πολυπλοκότητα των πραγματικών δεδομένων ενεργειακής κατανάλωσης, οι μη γραμμικές δομές, οι χρονικές μεταβολές και οι σύνθετες αλληλεπιδράσεις που τα χαρακτηρίζουν, καθιστούν δυσκολότερη την αποτύπωσή τους με την ίδια ακρίβεια.

Boosting (HistGB, GBDT)

Τα μοντέλα ενίσχυσης (Boosting) εμφανίζουν συνολικά την καλύτερη απόδοση ιδιαίτερα στους μεσοπρόθεσμους και μακροπόθεμσους χρονικούς ορίζοντες. Ειδικότερα το HistGB επιτυγχάνει την υψηλότερη επίδοση στη μακροπρόθεσμη πρόβλεψη με **R^2 = 0.802913** και ταυτόχρονα το χαμηλότερο **MAE = 10.760 KWh** και **RMSE = 60.911 KWh**. Αυτό πρακτικά σημαίνει ότι το μοντέλο εξηγεί περίπου το **80.3%** της διακύμανσης των συνολικών μηνιαίων καταναλώσεων ανά κατηγορία δημοτικής υποδομής. Ταυτόχρονα διατηρείται χαμηλά το απόλυτο σφάλμα πρόβλεψης (MAE - kWh/μήνα) αλλά και τα σφάλματα που οφείλονται σε μεγάλες αποκλίσεις (RMSE - kWh/μήνα) της ενεργειακής κατανάλωσης για κάποιους μήνες σε σχέση με τα υπόλοιπα μοντέλα. Ακολουθεί το GBDT γεγονός που επιβεβαιώνει ότι τα συγκεκριμένα μοντέλα ανταποκρίνονται πιο αποτελεσματικά στο συγκεκριμένο πρόβλημα πρόβλεψης.

Forests (ExtraTrees, RandomForest)

Στη μακροπρόθεσμη πρόβλεψη ακολουθούν σχετικά κοντά τα μοντέλα Forests με τιμές **R^2 ≈ 0.772 - 0777** τα οποία έχουν καλή επίδοση και σταθερότητα. Παρότι οι τιμές MAE και RMSE είναι λίγο υψηλότερες από αυτές των μοντέλων ενίσχυσης (boosting), εμφανίζουν ικανοποιητική ικανότητα γενίκευσης και στους τρείς χρονικούς ορίζοντες.

kNN (KNN_distance, KNN_uniform)

Τα μοντέλα kNN εμφανίζουν πολύ καλή συμπεριφορά στη βραχυπρόθεσμη διακύμανση των συνολικών μηνιαίων καταναλώσεων ανά κατηγορία υποδομής επιτυγχάνοντας τις καλύτερες επιδόσεις οι οποίες όμως μειώνονται όσο αυξάνεται ο χρονικός ορίζοντας. Στη μεσοπρόθεσμη και μακροπρόθεσμη πρόβλεψη εμφανίζουν επίσης καλή προσαρμογή και μάλιστα με χαμημότερο μέσο απόλυτο σφάλμα MAE σε σχέση με τα μοντέλα Forests. Αυτό πρακτικά σημαίνει ικανοποιητικές κατά μέσο όρο προβλέψεις αλλά οι υψηλότερες τιμές RMSE δείχνουν μεγάλες αποκλίσεις σε κάποιους μήνες ανά κατηγορία δημοτικης υποδομής.

Neural (MLP_LBFGS, MLP_Adam)

Τα μοντέλα νευρωνικών δικτύων (MLP_LBFGS, MLP_Adam) επιτυγχάνουν συντελεστή **R^2 ≈ 0.741 – 0.745** στη μακροπρόθεσμη πρόβλεψη αλλά με αισθητά μεγαλύτερο μέσο απόλυτο σφάλμα MAE και RMSE σε kWh/μήνα σε σχέση με τα μοντέλα που βασίζονται σε δέντρα απόφασης. Παρά την ικανότητα τους να εξηγούν σύνθετες και μη γραμμικές

σχέσεις στο συγκεκριμένο σχήμα δεδομένων δεν υπερέχουν των τεχνικών boosting και forests.

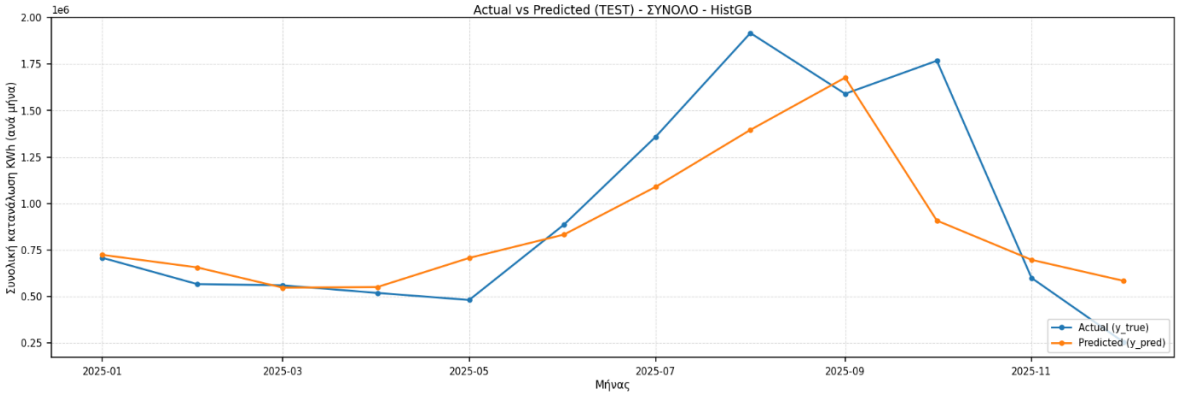
Linear (ElasticNet, Ridge)

Τα γραμμικά μοντέλα παρουσιάζουν ικανοποιητική επίδοση στη διακύμανση της μακροπρόθεσμης πρόβλεψης με τιμές **$R^2 \approx 0.708 – 0.732$.** Αυτό δείχνει ότι ένα μέρος της διακύμανσης της μηνιαίας κατανάλωσης μπορεί να αποδοθεί στην ύπαρξη γραμμικών σχέσεων μεταξύ των χαρακτηριστικών εισόδου X και της μεταβλητής-στόχου y = «Μηνιαία κατανάλωση KWh». Η χαμηλότερη ικανότητα γενίκευσης στους τρείς χρονικούς ορίζοντες σε σχέση με τα μοντέλα που βασίζονται σε δέντρα απόφασης (decision trees) σημαίνει ότι σε πραγματικά ενεργειακά δεδομένα εμφανίζονται έντονες μη γραμμικές δομές και αλληλεπιδράσεις μεταξύ των χαρακτηριστικών εισόδου X τις οποίες τα γραμμικά μοντέλα δεν μπορούν να αποτυπώσουν ικανοποιητικά.

SVR (LinearSVR, SVR_RBF)

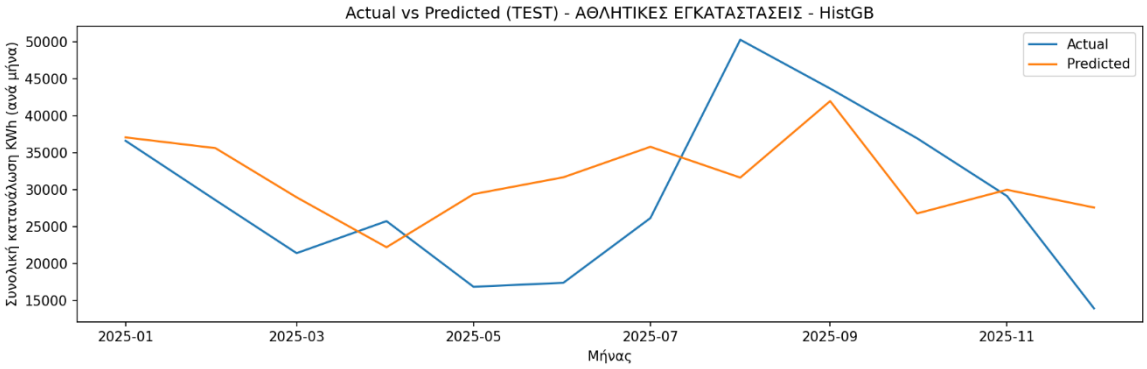
Από τα μοντέλα της οικογένειας μηχανών διανυσμάτων υποστήριξης SVM/SVR το LinearSVR παρουσιάζει ικανοποιητική επίδοση στον βραχυπρόθεσμο και μεσοπρόθεσμο ορίζοντα πρόβλεψης με χαμηλότερο μέσο απόλυτο σφάλμα MAE από τα γραμμικά μοντέλα χωρίς όμως να ξεπερνά τα decision trees μοντέλα. Αντίθετα το SVR_RBF παρουσιάζει πολύ χαμηλή απόδοση στη μακροπρόθεμση πρόβλεψη με τιμή **$R^2 \approx 0.165$** και πολύ υψηλά σφάλματα MAE και RMSE, γεγονός που δείχνει αδυναμία προσαρμογής στο συγκεκριμένο πρόβλημα πρόβλεψης.

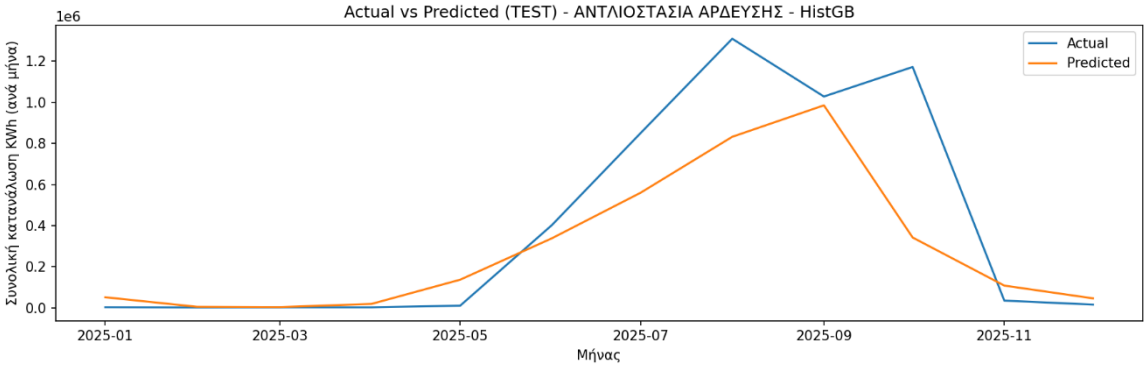
Στην Εικόνα 3.71 απεικονίζεται, για το μοντέλο με την καλύτερη απόδοση (HistGB) στον χρονικό ορίζοντα των 12 μηνών, το διάγραμμα της μηνιαίας πραγματικής και προβλεπόμενης τιμής της κατανάλωσης (kWh) για το σύνολο των δημοτικών υποδομών.

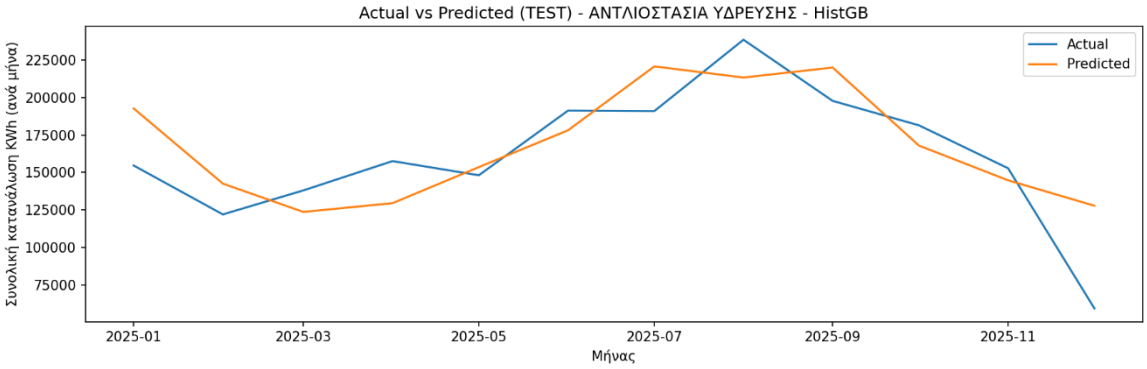


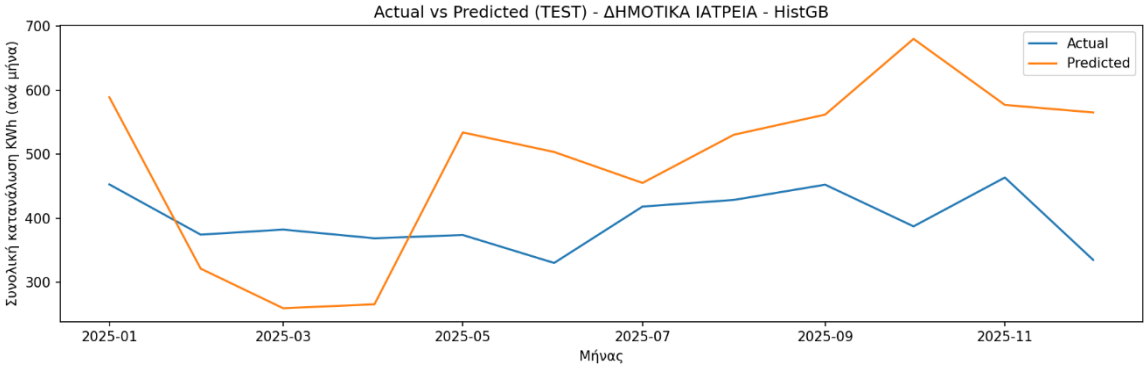
Εικόνα 3.71 Διάγραμμα μακροπρόθεσμης πρόβλεψης βέλτιστου μοντέλου

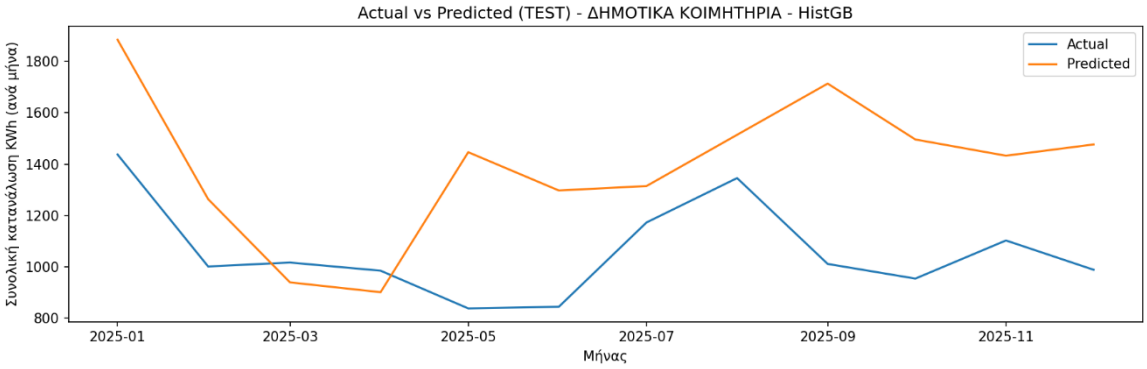
Στη Εικόνα 3.72 απεικονίζονται τα διαγράμματα των μηνιαίων πραγματικών και προβλεπόμενων τιμών όπως αυτά διαμορφώνονται για κάθε επιμέρους κατηγορία από το μοντέλο με την καλύτερη απόδοση (HistGB) στον χρονικό ορίζοντα των 12 μηνών.

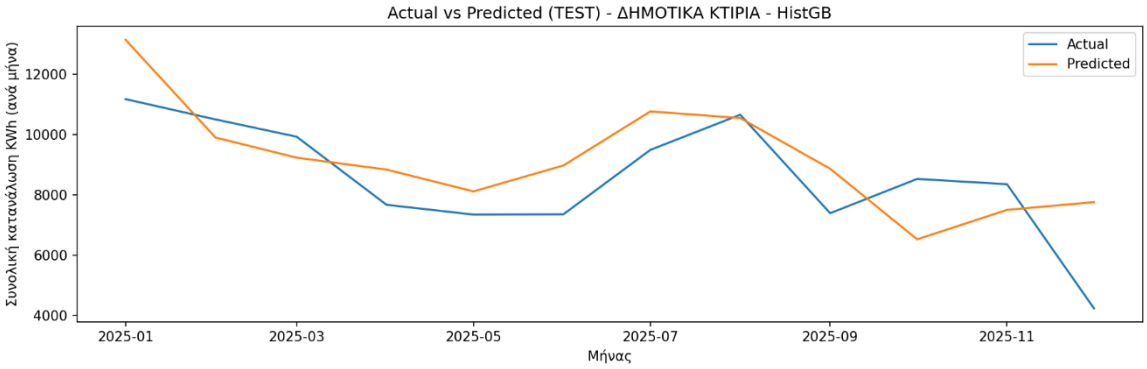


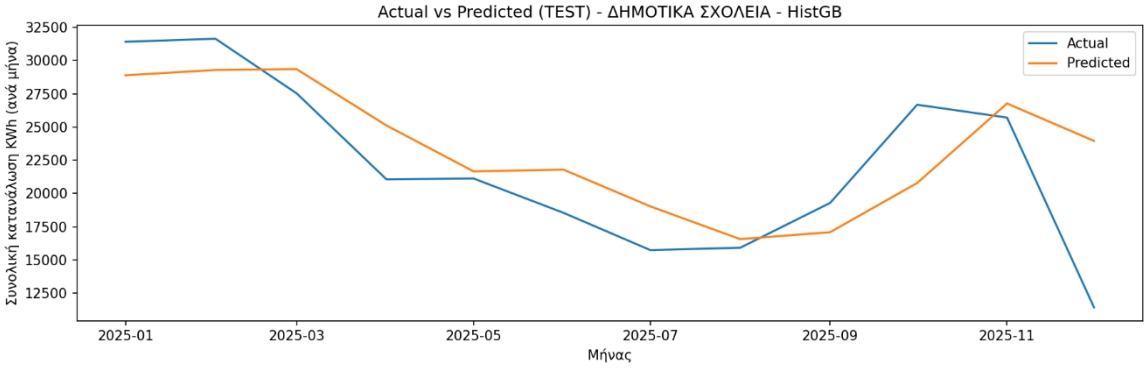


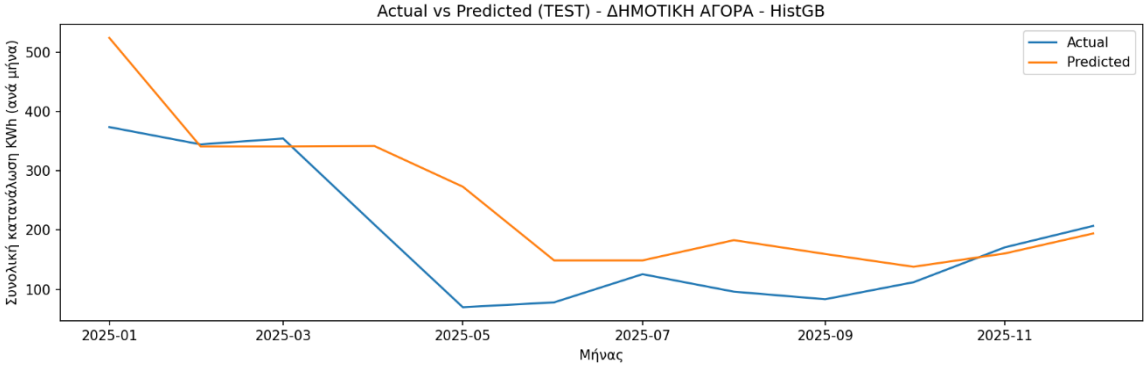


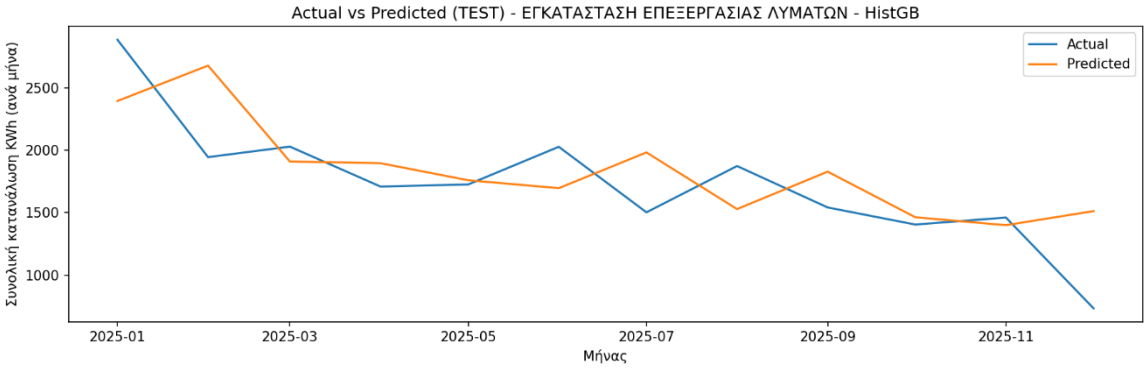


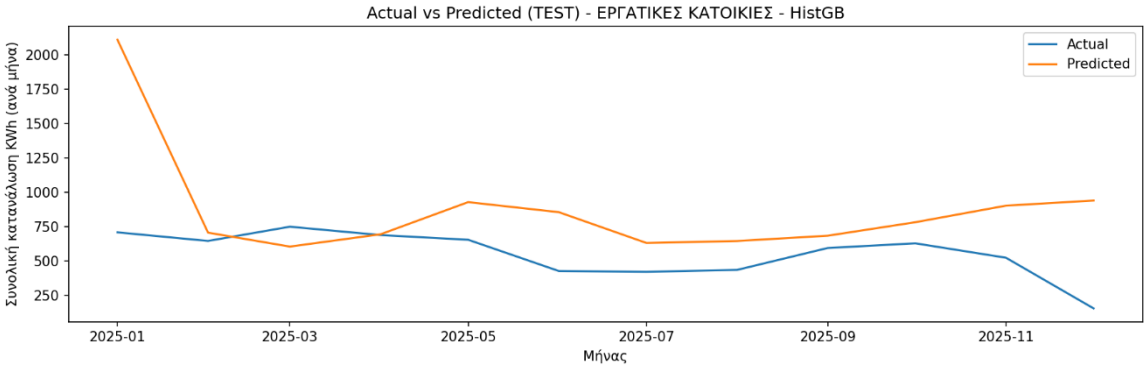

Actual vs Predicted (TEST) - ΔΗΜΟΤΙΚΑ ΚΟΙΜΗΤΗΡΙΑ - HistGB
Actual
Predicted
Συνολική καταναλωση KWh (ανά μήνα)
1800
1600
1400
1200
1000
800
2025-01
2025-03
2025-05
2025-07
2025-09
2025-11
Μήνας

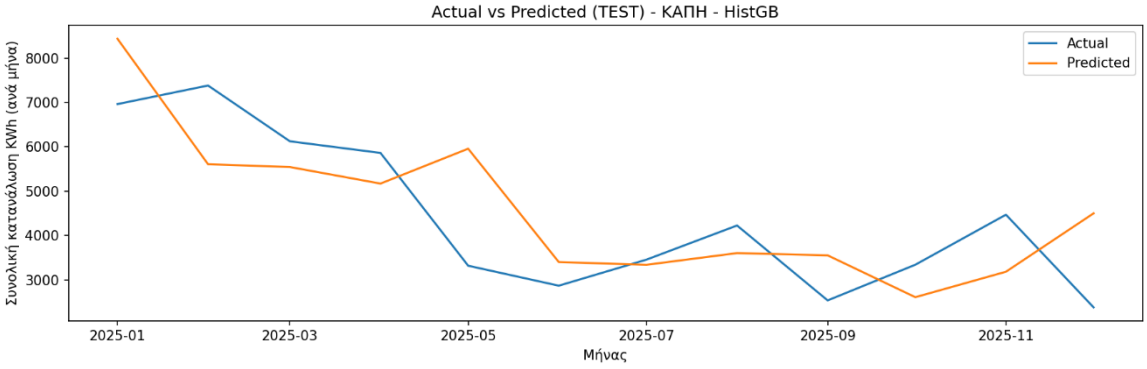

Actual vs Predicted (TEST) - ΔΗΜΟΤΙΚΑ ΚΤΙΡΙΑ - HistGB
Actual
Predicted
Συνολική καταναλωση KWh (ανά μήνα)
12000
10000
8000
6000
4000
2025-01
2025-03
2025-05
2025-07
2025-09
2025-11
Μήνας

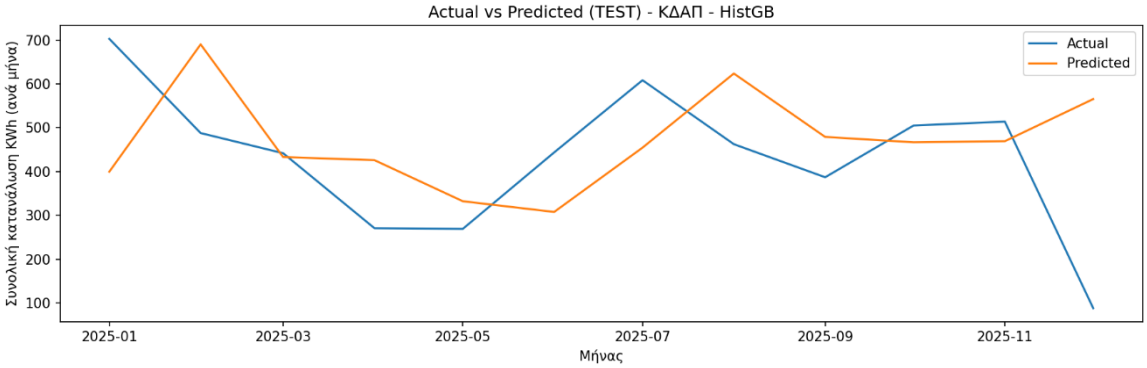

Actual vs Predicted (TEST) - ΔΗΜΟΤΙΚΑ ΣΧΟΛΕΙΑ - HistGB
Actual
Predicted
Συνολική καταναλωση KWh (ανά μήνα)
32500
30000
27500
25000
22500
20000
17500
15000
12500
2025-01
2025-03
2025-05
2025-07
2025-09
2025-11
Μήνας

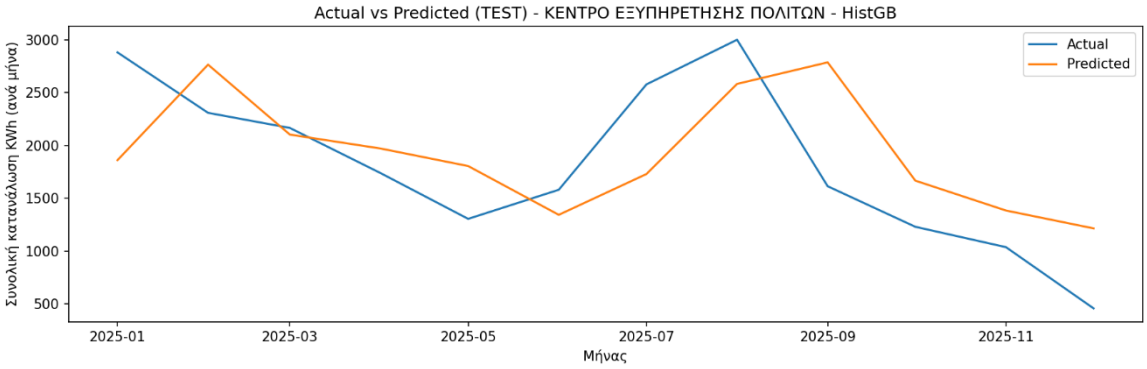

Actual vs Predicted (TEST) - ΔΗΜΟΤΙΚΗ ΑΓΟΡΑ - HistGB
Actual
Predicted
Συνολική καταναλωση KWh (ανά μήνα)
500
400
300
200
100
2025-01
2025-03
2025-05
2025-07
2025-09
2025-11
Μήνας

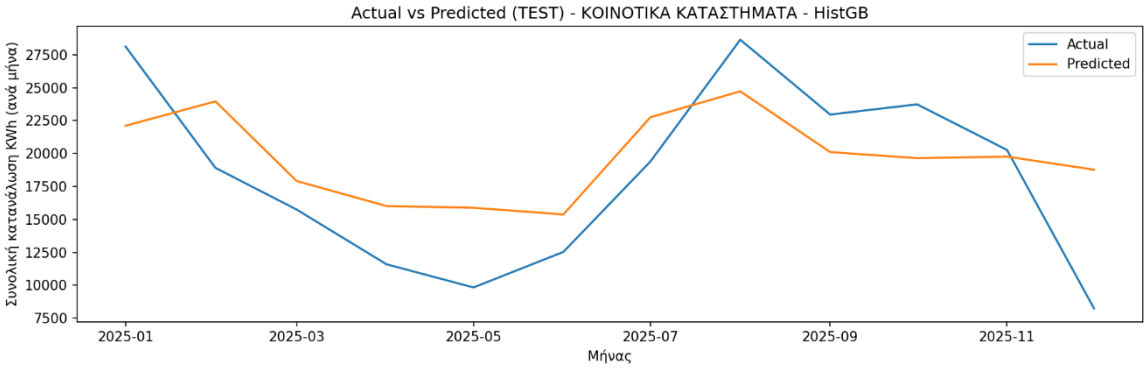

Actual vs Predicted (TEST) - ΕΓΚΑΤΑΣΤΑΣΗ ΕΠΕΞΕΡΓΑΣΙΑΣ ΛΥΜΑΤΩΝ - HistGB
Actual
Predicted
Συνολική καταναλωση KWh (ανά μήνα)
2500
2000
1500
1000
2025-01
2025-03
2025-05
2025-07
2025-09
2025-11
Μήνας

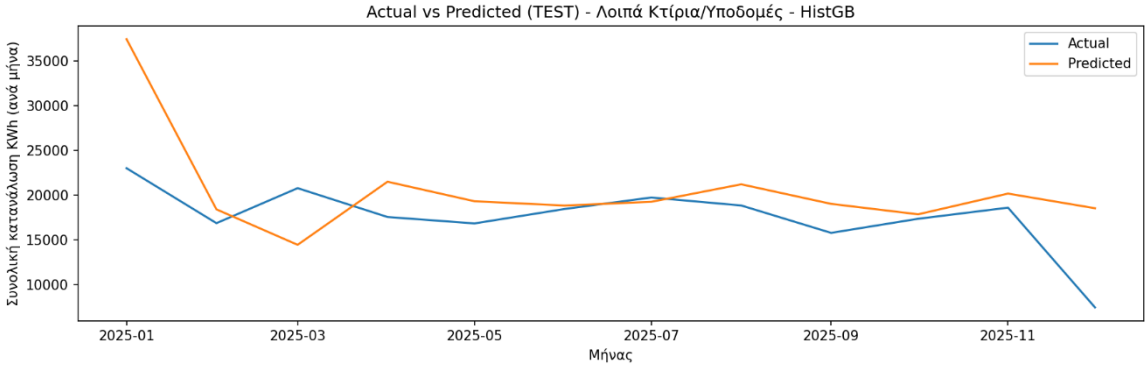

Actual vs Predicted (TEST) - ΕΡΓΑΤΙΚΕΣ ΚΑΤΟΙΚΙΕΣ - HistGB
Actual
Predicted
Συνολική καταναλωση KWh (ανά μήνα)
2000
1750
1500
1250
1000
750
500
250
2025-01
2025-03
2025-05
2025-07
2025-09
2025-11
Μήνας

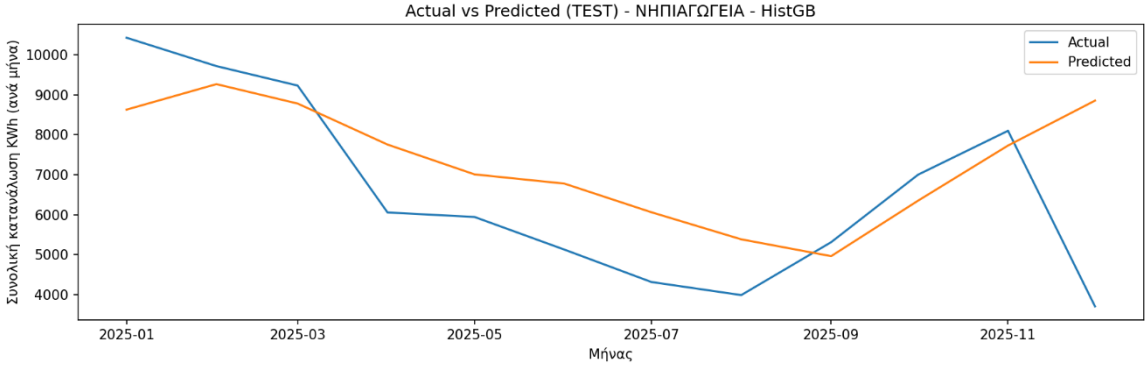

Actual vs Predicted (TEST) - ΚΑΠΗ - HistGB
Actual
Predicted
Συνολική καταναλωση KWh (ανά μήνα)
8000
7000
6000
5000
4000
3000
2025-01
2025-03
2025-05
2025-07
2025-09
2025-11
Μήνας

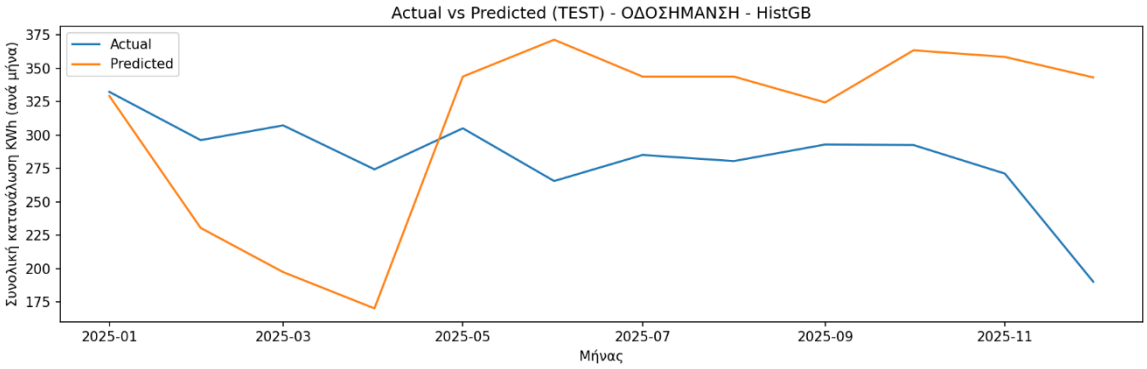

Actual vs Predicted (TEST) - ΚΔΑΠ - HistGB
Actual
Predicted
Συνολική καταναλωση KWh (ανά μήνα)
700
600
500
400
300
200
100
2025-01
2025-03
2025-05
2025-07
2025-09
2025-11
Μήνας

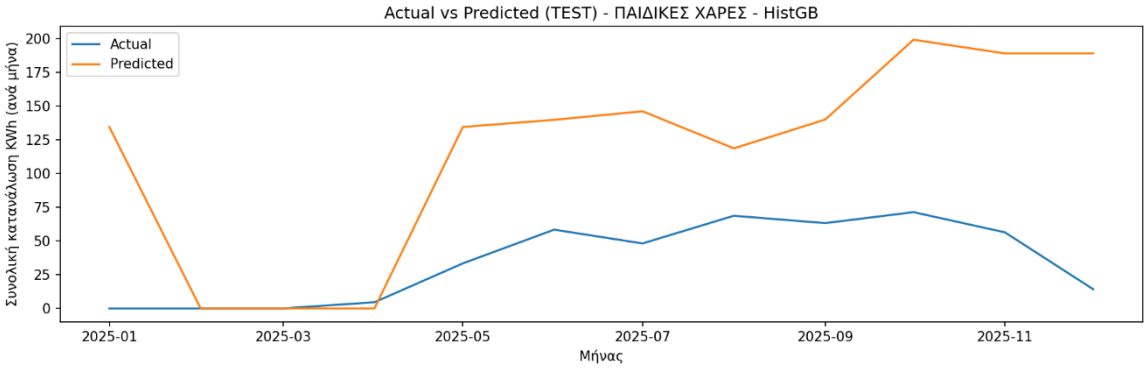

Actual vs Predicted (TEST) - ΚΕΝΤΡΟ ΕΞΥΠΗΡΕΤΗΣΗΣ ΠΟΛΙΤΩΝ - HistGB
Actual
Predicted
Συνολική κατανάλωση KWh (ανά μήνα)
3000
2500
2000
1500
1000
500
2025-01
2025-03
2025-05
2025-07
2025-09
2025-11
Μήνας

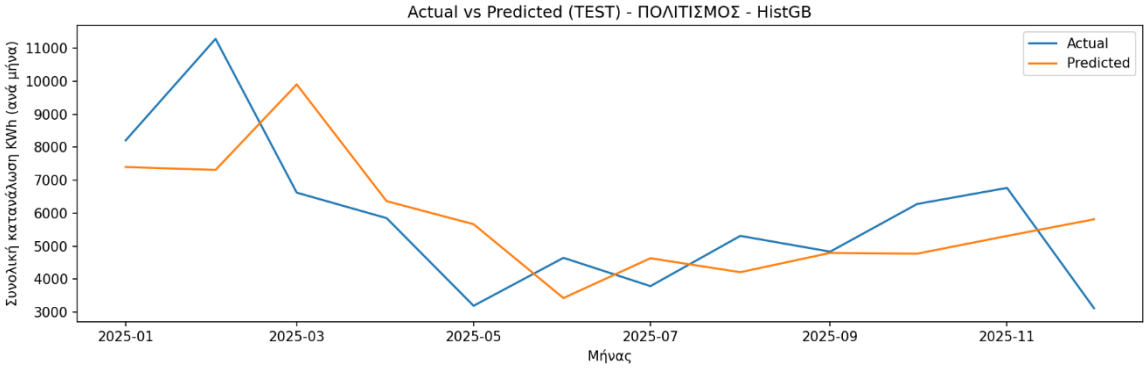

Actual vs Predicted (TEST) - ΚΟΙΝΟΤΙΚΑ ΚΑΤΑΣΤΗΜΑΤΑ - HistGB
Actual
Predicted
Συνολική κατανάλωση KWh (ανά μήνα)
27500
25000
22500
20000
17500
15000
12500
10000
7500
2025-01
2025-03
2025-05
2025-07
2025-09
2025-11
Μήνας

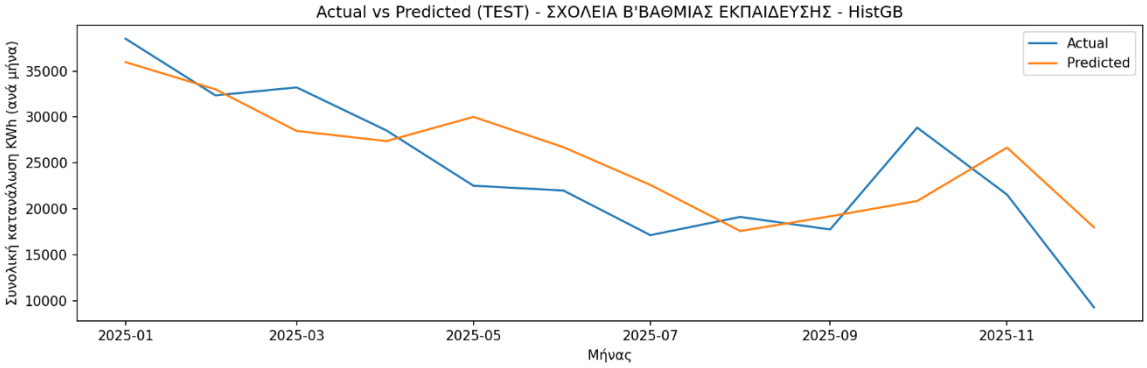

Actual vs Predicted (TEST) - Λοιπά Κτίρια/Υποδομές - HistGB
Actual
Predicted
Συνολική κατανάλωση KWh (ανά μήνα)
35000
30000
25000
20000
15000
10000
2025-01
2025-03
2025-05
2025-07
2025-09
2025-11
Μήνας


Actual vs Predicted (TEST) - ΝΗΠΙΑΓΩΓΕΙΑ - HistGB
Actual
Predicted
Συνολική κατανάλωση KWh (ανά μήνα)
10000
9000
8000
7000
6000
5000
4000
2025-01
2025-03
2025-05
2025-07
2025-09
2025-11
Μήνας


Actual vs Predicted (TEST) - ΟΔΟΣΗΜΑΝΣΗ - HistGB
Actual
Predicted
Συνολική κατανάλωση KWh (ανά μήνα)
375
350
325
300
275
250
225
200
175
2025-01
2025-03
2025-05
2025-07
2025-09
2025-11
Μήνας


Actual vs Predicted (TEST) - ΠΑΙΔΙΚΕΣ ΧΑΡΕΣ - HistGB
Actual
Predicted
Συνολική κατανάλωση KWh (ανά μήνα)
200
175
150
125
100
75
50
25
0
2025-01
2025-03
2025-05
2025-07
2025-09
2025-11
Μήνας


Actual vs Predicted (TEST) - ΠΟΛΙΤΙΣΜΟΣ - HistGB
Actual
Predicted
Συνολική κατανάλωση KWh (ανά μήνα)
11000
10000
9000
8000
7000
6000
5000
4000
3000
2025-01
2025-03
2025-05
2025-07
2025-09
2025-11
Μήνας


Actual vs Predicted (TEST) - ΣΧΟΛΕΙΑ Β'ΒΑΘΜΙΑΣ ΕΚΠΑΙΔΕΥΣΗΣ - HistGB
Actual
Predicted
Συνολική κατανάλωση KWh (ανά μήνα)
35000
30000
25000
20000
15000
10000
2025-01
2025-03
2025-05
2025-07
2025-09
2025-11
Μήνας

Εικόνα 3.72 Διαγράμματα μακροπρόθεσμης πρόβλεψης HistGB ανά κατηγορία

3.7.6 Συγκριτική Αξιολόγηση μοντέλων ανά κατηγορία Δημοτικής Υποδομής

Σε επίπεδο κατηγορίας διαπιστώνεται ότι το βέλτιστο μοντέλο σε κάθε χρονικό ορίζοντα διαφοροποιείται σε σχέση με αυτό που προέκυψε από την συγκεντρωτική αξιολόγηση. Στον Πίνακα 3.12 εμφανίζονται τα αποτελέσματα του βέλτιστου μοντέλου για την βραχυπρόθεσμη πρόβλεψη των 3 μηνών ανα κατηγορία δημοτικής υποδομής.

A/A	Κατηγορία Υποδομής	y_true_overall (kWh	y_pred_overall (kWh)	R2	MAE (kWh)	RMSE (kWh)	Model
1	ΑΘΛΗΤΙΚΕΣ ΕΓΚΑΤΑΣΤΑΣΕΙΣ	86.702,14	90.047,26	0.804813	2622.127889	2741.935434	KNN_uniform
2	ΑΝΤΛΙΟΣΤΑΣΙΑ ΑΡΔΕΥΣΗΣ	9.155,33	61.445,99	-5302.886296	17430.230115	27999.476294	HistGB
3	ΑΝΤΛΙΟΣΤΑΣΙΑ ΥΔΡΕΥΣΗΣ	414.633,95	427.292,50	0.482173	7667.687856	9576.532603	ExtraTrees
4	ΔΗΜΟΤΙΚΑ ΙΑΤΡΕΙΑ	1.209,30	1.190,01	-0.914490	39.756787	48.840526	LinearSVR
5	ΔΗΜΟΤΙΚΑ ΚΟΙΜΗΤΗΡΙΑ	3.455,17	3.322,71	0.577571	131.222616	131.345737	LinearSVR
6	ΔΗΜΟΤΙΚΑ ΚΤΙΡΙΑ	31.610,47	30.764,79	0.481118	281.895878	367.492210	ExtraTrees
7	ΔΗΜΟΤΙΚΑ ΣΧΟΛΕΙΑ	90.604,10	87.550,97	-0.424248	2232.698300	2252.548579	HistGB
8	ΔΗΜΟΤΙΚΗ ΑΓΟΡΑ	1.072,45	1.133,13	-2.871245	20.226511	23.851162	ExtraTrees
9	ΕΓΚΑΤΑΣΤΑΣΗ ΕΠΕΞΕΡΓΑΣΙΑΣ ΛΥΜΑΤΩΝ	6.855,86	6.297,02	0.541914	230.512078	287.247114	ExtraTrees
10	ΕΡΓΑΤΙΚΕΣ ΚΑΤΟΙΚΙΕΣ	2.100,92	2.348,66	-28.627340	179.708012	231.492495	KNN_distance
11	ΚΑΠΗ	20.472,22	20.334,89	0.547262	313.566833	351.921957	Ridge
12	ΚΔΑΠ	1.632,25	1.424,03	-0.646568	98.972370	145.983453	KNN_distance
13	ΚΕΝΤΡΟ ΕΞΥΠΗΡΕΤΗΣΗΣ ΠΟΛΙΤΩΝ	7.357,74	7.296,94	0.538431	185.272756	210.068466	ExtraTrees
14	ΚΟΙΝΟΤΙΚΑ ΚΑΤΑΣΤΗΜΑΤΑ	62.762,37	61.597,35	0.322276	3920.617095	4320.383585	MLP_LBFGS
15	Λοιπά Κτίρια/Υποδομές	60.601,54	49.639,75	-1.941966	3653.947173	4351.717825	KNN_distance
16	ΝΗΠΙΑΓΩΓΕΙΑ	29.364,18	28.789,64	0.019049	402.709417	486.283988	RandomForest
17	ΟΔΟΣΗΜΑΝΣΗ	935,75	899,60	-1.556400	15.662989	24.164961	ExtraTrees
18	ΠΑΙΔΙΚΕΣ ΧΑΡΕΣ	0,00	0,00	1.000000	0.000000	0.000000	Ridge
19	ΠΟΛΙΤΙΣΜΟΣ	26.110,90	24.610,56	-1.420822	2688.412153	3012.450232	HistGB
20	ΣΧΟΛΕΙΑ Β'ΒΑΘΜΙΑΣ ΕΚΠΑΙΔΕΥΣΗΣ	104.104,83	103.356,98	0.739499	1254.104283	1390.503670	RandomForest
21	ΥΠΟΔΟΜΕΣ ΠΡΟΣΧΟΛΙΚΗΣ ΑΓΩΓΗΣ	13.141,09	13.515,11	0.297344	136.501391	224.251422	LinearSVR
22	ΦΩΤΙΣΜΟΣ ΟΔΩΝ ΠΛΑΤΕΙΩΝ	723.261,84	684.278,80	0.459323	14798.475543	23536.865414	KNN_distance
23	ΦΩΤΙΣΜΟΣ ΠΕΖΟΔΡΟΜΙΩΝ	97.103,99	97.772,07	0.327428	3333.575917	3702.057005	RandomForest
24	ΦΩΤΙΣΜΟΣ ΠΛΑΤΕΙΩΝ	37.944,63	38.909,03	0.314299	1302.530448	1586.415951	HistGB

Πίνακας 3.12 Βέλτιστο μοντέλο βραχυπρόθεσμης πρόβλεψης ανά κατηγορία

Στον Πίνακα 3.13 εμφανίζονται τα αποτελέσματα του βέλτιστου μοντέλου για την μεσοπρόθεσμη πρόβλεψη των 6 μηνών ανα κατηγορία δημοτικής υποδομής.

A/A	Κατηγορία Υποδομής	y_true_overall (kWh)	y_pred_overall (kWh)	R2	MAE (kWh)	RMSE (kWh)	model
1	ΑΘΛΗΤΙΚΕΣ ΕΓΚΑΤΑΣΤΑΣΕΙΣ	146.813,42	163.913,12	0.362335	4722.617889	5483.845141	KNN_uniform
2	ΑΝΤΛΙΟΣΤΑΣΙΑ ΑΡΔΕΥΣΗΣ	425.524,55	556.576,65	0.828128	43166.319426	61466.371028	HistGB
3	ΑΝΤΛΙΟΣΤΑΣΙΑ ΥΔΡΕΥΣΗΣ	911.502,89	899.021,44	0.617154	11320.020400	13086.952141	ExtraTrees
4	ΔΗΜΟΤΙΚΑ ΙΑΤΡΕΙΑ	2.281,63	2.187,31	-0.265509	32.808494	41.088561	LinearSVR
5	ΔΗΜΟΤΙΚΑ ΚΟΙΜΗΤΗΡΙΑ	6.122,22	5.870,59	0.694465	104.777533	110.514016	LinearSVR
6	ΔΗΜΟΤΙΚΑ ΚΤΙΡΙΑ	53.973,04	53.853,96	0.856267	542.894898	601.505519	Ridge
7	ΔΗΜΟΤΙΚΑ ΣΧΟΛΕΙΑ	151.326,27	145.771,76	0.762712	1951.734695	2545.480395	KNN_distance
8	ΔΗΜΟΤΙΚΗ ΑΓΟΡΑ	1.429,51	1.487,79	0.814969	45.768871	54.958381	LinearSVR
9	ΕΓΚΑΤΑΣΤΑΣΗ ΕΠΕΞΕΡΓΑΣΙΑΣ ΛΥΜΑΤΩΝ	12.316,56	12.376,36	0.503010	221.870811	277.380265	ExtraTrees
10	ΕΡΓΑΤΙΚΕΣ ΚΑΤΟΙΚΙΕΣ	3.868,60	3.922,37	-2.203800	149.324281	185.234046	KNN_distance
11	ΚΑΠΗ	32.519,35	30.684,10	0.799635	516.763526	772.469526	Ridge
12	ΚΔΑΠ	2.615,34	2.612,34	0.186715	112.082891	132.325390	LinearSVR
13	ΚΕΝΤΡΟ ΕΞΥΠΗΡΕΤΗΣΗΣ ΠΟΛΙΤΩΝ	11.983,15	11.886,34	0.661152	224.171702	303.163386	KNN_distance
14	ΚΟΙΝΟΤΙΚΑ ΚΑΤΑΣΤΗΜΑΤΑ	96.707,01	101.211,15	0.689947	2905.171836	3408.207909	MLP_LBFGS
15	Λοιπά Κτίρια/Υποδομές	113.393,53	117.830,64	-2.107216	2808.887162	3992.238824	MLP_LBFGS
16	ΝΗΠΙΑΓΩΓΕΙΑ	46.485,81	45.006,89	0.755442	921.057806	1033.730239	KNN_distance
17	ΟΔΟΣΗΜΑΝΣΗ	1.780,74	1.819,37	-0.578973	20.501333	27.709036	ExtraTrees
18	ΠΑΙΔΙΚΕΣ ΧΑΡΕΣ	96,30	69,08	0.411973	13.487967	17.136221	RandomForest
19	ΠΟΛΙΤΙΣΜΟΣ	39.792,70	33.167,54	0.197914	1705.120278	2327.921323	KNN_uniform
20	ΣΧΟΛΕΙΑ Β'ΒΑΘΜΙΑΣ ΕΚΠΑΙΔΕΥΣΗΣ	177.164,31	181.304,35	0.872466	1821.664161	2109.058178	ExtraTrees
21	ΥΠΟΔΟΜΕΣ ΠΡΟΣΧΟΛΙΚΗΣ ΑΓΩΓΗΣ	24.079,38	25.179,84	0.579264	193.159058	290.117695	LinearSVR
22	ΦΩΤΙΣΜΟΣ ΟΔΩΝ ΠΛΑΤΕΙΩΝ	1.228.623,09	1.220.069,79	0.830413	12470.886853	18349.616941	KNN_distance
23	ΦΩΤΙΣΜΟΣ ΠΕΖΟΔΡΟΜΙΩΝ	164.599,80	161.220,83	0.673701	2317.126462	3381.107753	KNN_distance
24	ΦΩΤΙΣΜΟΣ ΠΛΑΤΕΙΩΝ	61.800,83	62.201,05	0.762152	1020.694654	1329.079804	LinearSVR

Πίνακας 3.13 Βέλτιστο μοντέλο μεσοπρόθεσμης πρόβλεψης ανά κατηγορία

Στον Πίνακα 3.14 εμφανίζονται τα αποτελέσματα του βέλτιστου μοντέλου για την μακροπρόθεσμη πρόβλεψη των 12 μηνών ανα κατηγορία δημοτικής υποδομής.

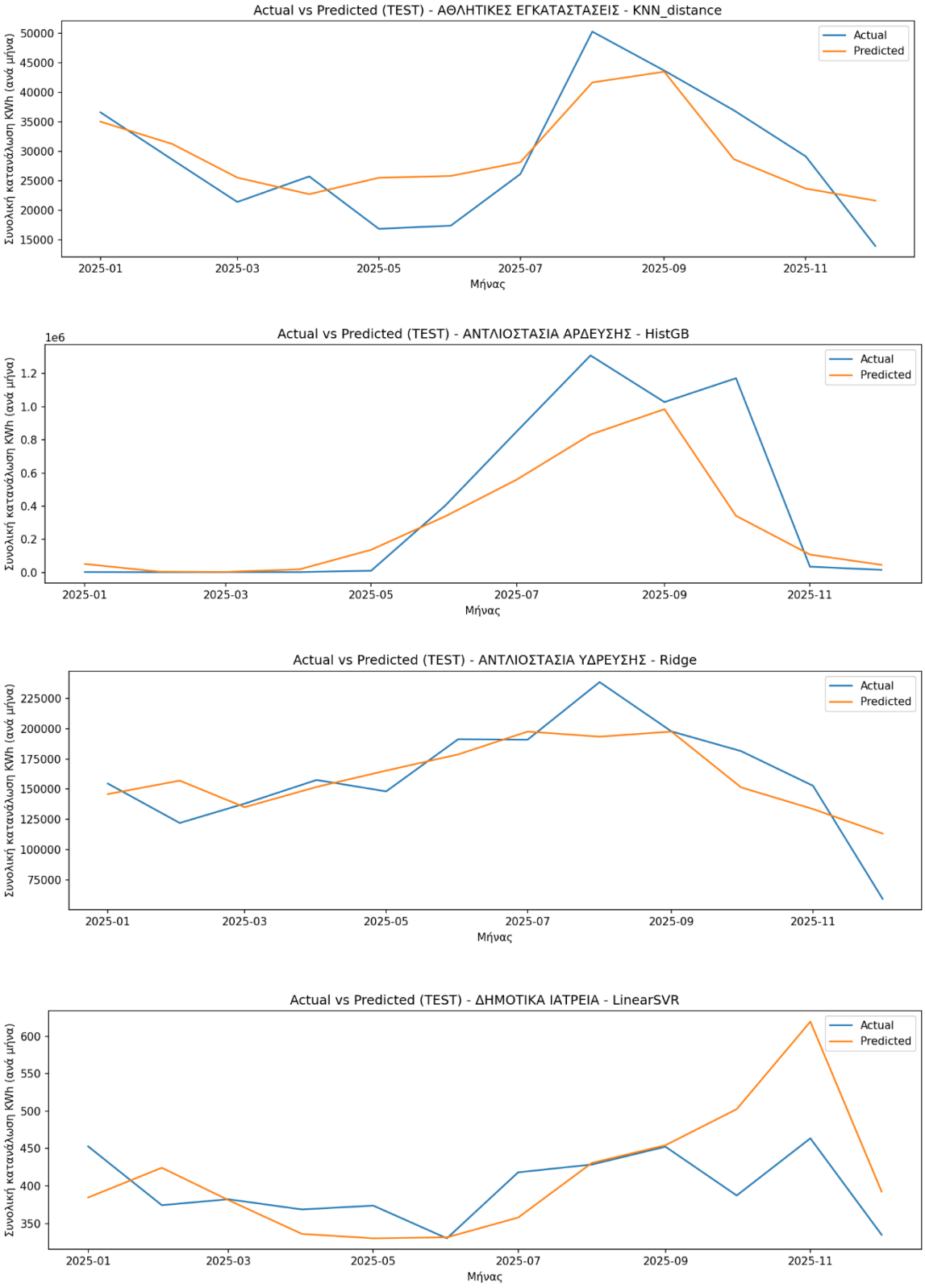
Α/Α	Κατηγορία Υποδομής	y_true_overall (kWh)	y_pred_overall (kWh)	R2	MAE (kWh)	RMSE (kWh)	model
1	ΑΘΛΗΤΙΚΕΣ ΕΓΚΑΤΑΣΤΑΣΕΙΣ	347.112,30	353.555,13	0.696327	5054.960611	5895.103709	KNN_distance
2	ΑΝΤΛΙΟΣΤΑΣΙΑ ΑΡΔΕΥΣΗΣ	4.834.823,24	3.430.778,20	0.665644	166735.955714	292880.097510	HistGB
3	ΑΝΤΛΙΟΣΤΑΣΙΑ ΥΔΡΕΥΣΗΣ	1.932.080,16	1.920.375,91	0.638084	19765.278667	25897.161213	Ridge
4	ΔΗΜΟΤΙΚΑ ΙΑΤΡΕΙΑ	4.766,26	4.944,45	-1.420237	49.177405	67.508959	LinearSVR
5	ΔΗΜΟΤΙΚΑ ΚΟΙΜΗΤΗΡΙΑ	12.696,73	12.798,97	-0.265585	154.665940	195.946138	LinearSVR
6	ΔΗΜΟΤΙΚΑ ΚΤΙΡΙΑ	102.618,77	107.580,20	0.496131	1101.716379	1323.447913	MLP_Adam
7	ΔΗΜΟΤΙΚΑ ΣΧΟΛΕΙΑ	266.062,13	261.948,13	0.652593	2750.808505	3638.095357	KNN_distance
8	ΔΗΜΟΤΙΚΗ ΑΓΟΡΑ	2.224,45	2.249,29	0.826808	35.134126	45.387548	LinearSVR
9	ΕΓΚΑΤΑΣΤΑΣΗ ΕΠΕΞΕΡΓΑΣΙΑΣ ΛΥΜΑΤΩΝ	20.831,88	21.522,67	0.616358	228.652425	300.938899	ExtraTrees
10	ΕΡΓΑΤΙΚΕΣ ΚΑΤΟΙΚΙΕΣ	6.622,09	7.669,94	-2.205738	210.002606	289.311257	LinearSVR
11	ΚΑΠΗ	52.920,52	49.370,91	0.597800	935.429865	1063.260626	KNN_distance
12	ΚΔΑΠ	5.179,52	5.163,85	0.002698	129.204569	156.259287	LinearSVR
13	ΚΕΝΤΡΟ ΕΞΥΠΗΡΕΤΗΣΗΣ ΠΟΛΙΤΩΝ	21.903,18	21.702,52	0.644288	337.957397	443.850947	LinearSVR
14	ΚΟΙΝΟΤΙΚΑ ΚΑΤΑΣΤΗΜΑΤΑ	219.911,43	236.961,00	0.433597	4315.288445	4942.071622	HistGB
15	Λοιπά Κτίρια/Υποδομές	211.062,13	172.322,01	-0.530707	3881.031739	4440.390546	Ridge
16	ΝΗΠΙΑΓΩΓΕΙΑ	78.902,80	82.500,71	0.647465	931.398246	1311.987333	KNN_distance
17	ΟΔΟΣΗΜΑΝΣΗ	3.393,23	3.604,01	-0.427119	25.254544	39.319873	ExtraTrees
18	ΠΑΙΔΙΚΕΣ ΧΑΡΕΣ	418,29	426,93	0.382331	17.025071	22.074203	RandomForest
19	ΠΟΛΙΤΙΣΜΟΣ	69.876,62	64.136,27	0.283261	1386.658639	1865.816048	KNN_uniform
20	ΣΧΟΛΕΙΑ Β'ΒΑΘΜΙΑΣ ΕΚΠΑΙΔΕΥΣΗΣ	290.890,21	311.046,77	0.806702	2814.024725	3470.552441	ExtraTrees
21	ΥΠΟΔΟΜΕΣ ΠΡΟΣΧΟΛΙΚΗΣ ΑΓΩΓΗΣ	45.611,80	47.247,62	0.122303	714.717659	846.077807	Ridge
22	ΦΩΤΙΣΜΟΣ ΟΔΩΝ ΠΛΑΤΕΙΩΝ	2.264.345,85	2.307.385,73	0.521689	18288.112351	31756.699396	KNN_distance
23	ΦΩΤΙΣΜΟΣ ΠΕΖΟΔΡΟΜΙΩΝ	283.565,96	316.485,44	0.343162	3521.003792	5389.591909	RandomForest
24	ΦΩΤΙΣΜΟΣ ΠΛΑΤΕΙΩΝ	127.118,95	123.635,60	0.352275	1600.509474	2036.814694	LinearSVR

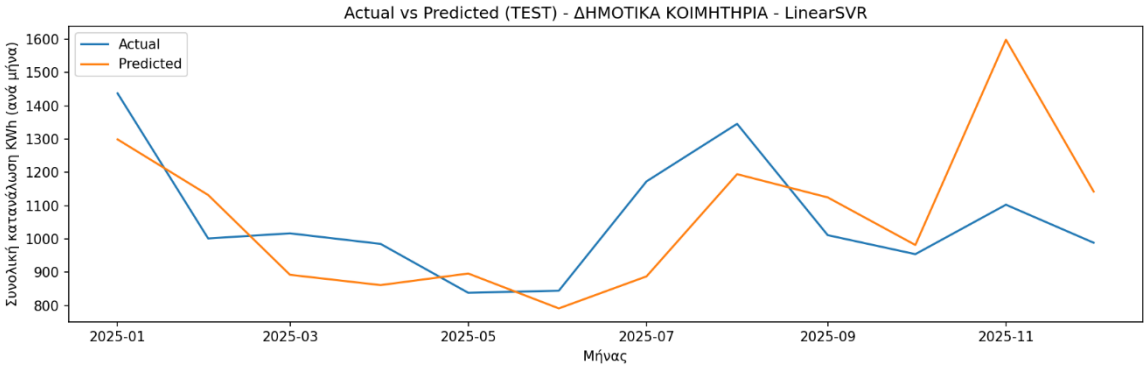
Πίνακας 3.14 Βέλτιστο μοντέλο μακροπρόθεσμης πρόβλεψης ανά κατηγορία

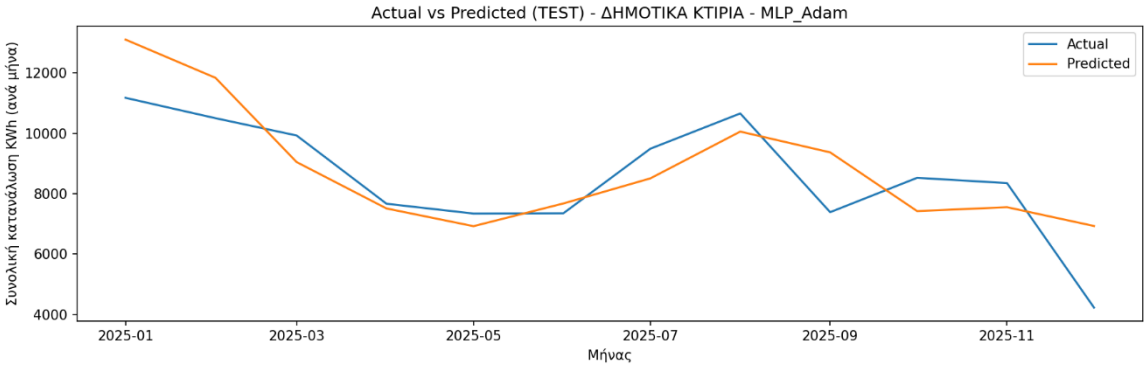
Κάθε λειτουργικός τομέας εμφανίζει διαφορετικά χαρακτηριστικά κατανάλωσης όπως διαφορετικό επίπεδο μεταβλητότητας, εποχικότητας, έντασης μη γραμμικότητας, παρουσία ακραίων τιμών (υποενότητα 3.6.3) και διαφορετικό πλήθος παροχών (Πίνακας 3.3). Αυτό έχει ως αποτέλεσμα οι κατηγορίες των δημοτικών υποδομών να εμφανίζουν διαφορετικό

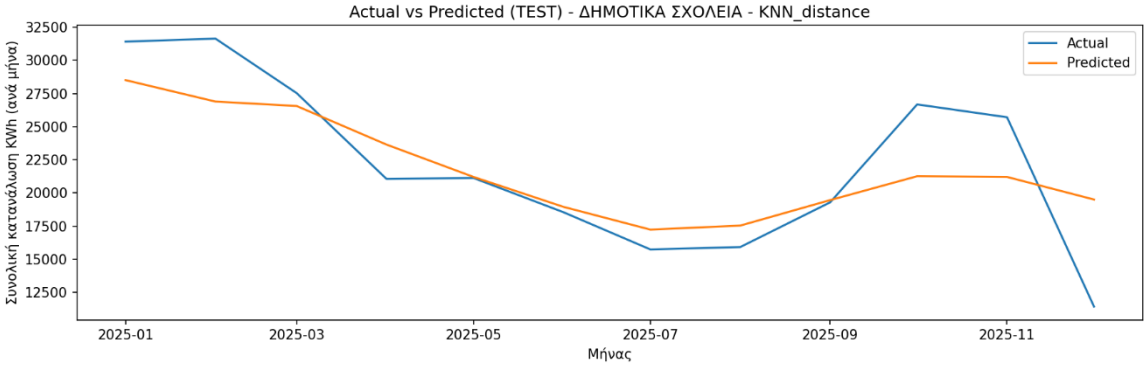
ενεργειακό προφίλ που εξαρτάται από τη δομή των επιμέρους δεδομένων. Στη βραχυπρόθεσμη πρόβλεψη των 3 μηνών (Πίνακας 3.12) το μοντέλο KNN_uniform που αναδείχθηκε ως το βέλτιστο στη συγκεντρωτική αξιολογηση του συνόλου των κατηγοριών εμφανίζεται πρώτο μόνο σε 1 από τις 24 κατηγορίες. Αντίστοιχα, στη μεσοπρόθεσμη πρόβλεψη των 6 μηνών το συνολικά βέλτιστο μοντέλο HistGB εμφανίζεται επίσης πρώτο μόνο σε 1 από τις 24 κατηγορίες ενώ στη μακροπρόθεσμη πρόβλεψη των 12 μηνών το HistGB διατηρεί την πρώτη θέση μόλις σε 2 από τις 24 κατηγορίες. Το εύρημα αυτό δείχνει ότι το μοντέλο που ελαχιστοποιεί το σφάλμα στη συνολική συγκεντρωτική αξιολογηση στο σύνολο των κατηγοριών διαφοροποιείται απο το μοντέλο που προσαρμόζεται καλύτερα στη δυναμική μιας συγκεκριμένης κατηγορίας. Η διαφοροποίηση του βέλτιστου μοντέλου σε επίπεδο κατηγορίας οφείλεται κυρίως στην ετερογένεια των προτύπων κατανάλωσης μεταξύ των κατηγοριών των δημοτικών υποδομών. Ειδικότερα, η συγκεντρωτική αξιολόγηση αναδεικνύει το μοντέλο που επιτυγχάνει τη βέλτιστη συνολική προσαρμογή στο άθροισμα όλων των μηνιαίων πραγματικών και προβλεπόμενων τιμών κατανάλωσης των παροχών για το σύνολο των κατηγοριών. Αντίθετα, η αξιολογηση σε επίπεδο επιμέρους κατηγορίας η οποία βασίζεται στο άθροισμα μόνο των μηνιαίων πραγματικών και προβλεπόμενων τιμών κατανάλωσης των παροχών της κατηγορίας αυτής, αποτυπώνει πιο έντονα τις ιδιαιτερότητες σχετικά με τη μεταβλητότητα, την εποχικότητα, τις μη γραμμικές σχέσεις και την παρουσία ακραίων τιμών του λειτουργικού τομέα που αντιπροσωπεύει η συγκεκριμένη κατηγορία. Επιπλέον, στην ενιαία συγκεντρωτική αξιολόγηση όλων των κατηγοριών οι διαθέσιμες μηνιαίες παρατηρήσεις των παροχων για κάθε χρονικό ορίζοντα είναι σημαντικά περισσότερες σε σχέση με τις αντίστοιχες μηνιαίες παρατηρήσεις των παροχών που προκύπτουν ανα κατηγορία δημοτικής υποδομής. Αυτό έχει ως αποτέλεσμα η άθροιση των μηνιαίων πραγματικών και προβλεπόμενων τιμών των παροχών στο σύνολο των κατηγοριών, να εξομαλύνει σε μεγαλύτερο βαθμό μέρος του θορύβου των επιμέρους χρονοσειρών ενώ η μετρική RMSE επηρεάζεται περισσότερο από κατηγορίες με υψηλότερη κατανάλωση και μεγαλύτερες αποκλίσεις. Το γεγονός αυτό δημιουργεί αστάθεια στις μετρικές αξιολόγησης ιδίως σε κατηγορίες με χαμηλή διακύμανση ή μεμονωμένες ακραίες τιμές. Αυτό επιβεβαιώνεται και από το ότι σε αρκετές κατηγορίες ακόμη και το καλύτερο διαθέσιμο μοντέλο εξακολουθεί να εμφανίζει αρνητικές τιμές του συντελεστή προσδιορισμού R^2 σε όλους τους χρονικούς ορίζοντες και ιδιαίτερα στη βραχυπρόθεσμη πρόβλεψη των 3 μηνών. Συνεπώς το μοντέλο που διαχειρίζεται αποτελεσματικότερα το συνολικό βάρος των μεγάλων σφαλμάτων αναδεικνύεται βέλτιστο

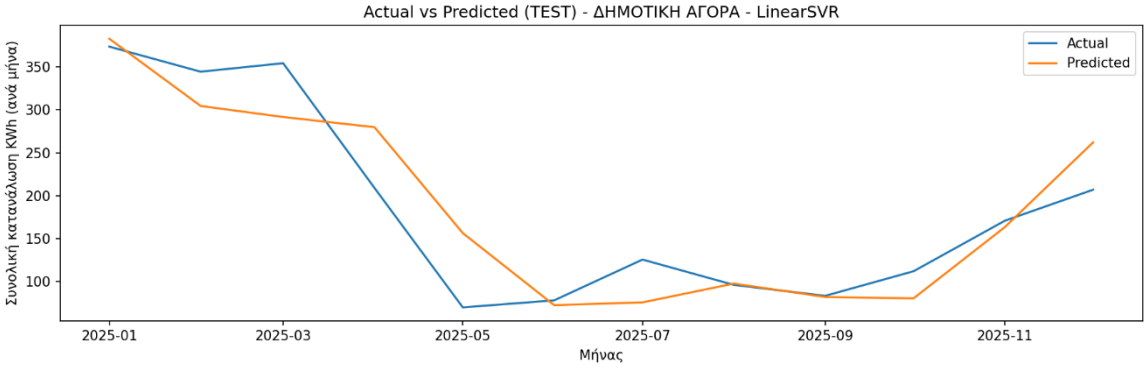
σε συγκεντρωτικό επίπεδο και όχι απαραίτητα το καταλληλότερο για κάθε επιμέρους κατηγορία. Στην Εικόνα 3.72 παρουσιάζονται τα διαγράμματα των μοντέλων που πέτυχαν την καλύτερη απόδοση σε κάθε κατηγορία ξεχωριστά στον χρονικό ορίζοντα των 12 μηνών.

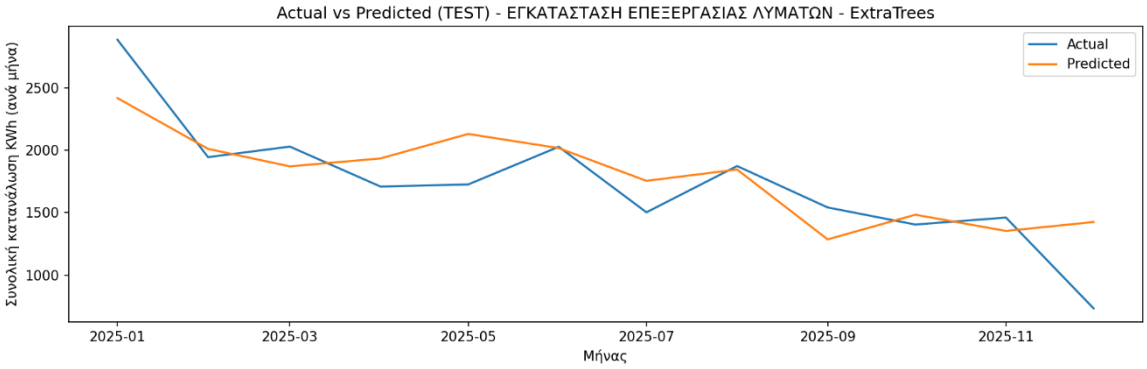


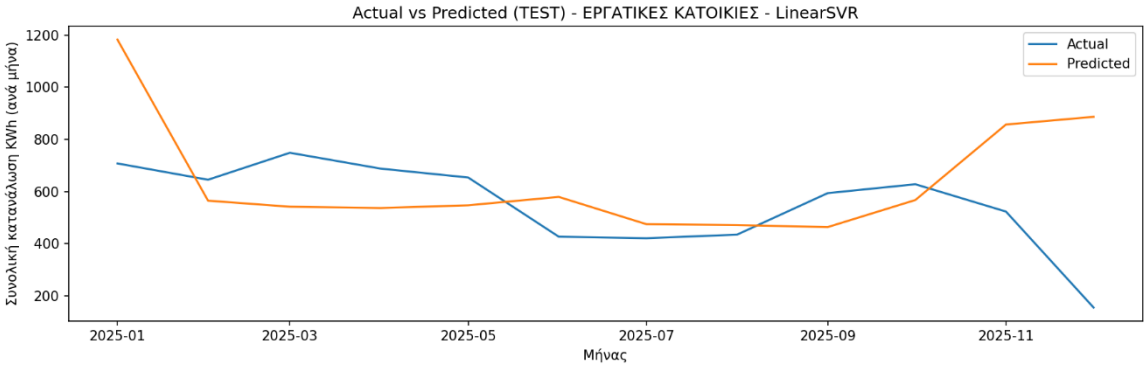

Actual vs Predicted (TEST) - ΔΗΜΟΤΙΚΑ ΚΟΙΜΗΤΗΡΙΑ - LinearSVR
Actual
Predicted
Συνολική καταναλωση KWh (ανά μήνα)
Μήνας

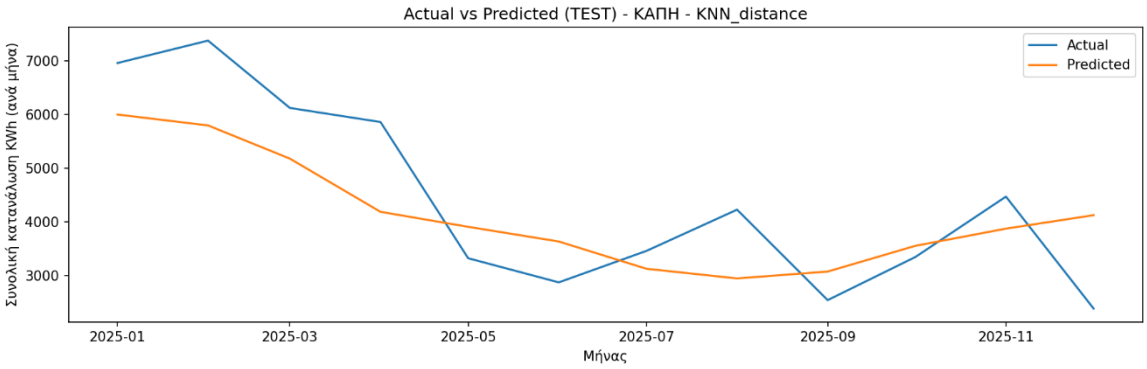

Actual vs Predicted (TEST) - ΔΗΜΟΤΙΚΑ ΚΤΙΡΙΑ - MLP_Adam
Actual
Predicted
Συνολική καταναλωση KWh (ανά μήνα)
Μήνας

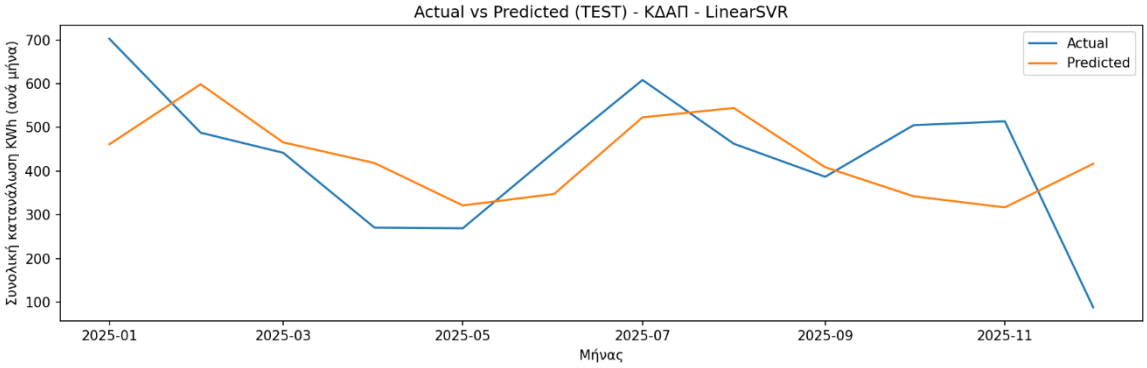

Actual vs Predicted (TEST) - ΔΗΜΟΤΙΚΑ ΣΧΟΛΕΙΑ - KNN_distance
Actual
Predicted
Συνολική καταναλωση KWh (ανά μήνα)
Μήνας

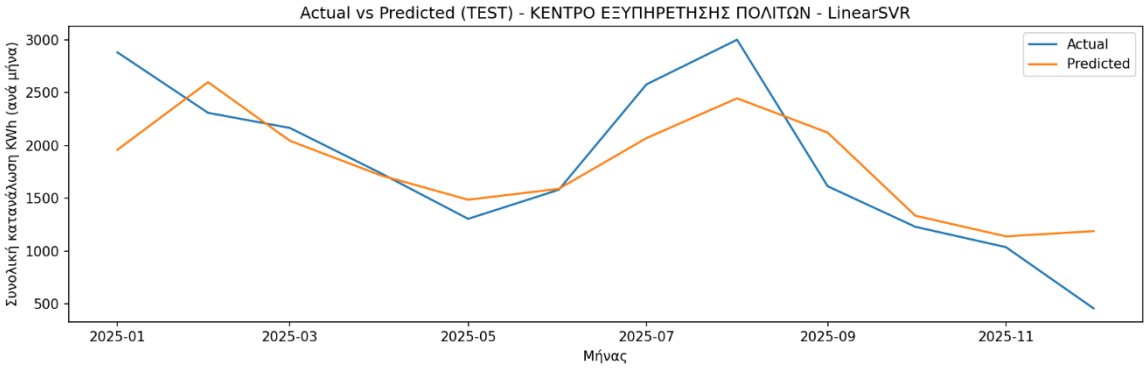

Actual vs Predicted (TEST) - ΔΗΜΟΤΙΚΗ ΑΓΟΡΑ - LinearSVR
Actual
Predicted
Συνολική καταναλωση KWh (ανά μήνα)
Μήνας

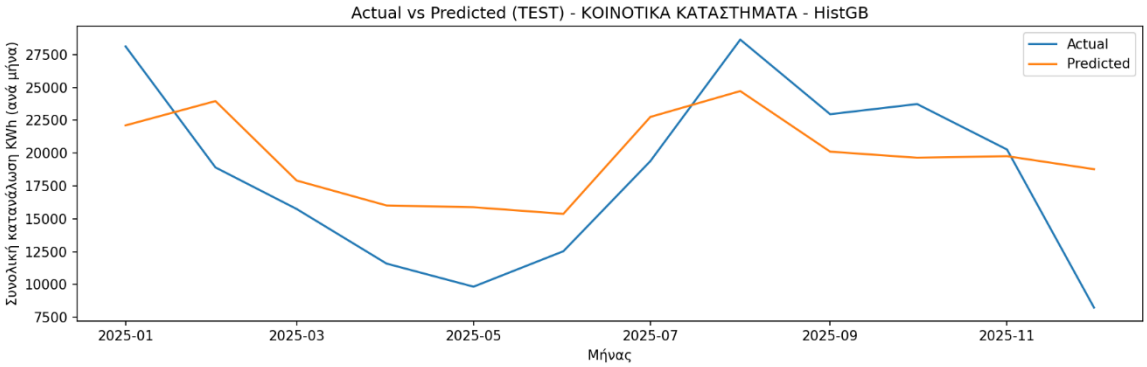

Actual vs Predicted (TEST) - ΕΓΚΑΤΑΣΤΑΣΗ ΕΠΕΞΕΡΓΑΣΙΑΣ ΛΥΜΑΤΩΝ - ExtraTrees
Actual
Predicted
Συνολική καταναλωση KWh (ανά μήνα)
2500
2000
1500
1000
2025-01
2025-03
2025-05
2025-07
2025-09
2025-11
Μήνας

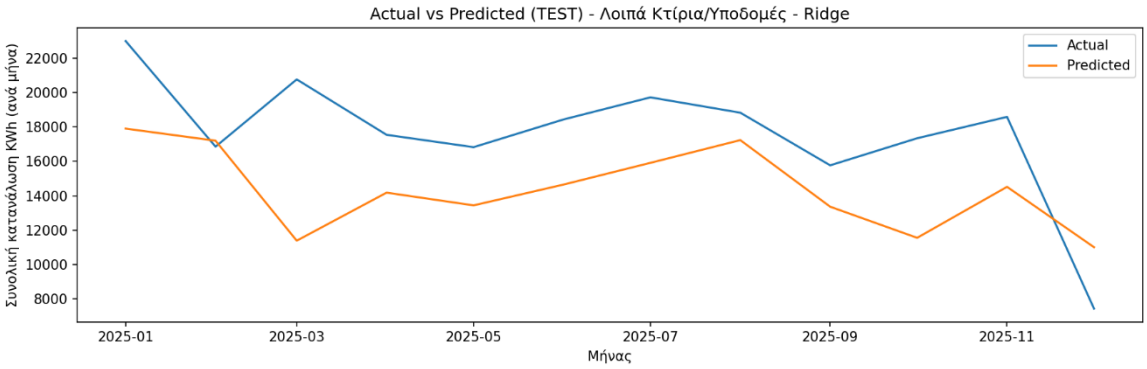

Actual vs Predicted (TEST) - ΕΡΓΑΤΙΚΕΣ ΚΑΤΟΙΚΙΕΣ - LinearSVR
Actual
Predicted
Συνολική καταναλωση KWh (ανά μήνα)
1200
1000
800
600
400
200
2025-01
2025-03
2025-05
2025-07
2025-09
2025-11
Μήνας

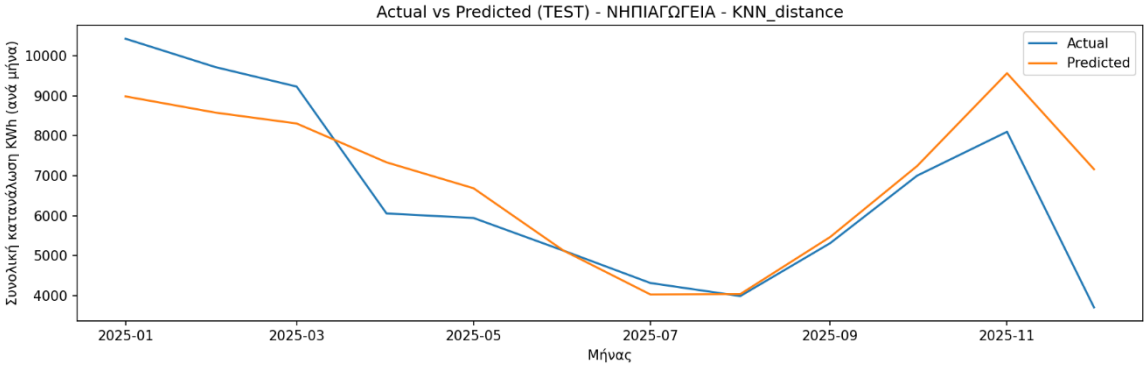

Actual vs Predicted (TEST) - ΚΑΠΗ - KNN_distance
Actual
Predicted
Συνολική καταναλωση KWh (ανά μήνα)
7000
6000
5000
4000
3000
2025-01
2025-03
2025-05
2025-07
2025-09
2025-11
Μήνας

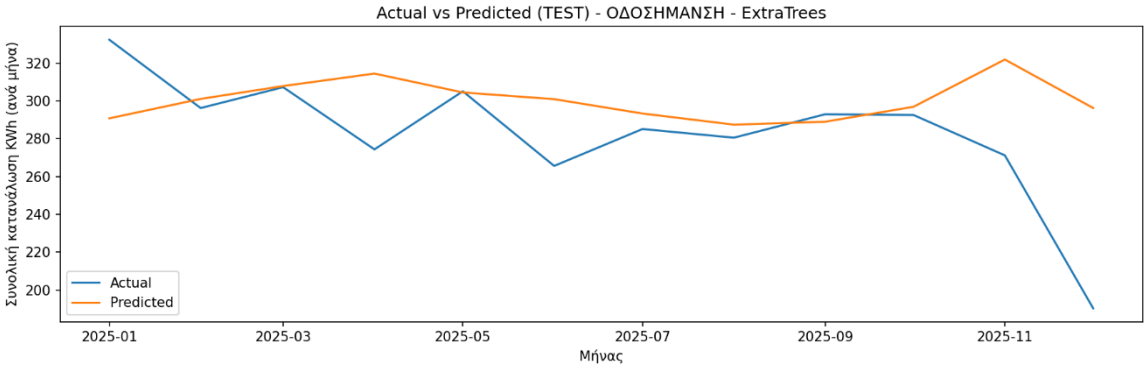

Actual vs Predicted (TEST) - ΚΔΑΠ - LinearSVR
Actual
Predicted
Συνολική καταναλωση KWh (ανά μήνα)
700
600
500
400
300
200
100
2025-01
2025-03
2025-05
2025-07
2025-09
2025-11
Μήνας

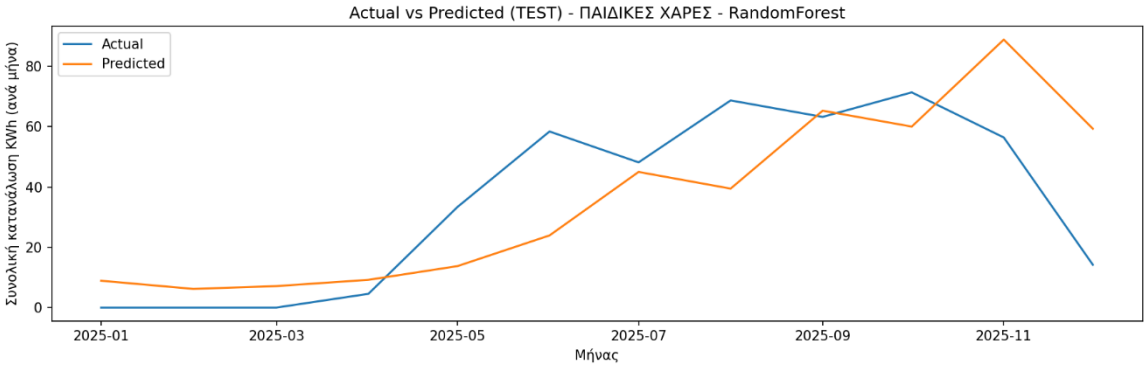

Actual vs Predicted (TEST) - ΚΕΝΤΡΟ ΕΞΥΠΗΡΕΤΗΣΗΣ ΠΟΛΙΤΩΝ - LinearSVR
Actual
Predicted
Συνολική κατανάλωση KWh (ανά μήνα)
3000
2500
2000
1500
1000
500
2025-01
2025-03
2025-05
2025-07
2025-09
2025-11
Μήνας

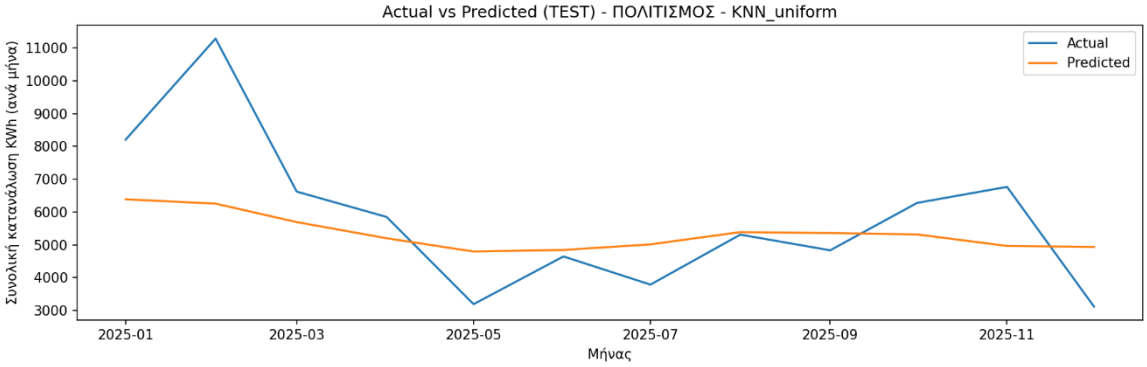

Actual vs Predicted (TEST) - ΚΟΙΝΟΤΙΚΑ ΚΑΤΑΣΤΗΜΑΤΑ - HistGB
Actual
Predicted
Συνολική κατανάλωση KWh (ανά μήνα)
27500
25000
22500
20000
17500
15000
12500
10000
7500
2025-01
2025-03
2025-05
2025-07
2025-09
2025-11
Μήνας

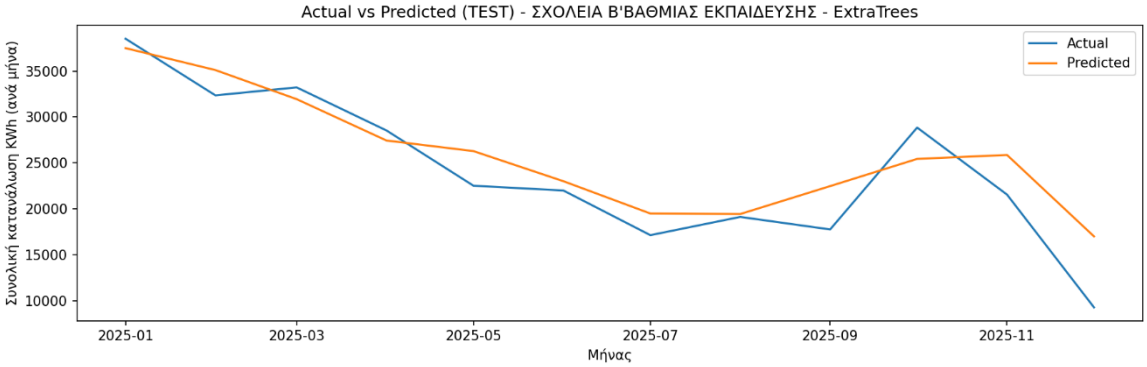

Actual vs Predicted (TEST) - Λοιπά Κτίρια/Υποδομές - Ridge
Actual
Predicted
Συνολική κατανάλωση KWh (ανά μήνα)
22000
20000
18000
16000
14000
12000
10000
8000
2025-01
2025-03
2025-05
2025-07
2025-09
2025-11
Μήνας

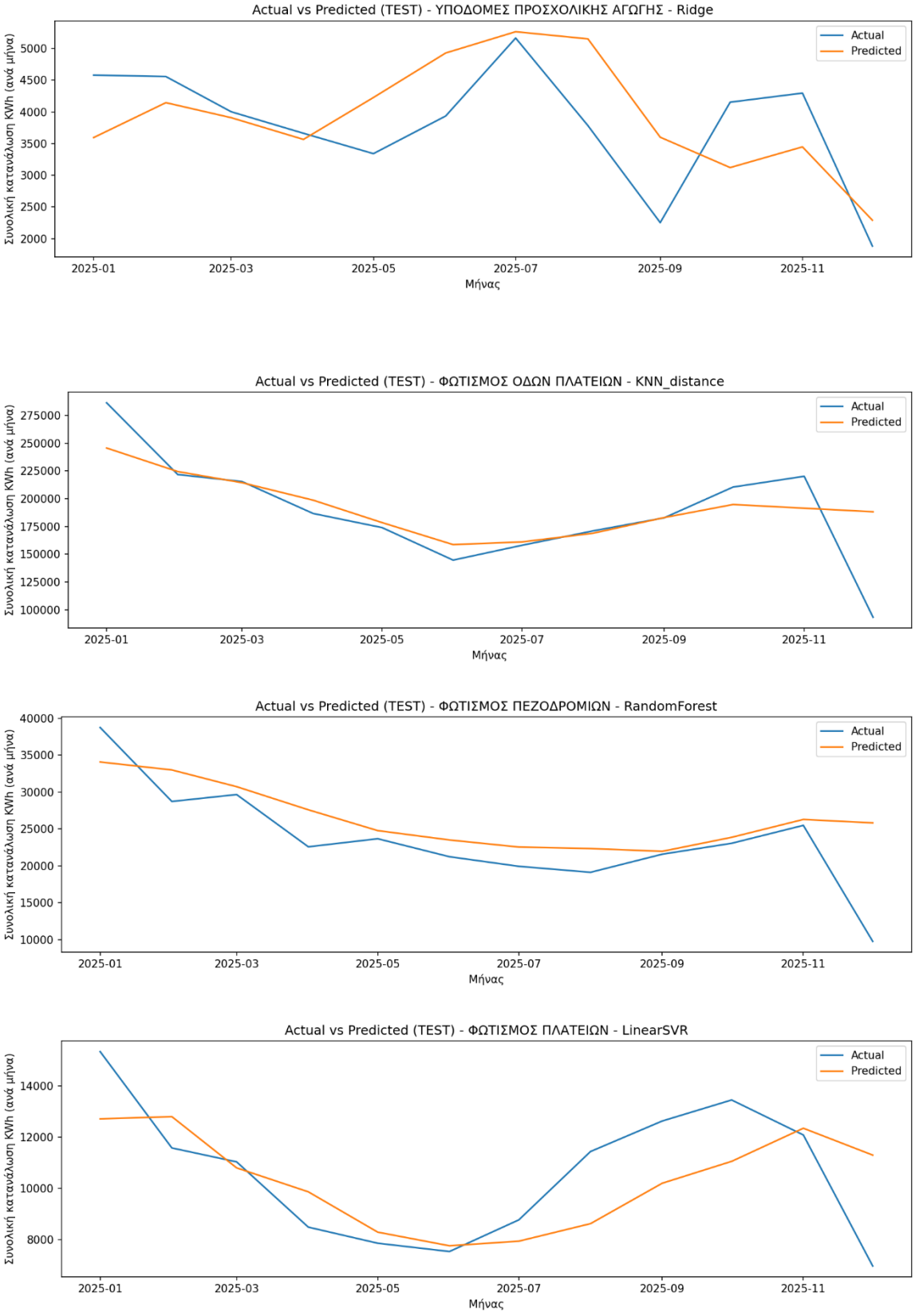

Actual vs Predicted (TEST) - ΝΗΠΙΑΓΩΓΕΙΑ - KNN_distance
Actual
Predicted
Συνολική κατανάλωση KWh (ανά μήνα)
10000
9000
8000
7000
6000
5000
4000
2025-01
2025-03
2025-05
2025-07
2025-09
2025-11
Μήνας


Actual vs Predicted (TEST) - ΟΔΟΣΗΜΑΝΣΗ - ExtraTrees
Συνολική καταναλωση KWh (ανά μήνα)
Μήνας
Actual
Predicted


Actual vs Predicted (TEST) - ΠΑΙΔΙΚΕΣ ΧΑΡΕΣ - RandomForest
Συνολική καταναλωση KWh (ανά μήνα)
Μήνας
Actual
Predicted


Actual vs Predicted (TEST) - ΠΟΛΙΤΙΣΜΟΣ - KNN_uniform
Συνολική καταναλωση KWh (ανά μήνα)
Μήνας
Actual
Predicted


Actual vs Predicted (TEST) - ΣΧΟΛΕΙΑ Β'ΒΑΘΜΙΑΣ ΕΚΠΑΙΔΕΥΣΗΣ - ExtraTrees
Συνολική καταναλωση KWh (ανά μήνα)
Μήνας
Actual
Predicted



Εικόνα 3.73 Διαγράμματα μακροπρόθεσμης πρόβλεψης βέλτιστου μοντέλου ανά κατηγορία

4. Συμπεράσματα και Μελλοντική Έρευνα

4.1 Συμπεράσματα

Στην παρούσα εργασία επιχειρήθηκε η ανάπτυξη και η αξιολόγηση ενός πλαισίου πρόβλεψης της **μηνιαίας κατανάλωσης ηλεκτρικής ενέργειας (kWh)** των δημοτικών κτιρίων και υποδομών ενός οργανισμού τοπικής αυτοδιοίκησης αξιοποιώντας τεχνικές μηχανικής μάθησης. Στο πλαίσιο αυτό αξιοποιήθηκαν πραγματικά δεδομένα ενεργειακής κατανάλωσης βασισμένα στους μηνιαίους λογαριασμούς του παρόχου ηλεκτρικής ενέργειας. Το γεγονός αυτό προσδίδει πρακτικό και επιχειρησιακό χαρακτήρα στην έρευνα αλλά ταυτόχρονα καθιστά ιδιαίτερα πολύπλοκη και απαιτητική τη διαδικασία προετοιμασίας και προεπεξεργασίας των πρωτογενών δεδομένων ώστε να καταστούν κατάλληλα για την είσοδό τους σε μοντέλα μηχανικής μάθησης.

Οι λογαριασμοί ηλεκτρικής ενέργειας εκδίδονται σε επίπεδο αριθμού παροχής των δημοτικών υποδομών που αποτελεί βασική παράμετρο για την ανάγκη της ομαδοποίησης τους ανά λειτουργικό τομέα (Αθλητισμός, Σχολεία, κτλ) λόγω του μεγάλου πλήθους. Η διαχείριση πραγματικών διοικητικών δεδομένων αυτού του τύπου απαιτεί ιδιαίτερη έμφαση στην ποιότητα, στη συνέπεια και στον κατάλληλο μετασχηματισμό τους καθώς αποτελούν βασικές παραμέτρους που επηρεάζουν άμεσα την αξιοπιστία των προβλεπτικών μοντέλων.

Στη συνέχεια εξετάστηκε η συμβολή των χαρακτηριστικών του αρχικού συνόλου δεδομένων στον προσδιορισμό της μεταβλητής-στόχου. Επιλέχθηκαν μόνο εκείνα τα χαρακτηριστικά που συντελούν στη διαμόρφωση της κατανάλωσης ηλεκτρικής ενέργειας ενώ δημιουργήθηκαν νέα χρονικά χαρακτηριστικά για την καλύτερη αποτύπωση της χρονικής αυτοσυσχέτισης των δεδομένων.

Μετά την επιλογή των αλγόριθμων μηχανικής μάθησης για προβλήματα παλινδρόμησης από διαφορετικές οικογένειες μοντέλων, η διαδικασία περιλάμβανε τη βελτιστοποίηση των βασικών υπερπαραμέτρων μέσω *grid search* στο αρχικό εκπαιδευτικό σύνολο σε κυλιόμενα χρονικά τμήματα εκπαίδευσης και επικύρωσης, με διατήρηση της χρονικής σειράς των παρατηρήσεων. Ακολούθησε επανεκπαίδευση (*refit*) των μοντέλων στο πλήρες εκπαιδευτικό σύνολο. Η αξιολόγηση της απόδοσής τους πραγματοποιήθηκε σε ένα πλήρως άγνωστο σύνολο ελέγχου για το διάστημα από 1/1/2025 έως 31/12/2025 σε τρεις χρονικούς

ορίζοντες: βραχυπρόθεσμο (3 μήνες), μεσοπρόθεσμο (6 μήνες) και μακροπρόθεσμο (12 μήνες).

Η διαδικασία εκπαίδευσης και πρόβλεψης πραγματοποιήθηκε σε επίπεδο παροχής και για λόγους επιχειρησιακής αξιολόγησης εφαρμόστηκε άθροιση (aggregation) των μηνιαίων πραγματικών τιμών και των προβλέψεων ανά μήνα και κατηγορία δημοτικής υποδομής. Με αυτή την προσέγγιση οι μετρικές απόδοσης αποτυπώνουν τις συνολικές μηνιαίες καταναλώσεις ανά κατηγορία με αποτέλεσμα να περιορίζεται ο θόρυβος που παρατηρείται σε επίπεδο μεμονωμένης παροχής. Η τελική αξιολόγηση στο σύνολο ελέγχου (test set) βασίστηκε σε 72 σημεία για τον χρονικό ορίζοντα των 3 μηνών, σε 144 σημεία για τον χρονικό ορίζοντα των 6 μηνών και σε 288 σημεία (24 κατηγορίες × 12 μήνες) για τον χρονικό ορίζοντα των 12 μηνών με χρήση των μετρικών R^2, MAE και RMSE.

Από τα αποτελέσματα της **συγκεντρωτικής αξιολόγησης** στο σύνολο ελέγχου προέκυψε ότι το βέλτιστο μοντέλο διαφοροποιείται ανάλογα με τον χρονικό ορίζοντα πρόβλεψης. Στον βραχυπρόθεσμο ορίζοντα των 3 μηνών οι παραλλαγές του μοντέλου kNN εμφάνισαν την καλύτερη απόδοση με το **KNN_uniform** να επιτυγχάνει το υψηλότερο **R^2 = 0.961858** και το χαμηλότερο **RMSE = 10424.20** ενώ το **KNN_distance** πέτυχε το χαμηλότερο **MAE = 3697.86**. Στον μεσοπρόθεσμο ορίζοντα των 6 μηνών το **HistGradientBoosting (HistGB)** κατέγραψε το καλύτερο **R^2 = 0.929015** και το χαμηλότερο **RMSE = 15620.79** ενώ το **LinearSVR** ακολούθησε πολύ κοντά ως προς το R^2 (**0.926642**) και πέτυχε το χαμηλότερο **MAE = 5110.17**. Στον μακροπρόθεσμο ορίζοντα των 12 μηνών το **HistGB** αναδείχθηκε σαφώς πρώτο επιτυγχάνοντας την καλύτερη επίδοση και στις τρεις μετρικές (**R^2 = 0.802913, MAE = 10760.46, RMSE = 60911.42**) (Πίνακας 3.11). Τα αποτελέσματα αυτά δείχνουν ότι όσο αυξάνεται ο χρονικός ορίζοντας πρόβλεψης η ακρίβεια όλων των μοντέλων μειώνεται, ωστόσο το HistGB εμφανίζει τη μεγαλύτερη σταθερότητα και ανθεκτικότητα στη μεσοπρόθεσμη και μακροπρόθεσμη πρόβλεψη.

Συνεπώς, σε επιχειρησιακό και διοικητικό επίπεδο για τον βραχυπρόθεσμο σχεδιασμό και τον άμεσο έλεγχο των αποκλίσεων ανά λειτουργικό τομέα τα μοντέλα kNN αποδείχθηκαν ιδιαίτερα αποτελεσματικά. Αντίθετα για τον μεσοπρόθεσμο και μακροπρόθεσμο ενεργειακό προγραμματισμό το μοντέλο HistGB αναδεικνύεται ως η καταλληλότερη επιλογή για ενιαία εφαρμογή πρόβλεψης.

Τα αποτελέσματα της **συγκριτικής αξιολόγησης ανά κατηγορία** δημοτικής υποδομής έδειξαν ότι το βέλτιστο μοντέλο διαφοροποιείται σε κάθε χρονικό ορίζοντα γεγονός που υποδηλώνει την ύπαρξη ετερογένειας στα πρότυπα κατανάλωσης κάθε λειτουργικού τομέα. Χαρακτηριστικά το μοντέλο που αναδείχθηκε καλύτερο στη συγκεντρωτική αξιολόγηση δεν επικράτησε στις περισσότερες επιμέρους κατηγορίες. Στον ορίζοντα των 3 μηνών το **KNN_uniform** παρότι πρώτο στη συνολική συγκεντρωτική αξιολόγηση, αναδείχθηκε πρώτο μόλις σε 1 από τις 24 κατηγορίες ενώ τα **ExtraTrees** σε 6 κατηγορίες, τα **KNN_distance** σε 4 κατηγορίες και το **HistGB** επίσης σε 4 κατηγορίες. Στον ορίζοντα των 6 μηνών το **HistGB** παρότι πρώτο συγκεντρωτικά, ανήλθε πρώτο μόλις σε 1 από τις 24 κατηγορίες ενώ τα **KNN_distance** και **LinearSVR** σε 6 κατηγορίες το καθένα. Στον ορίζοντα των 12 μηνών το **HistGB** ήταν πρώτο μόνο σε 2 από τις 24 κατηγορίες ενώ το **LinearSVR** αναδείχθηκε πρώτο σε 7 κατηγορίες και το **KNN_distance** σε 5 κατηγορίες. Το εύρημα αυτό επιβεβαιώνει ότι το μοντέλο που βελτιστοποιεί τη συνολική συγκεντρωτική επίδοση δεν συμπίπτει αναγκαστικά με το μοντέλο που αποτυπώνει καλύτερα τη δυναμική μιας συγκεκριμένης κατηγορίας υποδομής.

Παράλληλα τα αποτελέσματα ανά κατηγορία έδειξαν ότι υπάρχουν κατηγορίες για τις οποίες η πρόβλεψη είναι σαφώς πιο αξιόπιστη από ό,τι σε άλλες. Για παράδειγμα, τα **Σχολεία Β΄βάθμιας Εκπαίδευσης** εμφάνισαν σταθερά υψηλές τιμές R^2 και στους τρεις χρονικούς ορίζοντες ενώ αντίθετα κατηγορίες όπως τα **Δημοτικά Ιατρεία**, οι **Εργατικές Κατοικίες** και η **Οδοσήμανση,** παρουσίασαν αρνητικές ή πολύ χαμηλές τιμές R^2 ακόμη και με το κατά περίπτωση καλύτερο μοντέλο.

Με βάση τα παραπάνω εξάγονται δύο βασικά συμπεράσματα. Πρώτον, η μηχανική μάθηση μπορεί να υποστηρίξει αποτελεσματικά την πρόβλεψη της μηνιαίας ενεργειακής κατανάλωσης των δημοτικών υποδομών ιδιαίτερα σε συγκεντρωτικό επιχειρησιακό επίπεδο. Δεύτερον, η επιλογή του κατάλληλου μοντέλου εξαρτάται από το επίπεδο λήψης των αποφάσεων. Εάν ο στόχος είναι η υποστήριξη των αποφάσεων της Διοίκησης για τον συνολικό ενεργειακό προγραμματισμό του δήμου, το μοντέλο HistGB προκύπτει ως η πιο ισχυρή και σταθερή επιλογή για τους μεσοπρόθεσμους και μακροπρόθεσμους χρονικούς ορίζοντες. Εάν ο στόχος είναι η μέγιστη δυνατή ακρίβεια για την υποστήριξη και τον σχεδιασμό εξατομικευμένων τεχνικών παρεμβάσεων εξοικονόμησης ενέργειας ανά λειτουργικό τομέα, τότε απαιτείται διαφορετική προσέγγιση και επιλογή μοντέλου ανά κατηγορία προς υποστήριξη των αρμόδιων τεχνικών υπηρεσιών του δήμου.

Παρότι τα αποτελέσματα αναδεικνύουν την ικανότητα των προβλεπτικών μοντέλων η μελέτη υπόκειται σε ορισμένους περιορισμούς.

Η αξιολόγηση πραγματοποιήθηκε σε συγκεκριμένο οργανισμό τοπικής αυτοδιοίκησης και σε συγκεκριμένη χρονική περίοδο γεγονός που περιορίζει τη δυνατότητα άμεσης γενίκευσης των ευρημάτων χωρίς πρόσθετη διερεύνηση σε άλλους δήμους.

Η ανάλυση βασίστηκε αποκλειστικά στους μηνιαίους λογαριασμούς κατανάλωσης όπως αυτοί εκδίδονται από τον πάροχο ηλεκτρικής ενέργειας και όχι σε δεδομένα υψηλότερης χρονικής ανάλυσης όπως ημερήσιες ή ωριαίες μετρήσεις.

Τέλος, δεν υπήρχε δυνατότητα ενσωμάτωσης κρίσιμων εξωγενών μεταβλητών όπως μετεωρολογικά δεδομένα, χαρακτηριστικά κτιρίων, ωράρια λειτουργίας ή ειδικά τοπικά γεγονότα τα οποία θα μπορούσαν να βελτιώσουν περαιτέρω την ερμηνευτική και προβλεπτική ικανότητα των μοντέλων.

Παράλληλα η αυτοματοποιημένη λήψη αποφάσεων στον δημόσιο τομέα συνοδεύεται από κρίσιμα ηθικά διλήμματα όπως η αδιαφάνεια των αλγοριθμικών αποφάσεων, η δυσχέρεια απόδοσης λογοδοσίας και ο κίνδυνος μεροληψίας και διακρίσεων. Για τον λόγο αυτό η ενσωμάτωση τέτοιων τεχνολογιών στον δημόσιο τομέα προϋποθέτει όχι μόνο υψηλή προβλεπτική ακρίβεια αλλά και την ύπαρξη σαφούς θεσμικού πλαισίου που να διασφαλίζει τη διαφάνεια, τη λογοδοσία και τον ουσιαστικό ανθρώπινο έλεγχο.

Η αξιοποίηση μοντέλων μηχανικής μάθησης για την πρόβλεψη της ενεργειακής κατανάλωσης των δημοτικών υποδομών απαιτεί επαρκή ερμηνευσιμότητα ώστε τα αποτελέσματα να είναι κατανοητά και αξιοποιήσιμα από τη Διοίκηση. Για τον λόγο αυτό μελλοντική έρευνα μπορεί να εστιάσει στη συστηματική εφαρμογή τεχνικών ερμηνευσιμότητας (explainable AI), οι οποίες θα τεκμηριώνουν τη συμβολή συγκεκριμένων χαρακτηριστικών εισόδου απαραίτητων για την ενίσχυση της αξιοπιστίας των προβλεπτικών μοντέλων.

Παράλληλα, κρίσιμο πεδίο περαιτέρω διερεύνησης αποτελεί η μεθοδολογική τεκμηρίωση της αναγκαίας μοντελοποίησης των πραγματικών δεδομένων ενεργειακής κατανάλωσης για τη διασφάλιση της επιστημονικής εγκυρότητας των αποτελεσμάτων.

Τέλος, η ενσωμάτωση των προβλέψεων σε αναφορές ή και πίνακες ελέγχου (dashboards) μπορεί να μετατρέψει το προτεινόμενο μοντέλο σε εργαλείο υποστήριξης αποφάσεων

ενισχύοντας την αποδοχή και τη χρηστικότητα του σε πραγματικό επιχειρησιακό περιβάλλον από στελέχη της διοίκησης χωρίς εξειδικευμένη τεχνική γνώση.

4.2 Μελλοντική Έρευνα

Με βάση τα παραπάνω, προτείνονται οι ακόλουθες κατευθύνσεις για μελλοντική έρευνα και περαιτέρω βελτίωση του προτεινόμενου πλαισίου:

- **Εμπλουτισμός χαρακτηριστικών με εξωγενείς μεταβλητές**
  Η ενσωμάτωση μετεωρολογικών δεδομένων όπως η θερμοκρασία και οι βαθμοημέρες θέρμανσης/ψύξης καθώς και χαρακτηριστικών των κτιρίων όπως το εμβαδόν, η χρήση και το ωράριο λειτουργίας σε συνδυασμό με δεδομένα ειδικών γεγονότων όπως οι αργίες και οι σχολικές διακοπές, αναμένεται να βελτιώσει τόσο την ερμηνευσιμότητα όσο και την ακρίβεια των προβλέψεων.
- **Διερεύνηση εξειδικευμένης μοντελοποίησης ανά κατηγορία υποδομής**
  Η ανάπτυξη ξεχωριστών μοντέλων ανά κατηγορία ή ανά ομάδες κατηγοριών με παρόμοια ενεργειακή συμπεριφορά μπορεί να οδηγήσει σε καλύτερη προσαρμογή των προβλέψεων με βάση τις ιδιαιτερότητες κάθε λειτουργικού τομέα, δεδομένου ότι τα αποτελέσματα έδειξαν ότι διαφορετικοί αλγόριθμοι υπερέχουν σε διαφορετικές κατηγορίες δημοτικών υποδομών.
- **Ανίχνευση ανωμαλιών και επιχειρησιακή αξιοποίηση**
  Η αξιοποίηση των προβλέψεων για εντοπισμό αποκλίσεων ή ανωμαλιών όπως ασυνήθιστα υψηλές καταναλώσεις μπορεί να συμβάλει στην ανάπτυξη δεικτών έγκαιρης προειδοποίησης για την υποστήριξη στοχευμένων ενεργειακών παρεμβάσεων από τις αρμόδιες τεχνικές υπηρεσίες.

Συνολικά τα αποτελέσματα της παρούσας μελέτης δείχνουν ότι η πρόβλεψη της ενεργειακής κατανάλωσης των δημοτικών υποδομών με τεχνικές μηχανικής μάθησης είναι εφικτή και επιχειρησιακά χρήσιμη, με την προϋπόθεση ότι η επιλογή του κατάλληλου μοντέλου συνδέεται άμεσα με τον χρονικό ορίζοντα πρόβλεψης και το επίπεδο διοικητικής ή τεχνικής εφαρμογής στο οποίο πρόκειται να αξιοποιηθεί.

Βιβλιογραφία

Ahmad, M. W., Mourshed, M., & Rezgui, Y. (2017). Trees vs neurons: Comparison between random forest and ANN for high-resolution prediction of building energy consumption. *Energy and Buildings, 147*, 77–89. https://doi.org/10.1016/j.enbuild.2017.04.038

Athanasopoulos, G., & Kourentzes, N. (2023). On the evaluation of hierarchical forecasts. *International Journal of Forecasting, 39*(4), 1502–1511. https://doi.org/10.1016/j.ijforecast.2022.08.003

Cleveland, R. B., Cleveland, W. S., McRae, J. E., & Terpenning, I. J. (1990). STL: A seasonal-trend decomposition procedure based on loess. *Journal of Official Statistics, 6*(1), 3–33.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT Press. https://www.deeplearningbook.org/

Han, J., Kamber, M., & Pei, J. (2011). *Data mining: Concepts and techniques* (3rd ed.). Morgan Kaufmann.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (2nd ed.). Springer. https://link.springer.com/book/10.1007/978-0-387-84858-7

Hodrick, R. J., & Prescott, E. C. (1997). Postwar U.S. business cycles: An empirical investigation. *Journal of Money, Credit and Banking, 29*(1), 1–16. https://www.jstor.org/stable/2953682

Hyndman, R. J., & Athanasopoulos, G. (2018). *Forecasting: Principles and practice* (2nd ed.). OTexts. https://otexts.com/fpp2/

Irani, Z., Abril, R. M., Weerakkody, V., Omar, A., & Sivarajah, U. (2023). The impact of legacy systems on digital transformation in European public administration: Lesson learned from a multi case analysis. *Government Information Quarterly, 40*, Article 101784. https://doi.org/10.1016/j.giq.2022.101784

Janssen, M., van der Voort, H., & Wahyudi, A. (2017). Factors influencing big data decision-making quality. *Journal of Business Research, 70*, 338–345. https://doi.org/10.1016/j.jbusres.2016.08.007

Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. *Science, 349*(6245), 255–260. https://doi.org/10.1126/science.aaa8415

Kapoor, S., & Narayanan, A. (2023). Leakage and the reproducibility crisis in machine-learning-based science. *Patterns, 4*, Article 100804. https://doi.org/10.1016/j.patter.2023.100804

Kuhn, M., & Johnson, K. (2013). *Applied predictive modeling*. Springer. https://doi.org/10.1007/978-1-4614-6849-3

Lee, J., Kao, H.-A., & Yang, S. (2014). Service innovation and smart analytics for Industry 4.0 and big data environment. *Procedia CIRP, 16*, 3–8. https://doi.org/10.1016/j.procir.2014.02.001

Misuraca, G., & van Noordt, C. (2020). *Overview of the use and impact of AI in public services in the EU* (AI Watch, EUR 30255 EN, JRC120399). Luxembourg: Publications Office of the European Union. https://doi.org/10.2760/039619

Mitchell, T. M. (1997). *Machine learning*. McGraw-Hill.

Molnar, C. (2020). *Interpretable machine learning*. https://christophm.github.io/interpretable-ml-book/

Molnar, C., König, G., Bischl, B., & Casalicchio, G. (2024). Model-agnostic feature importance and effects with dependent features: A conditional subgroup approach. *Data Mining and Knowledge Discovery, 38*, 2903–2941. https://doi.org/10.1007/s10618-022-00901-9

OECD. (2020). *Regional policy for Greece post-2020 (OECD Territorial Reviews)*. OECD Publishing. https://doi.org/10.1787/cedf09a5-en

OECD. (2022). *Digital transformation projects in Greece's public sector: Governance, procurement and implementation (OECD Public Governance Reviews)*. OECD Publishing. https://doi.org/10.1787/33792fae-en

Plimakis, S. (2018). Strategic planning and service provision in Greek local government: A comparative assessment. *Advances in Social Sciences Research Journal, 5*(6), 470–486. https://doi.org/10.14738/assrj.56.4773

Ravn, M. O., & Uhlig, H. (2002). On adjusting the Hodrick–Prescott filter for the frequency of observations. *Review of Economics and Statistics, 84*(2), 371–380. https://doi.org/10.1162/003465302317411604

Shim, J. P., Warkentin, M., Courtney, J. F., Power, D. J., Sharda, R., & Carlsson, C. (2002). Past, present, and future of decision support technology. *Decision Support Systems, 33*(2), 111–126. https://doi.org/10.1016/S0167-9236(01)00139-7

Sprague, R. H. (1980). *A framework for the development of decision support systems*. *MIS Quarterly, 4*(4), 1-26. https://www.jstor.org/stable/248957

Wirtz, B. W., Weyerer, J. C., & Geyer, C. (2019). Artificial intelligence and the public sector—Applications and challenges. *International Journal of Public Administration, 42*(7), 596–615. https://doi.org/10.1080/01900692.2018.1498103

Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. (2016). *Data mining: Practical machine learning tools and techniques* (4th ed.). Morgan Kaufmann.

Βουλή των Ελλήνων. (2019). *Σύνταγμα της Ελλάδας (όπως αναθεωρήθηκε με το Ψήφισμα της 25ης Νοεμβρίου 2019, ΦΕΚ Α΄ 211/24.12.2019), άρθρο 102*.

Ελληνική Δημοκρατία. (2006). *Νόμος 3463/2006: Κύρωση του Κώδικα Δήμων και Κοινοτήτων (ΦΕΚ Α΄ 114/08.06.2006), άρθρο 75*.

Ελληνική Στατιστική Αρχή (ΕΛΣΤΑΤ). (2024). *Μητρώο φορέων γενικής κυβέρνησης – Κατάταξη κατά ESA 2010*. Αθήνα: Ελληνική Στατιστική Αρχή. https://www.statistics.gr/register-general-government-entities

Χλέπας, Ν.-Κ.Ε. (2020). *Ο Ευρωπαϊκός Χάρτης Τοπικής Αυτονομίας: Διαμόρφωση και Εφαρμογή μιας Διεθνούς Συνθήκης για την Τοπική Αυτοδιοίκηση*. Αθήνα: Εκδόσεις Παπαζήση.

Υπεύθυνη Δήλωση Συγγραφέα:

Δηλώνω ρητά ότι, σύμφωνα με το άρθρο 8 του Ν.1599/1986, η παρούσα εργασία αποτελεί αποκλειστικά προϊόν προσωπικής μου εργασίας, δεν προσβάλλει κάθε μορφής δικαιώματα διανοητικής ιδιοκτησίας, προσωπικότητας και προσωπικών δεδομένων τρίτων, δεν περιέχει έργα/εισφορές τρίτων για τα οποία απαιτείται άδεια των δημιουργών/δικαιούχων και δεν είναι προϊόν μερικής ή ολικής αντιγραφής, οι πηγές δε που χρησιμοποιήθηκαν περιορίζονται στις βιβλιογραφικές αναφορές και μόνον και πληρούν τους κανόνες της επιστημονικής παράθεσης.