ΧΑΡΟΚΟΠΕΙΟ ΠΑΝΕΠΙΣΤΗΜΙΟ
HAROKOPIO UNIVERSITY

Εξόρυξη Δεδομένων για την Ανάλυση και Πρόβλεψη Κατανάλωσης Ενέργειας σε ένα Έξυπνο Σπίτι

Ομάδα_7

Μπάλιου Ευαγγελία (it2022068@hua.gr),
Κουτσή Βασιλική - Μαρία (it2022149@hua.gr)

link : *Datamining Project*

ΠΕΡΙΕΧΟΜΕΝΑ

Εισαγωγή

Η παρούσα εργασία επικεντρώνεται στην εφαρμογή τεχνικών εξόρυξης δεδομένων για την ανάλυση και πρόβλεψη της κατανάλωσης ηλεκτρικής ενέργειας σε ένα έξυπνο οικιακό περιβάλλον. Στο πλαίσιο της αυξανόμενης ενεργειακής ζήτησης και της ανάπτυξης των έξυπνων δικτύων, η κατανόηση των προτύπων χρήσης ενέργειας σε επίπεδο νοικοκυριού αποτελεί κρίσιμο παράγοντα για τη βελτιστοποίηση της ενεργειακής αποδοτικότητας και την εξοικονόμηση πόρων. Για τον σκοπό αυτό, αξιοποιείται το σύνολο δεδομένων *Individual Household Electric Power Consumption*, το οποίο περιλαμβάνει εκτεταμένες μετρήσεις κατανάλωσης από πραγματικό νοικοκυριό σε βάθος χρόνου. Η ανάλυση πραγματοποιείται σε επίπεδο χρονοσειρών, αξιοποιώντας ημερήσια και υποημερήσια δεδομένα, με στόχο τόσο την πρόβλεψη όσο και την ερμηνεία της ενεργειακής συμπεριφοράς του νοικοκυριού.

Μέσα από διαδικασίες προεπεξεργασίας, καθαρισμού και εμπλουτισμού των δεδομένων, η εργασία αποσκοπεί στην εξαγωγή ουσιαστικής πληροφορίας και την ανακάλυψη κρυφών προτύπων κατανάλωσης. Παράλληλα, εφαρμόζονται μέθοδοι ταξινόμησης, παλινδρόμησης, ομαδοποίησης και εξόρυξης κανόνων συσχέτισης, με στόχο τόσο την πρόβλεψη μελλοντικής ενεργειακής χρήσης όσο και την ερμηνεία της συμπεριφοράς του νοικοκυριού. Τα αποτελέσματα της ανάλυσης φιλοδοξούν να συμβάλουν στην κατανόηση των ενεργειακών συνηθειών και να αναδείξουν δυνατότητες βελτίωσης της διαχείρισης της κατανάλωσης σε έξυπνα σπίτια.

Στόχος και μεθοδολογία

Ο βασικός στόχος της παρούσας εργασίας είναι η ανάλυση, η μοντελοποίηση και η πρόβλεψη της ημερήσιας κατανάλωσης ηλεκτρικής ενέργειας σε ένα έξυπνο οικιακό περιβάλλον, αξιοποιώντας τεχνικές εξόρυξης δεδομένων και ανάλυσης χρονοσειρών. Ιδιαίτερη έμφαση δόθηκε στην κατανόηση των προτύπων κατανάλωσης, στην αναγνώριση ημερών υψηλής ενεργειακής χρήσης και στη διερεύνηση της συμπεριφοράς του νοικοκυριού σε διαφορετικές χρονικές περιόδους. Ιδιαίτερη έμφαση δόθηκε στη χρονική διάσταση των δεδομένων, με τη χρήση χαρακτηριστικών υστέρησης (lag features), χρονολογικού διαχωρισμού εκπαίδευσης και ελέγχου, καθώς και εξειδικευμένων μεθόδων χρονοσειριακής πρόβλεψης.

Αρχικά, πραγματοποιήθηκε προσεκτική μελέτη και προεπεξεργασία του συνόλου δεδομένων, καθώς αυτό περιλάμβανε ελλιπείς τιμές, θόρυβο και ακανόνιστες χρονικές σφραγίδες. Στο στάδιο αυτό εφαρμόστηκαν τεχνικές καθαρισμού, μετασχηματισμού και εμπλουτισμού των δεδομένων, με σκοπό τη δημιουργία κατάλληλων χαρακτηριστικών που θα υποστήριζαν αποτελεσματικά τη μοντελοποίηση.

Στη συνέχεια, υιοθετήθηκε μια ολοκληρωμένη μεθοδολογική προσέγγιση, η οποία περιλάμβανε την εφαρμογή πολλαπλών τεχνικών εξόρυξης δεδομένων. Συγκεκριμένα, εφαρμόστηκαν μέθοδοι ταξινόμησης για τον εντοπισμό ημερών υψηλής κατανάλωσης ηλεκτρικής ενέργειας, παλινδρόμησης για την πρόβλεψη της κατανάλωσης της επόμενης ημέρας, καθώς και τεχνικές ομαδοποίησης για την αναγνώριση διακριτών προφίλ ημερήσιας κατανάλωσης.

Παράλληλα, υλοποιήθηκε εξόρυξη κανόνων συσχέτισης με σκοπό την ανακάλυψη σχέσεων μεταξύ της χρήσης υπομετρητών, της νυχτερινής κατανάλωσης και του αν μια ημέρα ανήκει σε Σαββατοκύριακο ή εργάσιμη. Τέλος, εφαρμόστηκαν μέθοδοι πρόβλεψης χρονοσειρών για την εκτίμηση μελλοντικών τιμών κατανάλωσης, επιτρέποντας τη σύγκριση στατιστικών και μη γραμμικών προσεγγίσεων.

Δομή της Εργασίας

Η εργασία θα ακολουθήσει την παρακάτω δομή:

- **Περιγραφή και Κατανόηση του Συνόλου Δεδομένων:**
  Παρουσίαση του συνόλου δεδομένων που χρησιμοποιήθηκε, των βασικών μεταβλητών του, καθώς και των κύριων προβλημάτων που εντοπίστηκαν, όπως ελλιπείς τιμές, θόρυβος και εποχικότητα.
- **Προεπεξεργασία Δεδομένων και Εξαγωγή Χαρακτηριστικών:**
  Ανάλυση των διαδικασιών καθαρισμού, μετασχηματισμού και εμπλουτισμού των δεδομένων, με στόχο τη δημιουργία κατάλληλων χαρακτηριστικών για τη μοντελοποίηση της κατανάλωσης ενέργειας.
- **Μεθοδολογία Εξόρυξης Δεδομένων & Πειραματική Διαδικασία και Μοντελοποίηση:**
  Παρουσίαση των τεχνικών εξόρυξης δεδομένων που εφαρμόστηκαν, συμπεριλαμβανομένης της ταξινόμησης, της παλινδρόμησης, της ομαδοποίησης και της εξόρυξης κανόνων συσχέτισης, καθώς και της λογικής πίσω από την επιλογή τους. Αναλυτική περιγραφή της διαδικασίας εκπαίδευσης και αξιολόγησης των μοντέλων, των μετρικών που χρησιμοποιήθηκαν και των παραμέτρων που εξετάστηκαν κατά τα πειράματα. Ιδιαίτερη έμφαση δίνεται στη σύγκριση διαφορετικών μοντέλων και στην επιλογή του βέλτιστου με βάση κατάλληλες μετρικές αξιολόγησης, λαμβάνοντας υπόψη τη χρονική φύση των δεδομένων.
- **Αποτελέσματα και Συζήτηση:**
  Παρουσίαση και ερμηνεία των αποτελεσμάτων που προέκυψαν από τις διάφορες τεχνικές εξόρυξης δεδομένων, με έμφαση στη σύγκριση των μοντέλων και στην κατανόηση των προτύπων κατανάλωσης ενέργειας.

Περιγραφή και Κατανόηση του Συνόλου Δεδομένων

Το σύνολο δεδομένων που χρησιμοποιήθηκε στην παρούσα εργασία είναι το Individual Household Electric Power Consumption, το οποίο προέρχεται από πραγματικές μετρήσεις κατανάλωσης ηλεκτρικής ενέργειας ενός οικιακού περιβάλλοντος στη Γαλλία.

Το σύνολο δεδομένων περιλαμβάνει μετρήσεις κατανάλωσης ηλεκτρικής ενέργειας ανά λεπτό και καλύπτει χρονικό διάστημα αρκετών ετών. Οι βασικές μεταβλητές περιλαμβάνουν τη συνολική ενεργή ισχύ (Global Active Power), η οποία εκφράζει τη συνολική στιγμιαία κατανάλωση του νοικοκυριού, καθώς και τρεις επιμέρους υπομετρητές κατανάλωσης (Sub_metering_1, Sub_metering_2 και Sub_metering_3). Οι υπομετρητές αυτοί αντιστοιχούν σε διαφορετικές ομάδες ηλεκτρικών συσκευών, όπως συσκευές κουζίνας, πλυντήρια και συστήματα θέρμανσης ή ψύξης, παρέχοντας μια πιο λεπτομερή εικόνα της κατανομής της κατανάλωσης ενέργειας στο οικιακό περιβάλλον.Η συνολική ενεργή ισχύς μετράται σε κιλοβάτ (kW) και καταγράφεται ανά λεπτό, ενώ οι υπομετρητές εκφράζουν επιμέρους καταναλώσεις σε βατώρες (Wh) ανά λεπτό, γεγονός που επιτρέπει τον υπολογισμό ημερήσιας κατανάλωσης ενέργειας σε κιλοβατώρες (kWh).

Κατά την αρχική διερεύνηση των δεδομένων διαπιστώθηκαν προβλήματα ελλιπών τιμών, παρουσίας μη έγκυρων συμβόλων, καθώς και ακανόνιστες χρονικές σφραγίδες. Επιπλέον, η χρονοσειριακή φύση των δεδομένων αναδεικνύει έντονα φαινόμενα εποχικότητας και περιοδικότητας, τα οποία καθιστούν απαραίτητη την κατάλληλη προεπεξεργασία πριν τη μοντελοποίηση.

Προεπεξεργασία Δεδομένων και Εξαγωγή Χαρακτηριστικών

Η προεπεξεργασία των δεδομένων αποτέλεσε κρίσιμο στάδιο της παρούσας εργασίας, καθώς το αρχικό σύνολο δεδομένων παρουσίαζε ελλιπείς τιμές, μη εγκυρες εγγραφές και ακανόνιστα χρονικά διαστήματα. Επιπλέον, η χρονοσειριακή φύση των δεδομένων απαιτούσε προσεκτικό χειρισμό, ώστε να διασφαλιστεί η χρονική συνέπεια πριν τη μοντελοποίηση.

1. Καθαρισμός και Μετασχηματισμός Δεδομένων

Αρχικά, τα δεδομένα φορτώθηκαν από το αρχείο κειμένου και πραγματοποιήθηκε έλεγχος της δομής τους, με στόχο τον εντοπισμό ελλιπών τιμών και ακατάλληλων τύπων δεδομένων. Οι μη έγκυρες τιμές, οι οποίες στο αρχικό σύνολο δεδομένων δηλώνονταν με ειδικά σύμβολα, μετατράπηκαν σε ελλείπουσες τιμές (NaN) ώστε να αντιμετωπιστούν με συστηματικό τρόπο.

	0
Global_active_power	25979
Sub_metering_2	25979
Global_reactive_power	25979
Voltage	25979
Global_intensity	25979
Sub_metering_3	25979
Sub_metering_1	25979
Date	0
Time	0

dtype: int64

Στη συνέχεια, οι μεταβλητές ημερομηνίας και ώρας συνενώθηκαν και μετατράπηκαν σε ενιαία χρονολογική μεταβλητή τύπου datetime, η οποία ορίστηκε ως χρονικός δείκτης (index) του συνόλου δεδομένων. Οι εγγραφές που δεν ήταν δυνατό να μετατραπούν σε έγκυρη χρονική μορφή απομακρύνθηκαν, διασφαλίζοντας τη σωστή χρονική σειρά των παρατηρήσεων.

Datetime	Global_active_power	Global_reactive_power	Voltage	Global_intensity	Sub_metering_1	Sub_metering_2	Sub_metering_3
2006-12-16 17:24:00	4.216	0.418	234.84	18.4	0.0	1.0	17.0
2006-12-16 17:25:00	5.360	0.436	233.63	23.0	0.0	1.0	16.0
2006-12-16 17:26:00	5.374	0.498	233.29	23.0	0.0	2.0	17.0
2006-12-16 17:27:00	5.388	0.502	233.74	23.0	0.0	1.0	17.0
2006-12-16 17:28:00	3.666	0.528	235.68	15.8	0.0	1.0	17.0

Όλες οι μεταβλητές κατανάλωσης μετατράπηκαν σε αριθμητική μορφή, ώστε να είναι δυνατός ο υπολογισμός στατιστικών μεγεθών και η εφαρμογή τεχνικών ανάλυσης χρονοσειρών.

```
<class 'pandas.core.frame.DataFrame'>
DatetimeIndex: 2075259 entries, 2006-12-16 17:24:00 to 2010-11-26 21:02:00
Data columns (total 7 columns):
 #   Column                 Dtype  
---  ------                 -----  
 0   Global_active_power    float64
 1   Global_reactive_power  float64
 2   Voltage                float64
 3   Global_intensity       float64
 4   Sub_metering_1         float64
 5   Sub_metering_2         float64
 6   Sub_metering_3         float64
dtypes: float64(7)
memory usage: 126.7 MB
```

2. Επαναδειγματοληψία (resampling) και Αντιμετώπιση Ελλιπών Τιμών

Για την αντιμετώπιση πιθανών χρονικών ασυνεπειών και κενών, εφαρμόστηκε επαναδειγματοληψία (resampling) σε βήμα ενός λεπτού. Η διαδικασία αυτή δημιούργησε ένα πλήρες χρονικό πλέγμα, στο οποίο οι αρχικές παρατηρήσεις ευθυγραμμίστηκαν χρονικά.

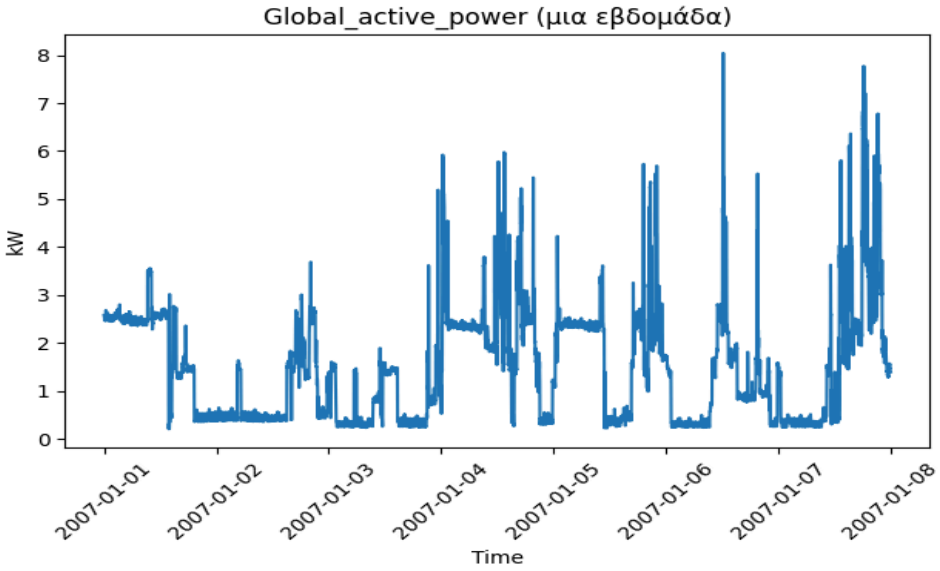
```
Αρχικές γραμμές: 2075259
Μετά το 1min resample: 2075259
```

	0
Global_active_power	0.012518
Global_reactive_power	0.012518
Voltage	0.012518
Global_intensity	0.012518
Sub_metering_1	0.012518
Sub_metering_2	0.012518
Sub_metering_3	0.012518

dtype: float64

Η επαναδειγματοληψία είχε ως αποτέλεσμα την εμφάνιση πρόσθετων ελλιπών τιμών, οι οποίες αντιμετωπίστηκαν με συνδυασμό τεχνικών παρεμβολής και συμπλήρωσης. Συγκεκριμένα, εφαρμόστηκε χρονική παρεμβολή για την κάλυψη μικρών κενών στη χρονοσειρά, ενώ για τις εναπομείνασες ελλείψεις χρησιμοποιήθηκαν τεχνικές προώθησης και οπισθοδρόμησης τιμών. Με τον τρόπο αυτό διασφαλίστηκε η συνέχεια των δεδομένων, χωρίς να εισαχθούν απότομα ή μη ρεαλιστικά άλματα στις μετρήσεις.

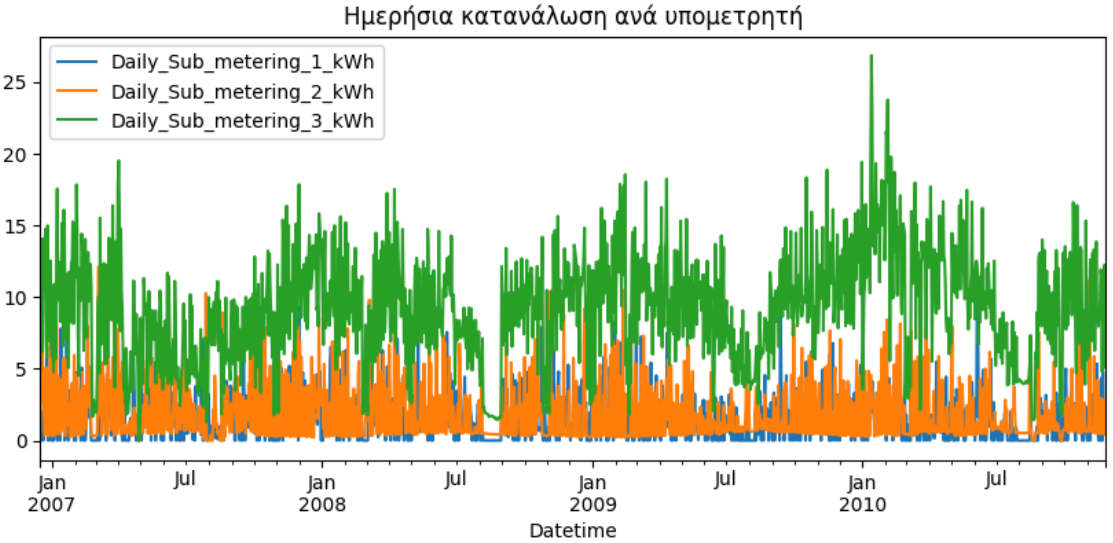
Η ορθότητα της διαδικασίας καθαρισμού επαληθεύτηκε μέσω γραφικής απεικόνισης της κατανάλωσης σε επιλεγμένα χρονικά διαστήματα, επιβεβαιώνοντας την ομαλή συμπεριφορά της χρονοσειράς.



3. Δημιουργία Ημερήσιων Μεταβλητών Κατανάλωσης

Μετά τον καθαρισμό των δεδομένων σε επίπεδο λεπτού, πραγματοποιήθηκε μετασχηματισμός σε ημερήσιο επίπεδο, ώστε να διευκολυνθεί η εφαρμογή των τεχνικών εξόρυξης δεδομένων. Η συνολική ημερήσια κατανάλωση ενέργειας υπολογίστηκε μετατρέποντας τη συνολική ενεργή ισχύ από κιλοβάτ (kW) σε κιλοβατώρες (kWh) και αθροίζοντας τις τιμές σε ημερήσια βάση.

Παράλληλα, υπολογίστηκε η ημερήσια κατανάλωση των επιμέρους υπομετρητών, οι οποίοι μετατράπηκαν από βατώρες (Wh) σε κιλοβατώρες (kWh). Ο έλεγχος συνέπειας μεταξύ της συνολικής κατανάλωσης και του αθροίσματος των υπομετρητών επιβεβαίωσε την ορθότητα των υπολογισμών.



```
(            Daily_total_kWh  Daily_Sub_metering_1_kWh  \
 Datetime
 2006-12-16        20.152933                     0.000
 2006-12-17        56.507667                     2.033
 2006-12-18        36.730433                     1.063
 2006-12-19        27.769900                     0.839
 2006-12-20        37.095800                     0.000

            Daily_Sub_metering_2_kWh  Daily_Sub_metering_3_kWh  Sub_sum_kWh
 Datetime
 2006-12-16                     0.546                     4.926        5.472
 2006-12-17                     4.187                    13.341       19.561
 2006-12-18                     2.621                    14.018       17.702
 2006-12-19                     7.602                     6.197       14.638
 2006-12-20                     2.648                    14.063       16.711
 (1442, 5))
```

4. Εμπλουτισμός Δεδομένων με Χρονικά και Συμπεριφορικά Χαρακτηριστικά

Για τον εμπλουτισμό του ημερήσιου συνόλου δεδομένων δημιουργήθηκαν επιπλέον χαρακτηριστικά. Συγκεκριμένα, προστέθηκαν μεταβλητές που περιγράφουν την ημέρα της εβδομάδας, το όνομα της ημέρας και την ένδειξη αν μια ημέρα ανήκει σε Σαββατοκύριακο ή εργάσιμη.

Datetime	Daily_total_kWh	Daily_Sub_metering_1_kWh	Daily_Sub_metering_2_kWh	Daily_Sub_metering_3_kWh	Sub_sum_kWh	day_of_week	day_name	is_weekend
2006-12-16	20.152933	0.000	0.546	4.926	5.472	5	Saturday	1
2006-12-17	56.507667	2.033	4.187	13.341	19.561	6	Sunday	1
2006-12-18	36.730433	1.063	2.621	14.018	17.702	0	Monday	0
2006-12-19	27.769900	0.839	7.602	6.197	14.638	1	Tuesday	0
2006-12-20	37.095800	0.000	2.648	14.063	16.711	2	Wednesday	0

Στη συνέχεια, βρέθηκε το Peak hour power, δηλαδή η μέγιστη μέση ωριαία ισχύς μέσα στη μέρα και προστέθηκε στον πίνακα.

Datetime	Daily_total_kWh	Daily_Sub_metering_1_kWh	Daily_Sub_metering_2_kWh	Daily_Sub_metering_3_kWh	Sub_sum_kWh	day_of_week	day_name	is_weekend	Nighttime_kWh	Peak_hour_power
2006-12-16	20.152933	0.000	0.546	4.926	5.472	5	Saturday	1	NaN	4.222889
2006-12-17	56.507667	2.033	4.187	13.341	19.561	6	Sunday	1	12.693833	3.697100
2006-12-18	36.730433	1.063	2.621	14.018	17.702	0	Monday	0	2.503900	3.050567
2006-12-19	27.769900	0.839	7.602	6.197	14.638	1	Tuesday	0	2.460200	3.879033
2006-12-20	37.095800	0.000	2.648	14.063	16.711	2	Wednesday	0	2.364600	3.646067

Επιπλέον, υπολογίστηκε η νυχτερινή κατανάλωση ενέργειας, η οποία ορίστηκε ως η κατανάλωση κατά το χρονικό διάστημα από τις 00:00 έως τις 06:00, παρέχοντας πληροφορία για τη χρήση ενέργειας εκτός ωρών αιχμής. Ως δείκτης έντασης χρήσης υπολογίστηκε επίσης η μέγιστη μέση ωριαία ισχύς κάθε ημέρας, η οποία χρησιμοποιήθηκε ως χαρακτηριστικό αιχμής κατανάλωσης.

Datetime	Daily_total_kWh	Daily_Sub_metering_1_kWh	Daily_Sub_metering_2_kWh	Daily_Sub_metering_3_kWh	Sub_sum_kWh	day_of_week	day_name	is_weekend	Nighttime_kWh
2006-12-16	20.152933	0.000	0.546	4.926	5.472	5	Saturday	1	NaN
2006-12-17	56.507667	2.033	4.187	13.341	19.561	6	Sunday	1	12.693833
2006-12-18	36.730433	1.063	2.621	14.018	17.702	0	Monday	0	2.503900
2006-12-19	27.769900	0.839	7.602	6.197	14.638	1	Tuesday	0	2.460200
2006-12-20	37.095800	0.000	2.648	14.063	16.711	2	Wednesday	0	2.364600

Επισης, βρέθηκε η κατανάλωση ενέργειας μόνο για Σαββατοκύριακα, οπου θεωρηθηκε τιμή ενέργειας όταν ειναι Σαββατοκυριακο αλλιως εμπαινε τιμή 0.

Datetime	Daily_total_kWh	Daily_Sub_metering_1_kWh	Daily_Sub_metering_2_kWh	Daily_Sub_metering_3_kWh	Sub_sum_kWh	day_of_week	day_name	is_weekend	Nighttime_kWh	Peak_hour_power	Weekend_kWh
2006-12-16	20.152933	0.000	0.546	4.926	5.472	5	Saturday	1	NaN	4.222889	20.152933
2006-12-17	56.507667	2.033	4.187	13.341	19.561	6	Sunday	1	12.693833	3.697100	56.507667
2006-12-18	36.730433	1.063	2.621	14.018	17.702	0	Monday	0	2.503900	3.050567	0.000000
2006-12-19	27.769900	0.839	7.602	6.197	14.638	1	Tuesday	0	2.460200	3.879033	0.000000
2006-12-20	37.095800	0.000	2.648	14.063	16.711	2	Wednesday	0	2.364600	3.646067	0.000000

Τέλος, δημιουργήθηκαν εποχικά χαρακτηριστικά με την κατηγοριοποίηση των ημερών σε εποχές (χειμώνας, άνοιξη, καλοκαίρι, φθινόπωρο), επιτρέποντας την ανάλυση της κατανάλωσης σε συνάρτηση με εποχικές μεταβολές.

Datetime	Daily_total_kWh	Daily_Sub_metering_1_kWh	Daily_Sub_metering_2_kWh	Daily_Sub_metering_3_kWh	Sub_sum_kWh	day_of_week	day_name	is_weekend	Nighttime_kWh	Peak_hour_power	Weekend_kWh	season
2006-12-16	20.152933	0.000	0.546	4.926	5.472	5	Saturday	1	NaN	4.222889	20.152933	Winter
2006-12-17	56.507667	2.033	4.187	13.341	19.561	6	Sunday	1	12.693833	3.697100	56.507667	Winter
2006-12-18	36.730433	1.063	2.621	14.018	17.702	0	Monday	0	2.503900	3.050567	0.000000	Winter
2006-12-19	27.769900	0.839	7.602	6.197	14.638	1	Tuesday	0	2.460200	3.879033	0.000000	Winter
2006-12-20	37.095800	0.000	2.648	14.063	16.711	2	Wednesday	0	2.364600	3.646067	0.000000	Winter

5. Τελικό Σύνολο Χαρακτηριστικών

Το τελικό σύνολο δεδομένων περιλαμβάνει ημερήσιες μεταβλητές κατανάλωσης, χρονικά χαρακτηριστικά, δείκτες συμπεριφοράς και εποχικότητας. Το σύνολο αυτό αποθηκεύτηκε και χρησιμοποιήθηκε ως είσοδος για τα επόμενα στάδια μοντελοποίησης, συμπεριλαμβανομένης της ταξινόμησης, της παλινδρόμησης, της ομαδοποίησης και της πρόβλεψης χρονοσειρών.

	Daily_total_kWh	Daily_Sub_metering_1_kWh	Daily_Sub_metering_2_kWh	Daily_Sub_metering_3_kWh	Sub_sum_kWh	day_of_week	is_weekend	Nighttime_kWh	Peak_hour_power	Weekend_kWh
count	1442.000000	1442.000000	1442.000000	1442.000000	1442.000000	1442.000000	1442.000000	1441.000000	1442.000000	1442.000000
mean	26.056761	1.598806	1.853658	9.235660	12.688124	3.000000	0.285714	3.033082	2.847893	8.334052
std	10.058666	1.589069	2.088267	3.802919	5.501466	2.000694	0.451911	2.106748	1.066953	14.696637
min	4.106612	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.992000	0.171109	0.000000
25%	19.444358	0.579750	0.430000	6.646250	9.076250	1.000000	0.000000	2.051867	2.154808	0.000000
50%	25.708100	1.110500	0.684000	9.262500	12.242000	3.000000	0.000000	2.492267	2.803800	0.000000
75%	31.565983	2.201750	2.732000	11.740000	16.115500	5.000000	1.000000	2.999067	3.483675	15.289225
max	79.556433	11.224000	12.109000	26.835527	32.485112	6.000000	1.000000	17.580733	6.560533	79.556433

Μεθοδολογία Εφαρμογής πολλαπλών τεχνικών Εξόρυξης Δεδομένων

Η ανάλυση πραγματοποιήθηκε κυρίως σε ημερήσιο επίπεδο, προκειμένου να μειωθεί ο θόρυβος των λεπτομερών μετρήσεων ανά λεπτό και να διευκολυνθεί η εφαρμογή μοντέλων εξόρυξης δεδομένων. Οι τεχνικές που εφαρμόστηκαν περιλαμβάνουν ταξινόμηση, παλινδρόμηση, ομαδοποίηση, εξόρυξη κανόνων συσχέτισης και πρόβλεψη χρονοσειρών.

1. Ταξινόμηση Ημερών Κατανάλωσης

Η ταξινόμηση ημερών κατανάλωσης στοχεύει στον εντοπισμό ημερών με ασυνήθιστα υψηλή ενεργειακή χρήση, σε σύγκριση με τη συνήθη συμπεριφορά του νοικοκυριού. Το πρόβλημα διατυπώθηκε ως δυαδική ταξινόμηση σε ημερήσιο επίπεδο, ώστε να είναι δυνατή τόσο η πρόβλεψη όσο και η ερμηνεία των αποτελεσμάτων.
Στόχος είναι, με βάση χαρακτηριστικά που περιγράφουν τη συμπεριφορά μιας ημέρας, να προβλεφθεί αν αυτή ανήκει στην κατηγορία κανονικής ή υψηλής κατανάλωσης.

Βήμα 1: Δημιουργία ημερήσιου συνόλου δεδομένων

Αρχικά, οι μετρήσεις κατανάλωσης ανά λεπτό μετασχηματίστηκαν σε ημερήσια δεδομένα, υπολογίζοντας τη συνολική ημερήσια κατανάλωση ενέργειας σε κιλοβατώρες (kWh). Η επιλογή ημερήσιου επιπέδου ανάλυσης μειώνει τον θόρυβο των αρχικών μετρήσεων υψηλής συχνότητας και επιτρέπει τη σύγκριση ημερών με διαφορετικά μοτίβα χρήσης.

Στο στάδιο αυτό δημιουργήθηκαν επίσης χρονικά χαρακτηριστικά, όπως η ημέρα της εβδομάδας, ο μήνας, η κατανάλωση κατά τις νυχτερινές ώρες, η μέγιστη ισχύς σε ώρες αιχμής, καθώς και η ένδειξη αν μια ημέρα ανήκει σε Σαββατοκύριακο.Τα χαρακτηριστικά αυτά περιγράφουν τη συνολική συμπεριφορά της ημέρας και είναι διαθέσιμα τη στιγμή της πρόβλεψης.

	date	Daily_total_kWh	day_of_week	month	day_of_month	is_weekend	lag1_kWh	lag7_kWh	roll7_mean_kWh
0	2006-12-23	79.556433	5	12	23	1	39.022300	20.152933	35.127062
1	2006-12-24	42.500200	6	12	24	1	79.556433	56.507667	43.613276
2	2006-12-25	45.718667	0	12	25	0	42.500200	36.730433	41.612210
3	2006-12-26	65.568500	1	12	26	0	45.718667	27.769900	42.896243
4	2006-12-27	25.479333	2	12	27	0	65.568500	37.095800	48.296043

Βήμα 2: Επιλογή χαρακτηριστικών συμπεριφοράς και ημερολογίου

Για την εκπαίδευση του μοντέλου ταξινόμησης χρησιμοποιήθηκαν χαρακτηριστικά που αποτυπώνουν:

- συμπεριφορικά μοτίβα κατανάλωσης (π.χ. νυχτερινή κατανάλωση, ισχύς αιχμής),
- ημερολογιακή πληροφορία (ημέρα εβδομάδας, Σαββατοκύριακο, εποχή).

Σημαντικό είναι ότι η ίδια η ημερήσια κατανάλωση (Daily_total_kWh) δεν χρησιμοποιείται ως χαρακτηριστικό, καθώς από αυτήν ορίζεται η μεταβλητή στόχος. Με τον τρόπο αυτό αποφεύγεται το φαινόμενο της διαρροής πληροφορίας (data leakage).

	day_of_week	month	day_of_month	is_weekend	lag1_kWh	lag7_kWh	roll7_mean_kWh
0	5	12	23	1	39.022300	20.152933	35.127062
1	6	12	24	1	79.556433	56.507667	43.613276
2	0	12	25	0	42.500200	36.730433	41.612210
3	1	12	26	0	45.718667	27.769900	42.896243
4	2	12	27	0	65.568500	37.095800	48.296043

Βήμα 3: Ορισμός μεταβλητής στόχου

Η μεταβλητή στόχος ορίστηκε με βάση τον μέσο όρο της ημερήσιας κατανάλωσης ενέργειας στο σύνολο δεδομένων. Ημέρες με κατανάλωση μεγαλύτερη από τον μέσο όρο χαρακτηρίστηκαν ως ημέρες υψηλής κατανάλωσης (κλάση 1), ενώ οι υπόλοιπες ως ημέρες κανονικής κατανάλωσης (κλάση 0). Ο ορισμός αυτός παρέχει ένα απλό και ερμηνεύσιμο κριτήριο διαχωρισμού.

	date	Daily_total_kWh	y_true
1148	2010-02-13	39.192600	1
1149	2010-02-14	28.111633	0
1150	2010-02-15	34.952200	1
1151	2010-02-16	29.962600	0
1152	2010-02-17	34.242667	1
1153	2010-02-18	29.352600	0
1154	2010-02-19	32.559533	1
1155	2010-02-20	33.062133	1
1156	2010-02-21	45.671900	1
1157	2010-02-22	24.567267	0

Εδώ παρουσιαζεται ενα συνολο 10 ημερων και πώς κατανέμεται η μεταβλητή στόχος (y) στο πρόβλημα της ταξινόμησης ημερών κατανάλωσης ανά ημέρα.

	Class (y)	Meaning	Train samples	Test samples
0	0	Low / Normal consumption (≤ train mean)	861	254
1	1	High consumption (> train mean)	287	33

Ο συγκεκριμένος πίνακας παρουσιάζει πώς ορίστηκε και πώς κατανέμεται η μεταβλητή στόχος (y) στο πρόβλημα της ταξινόμησης ημερών κατανάλωσης. Η κλάση 0 αντιστοιχεί σε ημέρες χαμηλής ή κανονικής κατανάλωσης, ενώ η κλάση 1 αντιστοιχεί σε ημέρες υψηλής κατανάλωσης, όπως αυτές ορίστηκαν βάσει του μέσου όρου κατανάλωσης στο σύνολο εκπαίδευσης.

Βήμα 4: Χρονολογικός διαχωρισμός δεδομένων

Τα δεδομένα χωρίστηκαν σε σύνολο εκπαίδευσης (80%) και σύνολο ελέγχου (20%) με τυχαίο αλλά στρωματοποιημένο διαχωρισμό (stratified split), ώστε να διατηρηθεί παρόμοια αναλογία ημερών υψηλής και κανονικής κατανάλωσης και στα δύο σύνολα.

Η ύπαρξη ανεξάρτητου συνόλου ελέγχου επιτρέπει την αντικειμενική αξιολόγηση της ικανότητας γενίκευσης του μοντέλου.

```
((1148, 9), (287, 9))
```

Βήμα 5: Εκπαίδευση και σύγκριση μοντέλων ταξινόμησης

Για την ταξινόμηση εκπαιδεύτηκαν τρία διαφορετικά μοντέλα:

- Logistic Regression,
- Random Forest και
- Gradient Boosting.

Τα μοντέλα ενσωματώθηκαν σε pipelines που περιλαμβάνουν συμπλήρωση ελλιπών τιμών, κανονικοποίηση αριθμητικών χαρακτηριστικών και κωδικοποίηση κατηγορικών μεταβλητών.Η σύγκριση γίνεται σε δύο επίπεδα:

- *Cross-validation (cv_*)* στο training set
- *Τελική αξιολόγηση (test_*)* στο test set

Στόχος: να επιλεγεί το καλύτερο μοντέλο γενίκευσης, όχι απλώς αυτό με την υψηλότερη accuracy. Η αξιολόγηση πραγματοποιήθηκε με:

- ➢ *cv_acc_mean:* Μέση ακρίβεια (Accuracy) στα folds του TimeSeries Cross-Validation.
- ➢ *cv_f1_mean:* Μέσο F1-score στο cross-validation.
- ➢ *cv_auc_mean:* Μέση τιμή ROC-AUC στο cross-validation
- ➢ *test_acc / test_f1 / test_auc:* Αντίστοιχες μετρικές, στο ανεξάρτητο test set, που δείχνουν: πόσο καλά γενικεύει το μοντέλο σε *απρόβλεπτα* δεδομένα.

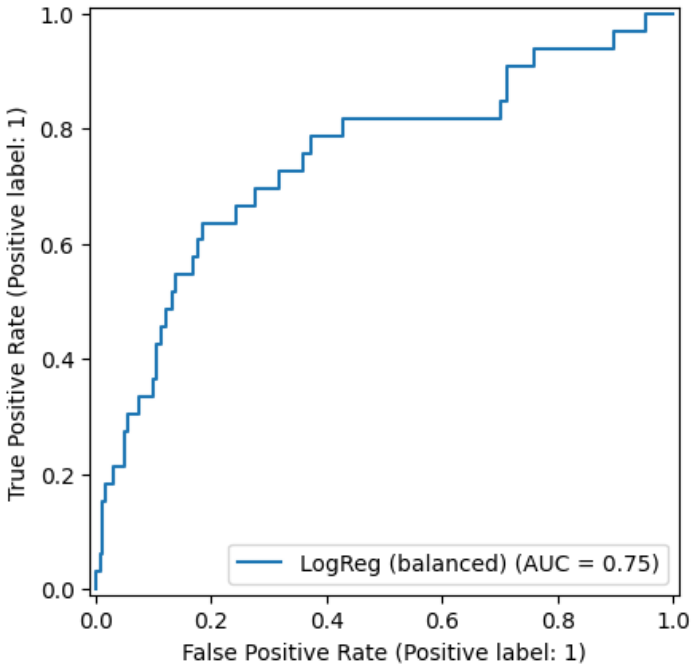
	model	cv_acc_mean	cv_f1_mean	cv_auc_mean	test_acc	test_f1	test_auc
1	RandomForest	0.792670	0.507310	0.832633	0.881533	0.413793	0.818599
2	GradientBoosting	0.785340	0.515080	0.823847	0.871080	0.327273	0.803985
0	LogReg (balanced)	0.747644	0.548206	0.809406	0.790941	0.400000	0.745168

Το μοντέλο Random Forest παρουσιάζει τη βέλτιστη συνολική επίδοση, με την υψηλότερη τιμή ROC-AUC τόσο στο cross-validation όσο και στο test set, καθιστώντας το καταλληλότερο για τον εντοπισμό ημερών υψηλής κατανάλωσης. Η επιλογή του μοντέλου βασίστηκε κυρίως στη μετρική ROC-AUC, η οποία είναι ιδιαίτερα κατάλληλη για προβλήματα δυαδικής ταξινόμησης με ανισορροπημένες κλάσεις.

Βήμα 6: Τελική αξιολόγηση στο σύνολο ελέγχου

Το βέλτιστο μοντέλο επιλέχθηκε με βάση τη μετρική ROC-AUC στο σύνολο ελέγχου. Για το μοντέλο αυτό παρουσιάζονται αναλυτικά η αναφορά ταξινόμησης, ο πίνακας σύγχυσης και η καμπύλη ROC, τα οποία επιτρέπουν την εις βάθος αξιολόγηση της απόδοσής του.

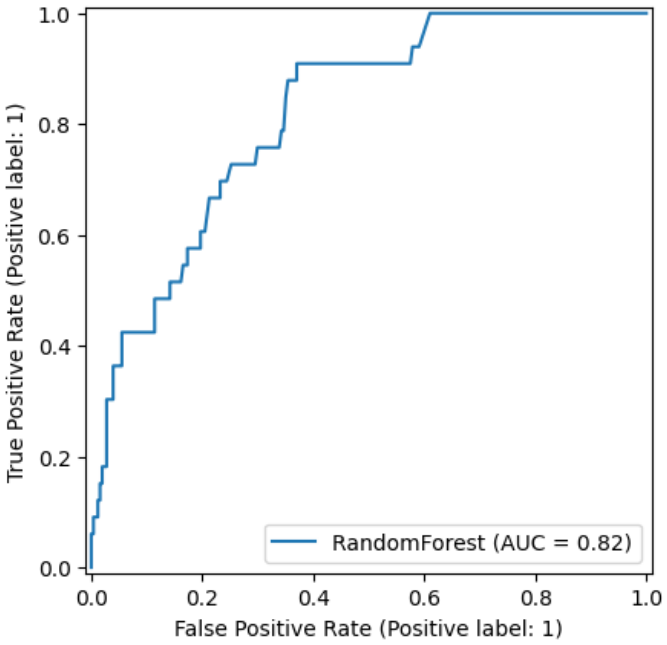
```
BEST MODEL: RandomForest

Classification report (TEST):
              precision    recall  f1-score   support

           0     0.9198    0.9488    0.9341       254
           1     0.4800    0.3636    0.4138        33

    accuracy                         0.8815       287
   macro avg     0.6999    0.6562    0.6740       287
weighted avg     0.8693    0.8815    0.8743       287

Confusion matrix (rows=true, cols=pred):
 [[241  13]
 [ 21  12]]
```



*ΣΧΟΛΙΑ:*Το γράφημα απεικονίζει την ROC καμπύλη (Receiver Operating Characteristic) για το μοντέλο Logistic Regression στο σύνολο ελέγχου. Η ROC καμπύλη δείχνει:

- την ικανότητα του μοντέλου να διακρίνει μεταξύ:
- κλάσης 0: χαμηλή/κανονική κατανάλωση
- κλάσης 1: υψηλή κατανάλωση

για όλα τα πιθανά κατώφλια απόφασης. *AUC = 0.75* → καλή διακριτική ικανότητα.

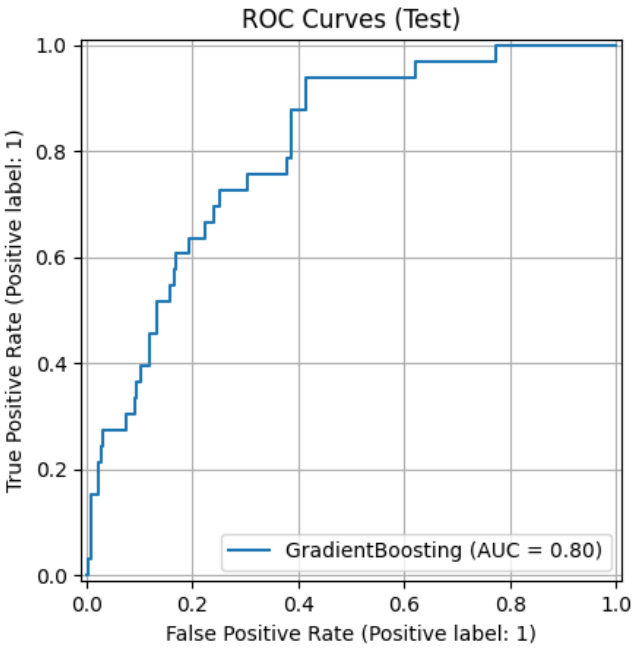


<u>ΣΧΟΛΙΑ:</u> Το γράφημα απεικονίζει την ROC καμπύλη (Receiver Operating Characteristic) για το μοντέλο Random Forest στο σύνολο ελέγχου.

Η ROC καμπύλη δείχνει την ικανότητα του μοντέλου να διαχωρίζει:

- ημέρες χαμηλής/κανονικής κατανάλωσης (κλάση 0)
- από ημέρες υψηλής κατανάλωσης (κλάση 1),

για όλα τα πιθανά κατώφλια απόφασης. AUC ≈ 0.82 υποδηλώνει πολύ καλή ικανότητα διαχωρισμού μεταξύ ημερών χαμηλής και υψηλής κατανάλωσης. Σε σύγκριση με τα υπόλοιπα μοντέλα που εξετάστηκαν, το Random Forest παρουσιάζει ανώτερη διακριτική ικανότητα, γεγονός που δικαιολογεί την επιλογή του ως τελικό μοντέλο ταξινόμησης.



ΣΧΟΛΙΑ: Το γράφημα απεικονίζει την ROC καμπύλη για το μοντέλο Gradient Boosting στο σύνολο ελέγχου. Δείχνει την ικανότητα του μοντέλου να διαχωρίζει ημέρες χαμηλής από ημέρες υψηλής κατανάλωσης, για όλα τα πιθανά κατώφλια απόφασης. AUC ≈ 0.80 σημαίνει ότι: Το μοντέλο έχει περίπου 80% πιθανότητα να δώσει υψηλότερο score σε μια ημέρα υψηλής κατανάλωσης απ' ό,τι σε μια ημέρα χαμηλής,

ΣΧΟΛΙΑ: Η σύγκριση των ROC καμπυλών δείχνει ότι όλα τα μοντέλα παρουσιάζουν ουσιαστική διακριτική ικανότητα, με το Random Forest να υπερέχει ελαφρώς στο σύνολο ελέγχου. Τα αποτελέσματα αυτά, σε συνδυασμό με τις μετρικές F1 και την ανάλυση του πίνακα σύγχυσης, οδήγησαν στην επιλογή του Random Forest ως τελικό μοντέλο ταξινόμησης.

2. Παλινδρόμηση για Πρόβλεψη Κατανάλωσης - REGRESSION

Η παλινδρόμηση χρησιμοποιήθηκε για την ποσοτική πρόβλεψη της συνολικής ημερήσιας κατανάλωσης ηλεκτρικής ενέργειας, με στόχο την εκτίμηση της κατανάλωσης της επόμενης ημέρας βάσει ιστορικών και χρονικών χαρακτηριστικών.

Βήμα 1: Ορισμός προβλήματος πρόβλεψης

Το πρόβλημα ορίστηκε ως πρόβλεψη της ημερήσιας κατανάλωσης της ημέρας t+1, χρησιμοποιώντας πληροφορία που είναι διαθέσιμη έως και την ημέρα t.

Βήμα 2: Φόρτωση τελικού feature dataset

Χρησιμοποιείται το ημερήσιο σύνολο δεδομένων που δημιουργήθηκε στο preprocessing στάδιο και αποθηκεύτηκε ως power_daily_features.parquet. Αυτό το dataset περιλαμβάνει ήδη "συμπεριφορικά" και χρονικά χαρακτηριστικά (π.χ. νυχτερινή κατανάλωση, peak hour, εποχικότητα).

Datetime	Daily_total_kWh	Daily_Sub_metering_1_kWh	Daily_Sub_metering_2_kWh	Daily_Sub_metering_3_kWh	Sub_sum_kWh	day_of_week	day_name	is_weekend	Nighttime_kWh	Peak_hour_power	Weekend_kWh	season
2006-12-16	20.152933	0.000	0.546	4.926	5.472	5	Saturday	1	NaN	4.222889	20.152933	Winter
2006-12-17	56.507667	2.033	4.187	13.341	19.561	6	Sunday	1	12.693833	3.697100	56.507667	Winter
2006-12-18	36.730433	1.063	2.621	14.018	17.702	0	Monday	0	2.503900	3.050567	0.000000	Winter
2006-12-19	27.769900	0.839	7.602	6.197	14.638	1	Tuesday	0	2.460200	3.879033	0.000000	Winter
2006-12-20	37.095800	0.000	2.648	14.063	16.711	2	Wednesday	0	2.364600	3.646067	0.000000	Winter

Βήμα 3: Δημιουργία μεταβλητής στόχου (target) με χρονική μετατόπιση

Για να προβλέψουμε την κατανάλωση της επόμενης ημέρας, δημιουργούμε στόχο με shift(-1):

- Η γραμμή της ημέρας t κρατά τα features της ημέρας t
- Η τιμή-στόχος είναι η κατανάλωση της ημέρας t+1

Έτσι διασφαλίζεται ότι το μοντέλο χρησιμοποιεί μόνο πληροφορία διαθέσιμη “μέχρι σήμερα” για να προβλέψει “αύριο”, δηλαδή δεν υπάρχει data leakage.

Βήμα 4: Επιλογή και ρόλος των χαρακτηριστικών εισόδου

Για την παλινδρόμηση χρησιμοποιήθηκαν χαρακτηριστικά που αποτυπώνουν τόσο τη συμπεριφορά κατανάλωσης όσο και τη χρονική δομή των δεδομένων. Συγκεκριμένα, συμπεριλήφθηκαν μεταβλητές όπως:

- ➢ *Nighttime_kWh:* ενέργεια που καταναλώθηκε στο διάστημα 00:00–06:00 (δείκτης νυχτερινής συμπεριφοράς)
- ➢ *Peak_hour_power:* μέγιστη μέση ωριαία ισχύς ανά ημέρα (δείκτης “αιχμής”)
- ➢ *day_of_week* και *is_weekend:* εβδομαδιακά μοτίβα (εργάσιμες vs ΣΚ)
- ➢ *season*: εποχικότητα (χειμώνας/άνοιξη/καλοκαίρι/φθινόπωρο)

Βήμα 5: Χρονολογικός διαχωρισμός δεδομένων Train/Test

Τα δεδομένα χωρίστηκαν χρονολογικά σε σύνολα εκπαίδευσης και ελέγχου, με αναλογία 80% train, 20% test. Ο διαχωρισμός αυτός είναι απαραίτητος σε προβλήματα χρονοσειρών, καθώς διασφαλίζει ότι το μοντέλο εκπαιδεύεται αποκλειστικά σε παρελθοντικές παρατηρήσεις και αξιολογείται σε μελλοντικές.

Βήμα 6: Preprocessing pipeline

Για να εφαρμοστεί σωστή προεπεξεργασία, χρησιμοποιήθηκε ColumnTransformer:

- *Numeric features:*
  - SimpleImputer(strategy="median") για συμπλήρωση ελλείψεων (robust σε outliers)
  - StandardScaler() για κλιμάκωση, ώστε μοντέλα όπως η γραμμική παλινδρόμηση να λειτουργούν σταθερά.
- *Categorical feature (season):*
  - OneHotEncoder(handle_unknown="ignore"), ώστε να μετατραπεί σε δυαδικά χαρακτηριστικά.

Το preprocessing ενσωματώθηκε σε Pipeline, ώστε να εφαρμόζεται με συνέπεια σε train/test.

Βήμα 7: Εκπαίδευση μοντέλων παλινδρόμησης

Εκπαιδεύτηκαν δύο μοντέλα:

- Linear Regression (Baseline): Χρησιμοποιήθηκε ως σημείο αναφοράς. Είναι απλό, ερμηνεύσιμο και δείχνει αν τα επιλεγμένα features αρκούν για μια περίπου γραμμική σχέση με το target.
- Random Forest Regressor: Μη γραμμικό μοντέλο ensemble, ικανό να συλλάβει αλληλεπιδράσεις και μη γραμμικότητες.

Βήμα 8: Μετρικές αξιολόγησης παλινδρόμησης

Η απόδοση των μοντέλων αξιολογήθηκε με χρήση τριών βασικών μετρικών:

- *MAE (Mean Absolute Error):* εκφράζει το μέσο απόλυτο σφάλμα πρόβλεψης σε kWh.
- *RMSE (Root Mean Squared Error):* τιμωρεί περισσότερο τα μεγάλα σφάλματα και αποκαλύπτει την παρουσία ακραίων αποκλίσεων.
- *R^2 (Συντελεστής Προσδιορισμού):* δείχνει το ποσοστό της διακύμανσης της κατανάλωσης που εξηγείται από το μοντέλο.

```
Linear Regression: 4.888039972213437 6.390366682976692 0.2799714744793147
Random Forest: 4.99572120838776 6.593086417680795 0.23356437906398386
```

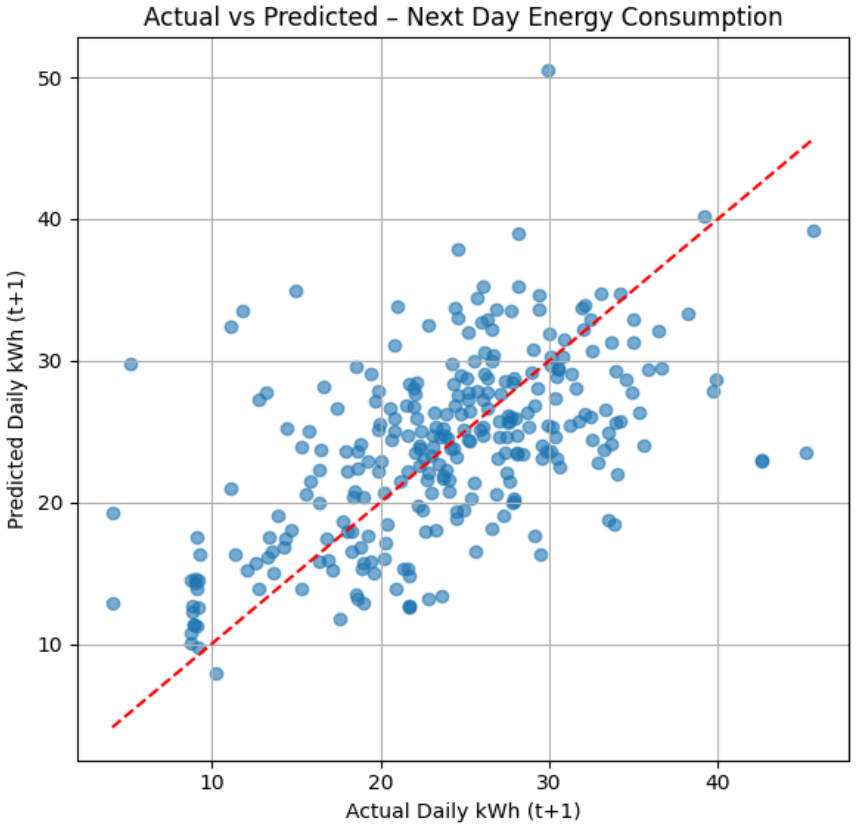
	Model	MAE	RMSE	R2
0	Linear Regression	4.888040	6.390367	0.279971
1	Random Forest	4.995721	6.593086	0.233564

ΣΧΟΛΙΑ: Ο Πίνακας παρουσιάζει τα αποτελέσματα της αξιολόγησης των μοντέλων παλινδρόμησης στο σύνολο ελέγχου. Παρατηρείται ότι το γραμμικό μοντέλο παρουσιάζει χαμηλότερο μέσο απόλυτο σφάλμα και μικρότερη τιμή RMSE, καθώς και υψηλότερο συντελεστή προσδιορισμού σε σύγκριση με το μοντέλο Random Forest. Το αποτέλεσμα αυτό υποδηλώνει ότι, για το συγκεκριμένο σύνολο χαρακτηριστικών, η σχέση μεταξύ των μεταβλητών και της ημερήσιας κατανάλωσης μπορεί να περιγραφεί ικανοποιητικά με γραμμικό τρόπο. Παράλληλα, αναδεικνύεται ότι η αύξηση της πολυπλοκότητας του μοντέλουδεν οδηγείΜαπαραίτητα σε καλύτερη γενίκευση.

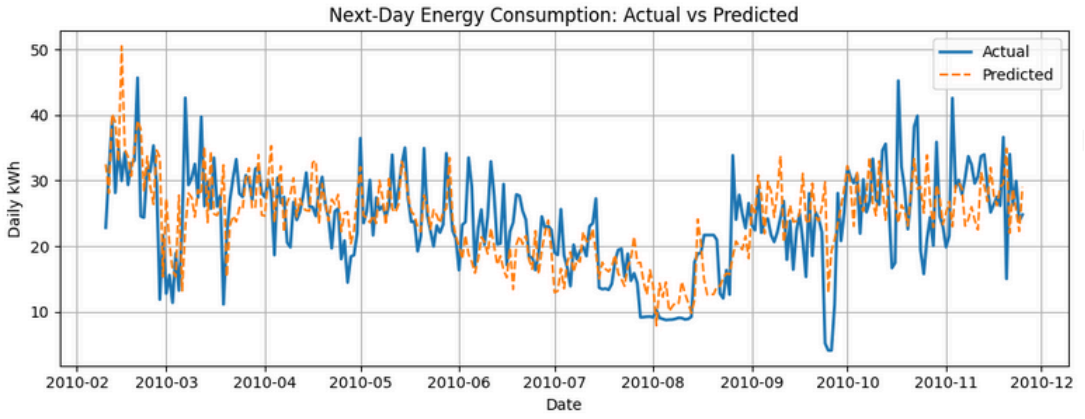
Βήμα 9: Οπτικοποίηση προβλέψεων και ανάλυση σφαλμάτων

Πέρα από τις αριθμητικές μετρικές, η απόδοση των μοντέλων αξιολογήθηκε και οπτικά. Παρουσιάστηκαν γραφήματα που συγκρίνουν τις πραγματικές και προβλεπόμενες τιμές κατανάλωσης στο σύνολο ελέγχου, επιτρέποντας την άμεση εκτίμηση της ποιότητας της πρόβλεψης στο χρόνο.

Επιπλέον, εξετάστηκε η κατανομή των υπολοίπων σφαλμάτων (residuals), προκειμένου να εντοπιστούν συστηματικές αποκλίσεις ή ακραία σφάλματα

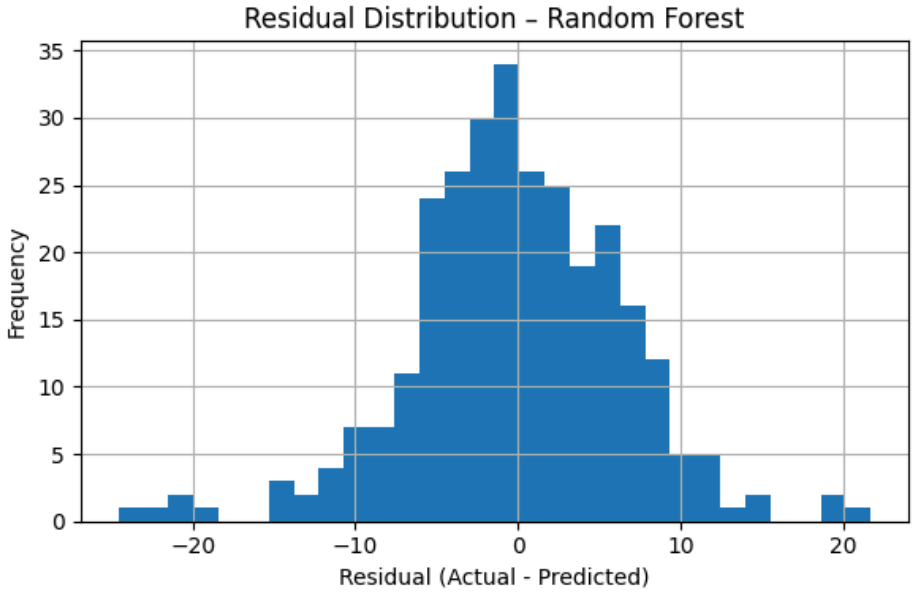


ΣΧΟΛΙΑ: Το Σχήμα απεικονίζει τη σχέση μεταξύ των πραγματικών και των προβλεπόμενων τιμών της ημερήσιας κατανάλωσης της επόμενης ημέρας στο σύνολο ελέγχου. Η κόκκινη διακεκομμένη γραμμή αντιστοιχεί στην ιδανική περίπτωση όπου η πρόβλεψη ταυτίζεται με την πραγματική τιμή. Παρατηρείται ισχυρή θετική συσχέτιση μεταξύ πραγματικών και προβλεπόμενων τιμών, γεγονός που υποδηλώνει ότι το μοντέλο έχει μάθει τη γενική τάση της κατανάλωσης. Ωστόσο, η σημαντική διασπορά γύρω από τη γραμμή y=x δείχνει ότι η ακρίβεια των προβλέψεων μειώνεται σε ημέρες με ακραία κατανάλωση. Ειδικότερα, το μοντέλο τείνει να υποεκτιμά την κατανάλωση σε ημέρες υψηλής ενεργειακής χρήσης και να υπερεκτιμά σε ημέρες πολύ χαμηλής κατανάλωσης.



ΣΧΟΛΙΑ: Το Σχήμα απεικονίζει τη σύγκριση της πραγματικής και της προβλεπόμενης ημερήσιας κατανάλωσης ηλεκτρικής ενέργειας της επόμενης ημέρας στο σύνολο ελέγχου. Παρατηρείται ότι η προβλεπόμενη χρονοσειρά ακολουθεί σε μεγάλο βαθμό τη γενική τάση της πραγματικής κατανάλωσης, αποτυπώνοντας σωστά τις εποχικές μεταβολές. Ωστόσο, η προβλεπόμενη καμπύλη

εμφανίζεται πιο ομαλή, γεγονός που υποδηλώνει ότι το μοντέλο τείνει να εξομαλύνει απότομες αιχμές και πτώσεις της κατανάλωσης.



ΣΧΟΛΙΑ: Το Σχήμα παρουσιάζει την κατανομή των υπολοίπων σφαλμάτων (residuals) του μοντέλου Random Forest στο σύνολο ελέγχου. Παρατηρείται ότι η πλειονότητα των residuals συγκεντρώνεται γύρω από το μηδέν, γεγονός που υποδηλώνει ότι οι περισσότερες προβλέψεις είναι κοντά στις πραγματικές τιμές κατανάλωσης.

Ομαδοποίηση (Clustering)

Για την ανάλυση των δεδομένων κατανάλωσης ενέργειας, εφαρμόσαμε την τεχνική ομαδοποίησης (clustering) χρησιμοποιώντας τον αλγόριθμο KMeans, με στόχο την αναγνώριση τυπικών προφίλ κατανάλωσης. Η διαδικασία περιλαμβάνει την επεξεργασία των δεδομένων, την επιλογή των χαρακτηριστικών, και τη χρήση κατάλληλων μεθόδων αξιολόγησης για τον προσδιορισμό του ιδανικού αριθμού συστάδων.

Βήμα 1:

- Επιλέξαμε τα χαρακτηριστικά που θα χρησιμοποιηθούν στην ομαδοποίηση (clustering), τα οποία είναι:
  - Daily_total_kWh**:** Η συνολική κατανάλωση ενέργειας ανά ημέρα.
  - Nighttime_kWh**:** Η κατανάλωση ενέργειας κατά τη διάρκεια της νύχτας.
  - Peak_hour_power**:** Η κατανάλωση ενέργειας κατά τις ώρες αιχμής.
- Χρησιμοποιήθηκε η μέθοδος του **SimpleImputer** για να αντικαταστήσουμε τις ελλιπείς τιμές με τον μέσο όρο κάθε χαρακτηριστικού.

- Εφαρμόστηκε η κανονικοποίηση των δεδομένων με τον **StandardScaler**, για να διασφαλίσουμε ότι όλα τα χαρακτηριστικά έχουν την ίδια κλίμακα και για να αποφύγουμε την

παραμόρφωση των αποτελεσμάτων του αλγορίθμου KMeans, καθώς ο αλγόριθμος βασίζεται στην απόσταση μεταξύ των σημείων δεδομένων.

Βήμα 2:

Επιλογή Τιμών για τον Αριθμό των Clusters (k)

- Ορίσαμε τις τιμές του k (αριθμός συστάδων) από 2 έως 10, για να δοκιμάσουμε διάφορους αριθμούς συστάδων και να επιλέξουμε τον καλύτερο.
- Για κάθε τιμή του k, εκτελούμε τον αλγόριθμο KMeans και υπολογίζουμε διάφορες μετρικές για να αξιολογήσουμε την ποιότητα της ομαδοποίησης.

Υπολογισμός Μετρικών Αξιολόγησης

Για κάθε τιμή του k, υπολογίζουμε τρεις βασικές μετρικές:

- *Silhouette Score:* Μετράει τη συνοχή των δεδομένων εντός της ίδιας συστάδας και την απόσταση από άλλες συστάδες. Υψηλότερες τιμές δείχνουν καλύτερη ποιότητα ομαδοποίησης.

- *Davies-Bouldin Index (DBI):* Αντιπροσωπεύει την απόσταση μεταξύ των συστάδων. Χαμηλότερες τιμές δείχνουν καλύτερη διαφοροποίηση των clusters.

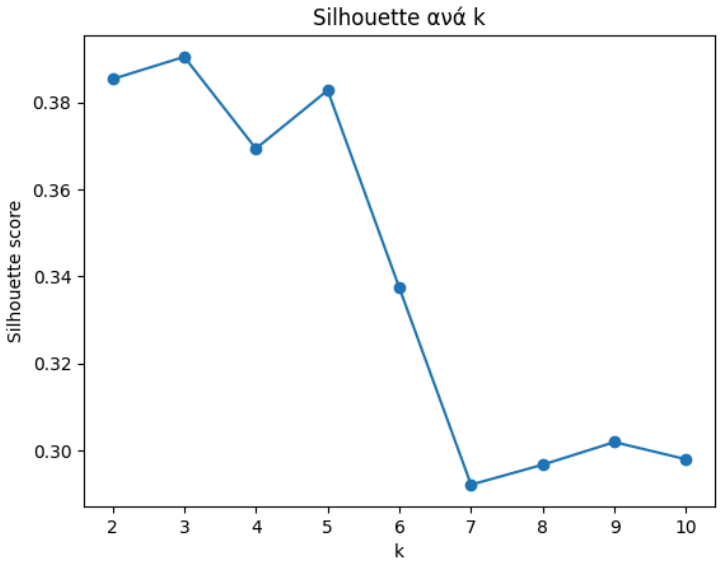
- *Inertia (WCSS):* Η συνολική απόκλιση εντός των clusters. Χαμηλότερες τιμές υποδεικνύουν ότι τα δεδομένα είναι πιο συγκεντρωμένα γύρω από τα κέντρα των clusters.

k	silhouette	davies_bouldin	inertia
2	0.3853	1.0549	2684.6572
3	0.3904	0.8504	1738.6010
4	0.3693	0.8502	1285.3780
5	0.3827	0.8139	1067.9840
6	0.3374	0.8692	882.3524
7	0.2922	0.9763	789.1519
8	0.2968	0.9804	724.3940
9	0.3020	0.9635	666.5135
10	0.2980	0.9556	615.9878

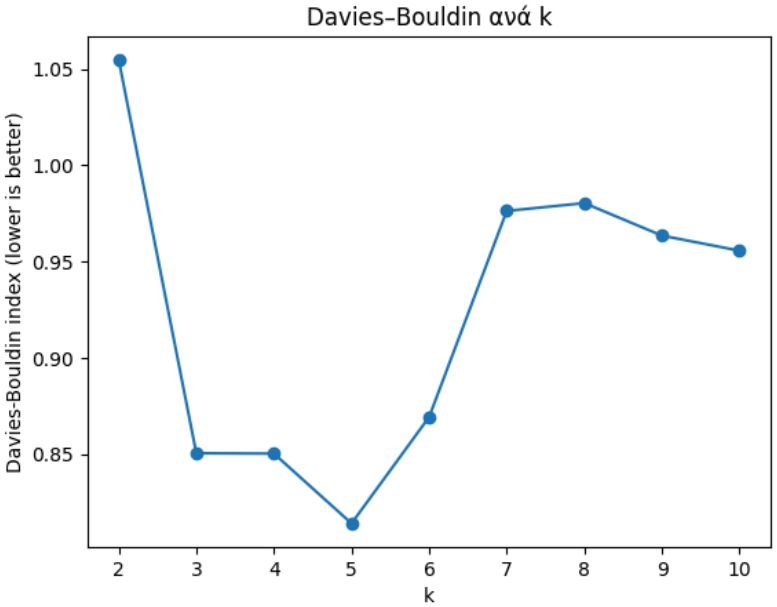
Οπτικοποίηση Μετρικών Αξιολόγησης

Δημιουργούμε τρεις γραφικές παραστάσεις για να απεικονίσουμε τις τιμές των μετρικών για κάθε k:

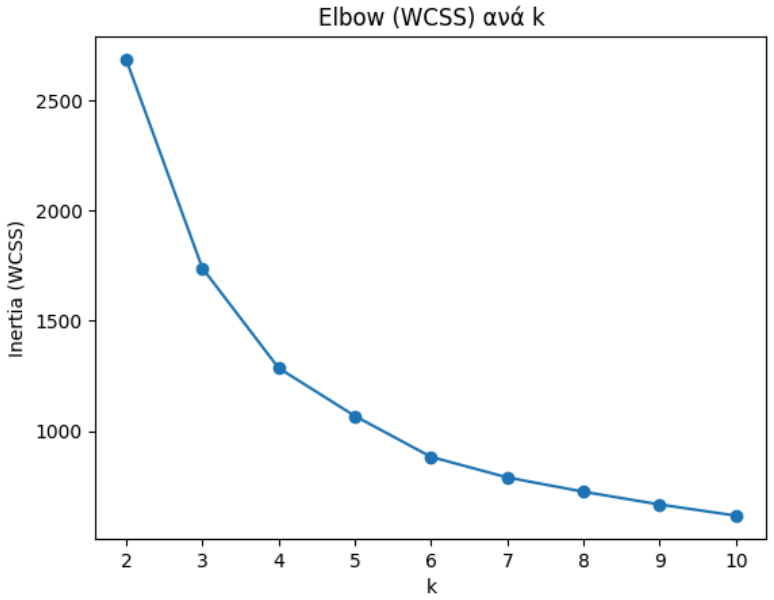
- *Silhouette Score:* Η καλύτερη τιμή για τον αριθμό των clusters φαίνεται να είναι k = 2 ή 3, καθώς εκεί έχουμε το υψηλότερο Silhouette Score.



➢ *Davies-Bouldin Index:* Η καλύτερη τιμή του k για τη διάκριση των συστάδων φαίνεται να είναι k = 3 ή 4, καθώς οι τιμές του DBI είναι οι χαμηλότερες και οι συστάδες είναι πιο διακριτές.



➢ *Inertia (WCSS):* Η καλύτερη τιμή του k φαίνεται να είναι k = 3 ή 4, καθώς εκεί παρατηρούμε τη μεγαλύτερη μείωση στο WCSS και μετά η βελτίωση επιβραδύνεται, πράγμα που υποδηλώνει ότι έχουμε βρει το πιο κατάλληλο σημείο για τον αριθμό των συστάδων.



Αυτόματη Επιλογή του Βέλτιστου k

Με βάση το Silhouette Score: Επιλέγουμε την τιμή του k που έχει τη μέγιστη τιμή του Silhouette Score, καθώς αυτό υποδεικνύει την καλύτερη συνοχή των clusters.

```
Best k by Silhouette: 3
Best k by trade-off (Silhouette+DBI ranks): 3
```

Με βάση τη συνολική ισορροπία Silhouette και Davies-Bouldin Index: Επιλέγουμε το k που έχει την καλύτερη ισορροπία μεταξύ υψηλού Silhouette Score και χαμηλού DBI. Για αυτό, χρησιμοποιούμε μια κατάταξη των τιμών και επιλέγουμε το k με το καλύτερο συνολικό σκορ.

k	silhouette	davies_bouldin	inertia	rank_sum
3	0.3904	0.8504	1738.6010	4.0
5	0.3827	0.8139	1067.9840	4.0
4	0.3693	0.8502	1285.3780	6.0
6	0.3374	0.8692	882.3524	9.0
2	0.3853	1.0549	2684.6572	11.0
9	0.3020	0.9635	666.5135	12.0
10	0.2980	0.9556	615.9878	12.0
7	0.2922	0.9763	789.1519	16.0
8	0.2968	0.9804	724.3940	16.0

Η καλύτερη τιμή για τον αριθμό των clusters είναι k = 3, καθώς έχει το υψηλότερο Silhouette Score(0.3904) και το χαμηλότερο Davies-Bouldin Index (0.8504).

Βήμα 3:

- *Ορισμός του Αριθμού των Clusters*
  - Ορίζουμε τον αριθμό των clusters (n_clusters) χρησιμοποιώντας την τιμή του best_k_tradeoff, η οποία επιλέχθηκε μέσω των προηγούμενων μετρικών αξιολόγησης (Silhouette Score και Davies-Bouldin Index).

  - Χρησιμοποιούμε τον αλγόριθμο KMeans με:

    - n_clusters = 3 (την τιμή του k που επιλέχθηκε).
    - init='k-means++' για καλύτερη αρχικοποίηση.
    - random_state=42 για αναπαραγωγιμότητα.
    - n_init=10 για εκτέλεση του αλγορίθμου 10 φορές με διαφορετικές αρχικές θέσεις, για να επιτευχθεί το βέλτιστο αποτέλεσμα.

- *Έλεγχος Σταθερότητας του KMeans με Διάφορους Σπόρους (Sanity Check)*
  - Για να ελέγξουμε τη σταθερότητα του αλγορίθμου, τρέχουμε τον KMeans με τρεις διαφορετικούς σπόρους (seeds = [0, 42, 123]).

  - Για κάθε σπόρο, εκτελούμε το KMeans και υπολογίζουμε δύο μετρικές αξιολόγησης:

    - Silhouette Score: Για να ελέγξουμε την ποιότητα της ομαδοποίησης (πώς οι σημεία της ίδιας συστάδας είναι κοντά μεταξύ τους σε σχέση με άλλες συστάδες).

    - Davies-Bouldin Index (DBI): Για να εξετάσουμε την απόσταση μεταξύ των clusters (χαμηλότερη τιμή δείχνει καλύτερη διαφοροποίηση).

  - Τα αποτελέσματα καταγράφονται σε έναν πίνακα για να ελέγξουμε αν τα αποτελέσματα είναι συνεπή με διαφορετικούς σπόρους.

Sanity check σταθερότητας KMeans:

	random_state	silhouette	davies_bouldin
0	0	0.3904	0.8500
1	42	0.3904	0.8504
2	123	0.3904	0.8506

- *Προσθήκη των Clusters στο Αρχικό DataFrame*

- Προσθέτουμε τις ετικέτες των clusters στο αρχικό DataFrame (daily), δημιουργώντας μια νέα στήλη 'Cluster'.

- Αυτή η στήλη περιέχει τις αντίστοιχες ετικέτες των clusters για κάθε ημέρα κατανάλωσης ενέργειας.

- *Υπολογισμός Μετρικών Αξιολόγησης για την Τελική Ομαδοποίηση*
  - Υπολογίζουμε ξανά τις μετρικές Silhouette Score και Davies-Bouldin Index για τα τελικά clusters για να επιβεβαιώσουμε την ποιότητα της ομαδοποίησης:
    - Silhouette Score: Μέτρηση της συνοχής των clusters.
    - Davies-Bouldin Index: Μέτρηση της απόστασης μεταξύ των clusters (χαμηλότερη τιμή είναι καλύτερη).
  - Εκτυπώνουμε τα αποτελέσματα με σκοπό την αξιολόγηση της ποιότητας των clusters.

```
Silhouette Score: 0.390
Davies-Bouldin Index: 0.850
```

- *Ανάλυση των Μέσων Όρων Ανά Σύσταση*
  - Υπολογίζουμε τους μέσους όρους των χαρακτηριστικών για κάθε cluster, ώστε να κατανοήσουμε καλύτερα τις διαφορές μεταξύ των clusters.
  - Χρησιμοποιούμε την groupby συνάρτηση για να υπολογίσουμε τους μέσους όρους για κάθε σύσταση (cluster) στα χαρακτηριστικά που επιλέξαμε (Daily_total_kWh, Nighttime_kWh, Peak_hour_power).

```
Μέσοι όροι ανά συστάδα:
         Daily_total_kWh  Nighttime_kWh  Peak_hour_power
Cluster
0              18.448712       2.474244         2.035547
1              31.950591       2.742668         3.573820
2              41.442798      10.367448         3.661537
```

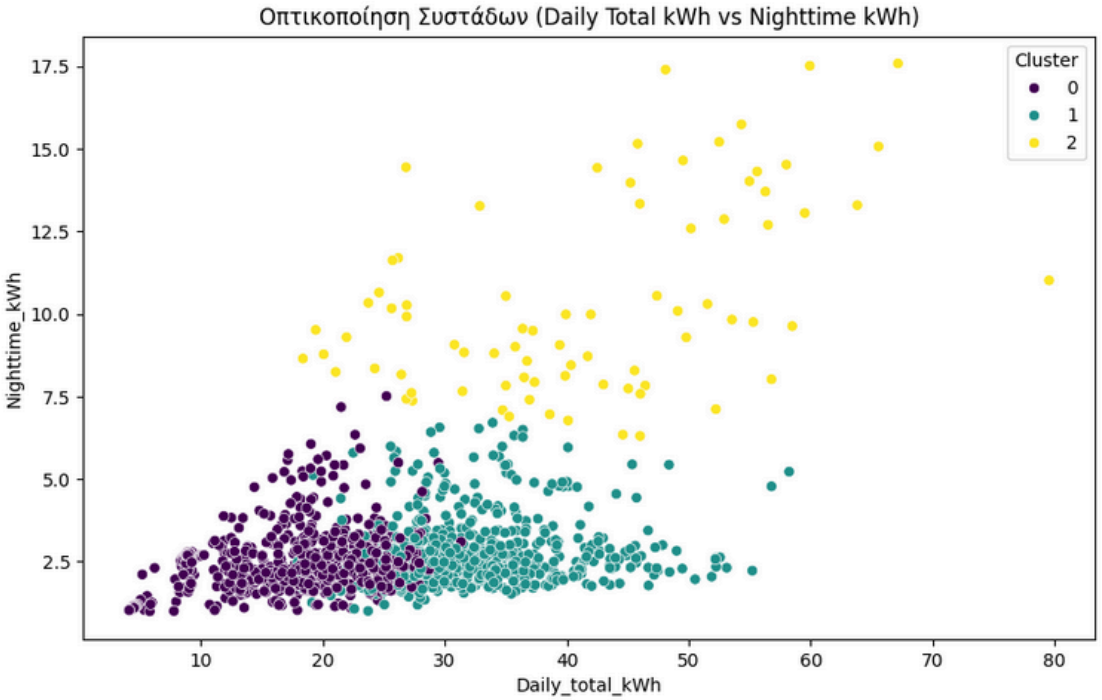
- *Ανάλυση Κατανάλωσης Σύμφωνα με το Σαββατοκύριακο (Weekend Analysis)*
  - Χρησιμοποιούμε την crosstab για να αναλύσουμε τη σχέση μεταξύ των clusters και του χαρακτηριστικού is_weekend (αν είναι σαββατοκύριακο ή όχι).
  - Υπολογίζουμε τη συχνότητα των clusters ανάλογα με το αν η ημέρα είναι σαββατοκύριακο ή καθημερινή.

is_weekend	0	1
Cluster		
0	0.792701	0.207299
1	0.671091	0.328909
2	0.405063	0.594937

Βήμα 4:

- *Οπτικοποίηση και Ερμηνεία των Συστάδων*

Δημιουργήθηκε διάγραμμα διασποράς (scatter plot) με άξονες τη συνολική ημερήσια κατανάλωση ενέργειας (Daily_total_kWh) και τη νυχτερινή κατανάλωση (Nighttime_kWh), με χρωματική διάκριση ανά cluster.



Σύσταση 0 (μωβ χρώμα):

- ➢ Οι περισσότεροι από τους μωβ σημείους (Cluster 0) βρίσκονται σε περιοχές με χαμηλή κατανάλωση τόσο την ημέρα όσο και τη νύχτα (χαμηλότερες τιμές και στους δύο άξονες).
- ➢ Αυτό δείχνει ότι το Cluster 0 αντιπροσωπεύει ημέρες με χαμηλή κατανάλωση, πιθανώς καθημερινές ημέρες με χαμηλότερη χρήση ενέργειας.

Σύσταση 1 (γαλαζοπράσινο χρώμα):

- ➢ Οι γαλαζοπράσινες κουκκίδες (Cluster 1) βρίσκονται στην περιοχή του γραφήματος με μεσαία κατανάλωση ενέργειας.
- ➢ Τα σημεία αυτά έχουν σχετικά υψηλότερη νυχτερινή κατανάλωση, ενώ η συνολική κατανάλωση είναι επίσης αυξημένη σε σχέση με το Cluster 0.

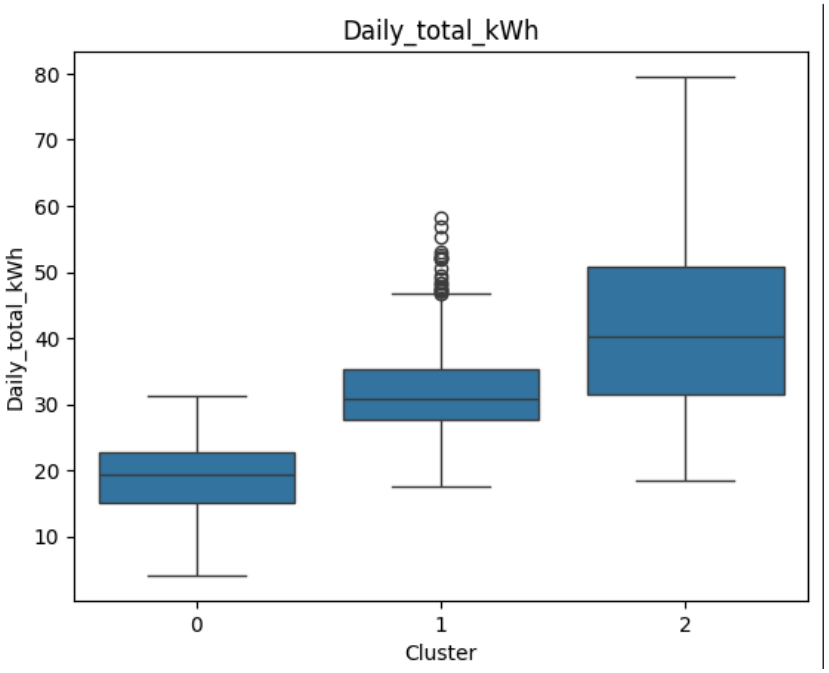
➢ Αυτό υποδεικνύει ημέρες με κανονική χρήση, ενδεχομένως κάποια καθημερινά ή εργάσιμα Σαββατοκύριακα.

Σύσταση 2 (κίτρινο χρώμα):

➢ Οι κίτρινες κουκκίδες (Cluster 2) βρίσκονται κυρίως στην περιοχή του γραφήματος με υψηλή κατανάλωση ενέργειας τόσο την ημέρα όσο και τη νύχτα.
➢ Αυτό δείχνει ότι το Cluster 2 αναπαριστά ημέρες υψηλής κατανάλωσης ενέργειας, πιθανώς Σαββατοκύριακα ή ημέρες με υψηλή χρήση κλιματιστικών ή άλλων ενεργοβόρων συσκευών.

- *Ανάλυση Κατανομής Χαρακτηριστικών ανά Cluster (Boxplots)*

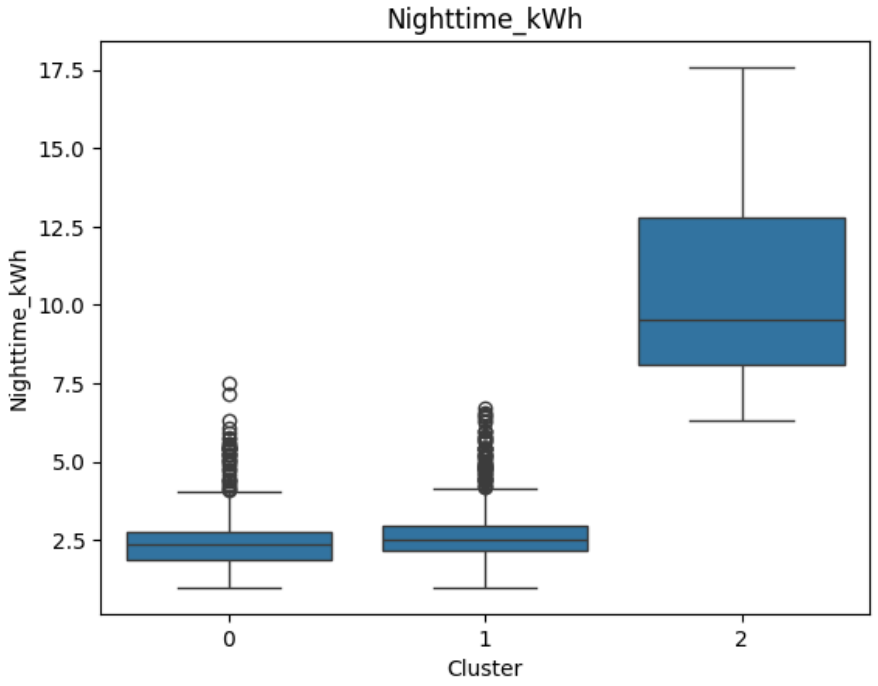
Δημιουργήθηκαν boxplots ανά cluster για τα χαρακτηριστικά: Daily_total_kWh, Nighttime_kWh και Peak_hour_power.



Cluster 0 (χαμηλή κατανάλωση): Η μέση τιμή βρίσκεται γύρω από 20 kWh, με χαμηλή διακύμανση και ελάχιστους ακραίους (outliers) πόντους. Αντιπροσωπεύει ημέρες με χαμηλή κατανάλωση ενέργειας, πιθανώς καθημερινές ημέρες.

Cluster 1 (μέτρια κατανάλωση): Η μέση τιμή κυμαίνεται γύρω από 30-35 kWh. Υπάρχουν αρκετοί ακραίοι πόντοι, υποδεικνύοντας κάποιες ημέρες με υψηλότερη κατανάλωση. Αυτό μπορεί να αφορά σαββατοκύριακα ή ημέρες με αυξημένη χρήση ενέργειας.

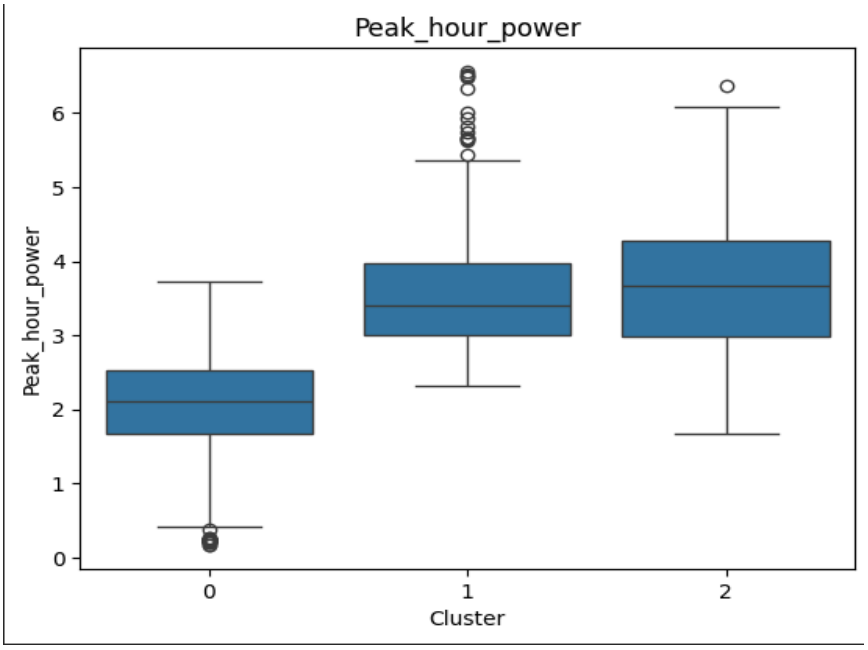
Cluster 2 (υψηλή κατανάλωση): Η μέση τιμή είναι πολύ υψηλότερη, γύρω από 40-50 kWh, και η διακύμανση είναι επίσης μεγαλύτερη. Αυτό το cluster αντιπροσωπεύει ημέρες υψηλής κατανάλωσης.



Από το boxplot για τη νυχτερινή κατανάλωση (Nighttime_kWh), παρατηρούμε τα εξής:

Cluster 0 και Cluster 1: Η νυχτερινή κατανάλωση είναι περίπου η ίδια, γύρω από 2.5-5 kWh, με αρκετούς ακραίους πόντους (outliers).

Cluster 2: Η νυχτερινή κατανάλωση είναι σημαντικά υψηλότερη, με τιμές γύρω από 10-12 kWh, υποδεικνύοντας υψηλότερη χρήση ενέργειας τη νύχτα, πιθανώς λόγω ενεργοβόρων συσκευών.



Cluster 0: Η κατανάλωση κατά τις ώρες αιχμής είναι χαμηλή, με μέση τιμή γύρω από 2 kWh, και περιορισμένη διακύμανση. Υπάρχουν μερικοί ακραίοι πόντοι, αλλά η κατανάλωση παραμένει γενικά σταθερή.

Cluster 1: Η κατανάλωση είναι μεσαία, με μέση τιμή γύρω από 3.5 kWh και κάποια ακραία σημεία που υποδηλώνουν μεγαλύτερη κατανάλωση κατά τις ώρες αιχμής.

Cluster 2: Η κατανάλωση είναι υψηλότερη, με μέση τιμή γύρω από 4 kWh και μεγαλύτερη διακύμανση. Υπάρχουν μερικά ακραία σημεία, υποδεικνύοντας ακόμα μεγαλύτερη κατανάλωση σε κάποιες περιπτώσεις.

APA:

Cluster 0 → Καθημερινές, Τυπικό Μοτίβο
Cluster 1 → Σαββατοκύριακα, Πρωινές/Βραδινές Αιχμές
Cluster 2 → Ανώμαλες/υψηλής κατανάλωσης ημέρες

Associations Rules - ΚΑΝΟΝΕΣ ΣΥΣΧΕΤΙΣΗΣ

Για την ανάλυση των δεδομένων κατανάλωσης ενέργειας, εφαρμόσαμε την τεχνική εξόρυξης κανόνων συσχέτισης (Association Rule Mining) χρησιμοποιώντας τον αλγόριθμο Apriori, με στόχο την αναγνώριση συχνών συνδυασμών χαρακτηριστικών κατανάλωσης. Η διαδικασία περιλαμβάνει τη διακριτοποίηση των δεδομένων, τη δημιουργία συναλλαγών και την εφαρμογή κανόνων συσχέτισης για την ανάλυση των σχέσεων μεταξύ των χαρακτηριστικών.

Βήμα 1:

- Έλεγχος σχήματος δεδομένων: Χρησιμοποιώ την εντολή daily.shape για να ελέγξω τις διαστάσεις του DataFrame και να δω πόσες γραμμές και στήλες περιέχει το σύνολο δεδομένων.

- Έλεγχος ονομάτων στηλών: Χρησιμοποιώ την εντολή daily.columns για να εξετάσω τα ονόματα των στηλών του DataFrame και να βεβαιωθώ ότι περιλαμβάνονται όλα τα πεδία που απαιτούνται για την εξόρυξη κανόνων συσχέτισης.

- Προεπισκόπηση των πρώτων γραμμών: Με την εντολή daily.head(), βλέπω τις πρώτες γραμμές του DataFrame για να επιβεβαιώσω τη μορφή των δεδομένων και να διασφαλίσω ότι δεν υπάρχουν προβλήματα με την εισαγωγή των δεδομένων.

Datetime	Daily_total_kWh	Daily_Sub_metering_1_kWh	Daily_Sub_metering_2_kWh	Daily_Sub_metering_3_kWh	Sub_sum_kWh	day_of_week	day_name	is_weekend	Nighttime_kWh	Peak_hour_power	Weekend_kWh	season	Daily_total_kWh_next	Cluster
2006-12-16	20.152933	0.000	0.546	4.926	5.472	5	Saturday	1	NaN	4.222889	20.152933	Winter	56.507667	1
2006-12-17	56.507667	2.033	4.187	13.341	19.561	6	Sunday	1	12.693833	3.697100	56.507667	Winter	36.730433	2
2006-12-18	36.730433	1.063	2.621	14.018	17.702	0	Monday	0	2.503900	3.050567	0.000000	Winter	27.769900	1
2006-12-19	27.769900	0.839	7.602	6.197	14.638	1	Tuesday	0	2.460200	3.879033	0.000000	Winter	37.095800	1
2006-12-20	37.095800	0.000	2.648	14.063	16.711	2	Wednesday	0	2.364600	3.646067	0.000000	Winter	28.618567	1

Βήμα 2:

Διακριτοποίηση Δεδομένων για Εξόρυξη Κανόνων Συσχέτισης

- Δημιουργία Συνάρτησης Διακριτοποίησης: Ορίζουμε τη συνάρτηση discretize_3levels που χρησιμοποιεί την pd.qcut για να διαχωρίσει τις τιμές σε τρία ισοδύναμα τμήματα (Low, Med, High), επιτρέποντας την κατηγοριοποίηση των δεδομένων σε τρεις κατηγορίες. Η παράμετρος duplicates="drop" εξασφαλίζει ότι δεν δημιουργούνται διπλές ετικέτες σε περιπτώσεις ισοδύναμων τιμών.

- Δημιουργία Αντιγράφου Δεδομένων: Κάνουμε ένα αντίγραφο του DataFrame (*daily_ar = daily.copy()*) για να εργαστούμε πάνω σε αυτό, χωρίς να επηρεάσουμε τα αρχικά δεδομένα.

- Συνολική Κατανάλωση (Total): Διακριτοποιούμε τη συνολική κατανάλωση ενέργειας (*Daily_total_kWh*) σε τρεις κατηγορίες (Low, Med, High) και αποθηκεύουμε τα αποτελέσματα στη νέα στήλη "Total".

- Νυχτερινή Κατανάλωση (Night): Διακριτοποιούμε τη νυχτερινή κατανάλωση ενέργειας (*Nighttime_kWh*) και αποθηκεύουμε τα αποτελέσματα στη νέα στήλη "Night".

- Κατανάλωση Κατά τις Ώρες Αιχμής (Peak): Διακριτοποιούμε την κατανάλωση κατά τις ώρες αιχμής (*Peak_hour_power*) και αποθηκεύουμε τα αποτελέσματα στη νέα στήλη "Peak".

- Σαββατοκύριακο (Weekend): Δημιουργούμε τη στήλη "Weekend", η οποία παίρνει την τιμή "Weekend" για τις ημέρες του Σαββατοκύριακου (*is_weekend = 1*) και "Weekday" για τις υπόλοιπες ημέρες.

- Υπομετρητές Κατανάλωσης (SM1, SM2, SM3): Διακριτοποιούμε τα Sub-meterings (η κατανάλωση από συσκευές) σε τρεις κατηγορίες και τις αποθηκεύουμε σε τρεις νέες στήλες: "SM1", "SM2", και "SM3", με βάση τα πεδία *Daily_Sub_metering_1_kWh*, *Daily_Sub_metering_2_kWh*, και *Daily_Sub_metering_3_kWh*.

- Προβολή των Επεξεργασμένων Δεδομένων: Εκτυπώνουμε τις πρώτες γραμμές του DataFrame με τα νέα διακριτοποιημένα χαρακτηριστικά για να επιβεβαιώσουμε ότι οι διακριτοποιήσεις έγιναν σωστά.

Datetime	Total	Night	Peak	Weekend	SM1	SM2	SM3
2006-12-16	Low	NaN	High	Weekend	Low	Med	Low
2006-12-17	High	High	High	Weekend	High	High	High
2006-12-18	High	Med	Med	Weekday	Med	High	High
2006-12-19	Med	Med	High	Weekday	Low	High	Low
2006-12-20	High	Med	High	Weekday	Low	High	High

Βήμα 3:

- *Επιλογή Στηλών για τα Items:*
  - Δημιουργούμε μια λίστα με τα items που θα χρησιμοποιηθούν στις συναλλαγές: *["Total", "Night", "Peak", "Weekend", "SM1", "SM2", "SM3"]*.
- *Διαγραφή NaN Τιμών:*
  - Χρησιμοποιούμε την εντολή dropna(*subset=item_cols*) για να διαγράψουμε τις γραμμές που περιέχουν ελλιπείς τιμές στα πεδία των items.

- *Δημιουργία Συναλλαγών (Transactions):*

  - Δημιουργούμε τις συναλλαγές χρησιμοποιώντας την εντολή apply() για κάθε γραμμή του DataFrame. Κάθε γραμμή μετατρέπεται σε λίστα items που περιέχουν τις κατηγορίες των χαρακτηριστικών (π.χ., *"Total=Low", "Night=Med"*).

  - Κάθε συναλλαγή είναι μια λίστα με τα items, τα οποία αντιστοιχούν στις τιμές των χαρακτηριστικών (π.χ., χαμηλή, μέτρια, υψηλή κατανάλωση).

- *Λεξικό Μετάφρασης (Pretty Labels):*

  - Δημιουργούμε ένα λεξικό pretty, το οποίο συνδέει τις κατηγορίες των χαρακτηριστικών με ετικέτες (π.χ., *"Total=Low" → "Χαμηλή Συνολική Κατανάλωση"*).

- *Συνάρτηση pretty_tx:*

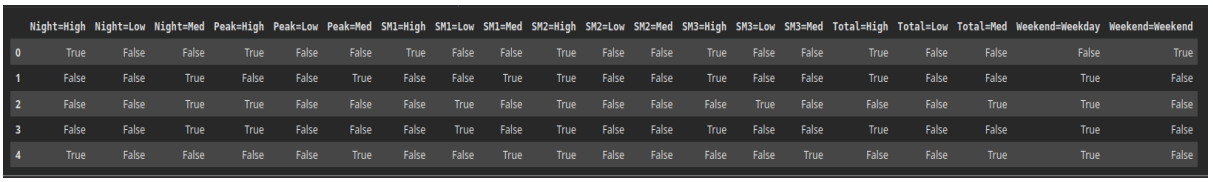
  - Ορίζουμε τη συνάρτηση *pretty_tx(tx)*, η οποία παίρνει μια συναλλαγή (μια λίστα από items) και επιστρέφει την εκτυπωμένη μορφή της συναλλαγής, με τις κατανοητές ετικέτες από το λεξικό pretty. Η συνάρτηση χρησιμοποιεί την join() για να συνδυάσει τα στοιχεία της συναλλαγής με το σύμβολο " + ".

Βήμα 4:

- Μετατροπή των συναλλαγών σε μορφή 0/1: Κάθε συναλλαγή (ημέρα) αντιστοιχεί σε μια σειρά στον πίνακα, ενώ κάθε κατηγορία χαρακτηριστικού (π.χ., *"Total=High", "Night=Low"*) αντιστοιχεί σε μία στήλη. Η τιμή 1 δείχνει ότι το χαρακτηριστικό εμφανίζεται στην εν λόγω συναλλαγή, ενώ η τιμή 0 δείχνει ότι δεν εμφανίζεται.

- Δημιουργία DataFrame: Μετατρέπουμε τα αποτελέσματα της μετατροπής σε DataFrame για ευκολότερη επεξεργασία και κατανόηση των συναλλαγών και χαρακτηριστικών.

- Παρουσιάζουμε τις πρώτες γραμμές του πίνακα, για να επιβεβαιώσουμε ότι η μετατροπή έγινε σωστά και ότι οι συναλλαγές έχουν τη σωστή μορφή για την ανάλυση.



	Night=High	Night=Low	Night=Med	Peak=High	Peak=Low	Peak=Med	SM1=High	SM1=Low	SM1=Med	SM2=High	SM2=Low	SM2=Med	SM3=High	SM3=Low	SM3=Med	Total=High	Total=Low	Total=Med	Weekend=Weekday	Weekend=Weekend
0	True	False	False	True	False	False	True	False	False	True	False	False	True	False	False	True	False	False	False	True
1	False	False	True	False	False	True	False	False	True	True	False	False	True	False	False	True	False	False	True	False
2	False	False	True	True	False	False	False	True	False	True	False	False	False	True	False	False	False	True	True	False
3	False	False	True	True	False	False	False	True	False	True	False	False	True	False	False	True	False	False	True	False
4	True	False	False	False	False	True	False	False	True	True	False	False	False	False	True	False	False	True	True	False

Βήμα 5:

- Εφαρμόζουμε τον αλγόριθμο Apriori για να εντοπίσουμε τους συχνότερους συνδυασμούς items στις συναλλαγές (δηλαδή τις ημέρες με συγκεκριμένα χαρακτηριστικά).

- Ορίζουμε την ελάχιστη υποστήριξη στο 10%, δηλαδή μόνο οι συνδυασμοί items που εμφανίζονται τουλάχιστον στο 10% των ημερών (συναλλαγών) θα θεωρούνται συχνά.

- Μετά την εκτέλεση του αλγορίθμου, ταξινομούμε τα αποτελέσματα κατά φθίνουσα υποστήριξη, ώστε οι πιο συχνοί συνδυασμοί να εμφανίζονται πρώτοι.

- Παρουσιάζουμε τους 10 πιο συχνούς συνδυασμούς items με την υψηλότερη υποστήριξη για να κατανοήσουμε ποιοι συνδυασμοί εμφανίζονται πιο συχνά στις συναλλαγές.

	support	itemsets
18	0.714781	(Weekend=Weekday)
10	0.335184	(SM2=Low)
6	0.333796	(SM1=High)
7	0.333796	(SM1=Low)
1	0.333796	(Night=Low)
4	0.333796	(Peak=Low)
14	0.333796	(SM3=Med)
15	0.333796	(Total=High)
2	0.333102	(Night=Med)
0	0.333102	(Night=High)

Βήμα 6:

- Φιλτράρισμα των Συνδυασμών: Επιλέγουμε μόνο τους συνδυασμούς που περιέχουν τουλάχιστον 2 items, χρησιμοποιώντας τη μέθοδο apply(len) για να μετρήσουμε τον αριθμό των items σε κάθε συνδυασμό.

- Ταξινομούμε τα αποτελέσματα κατά φθίνουσα υποστήριξη για να δούμε τους πιο συχνούς συνδυασμούς.

- Παρουσιάζουμε τους 10 πιο συχνούς συνδυασμούς με 2 ή περισσότερα items, για να εστιάσουμε στους πιο σημαντικούς και πολυάριθμους συνδυασμούς.

	support	itemsets
100	0.276891	(Weekend=Weekday, SM1=Med)
111	0.274809	(Weekend=Weekday, SM2=Low)
81	0.269951	(Peak=Med, Weekend=Weekday)
93	0.265094	(Weekend=Weekday, SM1=Low)
126	0.261624	(Weekend=Weekday, Total=Low)
123	0.260236	(Weekend=Weekday, SM3=Med)
46	0.260236	(Weekend=Weekday, Night=Low)
73	0.256072	(Peak=Low, Weekend=Weekday)
59	0.254684	(Night=Med, Weekend=Weekday)
127	0.253296	(Weekend=Weekday, Total=Med)

Βήμα 7:

- Χρησιμοποιούμε τον αλγόριθμο association_rules για να δημιουργήσουμε κανόνες συσχέτισης από τα συχνά itemsets που βρήκαμε νωρίτερα.

- Επιλέγουμε κανόνες που έχουν τουλάχιστον 60% αξιοπιστία (confidence). Αυτό σημαίνει ότι αν ισχύει το antecedent (προηγούμενο στοιχείο του κανόνα), τότε το consequent (επόμενο στοιχείο) θα ισχύει τουλάχιστον στο 60% των περιπτώσεων.

- Ταξινομούμε τους κανόνες κατά διάσταση "lift", "confidence" και "support" σε φθίνουσα σειρά για να εμφανιστούν πρώτα οι πιο ισχυροί κανόνες.

- Παρουσιάζουμε τους 10 πιο ισχυρούς κανόνες, με τα στοιχεία τους, όπως τα antecedents (προηγούμενα items), consequents (επόμενα items), support, confidence και lift.

	antecedents	consequents	support	confidence	lift
69	(Total=Low, SM1=Low)	(Peak=Low, SM3=Low)	0.140180	0.664474	3.395413
68	(Peak=Low, SM3=Low)	(Total=Low, SM1=Low)	0.140180	0.716312	3.395413
136	(SM3=High, Peak=High)	(SM2=High, Total=High)	0.111728	0.665289	3.317238
178	(SM3=High, Peak=High)	(SM1=High, Total=High)	0.101319	0.603306	3.196190
70	(SM3=Low, SM1=Low)	(Peak=Low, Total=Low)	0.140180	0.779923	3.174770
137	(SM3=High, SM2=High)	(Peak=High, Total=High)	0.111728	0.759434	3.153730
189	(SM2=Med, SM1=Low)	(Peak=Low, Total=Low)	0.100625	0.728643	2.966031
65	(Peak=Low, SM1=Low, SM3=Low)	(Total=Low)	0.140180	0.966507	2.901535
174	(SM3=High, SM1=High, Peak=High)	(Total=High)	0.101319	0.960526	2.877585
132	(SM3=High, SM2=High, Peak=High)	(Total=High)	0.111728	0.958333	2.871015

Support: Αντιπροσωπεύει το ποσοστό των συναλλαγών (ημερών) που περιλαμβάνουν τον συγκεκριμένο κανόνα. Όλοι οι κανόνες έχουν support γύρω από 0.1, δηλαδή εμφανίζονται σε περίπου 10% των συναλλαγών.

Confidence: Αντιπροσωπεύει την πιθανότητα ότι το consequent (επόμενο item) θα ισχύει δεδομένου ότι το antecedent (προηγούμενο item) ισχύει. Οι τιμές του confidence κυμαίνονται από 0.6 έως 0.96, με τον κανόνα 65 (*Peak=Low, SM1=Low, SM3=Low → Total=Low*) να έχει την υψηλότερη confidence (*0.966*).

Lift: Αντιπροσωπεύει το πόσο καλύτερα προγνωστικά το antecedent επηρεάζει το consequent σε σχέση με την τυχαία πιθανότητα. Οι κανόνες με υψηλότερο lift είναι πιο ενδιαφέροντες, καθώς υποδεικνύουν πιο ισχυρές συσχετίσεις. Ο κανόνας 69 (*Total=Low → Peak=Low, SM1=Low*) έχει το υψηλότερο lift (*3.395*).

TIME SERIES FORECASTING- Πρόβλεψη Χρονοσειρών

Για την πρόβλεψη χρονοσειρών, εφαρμόσαμε τρία διαφορετικά μοντέλα (ARIMA, Prophet και LSTM) για την πρόβλεψη της ημερήσιας κατανάλωσης ενέργειας. Η διαδικασία περιλαμβάνει τον διαχωρισμό των δεδομένων σε train και test set, την εκπαίδευση των μοντέλων και την αξιολόγηση της απόδοσής τους με δείκτες όπως το RMSE και το MAPE.

Βήμα 1:

- Επιλέγουμε τη στήλη "*Daily_total_kWh*" ως τη χρονοσειρά που θα προβλέψουμε.
- Ρυθμίζουμε τη συχνότητα της χρονοσειράς σε ημερήσια (daily frequency) χρησιμοποιώντας την εντολή *asfreq("D")*.
- Χρησιμοποιούμε την γραμμική παρεμβολή (*interpolate("time")*) για να γεμίσουμε τα κενά δεδομένα, ώστε να επιτρέπεται η ομαλή εφαρμογή των μοντέλων πρόβλεψης.
- Ελέγχουμε τα πρώτα δεδομένα της χρονοσειράς με την εντολή *head()* για να επιβεβαιώσουμε τη σωστή προετοιμασία.

Datetime	Daily_total_kWh
2006-12-16	20.152933
2006-12-17	56.507667
2006-12-18	36.730433
2006-12-19	27.769900
2006-12-20	37.095800

Βήμα 2:

- Χωρίζουμε τη χρονοσειρά σε 80% για εκπαίδευση (train) και 20% για δοκιμή (test).
- Ορίζουμε το σημείο διαχωρισμού με τη μέθοδο *int(len(ts) * 0.8)* για να υπολογίσουμε το 80% του μήκους της χρονοσειράς.
- Χρησιμοποιούμε την iloc για να χωρίσουμε τα δεδομένα σε train και test:
  - Το train set περιλαμβάνει το πρώτο 80% της χρονοσειράς.
  - Το test set περιλαμβάνει το υπόλοιπο 20%.
- Ελέγχουμε το μέγεθος των δεδομένων για κάθε σύνολο χρησιμοποιώντας την shape για να βεβαιωθούμε ότι ο διαχωρισμός έγινε σωστά.

```
(1153,) (289,)
```

Βήμα 3:

- Επιλέγεται η χρονοσειρά ημερήσιας συνολικής κατανάλωσης ενέργειας (*Daily_total_kWh*) ως βάση για όλα τα μοντέλα πρόβλεψης.

- Διασφαλίζεται η ημερήσια συχνότητα της χρονοσειράς και τα πιθανά κενά δεδομένα συμπληρώνονται με χρονική παρεμβολή, ώστε να εξασφαλιστεί ομαλή εφαρμογή των μοντέλων.

- Τα δεδομένα διαχωρίζονται χρονολογικά σε 80% σύνολο εκπαίδευσης (train) και 20% σύνολο δοκιμής (test), με σκοπό την αντικειμενική αξιολόγηση των προβλέψεων.

- Ορίζονται κοινές μετρικές αξιολόγησης για όλα τα μοντέλα:

    - *RMSE (Root Mean Squared Error)*, για την εκτίμηση του μέσου σφάλματος πρόβλεψης σε απόλυτες μονάδες (kWh).

    - *MAPE (Mean Absolute Percentage Error)*, για την εκτίμηση του μέσου ποσοστιαίου σφάλματος, με πρόβλεψη αριθμητικής σταθερότητας μέσω μικρής σταθεράς (eps).

Βήμα 4:

- Εφαρμόζεται το μοντέλο ARIMA(2,1,2) στο σύνολο εκπαίδευσης (train) για την πρόβλεψη της ημερήσιας κατανάλωσης ενέργειας.

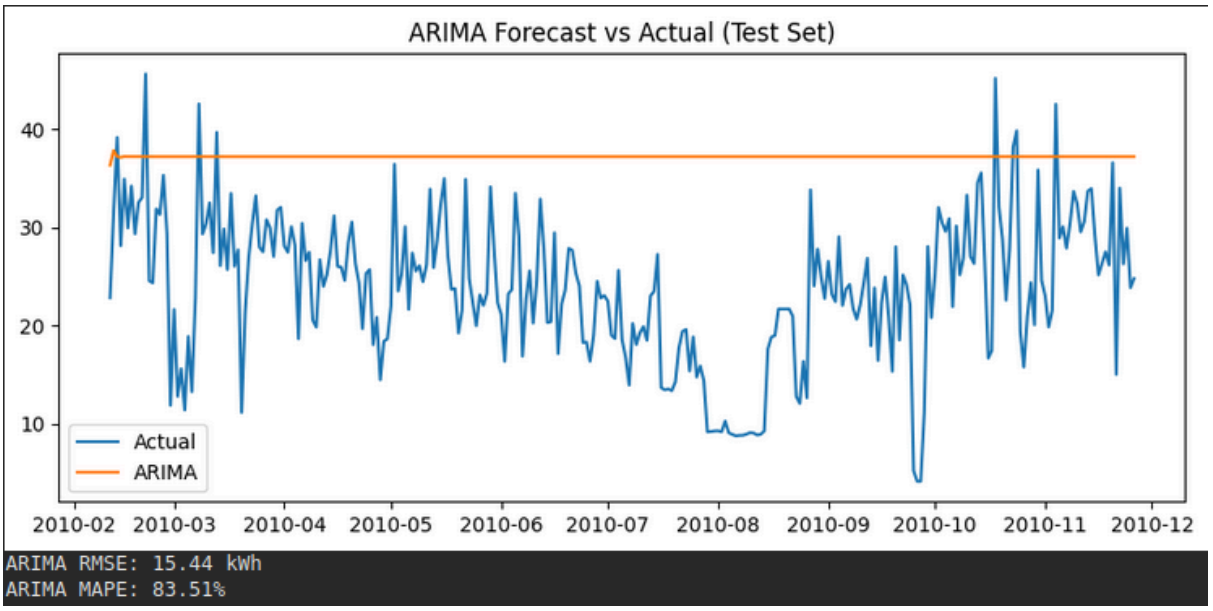
- Το εκπαιδευμένο μοντέλο χρησιμοποιείται για την παραγωγή προβλέψεων για ολόκληρο το test set, ώστε να είναι δυνατή η άμεση σύγκριση με τις πραγματικές τιμές.

- Η απόδοση του μοντέλου αξιολογείται με τους δείκτες:

    - *RMSE*, που εκφράζει το μέσο σφάλμα πρόβλεψης σε kWh.

    - *MAPE*, που εκφράζει το μέσο ποσοστιαίο σφάλμα πρόβλεψης.

- Παρουσιάζεται γραφικά η σύγκριση μεταξύ πραγματικών τιμών και προβλέψεων ARIMA, επιτρέποντας τον οπτικό έλεγχο της προσαρμογής του μοντέλου.



- Το μοντέλο ARIMA αποτυπώνει μόνο τη γενική μέση στάθμη της ημερήσιας κατανάλωσης, χωρίς να καταφέρνει να ακολουθήσει τη δυναμική και τις έντονες διακυμάνσεις της χρονοσειράς.
- Οι προβλέψεις εμφανίζονται σχεδόν επίπεδες, γεγονός που υποδηλώνει αδυναμία του μοντέλου να συλλάβει την εποχικότητα και τις βραχυχρόνιες μεταβολές της κατανάλωσης.
- Οι υψηλές τιμές RMSE (*15.44 kWh*) και ιδιαίτερα MAPE *(83.51%*) καταδεικνύουν χαμηλή προγνωστική ακρίβεια, επιβεβαιώνοντας ότι το ARIMA δεν είναι κατάλληλο για δεδομένα με έντονη εποχικότητα και μεταβλητότητα.

Βήμα 5:

- Εφαρμόζεται το μοντέλο Prophet, το οποίο είναι κατάλληλο για χρονοσειρές με τάση (trend) και εποχικότητα (seasonality).

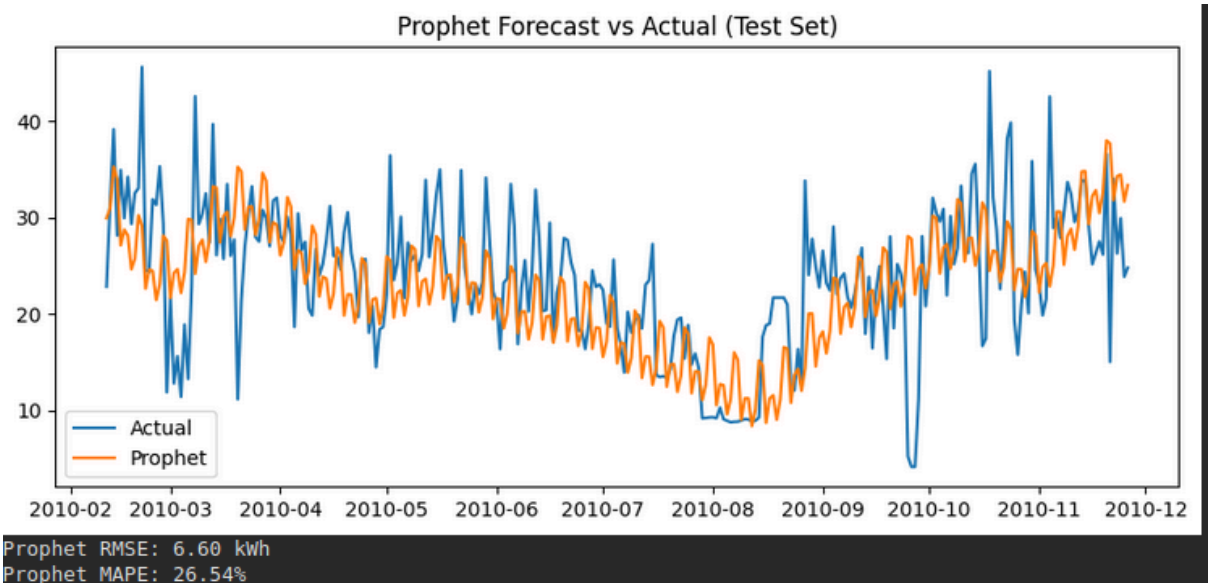
- Τα δεδομένα εκπαίδευσης μετασχηματίζονται στη μορφή που απαιτεί το Prophet, με τις στήλες:
  - *ds:* χρονική στιγμή (ημερομηνία),
  - *y:* τιμή της χρονοσειράς (ημερήσια κατανάλωση).

- Το μοντέλο εκπαιδεύεται λαμβάνοντας υπόψη:
  - ετήσια εποχικότητα (yearly seasonality),
  - εβδομαδιαία εποχικότητα (weekly seasonality),
  - χωρίς ημερήσια εποχικότητα, καθώς τα δεδομένα είναι ημερήσια.

- Παράγονται προβλέψεις για όλο το test set, ώστε να είναι δυνατή η άμεση σύγκριση με τις πραγματικές τιμές.

- Η απόδοση του μοντέλου αξιολογείται με τους δείκτες:
  - *RMSE*, για την εκτίμηση του μέσου σφάλματος πρόβλεψης σε kWh
  - *MAPE*, για την εκτίμηση του μέσου ποσοστιαίου σφάλματος

- Παρουσιάζεται γραφικά η σύγκριση μεταξύ πραγματικών τιμών και προβλέψεων Prophet, επιτρέποντας την οπτική αξιολόγηση της προσαρμογής του μοντέλου.



- Το μοντέλο Prophet αποτυπώνει ικανοποιητικά τη μακροχρόνια τάση και την εποχικότητα της ημερήσιας κατανάλωσης ενέργειας, παρουσιάζοντας σαφώς βελτιωμένη προσαρμογή σε σχέση με το ARIMA.
- Οι προβλέψεις ακολουθούν πιο πιστά τη γενική εξέλιξη της χρονοσειράς, αν και εξακολουθούν να εξομαλύνουν τις έντονες βραχυχρόνιες διακυμάνσεις.
- Οι χαμηλότερες τιμές RMSE (*6.60 kWh*) και MAPE (*26.54%*) υποδηλώνουν σημαντικά καλύτερη προγνωστική ικανότητα, επιβεβαιώνοντας την καταλληλότητα του Prophet για δεδομένα με έντονη εποχικότητα.

Βήμα 6:

- Εφαρμόζεται μοντέλο LSTM (Long Short-Term Memory), κατάλληλο για χρονοσειρές, με στόχο να «μάθει» χρονικές εξαρτήσεις από προηγούμενες ημέρες κατανάλωσης.

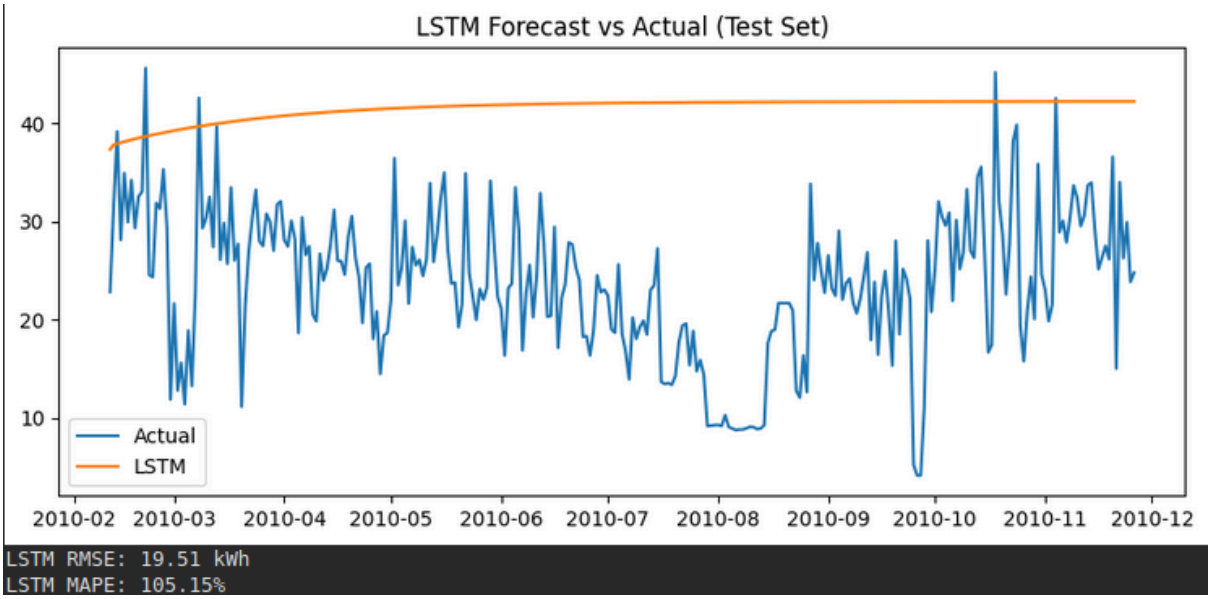
- Χρησιμοποιείται παράθυρο ιστορικού 30 ημερών (*lookback=30*), ώστε κάθε πρόβλεψη να βασίζεται στις 30 προηγούμενες τιμές της χρονοσειράς.

- Πραγματοποιείται κανονικοποίηση (MinMax scaling) μόνο στο train set για αποφυγή data leakage, και στη συνέχεια δημιουργούνται ακολουθίες εισόδου/στόχου (sequences) για εκπαίδευση του δικτύου.

- Εκπαιδεύεται ένα απλό δίκτυο LSTM(32) → Dense(1) με loss MSE και optimizer Adam, για 20 epochs.

- Οι προβλέψεις στο test set παράγονται με recursive multi-step forecasting, δηλαδή κάθε νέα πρόβλεψη τροφοδοτείται ως είσοδος για την επόμενη ημέρα.

- Οι προβλέψεις επαναφέρονται στην αρχική κλίμακα (inverse transform) και αξιολογούνται με RMSE και MAPE, ενώ παρουσιάζεται και γράφημα σύγκρισης Actual vs LSTM Forecast για οπτική αξιολόγηση.



- Το μοντέλο LSTM εμφανίζει έντονη εξομάλυνση στις προβλέψεις και αδυνατεί να ακολουθήσει τις βραχυχρόνιες διακυμάνσεις και τις αιχμές της ημερήσιας κατανάλωσης.
- Η προβλεπόμενη καμπύλη εξελίσσεται σχεδόν μονοτονικά, υποδηλώνοντας ανεπαρκή αξιοποίηση της χρονικής πληροφορίας του ιστορικού παραθύρου.
- Οι υψηλές τιμές RMSE και MAPE επιβεβαιώνουν τη χαμηλή προγνωστική απόδοση του LSTM στη συγκεκριμένη ρύθμιση.

Βήμα 7:

Δημιουργείται συγκεντρωτικός πίνακας αποτελεσμάτων που περιλαμβάνει τα τρία μοντέλα πρόβλεψης (ARIMA, Prophet, LSTM) και τις αντίστοιχες μετρικές αξιολόγησης RMSE και MAPE.

	Model	RMSE	MAPE%
1	Prophet	6.595113	26.541062
0	ARIMA(2,1,2)	15.442242	83.505033
2	LSTM	19.508587	105.146251

Best by RMSE: Prophet (RMSE=6.60, MAPE=26.54%

- Το Prophet παρουσιάζει τη βέλτιστη απόδοση, με το χαμηλότερο RMSE (*≈6.6 kWh*) και σαφώς μικρότερο MAPE (*≈26.5%*), υποδεικνύοντας υψηλότερη ακρίβεια πρόβλεψης.

- Το ARIMA(2,1,2) εμφανίζει σημαντικά υψηλότερα σφάλματα (*RMSE ≈15.4 kWh, MAPE ≈83.5%*), γεγονός που δείχνει περιορισμένη ικανότητα αποτύπωσης της μεταβλητότητας.

- Το LSTM παρουσιάζει τη χαμηλότερη απόδοση, με πολύ υψηλό MAPE, υποδηλώνοντας ανεπαρκή προσαρμογή στη συγκεκριμένη χρονοσειρά.

Συμπέρασμα: Το Prophet αποτελεί το καταλληλότερο μοντέλο για την πρόβλεψη της ημερήσιας κατανάλωσης ενέργειας στο συγκεκριμένο σύνολο δεδομένων.