

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

1. Εισαγωγή

Στο πλαίσιο της αυξανόμενης παγκόσμιας ζήτησης ενέργειας και της ανάπτυξης των έξυπνων δικτύων (smart grids), η ανάλυση της κατανάλωσης ηλεκτρικής ενέργειας στα νοικοκυριά καθίσταται ολοένα και πιο κρίσιμη.

Η παρούσα εργασία εφαρμόζει τεχνικές **εξόρυξης δεδομένων (data mining)** στο **σύνολο δεδομένων [Individual Household Electric Power Consumption](#)** (UCI Machine Learning Repository), με στόχο την ανακάλυψη προτύπων κατανάλωσης, την πρόβλεψη μελλοντικής χρήσης και την ταξινόμηση συμπεριφορών του νοικοκυριού.

Το σύνολο δεδομένων περιλαμβάνει πάνω από **2 εκατομμύρια μετρήσεις** που συλλέχθηκαν σε διάστημα σχεδόν 4 ετών (2006–2010) από ένα μόνο νοικοκυριό και περιλαμβάνει μετρήσεις υπομετρητών για συσκευές, τάσης και έντασης ρεύματος. Τα δεδομένα είναι **μη καθαρά (messy)**, με **ελλιπείς τιμές**, **ακανόνιστα χρονικά διαστήματα** και **μικτά αριθμητικά και κατηγορικά γνωρίσματα**, γεγονός που τα καθιστά ιδανικά για ρεαλιστική εξάσκηση στις τεχνικές εξόρυξης δεδομένων.

2. Στόχοι

Η εργασία αποσκοπεί στα εξής:

1. **Προεπεξεργασία και καθαρισμό** των πραγματικών δεδομένων κατανάλωσης ενέργειας, αντιμετωπίζοντας θόρυβο και ελλιπείς τιμές.
2. **Εξαγωγή ουσιαστικών χαρακτηριστικών**, όπως ημερήσια/εβδομαδιαία αθροίσματα, χρονικά και μετεωρολογικά χαρακτηριστικά.
3. Εφαρμογή πολλαπλών τεχνικών εξόρυξης δεδομένων:
 - ο **Ταξινόμηση (Classification)**: Εντοπισμός ημερών με ανώμαλη ή υψηλή κατανάλωση.
 - ο **Παλινδρόμηση (Regression)**: Πρόβλεψη της συνολικής κατανάλωσης της επόμενης ημέρας.
 - ο **Ομαδοποίηση (Clustering)**: Αναγνώριση τυπικών προφίλ ημερήσιας κατανάλωσης.
 - ο **Κανόνες Συσχέτισης (Association Rules)**: Ανακάλυψη συσχετίσεων μεταξύ χρήσης συσκευών και χρονικής στιγμής της ημέρας.
4. Προαιρετικά, να συμπεριληφθεί **πρόβλεψη χρονοσειρών** με προχωρημένα μοντέλα.

3. Περιγραφή Συνόλου Δεδομένων

Σύνολο Δεδομένων: *Individual household electric power consumption*

Χαρακτηριστικά:

Μεταβλητή	Περιγραφή
<code>Date, Time</code>	Χρονική σήμανση της μέτρησης (ανά λεπτό)
<code>Global_active_power</code>	Ολική ενεργή ισχύς του νοικοκυριού (kW)
<code>Global_reactive_power</code>	Αντιδραστική ισχύς (kW)
<code>Voltage</code>	Μέση τάση (V)
<code>Global_intensity</code>	Ένταση ρεύματος (A)
<code>Sub_metering_1</code>	Κατανάλωση κουζίνας (Wh)
<code>Sub_metering_2</code>	Κατανάλωση πλυντηρίου/συσκευών πλύσης (Wh)
<code>Sub_metering_3</code>	Κατανάλωση θερμοσίφωνα και κλιματισμού (Wh)

Ζητήματα:

- Ελλιπείς ή μη έγκυρες μετρήσεις (? ή μηδενικές τιμές)
- Ακανόνιστες χρονικές σφραγίδες
- Ισχυρά εποχικά πρότυπα (ημερήσια/εβδομαδιαία)
- Ανάγκη για ομαδοποίηση και δημιουργία νέων χαρακτηριστικών

4. Προεπεξεργασία Δεδομένων και Εμπλουτισμός Χαρακτηριστικών

Βήματα:

1. Καθαρισμός Δεδομένων

- Αφαίρεση ή συμπλήρωση ελλιπών τιμών.
- Μετατροπή των χρονικών στιγμών σε αντικείμενα `datetime` και επαναδειγματοληψία (resampling).

2. Δημιουργία Χαρακτηριστικών

- Συγκεντρωτικά χαρακτηριστικά:
 - `Daily_total_power = sum(Global_active_power)`
 - `Peak_hour_power, Nighttime_usage, Weekend_usage`.
- Κατηγορικά χαρακτηριστικά: ημέρα εβδομάδας, εποχή, εργάσιμη/μη εργάσιμη ημέρα.
- Προαιρετικά: ενσωμάτωση **εξωτερικών δεδομένων καιρού** (θερμοκρασία, υγρασία).

3. Κανονικοποίηση (Normalization)

- Κανονικοποίηση των συνεχών χαρακτηριστικών όταν απαιτείται.

5. Μεθοδολογία

5.1 Ταξινόμηση (Classification)

Στόχος: Πρόβλεψη εάν η κατανάλωση μιας ημέρας είναι “Υψηλή” ή “Χαμηλή” σε σχέση με τον μέσο όρο του νοικοκυριού.

- **Μεταβλητή στόχος:** Δυαδική (1 = Υψηλή κατανάλωση, 0 = Κανονική).
- **Αξιολόγηση:** Accuracy, F1-score, ROC-AUC.

5.2 Παλινδρόμηση (Regression)

Στόχος: Πρόβλεψη της συνολικής κατανάλωσης ενέργειας της επόμενης ημέρας (σε kWh).

- **Αξιολόγηση:** MAE, RMSE, R².

5.3 Ομαδοποίηση (Clustering)

Στόχος: Εντοπισμός ομάδων ημερών με παρόμοια προφίλ κατανάλωσης.

- **Αξιολόγηση:** Silhouette Score, Davies–Bouldin Index.
- **Ερμηνεία:** π.χ.
 - Συστάδα 1 → Καθημερινές, τυπικό μοτίβο.
 - Συστάδα 2 → Σαββατοκύριακα, πρωινές/βραδινές αιχμές.
 - Συστάδα 3 → Ανώμαλες/υψηλής κατανάλωσης ημέρες.

5.4 Εξόρυξη Κανόνων Συσχέτισης (Association Rule Mining)

Στόχος: Ανακάλυψη συχνών προτύπων μεταξύ χρήσης συσκευών και χρονικών περιόδων.

- **Προεπεξεργασία:**

- Διακριτοποίηση συνεχών χαρακτηριστικών (π.χ. Χαμηλή/Μέτρια/Υψηλή χρήση υπομετρητών).
- Κωδικοποίηση ως συναλλαγές (π.χ. “Υψηλό Sub_metering_1 + Νύχτα” → “Υψηλό Global_active_power”).
- **Αλγόριθμοι:** Apriori ή FP-Growth (μέσω `mlxtend`).
- **Μετρικές:** Support, Confidence, Lift.
- **Παράδειγμα Κανόνα:**
 - *AN (Υψηλό Sub_metering_3) ΚΑΙ (Σαββατοκύριακο) → Υψηλή Συνολική Κατανάλωση (Support=0.18, Confidence=0.74).*

5.5 (Προαιρετικά) Πρόβλεψη Χρονοσειρών

Στόχος: Πρόβλεψη μελλοντικής ωριαίας/ημερήσιας κατανάλωσης.

- **Μέθοδοι:** ARIMA, Prophet ή LSTM.
- **Αξιολόγηση:** RMSE, MAPE.

6. Αξιολόγηση & Ερμηνεία Αποτελεσμάτων

Εργασία	Μετρική	Αναμενόμενο Αποτέλεσμα
Ταξινόμηση	Accuracy, F1	Αναγνώριση ημερών υψηλής κατανάλωσης με ακρίβεια >85%
Παλινδρόμηση	RMSE	Πρόβλεψη επόμενης ημέρας με απόκλιση ± 0.3 kWh
Ομαδοποίηση	Silhouette	Διακριτές ομάδες καθημερινών/σαββατοκύριακων
Κανόνες Συσχέτισης	Lift, Confidence	Εντοπισμός ουσιαστικών συσχετίσεων στη χρήση

Ερμηνεία προτύπων σε σχέση με:

- Εποχικές και ημερήσιες επιδράσεις
- Απόδοση σε επίπεδο συσκευών
- Δυνατότητες εξοικονόμησης ενέργειας

7. Αναμενόμενα Παραδοτέα

Η ημερομηνία παράδοσης είναι στις 15/1/2026 στο eclass. Θα υπάρχει σύντομη παρουσίαση 10 λεπτών στην αίθουσα.

1. **Notebook προεπεξεργασίας δεδομένων** (καθαρισμός + εξαγωγή χαρακτηριστικών).
2. **Notebook μοντελοποίησης** (ταξινόμηση, παλινδρόμηση, ομαδοποίηση, συσχετίσεις).
3. **Τελική αναφορά**, η οποία θα περιλαμβάνει:

- ο Εισαγωγή
- ο Μεθοδολογία και πειραματική διαδικασία
- ο Αποτελέσματα και συζήτηση
- ο Συμπεράσματα και περιορισμοί

Οι εργασίες θα βαθμολογηθούν:

- i) για την προετοιμασία: η περιγραφή των ενεργειών (10%)
- ii) για την ταξινόμηση: η περιγραφή των ενεργειών και η αξιολόγηση των αλγορίθμων (20%),
- iii) για την παλινδρόμηση: οι επιδόσεις που είχατε στα δεδομένα εκπαίδευσης (20%)
- iv) για τη συσταδοποίηση: η μεθοδολογία που ακολουθήσατε και η ερμηνεία των συστάδων (20%)
- v) για την εξόρυξη κανόνων: η περιγραφή της μεθοδολογίας και η ερμηνεία των παραγόμενων κανόνων (20%)
- vi) η τελική παρουσίαση της εργασίας (10%)
- vii) οι προβλέψεις στις χρονοσειρές (10%)

Η εργασία είναι για 3 (το πολύ) άτομα.