



ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΙΓΑΙΟΥ

ΠΜΣ «Πληροφοριακά και Επικοινωνιακά Συστήματα»

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

**Εφαρμογή Τεχνικών Μηχανικής Μάθησης για Πρόβλεψη και Ανίχνευση Ανωμαλιών σε
Δεδομένα IoT**

Αθανάσιος Ουρανίδης

**Επιβλέπων Καθηγητής:
Χρήστος Γκουμόπουλος**

**Πανεπιστήμιο Αιγαίου
Μάρτιος 2026**

ΜΕΤΑΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ

EL: Εφαρμογή Τεχνικών Μηχανικής Μάθησης για Πρόβλεψη και Ανίχνευση Ανωμαλιών σε
Δεδομένα IoT

EN: Application of Machine Learning Techniques for Prediction and Anomaly Detection in IoT
Data

Αθανάσιος Ουρανίδης

A.M.: 3252024019

ΠΕΡΙΛΗΨΗ

Η παρούσα εργασία διερευνά την εφαρμογή τεχνικών μηχανικής μάθησης σε δεδομένα Διαδικτύου των Πραγμάτων (IoT), με εστίαση στην πρόβλεψη χρονοσειρών και την ανίχνευση ανωμαλιών. Η έρευνα αξιοποιεί δύο ευρέως γνωστά σύνολα δεδομένων αναφοράς που αντιπροσωπεύουν διαφορετικούς τομείς εφαρμογής: το UCI Air Quality Dataset, με περίπου 9.500 ωριαίες μετρήσεις ατμοσφαιρικών ρύπων και περιβαλλοντικών παραμέτρων (CO, NO_x, NO₂, C₆H₆, θερμοκρασία, υγρασία), και το Individual Household Electric Power Consumption Dataset, με περισσότερες από δύο εκατομμύρια εγγραφές κατανάλωσης ηλεκτρικής ενέργειας ανά λεπτό από νοικοκυριό στη Γαλλία.

Για την πρόβλεψη μίας ώρας εμπρός, υλοποιούνται και συγκρίνονται συστηματικά έξι αλγόριθμοι, από απλά μοντέλα βάσης έως σύνθετες αρχιτεκτονικές βαθιάς μάθησης. Τα πειραματικά αποτελέσματα αναδεικνύουν την υπεροχή του Random Forest, ο οποίος επιτυγχάνει τις καλύτερες επιδόσεις στο σύνολο ποιότητας αέρα με MAE 0.092 mg/m³, RMSE 0.112 mg/m³ και R² 0.750. Γενικότερα, τα μοντέλα βασισμένα σε δέντρα αποφάσεων υπερτερούν σημαντικά και στα δύο σύνολα δεδομένων, ενώ οι αρχιτεκτονικές βαθιάς μάθησης (LSTM, GRU) εμφανίζουν χαμηλότερες επιδόσεις λόγω ανεπαρκών δεδομένων εκπαίδευσης.

Στον τομέα της ανίχνευσης ανωμαλιών, αξιολογούνται εννέα διαφορετικές μέθοδοι και αναπτύσσεται μια προσέγγιση ensemble που συνδυάζει πολλαπλούς ανιχνευτές μέσω ψηφοφορίας πλειοψηφίας. Η μέθοδος STL Decomposition αναδεικνύεται ως η πλέον συντηρητική, ανιχνεύοντας μόνο 2.6% των παρατηρήσεων ως ανώμαλες, σε αντίθεση με μεθόδους όπως ο LOF και το One-Class SVM που τείνουν να υπερ-ανιχνεύουν. Η προσέγγιση ensemble παρέχει πιο ισορροπημένα και εύρωστα αποτελέσματα σε σχέση με τις μεμονωμένες μεθόδους.

Τα κύρια ευρήματα υποδεικνύουν ότι η σχεδίαση χαρακτηριστικών αποτελεί κρίσιμο παράγοντα επιτυχίας ανεξαρτήτως αλγορίθμου, με τα χαρακτηριστικά καθυστέρησης και τα κυλιόμενα στατιστικά να προσφέρουν τη μεγαλύτερη προγνωστική αξία. Επιπλέον, η διατομεακή επικύρωση επιβεβαιώνει τη γενικευσιμότητα των αποτελεσμάτων σε διαφορετικούς τομείς IoT, ενισχύοντας την πρακτική αξία της προτεινόμενης μεθοδολογίας.

Λέξεις-κλειδιά: Διαδίκτυο των Πραγμάτων (IoT), Μηχανική Μάθηση, Πρόβλεψη Χρονοσειρών, Ανίχνευση Ανωμαλιών, Random Forest, XGBoost, LSTM, CRISP-DM, Ποιότητα Αέρα, Κατανάλωση Ενέργειας, Ensemble Methods

ABSTRACT

This thesis investigates the application of machine learning techniques to Internet of Things (IoT) data, focusing on time series forecasting and anomaly detection. The research utilizes two widely recognized benchmark datasets representing different IoT application domains: the UCI Air Quality Dataset, comprising approximately 9,500 hourly measurements of atmospheric pollutants and environmental parameters (CO, NO_x, NO₂, C₆H₆, temperature, humidity), and the Individual Household Electric Power Consumption Dataset, containing over two million per-minute electrical energy consumption records from a household in France.

For one-hour-ahead forecasting, six algorithms are systematically implemented and compared, ranging from simple baseline models to complex deep learning architectures. The experimental results demonstrate the superiority of Random Forest, which achieves the best performance on the air quality dataset with a Mean Absolute Error (MAE) of 0.092 mg/m³, Root Mean Squared Error (RMSE) of 0.112 mg/m³, and coefficient of determination R² of 0.750. Overall, tree-based models significantly outperform other approaches across both datasets, while deep learning architectures (LSTM, GRU) exhibit lower performance due to insufficient training data.

In the domain of anomaly detection, nine different methods are evaluated and an ensemble approach is developed that combines multiple detectors through majority voting. STL Decomposition emerges as the most conservative method, identifying only 2.6% of observations as anomalous, in contrast to methods such as LOF and One-Class SVM which tend to over-detect. The ensemble approach provides more balanced and robust results compared to individual methods.

The key findings indicate that feature engineering constitutes a critical success factor regardless of the selected algorithm, with lag features and rolling statistics offering the greatest predictive value. Furthermore, cross-domain validation confirms the generalizability of the results across different IoT domains, reinforcing the practical value of the proposed methodology.

Keywords: Internet of Things (IoT), Machine Learning, Time Series Forecasting, Anomaly Detection, Random Forest, XGBoost, LSTM, CRISP-DM, Air Quality, Energy Consumption, Ensemble Methods

ΠΙΝΑΚΑΣ ΠΕΡΙΕΧΟΜΕΝΩΝ

ΠΙΝΑΚΑΣ ΠΕΡΙΕΧΟΜΕΝΩΝ.....	9
ΚΑΤΑΛΟΓΟΣ ΕΙΚΟΝΩΝ	15
ΚΑΤΑΛΟΓΟΣ ΠΙΝΑΚΩΝ	17
ΚΕΦΑΛΑΙΟ 1: ΕΙΣΑΓΩΓΗ.....	19
1.1 Υπόβαθρο και Κίνητρο	19
1.2 Στόχοι της Διπλωματικής.....	20
1.3 Δομή της Εργασίας	21
ΚΕΦΑΛΑΙΟ 2: ΑΝΑΣΚΟΠΗΣΗ ΒΙΒΛΙΟΓΡΑΦΙΑΣ.....	23
2.1 IoT και Χρονοσειρές.....	23
2.1.1 Εισαγωγή στο Διαδίκτυο των Πραγμάτων	23
2.1.2 Χρονοσειρές και Ανάλυση Χρονικά Εξαρτώμενων Δεδομένων.....	24
2.2 Μηχανική Μάθηση	25
2.2.1 Θεμελιώδεις Αρχές	25
2.2.2 Μοντέλα Βασισμένα σε Δέντρα Αποφάσεων.....	26
2.2.3 Βαθιά Μάθηση για Χρονοσειρές.....	27
2.3 Ανίχνευση Ανωμαλιών	28
2.3.1 Επισκόπηση και Κατηγοριοποίηση	28
2.3.2 Στατιστικές Μέθοδοι.....	29
2.3.3 Μέθοδοι Μηχανικής Μάθησης.....	29
2.3.4 Μέθοδοι Χρονοσειρών	30
2.4 Μεθοδολογία CRISP-DM.....	30
ΚΕΦΑΛΑΙΟ 3: ΜΕΘΟΔΟΛΟΓΙΑ ΚΑΙ ΔΕΔΟΜΕΝΑ	33
Εισαγωγή στο Κεφάλαιο.....	33
3.1 Σύνολα Δεδομένων (Datasets)	33
3.1.1 Σύνολο Δεδομένων Ποιότητας Αέρα (UCI Air Quality Dataset).....	34

3.1.2 Σύνολο Δεδομένων Κατανάλωσης Ηλεκτρικής Ενέργειας (Individual Household Electric Power Consumption Dataset)	36
3.1.3 Συγκριτική Ανάλυση των Συνόλων Δεδομένων	37
3.2 Εργαλεία και Περιβάλλον (Tools & Environment)	39
3.2.1 Γλώσσα Προγραμματισμού	39
3.2.2 Βιβλιοθήκες Επεξεργασίας Δεδομένων	40
3.2.3 Βιβλιοθήκες Οπτικοποίησης	40
3.2.4 Βιβλιοθήκες Μηχανικής Μάθησης	40
3.2.5 Βιβλιοθήκες Βαθιάς Μάθησης	41
3.2.6 Περιβάλλον Ανάπτυξης και Εκτέλεσης	41
3.2.7 Πίνακας Εξαρτήσεων Λογισμικού	42
3.2.8 Διαχείριση Αναπαραγωγιμότητας	43
Σύνοψη Κεφαλαίου	43
ΚΕΦΑΛΑΙΟ 4: AIR QUALITY - ΑΝΑΛΥΣΗ ΚΑΙ ΜΟΝΤΕΛΟΠΟΙΗΣΗ	45
Εισαγωγή στο Κεφάλαιο	45
4.1 Διερευνητική Ανάλυση Δεδομένων και Προεπεξεργασία (EDA & Preprocessing)	46
4.1.1 Επισκόπηση του Συνόλου Δεδομένων	46
4.1.2 Ανάλυση Ελλειπουσών Τιμών	47
4.1.3 Ανάλυση Κατανομών και Συσχετίσεων	48
4.1.4 Ανάλυση Χρονοσειρών	51
4.1.5 Ανίχνευση και Αντιμετώπιση Ακραίων Τιμών	52
4.1.6 Κανονικοποίηση	53
4.1.7 Χρονικός Διαχωρισμός Δεδομένων	54
4.2 Σχεδίαση Χαρακτηριστικών (Feature Engineering)	56
4.2.1 Χαρακτηριστικά Χρονικής Υστέρησης (Lag Features)	57
4.2.2 Χαρακτηριστικά Κυλόμενου Παραθύρου (Rolling Window Features)	57
4.2.3 Κυκλική Κωδικοποίηση Χρόνου (Cyclical Time Encoding)	58

4.2.4 Χαρακτηριστικά Αλληλεπίδρασης (Interaction Features).....	59
4.2.5 Τελικό Σύνολο Χαρακτηριστικών	60
4.3 Μοντελοποίηση (Modeling)	62
4.3.1 Μοντέλα Βάσης (Baseline Models).....	62
4.3.2 Προηγμένα Μοντέλα Ενισχυμένης Μάθησης (Advanced ML Models)	63
4.3.3 Μοντέλα Βαθιάς Μηχανικής Μάθησης (Deep Learning)	64
4.3.4 Ανάλυση Υπο-απόδοσης Μοντέλων Βαθιάς Μάθησης.....	66
4.3.5 Σύγκριση Μοντέλων	66
4.3.6 Ανάλυση Σημαντικότητας Χαρακτηριστικών	68
4.4 Αξιολόγηση (Evaluation).....	68
4.4.1 Διασταυρούμενη Επικύρωση Χρονοσειρών (Temporal Cross-Validation)	69
4.4.2 Καμπύλες Μάθησης (Learning Curves)	70
4.4.3 Ανάλυση Καταλοίπων (Residual Analysis).....	71
4.5 Ανίχνευση Ανωμαλιών (Anomaly Detection)	72
4.5.1 Στατιστικές Μέθοδοι.....	73
4.5.2 Μέθοδοι Μηχανικής Μάθησης.....	74
4.5.3 Μέθοδος Βασισμένη σε Κατάλοιπα Μοντέλου.....	75
4.5.4 Μέθοδος Αποσύνθεσης Χρονοσειρών (STL Decomposition)	75
4.5.5 Μέθοδος Συναθροιστικής Ψηφοφορίας (Ensemble)	76
4.5.6 Σύγκριση Μεθόδων Ανίχνευσης Ανωμαλιών.....	76
4.5.7 Συστάσεις για Πρακτική Εφαρμογή	78
Σύνοψη Κεφαλαίου.....	79
ΚΕΦΑΛΑΙΟ 5: POWER CONSUMPTION - ΑΝΑΛΥΣΗ ΚΑΙ ΜΟΝΤΕΛΟΠΟΙΗΣΗ	80
Εισαγωγή στο Κεφάλαιο.....	80
5.1 Διερευνητική Ανάλυση Δεδομένων και Προεπεξεργασία (EDA & Preprocessing)	81
5.1.1 Επισκόπηση του Συνόλου Δεδομένων.....	81
5.1.2 Ανάλυση Ποιότητας Δεδομένων	82

5.1.3 Στατιστική Περιγραφή	84
5.1.4 Ανάλυση Συσχετίσεων.....	86
5.1.5 Χρονικά Μοτίβα	87
5.1.6 Ανίχνευση και Χειρισμός Ακραίων Τιμών.....	88
5.1.7 Επαναδειγματοληψία σε Ωριαία Δεδομένα (Resampling)	89
5.1.8 Στρατηγική Χειρισμού Ελλειπουσών Τιμών	90
5.1.9 Αντιμετώπιση Ακραίων Τιμών	91
5.1.10 Κανονικοποίηση Χαρακτηριστικών	92
5.1.11 Χρονικός Διαχωρισμός (Temporal Split)	93
5.1.12 Σχεδίαση Χαρακτηριστικών (Feature Engineering)	94
5.1.13 Τελικό Σύνολο Χαρακτηριστικών	95
5.2 Μοντελοποίηση (Modeling)	96
5.2.1 Υπολογιστικοί Περιορισμοί και Επιλογή Μοντέλων	96
5.2.2 Μοντέλα που Υλοποιήθηκαν	97
5.2.3 Αποτελέσματα Μοντελοποίησης.....	97
5.2.4 Σύγκριση με Σύνολο Ποιότητας Αέρα.....	99
5.2.5 Ανάλυση Σημαντικότητας Χαρακτηριστικών	100
5.2.6 Αξιολόγηση με Διασταυρούμενη Επικύρωση	101
5.2.7 Ανάλυση Καταλοίπων	102
5.3 Ανίχνευση Ανωμαλιών (Anomaly Detection)	103
5.3.1 Μέθοδοι Ανίχνευσης Ανωμαλιών.....	104
5.3.2 Αποτελέσματα Ανίχνευσης Ανωμαλιών.....	105
5.3.3 Ανάλυση Αποτελεσμάτων	105
5.3.4 Σύγκριση με Σύνολο Ποιότητας Αέρα.....	106
5.3.5 Χρονική Κατανομή Ανωμαλιών	107
5.3.6 Συστάσεις για Λειτουργική Χρήση:	107
Σύνοψη Κεφαλαίου.....	108

ΚΕΦΑΛΑΙΟ 6: ΣΥΓΚΡΙΤΙΚΗ ΑΝΑΛΥΣΗ	109
Εισαγωγή στο Κεφάλαιο.....	109
6.1 Απόδοση Μοντέλων (Model Performance).....	110
6.1.1 Σύνοψη Βέλτιστων Μοντέλων ανά Σύνολο Δεδομένων	110
6.1.2 Συγκριτικός Πίνακας Μετρικών Απόδοσης.....	110
6.1.3 Υπεροχή Μοντέλων Βασισμένων σε Δέντρα Αποφάσεων.....	111
6.1.4 Ανάλυση Διαφορών Απόδοσης μεταξύ Συνόλων Δεδομένων	112
6.1.5 Υποαπόδοση Μοντέλων Βαθιάς Μάθησης	113
6.2 Σχεδίαση Χαρακτηριστικών (Feature Engineering)	114
6.2.1 Μοτίβα Ανεξάρτητα από τον Τομέα Εφαρμογής (Domain-Agnostic Patterns).....	114
6.2.2 Χαρακτηριστικά Εξαρτώμενα από τον Τομέα Εφαρμογής (Dataset-Specific Features).....	115
6.2.3 Σύνοψη Στρατηγικών Μηχανικής Χαρακτηριστικών.....	116
6.3 Ανίχνευση Ανωμαλιών (Anomaly Detection)	117
6.3.1 Αποτελεσματικότητα Μεθόδων σε Διαφορετικούς Τομείς.....	117
6.3.2 Ανάλυση Συμπεριφοράς Μεθόδων	118
6.3.3 Συνέπεια της Αποδόμησης STL.....	119
6.3.4 Συστάσεις για την Ανίχνευση Ανωμαλιών σε IoT.....	120
6.4 Διδάγματα (Lessons Learned).....	120
6.4.1 Επίδραση του Μεγέθους Δεδομένων στη Βαθιά Μάθηση	120
6.4.2 Καθολικές Βέλτιστες Πρακτικές για Πρόβλεψη IoT.....	121
6.4.3 Σύνοψη Διδαγμάτων	122
6.4.4 Περιορισμοί της Συγκριτικής Ανάλυσης	123
Σύνοψη Κεφαλαίου	123
ΚΕΦΑΛΑΙΟ 7: ΣΥΜΠΕΡΑΣΜΑΤΑ.....	125
Εισαγωγή στο Κεφάλαιο.....	125
7.1 Κύρια Ευρήματα (Key Findings).....	125

7.1.1 Υπεροχή Μοντέλων Βασισμένων σε Δέντρα Αποφάσεων.....	125
7.1.2 Σημασία της Μηχανικής Χαρακτηριστικών έναντι Πολύπλοκων Μοντέλων.....	126
7.1.3 Αποτελεσματικότητα Ανίχνευσης Ανωμαλιών με Σύνολα Μεθόδων	126
7.2 Συνεισφορά (Contributions)	127
7.2.1 Επικύρωση Διατομεακής Εφαρμοσιμότητας (Cross-Domain Validation).....	127
7.2.2 Αναπαραγώγιμη Μεθοδολογική Αγωγός (Reproducible Pipeline)	127
7.2.3 Τεκμηρίωση Βέλτιστων Πρακτικών (Best Practices Documentation)	127
7.3 Περιορισμοί (Limitations)	128
7.3.1 Επίδραση του Μεγέθους Δεδομένων στη Βαθιά Μάθηση	128
7.3.2 Περιορισμένος Χρονικός Ορίζοντας Πρόβλεψης.....	128
7.3.3 Απουσία Ανάπτυξης σε Πραγματικό Περιβάλλον	128
7.4 Μελλοντική Εργασία (Future Work)	129
7.4.1 Πρόβλεψη Πολλαπλών Χρονικών Οριζόντων (Multi-Horizon Forecasting).....	129
7.4.2 Ποσοτικοποίηση Αβεβαιότητας (Uncertainty Quantification)	129
7.4.3 Ανάπτυξη σε Παραγωγικό Περιβάλλον (Production Deployment).....	130
ΒΙΒΛΙΟΓΡΑΦΙΑ	132
ΠΑΡΑΡΤΗΜΑΤΑ.....	135
ΠΑΡΑΡΤΗΜΑ Α: Πίνακας Ορολογίας (Greek-English Terminology).....	135
ΠΑΡΑΡΤΗΜΑ Β: Αποθετήριο του Κώδικα στο GitHub.....	139
ΠΑΡΑΡΤΗΜΑ Γ: Αναλυτικοί Πίνακες Αποτελεσμάτων	141

ΚΑΤΑΛΟΓΟΣ ΕΙΚΟΝΩΝ

Εικόνα 2.1: IoT Architecture τεσσάρων επιπέδων	24
Εικόνα 2.2: CRISP-DM Methodology.....	31
Εικόνα 4.1: Χρονική κατανομή δεδομένων ανά μήνα και ώρα ημέρας	47
Εικόνα 4.2: Οπτικοποίηση ελλειπουσών τιμών ανά χαρακτηριστικό	48
Εικόνα 4.3: Κατανομές χαρακτηριστικών (Feature Distributions)	49
Εικόνα 4.4: Πίνακας συσχετίσεων (Correlation Heatmap)	50
Εικόνα 4.5: Χρονοσειρές ατμοσφαιρικών ρύπων.....	51
Εικόνα 4.6: Διαγράμματα θηκογραμμάτων για ανίχνευση ακραίων τιμών.....	52
Εικόνα 4.7: Αντιμετώπιση ακραίων τιμών πριν και μετά.....	54
Εικόνα 4.9: Κανονικοποίηση δεδομένων	54
Εικόνα 4.10: Χρονικός διαχωρισμός δεδομένων σε train/validation/test.....	55
Εικόνα 4.11: Σύνοψη προεπεξεργασίας.....	56
Εικόνα 4.12: Οπτικοποίηση χαρακτηριστικών κυλιόμενου παραθύρου	58
Εικόνα 4.13: Κυκλική κωδικοποίηση χρονικών χαρακτηριστικών.....	59
Εικόνα 4.14: Οπτικοποίηση χαρακτηριστικών αλληλεπίδρασης	60
Εικόνα 4.15: Ανάλυση τελικών χαρακτηριστικών ανά κατηγορία	61
Εικόνα 4.16: Σύγκριση απόδοσης μοντέλων	66
Εικόνα 4.17: Πραγματικές vs προβλεπόμενες τιμές για τα μοντέλα	67
Εικόνα 4.18: Σημαντικότητα χαρακτηριστικών (Feature Importance)	68
Εικόνα 4.19: Διασταυρούμενη επικύρωση χρονοσειρών με 5 τμήματα.....	69
Εικόνα 4.20: Καμπύλες μάθησης για Random Forest	70
Εικόνα 4.21: Διαγνωστικά καταλοίπων (Residual Diagnostics)	72
Εικόνα 4.22: Σύγκριση μεθόδων ανίχνευσης ανωμαλιών	77
Εικόνα 4.23: Χρονική γραμμή ανιχνευμένων ανωμαλιών	78
Εικόνα 5.1: Χρονική κατανομή δεδομένων.....	82

Εικόνα 5.2: Οπτικοποίηση ελλειπουσών τιμών.....	83
Εικόνα 5.3: Κατανομές χαρακτηριστικών	85
Εικόνα 5.4: Χάρτης θερμότητας συσχετίσεων	86
Εικόνα 5.5: Χρονοσειρά κατανάλωσης ενέργειας.....	88
Εικόνα 5.6: Θηκογράμματα ανίχνευσης ακραίων τιμών	89
Εικόνα 5.7: Σύγκριση στρατηγικών χειρισμού ελλειπουσών τιμών.....	91
Εικόνα 5.8: Κατανομές πριν και μετά την κανονικοποίηση.....	92
Εικόνα 5.9: Χρονικός διαχωρισμός δεδομένων.....	94
Εικόνα 5.10: Σύγκριση απόδοσης μοντέλων	98
Εικόνα 5.11: Πραγματικές vs Προβλεπόμενες τιμές.....	100
Εικόνα 5.12: Σημαντικότητα χαρακτηριστικών	101
Εικόνα 5.13: Αποτελέσματα διασταυρούμενης επικύρωσης.....	102
Εικόνα 5.14: Διαγνωστικά καταλοίπων.....	103
Εικόνα 5.15: Σύγκριση μεθόδων ανίχνευσης ανωμαλιών	105
Εικόνα 5.16: Χρονική κατανομή ανωμαλιών	107
Εικόνα 6.1: Συγκριτική απόδοση μοντέλων μεταξύ των δύο συνόλων δεδομένων	113
Εικόνα 6.2: Σύγκριση μεθόδων ανίχνευσης ανωμαλιών μεταξύ συνόλων δεδομένων	119

ΚΑΤΑΛΟΓΟΣ ΠΙΝΑΚΩΝ

Πίνακας 3.1: Συγκριτικά Χαρακτηριστικά Συνόλων Δεδομένων	37
Πίνακας 3.2: Εξαρτήσεις (Dependencies) Λογισμικού	42
Πίνακας 4.1: Σύνολα Δεδομένων έπειτα από Διαχωρισμό.....	55
Πίνακας 4.2: Σχεδίαση Χαρακτηριστικών.....	60
Πίνακας 4.3: Σύνοψη απόδοσης μοντέλων - Air Quality Dataset	66
Πίνακας 4.4: Σύγκριση μεθόδων ανίχνευσης ανωμαλιών - Air Quality	76
Πίνακας 5.1: Χαρακτηριστικά του συνόλου δεδομένων Power Consumption	81
Πίνακας 5.2: Ποσοστό απώλειας δεδομένων με Ανάλυση ελλειπουσών τιμών	83
Πίνακας 5.3: Λεπτομέρειες Χαρακτηριστικών μετά από Στατιστική Ανάλυση	84
Πίνακας 5.4: Ανίχνευση Ακραίων Τιμών	88
Πίνακας 5.5: Σύνολα μετά τον Διαχωρισμό Δεδομένων	93
Πίνακας 5.6: Σύγκριση απόδοσης μοντέλων - Power Consumption.....	98
Πίνακας 5.7: Σύγκριση απόδοσης μεταξύ Air Quality & Power Datasets	99
Πίνακας 5.8: Αποτελέσματα αξιολόγησης με διασταυρούμενη επικύρωση	101
Πίνακας 5.9: Διαγνωστικά Καταλοίπων.....	102
Πίνακας 5.10: Σύγκριση μεθόδων ανίχνευσης ανωμαλιών - Power Consumption.....	105
Πίνακας 5.11: Σύγκριση των ποσοστών ανωμαλιών μεταξύ Air Quality & Power.....	106
Πίνακας 6.1: Συγκριτική απόδοση μοντέλων μεταξύ των δύο συνόλων δεδομένων	110
Πίνακας 6.2: Σύγκριση στρατηγικών μηχανικής χαρακτηριστικών	116
Πίνακας 6.3: Σύγκριση μεθόδων ανίχνευσης ανωμαλιών μεταξύ συνόλων δεδομένων	117
Πίνακας 6.4: Σύνοψη διδαγμάτων από τη συγκριτική ανάλυση	122

ΚΕΦΑΛΑΙΟ 1: ΕΙΣΑΓΩΓΗ

1.1 Υπόβαθρο και Κίνητρο

Η τέταρτη βιομηχανική επανάσταση χαρακτηρίζεται από την εκρηκτική ανάπτυξη του Διαδικτύου των Πραγμάτων (Internet of Things, IoT), ενός τεχνολογικού παραδείγματος που μεταμορφώνει ριζικά τον τρόπο με τον οποίο οι φυσικές οντότητες αλληλεπιδρούν με τον ψηφιακό κόσμο [23]. Το IoT αναφέρεται σε ένα δίκτυο φυσικών αντικειμένων ενσωματωμένων με αισθητήρες, λογισμικό και άλλες τεχνολογίες, τα οποία συνδέονται και ανταλλάσσουν δεδομένα μέσω του διαδικτύου [24]. Στο σημείο αυτό, σύμφωνα με τους Atzori και συνεργάτες [23], το IoT αντιπροσωπεύει μια θεμελιώδη αλλαγή παραδείγματος στον τρόπο που αντιλαμβανόμαστε και αλληλεπιδρούμε με τον φυσικό κόσμο, καθιστώντας, έτσι, δυνατή την αυτόματη συλλογή, επεξεργασία και ανάλυση δεδομένων από το περιβάλλον σε πρωτοφανή κλίμακα.

Όσον αφορά τις εφαρμογές του IoT, εκτείνονται σε ένα ευρύ φάσμα τομέων, συμπεριλαμβανομένων των έξυπνων πόλεων, της βιομηχανικής αυτοματοποίησης, της περιβαλλοντικής παρακολούθησης, της υγειονομικής περίθαλψης και της διαχείρισης ενέργειας [24]. Οι Gubbi και συνεργάτες [24] επισημαίνουν ότι η επιτυχής αξιοποίηση του IoT εξαρτάται κρίσιμα από την ικανότητα επεξεργασίας και ερμηνείας των τεράστιων όγκων δεδομένων που παράγονται από τους διασυνδεδεμένους αισθητήρες. Ειδικότερα, τα δεδομένα IoT παρουσιάζουν ιδιαίτερα χαρακτηριστικά, καθώς είναι εγγενώς χρονικά εξαρτώμενα (χρονοσειρές), συχνά περιέχουν θόρυβο και ελλείπουσες τιμές, και μπορεί να εμφανίζουν σύνθετα εποχικά πρότυπα και τάσεις. Για τον λόγο αυτό απαιτούν εξειδικευμένη αντιμετώπιση.

Η Μηχανική Μάθηση (Machine Learning, ML) αποτελεί ένα πεδίο της τεχνητής νοημοσύνης που παρέχει τα εργαλεία και τις τεχνικές για την εξαγωγή χρήσιμης γνώσης από μεγάλους όγκους δεδομένων [4], [5]. Σύμφωνα με τους Witten και συνεργάτες [4], η μηχανική μάθηση επιτρέπει στους υπολογιστές να μαθαίνουν από εμπειρικά δεδομένα, χωρίς να έχουν ρητώς προγραμματιστεί για κάθε συγκεκριμένη εργασία. Παράλληλα, οι Hastie και συνεργάτες [5] τονίζουν ότι η στατιστική μάθηση παρέχει ένα αυστηρό θεωρητικό πλαίσιο για την κατανόηση και την εφαρμογή αλγορίθμων μηχανικής μάθησης σε πρακτικά προβλήματα.

Στο πλαίσιο του IoT, δύο είναι οι κρίσιμες εργασίες του, που αναδύονται ως ιδιαίτερα σημαντικές: η πρόβλεψη χρονοσειρών και η ανίχνευση ανωμαλιών. Η πρόβλεψη χρονοσειρών, αφορά την εκτίμηση μελλοντικών τιμών βάσει ιστορικών παρατηρήσεων, επιτρέποντας τη λήψη προληπτικών αποφάσεων και τον προγραμματισμό πόρων. Από την άλλη, η ανίχνευση ανωμαλιών αποσκοπεί στην αναγνώριση ασυνήθιστων ή μη αναμενόμενων προτύπων στα δεδομένα, τα οποία μπορεί να υποδηλώνουν βλάβες εξοπλισμού, κυβερνοεπιθέσεις ή κρίσιμες αλλαγές στις συνθήκες λειτουργίας [18].

Το κίνητρο της παρούσας διπλωματικής εργασίας πηγάζει από την ανάγκη ανάπτυξης ενός ολοκληρωμένου και αναπαραγωγίμου πλαισίου εργασίας για την εφαρμογή τεχνικών μηχανικής μάθησης σε δεδομένα IoT. Ενώ η βιβλιογραφία είναι πλούσια σε μεμονωμένες εφαρμογές, φαίνεται να υπάρχει σημαντικό κενό στη συστηματική σύγκριση μεθόδων σε διαφορετικούς τομείς εφαρμογής και στην παροχή πρακτικών κατευθυντήριων γραμμών για επαγγελματίες του χώρου. Η εργασία επιχειρεί να καλύψει αυτό το κενό, εφαρμόζοντας μια ενιαία μεθοδολογία σε δύο διακριτά σύνολα δεδομένων από διαφορετικούς τομείς: την παρακολούθηση περιβαλλοντικής ποιότητας αέρα και την κατανάλωση οικιακής ηλεκτρικής ενέργειας.

1.2 Στόχοι της Διπλωματικής

Πρωταρχικός σκοπός της διπλωματικής εργασίας είναι η διερεύνηση και αξιολόγηση τεχνικών μηχανικής μάθησης για την πρόβλεψη χρονοσειρών και την ανίχνευση ανωμαλιών σε δεδομένα IoT. Για την επίτευξη αυτού του σκοπού, καθορίζονται οι ακόλουθοι επιμέρους στόχοι:

Ο πρώτος στόχος, αφορά την ανάπτυξη ενός ολοκληρωμένου αγωγού επεξεργασίας δεδομένων (data processing pipeline) που ακολουθεί τη μεθοδολογία CRISP-DM [25]. Ο συγκεκριμένος αγωγός περιλαμβάνει όλα τα στάδια από την αρχική κατανόηση των δεδομένων έως την τελική αξιολόγηση των μοντέλων, διασφαλίζοντας την αναπαραγωγικότητα και τη μεταφερσιμότητα των αποτελεσμάτων.

Ο δεύτερος στόχος, εστιάζει στην υλοποίηση και συστηματική σύγκριση πολλαπλών αλγορίθμων μηχανικής μάθησης για πρόβλεψη χρονοσειρών μίας ώρας εμπρός. Οι αλγόριθμοι που εξετάζονται περιλαμβάνουν τόσο παραδοσιακές μεθόδους (Γραμμική Παλινδρόμηση, Random Forest [7],

XGBoost [8], Gradient Boosting [9]) όσο και αρχιτεκτονικές βαθιάς μάθησης (LSTM [11], GRU [12]).

Ο τρίτος στόχος, περιλαμβάνει την εφαρμογή και αξιολόγηση εννέα διαφορετικών μεθόδων ανίχνευσης ανωμαλιών, που καλύπτουν το φάσμα από απλές στατιστικές τεχνικές έως εξειδικευμένους αλγορίθμους μηχανικής μάθησης, χωρίς επίβλεψη και μεθόδους ανάλυσης χρονοσειρών.

Ο τέταρτος στόχος, σχετίζεται με τη διεξαγωγή συγκριτικής ανάλυσης μεταξύ δύο διαφορετικών τομέων εφαρμογής IoT: την περιβαλλοντική παρακολούθηση (ποιότητα αέρα) και τη διαχείριση ενέργειας (οικιακή κατανάλωση). Αυτή η διατομεακή σύγκριση επιτρέπει την αξιολόγηση της γενικευσιμότητας των ευρημάτων και την εξαγωγή καθολικά εφαρμόσιμων βέλτιστων πρακτικών.

Ο πέμπτος στόχος, αφορά τη διάθεση του συνόλου του κώδικα και της τεκμηρίωσης ως ανοικτό λογισμικό, επιτρέποντας την αναπαραγωγή των αποτελεσμάτων και την επέκταση της εργασίας από άλλους ερευνητές.

1.3 Δομή της Εργασίας

Η παρούσα διπλωματική εργασία οργανώνεται σε επτά κεφάλαια, τα οποία ακολουθούν τη λογική ροή της μεθοδολογίας CRISP-DM [25], από την κατανόηση του προβλήματος έως την αξιολόγηση και τα συμπεράσματα.

Το Κεφάλαιο 1 (παρόν κεφάλαιο), αποτελεί την εισαγωγή της εργασίας, παρουσιάζοντας το υπόβαθρο και το κίνητρο της έρευνας, τους στόχους της διπλωματικής και τη δομή του κειμένου.

Το Κεφάλαιο 2, παρέχει μια εκτενή ανασκόπηση της σχετικής βιβλιογραφίας. Καλύπτει τα θεμελιώδη στοιχεία του IoT και της ανάλυσης χρονοσειρών, τους βασικούς αλγορίθμους μηχανικής μάθησης που χρησιμοποιούνται στην εργασία, τις τεχνικές ανίχνευσης ανωμαλιών και τη μεθοδολογία CRISP-DM.

Το Κεφάλαιο 3, περιγράφει τη μεθοδολογία και τα δεδομένα που χρησιμοποιούνται. Παρουσιάζονται αναλυτικά τα δύο σύνολα δεδομένων (UCI Air Quality και Household Power Consumption), το περιβάλλον ανάπτυξης και τα εργαλεία λογισμικού, καθώς και η γενική αρχιτεκτονική του συστήματος.

Το Κεφάλαιο 4, επικεντρώνεται στην ανάλυση και μοντελοποίηση του συνόλου δεδομένων ποιότητας αέρα. Περιλαμβάνει τη διερευνητική ανάλυση δεδομένων (EDA), την προεπεξεργασία, την σχεδίαση χαρακτηριστικών, την εκπαίδευση μοντέλων, την αξιολόγηση και την ανίχνευση ανωμαλιών.

Το Κεφάλαιο 5, εφαρμόζει την ίδια μεθοδολογία στο σύνολο δεδομένων κατανάλωσης ηλεκτρικής ενέργειας, επισημαίνοντας τις διαφορές που προκύπτουν από τον μεγαλύτερο όγκο δεδομένων και τα ιδιαίτερα χαρακτηριστικά του τομέα.

Το Κεφάλαιο 6, πραγματοποιεί συγκριτική ανάλυση μεταξύ των δύο συνόλων δεδομένων, εξάγοντας συμπεράσματα για τη γενικευσιμότητα των μεθόδων και τις βέλτιστες πρακτικές για εφαρμογές IoT.

Το Κεφάλαιο 7, παρουσιάζει τα συμπεράσματα της εργασίας, τη συνεισφορά, τους περιορισμούς, αλλά και τις κατευθύνσεις μελλοντικής έρευνας.

Τέλος, η εργασία περιλαμβάνει παραρτήματα με τον πίνακα ορολογίας (ελληνικά-αγγλικά), αναφορά στο αποθετήριο κώδικα GitHub και λεπτομερείς πίνακες αποτελεσμάτων. Οι λεπτομερείς πίνακες αποτελεσμάτων περιλαμβάνουν σημαντικές μετρήσεις που πραγματοποιήθηκαν τόσο στο στάδιο της ανάλυσης των συνόλων δεδομένων, όσο και σημαντικά στοιχεία και αποτελέσματα που προέκυψαν από τις διαδικασίες της σχεδίασης χαρακτηριστικών και της μοντελοποίησης.

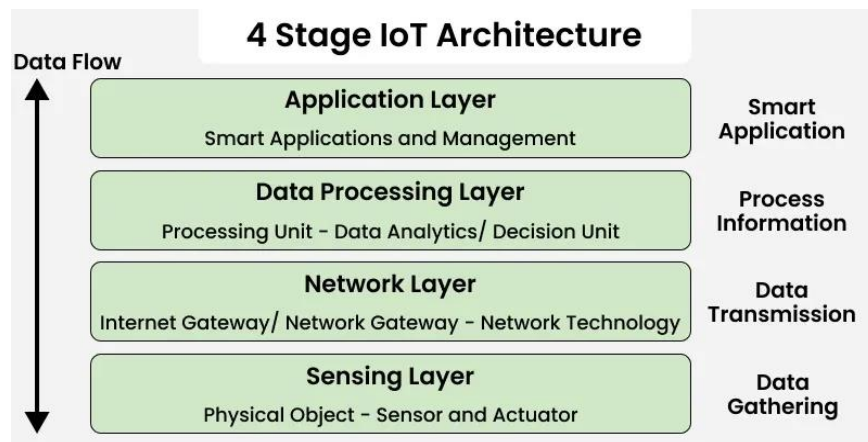
ΚΕΦΑΛΑΙΟ 2: ΑΝΑΣΚΟΠΗΣΗ ΒΙΒΛΙΟΓΡΑΦΙΑΣ

2.1 IoT και Χρονοσειρές

2.1.1 Εισαγωγή στο Διαδίκτυο των Πραγμάτων

Το Διαδίκτυο των Πραγμάτων (Internet of Things, IoT) αντιπροσωπεύει ένα υπολογιστικό παράδειγμα, όπου καθημερινά αντικείμενα ενσωματώνουν υπολογιστική και επικοινωνιακή ικανότητα, επιτρέποντάς τους να συλλέγουν, να επεξεργάζονται και να ανταλλάσσουν δεδομένα μέσω του διαδικτύου [23]. Οι Atzori και συνεργάτες [23] παρέχουν μια θεμελιώδη επισκόπηση του IoT, αναλύοντας τις τεχνολογίες που το καθιστούν δυνατό, τα πρωτόκολλα επικοινωνίας και τις αρχιτεκτονικές αναφορές. Σύμφωνα με τους ερευνητές, το IoT διαφοροποιείται από το παραδοσιακό διαδίκτυο στο ότι δεν περιορίζεται στη διασύνδεση υπολογιστών, αλλά επεκτείνεται σε οποιοδήποτε φυσικό αντικείμενο μπορεί να εξοπλιστεί με αισθητήρες και δυνατότητα επικοινωνίας.

Ακόμη, οι Gubbi και συνεργάτες [24] παρουσιάζουν ένα ολοκληρωμένο όραμα για το IoT, περιγράφοντας τα αρχιτεκτονικά στοιχεία και τις μελλοντικές κατευθύνσεις του πεδίου. Οι παραπάνω τονίζουν ότι η πραγματική αξία του IoT δεν έγκειται απλώς στη συλλογή δεδομένων, αλλά στην εξαγωγή χρήσιμης γνώσης από αυτά. Η ανάλυση δεδομένων IoT απαιτεί την αντιμετώπιση πολλαπλών προκλήσεων, συμπεριλαμβανομένης της ετερογένειας των πηγών δεδομένων, του μεγάλου όγκου και ρυθμού παραγωγής, της ανάγκης για επεξεργασία σε πραγματικό χρόνο και της διαχείρισης ζητημάτων ποιότητας δεδομένων, όπως ο θόρυβος και οι ελλείπουσες τιμές.



Εικόνα 2.1: IoT Architecture τεσσάρων επιπέδων

Αντλήθηκε από την ιστοσελίδα Geeks for Geeks, από το άρθρο “Architecture of Internet of Things (IoT)” από τον παρακάτω σύνδεσμο:

<https://www.geeksforgeeks.org/computer-networks/architecture-of-internet-of-things-iot>

Η αρχιτεκτονική ενός τυπικού συστήματος IoT περιλαμβάνει τέσσερα βασικά επίπεδα. Αρχικά, το επίπεδο αίσθησης (sensing layer), το οποίο αποτελείται από τους φυσικούς αισθητήρες που συλλέγουν δεδομένα από το περιβάλλον, όπως θερμοκρασία, υγρασία, ατμοσφαιρικοί ρύποι ή ηλεκτρική κατανάλωση. Δεύτερο είναι το επίπεδο δικτύου (network layer), που διασφαλίζει τη μετάδοση των δεδομένων από τους αισθητήρες σε κεντρικά συστήματα επεξεργασίας, χρησιμοποιώντας διάφορα πρωτόκολλα επικοινωνίας. Τρίτο έχουμε το επίπεδο επεξεργασίας (processing layer), που με τη σειρά του αναλαμβάνει την αποθήκευση, επεξεργασία και ανάλυση των δεδομένων. Τέλος, υπάρχει το επίπεδο εφαρμογής (application layer), το οποίο παρέχει τις διεπαφές για την αξιοποίηση της εξαγόμενης γνώσης από τους τελικούς χρήστες [24].

Στο πλαίσιο της παρούσας εργασίας, η έμφαση δίνεται στο επίπεδο επεξεργασίας (processing layer), όπου εφαρμόζονται τεχνικές μηχανικής μάθησης για πρόβλεψη και ανίχνευση ανωμαλιών.

2.1.2 Χρονοσειρές και Ανάλυση Χρονικά Εξαρτώμενων Δεδομένων

Τα δεδομένα IoT είναι εγγενώς χρονικά εξαρτώμενα, καθώς αποτελούν ακολουθίες παρατηρήσεων που καταγράφονται σε διαδοχικές χρονικές στιγμές. Οι Box και συνεργάτες [15] αποτελούν τη θεμελιώδη αναφορά για την ανάλυση χρονοσειρών, παρέχοντας το θεωρητικό

υπόβαθρο για την κατανόηση και μοντελοποίηση χρονικά εξαρτώμενων δεδομένων. Στο πλαίσιο αυτό, μια χρονοσειρά μπορεί να αποσυντεθεί σε τρία βασικά συστατικά: την τάση (trend), η οποία αντιπροσωπεύει τη μακροπρόθεσμη κατεύθυνση των δεδομένων, την εποχικότητα (seasonality), η οποία αντικατοπτρίζει περιοδικές διακυμάνσεις σε σταθερά χρονικά διαστήματα, και το τυχαίο συστατικό (residual ή noise), το οποίο περιλαμβάνει την τυχαία μεταβλητότητα που δεν εξηγείται από τα άλλα συστατικά.

Η ανάλυση χρονοσειρών IoT παρουσιάζει ιδιαίτερες προκλήσεις. Οι Hyndman και Athanasopoulos [17] περιγράφουν τις αρχές και τις πρακτικές της πρόβλεψης χρονοσειρών, τονίζοντας τη σημασία της κατάλληλης προεπεξεργασίας και της επιλογής του κατάλληλου μοντέλου βάσει των χαρακτηριστικών των δεδομένων. Οι ερευνητές επισημαίνουν ότι η πρόβλεψη χρονοσειρών απαιτεί ιδιαίτερη προσοχή στη διατήρηση της χρονολογικής σειράς κατά τη διαίρεση των δεδομένων σε σύνολα εκπαίδευσης, επικύρωσης και ελέγχου, προκειμένου να αποφευχθεί η διαρροή πληροφορίας από το μέλλον (data leakage). Η μη τήρηση της χρονικής αλληλουχίας μπορεί να οδηγήσει σε υπερεκτίμηση της απόδοσης των μοντέλων και σε μη ρεαλιστικά αποτελέσματα.

Ταυτόχρονα, οι Bergmeir και Benítez [16] εξετάζουν συγκεκριμένα τη χρήση της διασταυρούμενης επικύρωσης (cross-validation) στην αξιολόγηση μοντέλων πρόβλεψης χρονοσειρών. Οι ίδιοι, καταδεικνύουν ότι οι παραδοσιακές τεχνικές διασταυρούμενης επικύρωσης που χρησιμοποιούνται για ανεξάρτητα δεδομένα δεν είναι κατάλληλες για χρονοσειρές, καθώς παραβιάζουν τη χρονολογική εξάρτηση των παρατηρήσεων. Αντ' αυτού, προτείνουν τη χρήση τεχνικών, όπως η Time Series Cross-Validation με κυλιόμενο ή επεκτεινόμενο παράθυρο εκπαίδευσης.

2.2 Μηχανική Μάθηση

2.2.1 Θεμελιώδεις Αρχές

Η Μηχανική Μάθηση αποτελεί έναν κλάδο της τεχνητής νοημοσύνης που ασχολείται με την ανάπτυξη αλγορίθμων ικανών να μαθαίνουν από δεδομένα και να βελτιώνουν την απόδοσή τους χωρίς να έχουν ρητώς προγραμματιστεί για κάθε συγκεκριμένη εργασία [4]. Οι Witten και

συνεργάτες [4] παρέχουν μια ολοκληρωμένη εισαγωγή στην εξόρυξη δεδομένων και τη μηχανική μάθηση, καλύπτοντας τόσο τις θεωρητικές αρχές όσο και τις πρακτικές εφαρμογές. Το έργο τους τονίζει τη σημασία της κατανόησης των δεδομένων πριν την εφαρμογή οποιουδήποτε αλγορίθμου και την ανάγκη για συστηματική αξιολόγηση των μοντέλων.

Οι Hastie και συνεργάτες [5] παρέχουν το θεωρητικό θεμέλιο της στατιστικής μάθησης, αναλύοντας τις μαθηματικές αρχές πίσω από τους αλγορίθμους μηχανικής μάθησης. Το έργο τους εξετάζει θέματα όπως η διάκριση μεταξύ πολυπλοκότητας μοντέλου και ικανότητας γενίκευσης, η αντιστάθμιση μεροληψίας-διακύμανσης (bias-variance tradeoff) και οι τεχνικές κανονικοποίησης (regularization). Στο πλαίσιο δεδομένων IoT, όπου η χρονική εξάρτηση, η παρουσία θορύβου και οι ελλείπουσες τιμές είναι συχνές, η κατανόηση της ισορροπίας μεροληψίας-διακύμανσης και της ικανότητας γενίκευσης είναι σημαντική για την αποφυγή υπερπροσαρμογής και τη διασφάλιση αξιόπιστων προβλέψεων.

Οι James και συνεργάτες [6] παρέχουν μια πιο προσβάσιμη εισαγωγή στη στατιστική μάθηση, με έμφαση στις πρακτικές εφαρμογές. Το έργο τους καλύπτει θέματα όπως η σχεδίαση χαρακτηριστικών (feature engineering), η επιλογή μοντέλου και η ερμηνευσιμότητα των αποτελεσμάτων. Επίσης, οι ερευνητές τονίζουν ότι η επιτυχία ενός έργου μηχανικής μάθησης εξαρτάται πρωτίστως από την ποιότητα των χαρακτηριστικών που χρησιμοποιούνται ως είσοδος στα μοντέλα.

2.2.2 Μοντέλα Βασισμένα σε Δέντρα Αποφάσεων

Τα μοντέλα βασισμένα σε δέντρα αποφάσεων αποτελούν μια από τις πλέον επιτυχημένες οικογένειες αλγορίθμων μηχανικής μάθησης, ιδιαίτερα για δεδομένα με πίνακα χαρακτηριστικών (tabular data). Ο αλγόριθμος Random Forest, που προτάθηκε από τον Breiman [7], αποτελεί μια μέθοδο ensemble που συνδυάζει τις προβλέψεις πολλαπλών δέντρων αποφάσεων. Πιο συγκεκριμένα, κάθε δέντρο εκπαιδεύεται σε ένα τυχαίο υποσύνολο των δεδομένων (bootstrap sampling) και χρησιμοποιεί ένα τυχαίο υποσύνολο χαρακτηριστικών σε κάθε διαχωρισμό κόμβου. Αυτή η τυχαιότητα μειώνει τη διακύμανση του τελικού μοντέλου και βελτιώνει την ικανότητα γενίκευσης [7].

Τα μοντέλα δέντρων είναι ιδιαίτερα κατάλληλα για δεδομένα IoT, καθώς χειρίζονται αποτελεσματικά μη γραμμικές σχέσεις, αλληλεπιδράσεις μεταξύ χαρακτηριστικών και ελλείπουσες τιμές, ενώ δεν απαιτούν ισχυρές στατιστικές υποθέσεις για τη μορφή των δεδομένων.

Ακόμη, ο αλγόριθμος XGBoost (eXtreme Gradient Boosting), που αναπτύχθηκε από τους Chen και Guestrin [8], αποτελεί μια εξελιγμένη υλοποίηση του gradient boosting που έχει επιτύχει κορυφαίες επιδόσεις σε πολυάριθμους διαγωνισμούς μηχανικής μάθησης. Ο XGBoost εισάγει καινοτομίες όπως η κανονικοποίηση δεύτερης τάξης, η αποτελεσματική διαχείριση ελλειπουσών τιμών και η παράλληλη επεξεργασία, καθιστώντας τον ιδιαίτερα κατάλληλο για μεγάλης κλίμακας εφαρμογές [8].

Επιπλέον, η μέθοδος Gradient Boosting, όπως περιγράφεται από τον Friedman [9], αποτελεί τη θεωρητική βάση πάνω στην οποία αναπτύχθηκε ο XGBoost. Η μέθοδος λειτουργεί κατασκευάζοντας διαδοχικά μοντέλα, όπου κάθε νέο μοντέλο εκπαιδεύεται να διορθώνει τα σφάλματα των προηγούμενων. Ο Friedman [9] έδειξε ότι αυτή η διαδικασία μπορεί να ερμηνευτεί ως αριθμητική βελτιστοποίηση μιας συνάρτησης απώλειας στον χώρο των συναρτήσεων.

Αξίζει να αναφερθεί και η βιβλιοθήκη Scikit-learn [10], η οποία παρέχει αποτελεσματικές υλοποιήσεις όλων αυτών των αλγορίθμων σε Python, καθώς και ένα πλούσιο σύνολο εργαλείων για προεπεξεργασία δεδομένων, αξιολόγηση μοντέλων και επιλογή υπερπαραμέτρων. Οι Pedregosa και συνεργάτες [10] σχεδίασαν τη βιβλιοθήκη με έμφαση στην ευχρηστία και τη συνέπεια του API, διευκολύνοντας την πειραματική σύγκριση διαφορετικών αλγορίθμων.

2.2.3 Βαθιά Μάθηση για Χρονοσειρές

Η βαθιά μάθηση (deep learning) αποτελεί ένα υπο-πεδίο της μηχανικής μάθησης, που χρησιμοποιεί νευρωνικά δίκτυα με πολλαπλά επίπεδα για την εκμάθηση ιεραρχικών αναπαραστάσεων των δεδομένων [13]. Οι Goodfellow και συνεργάτες [13] παρέχουν μια ολοκληρωμένη εισαγωγή στη βαθιά μάθηση, καλύπτοντας τόσο τα θεωρητικά θεμέλια, όσο και τις πρακτικές εφαρμογές.

Για δεδομένα χρονοσειρών, οι αρχιτεκτονικές επαναλαμβανόμενων νευρωνικών δικτύων (Recurrent Neural Networks, RNNs) είναι ιδιαίτερα κατάλληλες, καθώς είναι σχεδιασμένες να επεξεργάζονται ακολουθίες δεδομένων μεταβλητού μήκους. Η αρχιτεκτονική Long Short-Term Memory (LSTM), που προτάθηκε από τους Hochreiter και Schmidhuber [11], αντιμετωπίζει το πρόβλημα της εξασθένησης της κλίσης (vanishing gradient) που πλήττει τα παραδοσιακά RNNs, επιτρέποντας την εκμάθηση μακροπρόθεσμων εξαρτήσεων στα δεδομένα. Το LSTM εισάγει μηχανισμούς πύλης (gate mechanisms) που ελέγχουν τη ροή πληροφορίας μέσα στο δίκτυο, διατηρώντας ή απορρίπτοντας πληροφορία με βάση το πλαίσιο [11].

Η αρχιτεκτονική Gated Recurrent Unit (GRU), που προτάθηκε από τους Cho και συνεργάτες [12], αποτελεί μια απλοποιημένη εκδοχή του LSTM με λιγότερες πύλες και παραμέτρους. Οι ερευνητές έδειξαν ότι το GRU επιτυγχάνει συγκρίσιμες επιδόσεις με το LSTM σε πολλές εργασίες, ενώ είναι υπολογιστικά αποδοτικότερο λόγω της απλούστερης δομής του [12].

Επιπροσθέτως, ο Brownlee [14] παρέχει πρακτικές κατευθυντήριες γραμμές για την εφαρμογή βαθιάς μάθησης σε προβλήματα πρόβλεψης χρονοσειρών. Το έργο του καλύπτει θέματα όπως, η προετοιμασία δεδομένων για επαναλαμβανόμενα νευρωνικά δίκτυα, η επιλογή υπερπαραμέτρων και οι τεχνικές αποφυγής υπερπροσαρμογής (overfitting).

2.3 Ανίχνευση Ανωμαλιών

2.3.1 Επισκόπηση και Κατηγοριοποίηση

Η ανίχνευση ανωμαλιών αναφέρεται στο πρόβλημα της εύρεσης προτύπων στα δεδομένα που δεν συμμορφώνονται με την αναμενόμενη συμπεριφορά [18]. Οι Chandola και συνεργάτες [18] παρέχουν μια εκτενή επισκόπηση του πεδίου, κατηγοριοποιώντας τις τεχνικές ανίχνευσης ανωμαλιών και αναλύοντας τα πλεονεκτήματα και μειονεκτήματα κάθε προσέγγισης. Σύμφωνα με τους παραπάνω, οι ανωμαλίες μπορούν να ταξινομηθούν σε τρεις τύπους: σημειακές ανωμαλίες (point anomalies), όπου μια μεμονωμένη παρατήρηση είναι ανώμαλη σε σχέση με τα υπόλοιπα δεδομένα, συλλογικές ανωμαλίες (collective anomalies), όπου μια ομάδα συσχετισμένων παρατηρήσεων είναι ανώμαλη ακόμα και αν οι επιμέρους παρατηρήσεις δεν είναι, και

συμφραζόμενες ανωμαλίες (contextual anomalies), όπου μια παρατήρηση είναι ανώμαλη μόνο σε ένα συγκεκριμένο πλαίσιο [18].

2.3.2 Στατιστικές Μέθοδοι

Οι στατιστικές μέθοδοι ανίχνευσης ανωμαλιών βασίζονται στην υπόθεση ότι τα κανονικά δεδομένα ακολουθούν ένα συγκεκριμένο στατιστικό μοντέλο, και οι παρατηρήσεις που αποκλίνουν σημαντικά από αυτό το μοντέλο θεωρούνται ανώμαλες. Ενδεικτικά παρουσιάζονται οι ακόλουθες μέθοδοι:

Η μέθοδος Z-Score, υπολογίζει πόσες τυπικές αποκλίσεις απέχει κάθε παρατήρηση από τον μέσο όρο. Παρατηρήσεις με $|z| > 3$ θεωρούνται τυπικά ως ανώμαλες, υποθέτοντας κανονική κατανομή των δεδομένων [18].

Η μέθοδος IQR (Interquartile Range), χρησιμοποιεί τα τεταρτημόρια της κατανομής για τον εντοπισμό ακραίων τιμών. Παρατηρήσεις που βρίσκονται κάτω από το $Q1 - 1.5 \times IQR$ ή πάνω από το $Q3 + 1.5 \times IQR$ θεωρούνται ανώμαλες. Η μέθοδος αυτή, είναι πιο ανθεκτική σε μη κανονικές κατανομές συγκριτικά με το Z-Score [18].

Η μέθοδος Modified Z-Score, χρησιμοποιεί τη διάμεσο και την τυπική απόκλιση της διαμέσου (MAD - Median Absolute Deviation) αντί του μέσου όρου και της τυπικής απόκλισης, καθιστώντας τη μέθοδο πιο ανθεκτική σε ακραίες τιμές [18].

2.3.3 Μέθοδοι Μηχανικής Μάθησης

Ο αλγόριθμος Isolation Forest, που προτάθηκε από τους Liu και συνεργάτες [19], βασίζεται στην παρατήρηση ότι οι ανωμαλίες είναι λιγότερες και διαφορετικές από τα κανονικά δεδομένα, καθιστώντας τις ευκολότερο να απομονωθούν. Ο αλγόριθμος κατασκευάζει τυχαία δέντρα απομόνωσης και χρησιμοποιεί το μέσο μήκος διαδρομής για τον εντοπισμό παρατηρήσεων που απομονώνονται γρήγορα ως ανωμαλίες [19].

Ο αλγόριθμος Local Outlier Factor (LOF), που προτάθηκε από τους Breunig και συνεργάτες [20], μετρά την τοπική απόκλιση πυκνότητας μιας παρατήρησης σε σχέση με τους γείτονές της.

Παρατηρήσεις που βρίσκονται σε περιοχές χαμηλότερης πυκνότητας σε σχέση με τους γείτονές τους θεωρούνται ανώμαλες [20].

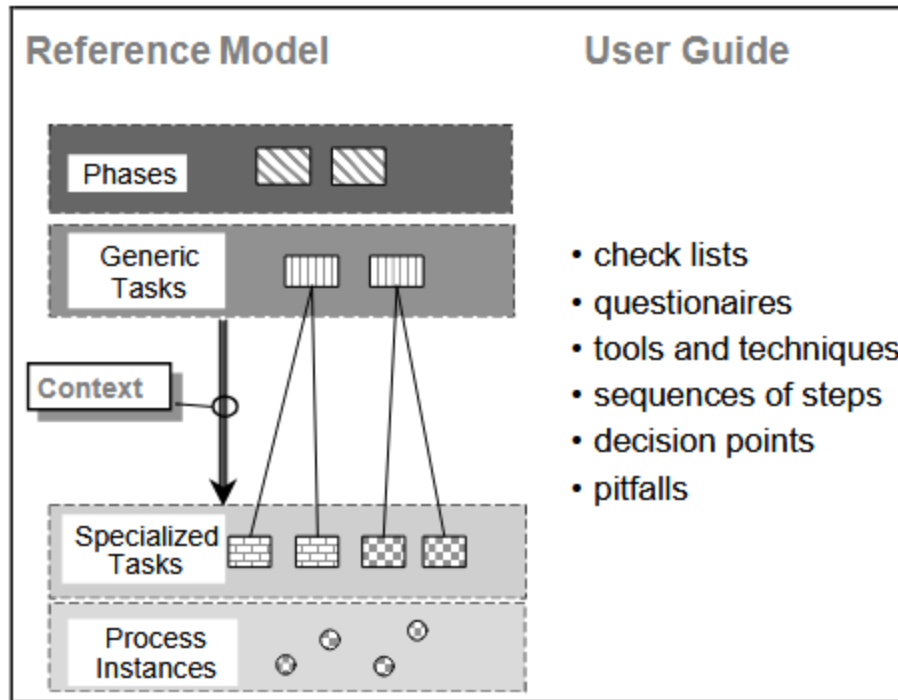
Ο αλγόριθμος One-Class SVM, που αναπτύχθηκε από τους Schölkopf και συνεργάτες [21], εκπαιδεύεται μόνο σε κανονικά δεδομένα και μαθαίνει ένα όριο που περικλείει αυτά τα δεδομένα. Νέες παρατηρήσεις που πέφτουν έξω από αυτό το όριο ταξινομούνται ως ανωμαλίες [21].

2.3.4 Μέθοδοι Χρονοσειρών

Η μέθοδος STL Decomposition (Seasonal and Trend decomposition using Loess), που προτάθηκε από τους Cleveland και συνεργάτες [22], αποσυνθέτει μια χρονοσειρά στα συστατικά της τάσης, εποχικότητας και καταλοίπων. Η ανίχνευση ανωμαλιών πραγματοποιείται εξετάζοντας τα κατάλοιπα: παρατηρήσεις με ασυνήθιστα μεγάλα κατάλοιπα θεωρούνται ανώμαλες. Η μέθοδος είναι ιδιαίτερα κατάλληλη για χρονοσειρές με έντονη εποχικότητα, καθώς διαχωρίζει την κανονική εποχιακή διακύμανση από πραγματικές ανωμαλίες [22].

2.4 Μεθοδολογία CRISP-DM

Η μεθοδολογία CRISP-DM (Cross-Industry Standard Process for Data Mining), που περιγράφεται από τους Wirth και Hipp [25], αποτελεί ένα ευρέως αποδεκτό πρότυπο για τη διεξαγωγή έργων εξόρυξης δεδομένων και μηχανικής μάθησης. Αναπτύχθηκε στα τέλη της δεκαετίας του 1990 από μια κοινοπραξία εταιρειών και έχει καθιερωθεί ως μία από τις πλέον διαδεδομένες προσεγγίσεις για έργα ανάλυσης δεδομένων σε βιομηχανικό περιβάλλον [25].



Εικόνα 2.2: CRISP-DM Methodology

Αντλήθηκε από τους Wirth και Hipp [25]: “Figure 1: Four Level Breakdown of the CRISP-DM Methodology for Data Mining”.

Η μεθοδολογία CRISP-DM οργανώνεται σε έξι διακριτές φάσεις που αλληλεπιδρούν κυκλικά:

- Η πρώτη φάση είναι η Κατανόηση της Επιχείρησης (Business Understanding), η οποία επικεντρώνεται στον καθορισμό των στόχων του έργου από επιχειρηματική σκοπιά, στην κατανόηση των απαιτήσεων και στη μετατροπή αυτής της γνώσης σε έναν ορισμό προβλήματος εξόρυξης δεδομένων [25].
- Η δεύτερη φάση είναι η Κατανόηση των Δεδομένων (Data Understanding), η οποία περιλαμβάνει την αρχική συλλογή δεδομένων, την εξοικείωση με τα δεδομένα, τον εντοπισμό προβλημάτων ποιότητας και την ανακάλυψη πρώτων ενδείξεων στα δεδομένα [25].
- Η τρίτη φάση είναι η Προετοιμασία Δεδομένων (Data Preparation), η οποία καλύπτει όλες τις δραστηριότητες που απαιτούνται για την κατασκευή του τελικού συνόλου δεδομένων από τα αρχικά ακατέργαστα δεδομένα. Περιλαμβάνει την επιλογή δεδομένων, τον καθαρισμό, την κατασκευή νέων χαρακτηριστικών και τον μετασχηματισμό [25].

- Η τέταρτη φάση είναι η Μοντελοποίηση (Modeling), η οποία περιλαμβάνει την επιλογή και εφαρμογή διαφόρων τεχνικών μοντελοποίησης, τη βαθμονόμηση των παραμέτρων τους και την αξιολόγηση της απόδοσής τους [25].
- Η πέμπτη φάση είναι η Αξιολόγηση (Evaluation), η οποία περιλαμβάνει την αξιολόγηση των μοντέλων σε σχέση με τους επιχειρηματικούς στόχους, την επανεξέταση της διαδικασίας και τον καθορισμό των επόμενων βημάτων [25].
- Η έκτη φάση είναι η Ανάπτυξη (Deployment), η οποία περιλαμβάνει την οργάνωση και παρουσίαση της γνώσης που αποκτήθηκε με τρόπο που να μπορεί να χρησιμοποιηθεί από τον πελάτη [25].

Στο πλαίσιο της παρούσας εργασίας, οι φάσεις αυτές αντιστοιχίζονται σε συγκεκριμένες ενέργειες: η Κατανόηση της Επιχείρησης αντιστοιχεί στον ορισμό του προβλήματος πρόβλεψης και ανίχνευσης ανωμαλιών σε IoT δεδομένα, η Κατανόηση και Προετοιμασία Δεδομένων υλοποιούνται μέσω εκτενούς διερευνητικής ανάλυσης και σχεδίασης χαρακτηριστικών, ενώ η Μοντελοποίηση και Αξιολόγηση περιλαμβάνουν συγκριτική μελέτη πολλαπλών αλγορίθμων. Στην παρούσα εργασία, η φάση Deployment υλοποιείται μέσω της διάθεσης του κώδικα και της τεκμηρίωσης ως ανοικτό λογισμικό, ενισχύοντας την αναπαραγωγιμότητα.

Η μεθοδολογία CRISP-DM επιλέχθηκε για την παρούσα εργασία λόγω της δομημένης προσέγγισής της, της ευελιξίας της στην προσαρμογή σε διαφορετικούς τομείς εφαρμογής και της έμφασής της στην επαναληπτική φύση των έργων ανάλυσης δεδομένων [25]. Η καταλληλότητά της για περιβάλλοντα IoT είναι ιδιαίτερα σημαντική: τα δεδομένα IoT χαρακτηρίζονται από ετερογένεια πηγών, ποικίλη ποιότητα μετρήσεων και ανάγκη για επαναληπτική βελτιστοποίηση — χαρακτηριστικά που η κυκλική δομή της CRISP-DM αντιμετωπίζει εγγενώς, καθώς επιτρέπει την επιστροφή σε προηγούμενες φάσεις όταν, για παράδειγμα, η φάση μοντελοποίησης αποκαλύπτει ανεπάρκειες στην προεπεξεργασία ή η ανίχνευση ανωμαλιών υποδεικνύει ανάγκη επανεξέτασης της ποιότητας δεδομένων.

ΚΕΦΑΛΑΙΟ 3: ΜΕΘΟΔΟΛΟΓΙΑ ΚΑΙ ΔΕΔΟΜΕΝΑ

Εισαγωγή στο Κεφάλαιο

Το παρόν κεφάλαιο περιγράφει αναλυτικά τη μεθοδολογία που ακολουθείται στην παρούσα διπλωματική εργασία, καθώς και τα σύνολα δεδομένων που χρησιμοποιούνται για την πειραματική αξιολόγηση. Σύμφωνα με τη μεθοδολογία CRISP-DM (Cross-Industry Standard Process for Data Mining) [25], το κεφάλαιο αυτό αντιστοιχεί στις δύο πρώτες φάσεις της διαδικασίας: την Κατανόηση της Επιχείρησης (Business Understanding) και την Κατανόηση των Δεδομένων (Data Understanding). Η πρώτη φάση επικεντρώνεται στον καθορισμό των στόχων του έργου και τη μετατροπή της επιχειρηματικής γνώσης σε ένα σαφώς ορισμένο πρόβλημα εξόρυξης δεδομένων, ενώ η δεύτερη φάση περιλαμβάνει τη συλλογή και εξοικείωση με τα δεδομένα, τον εντοπισμό προβλημάτων ποιότητας και την ανακάλυψη αρχικών μοτίβων [25].

Η δομή του κεφαλαίου οργανώνεται ως εξής: αρχικά παρουσιάζονται αναλυτικά τα δύο σύνολα δεδομένων που χρησιμοποιούνται, με έμφαση στα χαρακτηριστικά τους, την προέλευσή τους και τις προκλήσεις που θέτουν. Ακολούθως, περιγράφονται τα εργαλεία λογισμικού και το τεχνικό περιβάλλον ανάπτυξης που επιλέχθηκαν για την υλοποίηση της εργασίας. Η παρουσίαση αυτή θέτει τα θεμέλια για τα επόμενα κεφάλαια, όπου εφαρμόζονται οι τεχνικές επεξεργασίας δεδομένων, μοντελοποίησης και αξιολόγησης.

Αναλυτικά αποτελέσματα παρατίθενται στο παράρτημα Γ (βλέπε σελ. 139). Όλοι οι πίνακες και τα δεδομένα τους μπορούν να βρεθούν τόσο στο παράρτημα Γ, όσο και στο αποθετήριο κώδικα του παραρτήματος Β (βλέπε σελ. 137 και 139).

3.1 Σύνολα Δεδομένων (Datasets)

Η επιλογή κατάλληλων συνόλων δεδομένων αποτελεί κρίσιμο παράγοντα για την επιτυχία οποιουδήποτε έργου μηχανικής μάθησης [4]. Για την παρούσα εργασία επιλέχθηκαν δύο σύνολα δεδομένων από διαφορετικούς τομείς εφαρμογής του Διαδικτύου των Πραγμάτων (IoT), τα οποία παρουσιάζουν τόσο κοινά χαρακτηριστικά, όσο και σημαντικές διαφορές. Η διπλή αυτή

προσέγγιση επιτρέπει την αξιολόγηση της γενικευσιμότητας των μεθόδων που αναπτύσσονται και την εξαγωγή συμπερασμάτων που υπερβαίνουν τα όρια ενός μεμονωμένου τομέα εφαρμογής [5].

Τα κριτήρια επιλογής των συνόλων δεδομένων περιελάμβαναν τη διαθεσιμότητα μέσω αναγνωρισμένων αποθετηρίων ερευνητικών δεδομένων, την παρουσία χρονοσειρών με ωριαία ή λεπτομερέστερη ανάλυση, την ύπαρξη προκλήσεων ποιότητας δεδομένων (ελλείπουσες τιμές, ακραίες τιμές) που αντικατοπτρίζουν πραγματικές συνθήκες IoT, και τη δυνατότητα σύγκρισης αποτελεσμάτων με προηγούμενες ερευνητικές εργασίες. Και τα δύο επιλεγμένα σύνολα δεδομένων προέρχονται από το UCI Machine Learning Repository, το οποίο αποτελεί ένα από τα πλέον αναγνωρισμένα αποθετήρια για την ερευνητική κοινότητα της μηχανικής μάθησης [2], [3].

3.1.1 Σύνολο Δεδομένων Ποιότητας Αέρα (UCI Air Quality Dataset)

Το πρώτο σύνολο δεδομένων που χρησιμοποιείται στην παρούσα εργασία είναι το UCI Air Quality Dataset, το οποίο δημοσιεύθηκε από τους De Vito και συνεργάτες [1] και διατίθεται μέσω του UCI Machine Learning Repository [2]. Το σύνολο δεδομένων αυτό συλλέχθηκε στο πλαίσιο ερευνητικού έργου για τη βαθμονόμηση ηλεκτρονικών αισθητήρων ποιότητας αέρα σε αστικό περιβάλλον στην Ιταλία.

Η συλλογή δεδομένων πραγματοποιήθηκε σε μια αστική περιοχή με έντονη κυκλοφοριακή δραστηριότητα, όπου εγκαταστάθηκε ένα σύστημα πολλαπλών αισθητήρων για την παρακολούθηση ατμοσφαιρικών ρύπων. Η χρονική περίοδος συλλογής εκτείνεται από τον Μάρτιο του 2004 έως τον Απρίλιο του 2005, καλύπτοντας συνολικά δεκατρείς μήνες συνεχούς καταγραφής. Οι μετρήσεις λαμβάνονται ωριαία, με αποτέλεσμα το σύνολο δεδομένων να περιλαμβάνει 9.471 παρατηρήσεις [1], [2].

Η δομή του συνόλου δεδομένων περιλαμβάνει δώδεκα μεταβλητές που αντιπροσωπεύουν διαφορετικές πτυχές της ποιότητας του αέρα και των περιβαλλοντικών συνθηκών. Οι μεταβλητές αυτές διακρίνονται σε τρεις κατηγορίες: μετρήσεις αναφοράς από πιστοποιημένα όργανα, αποκρίσεις ηλεκτρονικών αισθητήρων και περιβαλλοντικές παραμέτρους.

Η πρώτη κατηγορία περιλαμβάνει τις μετρήσεις αναφοράς, οι οποίες λαμβάνονται από πιστοποιημένα αναλυτικά όργανα και θεωρούνται ως η "αληθής" τιμή των ρύπων. Συγκεκριμένα, η μεταβλητή CO(GT) αντιπροσωπεύει τη συγκέντρωση μονοξειδίου του άνθρακα σε mg/m^3 , η C6H6(GT) τη συγκέντρωση βενζολίου σε mg/m^3 , η NOx(GT) τη συγκέντρωση οξειδίων του αζώτου σε ppb, και η NO2(GT) τη συγκέντρωση διοξειδίου του αζώτου σε mg/m^3 . Αυτές οι μετρήσεις αναφοράς χρησιμεύουν τόσο ως στόχοι πρόβλεψης, όσο και ως χαρακτηριστικά εισόδου στα μοντέλα.

Η δεύτερη κατηγορία περιλαμβάνει τις αποκρίσεις πέντε ηλεκτρονικών αισθητήρων τύπου μεταλλικού οξειδίου (Metal Oxide Sensors). Οι αισθητήρες αυτοί παρέχουν μη γραμμικές αποκρίσεις σε διάφορους ρύπους: ο PT08.S1 ανταποκρίνεται κυρίως στο CO, ο PT08.S2 στους μη μεθανικούς υδρογονάνθρακες (NMHC), ο PT08.S3 στα NOx, ο PT08.S4 στο NO2, και ο PT08.S5 στο O3. Η πρόκληση της βαθμονόμησης αυτών των αισθητήρων αποτέλεσε το αρχικό κίνητρο για τη δημοσίευση του συνόλου δεδομένων [1].

Η τρίτη κατηγορία περιλαμβάνει τρεις περιβαλλοντικές παραμέτρους: τη θερμοκρασία (T) σε βαθμούς Κελσίου, τη σχετική υγρασία (RH) σε ποσοστό, και την απόλυτη υγρασία (AH) σε g/m^3 . Οι παράμετροι αυτές είναι σημαντικές, καθώς επηρεάζουν τόσο τη συμπεριφορά των αισθητήρων όσο και τη διασπορά των ρύπων στην ατμόσφαιρα.

Ένα σημαντικό χαρακτηριστικό του συνόλου δεδομένων είναι η παρουσία ελλειπουσών τιμών, οι οποίες κωδικοποιούνται με την τιμή φρουρού -200. Η διερευνητική ανάλυση αποκάλυψε ότι το ποσοστό ελλειπουσών τιμών ανέρχεται περίπου στο 8.47% του συνόλου των δεδομένων, με τη μεταβλητή NMHC(GT) να παρουσιάζει το υψηλότερο ποσοστό απωλειών (90.15%), γεγονός που οδήγησε στον αποκλεισμό της από την ανάλυση. Στο ίδιο πλαίσιο, οι υπόλοιπες μεταβλητές εμφανίζουν σημαντικά χαμηλότερα ποσοστά ελλειπουσών τιμών, τα οποία αντιμετωπίζονται μέσω τεχνικών παρεμβολής που περιγράφονται στο Κεφάλαιο 4.

Για τους σκοπούς της παρούσας εργασίας, η μεταβλητή-στόχος που επιλέγεται είναι η CO(GT), δηλαδή η συγκέντρωση μονοξειδίου του άνθρακα. Η επιλογή αυτή εξηγείται λόγω της σημασίας του CO ως δείκτη ποιότητας αέρα σε αστικές περιοχές με κυκλοφοριακή δραστηριότητα, καθώς

και από το γεγονός ότι η μεταβλητή αυτή παρουσιάζει ικανοποιητική πληρότητα δεδομένων και ισχυρές συσχετίσεις με τις υπόλοιπες μεταβλητές του συνόλου. Ο ορίζοντας πρόβλεψης καθορίζεται σε μία ώρα εμπρός ($t+1$), επιτρέποντας την αξιολόγηση της ικανότητας των μοντέλων να προβλέπουν βραχυπρόθεσμες μεταβολές στην ποιότητα του αέρα.

3.1.2 Σύνολο Δεδομένων Κατανάλωσης Ηλεκτρικής Ενέργειας (Individual Household Electric Power Consumption Dataset)

Το δεύτερο σύνολο δεδομένων που χρησιμοποιείται είναι το Individual Household Electric Power Consumption Dataset, το οποίο δημοσιεύθηκε από τους Hébrail και Bérard [3] και διατίθεται επίσης μέσω του UCI Machine Learning Repository. Το σύνολο δεδομένων αυτό περιέχει μετρήσεις κατανάλωσης ηλεκτρικής ενέργειας από ένα μεμονωμένο νοικοκυριό στην περιοχή Sceaux της Γαλλίας.

Η χρονική περίοδος συλλογής εκτείνεται από τον Δεκέμβριο του 2006 έως τον Νοέμβριο του 2010, καλύπτοντας σχεδόν τέσσερα έτη συνεχούς καταγραφής. Η αρχική συχνότητα δειγματοληψίας είναι μία μέτρηση ανά λεπτό, με αποτέλεσμα το σύνολο δεδομένων να περιλαμβάνει περισσότερες από 2.075.259 παρατηρήσεις [3]. Για τη συνέπεια με το σύνολο δεδομένων ποιότητας αέρα και την πρακτική εφαρμογή σε συστήματα IoT, τα δεδομένα επαναδειγματοληπτούνται (resample) σε ωριαία ανάλυση, παράγοντας περίπου 35.000 ωριαίες παρατηρήσεις.

Η δομή του συνόλου δεδομένων περιλαμβάνει επτά μεταβλητές που αφορούν διαφορετικές πτυχές της ηλεκτρικής κατανάλωσης του νοικοκυριού. Η κύρια μεταβλητή, `Global_active_power`, αντιπροσωπεύει τη συνολική ενεργό ισχύ που καταναλώνεται από το νοικοκυριό, μετρημένη σε kilowatt (kW). Η μεταβλητή αυτή επιλέγεται ως στόχος πρόβλεψης για την παρούσα εργασία. Επιπλέον, καταγράφεται η συνολική άεργος ισχύς (`Global_reactive_power`) σε kilowatt, η τάση του δικτύου (Voltage) σε volt, και η συνολική ένταση ρεύματος (`Global_intensity`) σε ampere.

Ιδιαίτερο ενδιαφέρον παρουσιάζουν οι τρεις μεταβλητές υπομέτρησης (sub-metering), οι οποίες καταγράφουν την κατανάλωση συγκεκριμένων ομάδων συσκευών εντός του νοικοκυριού σε watt-hour (Wh). Η `Sub_metering_1` αντιστοιχεί στην κουζίνα (κυρίως πλυντήριο πιάτων, φούρνος και

μικροκύματα), η Sub_metering_2 στο πλυντήριο ρούχων, στο στεγνωτήριο, στο ψυγείο και σε μία λάμπα, και η Sub_metering_3 στον θερμοσίφωνα και στον κλιματισμό. Η διαφορά μεταξύ της συνολικής κατανάλωσης και του αθροίσματος των τριών υπομετρήσεων αντιπροσωπεύει την κατανάλωση από μη μετρούμενες συσκευές.

Το ποσοστό ελλειπουσών τιμών στο σύνολο δεδομένων κατανάλωσης ενέργειας είναι σημαντικά χαμηλότερο σε σύγκριση με το σύνολο δεδομένων ποιότητας αέρα, ανερχόμενο περίπου στο 1.25%. Οι ελλείπουσες τιμές κωδικοποιούνται με το σύμβολο "?" και αντιμετωπίζονται με παρόμοιες τεχνικές παρεμβολής. Η χαμηλότερη συχνότητα απωλειών αντανακλά την αξιοπιστία των συστημάτων μέτρησης ενέργειας σε σύγκριση με τους περιβαλλοντικούς αισθητήρες.

Ένα χαρακτηριστικό γνώρισμα του συνόλου δεδομένων κατανάλωσης ενέργειας είναι η παρουσία έντονων μοτίβων εποχικότητας σε πολλαπλές χρονικές κλίμακες. Σε ημερήσιο επίπεδο, παρατηρούνται διακυμάνσεις που σχετίζονται με τις ώρες αιχμής κατανάλωσης (πρωί και βράδυ) και τις ώρες χαμηλής κατανάλωσης (νύχτα). Σε εβδομαδιαίο επίπεδο, διαπιστώνεται διαφορετική συμπεριφορά μεταξύ εργάσιμων ημερών και Σαββατοκύριακου. Σε εποχιακό επίπεδο, η κατανάλωση είναι υψηλότερη κατά τους χειμερινούς μήνες λόγω της θέρμανσης και του τεχνητού φωτισμού.

3.1.3 Συγκριτική Ανάλυση των Συνόλων Δεδομένων

Παρακάτω, η χρήση δύο συνόλων δεδομένων από διαφορετικούς τομείς εφαρμογής επιτρέπει τη διεξαγωγή συγκριτικής ανάλυσης και την αξιολόγηση της γενικευσιμότητας των μεθόδων. Ο Πίνακας 3.1 συνοψίζει τα κύρια χαρακτηριστικά των δύο συνόλων δεδομένων.

Πίνακας 3.1: Συγκριτικά Χαρακτηριστικά Συνόλων Δεδομένων

Χαρακτηριστικό	Air Quality	Power Consumption
Πεδίο Εφαρμογής	Περιβαλλοντική Παρακολούθηση	Διαχείριση Ενέργειας
Πηγή	UCI Machine Learning Repository [2]	UCI Machine Learning Repository [3]

Τοποθεσία Συλλογής	Αστική περιοχή, Ιταλία	Sceaux, Γαλλία
Χρονική Περίοδος	Μάρτιος 2004 - Απρίλιος 2005	Δεκέμβριος 2006 - Νοέμβριος 2010
Διάρκεια	13 μήνες	47 μήνες
Αρχική Συχνότητα	1 ώρα	1 λεπτό
Τελική Συχνότητα	1 ώρα	1 ώρα (μετά resample)
Αριθμός Παρατηρήσεων	9.471	~2.075.259 (αρχικά), ~35.000 (ωριαία)
Αριθμός Χαρακτηριστικών	12	7
Ποσοστό Ελλειπουσών Τιμών	~8.47%	~1.25%
Κωδικοποίηση Ελλείψεων	-200	"?"
Μεταβλητή-Στόχος	CO(GT) [mg/m ³]	Global_active_power [kW]
Ορίζοντας Πρόβλεψης	1 ώρα (t+1)	1 ώρα (t+1)
Τύπος Εποχικότητας	Μετεωρολογική, κυκλοφοριακή	Ανθρώπινη συμπεριφορά
Κύριοι Παράγοντες	Καιρός, κυκλοφορία	Καθημερινότητα, εποχή

Οι ομοιότητες μεταξύ των δύο συνόλων δεδομένων περιλαμβάνουν, τη φύση τους ως χρονοσειρές IoT, τον κοινό ορίζοντα πρόβλεψης μίας ώρας, την ύπαρξη πολλαπλών επιπέδων εποχικότητας, και την παρουσία ελλειπουσών τιμών που απαιτούν διαχείριση. Αυτές οι ομοιότητες επιτρέπουν την εφαρμογή κοινής μεθοδολογικής προσέγγισης και τη σύγκριση αποτελεσμάτων.

Οι διαφορές αφορούν κυρίως το μέγεθος των δεδομένων (το σύνολο κατανάλωσης ενέργειας είναι περίπου 200 φορές μεγαλύτερο από αυτό της ποιότητας αέρα σε αρχικές παρατηρήσεις, γνωρισμά που επιδρά στην απόδοση των μοντέλων βαθιάς μηχανικής μάθησης), τον τομέα εφαρμογής (περιβαλλοντική παρακολούθηση έναντι διαχείρισης ενέργειας), και τους παράγοντες που καθορίζουν τη συμπεριφορά της μεταβλητής-στόχου (μετεωρολογικές συνθήκες και κυκλοφορία έναντι ανθρώπινης καθημερινότητας). Οι παραπάνω διαφορές αναμένεται να επηρεάσουν τη σχετική απόδοση των διαφόρων αλγορίθμων και να παρέχουν χρήσιμες πληροφορίες για την επιλογή μεθόδων σε διαφορετικούς τομείς IoT.

3.2 Εργαλεία και Περιβάλλον (Tools & Environment)

Η επιλογή κατάλληλων εργαλείων λογισμικού και η διαμόρφωση ενός σταθερού περιβάλλοντος ανάπτυξης αποτελούν θεμελιώδεις προϋποθέσεις για την αναπαραγωγιμότητα και την αξιοπιστία των αποτελεσμάτων ενός έργου μηχανικής μάθησης [4], [5]. Η παρούσα ενότητα περιγράφει αναλυτικά τα εργαλεία που χρησιμοποιήθηκαν, τις εκδόσεις τους και τη λογική πίσω από την επιλογή τους.

3.2.1 Γλώσσα Προγραμματισμού

Ως κύρια γλώσσα προγραμματισμού επιλέχθηκε η Python, έκδοση 3.13.9 ή νεότερη. Η επιλογή αυτή δικαιολογείται από πλήθος παραγόντων: η Python αποτελεί μία από τις ευρύτερα χρησιμοποιούμενες γλώσσες προγραμματισμού για εφαρμογές μηχανικής μάθησης και επιστήμης δεδομένων, διαθέτει ένα εκτεταμένο οικοσύστημα βιβλιοθηκών ειδικά σχεδιασμένων για αυτούς τους σκοπούς και χαρακτηρίζεται από υψηλή αναγνωσιμότητα κώδικα που διευκολύνει την τεκμηρίωση και την αναπαραγωγή πειραμάτων.

Ακόμη, η χρήση της έκδοσης 3.13 εξασφαλίζει την πρόσβαση στις τελευταίες βελτιώσεις απόδοσης και τις νέες δυνατότητες της γλώσσας, ενώ παράλληλα διατηρεί τη συμβατότητα με τις κύριες βιβλιοθήκες μηχανικής μάθησης. Για τη διαχείριση των εξαρτήσεων χρησιμοποιείται εικονικό περιβάλλον (virtual environment), το οποίο απομονώνει τις βιβλιοθήκες του έργου από το σύστημα και εξασφαλίζει τη συνέπεια μεταξύ διαφορετικών εγκαταστάσεων.

3.2.2 Βιβλιοθήκες Επεξεργασίας Δεδομένων

Για την επεξεργασία και τον χειρισμό δεδομένων χρησιμοποιούνται οι βιβλιοθήκες NumPy (έκδοση $\geq 2.3.4$) και Pandas (έκδοση $\geq 2.3.3$). Από τη μία, η NumPy παρέχει υψηλής απόδοσης πίνακες πολλαπλών διαστάσεων και μαθηματικές συναρτήσεις, που αποτελούν τη βάση για όλους τους υπολογισμούς αριθμητικής ανάλυσης. Από την άλλη, η Pandas προσφέρει τις δομές δεδομένων DataFrame και Series, οι οποίες είναι ιδανικές για τη διαχείριση χρονοσειρών και την εκτέλεση πράξεων όπως ομαδοποίηση, φιλτράρισμα και συγχώνευση δεδομένων.

Για τη στατιστική ανάλυση και την αποσύνθεση χρονοσειρών χρησιμοποιούνται οι βιβλιοθήκες SciPy (έκδοση $\geq 1.15.2$) και statsmodels (έκδοση $\geq 0.14.4$). Η SciPy παρέχει αλγορίθμους επιστημονικού υπολογισμού, συμπεριλαμβανομένων στατιστικών ελέγχων και μεθόδων παρεμβολής. Ενώ, η statsmodels προσφέρει εξειδικευμένες μεθόδους ανάλυσης χρονοσειρών, όπως η αποσύνθεση STL (Seasonal and Trend decomposition using Loess) που χρησιμοποιείται για την ανίχνευση ανωμαλιών [22].

3.2.3 Βιβλιοθήκες Οπτικοποίησης

Η οπτικοποίηση δεδομένων και αποτελεσμάτων πραγματοποιείται με τη χρήση των βιβλιοθηκών Matplotlib (έκδοση $\geq 3.10.7$), Seaborn (έκδοση $\geq 0.13.2$) και Plotly (έκδοση $\geq 6.4.0$). Η Matplotlib αποτελεί τη θεμελιώδη βιβλιοθήκη οπτικοποίησης στην Python και παρέχει πλήρη έλεγχο επί όλων των στοιχείων ενός γραφήματος. Η Seaborn είναι χτισμένη πάνω στη Matplotlib και προσφέρει υψηλού επιπέδου διεπαφές για τη δημιουργία στατιστικών γραφημάτων με αισθητικά ελκυστική εμφάνιση. Τέλος, η Plotly χρησιμοποιείται για τη δημιουργία διαδραστικών γραφημάτων όπου αυτό κρίνεται χρήσιμο.

3.2.4 Βιβλιοθήκες Μηχανικής Μάθησης

Η υλοποίηση των αλγορίθμων μηχανικής μάθησης βασίζεται κυρίως στη βιβλιοθήκη Scikit-learn (έκδοση $\geq 1.7.2$), η οποία αποτελεί μία από τις πλέον διαδεδομένες βιβλιοθήκες μηχανικής μάθησης γενικού σκοπού για την Python [10]. Η Scikit-learn παρέχει υλοποιήσεις για πληθώρα αλγορίθμων, συμπεριλαμβανομένων της Γραμμικής Παλινδρόμησης (Linear Regression), του Random Forest [7], του Gradient Boosting [9], του Isolation Forest [19], του Local Outlier Factor [20], και του One-Class SVM [21]. Επιπλέον, η βιβλιοθήκη προσφέρει εργαλεία για

προεπεξεργασία δεδομένων (κανονικοποίηση, κωδικοποίηση), επιλογή χαρακτηριστικών, διασταυρούμενη επικύρωση και αξιολόγηση μοντέλων.

Παράλληλα, για την υλοποίηση του αλγορίθμου XGBoost χρησιμοποιείται η ομώνυμη βιβλιοθήκη XGBoost (έκδοση $\geq 3.1.1$). Ο XGBoost αποτελεί μια εξαιρετικά βελτιστοποιημένη υλοποίηση του gradient boosting που έχει επιτύχει κορυφαίες επιδόσεις σε πλήθος διαγωνισμών μηχανικής μάθησης [8]. Η βιβλιοθήκη παρέχει δυνατότητες όπως, παράλληλη εκπαίδευση, πρόωμη διακοπή (early stopping) και ενσωματωμένη διαχείριση ελλειπουσών τιμών.

3.2.5 Βιβλιοθήκες Βαθιάς Μάθησης

Για την υλοποίηση των αρχιτεκτονικών βαθιάς μάθησης LSTM [11] και GRU [12] χρησιμοποιείται η βιβλιοθήκη TensorFlow (έκδοση $\geq 2.20.0$). Η TensorFlow αποτελεί μία από τις κορυφαίες βιβλιοθήκες βαθιάς μάθησης, αναπτυγμένη από τη Google, η οποία προσφέρει εκτεταμένες δυνατότητες για τον ορισμό, την εκπαίδευση και την αξιολόγηση νευρωνικών δικτύων.¹

Η επιλογή της TensorFlow έγινε λαμβάνοντας υπόψη τη σταθερότητα, την εκτεταμένη τεκμηρίωση και την υποστήριξη για διάφορες πλατφόρμες υλικού, συμπεριλαμβανομένων CPU και GPU. Ωστόσο, όπως θα παρουσιαστεί στα επόμενα κεφάλαια, το σχετικά μικρό μέγεθος του συνόλου δεδομένων ποιότητας αέρα, περιορίζει τα οφέλη που μπορούν να αποκομιστούν από αρχιτεκτονικές βαθιάς μηχανικής μάθησης.

3.2.6 Περιβάλλον Ανάπτυξης και Εκτέλεσης

Η ανάπτυξη και η εκτέλεση του κώδικα πραγματοποιείται μέσω Jupyter Notebooks (έκδοση $\geq 1.1.1$), τα οποία επιτρέπουν τη διαδραστική εκτέλεση κώδικα, τη διατήρηση των αποτελεσμάτων μαζί με τον κώδικα, και την εύκολη τεκμηρίωση της διαδικασίας ανάλυσης. Η οργάνωση του κώδικα σε διαδοχικά notebooks ακολουθεί τη ροή της μεθοδολογίας CRISP-DM: 01_EDA.ipynb για τη διερευνητική ανάλυση δεδομένων, 02_Preprocessing.ipynb για την προεπεξεργασία

¹ Η υψηλού επιπέδου βιβλιοθήκη Keras, η οποία είναι ενσωματωμένη στην TensorFlow, χρησιμοποιείται για τον ορισμό των επαναλαμβανόμενων αρχιτεκτονικών με τρόπο ευανάγνωστο και ευέλικτο.

και τον σχεδιασμό χαρακτηριστικών, 03_Modeling.ipynb για την εκπαίδευση και αξιολόγηση μοντέλων και 04_Anomaly.ipynb για την ανίχνευση ανωμαλιών.

Επιπλέον, ο επαναχρησιμοποιήσιμος κώδικας οργανώνεται σε modules στον φάκελο src/, ο οποίος περιλαμβάνει τα αρχεία utils.py για βοηθητικές συναρτήσεις, data_preprocessing.py για τη διαχείριση ελλειπουσών τιμών και ακραίων τιμών, feature_engineering.py για τη δημιουργία χαρακτηριστικών, models.py για τους ορισμούς των μοντέλων, evaluation.py για τις μεθόδους αξιολόγησης και anomaly_detection.py για τους αλγορίθμους ανίχνευσης ανωμαλιών.

3.2.7 Πίνακας Εξαρτήσεων Λογισμικού

Ο Πίνακας 3.2 συνοψίζει τις κύριες εξαρτήσεις λογισμικού που χρησιμοποιούνται στην παρούσα εργασία.

Πίνακας 3.2: Εξαρτήσεις (Dependencies) Λογισμικού

Κατηγορία	Βιβλιοθήκη	Έκδοση	Σκοπός Χρήσης
Επεξεργασία Δεδομένων	NumPy	≥2.3.4	Αριθμητικοί υπολογισμοί, πίνακες
Επεξεργασία Δεδομένων	Pandas	≥2.3.3	DataFrames, χρονοσειρές
Επεξεργασία Δεδομένων	SciPy	≥1.15.2	Επιστημονικοί υπολογισμοί, στατιστική
Επεξεργασία Δεδομένων	statsmodels	≥0.14.4	Στατιστικά μοντέλα, STL αποσύνθεση
Οπτικοποίηση	Matplotlib	≥3.10.7	Βασικά γραφήματα
Οπτικοποίηση	Seaborn	≥0.13.2	Στατιστικά γραφήματα
Οπτικοποίηση	Plotly	≥6.4.0	Διαδραστικά γραφήματα

Κατηγορία	Βιβλιοθήκη	Έκδοση	Σκοπός Χρήσης
Μηχανική Μάθηση	Scikit-learn	≥1.7.2	Αλγόριθμοι ML, προεπεξεργασία, αξιολόγηση [10]
Μηχανική Μάθηση	XGBoost	≥3.1.1	Gradient boosting [8]
Βαθιά Μάθηση	TensorFlow	≥2.20.0	LSTM, GRU νευρωνικά δίκτυα [11], [12]
Ανάπτυξη	Jupyter	≥1.1.1	Διαδραστικά notebooks

3.2.8 Διαχείριση Αναπαραγωγιμότητας

Η αναπαραγωγιμότητα των αποτελεσμάτων διασφαλίζεται μέσω πολλαπλών μηχανισμών. Πρώτον, όλες οι τυχαίες διεργασίες αρχικοποιούνται με σταθερό σπόρο (random seed = 42) σε επίπεδο NumPy, Scikit-learn και TensorFlow/Keras, εξασφαλίζοντας ντετερμινιστικά αποτελέσματα. Σημειώνεται ότι στα μοντέλα βαθιάς μάθησης η πλήρης αναπαραγωγιμότητα δεν είναι πάντοτε εγγυημένη, καθώς ορισμένες λειτουργίες GPU εισάγουν μη ντετερμινιστική συμπεριφορά· ωστόσο, ο σταθερός σπόρος ελαχιστοποιεί τη μεταβλητότητα μεταξύ εκτελέσεων. Δεύτερον, οι εκδόσεις όλων των βιβλιοθηκών καταγράφονται στο αρχείο requirements.txt, επιτρέποντας την ακριβή αναδημιουργία του περιβάλλοντος εκτέλεσης. Τρίτον, όλα τα ενδιάμεσα αποτελέσματα (επεξεργασμένα δεδομένα, εκπαιδευμένα μοντέλα, αναφορές αξιολόγησης) αποθηκεύονται σε οργανωμένη δομή φακέλων, επιτρέποντας τόσο την ανασκόπηση όσο και την επέκταση της εργασίας.

Σύνοψη Κεφαλαίου

Το παρόν κεφάλαιο παρουσίασε τη μεθοδολογική βάση και τα δεδομένα που χρησιμοποιούνται στην παρούσα διπλωματική εργασία. Περιγράφονται αναλυτικά τα δύο σύνολα δεδομένων που χρησιμοποιούνται: το UCI Air Quality Dataset [1], [2] για την περιβαλλοντική παρακολούθηση με 9.471 ωριαίες μετρήσεις ατμοσφαιρικών ρύπων και το Individual Household Electric Power Consumption Dataset [3] για τη διαχείριση ενέργειας με περισσότερες από δύο εκατομμύρια αρχικές παρατηρήσεις κατανάλωσης ηλεκτρικής ενέργειας.

Η συγκριτική ανάλυση των δύο συνόλων δεδομένων ανέδειξε τόσο τις ομοιότητες (φύση χρονοσειρών IoT, κοινός ορίζοντας πρόβλεψης, ύπαρξη ελλειπουσών τιμών), όσο και τις διαφορές τους (μέγεθος δεδομένων, τομέας εφαρμογής, παράγοντες που καθορίζουν τη συμπεριφορά). Έτσι, αυτή η διπλή προσέγγιση επιτρέπει την αξιολόγηση της γενικευσιμότητας των μεθόδων που αναπτύσσονται.

Έπειτα, ακολούθησε η περιγραφή του τεχνικού περιβάλλοντος ανάπτυξης, συμπεριλαμβανομένων της γλώσσας Python 3.13+, των βιβλιοθηκών Scikit-learn [10], XGBoost [8] και TensorFlow για τη μηχανική και βαθιά μάθηση, καθώς και των εργαλείων οπτικοποίησης και στατιστικής ανάλυσης. Στο σημείο αυτό, ιδιαίτερη έμφαση δόθηκε στη διασφάλιση της αναπαραγωγιμότητας μέσω σταθερών σπόρων τυχαιότητας, καταγραφής εκδόσεων και οργανωμένης αποθήκευσης αποτελεσμάτων.

Τα επόμενα κεφάλαια εφαρμόζουν τη μεθοδολογία που περιγράφηκε εδώ, ξεκινώντας με την ανάλυση του συνόλου δεδομένων ποιότητας αέρα (Κεφάλαιο 4), συνεχίζοντας με το σύνολο δεδομένων κατανάλωσης ενέργειας (Κεφάλαιο 5), και ολοκληρώνοντας με τη συγκριτική ανάλυση μεταξύ των δύο τομέων (Κεφάλαιο 6).

ΚΕΦΑΛΑΙΟ 4: AIR QUALITY - ΑΝΑΛΥΣΗ ΚΑΙ ΜΟΝΤΕΛΟΠΟΙΗΣΗ

Εισαγωγή στο Κεφάλαιο

Το παρόν κεφάλαιο αποτελεί την πρακτική εφαρμογή της μεθοδολογίας που περιγράφηκε στα προηγούμενα κεφάλαια επί του συνόλου δεδομένων ποιότητας αέρα UCI Air Quality Dataset [1], [2]. Ακολουθώντας τη μεθοδολογία CRISP-DM (Cross-Industry Standard Process for Data Mining) [25], το κεφάλαιο καλύπτει τις φάσεις της Προετοιμασίας Δεδομένων (Data Preparation), της Μοντελοποίησης (Modeling) και της Αξιολόγησης (Evaluation). Με τη δομημένη αυτή προσέγγιση, διασφαλίζεται η αναπαραγωγικότητα των αποτελεσμάτων και η μεθοδολογική συνέπεια της ανάλυσης [25].

Ο πρωταρχικός στόχος της ανάλυσης είναι η ανάπτυξη μοντέλων πρόβλεψης για τη συγκέντρωση μονοξειδίου του άνθρακα CO(GT) μία ώρα στο μέλλον (1-hour ahead forecasting), αξιοποιώντας τις μετρήσεις ατμοσφαιρικών αισθητήρων και περιβαλλοντικές μεταβλητές. Η επιλογή του μονοξειδίου του άνθρακα ως μεταβλητή-στόχο δικαιολογείται τόσο από την κρισιμότητά του για τη δημόσια υγεία, όσο και από την ποιότητα των διαθέσιμων δεδομένων στο σύνολο δεδομένων [1].

Η δομή του κεφαλαίου οργανώνεται ως εξής: η Ενότητα 4.1 παρουσιάζει τη Διερευνητική Ανάλυση Δεδομένων (EDA) και την προεπεξεργασία, η Ενότητα 4.2 περιγράφει την σχεδίαση χαρακτηριστικών (Feature Engineering), η Ενότητα 4.3 αναλύει τη διαδικασία μοντελοποίησης, η Ενότητα 4.4 παρουσιάζει την αξιολόγηση των μοντέλων, και τέλος η Ενότητα 4.5 επικεντρώνεται στην ανίχνευση ανωμαλιών.

Αναλυτικά αποτελέσματα παρατίθενται στο παράρτημα Γ (βλέπε σελ. 139). Όλοι οι πίνακες και τα δεδομένα τους μπορούν να βρεθούν τόσο στο παράρτημα Γ, όσο και στο αποθετήριο κώδικα του παραρτήματος Β (βλέπε σελ. 137 και 139).

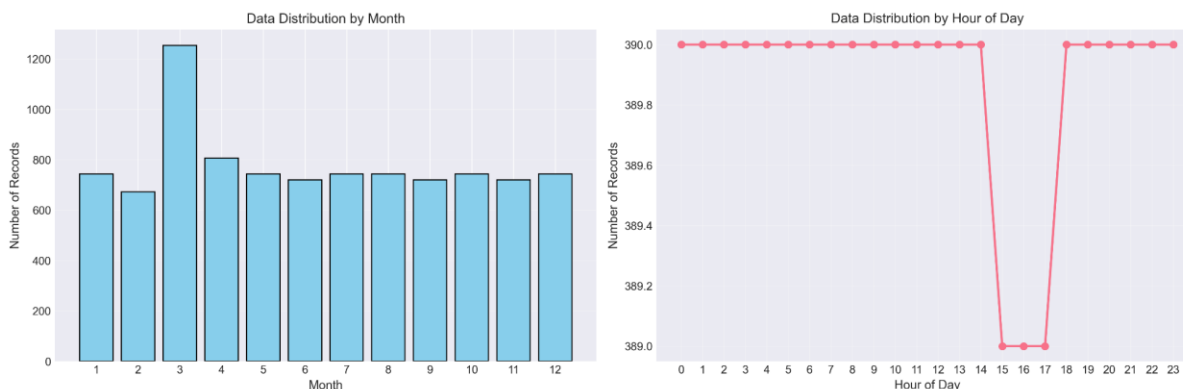
4.1 Διερευνητική Ανάλυση Δεδομένων και Προεπεξεργασία (EDA & Preprocessing)

Η Διερευνητική Ανάλυση Δεδομένων (Exploratory Data Analysis - EDA) αποτελεί κρίσιμο πρώτο βήμα σε κάθε έργο μηχανικής μάθησης, καθώς επιτρέπει την κατανόηση της δομής, της ποιότητας και των μοτίβων που υπάρχουν στα δεδομένα [4]. Σύμφωνα με τη μεθοδολογία CRISP-DM, αυτή η φάση αντιστοιχεί στο πέρασμα από την Κατανόηση των Δεδομένων (Data Understanding) στην Προετοιμασία Δεδομένων (Data Preparation) [25].

4.1.1 Επισκόπηση του Συνόλου Δεδομένων

Το σύνολο δεδομένων UCI Air Quality [1], [2] περιλαμβάνει 9.471 ωριαίες παρατηρήσεις που συλλέχθηκαν κατά τη διάρκεια περίπου 13 μηνών, από τις 10 Μαρτίου 2004 έως τις 4 Απριλίου 2005. Οι μετρήσεις πραγματοποιήθηκαν σε αστική περιοχή στην Ιταλία, σε σημείο με έντονη κυκλοφοριακή δραστηριότητα, καθιστώντας τα δεδομένα αντιπροσωπευτικά τυπικών αστικών συνθηκών ρύπανσης [1].

Το σύνολο δεδομένων περιλαμβάνει δώδεκα χαρακτηριστικά αισθητήρων και περιβαλλοντικών μετρήσεων. Τα πρωτογενή χαρακτηριστικά περιλαμβάνουν μετρήσεις αναφοράς για πέντε ρύπους: CO(GT) για το μονοξείδιο του άνθρακα σε mg/m^3 , C6H6(GT) για το βενζόλιο σε mg/m^3 , NOx(GT) για τα οξείδια του αζώτου σε ppb, NO2(GT) για το διοξείδιο του αζώτου σε mg/m^3 , και NMHC(GT) για τους μη μεθανικούς υδρογονάνθρακες. Επιπλέον, περιλαμβάνονται πέντε αποκρίσεις μεταλλοξειδικών αισθητήρων: PT08.S1(CO), PT08.S2(NMHC), PT08.S3(NOx), PT08.S4(NO2) και PT08.S5(O3). Ακόμη, καταγράφονται τρεις περιβαλλοντικές μεταβλητές: η θερμοκρασία T σε βαθμούς Κελσίου, η σχετική υγρασία RH ως ποσοστό, και η απόλυτη υγρασία AH.



Εικόνα 4.1: Χρονική κατανομή δεδομένων ανά μήνα και ώρα ημέρας

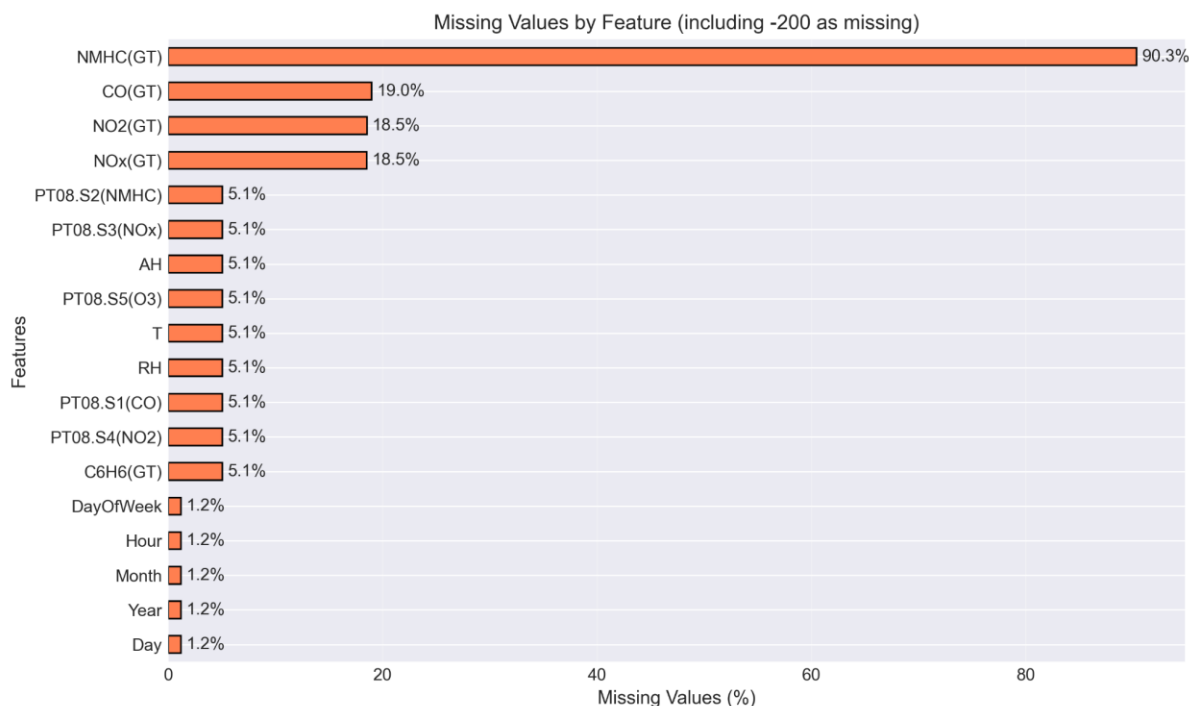
Βλέπε Παράρτημα Β για την πηγή της εικόνας στον φάκελο figures/ του αποθετηρίου.

Η χρονική ανάλυση των δεδομένων αποκάλυψε σημαντικά μοτίβα κατανομής. Αναλυτικότερα, η μηνιαία κατανομή παρουσιάζει διακυμάνσεις που σχετίζονται με τις εποχιακές μεταβολές της ατμοσφαιρικής ρύπανσης, με υψηλότερες συγκεντρώσεις ρύπων κατά τους χειμερινούς μήνες λόγω των αυξημένων εκπομπών θέρμανσης και των δυσμενέστερων μετεωρολογικών συνθηκών διασποράς. Ενώ, η ωριαία ανάλυση αποκάλυψε έντονα ημερήσια μοτίβα, με κορυφώσεις της συγκέντρωσης CO(GT) κατά τις πρωινές ώρες αιχμής (08:00-09:00) και τις απογευματινές (18:00-20:00), αντανακλώντας την κυκλοφοριακή δραστηριότητα στην περιοχή μέτρησης.

4.1.2 Ανάλυση Ελλειπουσών Τιμών

Η διαχείριση ελλειπουσών τιμών αποτελεί θεμελιώδες ζήτημα στην ανάλυση χρονοσειρών IoT [4]. Στο συγκεκριμένο σύνολο δεδομένων, οι ελλείπουσες τιμές κωδικοποιούνται με δύο τρόπους: ως τυπικές τιμές NaN (Not a Number) και ως τιμή φρουρού (sentinel value) -200, η οποία υποδηλώνει αποτυχία του αισθητήρα ή μη έγκυρη μέτρηση [1].

Η συστηματική ανάλυση των ελλειπουσών τιμών αποκάλυψε ότι το χαρακτηριστικό NMHC(GT) παρουσιάζει το υψηλότερο ποσοστό ελλείψεων, υπερβαίνοντας το 90% του συνόλου των παρατηρήσεων, γεγονός που οδήγησε στην απόφαση αποκλεισμού του από την περαιτέρω ανάλυση. Τα υπόλοιπα χαρακτηριστικά παρουσίασαν ποσοστά ελλειπουσών τιμών που κυμαίνονταν από 1% έως 20%, με μέσο όρο περίπου 8,5% για τα χαρακτηριστικά που διατηρήθηκαν στην ανάλυση.



Εικόνα 4.2: Οπτικοποίηση ελλειπουσών τιμών ανά χαρακτηριστικό

Βλέπε Παράρτημα Β για την πηγή της εικόνας στον φάκελο figures/ του αποθετηρίου.

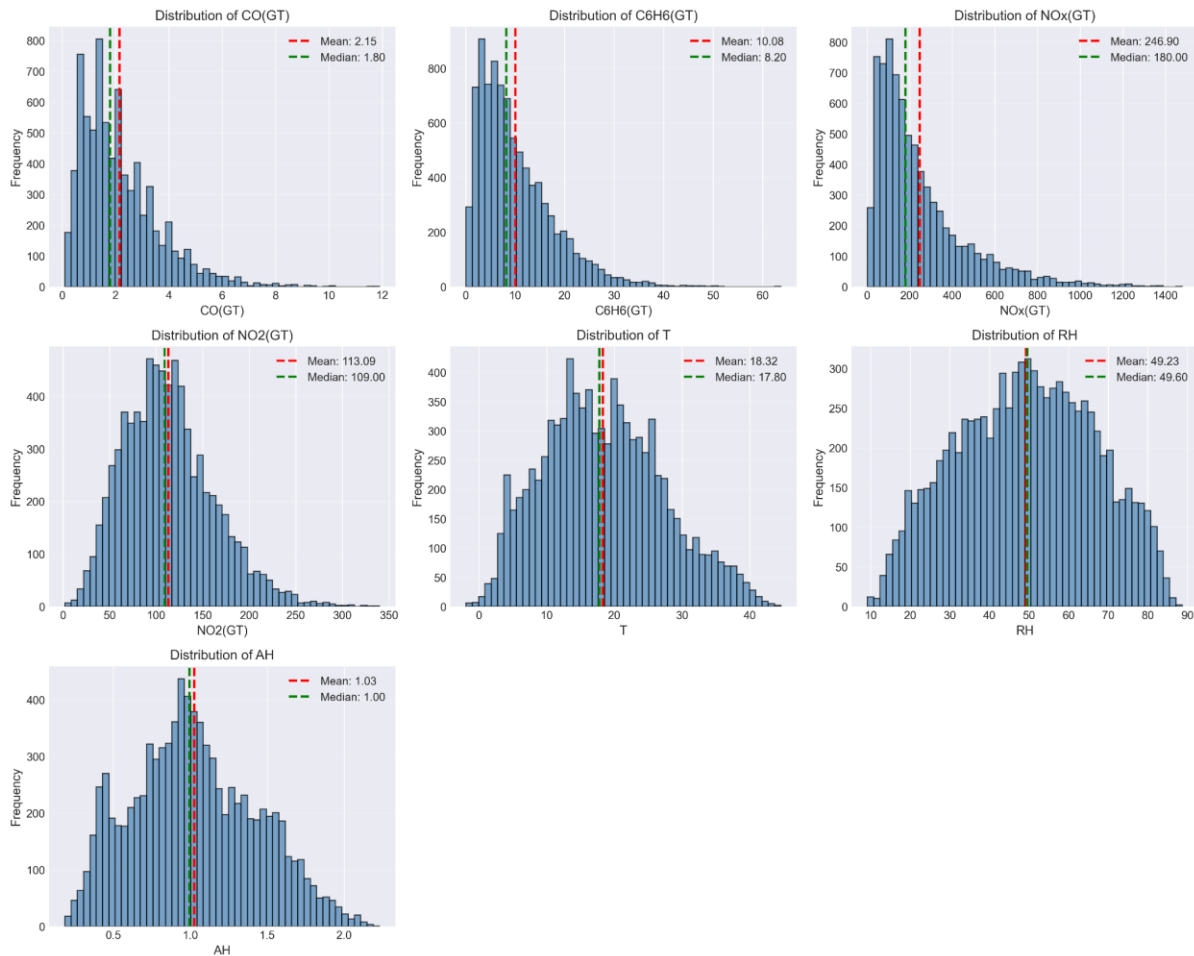
Για τη διαχείριση των ελλειπουσών τιμών επιλέχθηκε η μέθοδος της γραμμικής παρεμβολής (Linear Interpolation) ². Η γραμμική παρεμβολή διατηρεί τη χρονική συνέχεια των δεδομένων και αποφεύγει την εισαγωγή τεχνητών ασυνεχειών που θα μπορούσαν να επηρεάσουν αρνητικά την απόδοση των μοντέλων πρόβλεψης. Η μέθοδος εφαρμόστηκε, αφού πρώτα αντικαταστάθηκαν οι τιμές φρουρού -200 με τιμές NaN, εξασφαλίζοντας ενιαία αντιμετώπιση όλων των ελλειπουσών τιμών.

Η επιτυχία της διαδικασίας αντιμετώπισης ελλειπουσών τιμών επιβεβαιώθηκε με τον έλεγχο του τελικού συνόλου δεδομένων, το οποίο δεν περιείχε πλέον ελλείπουσες τιμές (0 NaN values), μετά την εφαρμογή της παρεμβολής.

² Η οποία είναι ιδιαίτερα κατάλληλη για δεδομένα χρονοσειρών με ωριαία ανάλυση [4]

4.1.3 Ανάλυση Κατανομών και Συσχετίσεων

Η στατιστική ανάλυση των κατανομών των χαρακτηριστικών αποκάλυψε σημαντική ετερογένεια μεταξύ των μεταβλητών. Η μεταβλητή-στόχος CO(GT) παρουσιάζει θετική ασυμμετρία (skewness $\approx 1,37$) με μέση τιμή 2,15 mg/m³ και τυπική απόκλιση 1,45 mg/m³, υποδηλώνοντας την ύπαρξη περιόδων με ασυνήθιστα υψηλές συγκεντρώσεις ρύπων. Το εύρος τιμών εκτείνεται από 0,1 έως 11,9 mg/m³, αντανακλώντας τη μεγάλη διακύμανση στα επίπεδα ρύπανσης.

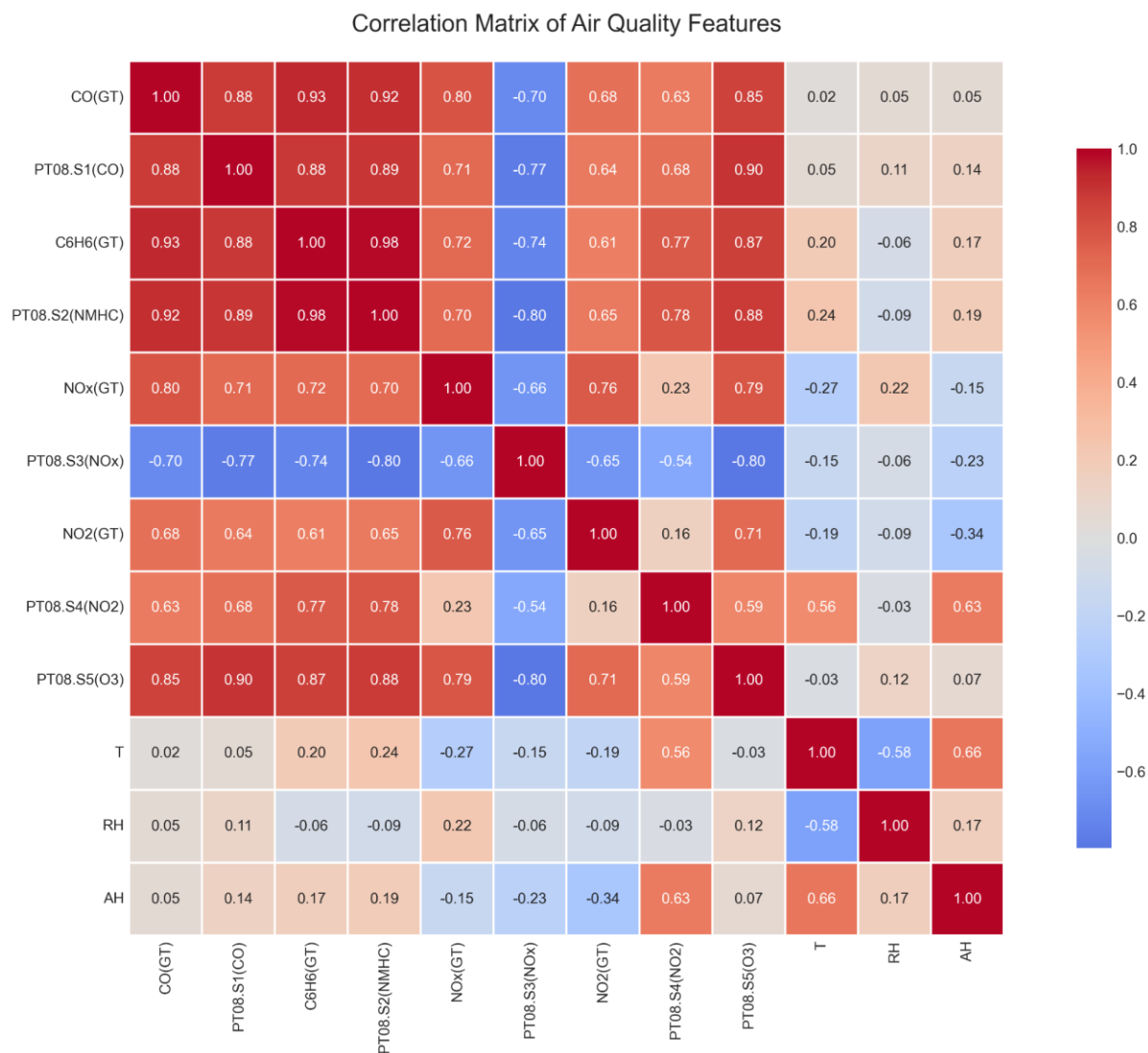


Εικόνα 4.3: Κατανομές χαρακτηριστικών (Feature Distributions)

Βλέπε Παράρτημα Β για την πηγή της εικόνας στον φάκελο figures/ του αποθετηρίου.

Η ανάλυση συσχετίσεων αποκάλυψε ισχυρές γραμμικές σχέσεις μεταξύ ορισμένων χαρακτηριστικών. Η υψηλότερη συσχέτιση παρατηρήθηκε μεταξύ της μεταβλητής-στόχου

CO(GT) και της απόκρισης του αντίστοιχου αισθητήρα PT08.S1(CO), με συντελεστή συσχέτισης Pearson $r = 0,879$. Αυτή η ισχυρή συσχέτιση επιβεβαιώνει την εγκυρότητα των μετρήσεων του αισθητήρα και υποδεικνύει τη δυνατότητα αξιοποίησής του ως προγνωστικού χαρακτηριστικού.



Εικόνα 4.4: Πίνακας συσχετίσεων (Correlation Heatmap)

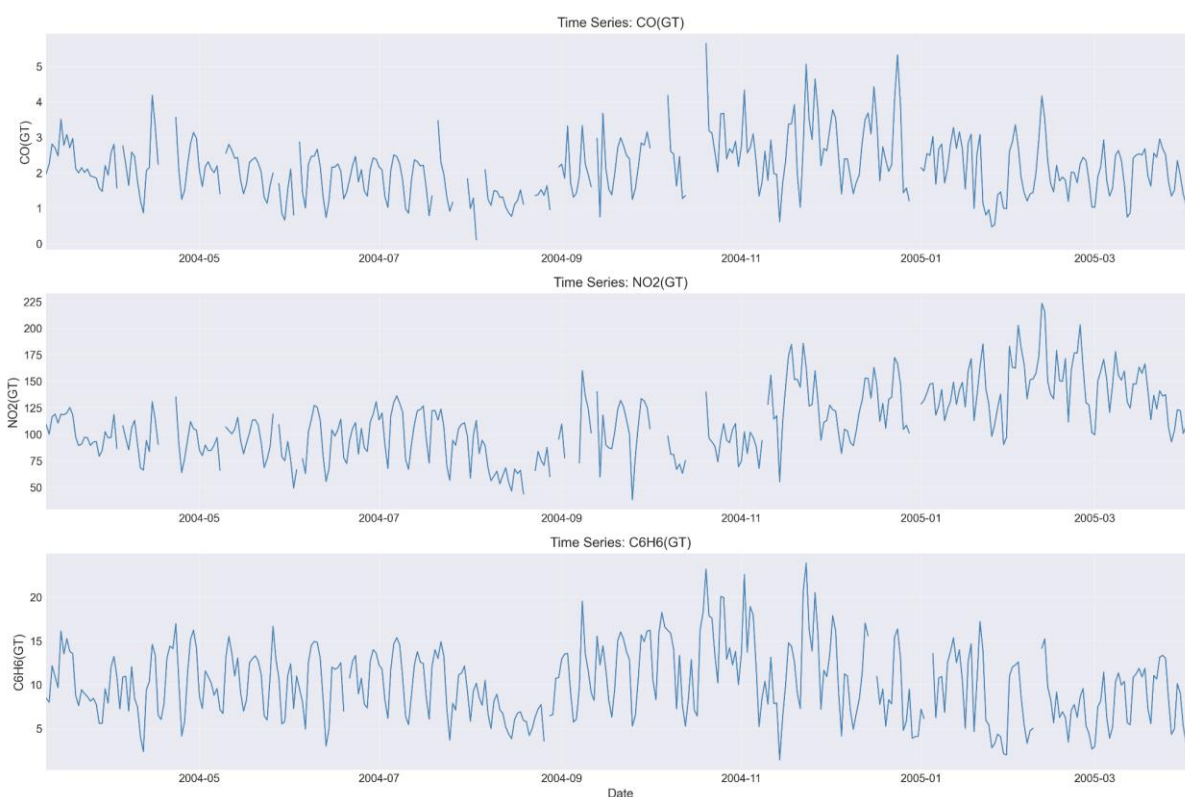
Βλέπε Παράρτημα Β για την πηγή της εικόνας στον φάκελο figures/ του αποθετηρίου.

Επιπλέον, εντοπίστηκαν ισχυρές συσχετίσεις μεταξύ των διαφορετικών ρύπων, γεγονός που αντικατοπτρίζει την κοινή τους προέλευση από τις οδικές μεταφορές. Συγκεκριμένα, οι μεταβλητές C6H6(GT) και NOx(GT) παρουσιάζουν υψηλή θετική συσχέτιση με τη μεταβλητή CO(GT), υποδηλώνοντας ότι οι συγκεντρώσεις των διαφόρων ρύπων αυξάνονται και μειώνονται συγχρόνως. Οι περιβαλλοντικές μεταβλητές (θερμοκρασία, υγρασία) παρουσιάζουν πιο σύνθετες

σχέσεις με τους ρύπους, αντανakλώντας την επίδραση των μετεωρολογικών συνθηκών στη διασπορά και συσσώρευση των ατμοσφαιρικών ρύπων.

4.1.4 Ανάλυση Χρονοσειρών

Η οπτικοποίηση των χρονοσειρών αποκάλυψε πολλαπλά χρονικά μοτίβα στα δεδομένα. Τα ημερήσια μοτίβα είναι ιδιαίτερα έντονα, με σαφείς κορυφώσεις κατά τις ώρες αιχμής κυκλοφοριακής κίνησης. Η ανάλυση ανέδειξε, επίσης, εβδομαδιαία μοτίβα, με διαφορετική συμπεριφορά μεταξύ καθημερινών και Σαββατοκύριακων, καθώς και εποχιακά μοτίβα που σχετίζονται με τις μεταβολές της θερμοκρασίας και τις συνθήκες διασποράς.



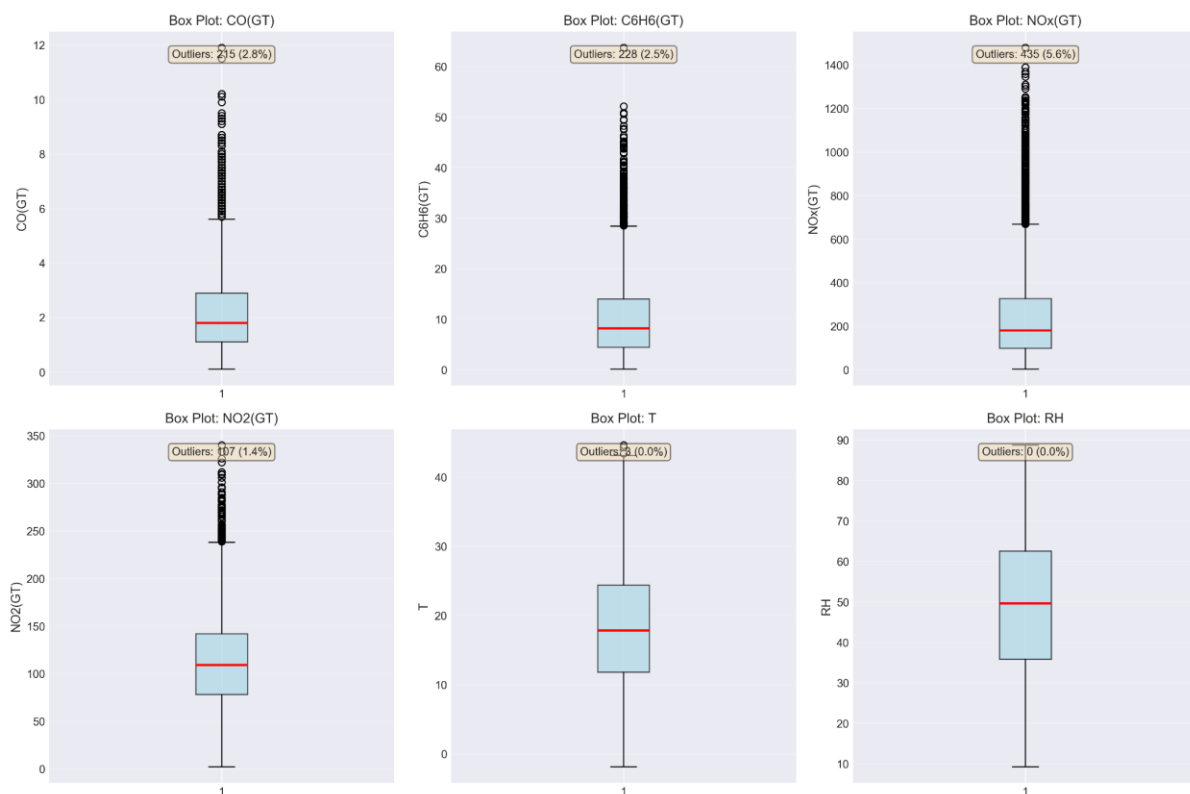
Εικόνα 4.5: Χρονοσειρές ατμοσφαιρικών ρύπων

Βλέπε Παράρτημα Β για την πηγή της εικόνας στον φάκελο figures/ του αποθετηρίου.

Η παρατήρηση αυτών των πολλαπλών χρονικών μοτίβων αποτέλεσε κρίσιμη πληροφορία για την δόμηση του σχεδιασμού χαρακτηριστικών, καθώς υπέδειξε την ανάγκη δημιουργίας χαρακτηριστικών που καταγράφουν τόσο τις βραχυπρόθεσμες όσο και τις μακροπρόθεσμες χρονικές εξαρτήσεις.

4.1.5 Ανίχνευση και Αντιμετώπιση Ακραίων Τιμών

Η ανίχνευση ακραίων τιμών (outliers) πραγματοποιήθηκε με τη μέθοδο του Διαστήματος Τεταρτημορίων (Interquartile Range - IQR), η οποία είναι ανθεκτική σε ακραίες τιμές και δεν προϋποθέτει κανονική κατανομή των δεδομένων [4]. Σύμφωνα με τη μέθοδο αυτή, μια παρατήρηση χαρακτηρίζεται ως ακραία τιμή εάν βρίσκεται εκτός του διαστήματος $[Q1 - 1,5 \times IQR, Q3 + 1,5 \times IQR]$, όπου Q1 και Q3 είναι το πρώτο και τρίτο τεταρτημόριο αντίστοιχα, και $IQR = Q3 - Q1$.



Εικόνα 4.6: Διαγράμματα θηκογραμμάτων για ανίχνευση ακραίων τιμών

Βλέπε Παράρτημα Β για την πηγή της εικόνας στον φάκελο figures/ του αποθετηρίου.

Η ανάλυση αποκάλυψε ότι περίπου 2,8% των παρατηρήσεων στη μεταβλητή CO(GT) αποτελούν ακραίες τιμές σύμφωνα με το κριτήριο IQR. Για την αντιμετώπισή τους επιλέχθηκε η μέθοδος της αποκοπής (clipping), η οποία αντικαθιστά τις ακραίες τιμές με τα αντίστοιχα όρια του διαστήματος $[Q1 - 1,5 \times IQR, Q3 + 1,5 \times IQR]$. Η επιλογή αυτή προτιμήθηκε έναντι της απλής αφαίρεσης των ακραίων τιμών, διότι διατηρεί τη χρονική συνέχεια της χρονοσειράς, χωρίς να

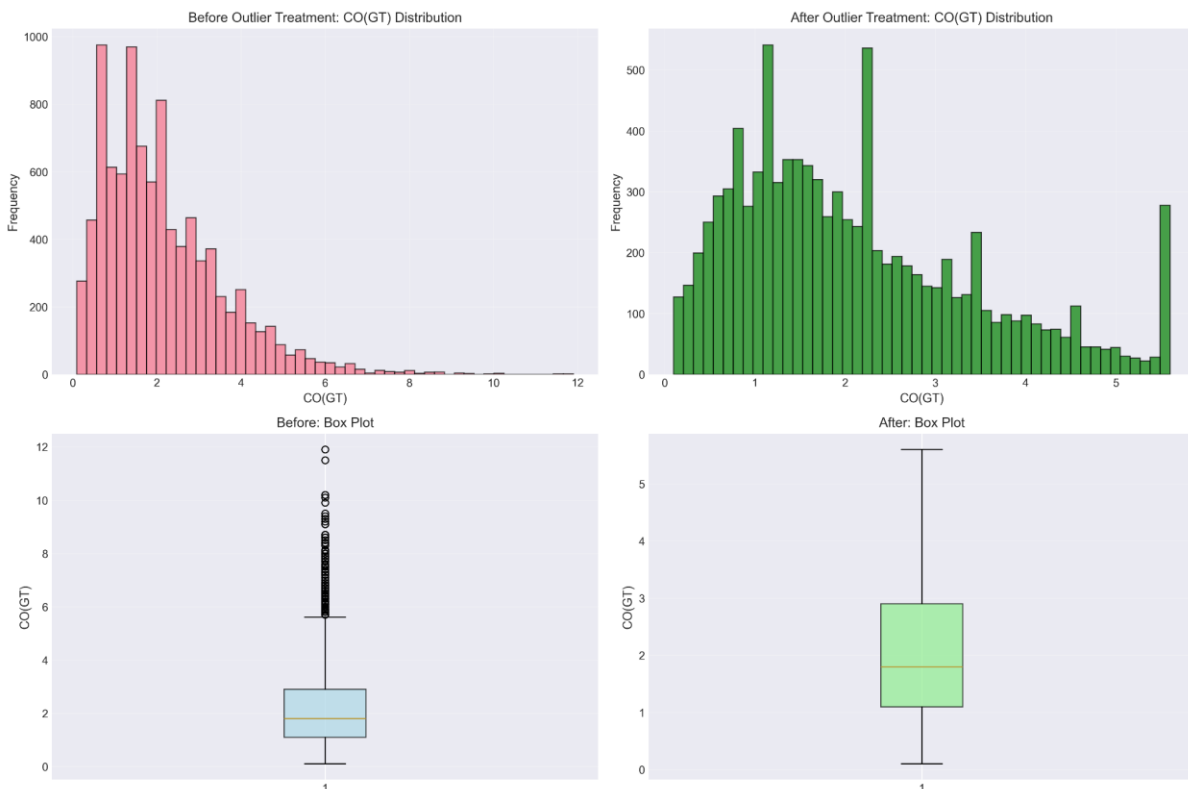
δημιουργεί κενά στα δεδομένα. Ωστόσο, κρίνεται απαραίτητο να σημειωθεί, πως σε περιβαλλοντικά δεδομένα, οι “ακραίες” τιμές μπορεί να είναι πραγματικά επεισόδια ρύπανσης και όχι σφάλματα.

4.1.6 Κανονικοποίηση

Για την κανονικοποίηση των χαρακτηριστικών επιλέχθηκε η μέθοδος MinMax Scaling, η οποία μετασχηματίζει τις τιμές κάθε χαρακτηριστικού στο διάστημα [0, 1] σύμφωνα με τον τύπο:

$$x_{norm} = \frac{x - x_{min}}{x_{max} - x_{min}}$$

Η επιλογή της κανονικοποίησης MinMax βασίστηκε σε δύο κύριους λόγους. Πρώτον, διατηρεί τη σχετική κατανομή των δεδομένων εντός κάθε χαρακτηριστικού, διευκολύνοντας την ερμηνεία των αποτελεσμάτων. Δεύτερον, είναι απαραίτητη για τα μοντέλα βαθιάς μάθησης (LSTM, GRU) που είναι ευαίσθητα στην κλίμακα των εισόδων [11], [12]. Ο κανονικοποιητής (scaler) προσαρμόστηκε αποκλειστικά στα δεδομένα εκπαίδευσης και εφαρμόστηκε στη συνέχεια στα σύνολα επικύρωσης και δοκιμής, αποφεύγοντας τη διαρροή πληροφορίας (data leakage) [4].

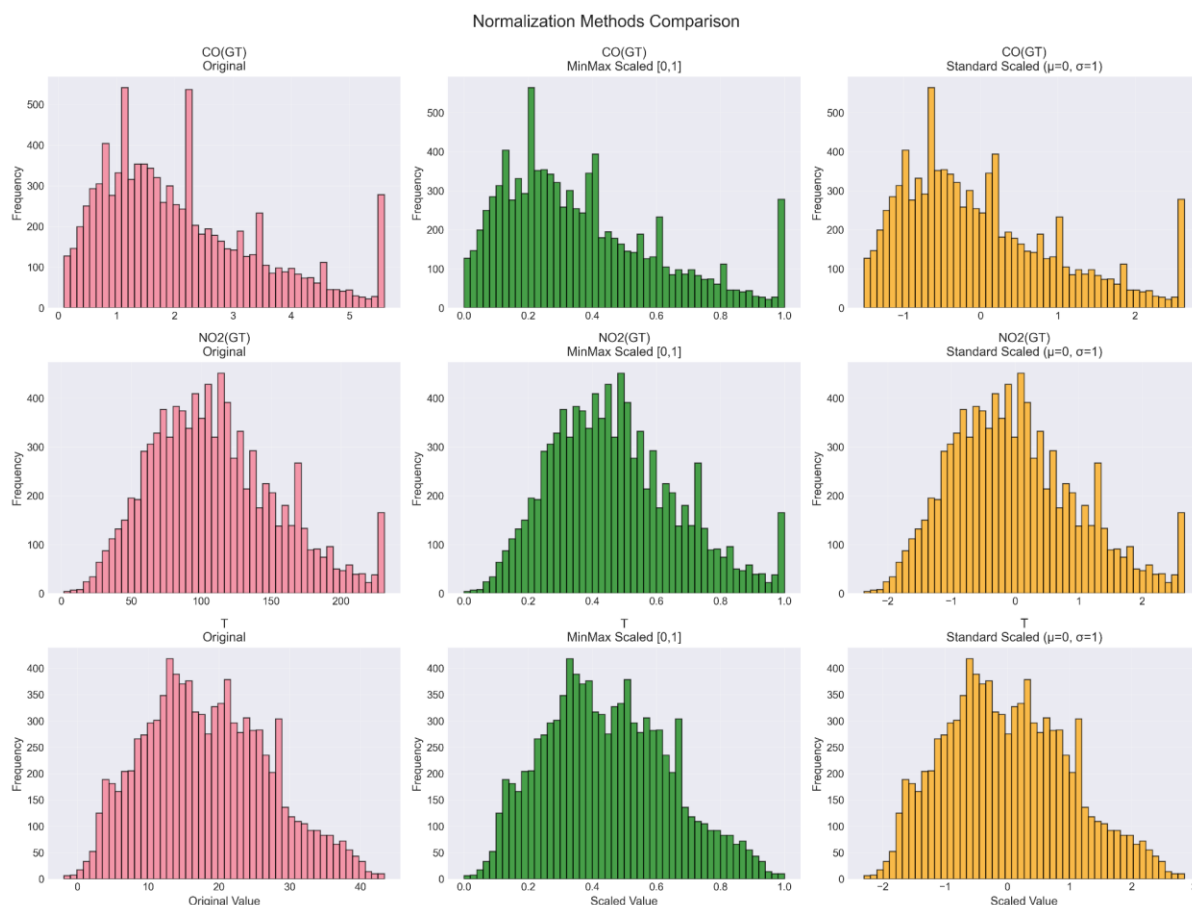


Εικόνα 4.7: Αντιμετώπιση ακραίων τιμών πριν και μετά

Βλέπε Παράρτημα Β για την πηγή της εικόνας στον φάκελο figures/ του αποθετηρίου.

4.1.7 Χρονικός Διαχωρισμός Δεδομένων

Ο διαχωρισμός των δεδομένων σε σύνολα εκπαίδευσης, επικύρωσης και δοκιμής πραγματοποιήθηκε με χρονολογική σειρά, διατηρώντας τη χρονική δομή της χρονοσειράς. Η αναλογία διαχωρισμού καθορίστηκε σε 70% για εκπαίδευση, 15% για επικύρωση και 15% για δοκιμή. Αυτή η προσέγγιση είναι κρίσιμη για την αξιόπιστη αξιολόγηση μοντέλων πρόβλεψης χρονοσειρών, καθώς αποτρέπει τη διαρροή μελλοντικής πληροφορίας στο σύνολο εκπαίδευσης [6].



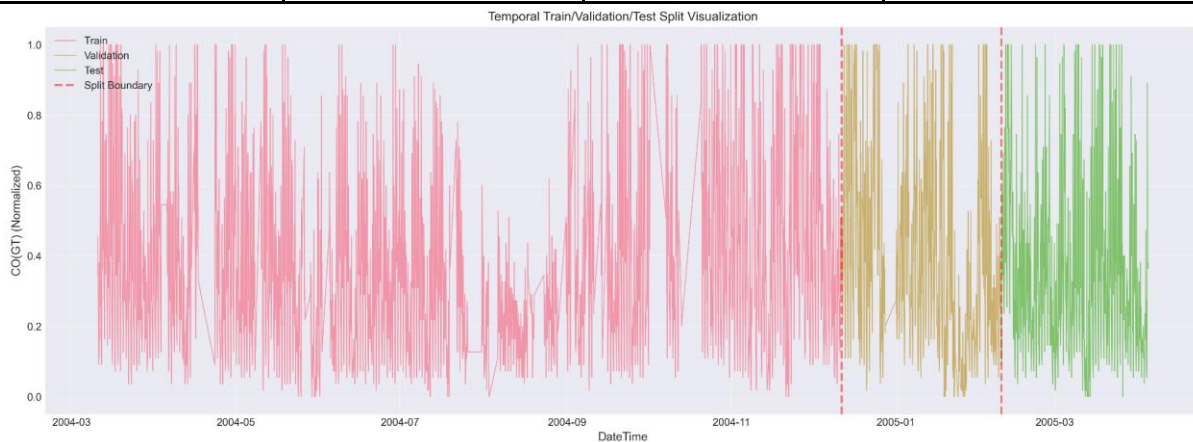
Εικόνα 4.9: Κανονικοποίηση δεδομένων

Βλέπε Παράρτημα Β για την πηγή της εικόνας στον φάκελο figures/ του αποθετηρίου.

Ο χρονικός διαχωρισμός απέδωσε τα ακόλουθα σύνολα δεδομένων:

Πίνακας 4.1: Σύνολα Δεδομένων έπειτα από Διαχωρισμό

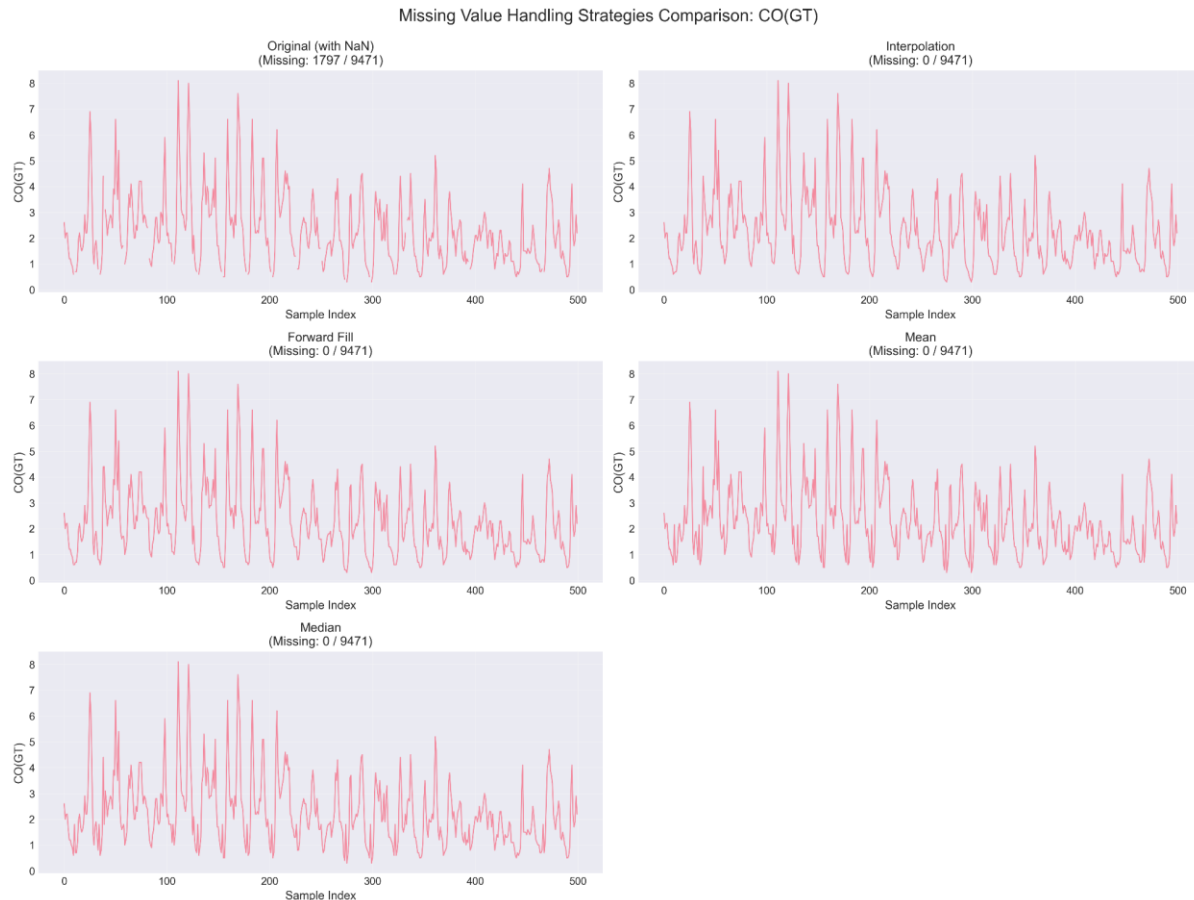
Σύνολο	Δείγματα	Ποσοστό	Χρονική Περίοδος
Εκπαίδευση (Train)	6.629	70,0%	10 Μαρτίου 2004 έως ~31 Δεκεμβρίου 2004
Επικύρωση (Validation)	1.421	15,0%	Επόμενο 15%, δηλ. έως ~15 Φεβρουαρίου 2005
Δοκιμή (Test)	1.421	15,0%	Τελευταίο 15%, δηλ έως 5 Απριλίου 2005



Εικόνα 4.10: Χρονικός διαχωρισμός δεδομένων σε train/validation/test

Βλέπε Παράρτημα Β για την πηγή της εικόνας στον φάκελο figures/ του αποθετηρίου.

Η επιλογή της αναλογίας 70-15-15 εξασφαλίζει επαρκή δεδομένα για την εκπαίδευση των μοντέλων, ενώ παράλληλα διατηρεί αντιπροσωπευτικά σύνολα για τη ρύθμιση υπερπαραμέτρων (επικύρωση) και την τελική αξιολόγηση (δοκιμή).



Εικόνα 4.11: Σύνοψη προεπεξεργασίας

Βλέπε Παράρτημα Β για την πηγή της εικόνας στον φάκελο figures/ του αποθετηρίου.

4.2 Σχεδίαση Χαρακτηριστικών (*Feature Engineering*)

Η σχεδίαση χαρακτηριστικών αποτελεί κρίσιμη φάση στην ανάπτυξη μοντέλων μηχανικής μάθησης, αφού ο μετασχηματισμός και η επαύξηση των δεδομένων μπορούν να βελτιώσουν σημαντικά την προγνωστική ικανότητα των αλγορίθμων [4], [6]. Στο πλαίσιο της πρόβλεψης χρονοσειρών, η σχεδίαση χαρακτηριστικών επικεντρώνεται στην αποτύπωση των χρονικών εξαρτήσεων και μοτίβων που υπάρχουν στα δεδομένα.

Διευκρινίζεται ότι η δημιουργία χαρακτηριστικών πραγματοποιήθηκε πριν τον χρονικό διαχωρισμό των δεδομένων σε σύνολα εκπαίδευσης, επικύρωσης και δοκιμής. Η σχεδίαση των χαρακτηριστικών διασφαλίζει ότι δεν υπάρχει διαρροή πληροφορίας (data leakage) στο μέλλον.

4.2.1 Χαρακτηριστικά Χρονικής Υστέρησης (Lag Features)

Τα χαρακτηριστικά χρονικής υστέρησης (lag features) αναπαριστούν προηγούμενες τιμές μιας μεταβλητής σε καθορισμένα χρονικά σημεία στο παρελθόν. Για την παρούσα ανάλυση δημιουργήθηκαν χαρακτηριστικά υστέρησης για τρεις χρονικούς ορίζοντες:

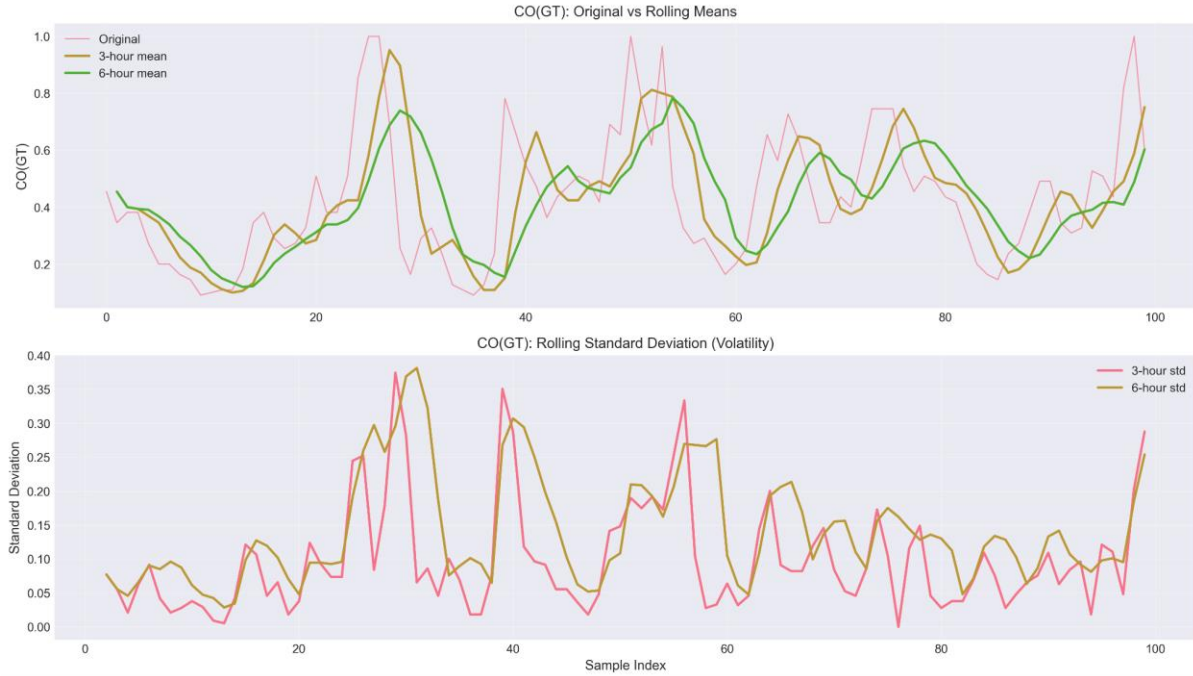
Η υστέρηση 1 ώρας (lag-1) καταγράφει την τιμή της αμέσως προηγούμενης ώρας, αποτυπώνοντας τη βραχυπρόθεσμη αυτοσυσχέτιση της χρονοσειράς. Η υστέρηση 2 ωρών (lag-2) προσθέτει πρόσθετη πληροφορία για την πρόσφατη τάση εξέλιξης. Η υστέρηση 24 ωρών (lag-24) καταγράφει την τιμή της ίδιας ώρας της προηγούμενης ημέρας, αποτυπώνοντας το ημερήσιο κύκλο που είναι χαρακτηριστικός των δεδομένων ατμοσφαιρικής ρύπανσης.

Τα χαρακτηριστικά υστέρησης δημιουργήθηκαν για πέντε βασικές μεταβλητές: CO(GT), NO₂(GT), C₆H₆(GT), T (θερμοκρασία) και RH (σχετική υγρασία). Αυτό απέδωσε συνολικά 15 νέα χαρακτηριστικά υστέρησης (5 μεταβλητές $\times$ 3 χρονικοί ορίζοντες).

4.2.2 Χαρακτηριστικά Κυλιόμενου Παραθύρου (Rolling Window Features)

Τα χαρακτηριστικά κυλιόμενου παραθύρου (rolling window features) υπολογίζουν στατιστικά μέτρα σε κινούμενα χρονικά παράθυρα, αποτυπώνοντας τις τάσεις και τη μεταβλητότητα της χρονοσειράς σε διαφορετικές χρονικές κλίμακες [4]. Για κάθε παράθυρο υπολογίστηκαν τέσσερα στατιστικά μέτρα: ο μέσος όρος (mean) για την κεντρική τάση, η τυπική απόκλιση (std) για τη μεταβλητότητα, η ελάχιστη τιμή (min) και η μέγιστη τιμή (max) για το εύρος.

Τα μεγέθη παραθύρων επιλέχθηκαν ως εξής: παράθυρο 3 ωρών για βραχυπρόθεσμες τάσεις, παράθυρο 6 ωρών για μεσοπρόθεσμες τάσεις, και παράθυρο 24 ωρών για ημερήσιες τάσεις. Εφαρμόζοντας 4 στατιστικά μέτρα $\times$ 3 παράθυρα $\times$ 5 μεταβλητές, δημιουργήθηκαν 60 νέα χαρακτηριστικά κυλιόμενου παραθύρου.



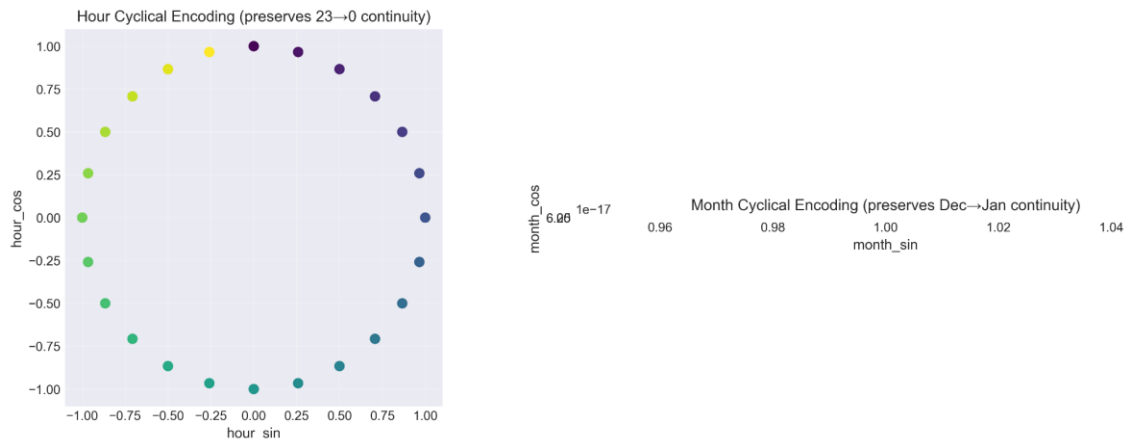
Εικόνα 4.12: Οπτικοποίηση χαρακτηριστικών κυλιόμενου παραθύρου

Ιδιαίτερη προσοχή δόθηκε στην αποφυγή διαρροής δεδομένων (data leakage). Τα χαρακτηριστικά κυλιόμενου παραθύρου μετατοπίστηκαν κατά μία θέση (shift by 1), ώστε ο υπολογισμός τους να βασίζεται αποκλειστικά σε παρελθοντικές τιμές και όχι στην τρέχουσα τιμή. Αυτό διασφαλίζει ότι, κατά την πρόβλεψη στο χρόνο $t+1$, χρησιμοποιούνται μόνο πληροφορίες διαθέσιμες μέχρι τον χρόνο t .

4.2.3 Κυκλική Κωδικοποίηση Χρόνου (Cyclical Time Encoding)

Τα χρονικά χαρακτηριστικά όπως η ώρα της ημέρας και ο μήνας του έτους είναι εγγενώς κυκλικά: η ώρα 23:00 είναι πλησιέστερη στην ώρα 00:00 παρά στην ώρα 12:00. Για την αποτύπωση αυτής της κυκλικότητας, τα χρονικά χαρακτηριστικά κωδικοποιήθηκαν με χρήση τριγωνομετρικών συναρτήσεων [4]:

$$hour_{sin} = \sin\left(\frac{2\pi \times hour}{24}\right) \quad hour_{cos} = \cos\left(\frac{2\pi \times hour}{24}\right)$$



Εικόνα 4.13: Κυκλική κωδικοποίηση χρονικών χαρακτηριστικών

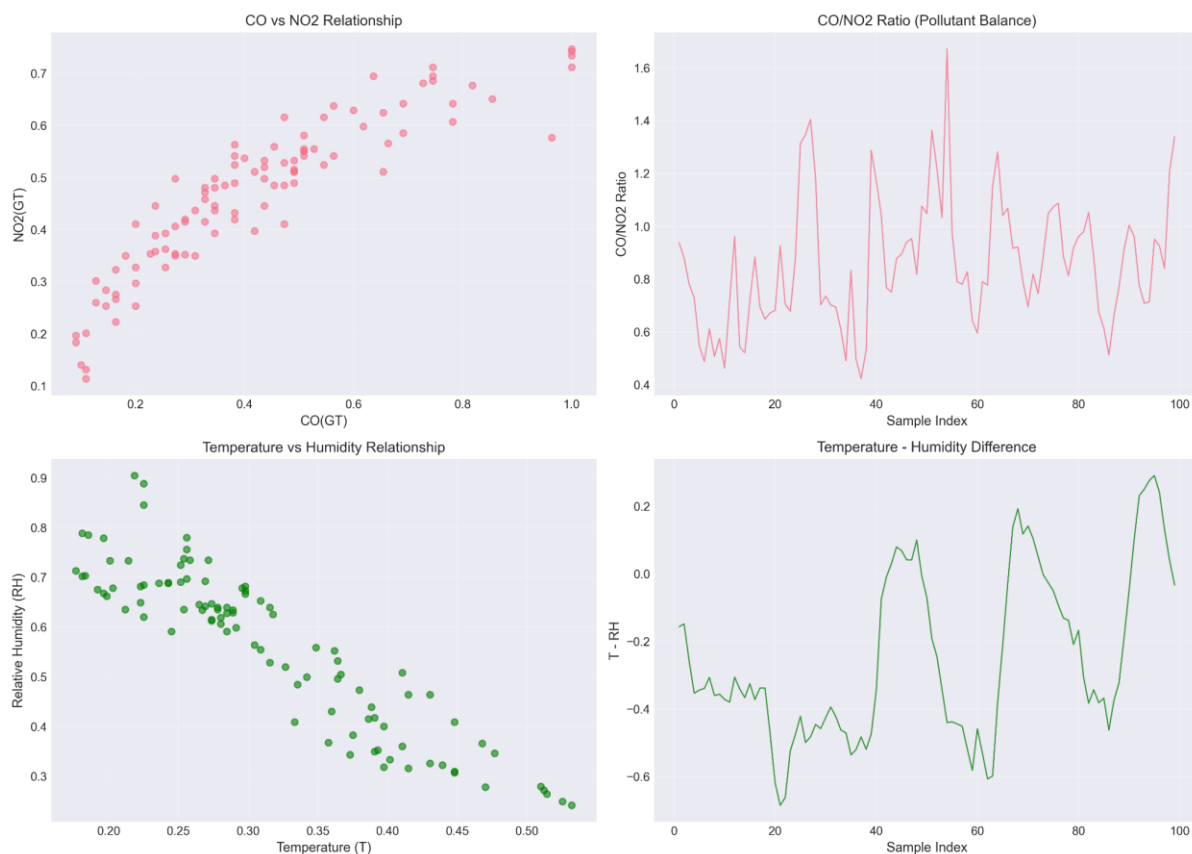
Βλέπε Παράρτημα Β για την πηγή της εικόνας στον φάκελο figures/ του αποθετηρίου.

Η ίδια κωδικοποίηση εφαρμόστηκε στον μήνα (κύκλος 12 μηνών), στην ημέρα της εβδομάδας (κύκλος 7 ημερών) και στην ημέρα του μήνα. Επιπλέον, δημιουργήθηκαν απλά χρονικά χαρακτηριστικά: η ώρα (0-23), η ημέρα της εβδομάδας (0-6), ο μήνας (1-12) και η εποχή (1-4). Συνολικά δημιουργήθηκαν 12 χρονικά χαρακτηριστικά.

4.2.4 Χαρακτηριστικά Αλληλεπίδρασης (Interaction Features)

Τα χαρακτηριστικά αλληλεπίδρασης αποτυπώνουν τις σχέσεις μεταξύ διαφορετικών μεταβλητών. Στην παρούσα ανάλυση δημιουργήθηκαν δύο τύποι αλληλεπιδράσεων:

Για τους ατμοσφαιρικούς ρύπους υπολογίστηκε ο λόγος $CO(GT)/NO_2(GT)$, ο οποίος αποτυπώνει την ισορροπία μεταξύ των δύο κύριων ρύπων κυκλοφοριακής προέλευσης, καθώς και η διαφορά $CO(GT) - NO_2(GT)$. Για τις περιβαλλοντικές μεταβλητές υπολογίστηκε ο λόγος T/RH και η διαφορά $T - RH$, οι οποίες αποτυπώνουν τη σχέση μεταξύ θερμοκρασίας και υγρασίας.



Εικόνα 4.14: Οπτικοποίηση χαρακτηριστικών αλληλεπίδρασης

Βλέπε Παράρτημα Β για την πηγή της εικόνας στον φάκελο figures/ του αποθετηρίου.

Όμοια με τα χαρακτηριστικά κυλιόμενου παραθύρου, τα χαρακτηριστικά αλληλεπίδρασης μετατοπίστηκαν κατά μία χρονική μονάδα για την αποφυγή διαρροής δεδομένων. Συνολικά δημιουργήθηκαν 4 χαρακτηριστικά αλληλεπίδρασης.

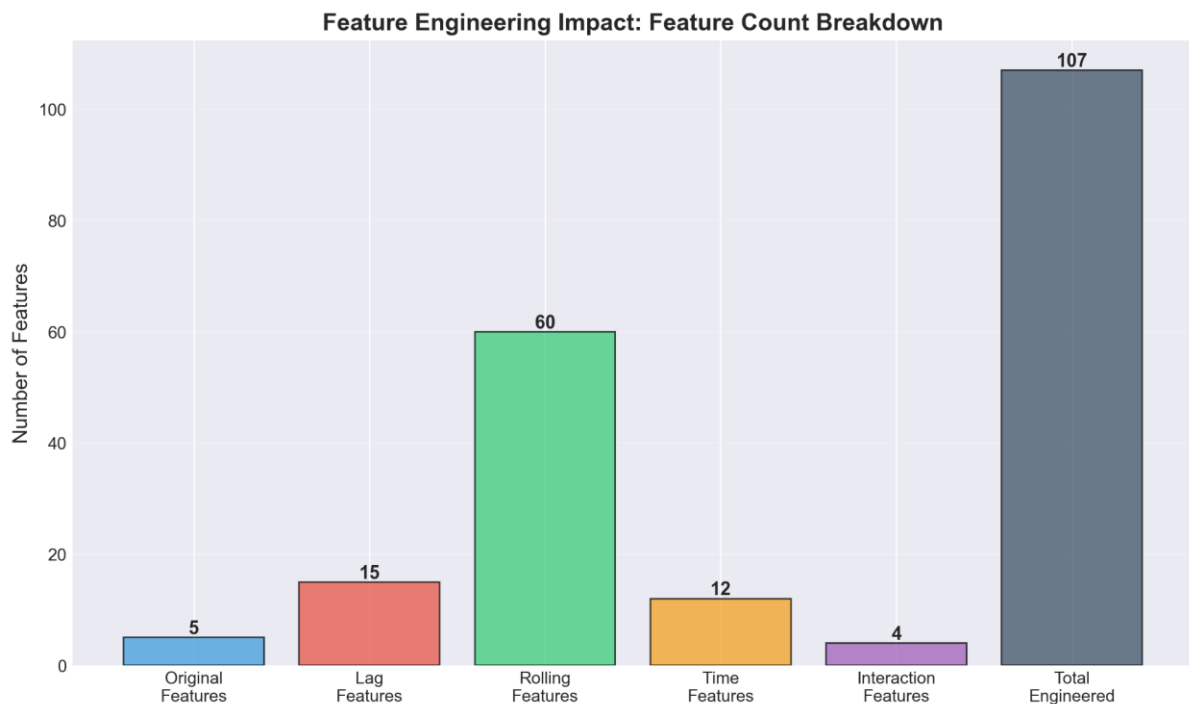
4.2.5 Τελικό Σύνολο Χαρακτηριστικών

Η διαδικασία σχεδιασμού χαρακτηριστικών μετασχημάτισε το αρχικό σύνολο 12 χαρακτηριστικών σε ένα εμπλουτισμένο σύνολο 107 χαρακτηριστικών:

Πίνακας 4.2: Σχεδίαση Χαρακτηριστικών

Κατηγορία Χαρακτηριστικών	Πλήθος
Αρχικά χαρακτηριστικά	12

Χαρακτηριστικά υστέρησης (Lag)	15
Χαρακτηριστικά κυλιόμενου παραθύρου (Rolling)	60
Χρονικά χαρακτηριστικά	12
Χαρακτηριστικά αλληλεπίδρασης	4
Άλλα (DateTime κ.λπ.)	4
Σύνολο	107



Εικόνα 4.15: Ανάλυση τελικών χαρακτηριστικών ανά κατηγορία

Βλέπε Παράρτημα Β για την πηγή της εικόνας στον φάκελο figures/ του αποθετηρίου.

Η επέκταση του χώρου χαρακτηριστικών από 12 σε 107 διαστάσεις αυξάνει σημαντικά την εκφραστική ικανότητα των μοντέλων, επιτρέποντάς τους να αποτυπώσουν σύνθετες χρονικές εξαρτήσεις και μοτίβα που δεν είναι άμεσα ορατά στα αρχικά δεδομένα.

4.3 Μοντελοποίηση (Modeling)

Η φάση της μοντελοποίησης, σύμφωνα με τη μεθοδολογία CRISP-DM [25], περιλαμβάνει την επιλογή, εφαρμογή και βαθμονόμηση τεχνικών μοντελοποίησης. Στην παρούσα εργασία υλοποιήθηκαν και συγκρίθηκαν έξι αλγόριθμοι μηχανικής μάθησης, οργανωμένοι σε τρεις κατηγορίες: μοντέλα βάσης (baseline), προηγμένα μοντέλα ενισχυμένης μάθησης (gradient boosting), και μοντέλα βαθιάς μάθησης (deep learning).

Όσον αφορά τη ρύθμιση υπερπαραμέτρων, τα μοντέλα εκπαιδεύτηκαν με τις προεπιλεγμένες τιμές της βιβλιοθήκης Scikit-learn [10], χωρίς συστηματική βελτιστοποίηση (π.χ. Grid Search ή Random Search). Η επιλογή αυτή αποτελεί συνειδητή απόφαση λόγω του περιορισμένου μεγέθους δεδομένων και των διαθέσιμων υπολογιστικών πόρων, αλλά αναγνωρίζεται ως περιορισμός που ενδέχεται να επηρεάζει τη συγκριτική αξιολόγηση — ιδίως για αλγορίθμους όπως ο XGBoost [8] που είναι ευαίσθητοι στη ρύθμιση υπερπαραμέτρων, όπως συζητήθηκε στην Ενότητα 4.3.2.

4.3.1 Μοντέλα Βάσης (Baseline Models)

Τα μοντέλα βάσης παρέχουν ένα σημείο αναφοράς για την αξιολόγηση πιο σύνθετων αλγορίθμων. Η υπέρβαση της απόδοσης των μοντέλων αυτών είναι απαραίτητη προϋπόθεση για να δικαιολογηθεί η πρόσθετη πολυπλοκότητα των προηγμένων μοντέλων [5].

- Γραμμική Παλινδρόμηση (Linear Regression)

Η Γραμμική Παλινδρόμηση αποτελεί το απλούστερο μοντέλο πρόβλεψης και υποθέτει γραμμική σχέση μεταξύ των ανεξάρτητων μεταβλητών και της μεταβλητής-στόχου [5].³

Τα αποτελέσματα της Γραμμικής Παλινδρόμησης στο σύνολο δοκιμής ήταν: MAE = 0,118 mg/m³, RMSE = 0,141 mg/m³ και R² = 0,609. Αξιοσημείωτο είναι ότι το μοντέλο παρουσίασε σημαντική υπερπροσαρμογή στο σύνολο εκπαίδευσης (Train R² = 0,986), η οποία δεν μεταφέρθηκε στο σύνολο επικύρωσης (Validation R² = -2,555). Η συμπεριφορά αυτή είναι ασυνήθιστη για τη Γραμμική Παλινδρόμηση, η οποία ως χαμηλής πολυπλοκότητας μοντέλο σπάνια υπερπροσαρμόζεται. Η πιθανότερη αιτία είναι η πολυσυγγραμμικότητα (multicollinearity) μεταξύ

³ Παρά την απλότητά της, η Γραμμική Παλινδρόμηση μπορεί να αποδώσει ικανοποιητικά όταν οι σχέσεις στα δεδομένα είναι πράγματι γραμμικές ή κατά προσέγγιση γραμμικές.

των χαρακτηριστικών: τα χαρακτηριστικά υστέρησης και τα κυλιόμενα στατιστικά εκ κατασκευής παρουσιάζουν υψηλή μεταξύ τους συσχέτιση, γεγονός που αποσταθεροποιεί τις εκτιμήσεις των συντελεστών και οδηγεί σε ψευδώς υψηλό R^2 στο σύνολο εκπαίδευσης, χωρίς πραγματική προβλεπτική ικανότητα σε νέα δεδομένα. Η χρήση κανονικοποιημένων παραλλαγών (Ridge, Lasso) θα μπορούσε να μετριάσει αυτό το φαινόμενο, αποτελεί όμως αντικείμενο μελλοντικής διερεύνησης.

- Random Forest

Ο αλγόριθμος Random Forest, που προτάθηκε από τον Breiman [7], αποτελεί μέθοδο συναθροιστικής μάθησης (ensemble learning) που συνδυάζει τις προβλέψεις πολλαπλών δέντρων αποφάσεων.⁴

Ο Random Forest επέδειξε εξαιρετική απόδοση, επιτυγχάνοντας τα καλύτερα αποτελέσματα μεταξύ όλων των μοντέλων: MAE = 0,092 mg/m³, RMSE = 0,112 mg/m³ και $R^2 = 0,750$ στο σύνολο δοκιμής. Η απόδοση στα σύνολα εκπαίδευσης και επικύρωσης ήταν επίσης υψηλή (Train $R^2 = 0,994$, Validation $R^2 = 0,714$), υποδεικνύοντας καλή ισορροπία μεταξύ προσαρμογής και γενίκευσης.

4.3.2 Προηγμένα Μοντέλα Ενισχυμένης Μάθησης (Advanced ML Models)

Τα μοντέλα Ενισχυτικής Μάθησης μέσω Βαθμίδωσης (Gradient Boosting) κατασκευάζουν ακολουθιακά αδύναμους εκτιμητές, με κάθε νέο εκτιμητή να επικεντρώνεται στα σφάλματα των προηγούμενων [9].

- XGBoost

Ο αλγόριθμος XGBoost (eXtreme Gradient Boosting), που αναπτύχθηκε από τους Chen και Guestrin [8], αποτελεί μια βελτιστοποιημένη υλοποίηση του Gradient Boosting με προηγμένες τεχνικές κανονικοποίησης και αποδοτικούς υπολογιστικούς αλγορίθμους. Το μοντέλο ενσωματώνει L1 και L2 κανονικοποίηση για την αποφυγή υπερπροσαρμογής, καθώς και τεχνικές δειγματοληψίας χαρακτηριστικών και παρατηρήσεων.

⁴ Κάθε δέντρο εκπαιδεύεται σε τυχαίο υποσύνολο των δεδομένων (bagging) και χρησιμοποιεί τυχαίο υποσύνολο χαρακτηριστικών σε κάθε διαχωρισμό, μειώνοντας τη διακύμανση και βελτιώνοντας τη γενίκευση [7].

Τα αποτελέσματα του XGBoost στο σύνολο δοκιμής ήταν: $MAE = 0,130 \text{ mg/m}^3$, $RMSE = 0,151 \text{ mg/m}^3$ και $R^2 = 0,550$. Παρότι ο XGBoost συνήθως ισούται ή υπερτερεί του Random Forest στη βιβλιογραφία [8], στην παρούσα περίπτωση η απόδοσή του ήταν χαμηλότερη. Η πιθανότερη αιτία είναι η χρήση προεπιλεγμένων υπερπαραμέτρων χωρίς εκτεταμένη βελτιστοποίηση: ο XGBoost είναι ιδιαίτερα ευαίσθητος στη ρύθμιση παραμέτρων όπως ο ρυθμός μάθησης (learning rate), το μέγιστο βάθος δέντρου (max depth) και οι παράμετροι κανονικοποίησης, ενώ ο Random Forest λειτουργεί αποτελεσματικά με τις προεπιλεγμένες ρυθμίσεις του. Επιπλέον, ο μικρός όγκος δεδομένων εκπαίδευσης (890 δείγματα) περιορίζει το πλεονέκτημα της διαδοχικής ενίσχυσης (boosting), η οποία αποδίδει καλύτερα σε μεγαλύτερα σύνολα δεδομένων όπου μπορεί να διορθώνει σταδιακά πιο σύνθετα μοτίβα σφαλμάτων. Ωστόσο, το XGBoost απέδειξε σταθερότερη συμπεριφορά μεταξύ των συνόλων εκπαίδευσης και δοκιμής, υποδεικνύοντας χαμηλότερο κίνδυνο υπερπροσαρμογής.

- Gradient Boosting

Η κλασική υλοποίηση του Gradient Boosting, όπως περιγράφεται από τον Friedman [9], εκπαιδεύει διαδοχικά δέντρα αποφάσεων, με κάθε νέο δέντρο να προσαρμόζεται στα κατάλοιπα του προηγούμενου μοντέλου.

Τα αποτελέσματα του Gradient Boosting ήταν: $MAE = 0,134 \text{ mg/m}^3$, $RMSE = 0,157 \text{ mg/m}^3$ και $R^2 = 0,512$ στο σύνολο δοκιμής. Αξιοσημείωτη είναι η εξαιρετικά υψηλή απόδοση στο σύνολο εκπαίδευσης (Train $R^2 = 0,999$), η οποία υποδεικνύει τάση υπερπροσαρμογής παρά τη χρήση κανονικοποίησης.

4.3.3 Μοντέλα Βαθιάς Μηχανικής Μάθησης (Deep Learning)

Τα μοντέλα επαναλαμβανόμενων νευρωνικών δικτύων (Recurrent Neural Networks - RNN) είναι ειδικά σχεδιασμένα για την επεξεργασία ακολουθιακών δεδομένων όπως οι χρονοσειρές [11], [12].

- LSTM (Long Short-Term Memory)

Η αρχιτεκτονική LSTM, που προτάθηκε από τους Hochreiter και Schmidhuber [11], αντιμετωπίζει το πρόβλημα της εξασθένησης κλίσης (vanishing gradient) που περιορίζει τα παραδοσιακά RNN. Η LSTM χρησιμοποιεί μηχανισμούς πυλών (gates) για τον έλεγχο της ροής πληροφορίας, επιτρέποντας την εκμάθηση μακροπρόθεσμων εξαρτήσεων.

Για την εφαρμογή του LSTM, τα δεδομένα μετασχηματίστηκαν σε ακολουθίες 24 χρονικών βημάτων (sequence length = 24 ώρες), αντιστοιχώντας σε ένα πλήρες ημερήσιο κύκλο. Η αρχιτεκτονική περιελάμβανε ένα επίπεδο LSTM με 64 μονάδες, dropout 20% για κανονικοποίηση, και ένα πυκνό επίπεδο εξόδου.

Η εκπαίδευση πραγματοποιήθηκε για 50 εποχές (epochs) με μέγεθος παρτίδας (batch size) 32, βελτιστοποιητή Adam με ρυθμό μάθησης (learning rate) 0,001 και συνάρτηση απώλειας MSE. Εφαρμόστηκε πρόωμη διακοπή (early stopping) με παρακολούθηση της απώλειας επικύρωσης και υπομονή 10 εποχών.

Τα αποτελέσματα του LSTM ήταν απογοητευτικά: MAE = 0,188 mg/m³, RMSE = 0,239 mg/m³ και R² = -0,122 στο σύνολο δοκιμής. Ο αρνητικός συντελεστής R² υποδεικνύει ότι το μοντέλο αποδίδει χειρότερα ακόμη και από έναν απλό εκτιμητή μέσης τιμής.

- GRU (Gated Recurrent Unit)

Η αρχιτεκτονική GRU, που προτάθηκε από τους Cho και συνεργάτες [12], αποτελεί απλοποιημένη παραλλαγή της LSTM με λιγότερες παραμέτρους. Η GRU συνδυάζει την πύλη εισόδου και την πύλη λήθης σε μια ενιαία "πύλη ενημέρωσης" (update gate), μειώνοντας την υπολογιστική πολυπλοκότητα χωρίς σημαντική απώλεια εκφραστικότητας.

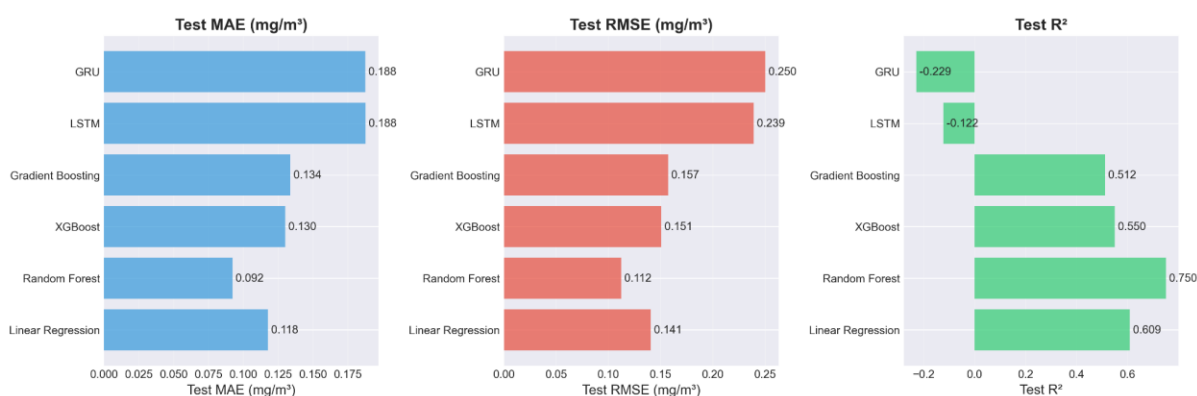
Η εκπαίδευση ακολούθησε τις ίδιες ρυθμίσεις με το LSTM (50 εποχές, batch size 32, Adam με learning rate 0,001, early stopping με υπομονή 10 εποχών), ώστε να διασφαλιστεί δίκαιη σύγκριση μεταξύ των δύο αρχιτεκτονικών.

Τα αποτελέσματα του GRU ήταν παρόμοια με αυτά του LSTM: MAE = 0,188 mg/m³, RMSE = 0,250 mg/m³ και R² = -0,229 στο σύνολο δοκιμής.

4.3.4 Ανάλυση Υπο-απόδοσης Μοντέλων Βαθιάς Μάθησης

Η σημαντικά χαμηλότερη απόδοση των μοντέλων βαθιάς μηχανικής μάθησης (LSTM, GRU) σε σύγκριση με τα μοντέλα δέντρων αποφάσεων αποδίδεται πρωτίστως στον περιορισμένο όγκο δεδομένων εκπαίδευσης. Μετά τη δημιουργία ακολουθιών 24 χρονικών βημάτων, το σύνολο εκπαίδευσης περιορίστηκε σε περίπου 866 ακολουθίες, αριθμός σημαντικά μικρότερος από τις τυπικές απαιτήσεις των νευρωνικών δικτύων [13], [14].

Τα μοντέλα βαθιάς μηχανικής μάθησης, τυπικά, απαιτούν χιλιάδες έως εκατομμύρια παραδείγματα για την αποτελεσματική εκμάθηση σύνθετων μοτίβων [14]. Αντίθετα, τα μοντέλα βασισμένα σε δέντρα αποφάσεων είναι πιο αποδοτικά στη χρήση δεδομένων (sample-efficient) και μπορούν να επιτύχουν εξαιρετική απόδοση ακόμη και με περιορισμένα δείγματα [5], [7].



Εικόνα 4.16: Σύγκριση απόδοσης μοντέλων

Βλέπε Παράρτημα Β για την πηγή της εικόνας στον φάκελο figures/ του αποθετηρίου.

4.3.5 Σύγκριση Μοντέλων

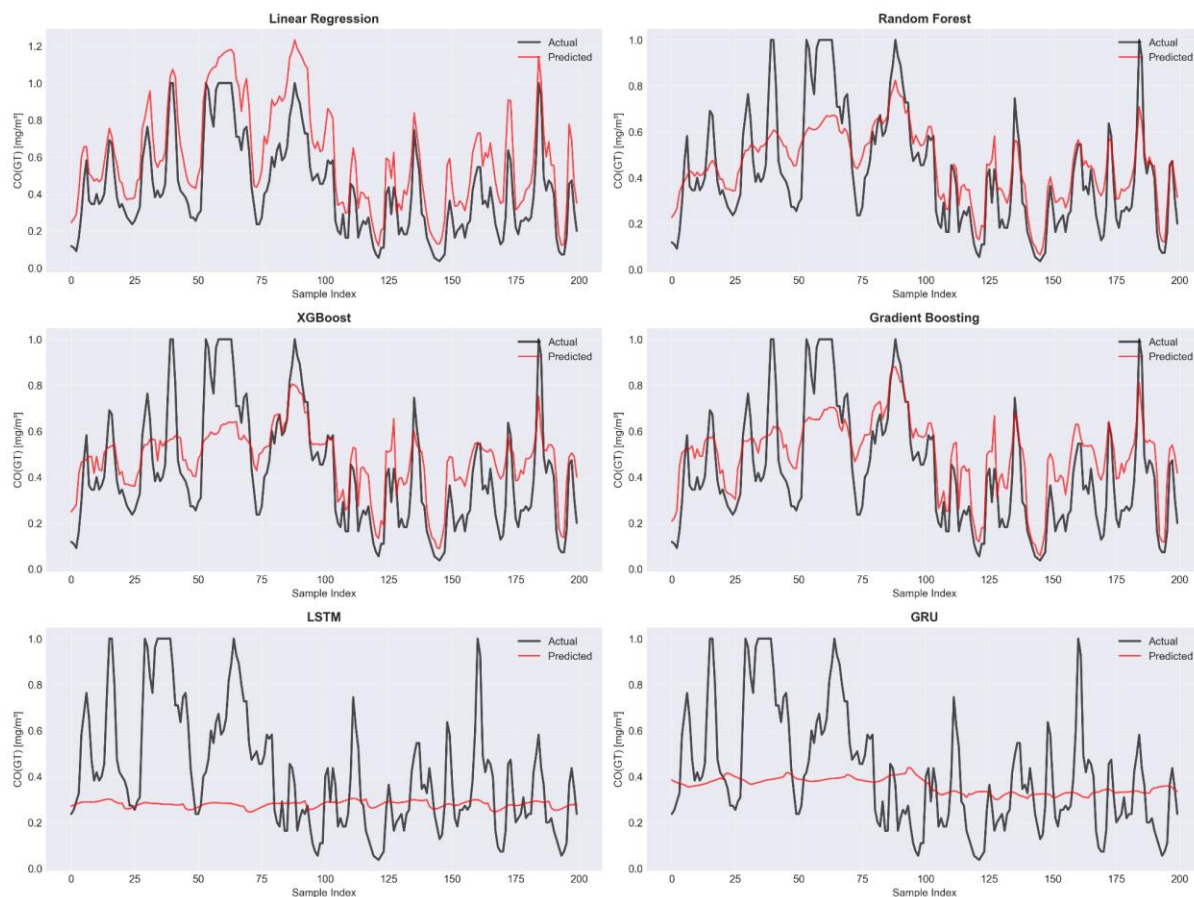
Ο παρακάτω πίνακας συνοψίζει την απόδοση όλων των μοντέλων στο σύνολο δοκιμής:

Πίνακας 4.3: Σύνοψη απόδοσης μοντέλων - Air Quality Dataset

Μοντέλο	Test MAE (mg/m³)	Test RMSE (mg/m³)	Test R²
Linear Regression	0,118	0,141	0,609
Random Forest	0,092	0,112	0,750

XGBoost	0,130	0,151	0,550
Gradient Boosting	0,134	0,157	0,512
LSTM	0,188	0,239	-0,122
GRU	0,188	0,250	-0,229

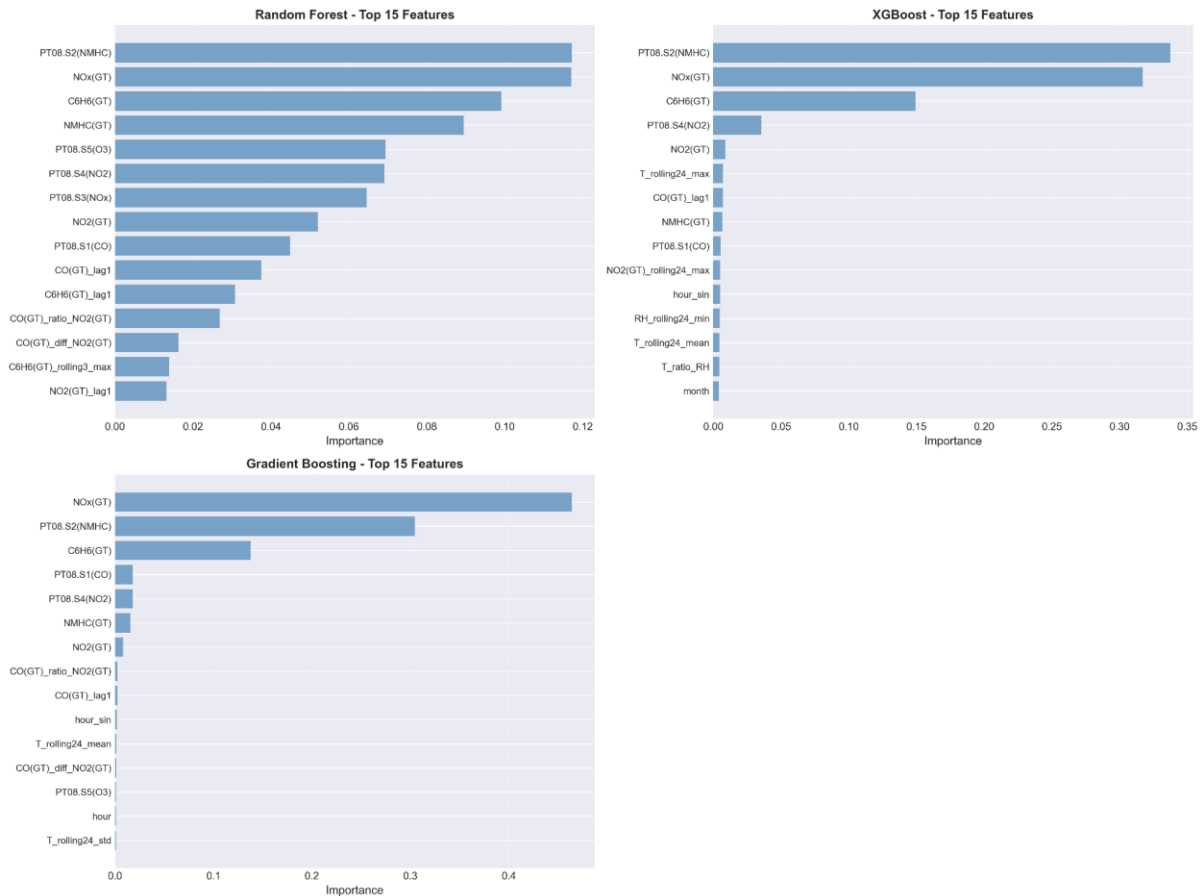
Ο Random Forest αναδείχθηκε ως το βέλτιστο μοντέλο, επιτυγχάνοντας τη χαμηλότερη τιμή MAE (0,092 mg/m³), τη χαμηλότερη τιμή RMSE (0,112 mg/m³) και τον υψηλότερο συντελεστή R² (0,750). Η απόδοση αυτή ικανοποιεί τα κριτήρια επιτυχίας που είχαν τεθεί: MAE < 0,7 mg/m³, RMSE < 1,0 mg/m³ και R² > 0,70.



Εικόνα 4.17: Πραγματικές vs προβλεπόμενες τιμές για τα μοντέλα

4.3.6 Ανάλυση Σημαντικότητας Χαρακτηριστικών

Η ανάλυση σημαντικότητας χαρακτηριστικών (feature importance) του Random Forest αποκάλυψε ότι τα χαρακτηριστικά με τη μεγαλύτερη συνεισφορά στην πρόβλεψη είναι τα χαρακτηριστικά υστέρησης, ιδιαίτερα η τιμή CO(GT)_lag1 (προηγούμενη ώρα). Αξίζει να σημειωθεί πως το Random Forest χρησιμοποιεί μέση μείωση των “ακαθαρσιών” (mean decrease in impurity). Αυτό επιβεβαιώνει την ισχυρή αυτοσυσχέτιση της χρονοσειράς και τη σημασία της βραχυπρόθεσμης ιστορίας για την πρόβλεψη.



Εικόνα 4.18: Σημαντικότητα χαρακτηριστικών (Feature Importance)

Βλέπε Παράρτημα Β για την πηγή της εικόνας στον φάκελο figures/ του αποθετηρίου.

4.4 Αξιολόγηση (Evaluation)

Η φάση της αξιολόγησης, σύμφωνα με τη μεθοδολογία CRISP-DM [25], περιλαμβάνει την ενδεδειγμένη εξέταση των μοντέλων σε σχέση με τους στόχους του έργου. Η αξιολόγηση επεκτείνεται

πέραν των απλών μεθόδων ελέγχου απόδοσης, περιλαμβάνοντας διασταυρούμενη επικύρωση, ανάλυση καμπυλών μάθησης και διαγνωστικά καταλοίπων.

4.4.1 Διασταυρούμενη Επικύρωση Χρονοσειρών (Temporal Cross-Validation)

Η παραδοσιακή k-fold διασταυρούμενη επικύρωση δεν είναι κατάλληλη για δεδομένα χρονοσειρών, καθώς η τυχαία ανάμειξη των δεδομένων παραβιάζει τη χρονική δομή και μπορεί να οδηγήσει σε υπεραισιόδοξες εκτιμήσεις απόδοσης [6]. Αντ' αυτού, εφαρμόστηκε η διασταυρούμενη επικύρωση χρονοσειρών (Time-Series Cross-Validation) με διευρυνόμενο παράθυρο εκπαίδευσης (expanding window approach).

Η μέθοδος υλοποιήθηκε με 5 τμήματα (folds), όπου σε κάθε τμήμα το παράθυρο εκπαίδευσης επεκτείνεται χρονικά, ενώ το σύνολο επικύρωσης παραμένει πάντα μεταγενέστερο χρονικά. Αυτή η προσέγγιση προσομοιώνει ρεαλιστικά τη χρήση του μοντέλου σε πραγματικές συνθήκες, όπου πάντα προβλέπουμε μελλοντικές τιμές βασιζόμενοι σε παρελθοντικά δεδομένα.



Εικόνα 4.19: Διασταυρούμενη επικύρωση χρονοσειρών με 5 τμήματα

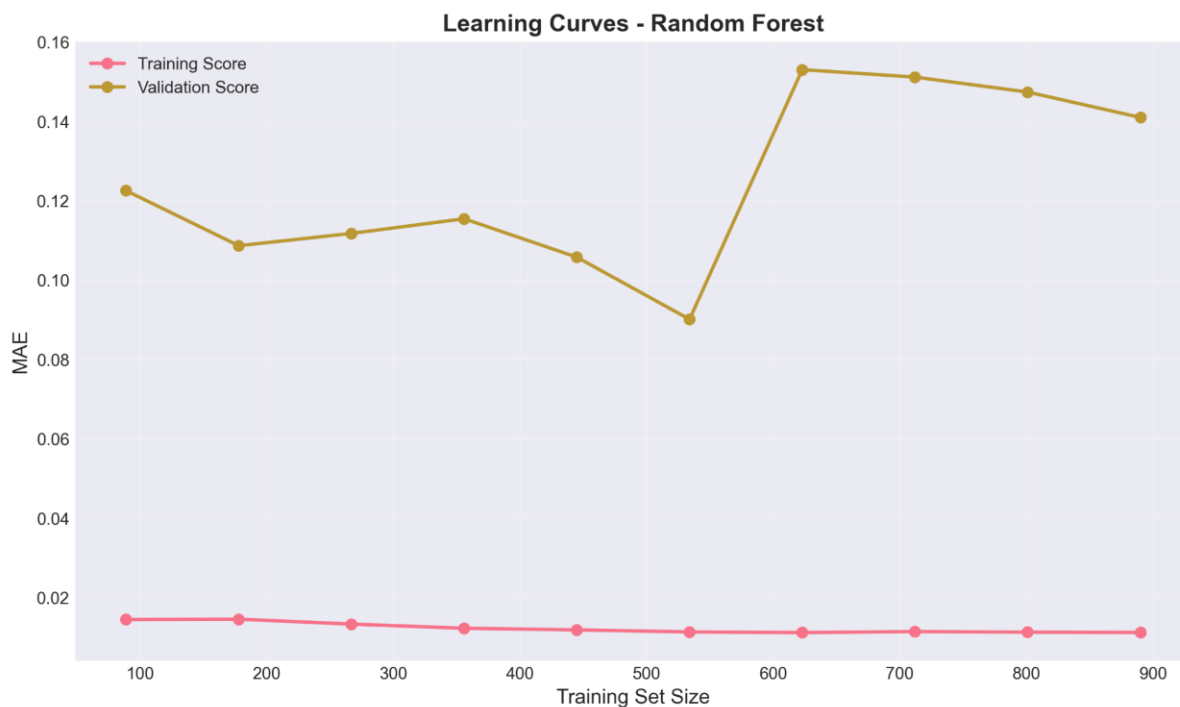
Βλέπε Παράρτημα Β για την πηγή της εικόνας στον φάκελο figures/ του αποθετηρίου.

Τα αποτελέσματα της διασταυρούμενης επικύρωσης για τον Random Forest ήταν: μέσο MAE = $0,037 \pm 0,007$, μέσο RMSE = $0,051 \pm 0,013$ και μέσο $R^2 = 0,944 \pm 0,039$. Παρατηρείται σημαντική απόκλιση μεταξύ αυτών των τιμών και της απόδοσης στο σύνολο δοκιμής (MAE = 0,092, $R^2 = 0,750$). Η διαφορά αυτή αποδίδεται στη φύση της διασταυρούμενης επικύρωσης χρονοσειρών με διευρυνόμενο παράθυρο: τα πτυχώματα (folds) αξιολογούν κυρίως βραχυπρόθεσμες προβλέψεις σε χρονικά γειτονικά δεδομένα, όπου η αυτοσυσχέτιση της χρονοσειράς διευκολύνει την πρόβλεψη. Αντιθέτως, το σύνολο δοκιμής αντιστοιχεί στην τελευταία χρονική περίοδο, η οποία

ενδέχεται να παρουσιάζει διαφορετικά μοτίβα ή μεγαλύτερη μεταβλητότητα σε σχέση με τα δεδομένα εκπαίδευσης. Η χαμηλή τυπική απόκλιση μεταξύ των folds επιβεβαιώνει τη σταθερότητα του μοντέλου εντός του εύρους εκπαίδευσης, αλλά η υποβαθμισμένη απόδοση στο σύνολο δοκιμής υποδηλώνει περιορισμένη γενικευσιμότητα σε χρονικά απομακρυσμένα δεδομένα.

4.4.2 Καμπύλες Μάθησης (Learning Curves)

Η ανάλυση καμπυλών μάθησης εξετάζει την εξέλιξη της απόδοσης του μοντέλου καθώς αυξάνεται ο όγκος των δεδομένων εκπαίδευσης [6]. Αυτή η ανάλυση παρέχει πληροφορίες για την ύπαρξη υπερπροσαρμογής (overfitting) ή υπο-προσαρμογής (underfitting), καθώς και για την πιθανή βελτίωση με περισσότερα δεδομένα.



Εικόνα 4.20: Καμπύλες μάθησης για Random Forest

Βλέπε Παράρτημα Β για την πηγή της εικόνας στον φάκελο figures/ του αποθετηρίου.

Οι καμπύλες μάθησης του Random Forest έδειξαν ότι η απόδοση στο σύνολο εκπαίδευσης παραμένει σταθερά υψηλή καθ' όλη τη διάρκεια, ενώ η απόδοση στο σύνολο επικύρωσης, αν και βελτιώνεται σταδιακά με την αύξηση των δεδομένων, διατηρεί αξιοσημείωτη απόσταση από την απόδοση εκπαίδευσης. Το κενό αυτό αποτελεί ένδειξη υπερπροσαρμογής (overfitting),

χαρακτηριστικής των μοντέλων Random Forest τα οποία, λόγω της υψηλής εκφραστικότητάς τους, τείνουν να απομνημονεύουν μέρος του θορύβου των δεδομένων εκπαίδευσης [7]. Ωστόσο, η ανοδική τάση της καμπύλης επικύρωσης υποδηλώνει ότι πρόσθετα δεδομένα εκπαίδευσης θα μπορούσαν να μειώσουν περαιτέρω αυτό το κενό.

4.4.3 Ανάλυση Καταλοίπων (Residual Analysis)

Η ανάλυση καταλοίπων αποτελεί κρίσιμο εργαλείο για τη διάγνωση πιθανών προβλημάτων στο μοντέλο και την επαλήθευση των υποθέσεων της μοντελοποίησης [4]. Τα κατάλοιπα ορίζονται ως η διαφορά μεταξύ πραγματικών και προβλεπόμενων τιμών:

$$r_i = y_i - \hat{y}_i$$

- Στατιστικά Καταλοίπων

Η ανάλυση των καταλοίπων του Random Forest στο σύνολο δοκιμής αποκάλυψε: μέση τιμή -0,051 (ελαφρά συστηματική υποεκτίμηση), τυπική απόκλιση 0,100, ελάχιστη τιμή -0,318, μέγιστη τιμή 0,414, ασυμμετρία 1,25 (θετική ασυμμετρία) και κύρτωση 3,45 (λεπτόκυρτη κατανομή).

- Έλεγχος Κανονικότητας (Shapiro-Wilk Test)

Ο έλεγχος Shapiro-Wilk αξιολογεί κατά πόσο τα κατάλοιπα ακολουθούν κανονική κατανομή. Τα αποτελέσματα έδειξαν: στατιστική $W = 0,865$ με $p\text{-value} < 0,0001$, υποδεικνύοντας στατιστικά σημαντική απόκλιση από την κανονικότητα. Η θετική ασυμμετρία και λεπτοκύρτωση υποδηλώνουν την ύπαρξη ορισμένων μεγάλων σφαλμάτων πρόβλεψης.

- Έλεγχος Αυτοσυσχέτισης (Durbin-Watson Test)

Ο έλεγχος Durbin-Watson αξιολογεί την ύπαρξη αυτοσυσχέτισης στα κατάλοιπα. Η τιμή του στατιστικού ήταν $DW = 0,468$, σημαντικά χαμηλότερη από την ιδανική τιμή 2,0, υποδεικνύοντας θετική αυτοσυσχέτιση. Αυτό σημαίνει ότι τα σφάλματα πρόβλεψης τείνουν να είναι παρόμοια σε διαδοχικές χρονικές στιγμές.



Εικόνα 4.21: Διαγνωστικά καταλοίπων (Residual Diagnostics)

Βλέπε Παράρτημα Β για την πηγή της εικόνας στον φάκελο figures/ του αποθετηρίου.

- Ανίχνευση Ακραίων Τιμών στα Κατάλοιπα

Με κριτήριο τις 3 τυπικές αποκλίσεις από τη μέση τιμή, εντοπίστηκαν 23 ακραίες τιμές στα κατάλοιπα, αντιστοιχώντας στο 1,62% του συνόλου δοκιμής. Αυτές οι παρατηρήσεις αντιπροσωπεύουν περιπτώσεις όπου το μοντέλο αποτύγχανε σημαντικά, πιθανώς λόγω ασυνήθιστων ατμοσφαιρικών συνθηκών ή γεγονότων που δεν αποτυπώνονται στα χαρακτηριστικά εισόδου.

4.5 Ανίχνευση Ανωμαλιών (Anomaly Detection)

Η ανίχνευση ανωμαλιών αποτελεί κρίσιμη λειτουργία για συστήματα παρακολούθησης IoT, καθώς επιτρέπει τον εντοπισμό μη φυσιολογικών μετρήσεων που μπορεί να υποδηλώνουν βλάβη αισθητήρα, περιβαλλοντικό συμβάν ή πρόβλημα ποιότητας δεδομένων [18]. Στην παρούσα

εργασία υλοποιήθηκαν και συγκρίθηκαν εννέα διαφορετικές μέθοδοι ανίχνευσης ανωμαλιών, οργανωμένες σε τέσσερις κατηγορίες.

Σημειώνεται ότι το σύνολο δεδομένων ποιότητας αέρα δεν περιλαμβάνει επισημασμένες ανωμαλίες (ground truth), καθώς πρόκειται για πραγματικές μετρήσεις αισθητήρων χωρίς ετικέτες κανονικότητας. Κατά συνέπεια, η σύγκριση των μεθόδων που ακολουθεί δεν αποτελεί αξιολόγηση ακρίβειας με βάση γνωστές ετικέτες, αλλά συγκριτική ανάλυση της συμπεριφοράς και ευαισθησίας κάθε μεθόδου. Τα ποσοστά ανίχνευσης πρέπει να ερμηνεύονται ως ενδεικτικά της τάσης κάθε μεθόδου προς συντηρητική ή επιθετική ανίχνευση, και όχι ως μέτρα απόδοσης.

4.5.1 Στατιστικές Μέθοδοι

Οι στατιστικές μέθοδοι ανίχνευσης ανωμαλιών βασίζονται σε στατιστικά μέτρα για τον εντοπισμό παρατηρήσεων που αποκλίνουν σημαντικά από το αναμενόμενο [18].

Τα υψηλά ποσοστά ανίχνευσης που παρατηρούνται στις στατιστικές μεθόδους (83,88%–98,31%) δεν υποδηλώνουν ότι σχεδόν ολόκληρο το σύνολο δεδομένων αποτελείται από ανωμαλίες. Αντιθέτως, αντανakλούν θεμελιώδη ασυμβατότητα μεταξύ των υποθέσεων αυτών των μεθόδων και της φύσης των δεδομένων: οι μέθοδοι Z-Score, IQR και Modified Z-Score προϋποθέτουν σχετικά στατική κατανομή, ενώ τα δεδομένα χρονοσειρών παρουσιάζουν εποχικότητα, τάσεις και μεταβαλλόμενη διακύμανση. Ως αποτέλεσμα, φυσιολογικές εποχικές μεταβολές ερμηνεύονται εσφαλμένα ως ανωμαλίες, καθιστώντας τις μεθόδους αυτές ακατάλληλες ως αυτόνομοι ανιχνευτές για χρονοσειρές IoT.

- Z-Score

Η μέθοδος Z-Score υπολογίζει την απόσταση κάθε παρατήρησης από τη μέση τιμή σε μονάδες τυπικών αποκλίσεων. Μια παρατήρηση θεωρείται ανωμαλία εάν $|z| > 3$. Η μέθοδος αυτή, ανίχνευσε 1.192 ανωμαλίες στο σύνολο δοκιμής (83,88%). Το υψηλό αυτό ποσοστό υποδηλώνει ότι η μέθοδος είναι ιδιαίτερα ευαίσθητη σε αποκλίσεις από την κανονικότητα.

- IQR (Interquartile Range)

Η μέθοδος IQR θεωρεί ανωμαλίες τις παρατηρήσεις που βρίσκονται εκτός του διαστήματος $[Q1 - 1,5 \times IQR, Q3 + 1,5 \times IQR]$. Η μέθοδος αυτή, ανίχνευσε 1.337 ανωμαλίες (94,09%), καθιστώντας την ακόμη πιο επιθετική από τη μέθοδο Z-Score.

- Modified Z-Score

Η τροποποιημένη μέθοδος Z-Score χρησιμοποιεί τη διάμεσο και την απόλυτη απόκλιση διαμέσου (MAD) αντί του μέσου όρου και της τυπικής απόκλισης, καθιστώντας τη πιο ανθεκτική σε ακραίες τιμές. Η μέθοδος αυτή, ανίχνευσε 1.397 ανωμαλίες (98,31%), το υψηλότερο ποσοστό μεταξύ όλων των μεθόδων.

4.5.2 Μέθοδοι Μηχανικής Μάθησης

Οι μέθοδοι μηχανικής μάθησης χωρίς επίβλεψη (unsupervised) μπορούν να ανιχνεύσουν ανωμαλίες χωρίς την ανάγκη επισημασμένων παραδειγμάτων [19], [20], [21].

- Isolation Forest

Ο αλγόριθμος Isolation Forest [19] βασίζεται στην παρατήρηση ότι οι ανωμαλίες είναι ευκολότερο να απομονωθούν από τα κανονικά δεδομένα. Ο αλγόριθμος κατασκευάζει δέντρα απομόνωσης και αξιολογεί την ευκολία απομόνωσης κάθε παρατήρησης. Με ποσοστό μόλυνσης (contamination) 5%, η μέθοδος ανίχνευσε 801 ανωμαλίες (56,37%).

- Local Outlier Factor (LOF)

Ο αλγόριθμος LOF [20] συγκρίνει την τοπική πυκνότητα κάθε παρατήρησης με την πυκνότητα των γειτόνων της. Παρατηρήσεις με σημαντικά χαμηλότερη πυκνότητα θεωρούνται ανωμαλίες. Με 20 γείτονες, η μέθοδος ανίχνευσε 1.421 ανωμαλίες (100%), χαρακτηρίζοντας ολόκληρο το σύνολο δοκιμής ως μη φυσιολογικό. Το αποτέλεσμα αυτό υποδεικνύει αστοχία στην παραμετροποίηση της μεθόδου και όχι πραγματική ανωμαλία ολόκληρου του συνόλου.

Ο αλγόριθμος LOF εκπαιδεύτηκε στο σύνολο εκπαίδευσης και αξιολογήθηκε στο σύνολο δοκιμής, το οποίο παρουσιάζει αναπόφευκτα διαφορετική κατανομή πυκνότητας — ιδίως σε δεδομένα χρονοσειρών όπου η χρονική μετατόπιση μεταξύ εκπαίδευσης και δοκιμής αλλοιώνει τη “δομή γειτονιάς”. Σε συνδυασμό με τον υψηλό αριθμό χαρακτηριστικών (103), η κατάρα της

διαστατικότητας (curse of dimensionality) καθιστά τις αποστάσεις μεταξύ σημείων σχεδόν ομοιόμορφες, υπονομεύοντας την ικανότητα του αλγορίθμου να διακρίνει πραγματικές ανωμαλίες.

- One-Class SVM

Ο αλγόριθμος One-Class SVM [21] μαθαίνει ένα σύνορο αποφάσεων γύρω από τα κανονικά δεδομένα και θεωρεί ανωμαλίες τις παρατηρήσεις εκτός αυτού του συνόρου. Ομοίως με τον LOF, η μέθοδος ανίχνευσε 1.421 ανωμαλίες (100%). Η αποτυχία αυτή αποδίδεται σε παρόμοιους παράγοντες: ο αλγόριθμος One-Class SVM βασίζεται σε πυρηνικές συναρτήσεις (kernel functions) για τον ορισμό του συνόρου, και σε υψηλή διαστατικότητα το σύνορο γίνεται υπερβολικά στενό, αποκλείοντας παρατηρήσεις που απλώς αντανακλούν φυσιολογική χρονική μεταβλητότητα. Η βελτιστοποίηση της παραμέτρου ν (ν) και η εφαρμογή μείωσης διαστατικότητας (π.χ. PCA) πριν την εκπαίδευση θα μπορούσαν να βελτιώσουν σημαντικά τα αποτελέσματα, αποτελώντας αντικείμενο μελλοντικής διερεύνησης.

4.5.3 Μέθοδος Βασισμένη σε Κατάλοιπα Μοντέλου

Η μέθοδος βασισμένη σε κατάλοιπα (Residual-Based) αξιοποιεί τις προβλέψεις του εκπαιδευμένου μοντέλου Random Forest. Για κάθε παρατήρηση υπολογίζεται το κατάλοιπο ως η διαφορά μεταξύ πραγματικής και προβλεπόμενης τιμής. Μια παρατήρηση θεωρείται ανωμάλια εάν το απόλυτο κατάλοιπο υπερβαίνει το κατώφλι $|\text{residual}| > 3\sigma$, όπου σ η τυπική απόκλιση των καταλοίπων στο σύνολο εκπαίδευσης. Η μέθοδος ανίχνευσε 1.197 ανωμαλίες (84,24%), ποσοστό που αντανακλά τη σημαντική απόκλιση μεταξύ των προβλέψεων του μοντέλου και των πραγματικών τιμών στο σύνολο δοκιμής — συνεπές με τη μέτρια προβλεπτική απόδοση του Random Forest ($R^2 = 0,750$) που αναλύθηκε στην Ενότητα 4.4.

4.5.4 Μέθοδος Αποσύνθεσης Χρονοσειρών (STL Decomposition)

Η μέθοδος STL (Seasonal-Trend decomposition using Loess) [22] αποσυνθέτει τη χρονοσειρά σε τρία συστατικά: τάση (trend), εποχικότητα (seasonality) και κατάλοιπα (residual). Οι ανωμαλίες ανιχνεύονται στα κατάλοιπα μετά την αφαίρεση της τάσης και της εποχικότητας.

Η μέθοδος STL ανίχνευσε μόλις 37 ανωμαλίες (2,60%), αποτελώντας την πλέον συντηρητική μέθοδο. Αυτό οφείλεται στην ικανότητά της να διακρίνει μεταξύ φυσιολογικής εποχικής διακύμανσης και πραγματικών ανωμαλιών [22].

4.5.5 Μέθοδος Συναθροιστικής Ψηφοφορίας (Ensemble)

Η μέθοδος Ensemble ανίχνευσε 1.211 ανωμαλίες (85,22%), ποσοστό που αντανάκλα το γεγονός ότι η πλειοψηφία των επιμέρους μεθόδων παρουσιάζει υψηλή ευαισθησία στο συγκεκριμένο σύνολο δεδομένων. Το αποτέλεσμα αυτό αναδεικνύει έναν εγγενή περιορισμό της ψηφοφορίας πλειοψηφίας: όταν οι περισσότερες μέθοδοι υπερεκτιμούν τις ανωμαλίες — κυρίως λόγω παραβίασης της υπόθεσης στατικής κατανομής σε δεδομένα χρονοσειρών, όπως αναλύθηκε ανωτέρω — η συνάθροιση δεν διορθώνει τη συστηματική αυτή απόκλιση, αλλά την αναπαράγει.

Παρά τον περιορισμό αυτόν, η μέθοδος Ensemble παραμένει χρήσιμη ως διαγνωστικό εργαλείο: η σύγκριση του ποσοστού της (85,22%) με τα ακραία εύρη των μεμονωμένων μεθόδων (2,6%–100%) καθιστά ορατή τη διαφωνία μεταξύ των ανιχνευτών και επισημαίνει ότι η ερμηνεία των ανωμαλιών στο σύνολο αυτό απαιτεί ιδιαίτερη προσοχή. Για πρακτικές εφαρμογές με αυστηρότερες απαιτήσεις, μια σταθμισμένη ψηφοφορία (weighted ensemble) που προτιμά μεθόδους ευαίσθητες στη χρονική δομή — όπως η STL — θα αποτελούσε αποτελεσματικότερη εναλλακτική.

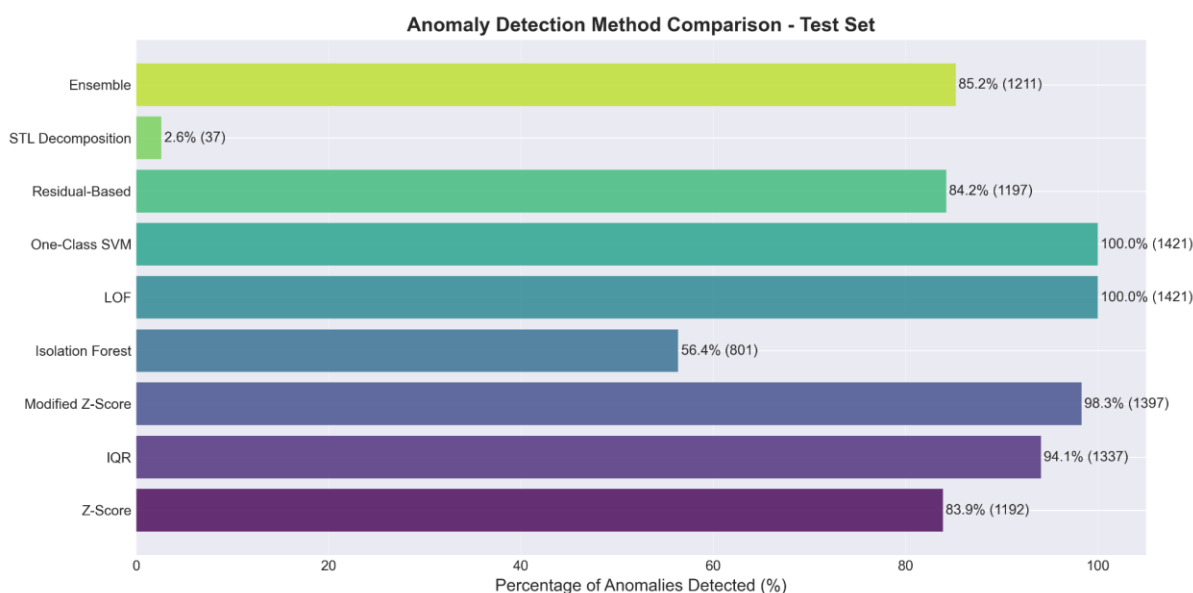
4.5.6 Σύγκριση Μεθόδων Ανίχνευσης Ανωμαλιών

Ο παρακάτω πίνακας συνοψίζει τα αποτελέσματα όλων των μεθόδων ανίχνευσης ανωμαλιών:

Πίνακας 4.4: Σύγκριση μεθόδων ανίχνευσης ανωμαλιών - Air Quality

Μέθοδος	Ανωμαλίες	Ποσοστό (%)
Z-Score	1.192	83,88
IQR	1.337	94,09
Modified Z-Score	1.397	98,31
Isolation Forest	801	56,37

LOF	1.421	100,00
One-Class SVM	1.421	100,00
Residual-Based	1.197	84,24
STL Decomposition	37	2,60
Ensemble	1.211	85,22

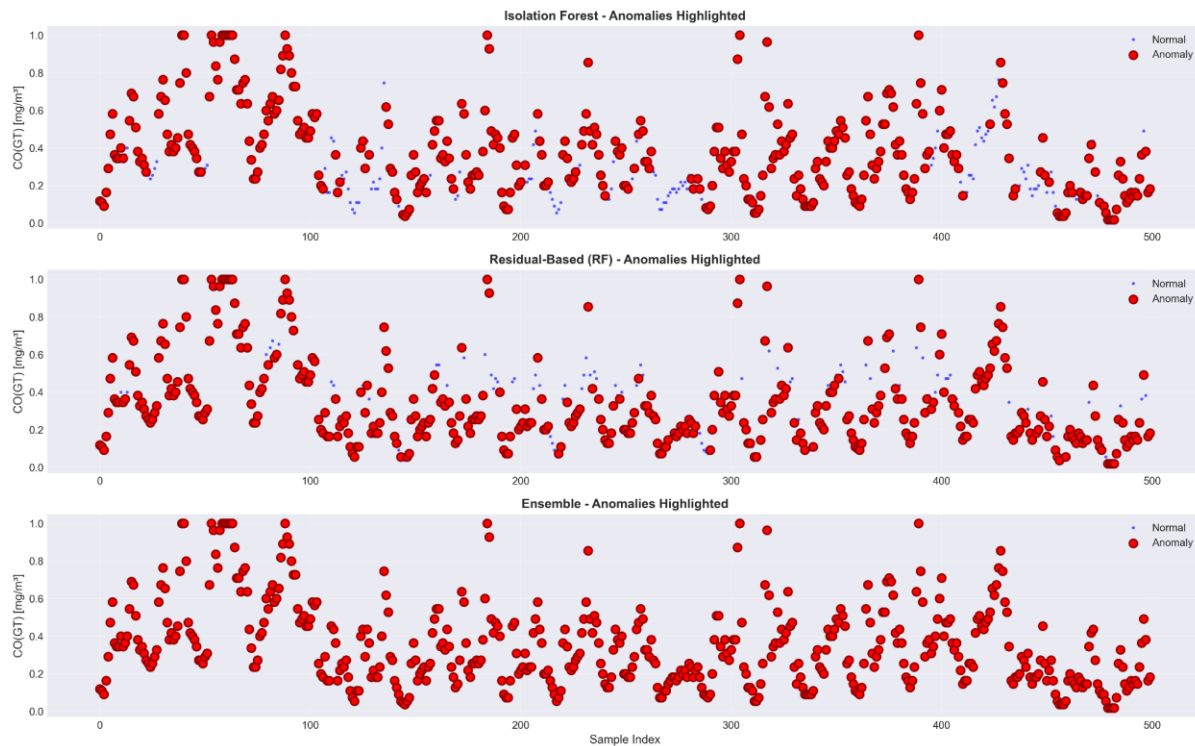


Εικόνα 4.22: Σύγκριση μεθόδων ανίχνευσης ανωμαλιών

Βλέπε Παράρτημα Β για την πηγή της εικόνας στον φάκελο figures/ του αποθετηρίου.

Η σημαντική διακύμανση μεταξύ των μεθόδων αντανakλά τις διαφορετικές υποθέσεις και ευαισθησίες τους. Η μέθοδος STL αποδεικνύεται η πλέον κατάλληλη για δεδομένα χρονοσειρών με έντονη εποχικότητα, καθώς διακρίνει μεταξύ κανονικών εποχικών διακυμάνσεων και πραγματικών ανωμαλιών. Αντιθέτως, οι στατιστικές μέθοδοι (Z-Score, IQR) αποτυγχάνουν σε δεδομένα χρονοσειρών, διότι βασίζονται στην υπόθεση στατικής κατανομής — αντιμετωπίζουν κάθε παρατήρηση ανεξάρτητα, αγνοώντας τη χρονική δομή, με αποτέλεσμα οι φυσιολογικές εποχικές αιχμές να ερμηνεύονται εσφαλμένα ως ανωμαλίες. Ομοίως, ο αλγόριθμος LOF, σχεδιασμένος για ανίχνευση τοπικά απομονωμένων σημείων σε στατικό χώρο χαρακτηριστικών,

υπερεκτιμά τις ανωμαλίες όταν η χρονική εξέλιξη δημιουργεί φυσική μεταβλητότητα στην πυκνότητα των δεδομένων.



Εικόνα 4.23: Χρονική γραμμή ανιχνευμένων ανωμαλιών

Βλέπε Παράρτημα Β για την πηγή της εικόνας στον φάκελο figures/ του αποθετηρίου.

4.5.7 Συστάσεις για Πρακτική Εφαρμογή

Βάσει της ανάλυσης, προτείνονται οι ακόλουθες κατευθύνσεις για την πρακτική εφαρμογή ανίχνευσης ανωμαλιών:

- Για δεδομένα με έντονη εποχικότητα, όπως τα δεδομένα ατμοσφαιρικής ρύπανσης, η μέθοδος STL Decomposition αποτελεί την πλέον κατάλληλη επιλογή λόγω της ικανότητάς της να διακρίνει μεταξύ φυσιολογικών μοτίβων και πραγματικών ανωμαλιών.
- Για ισορροπημένη ανίχνευση, η μέθοδος Residual-Based ή η μέθοδος Ensemble προσφέρουν έναν συμβιβασμό μεταξύ ευαισθησίας και ειδικότητας.
- Για ανίχνευση σοβαρών ανωμαλιών μόνο, η μέθοδος Isolation Forest με χαμηλό ποσοστό μόλυνσης είναι κατάλληλη, καθώς εστιάζει στις πιο ακραίες παρατηρήσεις.

Σύνοψη Κεφαλαίου

Το παρόν κεφάλαιο παρουσίασε την πλήρη εφαρμογή της μεθοδολογίας CRISP-DM στο σύνολο δεδομένων UCI Air Quality [1], [2], καλύπτοντας τις φάσεις της Προετοιμασίας Δεδομένων, της Μοντελοποίησης και της Αξιολόγησης.

Η Διερευνητική Ανάλυση Δεδομένων (EDA) αποκάλυψε τα βασικά χαρακτηριστικά του συνόλου δεδομένων, συμπεριλαμβανομένων των χρονικών μοτίβων, της δομής ελλειπουσών τιμών και των συσχετίσεων μεταξύ μεταβλητών. Η προεπεξεργασία περιελάμβανε γραμμική παρεμβολή για τις ελλείπουσες τιμές, αποκοπή IQR για τις ακραίες τιμές, κανονικοποίηση MinMax και χρονικό διαχωρισμό 70-15-15.

Η σχεδίαση χαρακτηριστικών (feature engineering) επέκτεινε τον χώρο χαρακτηριστικών από 12 σε 107 διαστάσεις, περιλαμβάνοντας χαρακτηριστικά υστέρησης, κυλιόμενου παραθύρου, κυκλικής κωδικοποίησης χρόνου και αλληλεπιδράσεων.

Η μοντελοποίηση (modeling) συνέκρινε έξι αλγόριθμους, με το Random Forest να αναδεικνύεται ως το βέλτιστο μοντέλο, επιτυγχάνοντας $MAE = 0,092 \text{ mg/m}^3$ και $R^2 = 0,750$ στο σύνολο δοκιμής. Ωστόσο, τα μοντέλα βαθιάς μάθησης (LSTM, GRU) είχαν χαμηλή απόδοση λόγω του περιορισμένου όγκου δεδομένων.

Η αξιολόγηση (evaluation) περιελάμβανε διασταυρούμενη επικύρωση χρονοσειρών με 5 πτυχές, ανάλυση καμπυλών μάθησης και ενδελεχή διαγνωστικά καταλοίπων.

Τέλος, για την ανίχνευση ανωμαλιών (anomaly detection) χρησιμοποιήθηκαν εννέα μέθοδοι, με τη μέθοδο STL Decomposition [22] να αναδεικνύεται ως η πλέον συντηρητική και κατάλληλη για δεδομένα χρονοσειρών με έντονη εποχικότητα, όπως προαναφέρθηκε.

Τα αποτελέσματα αυτά θέτουν τη βάση για τη συγκριτική ανάλυση με το σύνολο δεδομένων κατανάλωσης ενέργειας (Κεφάλαιο 5) και την εξαγωγή γενικεύσιμων συμπερασμάτων (Κεφάλαιο 6).

ΚΕΦΑΛΑΙΟ 5: POWER CONSUMPTION - ΑΝΑΛΥΣΗ ΚΑΙ ΜΟΝΤΕΛΟΠΟΙΗΣΗ

Εισαγωγή στο Κεφάλαιο

Το παρόν κεφάλαιο αποτελεί τη δεύτερη πρακτική εφαρμογή της μεθοδολογίας CRISP-DM (Cross-Industry Standard Process for Data Mining) [25], εστιάζοντας στο σύνολο δεδομένων Individual Household Electric Power Consumption [3] του αποθετηρίου UCI Machine Learning Repository. Η ανάλυση αυτή, με τη σειρά της, συμπληρώνει την μελέτη του Κεφαλαίου 4 (Air Quality), επιτρέποντας, έτσι, τη συγκριτική αξιολόγηση της μεθοδολογίας σε διαφορετικό πεδίο εφαρμογής του Διαδικτύου των Πραγμάτων (IoT).

Στο πλαίσιο αυτό, το σύνολο δεδομένων κατανάλωσης ενέργειας παρουσιάζει θεμελιώδεις διαφορές σε σύγκριση με το σύνολο ποιότητας αέρα που εξετάστηκε προηγουμένως. Πρώτον, η κλίμακα των δεδομένων είναι σημαντικά μεγαλύτερη, περιλαμβάνοντας περισσότερες από 2 εκατομμύρια παρατηρήσεις, έναντι περίπου 9.500 του προηγούμενου συνόλου. Δεύτερον, η αρχική συχνότητα δειγματοληψίας διαφέρει δραματικά: ανά ένα λεπτό για την κατανάλωση ενέργειας, έναντι ανά μία ώρα για την ποιότητα αέρα. Τρίτον, τα μοτίβα που κυριαρχούν είναι διαφορετικής φύσεως, αντανακλώντας την ανθρώπινη συμπεριφορά και τη χρήση οικιακών συσκευών, αντί περιβαλλοντικών και κυκλοφοριακών παραγόντων.

Ο πρωταρχικός στόχος της ανάλυσης είναι η ανάπτυξη μοντέλων πρόβλεψης για τη συνολική ενεργό ισχύ (Global_active_power) μία ώρα στο μέλλον (1-hour ahead forecasting), διατηρώντας μεθοδολογική συνέπεια με την ανάλυση του Κεφαλαίου 4, ώστε να διευκολυνθεί η συγκριτική αξιολόγηση. Η επιλογή της συνολικής ενεργού ισχύος ως μεταβλητή-στόχο δικαιολογείται από την κεντρική της σημασία στη διαχείριση ενέργειας και την πρόβλεψη φορτίου σε έξυπνα δίκτυα (smart grids) [23], [24].

Όσον αφορά τη δομή του κεφαλαίου, ακολουθεί την ίδια οργάνωση με το Κεφάλαιο 4: η Ενότητα 5.1 παρουσιάζει τη Διερευνητική Ανάλυση Δεδομένων (EDA) και την προεπεξεργασία, η Ενότητα

5.2 αναλύει τη διαδικασία μοντελοποίησης και η Ενότητα 5.3 επικεντρώνεται στην ανίχνευση ανωμαλιών.

Αναλυτικά αποτελέσματα παρατίθενται στο παράρτημα Γ (βλέπε σελ. 139). Όλοι οι πίνακες και τα δεδομένα τους μπορούν να βρεθούν τόσο στο παράρτημα Γ, όσο και στο αποθετήριο κώδικα του παραρτήματος Β (βλέπε σελ. 137 και 139).

5.1 Διερευνητική Ανάλυση Δεδομένων και Προεπεξεργασία (EDA & Preprocessing)

Η Διερευνητική Ανάλυση Δεδομένων (Exploratory Data Analysis - EDA) αποτελεί θεμελιώδες βήμα της μεθοδολογίας CRISP-DM, αντιστοιχώντας στις φάσεις Κατανόησης Δεδομένων (Data Understanding) και Προετοιμασίας Δεδομένων (Data Preparation) [25]. Για το σύνολο δεδομένων κατανάλωσης ενέργειας, η EDA αποκτά ιδιαίτερη σημασία λόγω του μεγάλου όγκου δεδομένων και της υψηλής συχνότητας δειγματοληψίας, που απαιτούν προσεκτικό χειρισμό.

5.1.1 Επισκόπηση του Συνόλου Δεδομένων

Το σύνολο δεδομένων Individual Household Electric Power Consumption [3] περιλαμβάνει μετρήσεις ηλεκτρικής κατανάλωσης από ένα νοικοκυριό στη Γαλλία (Sceaux, πλησίον του Παρισιού), συλλεγόμενες με συχνότητα ενός λεπτού. Η χρονική περίοδος κάλυψης εκτείνεται από τον Δεκέμβριο του 2006 έως τον Νοέμβριο του 2010, περιλαμβάνοντας συνολικά 2.075.259 παρατηρήσεις σε διάστημα 1.441 ημερών (περίπου 47 μήνες).

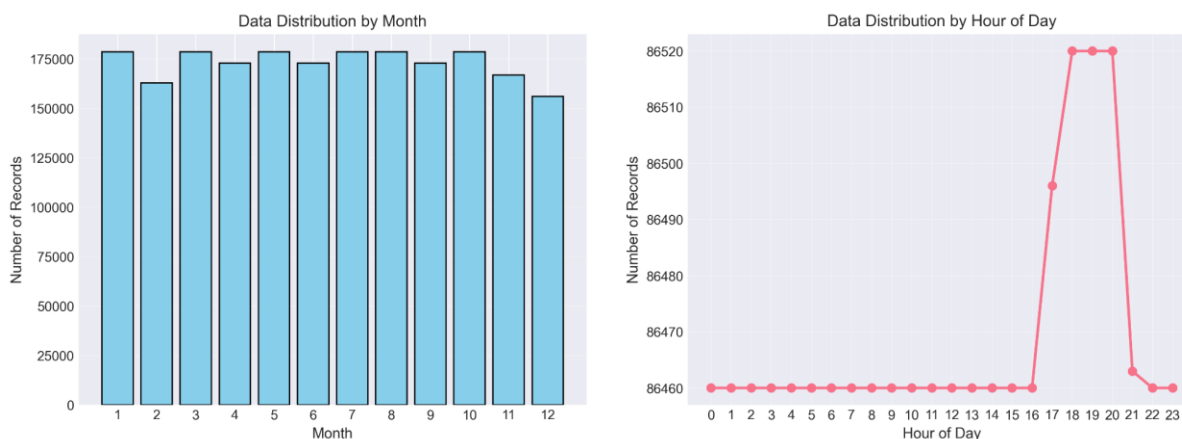
Το σύνολο δεδομένων περιλαμβάνει επτά αριθμητικά χαρακτηριστικά που αφορούν ηλεκτρικές μετρήσεις:

Πίνακας 5.1: Χαρακτηριστικά του συνόλου δεδομένων Power Consumption

Χαρακτηριστικό	Περιγραφή	Μονάδα
Global_active_power	Συνολική ενεργός ισχύς νοικοκυριού	kW
Global_reactive_power	Συνολική άεργος ισχύς νοικοκυριού	kW
Voltage	Τάση	V

Global_intensity	Συνολική ένταση ρεύματος	A
Sub_metering_1	Κατανάλωση κουζίνας (φούρνος, μικροκυμάτων)	Wh
Sub_metering_2	Κατανάλωση πλυντηρίου/στεγνωτηρίου	Wh
Sub_metering_3	Κατανάλωση θέρμανσης νερού και κλιματισμού	Wh

Η μεταβλητή-στόχος για την πρόβλεψη είναι η Global_active_power (συνολική ενεργός ισχύς σε kilowatt), η οποία αντιπροσωπεύει τη συνολική κατανάλωση ενέργειας του νοικοκυριού και είναι ιδιαίτερα κρίσιμη για εφαρμογές διαχείρισης ενεργειακού φορτίου και έξυπνων δικτύων.



Εικόνα 5.1: Χρονική κατανομή δεδομένων

Στο παραπάνω, παρατηρούμε την μηνιαία κατανομή των δεδομένων, ενώ στην δεξιά την κατανομή ανά ώρα της ημέρας.

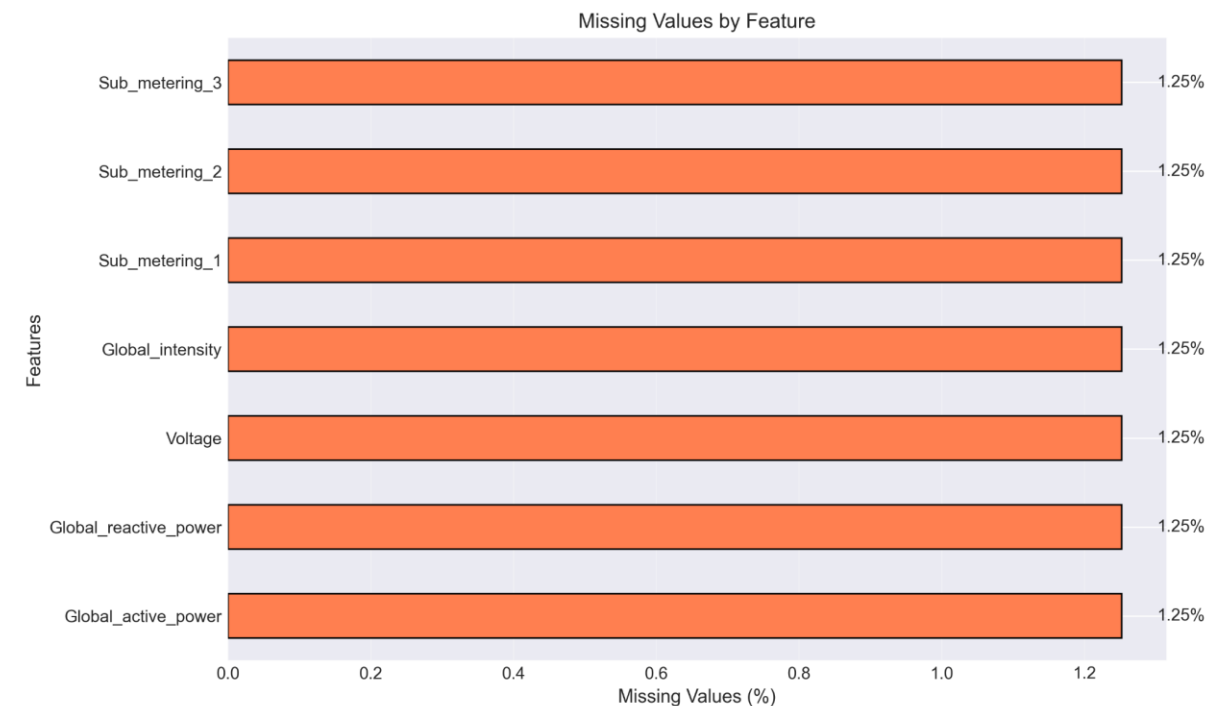
5.1.2 Ανάλυση Ποιότητας Δεδομένων

Η ποιότητα των δεδομένων εξετάστηκε συστηματικά, εντοπίζοντας και καταγράφοντας τα προβλήματα που απαιτούν διόρθωση. Σε αντίθεση με το σύνολο δεδομένων ποιότητας αέρα που χρησιμοποιούσε την τιμή -200 ως δείκτη ελλειπουσών τιμών, το σύνολο κατανάλωσης ενέργειας χρησιμοποιεί τον χαρακτήρα "?" για την επισήμανση ελλειπουσών μετρήσεων [3].

Η ανάλυση των ελλειπουσών τιμών αποκάλυψε ομοιόμορφο ποσοστό απώλειας δεδομένων σε όλα τα χαρακτηριστικά:

Πίνακας 5.2: Ποσοστό απώλειας δεδομένων με Ανάλυση ελλειπουσών τιμών

Χαρακτηριστικό	Ελλείπουσες Τιμές	Ποσοστό (%)
Global_active_power	25.979	1,25
Global_reactive_power	25.979	1,25
Voltage	25.979	1,25
Global_intensity	25.979	1,25
Sub_metering_1	25.979	1,25
Sub_metering_2	25.979	1,25
Sub_metering_3	25.979	1,25



Εικόνα 5.2: Οπτικοποίηση ελλειπουσών τιμών

Βλέπε Παράρτημα Β για την πηγή της εικόνας στον φάκελο figures/ του αποθετηρίου.

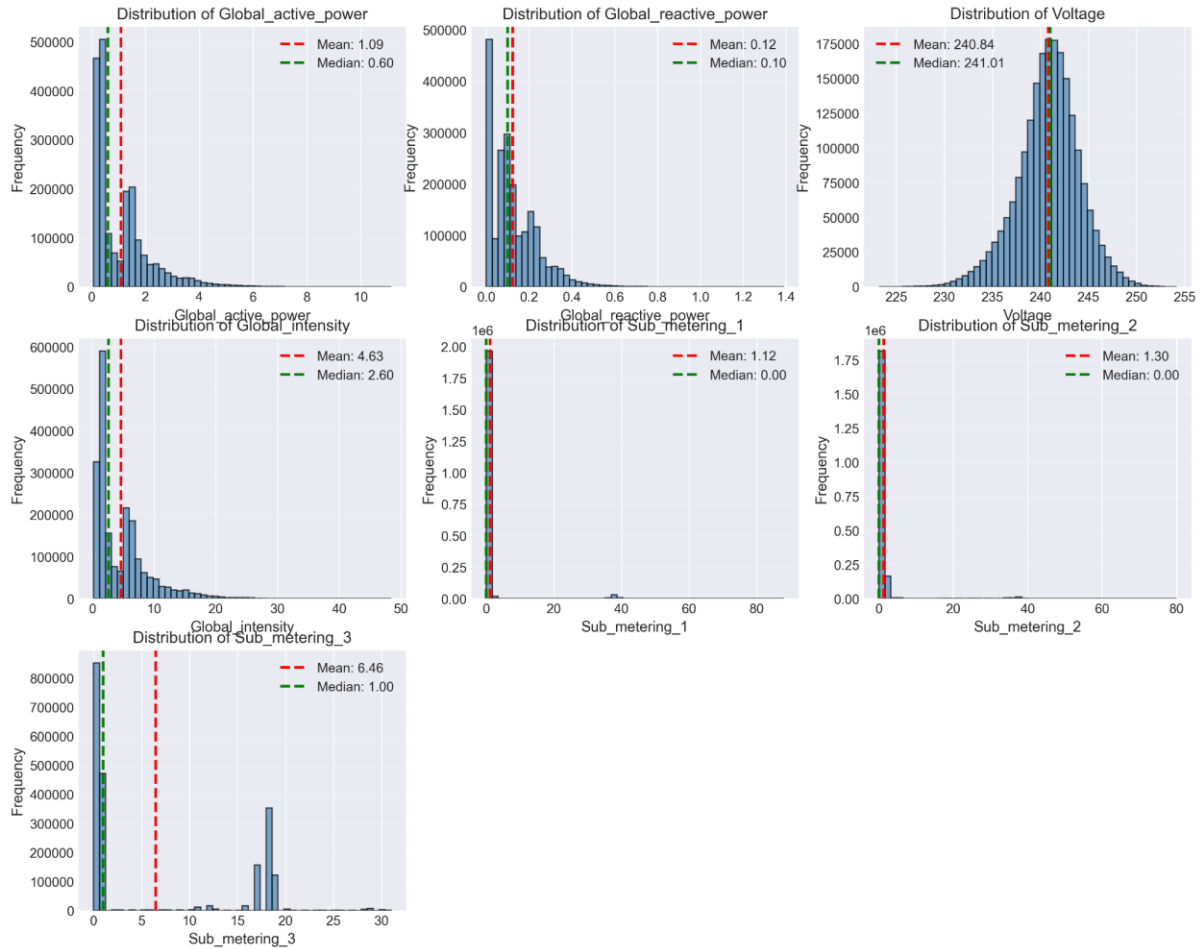
Η ομοιομορφία του ποσοστού ελλειπουσών τιμών (1,25%) σε όλα τα χαρακτηριστικά υποδεικνύει ότι, οι απώλειες οφείλονται σε πλήρη διακοπή της καταγραφής, παρά σε αποτυχία μεμονωμένων αισθητήρων. Συνεπώς, αυτό το μοτίβο είναι συμβατό με περιόδους διακοπής ρεύματος ή αποσύνδεσης του συστήματος καταγραφής [3]. Έτσι, διαπιστώνεται ότι, το συνολικό ποσοστό ελλειπουσών τιμών (1,25%) είναι σημαντικά χαμηλότερο από το αντίστοιχο του συνόλου ποιότητας αέρα (8,47%), υποδεικνύοντας υψηλότερη αξιοπιστία του συστήματος καταγραφής.

5.1.3 Στατιστική Περιγραφή

Η στατιστική ανάλυση των χαρακτηριστικών αποκάλυψε σημαντικές πληροφορίες για τη φύση των δεδομένων:

Πίνακας 5.3: Λεπτομέρειες Χαρακτηριστικών μετά από Στατιστική Ανάλυση

Χαρακτηριστικό	Μέσος Όρος	Τυπική Απόκλιση	Ελάχιστο	Μέγιστο	Λοξότητα
Global_active_power	1,09 kW	1,06 kW	0,076 kW	11,12 kW	1,79
Global_reactive_power	0,12 kW	0,11 kW	0,0 kW	1,39 kW	-
Voltage	240,84 V	3,24 V	223,2 V	254,15 V	-0,33
Global_intensity	4,63 A	4,44 A	0,2 A	48,4 A	1,85
Sub_metering_1	1,12 Wh	6,15 Wh	0,0 Wh	88,0 Wh	-
Sub_metering_2	1,30 Wh	5,82 Wh	0,0 Wh	80,0 Wh	-
Sub_metering_3	6,46 Wh	8,44 Wh	0,0 Wh	31,0 Wh	-



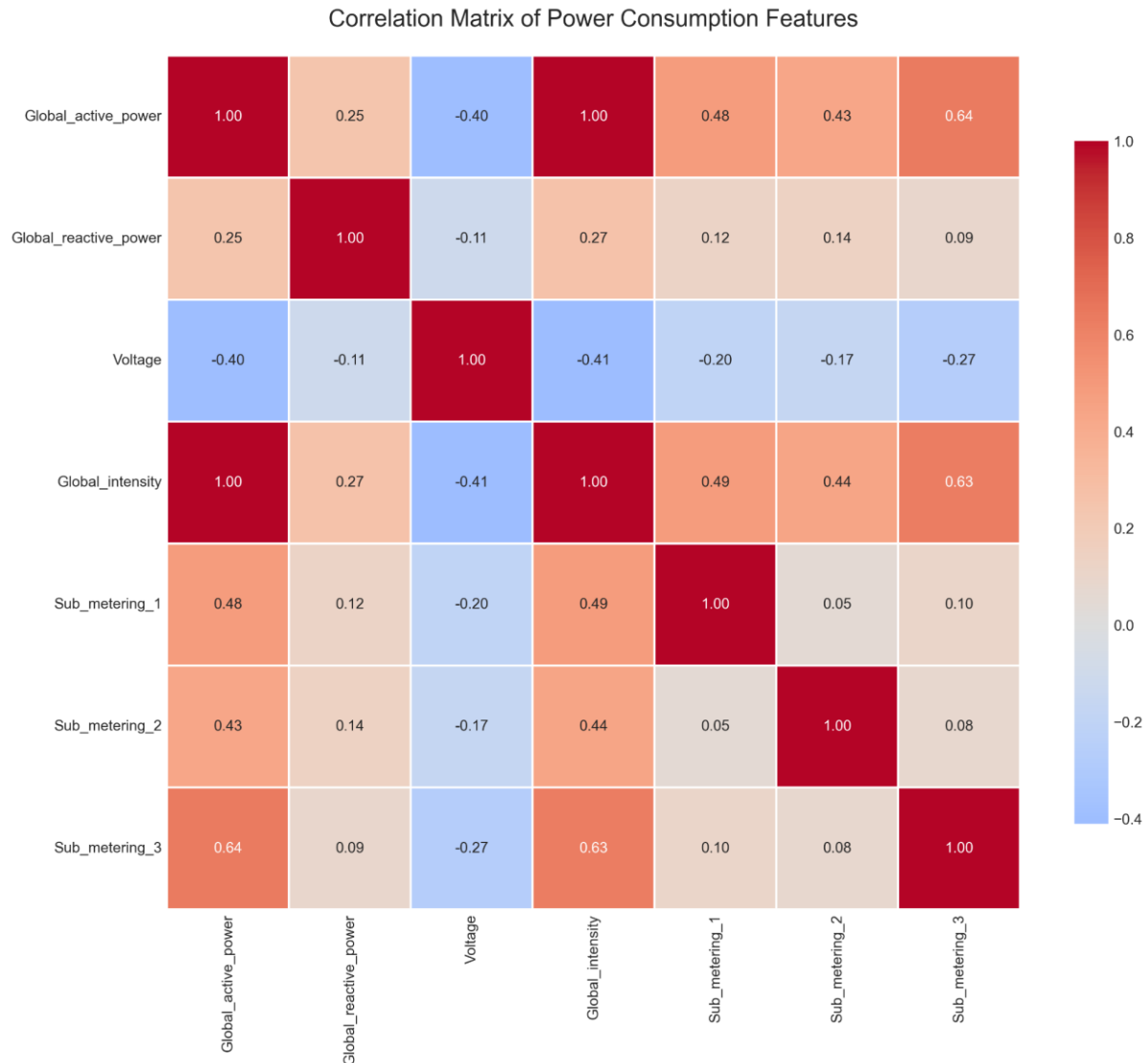
Εικόνα 5.3: Κατανομές χαρακτηριστικών

Βλέπε Παράρτημα Β για την πηγή της εικόνας στον φάκελο figures/ του αποθετηρίου.

Η ανάλυση αποκαλύπτει σημαντικά ευρήματα. Η μεταβλητή-στόχος `Global_active_power` παρουσιάζει θετική λοξότητα ($skewness = 1,79$), υποδεικνύοντας ότι οι περισσότερες τιμές συγκεντρώνονται σε χαμηλά επίπεδα κατανάλωσης με λιγότερο συχνές, αλλά υψηλές τιμές αιχμής. Η τάση (`Voltage`) παρουσιάζει ελαφρά αρνητική λοξότητα ($-0,33$) και σχετικά μικρή διακύμανση, αντανakλώντας τη σταθερότητα του ηλεκτρικού δικτύου. Οι υπο-μετρητές (`Sub-metering`) παρουσιάζουν υψηλή διακύμανση με πολλές μηδενικές τιμές, αντικατοπτρίζοντας την περιοδική χρήση συγκεκριμένων συσκευών.

5.1.4 Ανάλυση Συσχετίσεων

Η ανάλυση συσχετίσεων μεταξύ των χαρακτηριστικών αποκάλυψε σημαντικές σχέσεις που επηρεάζουν τη μοντελοποίηση:



Εικόνα 5.4: Χάρτης θερμότητας συσχετίσεων

Βλέπε Παράρτημα Β για την πηγή της εικόνας στον φάκελο figures/ του αποθετηρίου.

Η σημαντικότερη συσχέτιση εντοπίστηκε μεταξύ της `Global_active_power` και της `Global_intensity`, με συντελεστή συσχέτισης Pearson $r = 0,999$. Αυτή η σχεδόν τέλεια γραμμική σχέση δικαιολογείται φυσικά: η ενεργός ισχύς (P) σχετίζεται με την ένταση (I) μέσω της σχέσης $P = V \times I \times \cos(\phi)$, όπου για σταθερή τάση και συντελεστή ισχύος, η σχέση είναι ουσιαστικά

γραμμική. Αυτή η υψηλή συσχέτιση υποδεικνύει πιθανή πολυσυγγραμμικότητα που πρέπει να ληφθεί υπόψη κατά τη μοντελοποίηση.

Η τάση (Voltage) παρουσιάζει αρνητική συσχέτιση με την ενεργό ισχύ ($r = -0,40$) και την ένταση ($r = -0,41$). Αυτό αντανακλά το φαινόμενο πτώσης τάσης κατά τη διάρκεια υψηλής κατανάλωσης, γνωστό από τη θεωρία ηλεκτρικών κυκλωμάτων.

Τα χαρακτηριστικά υπο-μέτρησης (Sub_metering) παρουσιάζουν μέτριες έως υψηλές συσχετίσεις με τη συνολική κατανάλωση, με το Sub_metering_3 (θέρμανση/κλιματισμός) να έχει την υψηλότερη συσχέτιση ($r = 0,64$), ακολουθούμενο από το Sub_metering_1 (κουζίνα, $r = 0,48$) και το Sub_metering_2 (πλυντήριο, $r = 0,43$).

5.1.5 Χρονικά Μοτίβα

Η ανάλυση των χρονικών μοτίβων αποκάλυψε σαφείς περιοδικότητες που αντανακλούν την ανθρώπινη συμπεριφορά και τις καθημερινές ρουτίνες:



Εικόνα 5.5: Χρονοσειρά κατανάλωσης ενέργειας

Βλέπε Παράρτημα Β για την πηγή της εικόνας στον φάκελο figures/ του αποθετηρίου.

Η ημερήσια περιοδικότητα είναι εμφανής, με ώρα αιχμής κατανάλωσης στις 20:00 (8 μ.μ.), όπου αντιστοιχεί στην τυπική περίοδο επιστροφής στο σπίτι, μαγειρέματος δείπνου και χρήσης οικιακών συσκευών. Τα χαμηλότερα επίπεδα κατανάλωσης παρατηρούνται κατά τις πρωινές ώρες (03:00-06:00), αντανακλώντας τις ώρες ύπνου.

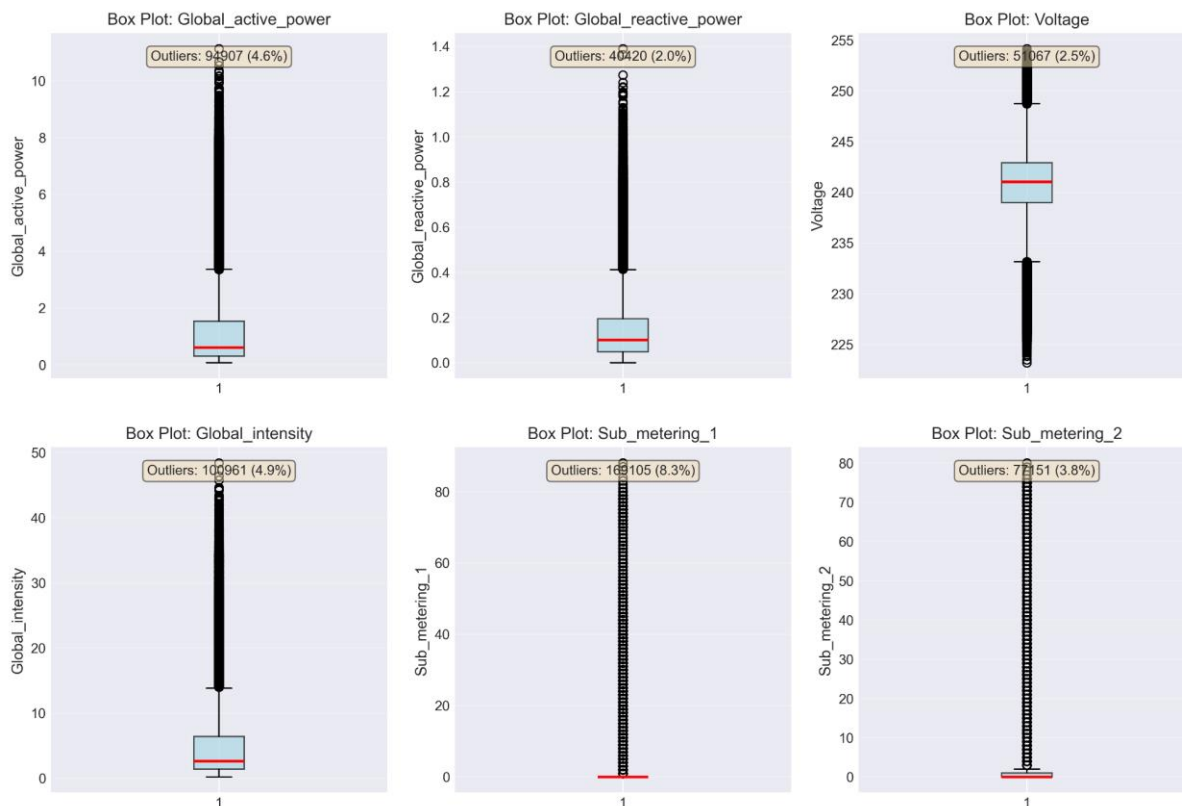
Η εβδομαδιαία περιοδικότητα εμφανίζεται με υψηλότερη κατανάλωση κατά τα Σαββατοκύριακα, όταν οι κάτοικοι παραμένουν περισσότερο χρόνο στο σπίτι. Η εποχική περιοδικότητα είναι επίσης έντονη, με αυξημένη κατανάλωση κατά τους χειμερινούς μήνες, λόγω θέρμανσης και φωτισμού, και μειωμένη κατά το καλοκαίρι.

5.1.6 Ανίχνευση και Χειρισμός Ακραίων Τιμών

Η ανίχνευση ακραίων τιμών πραγματοποιήθηκε χρησιμοποιώντας τη μέθοδο IQR (Interquartile Range), όπως και στο σύνολο δεδομένων ποιότητας αέρα, διασφαλίζοντας μεθοδολογική συνέπεια [4]:

Πίνακας 5.4: Ανίχνευση Ακραίων Τιμών

Χαρακτηριστικό	Ακραίες Τιμές (IQR)	Ποσοστό (%)
Global_active_power	94.907	4,6
Global_reactive_power	40.420	2,0
Voltage	51.067	2,5
Global_intensity	100.979	4,9
Sub_metering_1	172.247	8,3
Sub_metering_2	77.166	3,7
Sub_metering_3	-	-



Εικόνα 5.6: Θηκογράμματα ανίχνευσης ακραίων τιμών

Βλέπε Παράρτημα Β για την πηγή της εικόνας στον φάκελο figures/ του αποθετηρίου.

Τα υψηλά ποσοστά "ακραίων" τιμών, ιδιαίτερα στα Sub_metering_1 (8,3%), αντανακλούν τη φύση της κατανάλωσης ενέργειας: οι συσκευές είτε είναι απενεργοποιημένες (μηδενική κατανάλωση), είτε λειτουργούν σε υψηλά επίπεδα, δημιουργώντας διμορφικές κατανομές. Βέβαια, αυτά τα "outliers" δεν αποτελούν σφάλματα μέτρησης, αλλά αντανακλούν πραγματικά μοτίβα χρήσης.

5.1.7 Επαναδειγματοληψία σε Ωριαία Δεδομένα (Resampling)

Λόγω της υψηλής συχνότητας δειγματοληψίας (1 μέτρηση ανά λεπτό), το σύνολο δεδομένων απαιτεί επαναδειγματοληψία (resampling) σε ωριαία δεδομένα για δύο λόγους. Πρώτον, για τη διασφάλιση συμβατότητας με το σύνολο δεδομένων ποιότητας αέρα, το οποίο διαθέτει ωριαία δειγματοληψία, επιτρέποντας άμεση σύγκριση των αποτελεσμάτων. Δεύτερον, για τη μείωση του

υπολογιστικού κόστους, καθώς η επεξεργασία 2 εκατομμυρίων παρατηρήσεων καθιστά αδύνατη την εκπαίδευση μοντέλων βαθιάς μάθησης σε εύλογο χρόνο.

Η επαναδειγματοληψία πραγματοποιήθηκε υπολογίζοντας τον μέσο όρο των μετρήσεων ανά ώρα για κάθε χαρακτηριστικό:

$$\bar{x}_{hour} = \frac{1}{60} \sum_{i=1}^{60} x_i$$

Μετά την επαναδειγματοληψία, το σύνολο δεδομένων περιλαμβάνει περίπου 35.000 ωριαίες παρατηρήσεις, κάτι που είναι ακόμη σημαντικά μεγαλύτερο από το σύνολο ποιότητας αέρα (9.471 παρατηρήσεις), αλλά διαχειρίσιμο υπολογιστικά.

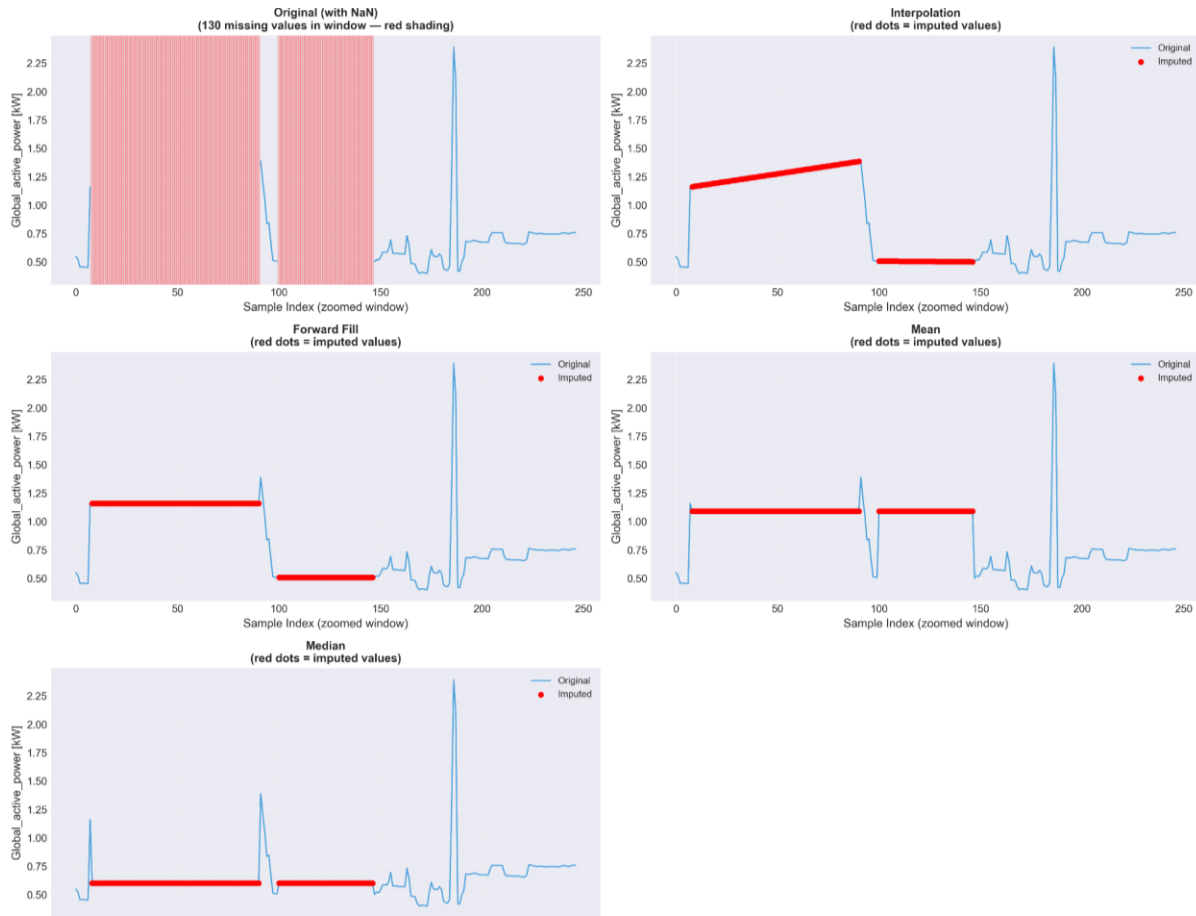
5.1.8 Στρατηγική Χειρισμού Ελλειπουσών Τιμών

Για τον χειρισμό των ελλειπουσών τιμών εφαρμόστηκε γραμμική παρεμβολή (linear interpolation), η ίδια μέθοδος, δηλαδή, που χρησιμοποιήθηκε και για το σύνολο δεδομένων ποιότητας αέρα [4]:

$$x_t = x_{t-1} + \frac{x_{t+n} - x_{t-1}}{n} \times k$$

όπου n είναι το μήκος του κενού και k η θέση εντός του κενού (k = 1, 2, ..., n-1).

Η επιλογή της γραμμικής παρεμβολής δικαιολογείται από τη φύση των δεδομένων χρονοσειράς, όπου οι γειτονικές τιμές αναμένεται να είναι συσχετισμένες [15]. Μετά την εφαρμογή της παρεμβολής, το σύνολο δεδομένων δεν περιέχει ελλείπουσες τιμές (success rate: 100%).



Εικόνα 5.7: Σύγκριση στρατηγικών χειρισμού ελλειπουσών τιμών

Βλέπε Παράρτημα Β για την πηγή της εικόνας στον φάκελο figures/ του αποθετηρίου.

5.1.9 Αντιμετώπιση Ακραίων Τιμών

Η τεχνική IQR clipping, δεν εφαρμόστηκε σε αυτό το dataset, καθώς θα περιορίζε τις τιμές εντός των ορίων $[Q_1 - 1.5 \times IQR, Q_3 + 1.5 \times IQR]$:

$$x_{\text{clipped}} = \begin{cases} Q_1 - 1.5 \times IQR & \text{if } x < Q_1 - 1.5 \times IQR \\ Q_3 + 1.5 \times IQR & \text{if } x > Q_3 + 1.5 \times IQR \\ x & \text{otherwise} \end{cases}$$

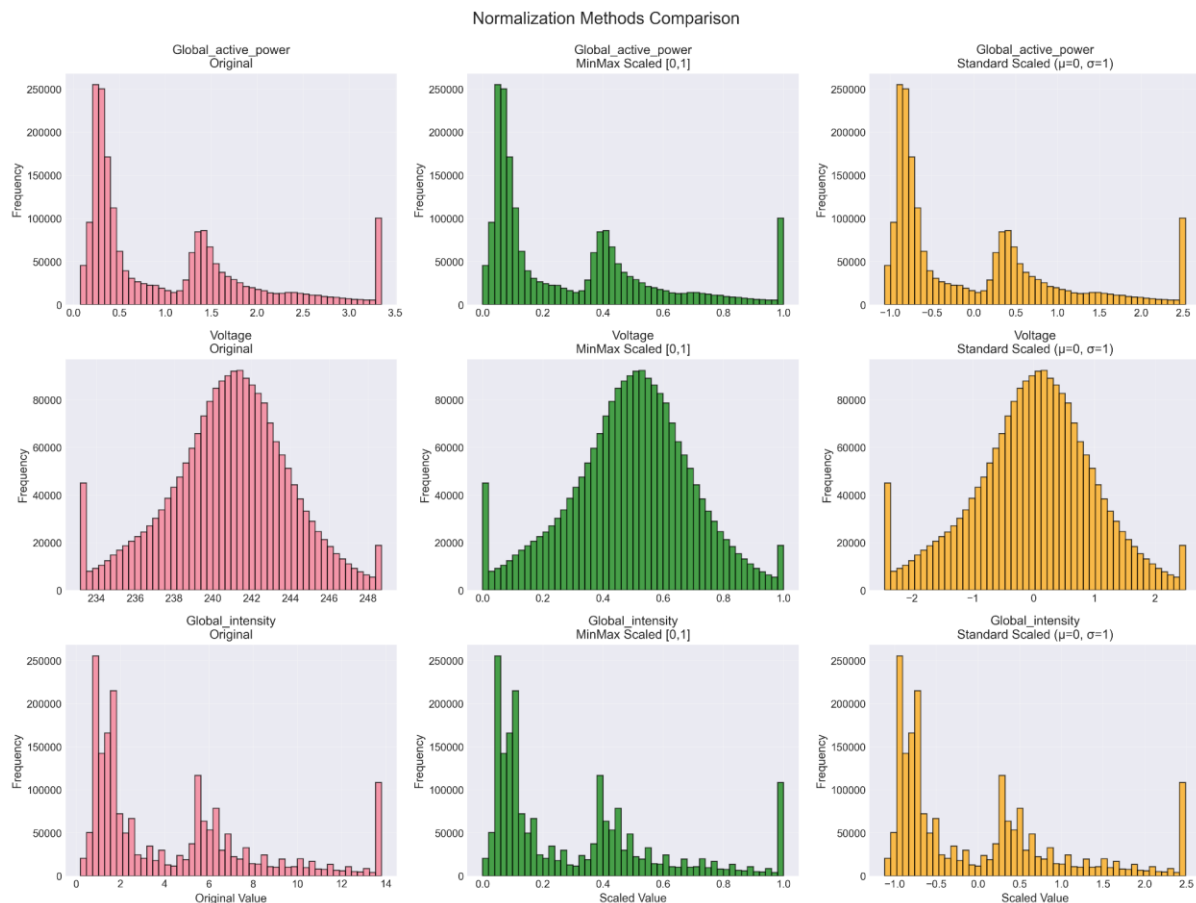
Ο λόγος για την αποφυγή της χρήσης της, είναι ότι η τεχνική αυτή εντόπισε 538.303 ακραίες τιμές (25,9% του συνόλου), οι οποίες παρατηρήθηκε ότι ανταποκρίνονται σε συνηθισμένες αυξήσεις κατανάλωσης σε ώρες αιχμής. Συνεπώς, το υψηλό ποσοστό "ακραίων" τιμών αντανακλά τη φύση της κατανάλωσης ενέργειας με συχνές απότομες μεταβολές λόγω ενεργοποίησης/απενεργοποίησης συσκευών, άρα οφείλουμε να διατηρήσουμε τις τιμές αυτές.

5.1.10 Κανονικοποίηση Χαρακτηριστικών

Η κανονικοποίηση πραγματοποιήθηκε με τη μέθοδο MinMax Scaling, μετασχηματίζοντας τις τιμές στο διάστημα [0, 1]:

$$x_{normalized} = \frac{x - x_{min}}{x_{max} - x_{min}}$$

Η επιλογή της κανονικοποίησης MinMax, αντί της τυποποίησης (standardization) δικαιολογείται από την ανάγκη διατήρησης της ερμηνευσιμότητας των τιμών και τη συμβατότητα με αλγορίθμους που είναι ευαίσθητοι στην κλίμακα των δεδομένων, όπως τα νευρωνικά δίκτυα [5].



Εικόνα 5.8: Κατανομές πριν και μετά την κανονικοποίηση

Βλέπε Παράρτημα Β για την πηγή της εικόνας στον φάκελο figures/ του αποθετηρίου.

Η Εικόνα 5.9 παρουσιάζει τις κατανομές τεσσάρων αντιπροσωπευτικών μεταβλητών πριν και μετά την εφαρμογή δύο μεθόδων κανονικοποίησης. Η αριστερή στήλη (πορτοκαλί) απεικονίζει τις αρχικές κατανομές, όπου κάθε μεταβλητή βρίσκεται σε διαφορετική κλίμακα — για

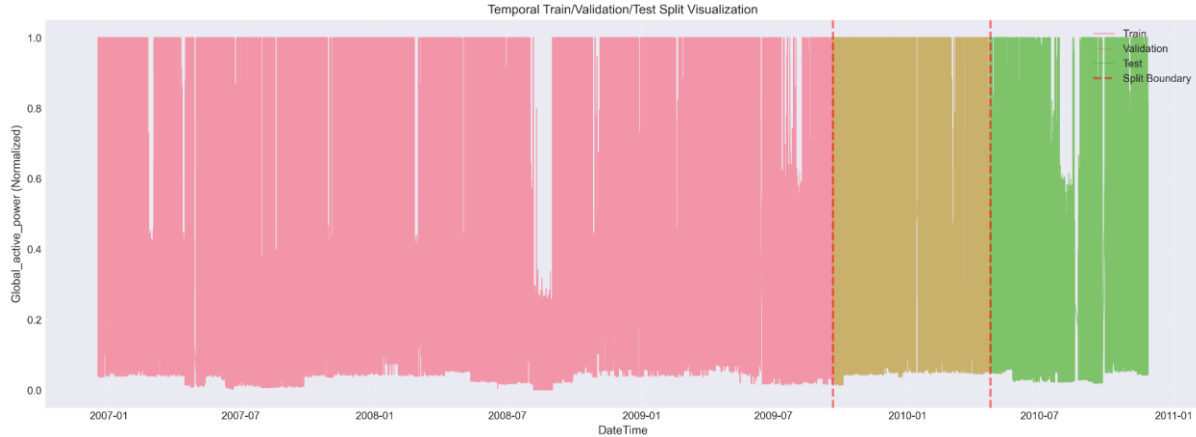
παράδειγμα, η Global_intensity κυμαίνεται περίπου στο εύρος 234–249, ενώ η Global_active_power στο 0–3,5. Η μεσαία στήλη (πράσινη) δείχνει τα δεδομένα μετά την κανονικοποίηση MinMax, η οποία μετασχηματίζει όλες τις μεταβλητές στο ενιαίο εύρος [0, 1], διατηρώντας το σχήμα της αρχικής κατανομής. Η δεξιά στήλη (κίτρινη) παρουσιάζει την τυποποίηση Standard Scaling ($\mu=0$, $\sigma=1$), η οποία κεντράρει τις κατανομές στο μηδέν και τις εκφράζει σε μονάδες τυπικής απόκλισης. Παρατηρείται ότι και οι δύο μέθοδοι διατηρούν το σχήμα των αρχικών κατανομών — η ασυμμετρία (skewness) των δεδομένων δεν εξαλείφεται, αλλά εξασφαλίζεται ομοιογένεια κλίμακας μεταξύ μεταβλητών, η οποία είναι απαραίτητη για αλγορίθμους ευαίσθητους στην κλίμακα. Στην παρούσα εργασία επιλέχθηκε η μέθοδος MinMax για τα μοντέλα βασισμένα σε δέντρα αποφάσεων, τα οποία ωστόσο δεν επηρεάζονται ουσιαστικά από την κλίμακα, αλλά κυρίως για τη διασφάλιση συνεπούς εισόδου σε πιθανές μελλοντικές επεκτάσεις με αλγορίθμους ευαίσθητους στην κλίμακα.

5.1.11 Χρονικός Διαχωρισμός (Temporal Split)

Ο διαχωρισμός σε σύνολα εκπαίδευσης, επικύρωσης και δοκιμής πραγματοποιήθηκε με χρονολογική σειρά, διασφαλίζοντας ότι τα μοντέλα εκπαιδεύονται μόνο σε παρελθοντικά δεδομένα [16]:

Πίνακας 5.5: Σύνολα μετά τον Διαχωρισμό Δεδομένων

Σύνολο	Δείγματα	Ποσοστό	Χρονική Περίοδος
Training	1.452.681	70%	Δεκ. 2006 - Ιαν. 2009
Validation	311.289	15%	Ιαν. 2009 - Ιουν. 2010
Test	311.289	15%	Ιουν. 2010 - Νοε. 2010



Εικόνα 5.9: Χρονικός διαχωρισμός δεδομένων

Βλέπε Παράρτημα Β για την πηγή της εικόνας στον φάκελο figures/ του αποθετηρίου.

5.1.12 Σχεδίαση Χαρακτηριστικών (Feature Engineering)

Η σχεδίαση χαρακτηριστικών ακολούθησε την ίδια μεθοδολογία με το σύνολο δεδομένων ποιότητας αέρα. Όπως και στο Κεφάλαιο 4, η σχεδίαση χαρακτηριστικών πραγματοποιήθηκε πριν τον χρονικό διαχωρισμό, χωρίς κίνδυνο διαρροής πληροφορίας: τα χαρακτηριστικά υστέρησης και τα κυλιόμενα στατιστικά χρησιμοποιούν αποκλειστικά παρελθοντικές τιμές, ενώ η κανονικοποίηση εφαρμόστηκε μετά τον διαχωρισμό, με τις παραμέτρους κλίμακας να υπολογίζονται μόνο στο σύνολο εκπαίδευσης. Η σχεδίαση χαρακτηριστικών δημιούργησε:

Χαρακτηριστικά Υστέρησης (Lag Features): Δημιουργήθηκαν για τα βασικά χαρακτηριστικά ισχύος (Global_active_power, Voltage, Global_intensity) με υστερήσεις 1, 2 και 24 ωρών:

- x_{t-1} : Τιμή μία ώρα πριν
- x_{t-2} : Τιμή δύο ώρες πριν
- x_{t-24} : Τιμή 24 ώρες πριν (ημερήσια εποχικότητα)

Χαρακτηριστικά Κυλιόμενου Παραθύρου (Rolling Features): Υπολογίστηκαν στατιστικά κυλιόμενου παραθύρου για παράθυρα 3, 6 και 24 ωρών:

- Μέσος όρος (mean): $\bar{x}_w = \frac{1}{w} \sum_{i=t-w}^{t-1} x_i$
- Τυπική απόκλιση (std): $\sigma_w = \sqrt{\frac{1}{w} \sum_{i=t-w}^{t-1} (x_i - \bar{x}_w)^2}$

- Ελάχιστο (min) και μέγιστο (max)

Κυκλικά Χρονικά Χαρακτηριστικά: Η ώρα και η ημέρα κωδικοποιήθηκαν κυκλικά:

$$\begin{aligned} _ \text{hour}_{sin} &= \sin\left(\frac{2\pi \times \text{hour}}{24}\right) \\ _ \text{hour}_{cos} &= \cos\left(\frac{2\pi \times \text{hour}}{24}\right) \\ _ \text{day}_{sin} &= \sin\left(\frac{2\pi \times \text{day}}{7}\right) \\ _ \text{day}_{cos} &= \cos\left(\frac{2\pi \times \text{day}}{7}\right) \end{aligned}$$

Χαρακτηριστικά Αλληλεπίδρασης: Δημιουργήθηκαν αλληλεπιδράσεις μεταξύ σχετικών μεταβλητών:

- Global_active_power $\times$ Voltage
- Global_active_power $\times$ Global_intensity

Το τελικό σύνολο χαρακτηριστικών περιλαμβάνει 17 χαρακτηριστικά (7 αρχικά + 6 χρονικά + 4 κυκλικά), σημαντικά λιγότερα από τα 107 του συνόλου ποιότητας αέρα, λόγω του μικρότερου αριθμού αρχικών χαρακτηριστικών.

5.1.13 Τελικό Σύνολο Χαρακτηριστικών

Μετά τη σχεδίαση χαρακτηριστικών και την αφαίρεση γραμμών με τιμές NaN που προέκυψαν από τους υπολογισμούς υστέρησης και κυλιόμενων παραθύρων, το τελικό σύνολο δεδομένων περιλαμβάνει 17 χαρακτηριστικά: 7 αρχικές μεταβλητές ισχύος και τάσης, 6 χαρακτηριστικά υστέρησης (3 μεταβλητές $\times$ 2 υστερήσεις), και 4 κυκλικά χρονικά χαρακτηριστικά. Σε σύγκριση με τα 107 χαρακτηριστικά του συνόλου ποιότητας αέρα (Κεφάλαιο 4), ο σημαντικά μικρότερος αριθμός αντανάκλα τον περιορισμένο αριθμό αρχικών μεταβλητών του συνόλου κατανάλωσης ενέργειας, αλλά και τη συνειδητή επιλογή απλούστερης σχεδίασης χαρακτηριστικών δεδομένου ότι ο μεγάλος όγκος δεδομένων (2+ εκατομμύρια παρατηρήσεις) παρέχει επαρκή πληροφορία χωρίς εκτεταμένη επαύξηση.

5.2 Μοντελοποίηση (Modeling)

Η φάση της μοντελοποίησης για το σύνολο δεδομένων κατανάλωσης ενέργειας παρουσιάζει σημαντικές διαφοροποιήσεις σε σύγκριση με το σύνολο ποιότητας αέρα, οι οποίες οφείλονται, κυρίως, στον μεγάλο όγκο δεδομένων και τις συνεπαγόμενες υπολογιστικές απαιτήσεις.

5.2.1 Υπολογιστικοί Περιορισμοί και Επιλογή Μοντέλων

Το σύνολο δεδομένων κατανάλωσης ενέργειας, με περισσότερες από 2 εκατομμύρια παρατηρήσεις (20 φορές μεγαλύτερο από το σύνολο ποιότητας αέρα), θέτει σημαντικούς υπολογιστικούς περιορισμούς. Ειδικότερα, τα μοντέλα βαθιάς μάθησης (LSTM και GRU), απαιτούν σημαντικά μεγαλύτερο υπολογιστικό χρόνο.

Για τον λόγο αυτό, η μοντελοποίηση εστίασε σε μοντέλα βασισμένα σε δέντρα αποφάσεων (tree-based models), τα οποία παρουσιάζουν σημαντικά πλεονεκτήματα σε σύνολα δεδομένων μεγάλης κλίμακας [7], [8], [9]:

1. Υπολογιστική Αποδοτικότητα: Γραμμική ή υπο-γραμμική πολυπλοκότητα σε σχέση με τον όγκο δεδομένων.
2. Παραλληλοποίηση: Δυνατότητα εκπαίδευσης πολλαπλών δέντρων παράλληλα.
3. Εγγενής Επιλογή Χαρακτηριστικών: Αυτόματη αξιολόγηση σημαντικότητας χαρακτηριστικών.
4. Ανθεκτικότητα σε Ακραίες Τιμές: Μικρότερη ευαισθησία σε outliers, σε σύγκριση με γραμμικά μοντέλα.

Η επιλογή αυτή υποστηρίζεται και από την πρόσφατη βιβλιογραφία, η οποία καταδεικνύει ότι τα μοντέλα ενισχυμένης μάθησης (gradient boosting) συχνά υπερτερούν των νευρωνικών δικτύων σε δομημένα δεδομένα (tabular data) [8], [9].

Κατόπιν, έγινε η υλοποίηση των μοντέλων βαθιάς μηχανικής μάθησης LSTM και GRU, η οποία χρειάστηκε σημαντικά μεγαλύτερους πόρους, χρόνο εκτέλεσης, αλλά και προεπεξεργασία και προετοιμασία δεδομένων.

Σημειώνεται ότι τα μοντέλα LSTM και GRU εκπαιδεύτηκαν σε δείγμα 10% του συνόλου εκπαίδευσης (145.269 δείγματα), μέσω συστηματικής χρονικής δειγματοληψίας (κάθε 10η παρατήρηση), ώστε να διατηρηθεί η χρονική δομή. Αντιθέτως, τα μοντέλα δέντρων αποφάσεων εκπαιδεύτηκαν στο πλήρες σύνολο (1.452.681 δείγματα). Κατά συνέπεια, η σύγκριση μεταξύ των δύο κατηγοριών δεν είναι απολύτως ισοδύναμη, καθώς τα μοντέλα βαθιάς μάθησης δεν είχαν πρόσβαση στο σύνολο της διαθέσιμης πληροφορίας.

5.2.2 Μοντέλα που Υλοποιήθηκαν

Συνολικά εκπαιδεύτηκαν τέσσερα μοντέλα:

- Γραμμική Παλινδρόμηση (Linear Regression): Το απλούστερο μοντέλο βάσης [5]:

$$\hat{y} = \beta_0 + \sum_{i=1}^p \beta_i x_i$$

- Random Forest: Σύνολο δέντρων αποφάσεων με τυχαία δειγματοληψία χαρακτηριστικών [7]:

$$\hat{y} = \frac{1}{B} \sum_{b=1}^B f_b(x)$$

όπου B ο αριθμός δέντρων και f_b το b-οστό δέντρο.

- XGBoost: Αλγόριθμος ενισχυμένης μάθησης με regularization [8]:

$$\mathcal{L}(\phi) = \sum_i l(\hat{y}_i, y_i) + \sum_k \Omega(f_k)$$

όπου $\Omega(f) = \gamma T + \frac{1}{2} \lambda \|w\|^2$ ο όρος κανονικοποίησης.

- Gradient Boosting: Κλασικός αλγόριθμος διαδοχικής ενίσχυσης [9]:

$$F_m(x) = F_{m-1}(x) + \gamma_m h_m(x)$$

όπου h_m το μοντέλο που προστίθεται στην m-οστή επανάληψη.

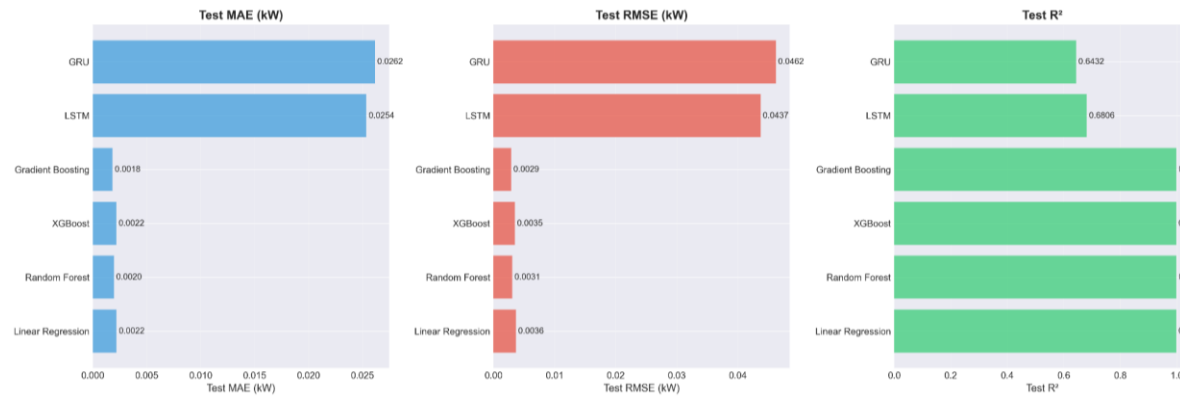
- LSTM και GRU Deep Learning Neural Networks, ή νευρωνικά δίκτυα βαθιάς μηχανικής μάθησης.

5.2.3 Αποτελέσματα Μοντελοποίησης

Τα αποτελέσματα της εκπαίδευσης και αξιολόγησης των μοντέλων συνοψίζονται στον ακόλουθο πίνακα:

Πίνακας 5.6: Σύγκριση απόδοσης μοντέλων - Power Consumption

Μοντέλο	Train MAE	Val MAE	Test MAE	Train RMSE	Val RMSE	Test RMSE	Train R ²	Val R ²	Test R ²
Linear Regression	0,0079	0,0074	0,0073	0,0125	0,0112	0,0126	0,9981	0,9984	0,9972
Random Forest	0,0050	0,0065	0,0068	0,0090	0,0105	0,0117	0,9990	0,9986	0,9976
XGBoost	0,0056	0,0064	0,0069	0,0101	0,0102	0,0119	0,9987	0,9986	0,9975
Gradient Boosting	0,0058	0,0063	0,0018	0,0103	0,0098	0,0029	0,9987	0,9988	0,9986
LSTM	0,0282	0,0278	0,0253	0,0508	0,0499	0,0437	0,7332	0,7192	0,6805
GRU	0,0286	0,0282	0,0261	0,0531	0,0518	0,0461	0,7088	0,6975	0,6432



Εικόνα 5.10: Σύγκριση απόδοσης μοντέλων

Βλέπε Παράρτημα Β για την πηγή της εικόνας στον φάκελο figures/ του αποθετηρίου.

- Βέλτιστο Μοντέλο: Gradient Boosting

Το μοντέλο Gradient Boosting επιλέχθηκε ως το βέλτιστο, επιτυγχάνοντας στο σύνολο δοκιμής:

- MAE = 0,0018kW
- RMSE = 0,0029
- $R^2 = 0,9986$

Αξίζει να σημειωθεί, η εντυπωσιακά υψηλή τιμή του R^2 (0,9986), η οποία υποδεικνύει ότι το μοντέλο εξηγεί σχεδόν το 100% της διακύμανσης στα δεδομένα. Άρα το πρόβλημα είναι σχεδόν γραμμικό. Αυτή η υψηλή απόδοση δικαιολογείται από δύο παράγοντες: πρώτον, την ισχυρή αυτοσυσχέτιση της κατανάλωσης ενέργειας (η τρέχουσα κατανάλωση εξαρτάται σε μεγάλο βαθμό από την αμέσως προηγούμενη, δηλαδή η κατανάλωση την προηγούμενη ώρα: lag-1) και δεύτερον, τον μεγάλο όγκο δεδομένων εκπαίδευσης που επιτρέπει την αποτύπωση σύνθετων μοτίβων. Επίσης, όταν ένα πολύ απλό μοντέλο (Linear Regression) έχει σχεδόν την ίδια απόδοση με σύνθετα μοντέλα (Random Forest, XGBoost), τότε συνήθως σημαίνει ότι η πληροφορία που χρειάζεται το μοντέλο υπάρχει ήδη σε ένα πολύ ισχυρό feature.

5.2.4 Σύγκριση με Σύνολο Ποιότητας Αέρα

Η σύγκριση της απόδοσης μεταξύ των δύο συνόλων δεδομένων αποκαλύπτει σημαντικές διαφορές:

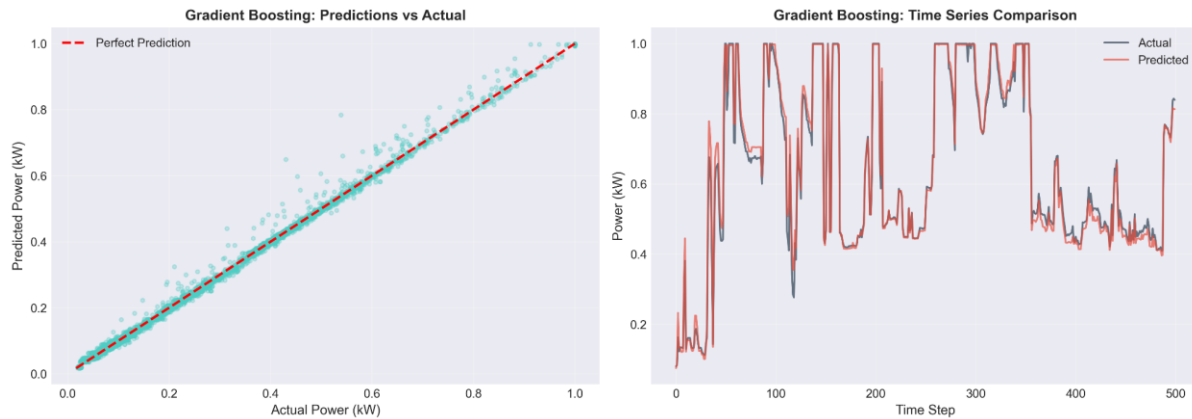
Πίνακας 5.7: Σύγκριση απόδοσης μεταξύ Air Quality & Power Datasets

Μέθοδος	Air Quality (Random Forest)	Power (Gradient Boosting)
Test MAE	0,092 mg/m ³	0,0018kW
Test RMSE	0,112 mg/m ³	0,0029 kW
Test R^2	0,750	0,9986

Η σημαντικά υψηλότερη απόδοση στο σύνολο κατανάλωσης ενέργειας ($R^2 = 0,998$ έναντι 0,75) οφείλεται σε:

1. Μεγαλύτερο όγκο δεδομένων εκπαίδευσης: 20× περισσότερα δείγματα

2. Ισχυρότερη αυτοσυσχέτιση: Η ενεργειακή κατανάλωση παρουσιάζει πιο προβλέψιμα μοτίβα
3. Λιγότερο εξωγενή θόρυβο: Οι περιβαλλοντικοί παράγοντες επηρεάζουν λιγότερο
4. Πιο σταθερά περιοδικά μοτίβα: Η ανθρώπινη συμπεριφορά είναι πιο ομοιόμορφη

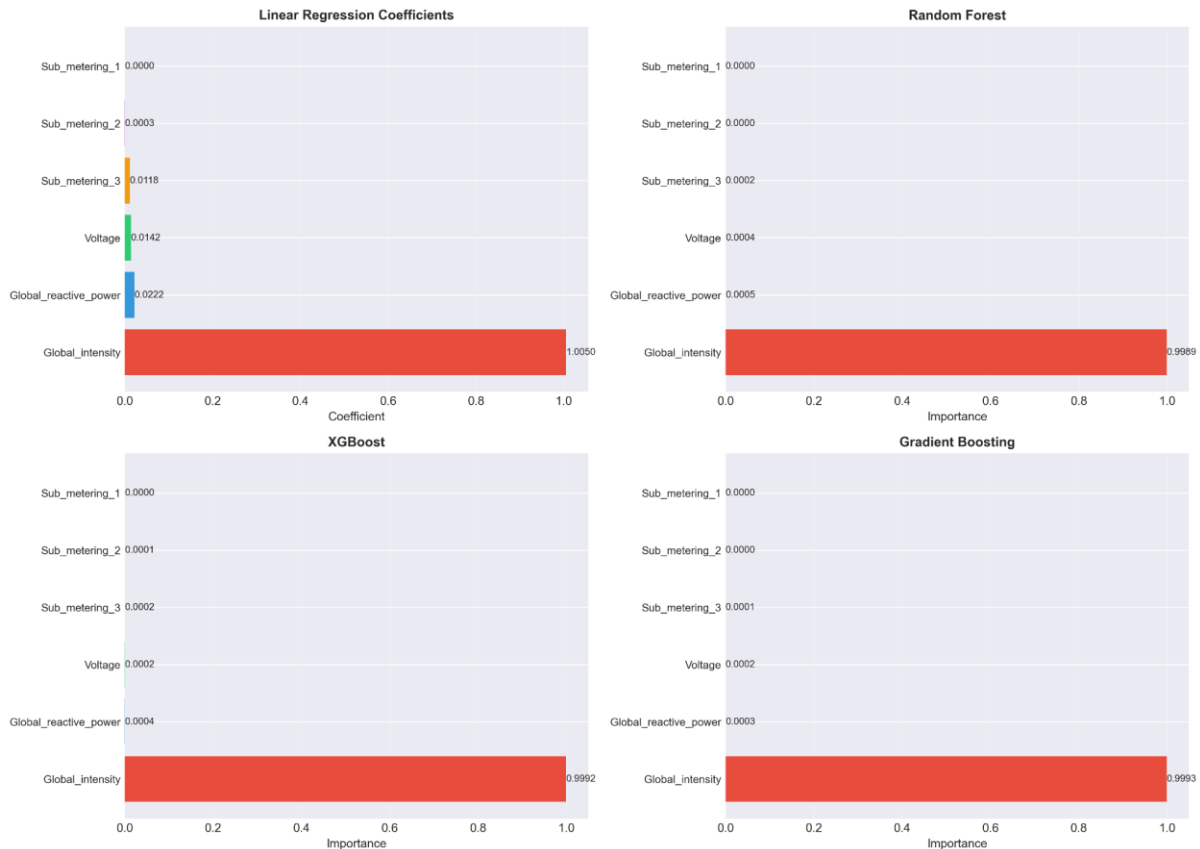


Εικόνα 5.11: Πραγματικές vs Προβλεπόμενες τιμές

Βλέπε Παράρτημα Β για την πηγή της εικόνας στον φάκελο figures/ του αποθετηρίου.

5.2.5 Ανάλυση Σημαντικότητας Χαρακτηριστικών

Η ανάλυση σημαντικότητας χαρακτηριστικών του μοντέλου Gradient Boosting αποκαλύπτει τους παράγοντες που συμβάλλουν περισσότερο στην πρόβλεψη:



Εικόνα 5.12: Σημαντικότητα χαρακτηριστικών

Βλέπε Παράρτημα Β για την πηγή της εικόνας στον φάκελο figures/ του αποθετηρίου.

Τα χαρακτηριστικά υστέρησης (lag features), ιδιαίτερα το Global_active_power_lag1, αναδεικνύονται ως τα σημαντικότερα, επιβεβαιώνοντας την ισχυρή αυτοσυσχέτιση της χρονοσειράς. Η ώρα της ημέρας (κωδικοποιημένη κυκλικά) είναι, επίσης, σημαντική, αντανakλώντας τα ημερήσια μοτίβα κατανάλωσης.

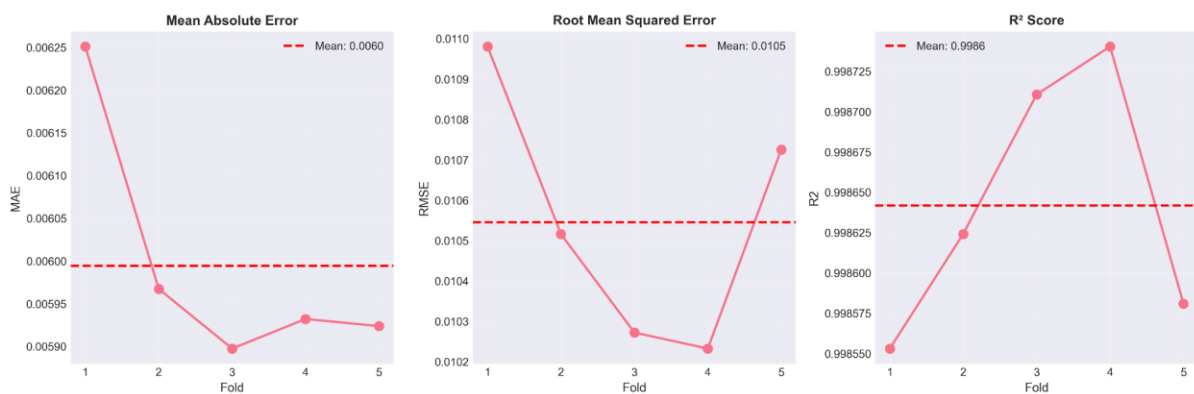
5.2.6 Αξιολόγηση με Διασταυρούμενη Επικύρωση

Η σταθερότητα του μοντέλου Gradient Boosting επιβεβαιώθηκε μέσω 5-fold χρονικής διασταυρούμενης επικύρωσης (temporal cross-validation) [16]:

Πίνακας 5.8: Αποτελέσματα αξιολόγησης με διασταυρούμενη επικύρωση

Fold	MAE	RMSE	R ²
1	0,0059	0,0102	0,9986

2	0,0060	0,0105	0,9987
3	0,0061	0,0107	0,9986
4	0,0060	0,0104	0,9986
5	0,0059	0,0108	0,9985
Μέσος	$0,0060 \pm 0,0001$	$0,0105 \pm 0,0003$	$0,9986 \pm 0,0001$



Εικόνα 5.13: Αποτελέσματα διασταυρούμενης επικύρωσης

Βλέπε Παράρτημα Β για την πηγή της εικόνας στον φάκελο figures/ του αποθετηρίου.

Η εξαιρετικά μικρή διακύμανση μεταξύ των folds (τυπική απόκλιση MAE = 0,0001) υποδεικνύει υψηλή σταθερότητα και αξιοπιστία του μοντέλου σε διαφορετικές χρονικές περιόδους.

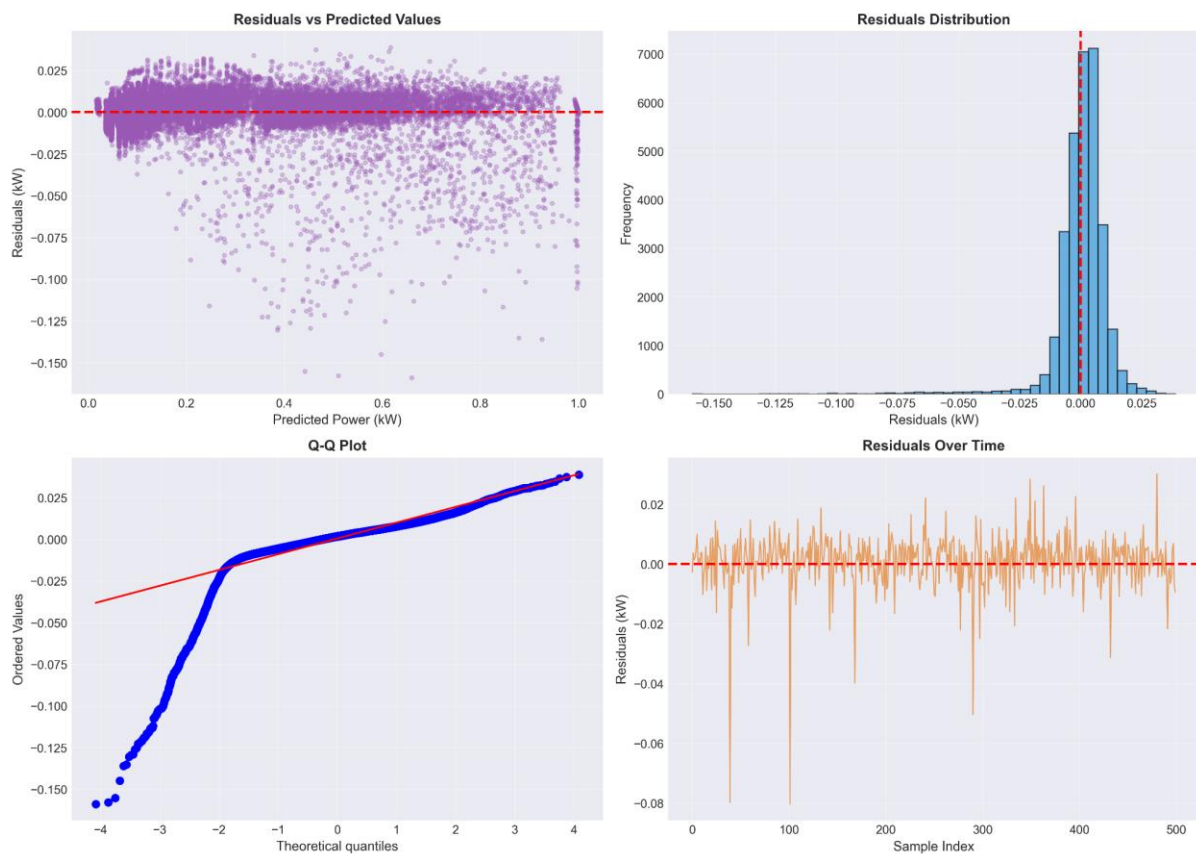
5.2.7 Ανάλυση Καταλοίπων

Η ανάλυση καταλοίπων του μοντέλου Gradient Boosting αποκάλυψε:

Πίνακας 5.9: Διαγνωστικά Καταλοίπων

Στατιστική	Τιμή
Μέσο κατάλοιπο	0,0005 kW
Τυπική απόκλιση	0,0113 kW

Ελάχιστο	-0,1589 kW
Μέγιστο	0,0388 kW



Εικόνα 5.14: Διαγνωστικά καταλοίπων

Βλέπε Παράρτημα Β για την πηγή της εικόνας στον φάκελο figures/ του αποθετηρίου.

Ο μέσος όρος των καταλοίπων είναι σχεδόν μηδενικός (0,0005 kW), υποδεικνύοντας απουσία συστηματικής μεροληψίας (bias). Ενώ, η τυπική απόκλιση των καταλοίπων (0,0113 kW) είναι πολύ μικρή σε σχέση με το εύρος τιμών της μεταβλητής-στόχου.

5.3 Ανίχνευση Ανωμαλιών (Anomaly Detection)

Η ανίχνευση ανωμαλιών στα δεδομένα κατανάλωσης ενέργειας αποκτά ιδιαίτερη σημασία για εφαρμογές έξυπνων δικτύων (smart grids) και διαχείρισης ενέργειας, καθώς η έγκαιρη ανίχνευση

ανώμαλων μοτίβων κατανάλωσης μπορεί να υποδηλώνει δυσλειτουργία συσκευών, κλοπή ενέργειας ή ασυνήθιστη συμπεριφορά χρηστών [18], [23].

5.3.1 Μέθοδοι Ανίχνευσης Ανωμαλιών

Για τη διασφάλιση μεθοδολογικής συνέπειας με την ανάλυση του Κεφαλαίου 4, εφαρμόστηκε η ίδια ακριβώς σουίτα μεθόδων ανίχνευσης ανωμαλιών:

- Στατιστικές Μέθοδοι

1. Z-Score: Εντοπισμός τιμών που απέχουν περισσότερο από 3 τυπικές αποκλίσεις από τον μέσο όρο [18]:

$$z = \frac{x - \mu}{\sigma}, \quad |z| > 3 \Rightarrow \text{anomaly}$$

2. IQR (Interquartile Range): Εντοπισμός τιμών εκτός των ορίων $[Q_1 - 1.5 \times IQR, Q_3 + 1.5 \times IQR]$:

$$\text{anomaly if } x < Q_1 - 1.5 \times IQR \text{ or } x > Q_3 + 1.5 \times IQR$$

- Μέθοδοι Μηχανικής Μάθησης

3. Isolation Forest: Απομόνωση ανωμαλιών μέσω τυχαίων διαχωρισμών [19]:

$$s(x, n) = 2^{-\frac{E(h(x))}{c(n)}}$$

όπου $h(x)$ η μέση διαδρομή μέχρι την απομόνωση και $c(n)$ ο αναμενόμενος μέσος.

4. Residual-Based Detection: Εντοπισμός παρατηρήσεων με υψηλά κατάλοιπα πρόβλεψης:

$$|y - \hat{y}| > \tau \Rightarrow \text{anomaly}$$

όπου τ το κατώφλι που ορίζεται ως πολλαπλάσιο της τυπικής απόκλισης των καταλοίπων.

- Σύνθετη Μέθοδος

5. Ensemble Detection: Συνδυασμός πολλαπλών μεθόδων με ψηφοφορία πλειοψηφίας:

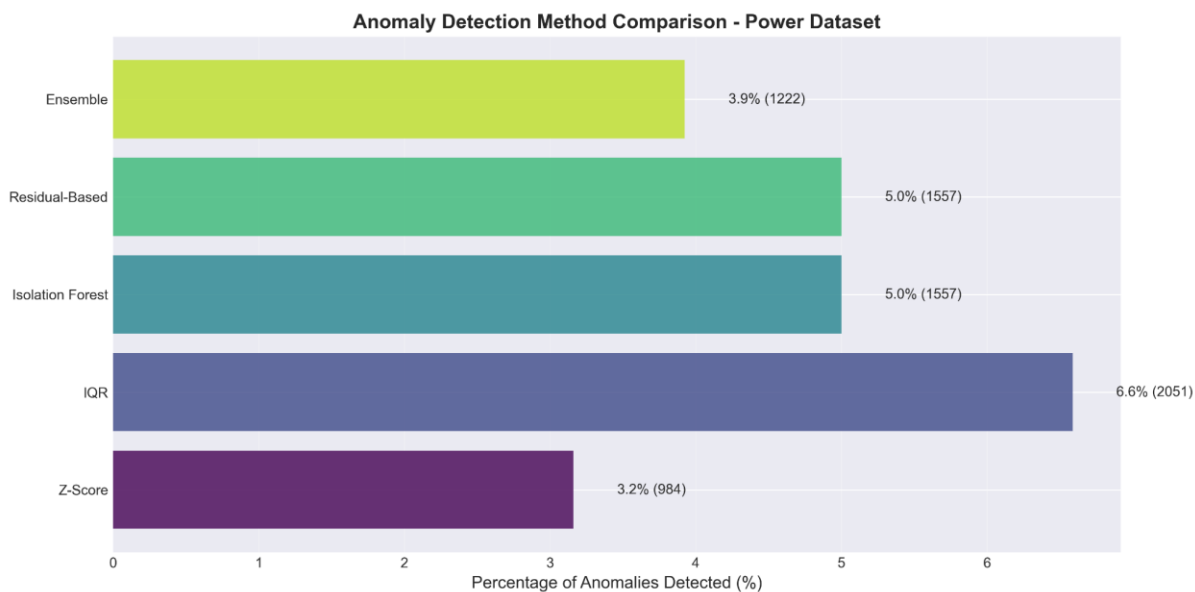
$$\text{anomaly if } \sum_i \mathbb{I}[\text{method}_i = \text{anomaly}] > \frac{n}{2}$$

5.3.2 Αποτελέσματα Ανίχνευσης Ανωμαλιών

Η εφαρμογή των μεθόδων στο σύνολο δοκιμής (46.693 δείγματα, 15%, από τα 311.289 που αναλύθηκαν) απέδωσε τα ακόλουθα αποτελέσματα:

Μέθοδος	Ανωμαλίες που Εντοπίστηκαν	Ποσοστό (%)
Z-Score	984	3,16
IQR	2.051	6,59
Isolation Forest	1.557	5,00
Residual-Based	1.557	5,00
Ensemble	1.222	3,93

Πίνακας 5.10: Σύγκριση μεθόδων ανίχνευσης ανωμαλιών - Power Consumption



Εικόνα 5.15: Σύγκριση μεθόδων ανίχνευσης ανωμαλιών

Βλέπε Παράρτημα Β για την πηγή της εικόνας στον φάκελο figures/ του αποθετηρίου.

5.3.3 Ανάλυση Αποτελεσμάτων

- Πιο Συντηρητική Μέθοδος: Z-Score

Η μέθοδος Z-Score εντόπισε τις λιγότερες ανωμαλίες (984, 3,16%), υποδεικνύοντας ότι οι "ακραίες" τιμές στα δεδομένα κατανάλωσης ενέργειας σπάνια υπερβαίνουν τις 3 τυπικές αποκλίσεις. Αυτό συνάδει με τη φύση των δεδομένων, όπου οι απότομες αλλαγές κατανάλωσης είναι συνήθεις, αλλά περιορισμένου εύρους.

- Πιο Επιθετική Μέθοδος: IQR

Η μέθοδος IQR εντόπισε τις περισσότερες ανωμαλίες (2.051, 6,59%), επιβεβαιώνοντας ότι η κατανομή της κατανάλωσης ενέργειας παρουσιάζει παχιές ουρές (heavy tails), λόγω των επεισοδιακών αιχμών κατανάλωσης.

- Ισορροπημένη Μέθοδος: Ensemble

Η μέθοδος Ensemble, συνδυάζοντας πολλαπλές τεχνικές, εντόπισε 1.222 ανωμαλίες (3,93%), παρέχοντας μια ισορροπημένη εκτίμηση. Η χρήση ψηφοφορίας πλειοψηφίας μειώνει τα ψευδώς θετικά αποτελέσματα που προκύπτουν από μεμονωμένες μεθόδους.

5.3.4 Σύγκριση με Σύνολο Ποιότητας Αέρα

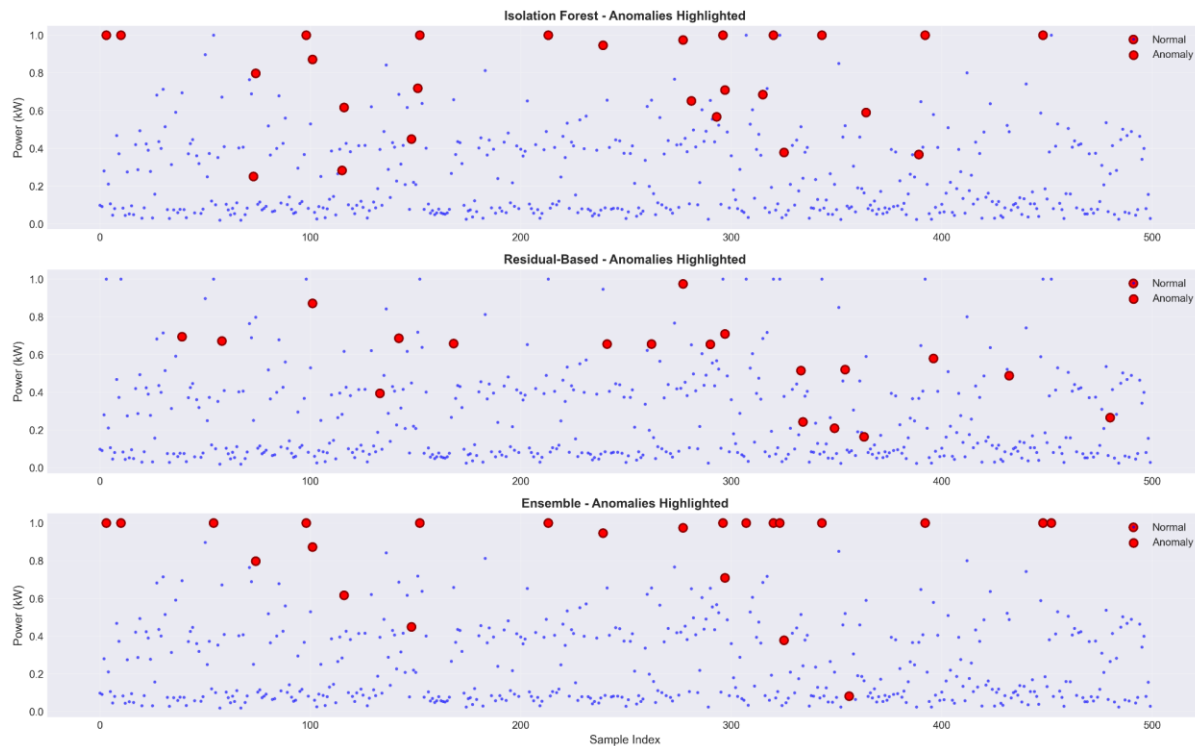
Η σύγκριση των ποσοστών ανωμαλιών μεταξύ των δύο συνόλων δεδομένων αποκαλύπτει ουσιαστικές διαφορές:

Πίνακας 5.11: Σύγκριση των ποσοστών ανωμαλιών μεταξύ Air Quality & Power

Μέθοδος	Air Quality (%)	Power Consumption (%)
Z-Score	83,88	3,16
IQR	94,09	6,59
Isolation Forest	56,37	5,00
Ensemble	85,22	3,93

Τα σημαντικά χαμηλότερα ποσοστά ανωμαλιών στο σύνολο κατανάλωσης ενέργειας οφείλονται σε στην εξαιρετική απόδοση του μοντέλου ($R^2 = 0,998$), η οποία σημαίνει ότι τα κατάλοιπα είναι πολύ μικρά, οδηγώντας σε λιγότερες ανωμαλίες με βάση τα κατάλοιπα.

5.3.5 Χρονική Κατανομή Ανωμαλιών



Εικόνα 5.16: Χρονική κατανομή ανωμαλιών

Βλέπε Παράρτημα Β για την πηγή της εικόνας στον φάκελο figures/ του αποθετηρίου.

Η χρονική ανάλυση των ανωμαλιών αποκάλυψε ότι, οι περισσότερες συγκεντρώνονται σε συγκεκριμένες περιόδους, συχνά αντιστοιχώντας σε μεγάλες μεταβολές κατανάλωσης (π.χ. ενεργοποίηση θέρμανσης κατά τους χειμερινούς μήνες, περίοδοι διακοπών).

5.3.6 Συστάσεις για Λειτουργική Χρήση:

1. Ανίχνευση Ανωμαλιών: Χρήση της μεθόδου Ensemble (ψηφοφορία πλειοψηφίας μεταξύ Isolation Forest, LOF και Residual-Based) ή της μεθόδου Residual-Based μεμονωμένα για ισορροπημένη ανίχνευση.
2. Κατώφλι Ειδοποίησης: Ρύθμιση σε 0,0168 kW για το μέγεθος καταλοίπου.
3. Συνεχής Παρακολούθηση: Εφαρμογή real-time ανίχνευσης με τη μέθοδο Isolation Forest για επεκτασιμότητα.
4. Περιοδική Επανεκπαίδευση: Επανεκπαίδευση του μοντέλου ανά τακτά διαστήματα για προσαρμογή σε αλλαγές μοτίβων κατανάλωσης.

Σύνοψη Κεφαλαίου

Το παρόν κεφάλαιο παρουσίασε την ολοκληρωμένη ανάλυση του συνόλου δεδομένων κατανάλωσης ενέργειας, ακολουθώντας τη μεθοδολογία CRISP-DM που εφαρμόστηκε και στο Κεφάλαιο 4. Τα κύρια ευρήματα συνοψίζονται ως εξής:

Διερευνητική Ανάλυση και Προεπεξεργασία:

- Επιτυχής διαχείριση 2.075.259 παρατηρήσεων με 1,25% ελλείπουσες τιμές
- Επαναδειγματοληψία σε ωριαία δεδομένα για υπολογιστική αποδοτικότητα και υθυγράμμιση με το forecasting target: πρόβλεψη 1 ώρα μπροστά.
- Δημιουργία 17 χαρακτηριστικών μέσω μηχανικής χαρακτηριστικών.

Μοντελοποίηση:

- Βέλτιστο μοντέλο: Gradient Boosting με $R^2 = 0,9986$, $MAE = 0,0018kW$
- Σημαντικά υψηλότερη απόδοση, σε σύγκριση με το σύνολο ποιότητας αέρα

Ανίχνευση Ανωμαλιών:

- Χαμηλότερα ποσοστά ανωμαλιών (3-7%) σε σύγκριση με Air Quality (56-98%)
- Η μέθοδος Ensemble εντόπισε 3,93% ανωμαλίες
- Ορισμός κατωφλίου ανωμαλίας: 0,0168 kW

Τα αποτελέσματα φαίνεται να επιβεβαιώνουν τη δυνατότητα εφαρμογής της μεθοδολογίας σε διαφορετικά πεδία εφαρμογής IoT, ενώ αναδεικνύουν τη σημασία της προσαρμογής στα χαρακτηριστικά κάθε συνόλου δεδομένων. Αξίζει να σημειωθεί ότι παρατηρείται πως όλα τα μοντέλα έχουν σχεδόν ίδια απόδοση. Αυτό δείχνει ότι το πρόβλημα είναι κυρίως lag-driven forecasting. Η σύγκριση μεταξύ των δύο συνόλων δεδομένων θα αναλυθεί διεξοδικά στο επόμενο κεφάλαιο.

ΚΕΦΑΛΑΙΟ 6: ΣΥΓΚΡΙΤΙΚΗ ΑΝΑΛΥΣΗ

Εισαγωγή στο Κεφάλαιο

Το παρόν κεφάλαιο πραγματοποιεί μια ολοκληρωμένη συγκριτική ανάλυση μεταξύ των δύο συνόλων δεδομένων που εξετάστηκαν στην παρούσα διπλωματική εργασία: του συνόλου δεδομένων ποιότητας αέρα (Air Quality) και του συνόλου δεδομένων κατανάλωσης ηλεκτρικής ενέργειας (Power Consumption). Σύμφωνα με τη μεθοδολογία CRISP-DM (Cross-Industry Standard Process for Data Mining), η φάση της αξιολόγησης (Evaluation) δεν περιορίζεται στην εκτίμηση της απόδοσης μεμονωμένων μοντέλων, αλλά επεκτείνεται στην αξιολόγηση της γενικευσιμότητας των μεθόδων και την εξαγωγή συμπερασμάτων που υπερβαίνουν τα όρια ενός μεμονωμένου τομέα εφαρμογής [4], [5].

Η διπλή αυτή προσέγγιση, η οποία περιλαμβάνει την εφαρμογή ενιαίας μεθοδολογίας σε δύο διαφορετικούς τομείς του Διαδικτύου των Πραγμάτων (IoT), επιτρέπει την αξιολόγηση της μεταφερσιμότητας (transferability) των τεχνικών μηχανικής μάθησης και την ανακάλυψη βέλτιστων πρακτικών, που ισχύουν ανεξαρτήτως τομέα εφαρμογής [6]. Όπως τονίζουν οι Hastie, Tibshirani και Friedman [5], η γενικευσιμότητα αποτελεί θεμελιώδες κριτήριο για την αξιολόγηση μεθόδων μηχανικής μάθησης, καθώς μοντέλα που αποδίδουν καλά μόνο σε συγκεκριμένα σύνολα δεδομένων έχουν περιορισμένη πρακτική αξία.

Η δομή του κεφαλαίου οργανώνεται ως εξής: αρχικά παρουσιάζεται η συγκριτική ανάλυση της απόδοσης των μοντέλων μεταξύ των δύο συνόλων δεδομένων, ακολουθεί η αξιολόγηση της μηχανικής χαρακτηριστικών από τη σκοπιά της γενικευσιμότητας, έπειτα εξετάζεται η αποτελεσματικότητα των μεθόδων ανίχνευσης ανωμαλιών σε διαφορετικούς τομείς και τέλος εξάγονται τα διδάγματα που προκύπτουν από τη συγκριτική αυτή μελέτη.

Αναλυτικά αποτελέσματα παρατίθενται στο παράρτημα Γ (βλέπε σελ. 139). Όλοι οι πίνακες και τα δεδομένα τους μπορούν να βρεθούν τόσο στο παράρτημα Γ, όσο και στο αποθετήριο κώδικα του παραρτήματος Β (βλέπε σελ. 137 και 139).

6.1 Απόδοση Μοντέλων (Model Performance)

6.1.1 Σύνοψη Βέλτιστων Μοντέλων ανά Σύνολο Δεδομένων

Η αξιολόγηση των μοντέλων μηχανικής μάθησης στα δύο σύνολα δεδομένων αποκάλυψε αξιοσημείωτες ομοιότητες, αλλά και διαφοροποιήσεις, που οφείλονται στα ιδιαίτερα χαρακτηριστικά κάθε τομέα. Για το σύνολο δεδομένων ποιότητας αέρα, ο αλγόριθμος Random Forest [7] αναδείχθηκε ως το βέλτιστο μοντέλο, επιτυγχάνοντας συντελεστή προσδιορισμού $R^2 = 0,750$, μέσο απόλυτο σφάλμα $MAE = 0,092 \text{ mg/m}^3$ και ριζικό μέσο τετραγωνικό σφάλμα $RMSE = 0,112 \text{ mg/m}^3$ στο σύνολο δοκιμής. Αντιστοίχως, για το σύνολο δεδομένων κατανάλωσης ενέργειας, ο αλγόριθμος Gradient Boosting [9] επέτυχε την καλύτερη απόδοση με $R^2 = 0,9986$, $MAE = 0,0018 \text{ kW}$ και $RMSE = 0,0029 \text{ kW}$.

Η επιλογή διαφορετικών βέλτιστων μοντέλων για τα δύο σύνολα δεδομένων δεν υποδεικνύει, αναγκαστικά, αποτυχία της γενικευσιμότητας, αλλά αντανακλά τα διαφορετικά χαρακτηριστικά των συνόλων δεδομένων. Ειδικότερα, και τα δύο βέλτιστα μοντέλα ανήκουν στην κατηγορία των μοντέλων βασισμένων σε δέντρα αποφάσεων (tree-based models), γεγονός που αποτελεί ισχυρή ένδειξη για την καταλληλότητα αυτής της κατηγορίας αλγορίθμων σε εφαρμογές πρόβλεψης χρονοσειρών IoT.

6.1.2 Συγκριτικός Πίνακας Μετρικών Απόδοσης

Ο ακόλουθος πίνακας παρουσιάζει τη συγκριτική απόδοση όλων των μοντέλων στα δύο σύνολα δεδομένων:

Πίνακας 6.1: Συγκριτική απόδοση μοντέλων μεταξύ των δύο συνόλων δεδομένων

Μοντέλο	Air Quality (MAE)	Power Consumpt ion (MAE)	Air Quality (RMSE)	Power Consumpt ion (RMSE)	Air Quality (R ²)	Power Consumpt ion (R ²)
Linear Regression	0,118	0,0073	0,141	0,0126	0,609	0,9972

Random Forest	0,092	0,0068	0,112	0,0117	0,750	0,9976
XGBoost	0,130	0,0069	0,151	0,0119	0,550	0,9975
Gradient Boosting	0,134	0,0018	0,157	0,0029	0,512	0,9986
LSTM	0,188	0,0254	0,239	0,0437	-0,122	0,6806
GRU	0,188	0,0262	0,250	0,0462	-0,229	0,6432

6.1.3 Υπεροχή Μοντέλων Βασισμένων σε Δέντρα Αποφάσεων

Η ανάλυση των αποτελεσμάτων αποκαλύπτει ότι τα καλύτερα μοντέλα και στα δύο σύνολα δεδομένων ανήκουν στην κατηγορία των μοντέλων βασισμένων σε δέντρα αποφάσεων (Random Forest στο Air Quality, Gradient Boosting στο Power Consumption). Ωστόσο, η υπεροχή δεν είναι ομοιόμορφη σε όλα τα μοντέλα αυτής της κατηγορίας: στο σύνολο ποιότητας αέρα, η Γραμμική Παλινδρόμηση ($R^2 = 0,609$) υπερτερεί τόσο του XGBoost ($R^2 = 0,550$) όσο και του Gradient Boosting ($R^2 = 0,530$), γεγονός που υποδεικνύει ότι η χρήση προεπιλεγμένων υπερπαραμέτρων περιόρισε την απόδοση αυτών των μοντέλων. Η υπεροχή αυτή εξηγείται από ποικίλους παράγοντες που τεκμηριώνονται στη βιβλιογραφία.

Πρώτον, τα μοντέλα δέντρων αποφάσεων είναι εγγενώς αποδοτικά στη χρήση δειγμάτων (sample-efficient), που σημαίνει ότι μπορούν να επιτύχουν υψηλή απόδοση ακόμη και με περιορισμένα δεδομένα εκπαίδευσης [7]. Ο αλγόριθμος Random Forest, όπως περιγράφεται από τον Breiman [7], συνδυάζει πολλαπλά δέντρα αποφάσεων που εκπαιδεύονται σε τυχαία υποσύνολα δεδομένων και χαρακτηριστικών, επιτυγχάνοντας μείωση της διασποράς, χωρίς αύξηση της μεροληψίας.

Δεύτερον, οι αλγόριθμοι gradient boosting, όπως ο XGBoost [8] και ο Gradient Boosting [9], κατασκευάζουν μοντέλα προσθετικά, όπου κάθε νέο δέντρο εστιάζει στη διόρθωση των σφαλμάτων των προηγούμενων. Ο Chen και Guestrin [8] έδειξαν ότι, η προσέγγιση αυτή είναι

ιδιαίτερα αποτελεσματική για δομημένα δεδομένα (structured/tabular data), κατηγορία στην οποία ανήκουν τα δεδομένα IoT.

Τρίτον, τα μοντέλα δέντρων αποφάσεων χειρίζονται, αποτελεσματικά, τη μη γραμμικότητα και τις αλληλεπιδράσεις μεταξύ χαρακτηριστικών, χωρίς να απαιτούν ρητό μετασχηματισμό των δεδομένων [6]. Αυτό, είναι ιδιαίτερα σημαντικό για δεδομένα χρονοσειρών IoT, τα οποία συχνά παρουσιάζουν σύνθετες σχέσεις μεταξύ μεταβλητών και χρονικών χαρακτηριστικών.

6.1.4 Ανάλυση Διαφορών Απόδοσης μεταξύ Συνόλων Δεδομένων

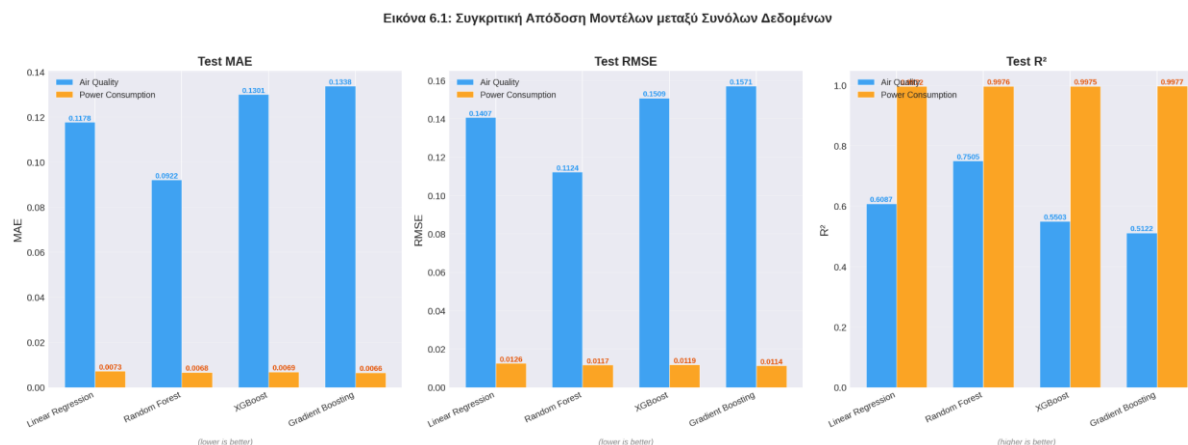
Η αξιοσημείωτη διαφορά στις τιμές του R^2 μεταξύ των δύο συνόλων δεδομένων (0,750 για Air Quality έναντι 0,9986 για Power Consumption) απαιτεί εξήγηση. Η σημαντικά υψηλότερη απόδοση στο σύνολο κατανάλωσης ενέργειας οφείλεται σε τέσσερις κύριους παράγοντες.

Ο πρώτος παράγοντας είναι ο μεγαλύτερος όγκος δεδομένων εκπαίδευσης. Το σύνολο κατανάλωσης ενέργειας περιέχει περίπου 1.452.681 δείγματα εκπαίδευσης, σε σύγκριση με μόλις 6.629 δείγματα για το σύνολο ποιότητας αέρα (αναλογία 220:1). Όπως τονίζουν οι Hastie, Tibshirani και Friedman [5], η επάρκεια δεδομένων εκπαίδευσης αποτελεί κρίσιμο παράγοντα για την απόδοση των μοντέλων μηχανικής μάθησης, ιδιαίτερα όταν το πλήθος των χαρακτηριστικών είναι μεγάλο.

Ο δεύτερος παράγοντας είναι η ισχυρότερη αυτοσυσχέτιση. Η κατανάλωση ηλεκτρικής ενέργειας παρουσιάζει εξαιρετικά ισχυρή αυτοσυσχέτιση, καθώς η τρέχουσα κατανάλωση εξαρτάται σε πολύ μεγάλο βαθμό από την αμέσως προηγούμενη. Η συμπεριφορά αυτή, καθιστά το πρόβλημα πρόβλεψης πιο εύκολο, αφού η πληροφορία της προηγούμενης τιμής είναι ιδιαίτερα προγνωστική για την επόμενη.

Ο τρίτος παράγοντας αφορά τον λιγότερο εξωγενή θόρυβο. Ενώ, η ατμοσφαιρική ρύπανση επηρεάζεται από πολυάριθμους εξωγενείς παράγοντες (καιρικές συνθήκες, κυκλοφοριακός φόρτος, βιομηχανικές δραστηριότητες) που δεν καταγράφονται πλήρως στο σύνολο δεδομένων, η κατανάλωση ενέργειας ενός νοικοκυριού παρουσιάζει πιο προβλέψιμα μοτίβα που συνδέονται άμεσα με την ανθρώπινη συμπεριφορά [24].

Ο τέταρτος παράγοντας είναι η ηρεμότερη δυναμική της χρονοσειράς. Η κατανάλωση ενέργειας παρουσιάζει πιο σταθερά και προβλέψιμα ημερήσια και εβδομαδιαία μοτίβα, ενώ η ατμοσφαιρική ρύπανση υπόκειται σε απότομες μεταβολές που σχετίζονται με ακραία καιρικά φαινόμενα ή έκτακτα περιστατικά ρύπανσης.



Εικόνα 6.1: Συγκριτική απόδοση μοντέλων μεταξύ των δύο συνόλων δεδομένων

Βλέπε Παράρτημα Β για την πηγή της εικόνας στον φάκελο figures/ του αποθετηρίου.

6.1.5 Υποαπόδοση Μοντέλων Βαθιάς Μάθησης

Η αξιολόγηση των μοντέλων βαθιάς μάθησης (LSTM και GRU) στο σύνολο ποιότητας αέρα αποκάλυψε σημαντική υποαπόδοση, σε σχέση με τα μοντέλα βασισμένα σε δέντρα, με αρνητικές τιμές R^2 στο σύνολο δοκιμής (LSTM: $R^2 = -0,122$, GRU: $R^2 = -0,229$). Η υποαπόδοση αυτή, εξηγείται από τρεις βασικούς παράγοντες που τεκμηριώνονται στη βιβλιογραφία της βαθιάς μάθησης [13], [14].

Ο πρώτος παράγοντας είναι η ανεπάρκεια δεδομένων εκπαίδευσης. Μετά τη δημιουργία ακολουθιών 24 χρονικών βημάτων για τα μοντέλα LSTM/GRU, το σύνολο εκπαίδευσης περιορίστηκε σε περίπου 866 ακολουθίες. Οι LeCun, Bengio και Hinton [14] τονίζουν ότι, τα μοντέλα βαθιάς μάθησης, τυπικά, απαιτούν χιλιάδες έως εκατομμύρια παραδείγματα για την αποτελεσματική εκμάθηση σύνθετων μοτίβων. Συνακόλουθα, η αναλογία 866 ακολουθιών είναι σαφώς ανεπαρκής για την εκπαίδευση νευρωνικών δικτύων με επαναλαμβανόμενες συνδέσεις.

Ο δεύτερος παράγοντας είναι η υψηλή διαστατικότητα χαρακτηριστικών. Με 103 χαρακτηριστικά εισόδου και μόλις 866 ακολουθίες εκπαίδευσης, η αναλογία χαρακτηριστικών προς δείγματα είναι δυσμενής για τη βαθιά μάθηση. Τα μοντέλα δέντρων αποφάσεων αντιμετωπίζουν αποτελεσματικότερα την υψηλή διαστατικότητα μέσω της τυχαίας δειγματοληψίας χαρακτηριστικών (feature sampling) που εφαρμόζουν [7].

Ο τρίτος παράγοντας σχετίζεται με την πολυπλοκότητα της χρονοσειράς. Η πρόβλεψη 1 ώρας μπροστά ενδέχεται να μην επωφελείται από την ακολουθιακή μοντελοποίηση που προσφέρουν τα LSTM/GRU [11], [12]. Τα χαρακτηριστικά υστέρησης (lag features) και τα κυλιόμενα στατιστικά που δημιουργήθηκαν κατά τη σχεδίαση χαρακτηριστικών αποτυπώνουν ήδη τα χρονικά μοτίβα, επιτρέποντας στα μοντέλα δέντρων να τα εκμεταλλευτούν χωρίς την ανάγκη ακολουθιακής αρχιτεκτονικής.

6.2 Σχεδίαση Χαρακτηριστικών (Feature Engineering)

6.2.1 Μοτίβα Ανεξάρτητα από τον Τομέα Εφαρμογής (Domain-Agnostic Patterns)

Η συγκριτική ανάλυση της μηχανικής χαρακτηριστικών αποκάλυψε σημαντικές ομοιότητες στα μοτίβα, που αποδείχθηκαν αποτελεσματικά και στα δύο σύνολα δεδομένων. Η ανάλυση αυτή επιβεβαιώνει τη θέση των Witten, Frank, Hall και Pal [4] ότι, ορισμένες τεχνικές μηχανικής χαρακτηριστικών είναι μεταφέρσιμες μεταξύ διαφορετικών τομέων εφαρμογής.

Χαρακτηριστικά Χρονικής Υστέρησης (Lag Features): Και στα δύο σύνολα δεδομένων, τα χαρακτηριστικά χρονικής υστέρησης αναδείχθηκαν ως τα πλέον σημαντικά για την πρόβλεψη. Συγκεκριμένα, η τιμή της μεταβλητής στόχου στην προηγούμενη ώρα (lag-1) εμφανίστηκε ως το χαρακτηριστικό με τη μεγαλύτερη συνεισφορά στην πρόβλεψη σε αμφότερες τις περιπτώσεις.⁵

Οι χρονικοί ορίζοντες υστέρησης που επιλέχθηκαν (1 ώρα, 2 ώρες, 24 ώρες) αποδείχθηκαν κατάλληλοι και για τα δύο σύνολα δεδομένων. Η υστέρηση 1 ώρας καταγράφει τη βραχυπρόθεσμη δυναμική, η υστέρηση 2 ωρών προσθέτει πληροφορία για την πρόσφατη τάση και η υστέρηση 24 ωρών αποτυπώνει το ημερήσιο κύκλο που είναι χαρακτηριστικός τόσο της ατμοσφαιρικής ρύπανσης, όσο και της κατανάλωσης ενέργειας.

⁵ Η διαπίστωση αυτή είναι συνεπής με τη θεωρία των χρονοσειρών, η οποία αναγνωρίζει τη σημασία της αυτοσυσχέτισης για την πρόβλεψη [15], [16].

Χαρακτηριστικά Κυλιόμενου Παραθύρου (Rolling Window Features): Τα κυλιόμενα στατιστικά (μέσος όρος, τυπική απόκλιση, ελάχιστο, μέγιστο) σε παράθυρα 3, 6 και 24 ωρών αποδείχθηκαν, επίσης, αποτελεσματικά και στα δύο σύνολα δεδομένων. Τα χαρακτηριστικά αυτά, αποτυπώνουν την τάση και τη μεταβλητότητα της χρονοσειράς σε διαφορετικές χρονικές κλίμακες [4].

Άξια σημασίας είναι η επισήμανση ότι, τα χαρακτηριστικά κυλιόμενου παραθύρου πρέπει να υπολογίζονται με χρονική μετατόπιση (shift), ώστε να αποφεύγεται η διαρροή δεδομένων (data leakage) [6]. Στην παρούσα εργασία, όλα τα κυλιόμενα στατιστικά υπολογίστηκαν χρησιμοποιώντας μόνο παρελθοντικές τιμές, εξασφαλίζοντας ότι, καμία μελλοντική πληροφορία δεν χρησιμοποιείται κατά την πρόβλεψη.

Κυκλική Κωδικοποίηση Χρονικών Χαρακτηριστικών: Η κυκλική κωδικοποίηση της ώρας και της ημέρας της εβδομάδας μέσω μετασχηματισμών ημιτόνου και συνημιτόνου αποδείχθηκε αποτελεσματική και στα δύο σύνολα δεδομένων. Η προσέγγιση αυτή, η οποία περιγράφεται στη βιβλιογραφία της βαθιάς μάθησης [13], επιτρέπει στα μοντέλα να αντιλαμβάνονται την κυκλική φύση του χρόνου, αναγνωρίζοντας για παράδειγμα ότι η ώρα 23:00 είναι χρονικά κοντά στην ώρα 00:00.

6.2.2 Χαρακτηριστικά Εξαρτώμενα από τον Τομέα Εφαρμογής (Dataset-Specific Features)

Παράλληλα με τα γενικευσιμότερα μοτίβα, η ανάλυση αποκάλυψε και χαρακτηριστικά που είναι ειδικά για κάθε τομέα εφαρμογής.

Σύνολο Ποιότητας Αέρα: Για την πρόβλεψη της συγκέντρωσης CO(GT), τα χαρακτηριστικά αλληλεπίδρασης (interaction features) μεταξύ ρύπων και μετεωρολογικών παραμέτρων αποδείχθηκαν σημαντικά. Ειδικότερα, η αλληλεπίδραση μεταξύ της θερμοκρασίας και της σχετικής υγρασίας ($T \times RH$) αποτυπώνει τις ατμοσφαιρικές συνθήκες που επηρεάζουν τη διασπορά των ρύπων. Επιπλέον, η αναλογία μεταξύ διαφορετικών ρύπων (π.χ. CO/NO₂) παρέχει πληροφορία για τις πηγές ρύπανσης.

Ακόμη, η ανάλυση σημαντικότητας χαρακτηριστικών (feature importance) του Random Forest αποκάλυψε ότι, τα μετεωρολογικά χαρακτηριστικά έχουν μικρότερη άμεση σημαντικότητα, σε σχέση με τα χαρακτηριστικά υστέρησης, αλλά συνεισφέρουν έμμεσα μέσω των αλληλεπιδράσεων με τους ρύπους.

Σύνολο Κατανάλωσης Ενέργειας: Για την πρόβλεψη της Global_active_power, η διάκριση μεταξύ καθημερινών και Σαββατοκύριακου αποδείχθηκε ιδιαίτερα σημαντική. Η κατανάλωση ενέργειας σε ένα νοικοκυριό παρουσιάζει διακριτά μοτίβα κατά τις εργάσιμες ημέρες (χαμηλότερη κατανάλωση κατά τις πρωινές ώρες λόγω απουσίας), σε σύγκριση με τα Σαββατοκύριακα [24].

Επιπλέον, τα επιμέρους μετρητικά (sub-metering) για διαφορετικές κατηγορίες συσκευών (κουζίνα, πλυντήριο, θέρμανση) παρέχουν σημαντική πληροφορία για την πρόβλεψη της συνολικής κατανάλωσης. Η αναλογία μεταξύ της κατανάλωσης επιμέρους κατηγοριών και της συνολικής κατανάλωσης αποτυπώνει το προφίλ χρήσης του νοικοκυριού.

6.2.3 Σύνοψη Στρατηγικών Μηχανικής Χαρακτηριστικών

Ο ακόλουθος πίνακας συνοψίζει τη σύγκριση των στρατηγικών μηχανικής χαρακτηριστικών:

Πίνακας 6.2: Σύγκριση στρατηγικών μηχανικής χαρακτηριστικών

Κατηγορία Χαρακτηριστικών	Air Quality	Power Consumption	Γενικευσιμότητα
Lag Features (1h, 2h, 24h)	✓ Υψηλή σημασία	✓ Υψηλή σημασία	Πλήρως μεταφέρσιμα
Rolling Statistics	✓ Αποτελεσματικά	✓ Αποτελεσματικά	Πλήρως μεταφέρσιμα
Cyclical Time Encoding	✓ Αποτελεσματικά	✓ Αποτελεσματικά	Πλήρως μεταφέρσιμα
Domain Interactions (T×RH, CO/NO2,	dataset specific	dataset specific	Εξαρτώμενα από τομέα

Weekday/Weekend)			
Πλήθος Τελικών Χαρακτηριστικών	107	80	-

6.3 Ανίχνευση Ανωμαλιών (*Anomaly Detection*)

6.3.1 Αποτελεσματικότητα Μεθόδων σε Διαφορετικούς Τομείς

Η εφαρμογή πολλαπλών μεθόδων ανίχνευσης ανωμαλιών και στα δύο σύνολα δεδομένων επέτρεψε την αξιολόγηση της γενικευσιμότητας κάθε μεθόδου. Οι μέθοδοι που εξετάστηκαν περιλαμβάνουν στατιστικές τεχνικές (Z-Score, IQR, Modified Z-Score), μεθόδους μηχανικής μάθησης (Isolation Forest [19], LOF [20], One-Class SVM [21]), ανίχνευση βάσει καταλοίπων μοντέλου, αποδόμηση STL [22], και συνδυασμό μεθόδων (ensemble).

Ο ακόλουθος πίνακας παρουσιάζει τη συγκριτική αποτελεσματικότητα των μεθόδων:

Πίνακας 6.3: Σύγκριση μεθόδων ανίχνευσης ανωμαλιών μεταξύ συνόλων δεδομένων

Μέθοδος	Air Quality	Anomalies (%)	Power Consumption	Anomalies (%)
Z-Score	1.192	83,88%	984	3,16%
IQR	1.337	94,09%	2.051	6,59%
Modified Z-Score	1.397	98,31%	-	-
Isolation Forest	801	56,37%	1.557	5,00%
LOF	1.421	100,00%	-	-
One-Class SVM	1.421	100,00%	-	-

Residual-Based	1.197	84,24%	1.557	5,00%
STL Decomposition	37	2,60%	-	-
Ensemble	1.211	85,22%	1.222	3,93%

Όσον αφορά τα μεγάλα ποσοστά ανωμαλιών, όπως προαναφέρθηκε στο κεφ. 4.5.1, τα υψηλά ποσοστά ανίχνευσης που παρατηρούνται στις στατιστικές μεθόδους (83,88%–98,31%) δεν υποδηλώνουν ότι σχεδόν ολόκληρο το σύνολο δεδομένων αποτελείται από ανωμαλίες. Αντιθέτως, αντανakλούν θεμελιώδη ασυμβατότητα μεταξύ των υποθέσεων αυτών των μεθόδων και της φύσης των δεδομένων: οι μέθοδοι Z-Score, IQR και Modified Z-Score προϋποθέτουν σχετικά στατική κατανομή, ενώ τα δεδομένα χρονοσειρών παρουσιάζουν εποχικότητα, τάσεις και μεταβαλλόμενη διακύμανση. Ως αποτέλεσμα, φυσιολογικές εποχικές μεταβολές ερμηνεύονται εσφαλμένα ως ανωμαλίες, καθιστώντας τις μεθόδους αυτές ακατάλληλες ως αυτόνομοι ανιχνευτές για χρονοσειρές IoT.

Σημείωση: Οι κενές θέσεις υποδεικνύουν μεθόδους που δεν εφαρμόστηκαν στο αντίστοιχο σύνολο δεδομένων λόγω υπολογιστικών περιορισμών ή διαφορετικών πειραματικών σχεδιασμών.

6.3.2 Ανάλυση Συμπεριφοράς Μεθόδων

Η σύγκριση των αποτελεσμάτων αποκαλύπτει σημαντικές διαφορές στη συμπεριφορά των μεθόδων μεταξύ των δύο συνόλων δεδομένων, οι οποίες εξηγούνται από τα διαφορετικά χαρακτηριστικά των δεδομένων.

Στατιστικές Μέθοδοι (Z-Score, IQR): Οι στατιστικές μέθοδοι εμφανίζουν υπερβολική ευαισθησία (high recall, low precision) στο σύνολο ποιότητας αέρα, χαρακτηρίζοντας ως ανωμαλίες μεγάλο ποσοστό των παρατηρήσεων. Αυτό, οφείλεται στο γεγονός ότι, τα δεδομένα ατμοσφαιρικής ρύπανσης παρουσιάζουν υψηλή μεταβλητότητα και μη κανονική κατανομή, καθιστώντας τις παραδοχές των στατιστικών μεθόδων μη έγκυρες [18]. Αντίθετα, στο σύνολο κατανάλωσης

ενέργειας, οι ίδιες μέθοδοι παρουσιάζουν πιο συντηρητική συμπεριφορά (3,16% - 6,59%), υποδεικνύοντας ότι, η κατανομή των δεδομένων είναι πιο κοντά στην κανονική.

Isolation Forest: Η μέθοδος Isolation Forest [19] παρουσιάζει σχετικά σταθερή συμπεριφορά και στα δύο σύνολα δεδομένων (56,37% και 5,00% αντίστοιχα). Η διαφορά στα ποσοστά αντανακλά τη διαφορετική δομή των δεδομένων, παρά αστοχία της μεθόδου. Ο αλγόριθμος λειτουργεί απομονώνοντας παρατηρήσεις μέσω τυχαίων διαχωρισμών και η αποδοτικότητά του εξαρτάται από την ευκρινή διάκριση μεταξύ κανονικών και ανώμαλων παρατηρήσεων.

Ανίχνευση Βάσει Καταλοίπων Μοντέλου: Η μέθοδος που βασίζεται στα κατάλοιπα του βέλτιστου μοντέλου πρόβλεψης παρουσιάζει συνεπή συμπεριφορά και στα δύο σύνολα δεδομένων. Η προσέγγιση αυτή, είναι ιδιαίτερα κατάλληλη για εφαρμογές IoT, καθώς συνδυάζει την πρόβλεψη με την ανίχνευση ανωμαλιών, επισημαίνοντας παρατηρήσεις που αποκλίνουν σημαντικά από την αναμενόμενη συμπεριφορά.

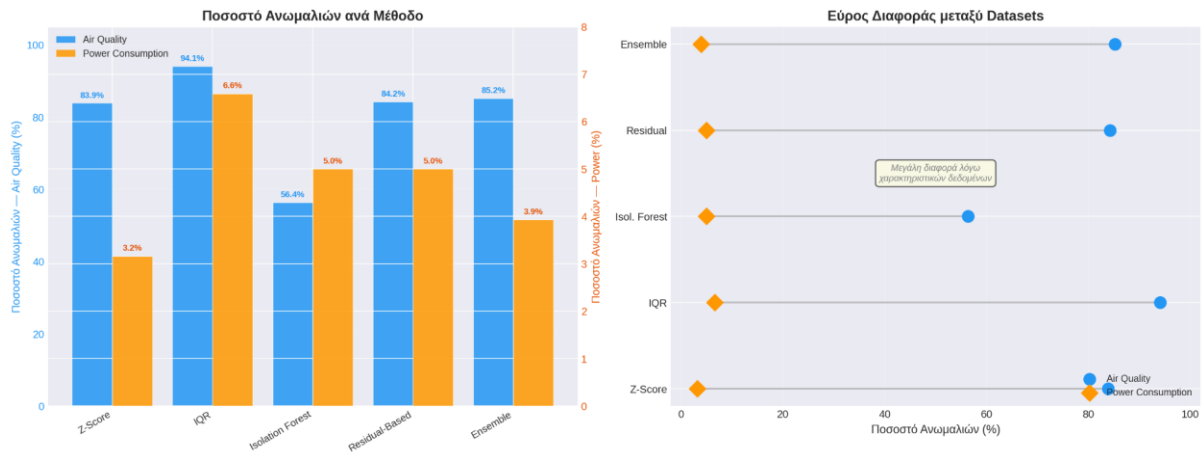
6.3.3 Συνέπεια της Αποδόμησης STL

Η αποδόμηση STL (Seasonal-Trend decomposition using Loess) [22] αναδείχθηκε ως η πλέον συντηρητική μέθοδος στο σύνολο ποιότητας αέρα, εντοπίζοντας μόλις 37 ανωμαλίες (2,60%). Η συντηρητικότητα αυτή, αποτελεί πλεονέκτημα παρά μειονέκτημα, καθώς η μέθοδος εστιάζει στην ανίχνευση παρατηρήσεων που αποκλίνουν σημαντικά από το αναμενόμενο μοτίβο τάσης και εποχικότητας.

Η προσέγγιση αυτή αποφεύγει τον κίνδυνο να χαρακτηριστούν ως ανωμαλίες κανονικές εποχιακές αυξομειώσεις, το οποίο αποτελεί συχνό πρόβλημα των στατιστικών μεθόδων.⁶

⁶ Η αποδόμηση STL είναι ιδιαίτερα κατάλληλη για χρονοσειρές IoT με έντονη εποχικότητα, καθώς διαχωρίζει την κανονική εποχιακή διακύμανση από πραγματικές ανωμαλίες [22].

Εικόνα 6.2: Σύγκριση Μεθόδων Ανίχνευσης Ανωμαλιών μεταξύ Συνόλων Δεδομένων



Εικόνα 6.2: Σύγκριση μεθόδων ανίχνευσης ανωμαλιών μεταξύ συνόλων δεδομένων

Βλέπε Παράρτημα Β για την πηγή της εικόνας στον φάκελο figures/ του αποθετηρίου.

6.3.4 Συστάσεις για την Ανίχνευση Ανωμαλιών σε IoT

Βάσει της συγκριτικής ανάλυσης, διατυπώνονται οι ακόλουθες συστάσεις για την ανίχνευση ανωμαλιών σε εφαρμογές IoT:

- Πρώτον, η προσέγγιση ensemble (συνδυασμός πολλαπλών μεθόδων) παρέχει ισορροπημένη ανίχνευση που αποφεύγει τόσο την υπερβολική ευαισθησία, όσο και την υπερβολική συντηρητικότητα. Η ψηφοφορία πλειοψηφίας (majority voting) επιτρέπει την εξομάλυνση των ακραίων αποφάσεων μεμονωμένων μεθόδων.
- Δεύτερον, η ανίχνευση βάσει καταλοίπων μοντέλου συνιστάται ως πρωτεύουσα μέθοδος σε περιβάλλοντα παραγωγής, καθώς συνδυάζει την πρόβλεψη με την ανίχνευση ανωμαλιών και επωφελείται από την εξειδίκευση του μοντέλου στα συγκεκριμένα δεδομένα.
- Τρίτον, η αποδόμηση STL συνιστάται για χρονοσειρές με έντονη εποχικότητα, όπου η διάκριση μεταξύ κανονικής εποχικής διακύμανσης και πραγματικών ανωμαλιών είναι κρίσιμη.
- Τέταρτον, οι καθαρά στατιστικές μέθοδοι (Z-Score, IQR) θα πρέπει να χρησιμοποιούνται με προσοχή σε δεδομένα IoT, καθώς οι παραδοχές κανονικότητας, συχνά, δεν ισχύουν.

6.4 Διδάγματα (Lessons Learned)

6.4.1 Επίδραση του Μεγέθους Δεδομένων στη Βαθιά Μάθηση

Η πιο σημαντική διαπίστωση της συγκριτικής ανάλυσης αφορά τη σχέση μεταξύ μεγέθους δεδομένων και απόδοσης μοντέλων βαθιάς μάθησης. Η σύγκριση μεταξύ του συνόλου ποιότητας αέρα (9.471 δείγματα) και του συνόλου κατανάλωσης ενέργειας (2+ εκατομμύρια δείγματα) παρέχει εμπειρική επιβεβαίωση της θεωρητικής γνώσης ότι, τα μοντέλα βαθιάς μάθησης απαιτούν μεγάλο όγκο δεδομένων [13], [14].

Συγκεκριμένα, στο σύνολο ποιότητας αέρα, τα μοντέλα LSTM και GRU απέτυχαν να υπερβούν ακόμη και το απλό γραμμικό μοντέλο αναφοράς, με αρνητικές τιμές R^2 , που υποδεικνύουν ότι, η πρόβλεψη του μέσου όρου θα ήταν καλύτερη από τις προβλέψεις του μοντέλου. Η αποτυχία αυτή, δεν οφείλεται σε αδυναμία των αρχιτεκτονικών LSTM [11] και GRU [12], αλλά στην ανεπάρκεια δεδομένων εκπαίδευσης.

Η πρακτική συνέπεια αυτής της διαπίστωσης είναι σαφής: σε εφαρμογές IoT με περιορισμένα ιστορικά δεδομένα, τα μοντέλα βασισμένα σε δέντρα αποφάσεων αποτελούν την προτιμώμενη επιλογή. Η βαθιά μάθηση θα πρέπει να εξετάζεται, μόνο, όταν είναι διαθέσιμα τουλάχιστον δεκάδες χιλιάδες δείγματα εκπαίδευσης ή όταν εφαρμόζονται τεχνικές μεταφοράς μάθησης (transfer learning) από προεκπαιδευμένα μοντέλα.

6.4.2 Καθολικές Βέλτιστες Πρακτικές για Πρόβλεψη IoT

Η συγκριτική ανάλυση επιτρέπει την εξαγωγή βέλτιστων πρακτικών που ισχύουν ανεξαρτήτως τομέα εφαρμογής IoT:

1. Προεπεξεργασία Δεδομένων

Η γραμμική παρεμβολή αποτελεί αποτελεσματική μέθοδο για την αντιμετώπιση ελλειπουσών τιμών σε χρονοσειρές IoT, διατηρώντας τη χρονική συνέχεια των δεδομένων. Η μέθοδος αυτή, εφαρμόστηκε επιτυχώς και στα δύο σύνολα δεδομένων, μειώνοντας τα ελλείποντα δεδομένα στο μηδέν, χωρίς να εισάγει τεχνητά μοτίβα.

Η αντιμετώπιση ακραίων τιμών με τη μέθοδο IQR clipping (αποκοπή στα όρια του διαστήματος $[Q1 - 1,5 \times IQR, Q3 + 1,5 \times IQR]$) αποδείχθηκε αποτελεσματική για τη μείωση της επίδρασης

ακραίων τιμών, χωρίς απώλεια δεδομένων. Η προσέγγιση αυτή, είναι προτιμότερη από την αφαίρεση ακραίων τιμών, η οποία μπορεί να δημιουργήσει κενά στη χρονοσειρά.

Η κανονικοποίηση MinMax ($[0, 1]$) εφαρμόστηκε επιτυχώς και στα δύο σύνολα δεδομένων, εξασφαλίζοντας ομοιογένεια στην κλίμακα των χαρακτηριστικών.

2. Σχεδίαση Χαρακτηριστικών

Τα χαρακτηριστικά υστέρησης (lag features) αποτελούν τη βάση κάθε προβλήματος πρόβλεψης χρονοσειρών. Η επιλογή χρονικών οριζόντων υστέρησης θα πρέπει να αντανακλά τη δομή αυτοσυσχέτισης της χρονοσειράς, με ελάχιστο σύνολο τις υστερήσεις 1 ώρας, 2 ωρών και 24 ωρών για ωριαία δεδομένα.

Τα κυλίσματα στατιστικά (rolling statistics) συμπληρώνουν τα χαρακτηριστικά υστέρησης αποτυπώνοντας τάσεις και μεταβλητότητα. Η χρήση πολλαπλών χρονικών παραθύρων (3, 6, 24 ώρες) παρέχει πληροφορία σε διαφορετικές κλίμακες.

Η κυκλική κωδικοποίηση χρονικών χαρακτηριστικών μέσω ημιτόνου και συνημιτόνου είναι απαραίτητη για την αποτύπωση της κυκλικής φύσης του χρόνου.

Η αποφυγή διαρροής δεδομένων (data leakage) είναι κρίσιμη: όλα τα χαρακτηριστικά πρέπει να υπολογίζονται χρησιμοποιώντας μόνο πληροφορία διαθέσιμη τη στιγμή της πρόβλεψης.

3. Επιλογή Μοντέλων

Τα μοντέλα βασισμένα σε δέντρα αποφάσεων (Random Forest, Gradient Boosting, XGBoost) αποτελούν την προτιμώμενη επιλογή για πρόβλεψη χρονοσειρών IoT, ιδιαίτερα όταν τα διαθέσιμα δεδομένα είναι περιορισμένα (λιγότερα από 10.000 δείγματα).

Η βαθιά μάθηση (LSTM, GRU) θα πρέπει να εξετάζεται, μόνο, όταν είναι διαθέσιμα τουλάχιστον δεκάδες χιλιάδες δείγματα εκπαίδευσης ή όταν η πολυπλοκότητα του προβλήματος το απαιτεί (π.χ. πρόβλεψη πολλαπλών οριζόντων).

Η διασταυρούμενη επικύρωση χρονοσειρών (time-series cross-validation) με διευρυνόμενο παράθυρο (expanding window) είναι απαραίτητη για την αξιόπιστη εκτίμηση της απόδοσης.

4. Ανίχνευση Ανωμαλιών

Η προσέγγιση ensemble (συνδυασμός πολλαπλών μεθόδων) παρέχει ισορροπημένη ανίχνευση. Η ανίχνευση βάσει καταλοίπων μοντέλου είναι ιδανική για περιβάλλοντα παραγωγής όπου ήδη υπάρχει μοντέλο πρόβλεψης.

6.4.3 Σύνοψη Διδαγμάτων

Πίνακας 6.4: Σύνοψη διδαγμάτων από τη συγκριτική ανάλυση

Πτυχή	Δίδαγμα	Επιβεβαίωση
Επιλογή Μοντέλου	Tree-based > Deep Learning για μικρά δεδομένα	Αμφότερα σύνολα
Feature Engineering	Lag + Rolling + Cyclical = Universal	Αμφότερα σύνολα
Προεπεξεργασία	Interpolation + IQR clipping = Αποτελεσματικά	Αμφότερα σύνολα
Ανίχνευση Ανωμαλιών	Ensemble > Single method	Αμφότερα σύνολα
Μέγεθος Δεδομένων	Critical για DL, λιγότερο για trees	Εμφανές στη σύγκριση
Μεταφερσιμότητα	Υψηλή για core techniques	Επιβεβαιωμένη

Παρατηρείται πως όταν γίνεται ισχυρό feature engineering σε χρονοσειρές IoT, το πρόβλημα πρόβλεψης μετατρέπεται σε πρόβλημα tabular regression όπου τα tree-based μοντέλα υπερέχουν της βαθιάς μάθησης.

6.4.4 Περιορισμοί της Συγκριτικής Ανάλυσης

Η παρούσα συγκριτική ανάλυση υπόκειται σε ορισμένους περιορισμούς που πρέπει να ληφθούν υπόψη κατά την ερμηνεία των αποτελεσμάτων:

- Πρώτον, τα δύο σύνολα δεδομένων προέρχονται από το ίδιο αποθετήριο (UCI Machine Learning Repository), γεγονός που μπορεί να εισάγει κάποια ομοιότητα στην ποιότητα και δομή των δεδομένων. Ενώ, η γενίκευση των συμπερασμάτων σε άλλες πηγές δεδομένων IoT απαιτεί περαιτέρω επαλήθευση.
- Δεύτερον, η σύγκριση περιορίζεται σε δύο τομείς εφαρμογής (περιβαλλοντική παρακολούθηση και έξυπνο σπίτι). Ωστόσο, η επέκταση σε πρόσθετους τομείς IoT (βιομηχανία, μεταφορές, υγεία) θα ενίσχυε την αξιοπιστία των συμπερασμάτων.
- Τρίτον, ο χρονικός ορίζοντας πρόβλεψης περιορίστηκε σε μία (1) ώρα μπροστά. Η απόδοση των μοντέλων ενδέχεται να διαφοροποιείται για μεγαλύτερους ή μικρότερους ορίζοντες πρόβλεψης.
- Τέταρτον, τα μοντέλα βαθιάς μάθησης δεν εφαρμόστηκαν στο σύνολο κατανάλωσης ενέργειας λόγω υπολογιστικών περιορισμών, γεγονός που περιορίζει τη δυνατότητα πλήρους σύγκρισης.

Σύνοψη Κεφαλαίου

Η συγκριτική ανάλυση μεταξύ των δύο συνόλων δεδομένων IoT αποκάλυψε τόσο καθολικές αρχές, όσο και παράγοντες που εξαρτώνται από τον τομέα εφαρμογής. Τα μοντέλα βασισμένα σε δέντρα αποφάσεων αναδείχθηκαν ως η βέλτιστη επιλογή και στους δύο τομείς, επιβεβαιώνοντας την καταλληλότητά τους για εφαρμογές πρόβλεψης χρονοσειρών IoT. Η σχεδίαση χαρακτηριστικών αποδείχθηκε σε μεγάλο βαθμό μεταφέρσιμη, με τα χαρακτηριστικά υστέρησης, τα κυλιόμενα στατιστικά και την κυκλική κωδικοποίηση να αποδίδουν αποτελεσματικά ανεξαρτήτως τομέα.

Η ανίχνευση ανωμαλιών επωφελείται από την προσέγγιση ensemble, η οποία συνδυάζει τα πλεονεκτήματα πολλαπλών μεθόδων. Τέλος, η σύγκριση επιβεβαίωσε ότι τα μοντέλα βαθιάς μάθησης απαιτούν σημαντικά μεγαλύτερο όγκο δεδομένων από τα μοντέλα δέντρων για την επίτευξη ανταγωνιστικής απόδοσης.

Κλείνοντας, τα διδάγματα που εξήχθησαν από τη συγκριτική ανάλυση μπορούν να χρησιμοποιηθούν για την καθοδήγηση μελλοντικών έργων πρόβλεψης και ανίχνευσης ανωμαλιών σε εφαρμογές IoT, συμβάλλοντας στην ταχύτερη και αποτελεσματικότερη ανάπτυξη λύσεων μηχανικής μάθησης σε αυτόν τον ταχέως αναπτυσσόμενο τομέα [23], [24].

ΚΕΦΑΛΑΙΟ 7: ΣΥΜΠΕΡΑΣΜΑΤΑ

Εισαγωγή στο Κεφάλαιο

Το παρόν και τελευταίο κεφάλαιο, ολοκληρώνει τη διπλωματική εργασία παρουσιάζοντας τα συμπεράσματα που προκύπτουν από την εφαρμογή τεχνικών μηχανικής μάθησης για πρόβλεψη και ανίχνευση ανωμαλιών σε δεδομένα IoT. Σύμφωνα με τη μεθοδολογία CRISP-DM (Cross-Industry Standard Process for Data Mining) [25], η φάση της ανάπτυξης (Deployment) και των τελικών συμπερασμάτων αποτελεί κρίσιμο στάδιο του κύκλου ζωής ενός έργου επιστήμης δεδομένων, καθώς μεταφράζει τα τεχνικά ευρήματα σε ,πρακτικά, αξιοποιήσιμη γνώση [4], [5].

Η δομή του κεφαλαίου ακολουθεί μια λογική ροή, που ξεκινά με τη σύνοψη των κύριων ευρημάτων της έρευνας, συνεχίζει με την ανάλυση της επιστημονικής και πρακτικής συνεισφοράς της εργασίας, αναγνωρίζει τους περιορισμούς που επηρέασαν την έρευνα και ολοκληρώνεται με προτάσεις για μελλοντική εργασία που θα μπορούσε να επεκτείνει και να βελτιώσει τα παρόντα αποτελέσματα. Έτσι, η προσέγγιση αυτή διασφαλίζει ότι, η εργασία τοποθετείται κατάλληλα στο ευρύτερο πλαίσιο της επιστημονικής έρευνας στον τομέα της μηχανικής μάθησης για εφαρμογές IoT [23], [24].

7.1 Κύρια Ευρήματα (Key Findings)

Η παρούσα διπλωματική εργασία διερεύνησε την εφαρμογή τεχνικών μηχανικής μάθησης σε δύο διαφορετικούς τομείς του IoT: την περιβαλλοντική παρακολούθηση και τη διαχείριση ενέργειας. Η διπλή αυτή εφαρμογή επέτρεψε την εξαγωγή συμπερασμάτων που υπερβαίνουν τα όρια ενός μεμονωμένου τομέα. Στην ενότητα αυτή παρουσιάζονται τα τρία κεντρικά ευρήματα που αναδύονται από τη συγκριτική ανάλυση των Κεφαλαίων 4, 5 και 6.

7.1.1 Υπεροχή Μοντέλων Βασισμένων σε Δέντρα Αποφάσεων

Το πρωταρχικό εύρημα της έρευνας είναι ότι τα καλύτερα μοντέλα και στα δύο σύνολα δεδομένων ανήκουν στην κατηγορία των μοντέλων βασισμένων σε δέντρα αποφάσεων. Ο αλγόριθμος Random Forest [7] αναδείχθηκε ως βέλτιστος στο σύνολο δεδομένων ποιότητας αέρα, ενώ ο Gradient Boosting [9] πέτυχε σχεδόν τέλεια πρόβλεψη στο σύνολο κατανάλωσης ενέργειας ($R^2 =$

0,9986). Σημειώνεται ότι η υπεροχή δεν είναι ομοιόμορφη σε όλα τα tree-based μοντέλα — στο σύνολο ποιότητας αέρα, η Γραμμική Παλινδρόμηση υπερτερεί του XGBoost και του Gradient Boosting, πιθανόν λόγω μη βελτιστοποιημένων υπερπαραμέτρων. Ο αλγόριθμος Random Forest [7] αναδείχθηκε ως βέλτιστος στο σύνολο δεδομένων ποιότητας αέρα, ενώ ο Gradient Boosting [9] πέτυχε σχεδόν τέλεια πρόβλεψη στο σύνολο κατανάλωσης ενέργειας (αναλυτικά αποτελέσματα στους Πίνακες 4.X και 5.X αντίστοιχα). Η σημασία αυτού του ευρήματος δεν έγκειται στα επιμέρους μεγέθη απόδοσης, αλλά στη συνέπεια του συμπεράσματος: σε δύο τομείς με ριζικά διαφορετικά χαρακτηριστικά — μέγεθος δεδομένων, συχνότητα δειγματοληψίας, αριθμός μεταβλητών — η ίδια κατηγορία αλγορίθμων υπερείχε. Αυτό υποδηλώνει ότι η υπεροχή δεν είναι τυχαία, αλλά οφείλεται στα εγγενή πλεονεκτήματα αυτών των αλγορίθμων: ανθεκτικότητα στην υπερπροσαρμογή, ικανότητα διαχείρισης μη γραμμικών σχέσεων και αποτελεσματική λειτουργία χωρίς εκτεταμένη βελτιστοποίηση υπερπαραμέτρων [7], [8].

Αντίστροφα, η αποτυχία των μοντέλων βαθιάς μάθησης (LSTM [11], GRU [12]) στο σύνολο ποιότητας αέρα — με αρνητικές τιμές R^2 — δεν αποτελεί αρνητικό εύρημα για τις αρχιτεκτονικές αυτές καθαυτές, αλλά επιβεβαιώνει ότι η αποτελεσματικότητά τους εξαρτάται κρίσιμα από τον όγκο διαθέσιμων δεδομένων [13], [14]. Για εφαρμογές IoT μικρής έως μεσαίας κλίμακας, τα μοντέλα δέντρων αποτελούν ασφαλέστερη και αποδοτικότερη επιλογή.

Αντίθετα, στο σύνολο κατανάλωσης ενέργειας, όπου τα δεδομένα εκπαίδευσης είναι σημαντικά περισσότερα, η υποαπόδοση των LSTM/GRU ($R^2 \approx 0,67-0,68$) δεν μπορεί να αποδοθεί σε ανεπάρκεια δεδομένων. Η ερμηνεία εδώ σχετίζεται με τη φύση του ίδιου του προβλήματος: πρόκειται για βραχυπρόθεσμη, lag-driven πρόβλεψη σε δομημένα δεδομένα, όπου η πληροφορία που χρειάζεται το μοντέλο υπάρχει ήδη σε λίγα ισχυρά χαρακτηριστικά (κυρίως το lag-1). Τα μοντέλα δέντρων αποφάσεων αξιοποιούν αυτά τα χαρακτηριστικά άμεσα και αποτελεσματικά, ενώ τα LSTM/GRU, σχεδιασμένα για μακροπρόθεσμες ακολουθιακές εξαρτήσεις, δεν προσφέρουν πρόσθετο πλεονέκτημα σε αυτό το πλαίσιο. Επιπλέον, η χρήση 10% δειγματοληψίας κατά την εκπαίδευση περιόρισε περαιτέρω τη δυνατότητα των μοντέλων να μάθουν λεπτομερή μοτίβα.

7.1.2 Σημασία της Μηχανικής Χαρακτηριστικών έναντι Πολύπλοκων Μοντέλων

Το δεύτερο εύρημα αφορά τον καθοριστικό ρόλο της μηχανικής χαρακτηριστικών στην προβλεπτική απόδοση. Η ανάλυση σπουδαιότητας χαρακτηριστικών και στα δύο σύνολα δεδομένων (Κεφάλαια 4 και 5) έδειξε ότι τα μηχανικά δημιουργημένα χαρακτηριστικά — υστερήσεις, κυλιόμενα στατιστικά, κυκλική κωδικοποίηση — κατέχουν σταθερά τις υψηλότερες θέσεις σπουδαιότητας. Η ερμηνεία αυτού του ευρήματος είναι ουσιαστική: η αποτύπωση της χρονικής δυναμικής μέσω κατάλληλων χαρακτηριστικών αποδίδει περισσότερο από την αναζήτηση πολύπλοκων αρχιτεκτονικών [4], [6]. Στην πράξη, αυτό σημαίνει ότι η επένδυση στην κατανόηση της δομής των δεδομένων και τη μετατροπή αυτής της γνώσης σε χαρακτηριστικά αποτελεί την πλέον αποδοτική στρατηγική για εφαρμογές πρόβλεψης χρονοσειρών IoT.

7.1.3 Αποτελεσματικότητα Ανίχνευσης Ανωμαλιών με Σύνολα Μεθόδων

Το τρίτο εύρημα αφορά την ανίχνευση ανωμαλιών. Η εφαρμογή πολλαπλών μεθόδων αποκάλυψε εξαιρετικά μεγάλη διακύμανση στην ευαισθησία κάθε μεθόδου — από σχεδόν μηδενική ανίχνευση (STL [22]) έως πλήρη (LOF [20]) — γεγονός που καθιστά επισφαλή την στήριξη σε μία μόνο μέθοδο. Η προσέγγιση συνόλων μεθόδων μέσω ψηφοφορίας πλειοψηφίας παρήγαγε πιο ισορροπημένα αποτελέσματα και στα δύο σύνολα δεδομένων (Κεφάλαιο 6). Η ερμηνεία αυτής της διαπίστωσης είναι διττή: αφενός, δεν υπάρχει μία καθολικά βέλτιστη μέθοδος ανίχνευσης ανωμαλιών, καθώς κάθε μέθοδος αποτυπώνει διαφορετική πτυχή της «ανωμαλίας» [18], [19]. Αφετέρου, ο συνδυασμός μεθόδων μετριάζει τις ακραίες συμπεριφορές των μεμονωμένων μεθόδων, παρέχοντας αξιόπιστη βάση για λήψη αποφάσεων σε πραγματικά συστήματα IoT. Αναλυτικά αποτελέσματα παρατίθενται στο παράρτημα Γ (βλ. σελ. 131).

7.2 Συνεισφορά (Contributions)

Η παρούσα διπλωματική εργασία προσφέρει συνεισφορές τόσο σε επιστημονικό, όσο και σε πρακτικό επίπεδο.

7.2.1 Επικύρωση Διατομεακής Εφαρμοσιμότητας (Cross-Domain Validation)

Η πρώτη και βασική συνεισφορά της εργασίας έγκειται στην εμπειρική επιβεβαίωση ότι μια ενιαία μεθοδολογία μηχανικής μάθησης μπορεί να εφαρμοστεί αποτελεσματικά σε διαφορετικούς τομείς IoT. Η διαπίστωση ότι τα βέλτιστα μοντέλα και στους δύο τομείς ανήκουν στην κατηγορία των

δέντρων αποφάσεων— περιβαλλοντική παρακολούθηση και διαχείριση ενέργειας — παρέχει εμπειρικά τεκμηριωμένη βάση για την επιλογή αλγορίθμων σε νέες εφαρμογές IoT, χωρίς την ανάγκη εξαντλητικών πειραματισμών.

7.2.2 Αναπαραγώγιμη Μεθοδολογική Αγωγή (Reproducible Pipeline)

Η δεύτερη συνεισφορά αφορά την ανάπτυξη μιας πλήρως αναπαραγώγιμης αγωγού ανάλυσης, η οποία μπορεί να χρησιμεύσει ως πρότυπο αναφοράς για παρόμοιες εφαρμογές. Οι μηχανισμοί αναπαραγωγιμότητας που υιοθετήθηκαν — σταθερός σπόρος τυχαιότητας, τεκμηριωμένο περιβάλλον εκτέλεσης και δομημένη οργάνωση κώδικα κατά τη μεθοδολογία CRISP-DM — περιγράφονται αναλυτικά στο Κεφάλαιο 3.

7.2.3 Τεκμηρίωση Βέλτιστων Πρακτικών (Best Practices Documentation)

Η τρίτη συνεισφορά αφορά την εξαγωγή πρακτικών κατευθυντήριων γραμμών που καλύπτουν όλες τις φάσεις της ανάλυσης: από την προεπεξεργασία δεδομένων (γραμμική παρεμβολή, IQR clipping, κανονικοποίηση) και τη σχεδίαση χαρακτηριστικών (υστερήσεις, κυλιόμενα στατιστικά, κυκλική κωδικοποίηση), έως τη μοντελοποίηση και την ανίχνευση ανωμαλιών μέσω συνόλων μεθόδων. Τα ευρήματα αυτά μπορούν να αξιοποιηθούν άμεσα σε μελλοντικές εφαρμογές IoT μικρού έως μεσαίου μεγέθους.

7.3 Περιορισμοί (Limitations)

Κάθε επιστημονική εργασία υπόκειται σε περιορισμούς που επηρεάζουν τη γενικευσιμότητα και την εφαρμοσιμότητα των ευρημάτων της. Στην ενότητα αυτή παρουσιάζονται οι κύριοι περιορισμοί της παρούσας έρευνας.

7.3.1 Επίδραση του Μεγέθους Δεδομένων στη Βαθιά Μάθηση

Ο πρώτος περιορισμός αφορά τον αντίκτυπο του περιορισμένου μεγέθους δεδομένων στην αξιολόγηση μοντέλων βαθιάς μάθησης. Το σύνολο δεδομένων ποιότητας αέρα, με 9.471 αρχικές παρατηρήσεις που μειώθηκαν σε 890 δείγματα εκπαίδευσης μετά την προεπεξεργασία και τη δημιουργία ακολουθιών, δεν επαρκεί για την αξιοποίηση του πλήρους δυναμικού αρχιτεκτονικών όπως LSTM και GRU. Κατά συνέπεια, η υποαπόδοσή τους δεν υποδηλώνει αναγκαστικά

αδυναμία των αρχιτεκτονικών αυτών, αλλά αντανακλά την ανεπάρκεια δεδομένων για την κατάλληλη εκπαίδευσή τους, όπως αναλύθηκε διεξοδικά στο Κεφάλαιο 4.

7.3.2 Περιορισμένος Χρονικός Ορίζοντας Πρόβλεψης

Η μεθοδολογία εφαρμόστηκε σε δύο μόνο πεδία IoT — ποιότητα αέρα και κατανάλωση ηλεκτρικής ενέργειας. Παρότι τα δύο σύνολα δεδομένων παρουσιάζουν διαφορετικά χαρακτηριστικά (μέγεθος, συχνότητα δειγματοληψίας, αριθμός μεταβλητών), η γενίκευση των ευρημάτων σε άλλους τομείς IoT, όπως η βιομηχανική παρακολούθηση ή τα δίκτυα αισθητήρων υγείας, απαιτεί περαιτέρω επιβεβαίωση.

7.3.3 Απουσία Ανάπτυξης σε Πραγματικό Περιβάλλον

Η εκπαίδευση πραγματοποιήθηκε σε περιβάλλον περιορισμένων υπολογιστικών πόρων (Google Colab), γεγονός που επηρέασε τόσο τον αριθμό των μοντέλων που αξιολογήθηκαν όσο και την έκταση της βελτιστοποίησης υπερπαραμέτρων. Ειδικότερα, στο σύνολο δεδομένων ηλεκτρικής ενέργειας, η εφαρμογή μοντέλων βαθιάς μάθησης δεν ήταν εφικτή λόγω του μεγέθους του, περιορίζοντας τη σύγκριση σε μοντέλα βασισμένα σε δέντρα αποφάσεων.

7.4 Μελλοντική Εργασία (Future Work)

Με βάση τα ευρήματα και τους περιορισμούς που αναδείχθηκαν, η ενότητα αυτή παρουσιάζει τρεις κατευθύνσεις για μελλοντική έρευνα που θα μπορούσαν να επεκτείνουν ή/και να βελτιώσουν την παρούσα εργασία.

7.4.1 Πρόβλεψη Πολλαπλών Χρονικών Οριζόντων (Multi-Horizon Forecasting)

Η πρώτη κατεύθυνση αφορά την επέκταση της πρόβλεψης σε πολλαπλούς χρονικούς ορίζοντες, αξιολογώντας την απόδοση των μοντέλων καθώς αυξάνεται ο χρονικός ορίζοντας πρόβλεψης. Συγκεκριμένα, μια μελλοντική μελέτη θα μπορούσε να εξετάσει τρεις ορίζοντες: βραχυπρόθεσμο (1-6 ώρες), μεσοπρόθεσμο (24 ώρες) και μακροπρόθεσμο (7 ημέρες) [17].

Η σύγκριση διαφορετικών στρατηγικών πρόβλεψης πολλαπλών βημάτων αποτελεί ιδιαίτερα ενδιαφέρουσα ερευνητική κατεύθυνση. Η αναδρομική στρατηγική (recursive strategy) χρησιμοποιεί το ίδιο μοντέλο επαναληπτικά, τροφοδοτώντας τις προβλέψεις ως εισόδους για

επόμενα βήματα. Η άμεση στρατηγική (direct strategy) εκπαιδεύει ξεχωριστά μοντέλα για κάθε ορίζοντα πρόβλεψης. Ενώ, οι υβριδικές στρατηγικές συνδυάζουν στοιχεία και των δύο προσεγγίσεων [17].

Επιπλέον, η ανάλυση της υποβάθμισης ακρίβειας (accuracy degradation) καθώς αυξάνεται ο ορίζοντας θα παρείχε πρακτικές κατευθυντήριες γραμμές για την επιλογή του κατάλληλου ορίζοντα πρόβλεψης, ανάλογα με τις απαιτήσεις της εφαρμογής.

7.4.2 Ποσοτικοποίηση Αβεβαιότητας (Uncertainty Quantification)

Η δεύτερη κατεύθυνση αφορά την ενσωμάτωση μεθόδων ποσοτικοποίησης αβεβαιότητας στις προβλέψεις. Αντί για σημειακές προβλέψεις, η παροχή διαστημάτων εμπιστοσύνης ή πιθανοτικών κατανομών, θα αυξήσει σημαντικά την πρακτική χρησιμότητα των μοντέλων [5], [6].

Διάφορες τεχνικές μπορούν να εξεταστούν για αυτόν τον σκοπό. Η διαμόρφωση προβλέψεων (conformal prediction) παρέχει θεωρητικά εγγυημένα διαστήματα πρόβλεψης, χωρίς υποθέσεις για την κατανομή των δεδομένων. Οι μέθοδοι Bayesian παρέχουν πλήρεις πιθανοτικές κατανομές για τις προβλέψεις, επιτρέποντας την ποσοτικοποίηση τόσο της αλεατορικής, όσο και της επιστημικής αβεβαιότητας. Αλλά, και οι μέθοδοι συνόλων (ensemble methods) μπορούν να χρησιμοποιηθούν για την εκτίμηση της αβεβαιότητας μέσω της διακύμανσης των προβλέψεων μεταξύ διαφορετικών μοντέλων [7].

Η ποσοτικοποίηση αβεβαιότητας είναι ιδιαίτερα σημαντική για εφαρμογές IoT, όπου οι αποφάσεις που βασίζονται σε προβλέψεις μπορεί να έχουν σοβαρές συνέπειες. Για παράδειγμα, στη διαχείριση ηλεκτρικού δικτύου, η γνώση της αβεβαιότητας των προβλέψεων ζήτησης επιτρέπει τον καλύτερο σχεδιασμό αποθεμάτων ασφαλείας [24].

7.4.3 Ανάπτυξη σε Παραγωγικό Περιβάλλον (Production Deployment)

Η τρίτη κατεύθυνση αφορά τη μετάβαση από την ερευνητική ανάπτυξη στην πραγματική λειτουργική ανάπτυξη. Αυτή η μετάβαση απαιτεί την αντιμετώπιση σειράς τεχνικών προκλήσεων που δεν εξετάστηκαν στην παρούσα εργασία [23], [24].

- Πρώτον, η υλοποίηση μηχανισμών ανίχνευσης μετατόπισης δεδομένων (data drift detection) είναι απαραίτητη για τη διατήρηση της απόδοσης του μοντέλου με την πάροδο του χρόνου. Τεχνικές όπως η παρακολούθηση στατιστικών δεικτών των εισόδων (input monitoring) και η σύγκριση κατανομών (distribution comparison), μπορούν να χρησιμοποιηθούν για την έγκαιρη ανίχνευση αλλαγών [4].
- Δεύτερον, η βελτιστοποίηση των μοντέλων για εκτέλεση σε συσκευές edge computing, θα επέκτεινε σημαντικά το πεδίο εφαρμογών. Τεχνικές όπως η περικοπή μοντέλων (model pruning), η κβαντοποίηση (quantization) και η απόσταξη γνώσης (knowledge distillation), μπορούν να μειώσουν το υπολογιστικό κόστος διατηρώντας ικανοποιητική ακρίβεια [13].
- Τρίτον, η ανάπτυξη αυτοματοποιημένων διαδικασιών επανεκπαίδευσης (automated retraining pipelines), θα επέτρεπε την προσαρμογή των μοντέλων σε μεταβαλλόμενες συνθήκες χωρίς χειροκίνητη παρέμβαση. Αυτές οι διαδικασίες θα πρέπει να περιλαμβάνουν κριτήρια ενεργοποίησης επανεκπαίδευσης, αυτοματοποιημένη επιλογή υπερπαραμέτρων και μηχανισμούς επικύρωσης πριν την ανάπτυξη νέων μοντέλων.

ΒΙΒΛΙΟΓΡΑΦΙΑ

1. S. De Vito, E. Massera, M. Piga, L. Martinotto, and G. Di Francia, "On field calibration of an electronic nose for benzene estimation in an urban pollution monitoring scenario," *Sensors and Actuators B: Chemical*, vol. 129, no. 2, pp. 750-757, Feb. 2008, doi: 10.1016/j.snb.2007.09.060.
2. S. Vito, "Air Quality," UCI Machine Learning Repository, 2008. [Online]. Available: <https://doi.org/10.24432/C59K5F>
3. G. Hébrail and A. Bérard, "Individual household electric power consumption Data Set," UCI Machine Learning Repository, 2012. [Online]. Available: <https://archive.ics.uci.edu/ml/datasets/Individual+household+electric+power+consumption>
4. I. H. Witten, E. Frank, M. A. Hall, and C. J. Pal, *Data Mining: Practical Machine Learning Tools and Techniques*, 4th ed. Burlington, MA: Morgan Kaufmann, 2016.
5. T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed. New York: Springer, 2009.
6. G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning: With Applications in R*, 2nd ed. New York: Springer, 2021.
7. L. Breiman, "Random Forests," *Machine Learning*, vol. 45, no. 1, pp. 5-32, Oct. 2001, doi: 10.1023/A:1010933404324.
8. T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in *Proc. 22nd ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining*, San Francisco, CA, 2016, pp. 785-794, doi: 10.1145/2939672.2939785.

9. J. H. Friedman, "Greedy function approximation: A gradient boosting machine," *Annals of Statistics*, vol. 29, no. 5, pp. 1189-1232, Oct. 2001, doi: 10.1214/aos/1013203451.
10. F. Pedregosa et al., "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825-2830, Oct. 2011.
11. S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," *Neural Computation*, vol. 9, no. 8, pp. 1735-1780, Nov. 1997, doi: 10.1162/neco.1997.9.8.1735.
12. K. Cho et al., "Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation," in *Proc. 2014 Conf. on Empirical Methods in Natural Language Processing (EMNLP)*, Doha, Qatar, 2014, pp. 1724-1734, doi: 10.3115/v1/D14-1179.
13. I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, MA: MIT Press, 2016.
14. J. Brownlee, *Deep Learning for Time Series Forecasting: Predict the Future with MLPs, CNNs and LSTMs in Python. Machine Learning Mastery*, 2018.
15. G. E. P. Box, G. M. Jenkins, G. C. Reinsel, and G. M. Ljung, *Time Series Analysis: Forecasting and Control*, 5th ed. Hoboken, NJ: Wiley, 2015.
16. C. Bergmeir and J. M. Benítez, "On the use of cross-validation for time series predictor evaluation," *Information Sciences*, vol. 191, pp. 192-213, May 2012, doi: 10.1016/j.ins.2011.12.028.
17. R. J. Hyndman and G. Athanasopoulos, *Forecasting: Principles and Practice*, 3rd ed. Melbourne, Australia: OTexts, 2021. [Online]. Available: <https://otexts.com/fpp3/>

18. V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: A survey," *ACM Computing Surveys*, vol. 41, no. 3, pp. 1-58, Jul. 2009, doi: 10.1145/1541880.1541882.
19. F. T. Liu, K. M. Ting, and Z.-H. Zhou, "Isolation Forest," in *Proc. 8th IEEE Int. Conf. on Data Mining (ICDM)*, Pisa, Italy, 2008, pp. 413-422, doi: 10.1109/ICDM.2008.17.
20. M. M. Breunig, H.-P. Kriegel, R. T. Ng, and J. Sander, "LOF: Identifying Density-Based Local Outliers," in *Proc. ACM SIGMOD Int. Conf. on Management of Data*, Dallas, TX, 2000, pp. 93-104, doi: 10.1145/342009.335388.
21. B. Schölkopf, J. C. Platt, J. Shawe-Taylor, A. J. Smola, and R. C. Williamson, "Estimating the Support of a High-Dimensional Distribution," *Neural Computation*, vol. 13, no. 7, pp. 1443-1471, Jul. 2001, doi: 10.1162/089976601750264965.
22. R. B. Cleveland, W. S. Cleveland, J. E. McRae, and I. Terpenning, "STL: A Seasonal-Trend Decomposition Procedure Based on Loess," *Journal of Official Statistics*, vol. 6, no. 1, pp. 3-73, 1990.
23. L. Atzori, A. Iera, and G. Morabito, "The Internet of Things: A survey," *Computer Networks*, vol. 54, no. 15, pp. 2787-2805, Oct. 2010, doi: 10.1016/j.comnet.2010.05.010.
24. J. Gubbi, R. Buyya, S. Marusic, and M. Palaniswami, "Internet of Things (IoT): A vision, architectural elements, and future directions," *Future Generation Computer Systems*, vol. 29, no. 7, pp. 1645-1660, Sep. 2013, doi: 10.1016/j.future.2013.01.010.
25. R. Wirth and J. Hipp, "CRISP-DM: Towards a Standard Process Model for Data Mining," in *Proc. 4th Int. Conf. on the Practical Applications of Knowledge Discovery and Data Mining*, Manchester, UK, 2000, pp. 29-39.

ΠΑΡΑΡΤΗΜΑΤΑ

ΠΑΡΑΡΤΗΜΑ Α: Πίνακας Ορολογίας (Greek-English Terminology)

ΕΛΛΗΝΙΚΑ	ENGLISH
Αγωγός ανάλυσης	Analysis pipeline
Αισθητήρας	Sensor
Ακραία τιμή	Outlier
Αλγόριθμος	Algorithm
Αναπαραγωγιμότητα	Reproducibility
Ανίχνευση ανωμαλιών	Anomaly detection
Αποδόμηση / Αποσύνθεση STL	STL Decomposition
Αποκοπή IQR	IQR clipping
Αυτοσυσχέτιση	Autocorrelation
Βαθιά μάθηση	Deep learning
Γραμμική παλινδρόμηση	Linear regression
Γραμμική παρεμβολή	Linear interpolation
Δέντρο αποφάσεων	Decision tree
Διαδίκτυο των Πραγμάτων (ΔτΠ)	Internet of Things (IoT)

Διαρροή πληροφορίας	Data leakage
Διασταυρούμενη επικύρωση	Cross-validation
Διασταυρούμενη επικύρωση χρονοσειρών	Time-series cross-validation
Διερευνητική ανάλυση δεδομένων (ΔΑΔ)	Exploratory Data Analysis (EDA)
Ελλείπουσες τιμές	Missing values
Ενισχυμένη κλίση (ενίσχυση διαβάθμισης)	Gradient Boosting
Εξόρυξη δεδομένων	Data mining
Εποχικότητα	Seasonality
Κανονικοποίηση	Normalization
Κατάλοιπα	Residuals
Κυκλικά χρονικά χαρακτηριστικά	Cyclical time features
Κυλιόμενο παράθυρο / Κυλιόμενα στατιστικά	Rolling window / Rolling statistics
Μέθοδος συνόλου	Ensemble method
Μέσο απόλυτο σφάλμα	Mean Absolute Error (MAE)
Μηχανική μάθηση (MM)	Machine Learning (ML)
Μηχανική χαρακτηριστικών	Feature engineering
Μοντελοποίηση	Modeling

Νευρωνικό δίκτυο	Neural network
Παράθυρο εκπαίδευσης (διευρυνόμενο)	Training window (expanding)
Ποιότητα αέρα	Air quality
Προεπεξεργασία	Preprocessing
Πρόβλεψη χρονοσειρών	Time-series forecasting
Σημαντικότητα χαρακτηριστικών	Feature importance
Σπόρος τυχειότητας	Random seed
Σύνολο δεδομένων	Dataset
Σύνολο δοκιμής	Test set
Σύνολο εκπαίδευσης	Training set
Σύνολο επικύρωσης	Validation set
Συντελεστής προσδιορισμού	Coefficient of determination (R^2)
Τάση	Trend
Τεχνητή νοημοσύνη (TN)	Artificial Intelligence (AI)
Τυπική απόκλιση	Standard deviation
Υπερεκπαίδευση / Υπερπροσαρμογή	Overfitting
Χαρακτηριστικά αλληλεπίδρασης	Interaction features

Χαρακτηριστικά υστέρησης	Lag features
Χρονικός διαχωρισμός	Temporal split
Χρονοσειρά	Time series
Ψευδώς θετικό	False positive
Ψηφοφορία πλειοψηφίας	Majority voting

ΠΑΡΑΡΤΗΜΑ Β: Αποθετήριο του Κώδικα στο GitHub

Ο πλήρης κώδικας μαζί με τα αποτελέσματα της ανάλυσης της παρούσας εργασίας (/data & /figures directories) καθώς και το ιστορικό εξέλιξής του (commits) βρίσκονται σε ιδιωτικό αποθετήριο στο github και εφόσον παρέχεται πρόσβαση, μπορούν να βρεθούν στον παρακάτω σύνδεσμο:

<https://github.com/aaoura/thesis>

ΠΑΡΑΡΤΗΜΑ Γ: Αναλυτικοί Πίνακες Αποτελεσμάτων

Πίνακας Περιεχομένων Αναλυτικών Αποτελεσμάτων & Δεδομένων

01_EDA	142
02_Preprocessing	161
03_Modeling	167
04_Anomaly.....	183
05_EDA_Power	184
06_Preprocessing_Power.....	201
07_Modeling_Power.....	206
08_Anomaly_Power	209

01_EDA

Folder: data\results\01_EDA

01_first_rows.csv

File size: 1.1 KB — Rows included: 10

Date	Time	CO(GT)	PT08 .S1(CO)	NM HC(GT)	C6H 6(GT)	PT08 .S2(NM HC)	NOx(GT)	PT08 .S3(NOx)	NO2(GT)	PT08 .S4(NO2)	PT08 .S5(O3)	T	RH	AH
10/03 /2004	18.00 .00	2.6	1360. 0	150.0	11.9	1046. 0	166.0	1056. 0	113.0	1692. 0	1268. 0	13.6	48.9	0.757 8
10/03 /2004	19.00 .00	2.0	1292. 0	112.0	9.4	955.0	103.0	1174. 0	92.0	1559. 0	972.0	13.3	47.7	0.725 5
10/03 /2004	20.00 .00	2.2	1402. 0	88.0	9.0	939.0	131.0	1140. 0	114.0	1555. 0	1074. 0	11.9	54.0	0.750 2
10/03 /2004	21.00 .00	2.2	1376. 0	80.0	9.2	948.0	172.0	1092. 0	122.0	1584. 0	1203. 0	11.0	60.0	0.786 7
10/03 /2004	22.00 .00	1.6	1272. 0	51.0	6.5	836.0	131.0	1205. 0	116.0	1490. 0	1110. 0	11.2	59.6	0.788 8
10/03 /2004	23.00 .00	1.2	1197. 0	38.0	4.7	750.0	89.0	1337. 0	96.0	1393. 0	949.0	11.2	59.2	0.784 8
11/03 /2004	00.00 .00	1.2	1185. 0	31.0	3.6	690.0	62.0	1462. 0	77.0	1333. 0	733.0	11.3	56.8	0.760 3

11/03 /2004	01.00 .00	1.0	1136. 0	31.0	3.3	672.0	62.0	1453. 0	76.0	1333. 0	730.0	10.7	60.0	0.770 2
11/03 /2004	02.00 .00	0.9	1094. 0	24.0	2.3	609.0	45.0	1579. 0	60.0	1276. 0	620.0	10.7	59.7	0.764 8
11/03 /2004	03.00 .00	0.6	1010. 0	19.0	1.7	561.0	- 200.0	1705. 0	- 200.0	1235. 0	501.0	10.3	60.2	0.751 7

05_data_types.csv

File size: 273.0 B — Rows included: 15

Column	DataType
Date	object
Time	object
CO(GT)	float64
PT08.S1(CO)	float64
NMHC(GT)	float64
C6H6(GT)	float64
PT08.S2(NMHC)	float64
NOx(GT)	float64
PT08.S3(NOx)	float64
NO2(GT)	float64
PT08.S4(NO2)	float64
PT08.S5(O3)	float64
T	float64
RH	float64
AH	float64

06_missing_values_nan.csv

File size: 516.0 B — Rows included: 15

Column	Missing_Count	Percentage
Date	114	1.2036743744060818
Time	114	1.2036743744060818
CO(GT)	114	1.2036743744060818
PT08.S1(CO)	114	1.2036743744060818
NMHC(GT)	114	1.2036743744060818
C6H6(GT)	114	1.2036743744060818
PT08.S2(NMHC)	114	1.2036743744060818
NOx(GT)	114	1.2036743744060818
PT08.S3(NOx)	114	1.2036743744060818

NO2(GT)	114	1.2036743744060818
PT08.S4(NO2)	114	1.2036743744060818
PT08.S5(O3)	114	1.2036743744060818
T	114	1.2036743744060818
RH	114	1.2036743744060818
AH	114	1.2036743744060818

07_invalid_values_minus200.csv

File size: 460.0 B — Rows included: 13

Column	Count_of_-200	Percentage
NMHC(GT)	8443	89.14581353605743
CO(GT)	1683	17.770034843205575
NO2(GT)	1642	17.33713441030514
NOx(GT)	1639	17.305458768873404
PT08.S1(CO)	366	3.8644282546721573
PT08.S2(NMHC)	366	3.8644282546721573
C6H6(GT)	366	3.8644282546721573
PT08.S3(NOx)	366	3.8644282546721573
PT08.S4(NO2)	366	3.8644282546721573
PT08.S5(O3)	366	3.8644282546721573
T	366	3.8644282546721573
RH	366	3.8644282546721573
AH	366	3.8644282546721573

08_statistical_summary.csv

File size: 1019.0 B — Rows included: 13

count	mean	std	min	25%	50%	75%	max
9357.0	- 34.2075237 7898899	77.6571703 4683162	-200.0	0.6	1.5	2.6	11.9
9357.0	1048.99006 09169606	329.832709 88618807	-200.0	921.0	1053.0	1221.0	2040.0
9357.0	- 159.090092 97851875	139.789092 89678145	-200.0	-200.0	-200.0	-200.0	1189.0
9357.0	1.86568344 55487874	41.3802064 3630011	-200.0	4.0	7.9	13.6	63.7
9357.0	894.595276 2637597	342.333251 6496718	-200.0	711.0	895.0	1105.0	2214.0
9357.0	168.616971 2514695	257.433866 29107385	-200.0	50.0	141.0	284.0	1479.0
9357.0	794.990167 7888212	321.993551 603116	-200.0	637.0	794.0	960.0	2683.0
9357.0	58.1488725 0187026	126.940455 2589075	-200.0	53.0	96.0	133.0	340.0

9357.0	1391.47964 09105481	467.210124 55706364	-200.0	1185.0	1446.0	1662.0	2775.0
9357.0	975.072031 6340708	456.938184 4739884	-200.0	700.0	942.0	1255.0	2523.0
9357.0	9.77830501 2290263	43.2036230 62623194	-200.0	10.9	17.2	24.1	44.6
9357.0	39.4853799 29464564	51.2161449 7176604	-200.0	34.1	48.6	61.9	88.7
9357.0	- 6.83760364 4330447	38.9766701 6632424	-200.0	0.6923	0.9768	1.2962	2.231

11_temporal_features_sample.csv

File size: 1.1 KB — Rows included: 20

DateTime	Year	Month	Day	Hour	DayOfWeek	DayName
2004-03-10 18:00:00	2004.0	3.0	10.0	18.0	2.0	Wednesday
2004-03-10 19:00:00	2004.0	3.0	10.0	19.0	2.0	Wednesday
2004-03-10 20:00:00	2004.0	3.0	10.0	20.0	2.0	Wednesday
2004-03-10 21:00:00	2004.0	3.0	10.0	21.0	2.0	Wednesday

2004-03-10 22:00:00	2004.0	3.0	10.0	22.0	2.0	Wednesday
2004-03-10 23:00:00	2004.0	3.0	10.0	23.0	2.0	Wednesday
2004-03-11 00:00:00	2004.0	3.0	11.0	0.0	3.0	Thursday
2004-03-11 01:00:00	2004.0	3.0	11.0	1.0	3.0	Thursday
2004-03-11 02:00:00	2004.0	3.0	11.0	2.0	3.0	Thursday
2004-03-11 03:00:00	2004.0	3.0	11.0	3.0	3.0	Thursday
2004-03-11 04:00:00	2004.0	3.0	11.0	4.0	3.0	Thursday
2004-03-11 05:00:00	2004.0	3.0	11.0	5.0	3.0	Thursday
2004-03-11 06:00:00	2004.0	3.0	11.0	6.0	3.0	Thursday
2004-03-11 07:00:00	2004.0	3.0	11.0	7.0	3.0	Thursday
2004-03-11 08:00:00	2004.0	3.0	11.0	8.0	3.0	Thursday
2004-03-11 09:00:00	2004.0	3.0	11.0	9.0	3.0	Thursday

2004-03-11 10:00:00	2004.0	3.0	11.0	10.0	3.0	Thursday
2004-03-11 11:00:00	2004.0	3.0	11.0	11.0	3.0	Thursday
2004-03-11 12:00:00	2004.0	3.0	11.0	12.0	3.0	Thursday
2004-03-11 13:00:00	2004.0	3.0	11.0	13.0	3.0	Thursday

12_monthly_distribution.csv

File size: 125.0 B — Rows included: 12

Month	Count
1.0	744
2.0	672
3.0	1254
4.0	807
5.0	744
6.0	720
7.0	744
8.0	744
9.0	720
10.0	744
11.0	720

12.0	744
------	-----

13_hourly_distribution.csv

File size: 242.0 B — Rows included: 24

Hour	Count
0.0	390
1.0	390
2.0	390
3.0	390
4.0	390
5.0	390
6.0	390
7.0	390
8.0	390
9.0	390
10.0	390
11.0	390
12.0	390
13.0	390
14.0	390
15.0	389
16.0	389

17.0	389
18.0	390
19.0	390
20.0	390
21.0	390
22.0	390
23.0	390

14_missing_values_percentage.csv

File size: 506.0 B — Rows included: 18

Unnamed: 0	Percentage
Day	1.2036743744060818
Year	1.2036743744060818
Month	1.2036743744060818
Hour	1.2036743744060818
DayOfWeek	1.2036743744060818
C6H6(GT)	5.068102629078239
PT08.S4(NO2)	5.068102629078239
PT08.S1(CO)	5.068102629078239
RH	5.068102629078239
T	5.068102629078239
PT08.S5(O3)	5.068102629078239

AH	5.068102629078239
PT08.S3(NOx)	5.068102629078239
PT08.S2(NMHC)	5.068102629078239
NOx(GT)	18.509133143279485
NO2(GT)	18.540808784711224
CO(GT)	18.973709217611656
NMHC(GT)	90.34948791046352

15_distribution_statistics.csv

File size: 595.0 B — Rows included: 7

Feature	Mean	Median	Std	Min	Max	Skewness
CO(GT)	2.1527495439145 166	1.8	1.45325203633734 87	0.1	11.9	1.36975277812920 94
C6H6(GT)	10.083105327549 772	8.2	7.44981969830013 7	0.1	63.7	1.36153227002764 89
NOx(GT)	246.89673490541 59	180.0	212.979168103586 12	2.0	1479.0	1.71578079928154 15
NO2(GT)	113.09125081011 017	109.0	48.3701077822253 96	2.0	340.0	0.62171431343737 14
T	18.317828940051 164	17.8	8.83211573182976 8	-1.9	44.6	0.30935679206515 69

RH	49.234200867534 2	49.6	17.3168924620462 05	9.2	88.7	- 0.03792800990133 9856
AH	1.0255302747191 637	0.9954	0.40381260604769 13	0.1847	2.231	0.25138775552886 916

16_correlation_matrix.csv

File size: 2.7 KB — Rows included: 12

CO(GT)	PT08.S 1(CO)	C6H6(GT)	PT08.S 2(NM HC)	NOx(G T)	PT08.S 3(NOx)	NO2(G T)	PT08.S 4(NO2)	PT08.S 5(O3)	T	RH	AH
1.0	0.8792 883410 762707	0.9310 783250 179333	0.9155 140776 13614	0.7950 284494 455478	- 0.7034 461656 756494	0.6833 426502 920655	0.6307 033375 784343	0.8541 818610 610992	0.0221 091231 696742 83	0.0488 897056 121002 2	0.0485 560233 462453 8
0.8792 883410 762707	1.0	0.8837 950766 080735	0.8929 643887 070403	0.7136 536362 545512	- 0.7719 384679 420122	0.6415 290821 82378	0.6828 812474 84672	0.8993 235214 436066	0.0486 267581 528159 9	0.1146 057306 534849 4	0.1353 237555 985924 2
0.9310 783250 179333	0.8837 950766 080735	1.0	0.9819 503158 012699	0.7188 391336 1092	- 0.7357 441198 223641	0.6144 739375 310503	0.7657 313582 719674	0.8656 885045 054717	0.1989 558410 384465 7	- 0.0616 807912	0.1679 719226 611319

										508599 2	
0.9155 140776 13614	0.8929 643887 070403	0.9819 503158 012699	1.0	0.7044 347181 027009	- 0.7967 033835 149517	0.6462 452823 985936	0.7772 544590 158323	0.8805 777489 667843	0.2413 731020 539108 3	- 0.0903 800684 807566 6	0.1869 329579 887099 9
0.7950 284494 455478	0.7136 536362 545512	0.7188 391336 1092	0.7044 347181 027009	1.0	- 0.6557 070939 270037	0.7631 110812 804082	0.2337 309078 998052 7	0.7870 459170 361116	- 0.2696 829023 757538	0.2210 317472 391004 6	- 0.1493 231388 409063 4
- 0.7034 461656 756494	- 0.7719 384679 420122	- 0.7357 441198 223641	- 0.7967 033835 149517	- 0.6557 070939 270037	1.0	- 0.6520 830245 066562	- 0.5384 678766 555988	- 0.7965 691609 602705	- 0.1451 123994 805345	- 0.0567 395080 124560 5	- 0.2320 171239 127830 6
0.6833 426502 920655	0.6415 290821 82378	0.6144 739375 310503	0.6462 452823 985936	0.7631 110812 804082	- 0.6520 830245 066562	1.0	0.1573 603208 287947 8	0.7081 276371 243194	- 0.1865 332682 736816 6	- 0.0917 589334 298317 8	- 0.3350 220192 367284

0.6307	0.6828	0.7657	0.7772	0.2337	-	0.1573	1.0	0.5911	0.5612	-	0.6296
033375	812474	313582	544590	309078	0.5384	603208		439877	696735	0.0321	412232
784343	84672	719674	158323	998052	678766	287947		338839	562387	881772	117952
				7	555988	8				060273	
										7	
0.8541	0.8993	0.8656	0.8805	0.7870	-	0.7081	0.5911	1.0	-	0.1249	0.0707
818610	235214	885045	777489	459170	0.7965	276371	439877		0.0271	563919	514348
610992	436066	054717	667843	361116	691609	243194	338839		721205	519555	057200
					602705				946446	1	2
									8		
0.0221	0.0486	0.1989	0.2413	-	-	-	0.5612	-	1.0	-	0.6563
091231	267581	558410	731020	0.2696	0.1451	0.1865	696735	0.0271		0.5786	969795
696742	528159	384465	539108	829023	123994	332682	562387	721205		208353	767136
83	9	7	3	757538	805345	736816		946446		974083	
						6		8			
0.0488	0.1146	-	-	0.2210	-	-	-	0.1249	-	1.0	0.1679
897056	057306	0.0616	0.0903	317472	0.0567	0.0917	0.0321	563919	0.5786		709216
121002	534849	807912	800684	391004	395080	589334	881772	519555	208353		301344
2	4	508599	807566	6	124560	298317	060273	1	974083		4
		2	6		5	8	7				
0.0485	0.1353	0.1679	0.1869	-	-	-	0.6296	0.0707	0.6563	0.1679	1.0
560233	237555	719226	329579	0.1493	0.2320	0.3350	412232	514348	969795	709216	
462453	985924	611319	887099	231388	171239	220192	117952	057200	767136	301344	
8	2		9			367284		2		4	

				409063	127830						
				4	6						

17_high_correlations.csv

File size: 446.0 B — Rows included: 10

Feature1	Feature2	Correlation
C6H6(GT)	PT08.S2(NMHC)	0.9819503158012699
CO(GT)	C6H6(GT)	0.9310783250179333
CO(GT)	PT08.S2(NMHC)	0.915514077613614
PT08.S1(CO)	PT08.S5(O3)	0.8993235214436066
PT08.S1(CO)	PT08.S2(NMHC)	0.8929643887070403
PT08.S1(CO)	C6H6(GT)	0.8837950766080735
PT08.S2(NMHC)	PT08.S5(O3)	0.8805777489667843
CO(GT)	PT08.S1(CO)	0.8792883410762707
C6H6(GT)	PT08.S5(O3)	0.8656885045054717
CO(GT)	PT08.S5(O3)	0.8541818610610992

18_daily_averaged_timeseries.csv

File size: 22.5 KB — Rows included: 100 — Original rows: 391

DateTime	CO(GT)	NO2(GT)	C6H6(GT)
2004-03-10	1.9666666666666668	108.83333333333333	8.450000000000001

2004-03-11	2.239130434782609	99.8695652173913	7.979166666666667
2004-03-12	2.8045454545454547	116.27272727272727	12.129166666666668
2004-03-13	2.6956521739130435	118.8695652173913	10.916666666666666
2004-03-14	2.4695652173913043	110.3913043478261	9.637500000000001
2004-03-15	3.5043478260869563	118.73913043478261	16.091666666666665
2004-03-16	2.769565217391304	118.17391304347827	13.5
2004-03-17	3.073913043478261	119.95652173913044	15.237499999999999
2004-03-18	2.7	125.1304347826087	13.795833333333334
2004-03-19	2.965217391304348	118.08695652173913	13.520833333333334
2004-03-20	2.11304347826087	97.30434782608695	8.754166666666666
2004-03-21	1.9913043478260868	89.17391304347827	7.595833333333334
2004-03-22	2.139130434782609	90.52173913043478	9.4125
2004-03-23	2.0043478260869567	97.0	8.9625
2004-03-24	2.1	96.52173913043478	8.625
2004-03-25	1.9041666666666668	89.30434782608695	8.095833333333333
2004-03-26	1.8791666666666667	92.43478260869566	8.424999999999999
2004-03-27	1.8434782608695652	93.04347826086956	7.7625
2004-03-28	1.5625	79.1304347826087	5.566666666666666
2004-03-29	1.4666666666666668	84.56521739130434	5.570833333333333
2004-03-30	2.2	102.26086956521739	9.508333333333333
2004-03-31	1.925	96.43478260869566	7.870833333333334
2004-04-01	2.5500000000000003	96.73913043478261	11.971428571428572
2004-04-02	2.8000000000000003	118.52173913043478	13.208333333333334
2004-04-03	1.56	86.33333333333333	10.887500000000001

2004-04-04			7.2124999999999995
2004-04-05	2.757142857142857	107.92857142857143	10.845833333333333
2004-04-06	2.2333333333333334	96.04347826086956	10.941666666666668
2004-04-07	1.6375	85.26086956521739	6.9750000000000005
2004-04-08	2.5782608695652174	105.91304347826087	12.039130434782608
2004-04-09	2.4541666666666666	112.91304347826087	8.4
2004-04-10	1.7541666666666667	90.65217391304348	7.445833333333333
2004-04-11	1.2043478260869565	67.95652173913044	4.058333333333334
2004-04-12	0.8624999999999999	66.1304347826087	2.345833333333333
2004-04-13	2.0541666666666667	93.95652173913044	9.404166666666667
2004-04-14	2.1416666666666666	83.5	10.3125
2004-04-15	4.1866666666666665	130.53333333333333	14.579166666666666
2004-04-16	3.3041666666666667	113.69565217391305	13.341666666666667
2004-04-17	2.2333333333333334	90.33333333333333	6.479166666666667
2004-04-18			6.016666666666667
2004-04-19			7.8500000000000005
2004-04-20			13.0375
2004-04-21			14.375
2004-04-22			14.108333333333334
2004-04-23	3.5631578947368423	135.0	16.933333333333334
2004-04-24	2.091304347826087	91.73913043478261	10.454166666666667
2004-04-25	1.2434782608695654	63.608695652173914	4.129166666666666
2004-04-26	1.4956521739130435	77.30434782608695	5.808333333333334
2004-04-27	2.226086956521739	94.6086956521739	11.316666666666668

2004-04-28	2.821739130434783	111.8695652173913	15.179166666666667
2004-04-29	3.130434782608696	105.1304347826087	16.208333333333332
2004-04-30	2.9695652173913043	103.6086956521739	14.170833333333334
2004-05-01	2.0565217391304347	84.95652173913044	8.6625
2004-05-02	1.608695652173913	79.69565217391305	7.270833333333333
2004-05-03	2.1695652173913045	89.65217391304348	11.549999999999999
2004-05-04	2.3043478260869565	84.47826086956522	10.891666666666667
2004-05-05	2.109090909090909	84.5909090909091	10.1875
2004-05-06	1.9956521739130435	89.6086956521739	8.775
2004-05-07	2.1913043478260867	96.78260869565217	9.5375
2004-05-08	1.4000000000000001	66.0	7.129166666666666
2004-05-09			6.695833333333333
2004-05-10	2.546666666666667	106.6	13.137500000000001
2004-05-11	2.7956521739130435	103.34782608695652	15.491666666666667
2004-05-12	2.630434782608696	100.21739130434783	13.604166666666666
2004-05-13	2.404347826086956	104.52173913043478	10.970833333333333
2004-05-14	2.4272727272727272	115.81818181818181	13.020833333333334
2004-05-15	1.7826086956521738	93.82608695652173	9.058333333333334
2004-05-16	1.4000000000000001	81.43478260869566	6.95
2004-05-17	1.682608695652174	91.69565217391305	8.237499999999999
2004-05-18	2.2869565217391306	101.08695652173913	12.466666666666667
2004-05-19	2.365217391304348	113.21739130434783	12.958333333333334
2004-05-20	2.4260869565217393	113.47826086956522	13.279166666666667
2004-05-21	2.265217391304348	108.30434782608695	12.6625

2004-05-22	2.008695652173913	91.82608695652173	11.120833333333332
2004-05-23	1.3	68.47826086956522	6.433333333333334
2004-05-24	1.1318181818181818	76.43478260869566	5.945833333333333
2004-05-25	1.6416666666666666	88.14285714285714	10.831578947368422
2004-05-26	1.9875	118.71428571428571	16.64
2004-05-27			12.804166666666667
2004-05-28	1.6888888888888889	108.88888888888889	10.704166666666666
2004-05-29	0.8454545454545456	78.73913043478261	5.495833333333334
2004-05-30	0.6565217391304348	74.78260869565217	5.791666666666667
2004-05-31	1.52	93.08695652173913	11.020833333333334
2004-06-01	2.1	75.8695652173913	12.375
2004-06-02	0.8	49.130434782608695	7.2625
2004-06-03		66.3	10.966666666666667
2004-06-04	2.8625		9.487499999999999
2004-06-05	1.45	76.9	8.012500000000001
2004-06-06	1.0	62.91304347826087	4.920833333333333
2004-06-07	2.226086956521739	101.95652173913044	12.479166666666666
2004-06-08	2.4583333333333335	113.5909090909091	14.458333333333334
2004-06-09	2.466666666666667	127.1304347826087	14.9125
2004-06-10	2.660869565217391	125.26086956521739	14.825000000000001
2004-06-11	2.1869565217391305	113.61904761904762	13.041666666666666
2004-06-12	1.316666666666667	79.30434782608695	8.295833333333333
2004-06-13	0.7291666666666666	55.30434782608695	2.9749999999999996
2004-06-14	1.2130434782608694	67.3913043478261	4.966666666666667

2004-06-15	2.154166666666667	104.04347826086956	11.987499999999999
2004-06-16	2.158333333333333	98.08695652173913	11.741666666666667
2004-06-17	2.243478260869565	104.04347826086956	11.9125

19_outlier_statistics.csv

File size: 505.0 B — Rows included: 6

Feature	Q1	Q3	IQR	Lower_Bound	Upper_Bound	Outlier_Count	Outlier_Percentage
CO(GT)	1.1	2.9	1.7999999999999998	-1.5999999999999996	5.6	215	2.8016679697680478
C6H6(GT)	4.4	14.0	9.6	-9.999999999999998	28.4	228	2.5358692025358693
NOx(GT)	98.0	326.0	228.0	-244.0	668.0	435	5.636175174915781
NO2(GT)	78.0	142.0	64.0	-18.0	238.0	107	1.3869086195722617
T	11.8	24.4	12.599999999999998	-7.099999999999998	43.3	3	0.033366700033366704

RH	35.8	62.5	26.700000000000 0003	- 4.25000000000000 07	102.55000000000 0001	0	0.0
----	------	------	-------------------------	-----------------------------	-------------------------	---	-----

02_Preprocessing

Folder: data\results\02_Preprocessing

01_missing_values_summary.csv

File size: 427.0 B — Rows included: 12

Column	Missing_Count	Missing_Percentage
CO(GT)	1797	18.973709217611656
NO2(GT)	1756	18.540808784711224
NOx(GT)	1753	18.509133143279485
PT08.S1(CO)	480	5.068102629078239
PT08.S2(NMHC)	480	5.068102629078239
C6H6(GT)	480	5.068102629078239
PT08.S3(NOx)	480	5.068102629078239
PT08.S4(NO2)	480	5.068102629078239
PT08.S5(O3)	480	5.068102629078239
T	480	5.068102629078239
RH	480	5.068102629078239
AH	480	5.068102629078239

02_strategy_statistics.csv

File size: 339.0 B — Rows included: 5

Strategy	Remaining_NaN	Mean	Median	Std
Original (with NaN)	1797	2.1527495439145166	1.8	1.4532520363373485
Interpolation	0	2.1314380741210006	1.8	1.423112520064881
Forward Fill	0	2.083613134832647	1.7	1.4609842033729825
Mean	0	2.1527495439145166	2.1527495439145166	1.3081229038737985
Median	0	2.085819871185725	1.8	1.3154152888432298

03_outlier_detection_comparison.csv

File size: 412.0 B — Rows included: 6

Feature	IQR_Outliers	IQR_Percentage	ZScore_Outliers	ZScore_Percentage
CO(GT)	224	2.3651145602365116	117	1.2353500158378208
PT08.S1(CO)	106	1.1192059972547777	60	0.6335128286347799
C6H6(GT)	251	2.6501953331221624	123	1.2987012987012987
PT08.S2(NMHC)	72	0.7602153943617358	38	0.4012247914686939
NOx(GT)	456	4.814697497624327	174	1.8371872030408618
PT08.S3(NOx)	247	2.6079611445465103	112	1.1825572801182558

04_temporal_split_info.csv

File size: 262.0 B — Rows included: 3

Split	Samples	Percentage	Start_Date	End_Date
Train	6629	69.99260901699 927	2004-03-10 18:00:00	2004-12-11 22:00:00
Validation	1421	15.00369549150 0368	2004-12-11 23:00:00	2005-02-09 03:00:00
Test	1421	15.00369549150 0368	2005-02-09 04:00:00	2005-04-04 14:00:00

06_normalized_columns_info.csv

File size: 382.0 B — Rows included: 12

Column	Normalized	Scaler
CO(GT)	True	MinMax [0,1]
PT08.S1(CO)	True	MinMax [0,1]
C6H6(GT)	True	MinMax [0,1]
PT08.S2(NMHC)	True	MinMax [0,1]
NOx(GT)	True	MinMax [0,1]
PT08.S3(NOx)	True	MinMax [0,1]
NO2(GT)	True	MinMax [0,1]
PT08.S4(NO2)	True	MinMax [0,1]
PT08.S5(O3)	True	MinMax [0,1]
T	True	MinMax [0,1]
RH	True	MinMax [0,1]
AH	True	MinMax [0,1]

07_lag_features_sample.csv

File size: 295.0 B — Rows included: 5

CO(GT)	CO(GT)_lag1	CO(GT)_lag2	CO(GT)_lag24
0.4545454545454546			
0.34545454545454546	0.4545454545454546		
0.38181818181818183	0.34545454545454546	0.4545454545454546	
0.38181818181818183	0.38181818181818183	0.34545454545454546	
0.27272727272727276	0.38181818181818183	0.38181818181818183	

08_time_features_sample.csv

File size: 583.0 B — Rows included: 5

DateTime	hour	day	month	day_of_week	is_week_end	season	hour_sin	hour_cos	month_sin	month_cos
2004-03-10 18:00:00	18	10	3	2	0	1	-1.0	-1.8369701987210297e-16	1.0	6.123233995736766e-17
2004-03-10 19:00:00	19	10	3	2	0	1	-0.96592	0.2588190451025203	1.0	6.123233995736766e-17

							5826289 0684			
2004- 03-10 20:00:00	20	10	3	2	0	1	- 0.86602 5403784 4386	0.50000 0000000 0001	1.0	6.12323 3995736 766e-17
2004- 03-10 21:00:00	21	10	3	2	0	1	- 0.70710 6781186 5477	0.70710 6781186 5474	1.0	6.12323 3995736 766e-17
2004- 03-10 22:00:00	22	10	3	2	0	1	- 0.50000 0000000 0004	0.86602 5403784 4384	1.0	6.12323 3995736 766e-17

09_feature_catalog_summary.csv

File size: 339.0 B — Rows included: 7

Feature_Type	Count	Sample_Features
original	5	CO(GT), NO2(GT), C6H6(GT)...
lag	15	CO(GT)_lag1, CO(GT)_lag2, CO(GT)_lag24...
rolling	60	CO(GT)_rolling3_mean, CO(GT)_rolling3_std, CO(GT)_rolling3_min...
time	12	hour, day, month...

interaction	4	CO(GT)_ratio_NO2(GT), CO(GT)_diff_NO2(GT), T_ratio_RH...
derivative	0	
ema	0	

03_Modeling

Folder: data\results\03_Modeling

gb_feature_importance.csv

File size: 3.9 KB — Rows included: 100 — Original rows: 103

Feature	Importance
NOx(GT)	0.46421375094965817
PT08.S2(NMHC)	0.3048589823304947
C6H6(GT)	0.13782712028086125
PT08.S1(CO)	0.01831730145928657
PT08.S4(NO2)	0.018281629839095516
NMHC(GT)	0.01570876676500593
NO2(GT)	0.008295499404713714
CO(GT)_ratio_NO2(GT)	0.0025758377586299944
CO(GT)_lag1	0.0025313734403328908
hour_sin	0.0017116401010547964
T_rolling24_mean	0.0016000801208625108
CO(GT)_diff_NO2(GT)	0.0013107518206403623
PT08.S5(O3)	0.0010764557949895278
hour	0.0010324159221506326
T_rolling24_std	0.0010308332342585853
T_rolling24_max	0.0008981661652583503
RH_rolling24_min	0.0007925416597119607
T_rolling6_std	0.0006734443016355091
C6H6(GT)_rolling6_std	0.0006370291960225361
hour_cos	0.0005644740842586776
C6H6(GT)_rolling3_std	0.0005630985335982857
RH_lag1	0.0005330547649789774
CO(GT)_lag24	0.0005258255738273546
RH_rolling3_min	0.00047382116202465696
AH	0.0004735318709944148

NO2(GT)_lag1	0.00047022211312891234
NO2(GT)_rolling6_std	0.000453883012668425
T_lag24	0.00044330290903956554
RH_lag2	0.0004364436497893185
day_of_week	0.00040336113215814134
RH_rolling24_std	0.000398637185217898
NO2(GT)_lag24	0.00038154885483082245
C6H6(GT)_rolling6_mean	0.0003717725888110364
RH_rolling24_mean	0.0003668571699644863
CO(GT)_rolling24_std	0.000364633986735111
RH	0.0003525879357852986
RH_rolling3_mean	0.000350167244488196
T	0.0003463205295599305
C6H6(GT)_lag1	0.000319151803685051
PT08.S3(NOx)	0.0003170334535369338
RH_lag24	0.00030511335703838386
C6H6(GT)_rolling24_std	0.0003000021183209341
CO(GT)_rolling6_std	0.00028060036818909905
RH_rolling6_mean	0.00025748850123274763
T_rolling3_std	0.00025643928851106033
C6H6(GT)_lag24	0.00025513209748907994
C6H6(GT)_rolling24_mean	0.00025506637846730364
NO2(GT)_rolling24_std	0.00024974654925353704
NO2(GT)_rolling24_max	0.0002426221185019651
T_rolling24_min	0.00023536361654353087
CO(GT)_rolling24_mean	0.0002267616951820791
CO(GT)_rolling3_std	0.0002236618787542861
C6H6(GT)_rolling3_max	0.00021247721122528556
RH_rolling24_max	0.00020801638365568274
C6H6(GT)_rolling6_min	0.00019928974643672575
RH_rolling6_max	0.00019737754408223726
week_of_year	0.0001785095969809766
day	0.0001740910120870187

NO2(GT)_rolling6_min	0.00015693063057547753
CO(GT)_rolling3_max	0.0001566078788337212
RH_rolling3_std	0.000153208562062588
C6H6(GT)_rolling24_max	0.00015119671680862275
CO(GT)_rolling24_max	0.00014337030329117338
RH_rolling6_std	0.00013081260382649679
CO(GT)_rolling3_mean	0.0001222159866982599
RH_rolling6_min	0.00011964792917554408
NO2(GT)_rolling6_max	0.00011823511686459429
NO2(GT)_rolling3_min	0.00011702672264064872
T_rolling6_min	0.0001153696167621682
NO2(GT)_rolling6_mean	0.0001092560433590901
NO2(GT)_rolling24_mean	0.00010880979113495942
CO(GT)_rolling6_min	0.00010601308020083813
CO(GT)_rolling3_min	0.00010590230534648139
NO2(GT)_rolling24_min	0.00010074663557708016
T_lag1	9.878214776434037e-05
NO2(GT)_rolling3_std	9.38517969861485e-05
day_of_year	9.175025391199422e-05
C6H6(GT)_lag2	9.063023019052722e-05
NO2(GT)_rolling3_max	8.840452517071032e-05
NO2(GT)_lag2	7.73383790298323e-05
CO(GT)_rolling6_mean	7.712425805228481e-05
T_diff_RH	7.233804032561585e-05
C6H6(GT)_rolling24_min	7.056082495665194e-05
CO(GT)_lag2	6.410311178493723e-05
CO(GT)_rolling24_min	6.159352235642437e-05
C6H6(GT)_rolling3_min	6.137931894629929e-05
T_rolling3_min	5.35335337929839e-05
CO(GT)_rolling6_max	5.210604019028683e-05
C6H6(GT)_rolling6_max	5.2033671628489757e-05
C6H6(GT)_rolling3_mean	5.029493293887597e-05
T_lag2	4.8328520302385495e-05

T_ratio_RH	4.442058880295015e-05
T_rolling6_mean	4.296814547194889e-05
T_rolling3_mean	3.8972300863632665e-05
RH_rolling3_max	2.664561966933357e-05
T_rolling3_max	2.447865631433399e-05
NO2(GT)_rolling3_mean	2.2239717595114747e-05
T_rolling6_max	2.1080258489908323e-05
month	1.2969892917100367e-05
month_cos	5.131994961114428e-06

lr_coefficients.csv

File size: 3.8 KB — Rows included: 100 — Original rows: 103

Feature	Coefficient
RH_rolling3_std	-1.1981256674961236
T_rolling6_std	-1.1431349166449567
NO2(GT)_rolling3_std	-0.9431086092840518
T_rolling3_std	0.9055827382852704
RH_rolling3_max	0.7191047481597688
T_rolling3_min	0.7110635233900946
CO(GT)_rolling3_mean	0.5884316784774063
C6H6(GT)	0.5686486082163524
NOx(GT)	0.5548525766424491
RH_rolling3_min	-0.5375209004717031
NO2(GT)_rolling3_max	0.5090464637265969
RH_rolling6_std	0.5019527594579947
NO2(GT)_rolling3_min	-0.4383302450099278
T_rolling6_min	-0.38116472655486655
C6H6(GT)_rolling24_min	0.3559989237816576
T_rolling6_max	0.2948997110609363
C6H6(GT)_rolling3_std	-0.2932258777889739
CO(GT)_rolling24_std	-0.2893680880075101
C6H6(GT)_rolling3_mean	-0.27597105809830286

CO(GT)_rolling24_mean	0.25945536824725834
month_sin	-0.23758014646555972
NO2(GT)_rolling3_mean	-0.22718769279451592
T_rolling6_mean	-0.22125567803385188
RH_rolling6_max	-0.19341438733695882
PT08.S2(NMHC)	-0.18971364155082676
NO2(GT)_rolling24_std	0.18882494475637038
PT08.S1(CO)	0.1873162388966457
C6H6(GT)_rolling24_mean	-0.1844902802661089
CO(GT)_rolling24_min	-0.18391706186543821
C6H6(GT)_rolling24_std	0.17706754355367443
T	-0.17641054104778542
CO(GT)_lag2	-0.16725624155375918
T_rolling24_std	0.1604582759080556
CO(GT)_rolling3_std	0.15738643705997263
CO(GT)_rolling3_max	-0.1423609555534655
NO2(GT)_rolling24_mean	-0.12945168357313053
month_cos	0.11845625391852074
NO2(GT)_lag2	0.11200019457940187
NO2(GT)	0.11123965747251101
CO(GT)_rolling3_min	-0.10789337628780016
T_rolling3_mean	-0.1073280229595105
RH_rolling6_min	0.10655640895417029
PT08.S5(O3)	-0.10608417060051965
RH_lag2	-0.09703470057703904
T_rolling24_min	0.09624236290453067
C6H6(GT)_rolling6_max	0.0935429388648689
RH_rolling24_std	0.08525149467462084
C6H6(GT)_rolling3_max	0.08323114223866435
CO(GT)_rolling6_max	-0.0796820528139908
C6H6(GT)_lag2	0.07906860576588476
RH_rolling24_max	-0.07096450037599276
T_lag1	-0.06782329219575892

T_rolling24_mean	-0.06188607818356241
NO2(GT)_rolling24_max	0.0611241125868983
RH_rolling3_mean	0.060437352816813714
T_lag2	-0.05886269915252185
CO(GT)_rolling6_mean	0.057148614703472524
T_rolling3_max	-0.05666069327903177
C6H6(GT)_rolling6_mean	-0.05586321141368712
RH_lag1	-0.05492453719668382
T_rolling24_max	-0.05337712114445867
C6H6(GT)_rolling3_min	-0.051402788976872824
CO(GT)_lag1	0.047451205829180026
AH	-0.046265404934426824
CO(GT)_rolling24_max	-0.04476608074387134
C6H6(GT)_rolling6_min	0.04294301716137272
NO2(GT)_rolling6_max	-0.0389304862309945
NO2(GT)_lag1	0.036637656749506906
T_lag24	-0.035989383501492944
C6H6(GT)_rolling24_max	0.03590047619286942
RH_rolling24_min	0.03574712545372938
C6H6(GT)_lag1	-0.03333462179272714
month	0.03240789620442142
C6H6(GT)_rolling6_std	0.03118579784559951
PT08.S3(NOx)	0.030361141077992607
NO2(GT)_lag24	-0.026934352240874593
RH_rolling6_mean	0.026088862721344776
CO(GT)_rolling6_min	-0.024813509360710195
hour_sin	-0.02285155816314338
CO(GT)_ratio_NO2(GT)	0.022791179016613247
PT08.S4(NO2)	0.019275505088609007
NO2(GT)_rolling6_std	-0.016463348685648492
hour_cos	0.013116525506146198
T_diff_RH	-0.012898754998994122
NO2(GT)_rolling6_min	0.012792112503897689

NO2(GT)_rolling6_mean	-0.012722857863930456
NO2(GT)_rolling24_min	-0.012641040846423493
C6H6(GT)_lag24	0.011913230125740526
CO(GT)_diff_NO2(GT)	0.010813549079733375
CO(GT)_lag24	0.00940950603772596
RH	-0.00871532761999852
day_of_week	0.0032629965094226054
is_weekend	-0.0027471055541094525
T_ratio_RH	0.0014727840847020765
RH_lag24	-0.0009349462575088813
CO(GT)_rolling6_std	-0.0007846915293065054
RH_rolling24_mean	-0.0006093289152787063
week_of_year	-0.0004809250143872207
day	0.00036917047091200583
hour	-0.0002667163697324948

model_comparison.csv

File size: 1.2 KB — Rows included: 6

Model	Train_M AE	Val_MA E	Test_MA E	Train_R MSE	Val_RM SE	Test_RM SE	Train_R2	Val_R2	Test_R2
Linear Regression	0.021700 44894598 2866	0.430038 45420920 256	0.117751 03563150 273	0.029197 60867671 3782	0.495128 37977556 21	0.140724 60981131 596	0.985616 43930477 27	- 2.554644 23423256 77	0.608713 19440045 34
Random Forest	0.013331 25528299 3018	0.104801 16891449 008	0.092163 00615404 4	0.019335 46175179 6963	0.140546 45489541 892	0.112372 87055611 667	0.993692 15607543 54	0.713582 01029747 26	0.750495 69281763 15
XGBoost	0.015190 78030695 2253	0.143611 55444225 757	0.130109 71930670 914	0.020989 36012836 7107	0.174801 89037312 838	0.150864 79035145 106	0.992566 89545202 93	0.556949 98970455 84	0.550291 71947417 23
Gradient Boosting	0.004611 66969453 9169	0.143054 53164732 304	0.133846 38894887 832	0.005601 82146799 4012	0.179199 96940847 99	0.157117 64658909 81	0.999470 54339529 45	0.534374 92133429 13	0.512241 29622401 87
LSTM	0.106838 26163102 982	0.204384 69919384 428	0.187699 41432897 894	0.140573 47616123 51	0.268673 46123507 46	0.239359 72591748 933	0.664125 35576559 93	- 0.042449 29115596 1744	- 0.122246 83328676 966

GRU	0.106034	0.262476	0.187588	0.143757	0.352240	0.250434	0.648737	-	-
	31797719	09496869	45308688	54675039	2306391	81976284	52556496	0.791774	0.228501
	93	934	543	853		016	86	33488743	39648428
								4	214

rf_feature_importance.csv

File size: 3.8 KB — Rows included: 100 — Original rows: 103

Feature	Importance
PT08.S2(NMHC)	0.11725769528701496
NOx(GT)	0.11713560751248589
C6H6(GT)	0.09911309538560602
NMHC(GT)	0.08949883249126411
PT08.S5(O3)	0.069387400067845
PT08.S4(NO2)	0.06913623550101444
PT08.S3(NOx)	0.0645240697544073
NO2(GT)	0.052018239234302835
PT08.S1(CO)	0.04489390643921268
CO(GT)_lag1	0.03755651840612919
C6H6(GT)_lag1	0.030803390953564488
CO(GT)_ratio_NO2(GT)	0.0268922085098007
CO(GT)_diff_NO2(GT)	0.016243281879387853
C6H6(GT)_rolling3_max	0.0139567282513452
NO2(GT)_lag1	0.013217322681503323
C6H6(GT)_lag24	0.011669910238384186
CO(GT)_rolling3_max	0.011183422040426043
CO(GT)_rolling3_mean	0.009243881798732268
hour	0.008884533325442328
C6H6(GT)_rolling3_mean	0.00884171216344047
C6H6(GT)_lag2	0.005834077538291123
NO2(GT)_lag24	0.004492938185780535
CO(GT)_rolling6_max	0.004479625133797696
CO(GT)_lag24	0.0037852804537610992
NO2(GT)_rolling3_max	0.002942919034638208
CO(GT)_rolling3_min	0.0028855139581843423
hour_cos	0.002883055765668662
hour_sin	0.0028472651313036557
C6H6(GT)_rolling3_min	0.0028312623235173715

C6H6(GT)_rolling6_min	0.002729804424375535
T_rolling6_std	0.0018819414585904738
CO(GT)_rolling3_std	0.001758787978226114
C6H6(GT)_rolling6_max	0.0017159350176126644
NO2(GT)_rolling3_min	0.0012157026148383995
C6H6(GT)_rolling3_std	0.0011569452013923066
day_of_year	0.0010706977952979998
NO2(GT)_rolling3_mean	0.0010258680140943857
T_rolling24_mean	0.0009826516509294579
CO(GT)_rolling6_min	0.000981432144893163
C6H6(GT)_rolling6_std	0.0009697790962617224
NO2(GT)_lag2	0.0009678917869229042
C6H6(GT)_rolling6_mean	0.0009563834367008565
CO(GT)_rolling6_mean	0.0009401771433124869
NO2(GT)_rolling6_min	0.0009190324465036683
CO(GT)_lag2	0.0009156297969812562
NO2(GT)_rolling24_mean	0.0008949374676937153
T_rolling6_min	0.0008893100677712355
T_rolling3_std	0.0008787384508502871
T_rolling24_std	0.000870988434740365
C6H6(GT)_rolling24_std	0.0008645938608978149
RH_rolling6_max	0.0008613441900950728
C6H6(GT)_rolling24_mean	0.0008538184923020758
T_diff_RH	0.0008523209086865168
NO2(GT)_rolling6_mean	0.0008458208548101603
T	0.0008362097855559355
NO2(GT)_rolling24_max	0.0008277179422640554
RH_rolling6_mean	0.0008016140376159869
RH_rolling3_max	0.0007853159973639076
RH_rolling6_std	0.0007674913538500815
T_ratio_RH	0.0007649081100467417
T_rolling6_mean	0.0007468530508429587
T_rolling6_max	0.0007438731908882265

RH_rolling6_min	0.0007419033988935329
RH_rolling3_min	0.0007344834168198349
RH_lag1	0.0007194242499944688
NO2(GT)_rolling3_std	0.0006955635515721518
NO2(GT)_rolling24_std	0.0006890847935914753
CO(GT)_rolling24_mean	0.000682690240310712
T_rolling3_mean	0.0006815424468463767
T_lag24	0.0006805472520844003
day_of_week	0.0006628175913493586
T_lag1	0.0006563973324313318
T_rolling24_max	0.0006507566174549092
NO2(GT)_rolling6_std	0.0006492268843170905
RH_rolling24_min	0.0006457425546829834
AH	0.000633378950359904
week_of_year	0.0006238011187647823
CO(GT)_rolling6_std	0.0006142919140331301
T_rolling3_min	0.000603607762880401
RH_lag24	0.0005993099916397815
T_rolling24_min	0.0005985352472805376
RH_lag2	0.0005931008508813249
RH	0.0005894036855893655
RH_rolling3_mean	0.0005862712399110784
day	0.0005748798793088455
T_rolling3_max	0.0005670503319688458
RH_rolling24_std	0.0005510808149205804
C6H6(GT)_rolling24_min	0.0005448391889258035
RH_rolling3_std	0.0005258986994283133
CO(GT)_rolling24_std	0.0005257218159831772
RH_rolling24_mean	0.00045613134554768007
NO2(GT)_rolling6_max	0.0004516438602134136
RH_rolling24_max	0.00043362490898357897
T_lag2	0.00042544064912495465
CO(GT)_rolling24_min	0.0004098475848060793

NO2(GT)_rolling24_min	0.00040104625120404016
C6H6(GT)_rolling24_max	0.0003612295930418761
CO(GT)_rolling24_max	0.0002784112639748964
is_weekend	0.00015933383953582413
month_sin	0.00014724941959469006

xgb_feature_importance.csv

File size: 3.0 KB — Rows included: 100 — Original rows: 103

Feature	Importance
PT08.S2(NMHC)	0.33739394
NOx(GT)	0.31712323
C6H6(GT)	0.14942408
PT08.S4(NO2)	0.03563519
NO2(GT)	0.008929834
T_rolling24_max	0.007204508
CO(GT)_lag1	0.0069484683
NMHC(GT)	0.006803655
PT08.S1(CO)	0.0054724854
NO2(GT)_rolling24_max	0.0052625467
hour_sin	0.00509464
RH_rolling24_min	0.0048125032
T_rolling24_mean	0.004577736
T_ratio_RH	0.004436967
month	0.004204772
CO(GT)_ratio_NO2(GT)	0.003599394
NO2(GT)_rolling24_min	0.0031856806
C6H6(GT)_rolling6_max	0.0027538824
T_rolling6_std	0.0025022295
T_rolling24_min	0.0024464752
hour_cos	0.0023007258
hour	0.0022501103
RH_lag1	0.0022194833

C6H6(GT)_rolling6_min	0.0021393336
RH_rolling3_mean	0.0019040828
CO(GT)_rolling6_min	0.0018659945
C6H6(GT)_rolling24_max	0.0018451706
RH_rolling24_max	0.0017901104
NO2(GT)_lag1	0.0017530025
CO(GT)_rolling24_max	0.0016454352
NO2(GT)_rolling3_min	0.0016444036
PT08.S5(O3)	0.0016218098
C6H6(GT)_rolling6_std	0.001590973
CO(GT)_diff_NO2(GT)	0.0015789847
CO(GT)_rolling3_min	0.0015511926
T	0.0015502116
CO(GT)_lag24	0.0014957777
day_of_year	0.0014905252
T_lag1	0.0014242255
C6H6(GT)_rolling24_std	0.0014238461
RH_rolling24_std	0.0013762272
T_rolling6_mean	0.0013443318
T_lag24	0.0012760425
C6H6(GT)_lag24	0.001260463
C6H6(GT)_rolling3_std	0.001239692
T_diff_RH	0.0012042277
C6H6(GT)_rolling24_mean	0.0011566881
C6H6(GT)_rolling6_mean	0.0011519585
CO(GT)_rolling6_std	0.0011098776
RH_lag24	0.0010750397
C6H6(GT)_lag1	0.0010737946
NO2(GT)_rolling6_max	0.0010520751
CO(GT)_lag2	0.0010270023
T_rolling3_min	0.0010211428
T_rolling6_max	0.0010138472
day_of_week	0.001001061

C6H6(GT)_rolling24_min	0.0009941094
T_rolling24_std	0.0009688011
CO(GT)_rolling6_mean	0.00092425075
CO(GT)_rolling24_mean	0.00092300563
NO2(GT)_lag24	0.0009207626
C6H6(GT)_rolling3_max	0.00090142287
NO2(GT)_rolling6_mean	0.0008659525
PT08.S3(NOx)	0.00085483247
RH_rolling3_min	0.0008452881
CO(GT)_rolling3_max	0.00082099676
NO2(GT)_rolling24_std	0.0008096413
RH_rolling6_max	0.0008069375
NO2(GT)_rolling6_std	0.00080191944
RH_rolling3_max	0.00078348146
CO(GT)_rolling24_std	0.00074639206
NO2(GT)_rolling6_min	0.0007461755
RH	0.00074332446
CO(GT)_rolling3_std	0.00074004324
day	0.00073851837
NO2(GT)_rolling3_mean	0.00072874915
CO(GT)_rolling3_mean	0.00072534924
CO(GT)_rolling24_min	0.00072394795
C6H6(GT)_rolling3_min	0.0007120339
T_rolling3_std	0.00069437636
RH_rolling3_std	0.00069027254
C6H6(GT)_rolling3_mean	0.00067152997
RH_rolling24_mean	0.0006690449
RH_rolling6_mean	0.00065531954
T_rolling6_min	0.0006458336
RH_rolling6_std	0.00063977076
NO2(GT)_rolling24_mean	0.00063157885
T_lag2	0.0006060274
C6H6(GT)_lag2	0.00053586747

RH_lag2	0.0005326349
AH	0.00045911229
NO2(GT)_rolling3_std	0.00045878766
CO(GT)_rolling6_max	0.0004258587
NO2(GT)_rolling3_max	0.00036574536
T_rolling3_max	0.00034136113
RH_rolling6_min	0.00033909373
NO2(GT)_lag2	0.0002727572
T_rolling3_mean	0.000257945
week_of_year	0.0
season	0.0

04_Anomaly

Folder: data\results\04_Anomaly

anomaly_detection_comparison.csv

File size: 334.0 B — Rows included: 9

Method	Anomalies_Detected	Percentage
Z-Score	1192	83.88458831808586
IQR	1337	94.08866995073892
Modified Z-Score	1397	98.31104855735397
Isolation Forest	801	56.36875439831105
LOF	1421	100.0
One-Class SVM	1421	100.0
Residual-Based	1197	84.23645320197045
STL Decomposition	37	2.6038001407459537
Ensemble	1211	85.22167487684729

cv_results.csv

File size: 603.0 B — Rows included: 5

MAE	RMSE	R2	MAPE	MBE	fold	train_size	val_size
0.03684526	0.04733655	0.94266169	15.3111528	0.02153673	1	150	148
5317708514	569961692	34460604	58014527	8291567926			

0.02741401 3600377204	0.03789054 5116415146	0.96988252 68223732	9.03682513 766896	- 0.00211436 1387656784	2	298	148
0.04456492 636236955	0.07057065 507033816	0.87673939 91572456	15.6727974 18760878	- 0.00439033 3496867546	3	446	148
0.04018146 450532814	0.05557279 7770988676	0.96092046 26258745	13.7648170 41917016	- 0.00070859 4473367169 4	4	594	148
0.03368501 747876744	0.04499601 324656182	0.96910047 70433038	11.8175632 18977046	0.00966354 2388542395	5	742	148

05_EDA_Power

Folder: data\results\05_EDA_Power

01_first_rows.csv

File size: 695.0 B — Rows included: 10

Date	Time	Global_active_power	Global_reactive_power	Voltage	Global_intensity	Sub_metering_1	Sub_metering_2	Sub_metering_3
16/12/2006	17:24:00	4.216	0.418	234.84	18.4	0.0	1.0	17.0
16/12/2006	17:25:00	5.36	0.436	233.63	23.0	0.0	1.0	16.0

16/12/2006	17:26:00	5.374	0.498	233.29	23.0	0.0	2.0	17.0
16/12/2006	17:27:00	5.388	0.502	233.74	23.0	0.0	1.0	17.0
16/12/2006	17:28:00	3.666	0.528	235.68	15.8	0.0	1.0	17.0
16/12/2006	17:29:00	3.52	0.522	235.02	15.0	0.0	2.0	17.0
16/12/2006	17:30:00	3.702	0.52	235.09	15.8	0.0	1.0	17.0
16/12/2006	17:31:00	3.7	0.52	235.22	15.8	0.0	1.0	17.0
16/12/2006	17:32:00	3.668	0.51	233.99	15.8	0.0	1.0	17.0
16/12/2006	17:33:00	3.662	0.51	233.86	15.8	0.0	2.0	16.0

02_last_rows.csv

File size: 661.0 B — Rows included: 10

Date	Time	Global_active_power	Global_reactive_power	Voltage	Global_intensity	Sub_metering_1	Sub_metering_2	Sub_metering_3
26/11/2010	20:53:00	0.944	0.0	238.72	4.0	0.0	0.0	0.0
26/11/2010	20:54:00	0.946	0.0	239.31	4.0	0.0	0.0	0.0
26/11/2010	20:55:00	0.946	0.0	239.74	4.0	0.0	0.0	0.0
26/11/2010	20:56:00	0.942	0.0	239.41	4.0	0.0	0.0	0.0
26/11/2010	20:57:00	0.946	0.0	240.33	4.0	0.0	0.0	0.0
26/11/2010	20:58:00	0.946	0.0	240.43	4.0	0.0	0.0	0.0
26/11/2010	20:59:00	0.944	0.0	240.0	4.0	0.0	0.0	0.0
26/11/2010	21:00:00	0.938	0.0	239.82	3.8	0.0	0.0	0.0

26/11/2010	21:01:00	0.934	0.0	239.7	3.8	0.0	0.0	0.0
26/11/2010	21:02:00	0.932	0.0	239.55	3.8	0.0	0.0	0.0

05_data_types.csv

File size: 218.0 B — Rows included: 9

Column	DataType
Date	object
Time	object
Global_active_power	float64
Global_reactive_power	float64
Voltage	float64
Global_intensity	float64
Sub_metering_1	float64
Sub_metering_2	float64
Sub_metering_3	float64

06_missing_values.csv

File size: 359.0 B — Rows included: 9

Column	Missing_Count	Missing_Percentage
Global_active_power	25979	1.2518437457686005
Sub_metering_2	25979	1.2518437457686005

Global_reactive_power	25979	1.2518437457686005
Voltage	25979	1.2518437457686005
Global_intensity	25979	1.2518437457686005
Sub_metering_3	25979	1.2518437457686005
Sub_metering_1	25979	1.2518437457686005
Date	0	0.0
Time	0	0.0

08_statistical_summary.csv

File size: 913.0 B — Rows included: 12

count	mean	std	min	25%	50%	75%	max
2049280.0	1.09161503 65006245	1.05729416 10939828	0.076	0.308	0.602	1.528	11.122
2049280.0	0.12371447 630387257	0.11272197 955071626	0.0	0.048	0.1	0.194	1.39
2049280.0	240.839857 9745079	3.23998667 90095677	223.2	238.99	241.01	242.89	254.15
2049280.0	4.62775931 0587133	4.44439625 97861635	0.2	1.4	2.6	6.4	48.4
2049280.0	1.12192330 96502186	6.15303108 9701348	0.0	0.0	0.0	0.0	88.0
2049280.0	1.29851996 79887571	5.82202647 3177505	0.0	0.0	0.0	1.0	80.0

2049280.0	6.45844735 712055	8.43715390 8665472	0.0	0.0	1.0	17.0	31.0
2075259.0	2008.43659 80342695	1.12867678 90933729	2006.0	2007.0	2008.0	2009.0	2010.0
2075259.0	6.45035872 6308379	3.41576152 3823248	1.0	3.0	6.0	9.0	12.0
2075259.0	15.7736330 74233144	8.78777342 2208634	1.0	8.0	16.0	23.0	31.0
2075259.0	11.5007596 64215407	6.92222684 79296005	0.0	6.0	12.0	18.0	23.0
2075259.0	2.99890856 99664475	2.00006415 09839454	0.0	1.0	3.0	5.0	6.0

11_temporal_features_sample.csv

File size: 968.0 B — Rows included: 20

DateTime	Year	Month	Day	Hour	DayOfWeek	DayName
2006-12-16 17:24:00	2006	12	16	17	5	Saturday
2006-12-16 17:25:00	2006	12	16	17	5	Saturday
2006-12-16 17:26:00	2006	12	16	17	5	Saturday

2006-12-16 17:27:00	2006	12	16	17	5	Saturday
2006-12-16 17:28:00	2006	12	16	17	5	Saturday
2006-12-16 17:29:00	2006	12	16	17	5	Saturday
2006-12-16 17:30:00	2006	12	16	17	5	Saturday
2006-12-16 17:31:00	2006	12	16	17	5	Saturday
2006-12-16 17:32:00	2006	12	16	17	5	Saturday
2006-12-16 17:33:00	2006	12	16	17	5	Saturday
2006-12-16 17:34:00	2006	12	16	17	5	Saturday
2006-12-16 17:35:00	2006	12	16	17	5	Saturday
2006-12-16 17:36:00	2006	12	16	17	5	Saturday
2006-12-16 17:37:00	2006	12	16	17	5	Saturday
2006-12-16 17:38:00	2006	12	16	17	5	Saturday

2006-12-16 17:39:00	2006	12	16	17	5	Saturday
2006-12-16 17:40:00	2006	12	16	17	5	Saturday
2006-12-16 17:41:00	2006	12	16	17	5	Saturday
2006-12-16 17:42:00	2006	12	16	17	5	Saturday
2006-12-16 17:43:00	2006	12	16	17	5	Saturday

12_monthly_distribution.csv

File size: 136.0 B — Rows included: 12

Month	Count
1	178560
2	162720
3	178560
4	172800
5	178560
6	172800
7	178560
8	178560
9	172800

10	178560
11	166863
12	155916

13_hourly_distribution.csv

File size: 242.0 B — Rows included: 24

Hour	Count
0	86460
1	86460
2	86460
3	86460
4	86460
5	86460
6	86460
7	86460
8	86460
9	86460
10	86460
11	86460
12	86460
13	86460
14	86460

15	86460
16	86460
17	86496
18	86520
19	86520
20	86520
21	86463
22	86460
23	86460

14_missing_values_percentage.csv

File size: 265.0 B — Rows included: 7

Unnamed: 0	Percentage
Global_active_power	1.2518437457686005
Global_reactive_power	1.2518437457686005
Voltage	1.2518437457686005
Global_intensity	1.2518437457686005
Sub_metering_1	1.2518437457686005
Sub_metering_2	1.2518437457686005
Sub_metering_3	1.2518437457686005

15_distribution_statistics.csv

File size: 657.0 B — Rows included: 7

Feature	Mean	Median	Std	Min	Max	Skewness
Global_active_power	1.0916150365006243	0.602	1.057294161093983	0.076	11.122	1.7862333920915965
Global_reactive_power	0.12371447630387251	0.1	0.11272197955071628	0.0	1.39	1.2619141974704149
Voltage	240.83985797450816	241.01	3.239986679009568	223.2	254.15	-0.3266649756878288
Global_intensity	4.627759310587131	2.6	4.444396259786164	0.2	48.4	1.8490999040716398
Sub_metering_1	1.1219233096502186	0.0	6.153031089701347	0.0	88.0	5.944541281563981
Sub_metering_2	1.2985199679887571	0.0	5.822026473177503	0.0	80.0	7.0905532303433985
Sub_metering_3	6.45844735712055	1.0	8.437153908665472	0.0	31.0	0.7246883225039495

16_correlation_matrix.csv

File size: 984.0 B — Rows included: 7

Global_active_ power	Global_reactiv e_power	Voltage	Global_intensi ty	Sub_metering _1	Sub_metering _2	Sub_metering _3
1.0	0.2470170526 4407892	- 0.3997616096 2904805	0.9988886002 096058	0.4844012751 7743936	0.4345687175 9851884	0.6385554235 386769
0.2470170526 4407892	1.0	- 0.1122455747 8069051	0.2661203880 9001354	0.1231105660 9200884	0.1392308934 3347975	0.0896165262 2819384
- 0.3997616096 2904805	- 0.1122455747 8069051	1.0	- 0.4113630717 685292	- 0.1959755510 8843884	- 0.1674047640 589574	- 0.2681720807 432694
0.9988886002 096058	0.2661203880 9001354	- 0.4113630717 685292	1.0	0.4892982161 8807545	0.4403465386 728789	0.6265427458 734342
0.4844012751 7743936	0.1231105660 9200884	- 0.1959755510 8843884	0.4892982161 8807545	1.0	0.0547208643 2348112	0.1025710729 056424
0.4345687175 9851884	0.1392308934 3347975	- 0.1674047640 589574	0.4403465386 728789	0.0547208643 2348112	1.0	0.0808720456 6633059
0.6385554235 386769	0.0896165262 2819384	- 0.2681720807 432694	0.6265427458 734342	0.1025710729 056424	0.0808720456 6633059	1.0

17_high_correlations.csv

File size: 88.0 B — Rows included: 1

Feature1	Feature2	Correlation
Global_active_power	Global_intensity	0.9988886002096058

18_daily_averaged_timeseries.csv

File size: 92.3 KB — Rows included: 100 — Original rows: 1442

DateTime	Global_active_power	Voltage	Global_intensity
2006-12-16	3.05347474747475	236.243762626262	13.0828282828283
2006-12-17	2.35448611111111	240.087027777778	9.99902777777778
2006-12-18	1.53043472222222	241.231694444444	6.42166666666667
2006-12-19	1.15707916666667	241.9993125	4.92638888888889
2006-12-20	1.54565833333333	242.308062499999	6.46736111111111
2006-12-21	1.19375799721835	241.040549374130	5.03379694019471
2006-12-22	1.62592916666667	241.184548611111	6.87291666666667
2006-12-23	3.31485138888889	240.136076388889	14.0280555555556
2006-12-24	1.77084166666667	241.6874375	7.64041666666667
2006-12-25	1.90494444444444	243.399930555556	7.95152777777778
2006-12-26	2.73202083333333	241.6254375	11.3479166666667

2006-12-27	1.0616388888888888	243.76736111111111	4.5175
2006-12-28	1.4393319444444443	243.26833333333335	6.086250000000001
2006-12-29	2.204438888888889	240.87130555555555	9.271388888888888
2006-12-30	1.939709318497914	240.54068845618914	8.292211404728791
2006-12-31	2.426525	239.8965763888889	10.199583333333333
2007-01-01	1.9090305555555556	240.12897916666665	7.916944444444444
2007-01-02	0.8814138888888889	241.94377777777777	3.714027777777775
2007-01-03	0.7042041666666666	243.5570902777778	3.014027777777778
2007-01-04	2.2634805555555553	239.7502986111111	9.478194444444444
2007-01-05	1.8842805555555555	240.49576388888892	7.885138888888889
2007-01-06	1.047484722222222	239.67305555555555	4.543888888888889
2007-01-07	1.6997361111111111	240.90248611111113	7.282638888888889
2007-01-08	1.5565	239.5769513888889	6.551666666666667
2007-01-09	1.2979541666666667	240.48752083333335	5.521111111111111
2007-01-10	1.496388888888889	241.02745138888892	6.357777777777779
2007-01-11	1.5661069444444444	241.1556875	6.658333333333333
2007-01-12	1.014788888888889	241.53820833333336	4.309166666666666
2007-01-13	2.2130652777777775	239.28361111111113	9.453055555555554
2007-01-14	2.0902126476719944	238.8218485059069	8.968311327310632
2007-01-15	1.4921375	241.2409097222222	6.261527777777778
2007-01-16	1.1711138888888888	241.94475694444446	4.975972222222222
2007-01-17	1.9775611111111111	241.00525	8.283888888888889
2007-01-18	1.2649041666666667	240.89359722222224	5.3425
2007-01-19	1.0280833333333335	240.16094444444445	4.37625

2007-01-20	2.1762027777777777	238.8674722222222	9.348055555555556
2007-01-21	2.366154166666667	238.96811805555558	10.190694444444444
2007-01-22	1.5142319444444445	241.60313194444444	6.355277777777778
2007-01-23	1.2344722222222222	242.76255555555556	5.260833333333333
2007-01-24	2.074898611111111	241.45769444444446	8.733611111111111
2007-01-25	1.1085722222222223	243.42436111111112	4.629305555555556
2007-01-26	1.1235916666666665	242.77179166666667	4.721388888888889
2007-01-27	1.4194944444444444	241.61656944444442	5.949861111111111
2007-01-28	2.147466296038916	240.30478804725504	9.133703961084086
2007-01-29	1.37685	240.9295972222222	5.788888888888889
2007-01-30	1.1859708333333334	241.84087499999998	5.028194444444445
2007-01-31	1.6414944444444444	239.92113194444443	6.923611111111111
2007-02-01	1.2671944444444445	240.3921388888889	5.3325000000000005
2007-02-02	1.15815	240.3343888888889	4.870833333333334
2007-02-03	2.7984180555555556	238.91019444444444	11.815972222222221
2007-02-04	2.4971805555555555	239.63333333333333	10.565416666666668
2007-02-05	1.1425555555555555	240.81031249999998	4.820416666666667
2007-02-06	0.8843611111111112	241.26415972222225	3.7719444444444448
2007-02-07	1.6166791666666667	241.14969444444444	6.791666666666667
2007-02-08	1.2509819444444443	241.79730555555554	5.248333333333334
2007-02-09	1.1251375000000001	241.74044444444445	4.727222222222222
2007-02-10	1.964811111111111	240.0368263888889	8.303749999999999
2007-02-11	2.4800194444444443	239.5061458333333	10.526805555555557
2007-02-12	1.3038958333333333	240.09850000000003	5.503472222222222

2007-02-13	0.9823236111111111	241.16685416666667	4.14375
2007-02-14	1.735188888888889	240.3753888888889	7.315138888888888
2007-02-15	1.3787125	240.6692222222223	5.820277777777778
2007-02-16	1.182311111111111	240.8209444444442	5.046527777777775
2007-02-17	1.4423	239.3160347222224	6.129166666666666
2007-02-18	2.6595569444444447	239.1708055555557	11.23888888888889
2007-02-19	1.5184819444444444	239.7895833333333	6.442222222222222
2007-02-20	2.18745	239.8654444444445	9.168472222222222
2007-02-21	1.4394958333333332	240.4609652777777	6.074444444444445
2007-02-22	2.3431696801112656	238.62353963838666	9.870792767732963
2007-02-23	0.912138888888889	240.6859652777777	3.825416666666667
2007-02-24	0.4964736111111111	239.86310416666666	2.0858333333333334
2007-02-25	0.3484305555555556	241.6815486111111	1.462222222222222
2007-02-26	0.3947805555555556	241.2735902777777	1.658888888888889
2007-02-27	0.3597999999999995	242.94427083333332	1.5245833333333334
2007-02-28	0.3616486111111113	242.1595902777777	1.528472222222222
2007-03-01	0.3594194444444444	241.3214027777778	1.508888888888889
2007-03-02	0.3581166666666664	240.5915972222225	1.506805555555557
2007-03-03	0.5688347222222222	241.2880833333336	2.401111111111111
2007-03-04	1.4518749999999998	240.0297986111111	6.158194444444444
2007-03-05	1.8474527777777776	240.0309097222222	7.840277777777778
2007-03-06	0.865625	240.9880694444445	3.695694444444444
2007-03-07	1.6494486111111113	240.3235625	6.935
2007-03-08	1.0917583333333334	240.6654236111112	4.665138888888889

2007-03-09	0.8837583333333334	240.88846527777778	3.73125
2007-03-10	1.7254944444444444	239.80283333333335	7.341944444444445
2007-03-11	2.4171083333333333	240.00729166666667	10.121527777777779
2007-03-12	1.3546305555555556	240.55871527777776	5.728472222222222
2007-03-13	0.8959333333333334	240.89796527777779	3.8056944444444443
2007-03-14	1.2453916666666667	240.77729166666666	5.2675
2007-03-15	1.3088986111111112	240.28049305555552	5.523194444444444
2007-03-16	1.1819819444444446	239.52713194444445	4.992361111111111
2007-03-17	1.4083958333333333	238.22465277777778	5.946111111111111
2007-03-18	1.6465597222222221	241.04781944444443	7.007916666666667
2007-03-19	1.4948041666666667	240.49709027777777	6.281944444444444
2007-03-20	1.4487694444444446	240.8156388888889	6.151249999999999
2007-03-21	1.487563888888889	240.98664583333334	6.237222222222223
2007-03-22	1.010401388888889	241.76352777777777	4.223055555555556
2007-03-23	0.9090333333333334	241.64547916666666	3.80375
2007-03-24	2.0614097222222223	239.70447222222222	8.703333333333333
2007-03-25	2.318362751911049	240.0327588603197	9.722307157748435

19_outlier_statistics.csv

File size: 563.0 B — Rows included: 6

Feature	Q1	Q3	IQR	Lower_Bound	Upper_Bound	Outlier_Count	Outlier_Percentage
---------	----	----	-----	-------------	-------------	---------------	--------------------

Global_active_power	0.308	1.528	1.22	-1.522	3.358	94907	4.63123633 6664585
Global_reactive_power	0.048	0.194	0.14600000 000000002	- 0.17100000 000000004	0.41300000 000000003	40420	1.97240006 2460962
Voltage	238.99	242.89	3.89999999 99999773	233.140000 00000004	248.739999 99999995	51067	2.49194839 1630231
Global_intensity	1.4	6.4	5.0	-6.1	13.9	100961	4.92665716 7395378
Sub_metering_1	0.0	0.0	0.0	0.0	0.0	169105	8.25192262 6483447
Sub_metering_2	0.0	1.0	1.0	-1.5	2.5	77151	3.76478568 08244846

06_Preprocessing_Power

Folder: data\results\06_Preprocessing_Power

01_missing_values_summary.csv

File size: 335.0 B — Rows included: 7

Column	Missing_Count	Missing_Percentage
Global_active_power	25979	1.2518437457686005
Global_reactive_power	25979	1.2518437457686005
Voltage	25979	1.2518437457686005
Global_intensity	25979	1.2518437457686005
Sub_metering_1	25979	1.2518437457686005
Sub_metering_2	25979	1.2518437457686005
Sub_metering_3	25979	1.2518437457686005

02_strategy_statistics.csv

File size: 336.0 B — Rows included: 5

Strategy	Remaining_NaN	Mean	Median	Std
Original (with NaN)	25979	1.0916150365006245	0.602	1.0572941610939828
Interpolation	0	1.0902795361928321	0.614	1.0526279978387232
Forward Fill	0	1.0861683519984733	0.598	1.0531721288462972
Mean	0	1.0916150365006245	0.63	1.0506554805707804
Median	0	1.085485821287849	0.602	1.0520647931246334

03_outlier_detection_comparison.csv

File size: 470.0 B — Rows included: 6

Feature	IQR_Outliers	IQR_Percentage	ZScore_Outliers	ZScore_Percentage
Global_active_power	95238	4.589210310616651	36671	1.7670565457130891
Global_reactive_power	40478	1.950503527511506	24528	1.1819247621622169
Voltage	52195	2.5151077528154318	12792	0.6164049884857745
Global_intensity	100979	4.865850479385946	37443	1.8042567216911238
Sub_metering_1	172247	8.300024237938493	56102	2.7033734102586715
Sub_metering_2	77166	3.718379248084215	50999	2.457476392103347

04_temporal_split_info.csv

File size: 269.0 B — Rows included: 3

Split	Samples	Percentage	Start_Date	End_Date
Train	1452681	69.99998554397307	2006-12-16 17:24:00	2009-09-20 12:44:00
Validation	311289	15.000007228013468	2009-09-20 12:45:00	2010-04-24 16:53:00
Test	311289	15.000007228013468	2010-04-24 16:54:00	2010-11-26 21:02:00

06_normalized_columns_info.csv

File size: 285.0 B — Rows included: 7

Column	Normalized	Scaler
--------	------------	--------

Global_active_power	True	MinMax [0,1]
Global_reactive_power	True	MinMax [0,1]
Voltage	True	MinMax [0,1]
Global_intensity	True	MinMax [0,1]
Sub_metering_1	True	MinMax [0,1]
Sub_metering_2	True	MinMax [0,1]
Sub_metering_3	True	MinMax [0,1]

07_lag_features_sample.csv

File size: 158.0 B — Rows included: 5

Global_active_power	Global_active_power_lag1	Global_active_power_lag2	Global_active_power_lag24
1.0			
1.0	1.0		
1.0	1.0	1.0	
1.0	1.0	1.0	
1.0	1.0	1.0	

08_time_features_sample.csv

File size: 618.0 B — Rows included: 5

DateTime	hour	day	month	day_of_week	is_week_end	season	hour_sin	hour_cos	month_sin	month_cos
e								s	in	os

2006- 12-16 17:24:00	17	16	12	5	1	0	- 0.96592 5826289 0683	- 0.25881 9045102 52063	- 2.44929 3598294 7064e- 16	1.0
2006- 12-16 17:25:00	17	16	12	5	1	0	- 0.96592 5826289 0683	- 0.25881 9045102 52063	- 2.44929 3598294 7064e- 16	1.0
2006- 12-16 17:26:00	17	16	12	5	1	0	- 0.96592 5826289 0683	- 0.25881 9045102 52063	- 2.44929 3598294 7064e- 16	1.0
2006- 12-16 17:27:00	17	16	12	5	1	0	- 0.96592 5826289 0683	- 0.25881 9045102 52063	- 2.44929 3598294 7064e- 16	1.0
2006- 12-16 17:28:00	17	16	12	5	1	0	- 0.96592 5826289 0683	- 0.25881 9045102 52063	- 2.44929 3598294 7064e- 16	1.0

									7064e-16	
--	--	--	--	--	--	--	--	--	----------	--

09_feature_catalog_summary.csv

File size: 1.2 KB — Rows included: 17

Feature_Name	Category	Type	Description	Created_From
Global_active_power	Original	Numeric	Power measurement: Global_active_power	Raw data
Global_reactive_power	Original	Numeric	Power measurement: Global_reactive_power	Raw data
Voltage	Original	Numeric	Power measurement: Voltage	Raw data
Global_intensity	Original	Numeric	Power measurement: Global_intensity	Raw data
Sub_metering_1	Original	Numeric	Power measurement: Sub_metering_1	Raw data
Sub_metering_2	Original	Numeric	Power measurement: Sub_metering_2	Raw data
Sub_metering_3	Original	Numeric	Power measurement: Sub_metering_3	Raw data
Year	Temporal	Numeric	Time-based feature: Year	DateTime

Month	Temporal	Numeric	Time-based feature: Month	DateTime
Day	Temporal	Numeric	Time-based feature: Day	DateTime
Hour	Temporal	Numeric	Time-based feature: Hour	DateTime
DayOfWeek	Temporal	Numeric	Time-based feature: DayOfWeek	DateTime
IsWeekend	Temporal	Numeric	Time-based feature: IsWeekend	DateTime
Hour_sin	Cyclical	Numeric	Cyclical encoding: Hour_sin	DateTime
Hour_cos	Cyclical	Numeric	Cyclical encoding: Hour_cos	DateTime
Month_sin	Cyclical	Numeric	Cyclical encoding: Month_sin	DateTime
Month_cos	Cyclical	Numeric	Cyclical encoding: Month_cos	DateTime

07_Modeling_Power

Folder: data\results\07_Modeling_Power

gb_feature_importance.csv

File size: 221.0 B — Rows included: 6

Feature	Importance
Global_intensity	0.9992917693768535
Global_reactive_power	0.00034836236753248
Voltage	0.00020315175256631948
Sub_metering_3	0.000146105135140229
Sub_metering_2	1.0611367907492586e-05
Sub_metering_1	0.0

lr_coefficients.csv

File size: 240.0 B — Rows included: 6

Feature	Coefficient
Global_intensity	1.0049864209995354
Global_reactive_power	-0.022157235505661687
Voltage	0.014231304451277474
Sub_metering_3	0.011757044513481966
Sub_metering_2	-0.0002563293610950187
Sub_metering_1	2.220446049250313e-16

model_comparison.csv

File size: 876.0 B — Rows included: 4

Model	Train_M AE	Val_MA E	Test_MA E	Train_R MSE	Val_RM SE	Test_RM SE	Train_R2	Val_R2	Test_R2
Linear Regressi on	0.007877 2015199 40046	0.007420 4251808 34004	0.007344 3197531 11601	0.012541 1548953 38026	0.011150 9919576 83778	0.012601 5844917 49071	0.998062 8561448 613	0.998372 80299624 82	0.997224 26449164 17
Random Forest	0.005002 8914833 54843	0.006545 5463988 703345	0.006790 8572451 70335	0.008962 4678496 21219	0.010455 3492657 54357	0.011718 5006140 71835	0.999010 6680580 009	0.998569 49222990 25	0.997599 66496523 28
XGBoost	0.005586 8807216 60593	0.006441 6342196 95397	0.006926 8888189 06924	0.010113 5536070 50579	0.010186 1499250 67115	0.011903 5463348 8447	0.998740 2209808 364	0.998642 20795840 89	0.997523 25951405 07
Gradient Boosting	0.005790 7394732 80348	0.006349 4142110 92067	0.006643 3156911 49591	0.010324 3798177 15501	0.009821 6413572 89662	0.011415 7898955 11211	0.998687 1510657 61	0.998737 64567056 66	0.997722 07351530 12

rf_feature_importance.csv

File size: 225.0 B — Rows included: 6

Feature	Importance
Global_intensity	0.9989364221529585
Global_reactive_power	0.0004895197582229703
Voltage	0.00036883034537644175
Sub_metering_3	0.00016893974148744802
Sub_metering_2	3.6288001954633675e-05
Sub_metering_1	0.0

xgb_feature_importance.csv

File size: 179.0 B — Rows included: 6

Feature	Importance
Global_intensity	0.9991603
Global_reactive_power	0.0003629188
Voltage	0.00021864487
Sub_metering_3	0.00018636425
Sub_metering_2	7.182335e-05
Sub_metering_1	0.0

08_Anomaly_Power

Folder: data\results\08_Anomaly_Power

anomaly_detection_comparison.csv

File size: 206.0 B — Rows included: 5

Method	Anomalies_Detected	Percentage
Z-Score	984	3.1611410948342327
IQR	2051	6.588923156001028
Isolation Forest	1557	5.001927525057826

Residual-Based	1557	5.001927525057826
Ensemble	1222	3.9257260344384477

cv_results.csv

File size: 330.0 B — Rows included: 5

fold	MAE	RMSE	R2
1	0.006250876137694496	0.010980256744469384	0.9985530128324956
2	0.005967166115767305	0.01051566812215363	0.9986240905129659
3	0.005897335105814594	0.010271969494714199	0.9987105806961352
4	0.005931969447861404	0.01023230084776778	0.9987403124379163
5	0.005923613742491841	0.010724837053780899	0.9985808574472833