2025-02

Customer Segmentation and Behavior Prediction in Energy Using Clustering and Predictive Analytics

þÿ “ º ± ½ ¯ ´ ¿ Å , ” ® ¼ · Ä Á ±

ΜΕΤΑΠΤΥΧΙΑΚΟ ΠΡΟΓΡΑΜΜΑ ΣΠΟΥΔΩΝ ΣΤΗΝ ΑΝΑΛΥΣΗ ΔΕΔΟΜΕΝΩΝ ΚΑΙ ΧΡΗΜΑΤΟΟΙΚΟΝΟΜΙΚΗ ΤΕΧΝΟΛΟΓΙΑ

Πανεπιστήμιο Νεάπολις Πάφος

Customer Segmentation and Behavior Prediction in Energy Using Clustering and Predictive Analytics

ΔΗΜΗΤΡΑ ΓΚΑΝΙΔΟΥ

ΓΕΩΡΓΙΟΣ ΜΑΣΤΟΡΑΚΗΣ

Φεβρουάριος, 2025

ΜΕΤΑΠΤΥΧΙΑΚΟ ΠΡΟΓΡΑΜΜΑ ΣΠΟΥΔΩΝ ΣΤΗΝ ΑΝΑΛΥΣΗ ΔΕΔΟΜΕΝΩΝ ΚΑΙ ΧΡΗΜΑΤΟΟΙΚΟΝΟΜΙΚΗ ΤΕΧΝΟΛΟΓΙΑ

Customer Segmentation and Behavior Prediction in Energy Using Clustering and Predictive Analytics

Διπλωματική Εργασία η οποία υποβλήθηκε προς απόκτηση εξ αποστάσεως μεταπτυχιακού τίτλου σπουδών στην Ανάλυση Δεδομένων και Χρηματοοικονομική Τεχνολογία στο Πανεπιστήμιο Νεάπολις Πάφος

ΔΗΜΗΤΡΑ ΓΚΑΝΙΔΟΥ

ΓΕΩΡΓΙΟΣ ΜΑΣΤΟΡΑΚΗΣ

Φεβρουάριος, 2025

Πνευματικά δικαιώματα

Σελίδα Εγκυρότητας

Ονοματεπώνυμο Φοιτητή: Δήμητρα Γκανίδου
Τίτλος Διπλωματικής Εργασίας: Customer Segmentation and Behavior Prediction in Energy Using Clustering and Predictive Analytics

Η παρούσα Διπλωματική Εργασία εκπονήθηκε στο πλαίσιο των σπουδών για την απόκτηση εξ αποστάσεως μεταπτυχιακού τίτλου στο Πανεπιστήμιο Νεάπολις και εγκρίθηκε στις ..... [ημερομηνία έγκρισης] από τα μέλη της Εξεταστικής Επιτροπής.

Εξεταστική Επιτροπή:
Πρώτος επιβλέπων (Πανεπιστήμιο Νεάπολις Πάφος)……[ονοματεπώνυμο, βαθμίδα]
Μέλος Εξεταστικής Επιτροπής: ……[ονοματεπώνυμο, βαθμίδα]
Μέλος Εξεταστικής Επιτροπής: ……[ονοματεπώνυμο, βαθμίδα]

Ή ΥΠΕΥΘΥΝΗ ΔΗΛΩΣΗ

Η Δήμητρα Γκανίδου, γνωρίζοντας τις συνέπειες της λογοκλοπής, δηλώνω υπεύθυνα ότι η παρούσα εργασία με τίτλο «Customer Segmentation and Behavior Prediction in Energy Using Clustering and Predictive Analytics», αποτελεί προϊόν αυστηρά προσωπικής εργασίας και όλες οι πηγές που έχω χρησιμοποιήσει, έχουν δηλωθεί κατάλληλα στις βιβλιογραφικές παραπομπές και αναφορές. Τα σημεία όπου έχω χρησιμοποιήσει ιδέες, κείμενο ή/και πηγές άλλων συγγραφέων, αναφέρονται ευδιάκριτα στο κείμενο με την κατάλληλη παραπομπή και η σχετική αναφορά περιλαμβάνεται στο τμήμα των βιβλιογραφικών αναφορών με πλήρη περιγραφή.

Ο/Η Δηλών /σα

Δήμητρα Γκανίδου

Πίνακας περιεχομένων

**Περίληψη……
……8**

**Abstract……
……..9**

1.Εισαγωγή

1.1 Αντικείμενο και
Σκοπός……..……10

1.2 Ερευνητικά Ερωτήματα και
Στόχοι……..……10

1.3 Δομή
Εργασίας……...……
……11

2. Θεωρητική Θεμελίωση

2.1 Σχετικές
Έρευνες…….……..
12

2.2 Ανάλυση πελατών και τμηματοποίηση (Customer
Segmentation)……12

2.3 Συσταδοποίηση (Clustering
Analysis)……..…12

2.4 Τεχνικές ομαδοποίησης
(Clustering)……13

2.4.1 Ανάλυση
RFM……..13

2.4.2 K-Means Clustering
…….14

2.5 Μέθοδοι αξιολόγησης συστάδων (Elbow,
Silhouette)……14

2.6 Πρόβλεψη Αποχώρησης Πελατών (Churn Modeling) στον Ενεργειακό
Τομέα……..15

2.7 Αντιμετώπιση Ανισορροπίας Κλάσεων (Class Imbalance) –
SMOTE……15

2.8 Προγνωστικά μοντέλα στη Συμπεριφορική Ανάλυση Πελατών (Behavior
Prediction
Analysis)……
……16

2.8.1 Λογιστική Παλινδρόμηση (Logistic
Regression)……16

2.8.2 Random Forest…….17

2.9 Επιχειρησιακή αξιοποίηση και στρατηγικές αποφάσεων……18

2.10 Σύνοψη βιβλιογραφικής ανασκόπησης και ερευνητικά κενά……..19

3. Μεθοδολογία Έρευνας και Περιγραφή Δεδομένων

3.1 Εισαγωγή στη μεθοδολογία……19

3.2 Ερευνητικός σχεδιασμός……20

3.3 Περιγραφή Δεδομένων. ……..…….21

3.4 Προεπεξεργασία Δεδομένων……..23

3.5 Δημιουργία Χαρακτηριστικών - RFM Features Engineerng. …….27

3.6 Τυποποίηση Ομαδοποίησης και Μείωση Διάστασης (Clustering Scaling & PCA)……28

3.7 Επιλογή αριθμού συστάδων (Elbow & Silhouette)…….30

3.8 Εφαρμογή K-Means Clustering……31

3.9 Ορισμός Μεταβλητής Αποχώρησης – Churn…….32

3.10 Διαχωρισμός Δεδομένων (Trin/Test Split)……34

3.11 Εξισορρόπηση Δεδομένων με SMOTE……..35

3.12 Εκπαίδευση Προγνωστικών Μοντέλων……37

3.12.1 Τυποποίηση Πρόβλεψης (Prediction Scaling)……37

3.12.2 Υλοποίηση Εκπαίδευσης Προγνωστικών Μοντέλων……..38

4. Αποτελέσματα και Ανάλυση

4.1 Οπτικοποίηση και Προφίλ Συστάδων……..39

4.2 Συστάδες και Αποχώρηση…….42

4.2.1 Churn Rate ανά Cluster……..42

4.2.2 Ερμηνεία Συστάδων Πελατών με Επιχειρησιακή Προσέγγιση…….43

4.3 Αξιολόγηση Προγνωστικών Μοντέλων (Evaluation)…….43

4.3.1 Απόδοση Λογιστικής Παλινδρόμησης - Logistic Regression……..….44

4.3.2 Απόδοση Τυχαίου Δάσους – Random Forest……..45

4.5 Σημαντικότητα Χαρακτηριστικών……..46

4.5.1 Ανάλυση Σημαντικότητας Χαρακτηριστικών (Feature Importance)……46

4.5.2 Σημαντικότητα Χαρακτηριστικών στο Random Forest……47

4.5.3 Συντελεστές Logistic Regression……48

4.5.4 Σύνδεση Τμηματοποίησης με Σημαντικότητα Χαρακτηριστικών (Feature Importance)……49

4.8 Συγκριτική Αξιολόγηση Μοντέλων – Final Model……50

5. Συμπεράσματα

5.1 Τελικό Συμπέρασμα Σύγκρισης Μοντέλων & Επιχειρησιακή Ερμηνεία…….50

5.2 Συμπεράσματα ……..51

5.3 Περιορισμοί και Μελλοντική Έρευνα……52

6. Βιβλιογραφικές Αναφορές

6.1 Βιβλιογραφία……53

6.1 Παράρτημα Α – Κώδικας…….59

6.2 Παράρτημα Β –
Γραφήματα……..64

Περίληψη

Η απελευθέρωση της αγοράς ηλεκτρικής ενέργειας και η αυξανόμενη διαθεσιμότητα δεδομένων κατανάλωσης από έξυπνους μετρητές έχουν εντείνει τον ανταγωνισμό μεταξύ παρόχων και την ανάγκη για αποτελεσματική κατανόηση της συμπεριφοράς των πελατών. Οι εταιρείες ενέργειας καλούνται να εντοπίζουν εγκαίρως πελάτες υψηλού κινδύνου αποχώρησης και να προσαρμόζουν στοχευμένες στρατηγικές διαχείρισης και διατήρησης.

Η παρούσα διπλωματική εργασία προτείνει ένα ολοκληρωμένο αναλυτικό πλαίσιο τμηματοποίησης και πρόβλεψης συμπεριφοράς πελατών ηλεκτρικής ενέργειας χαμηλής τάσης, αξιοποιώντας τεχνικές μηχανικής μάθησης. Αρχικά εφαρμόζεται ανάλυση RFM για την εξαγωγή χαρακτηριστικών συμπεριφοράς και στη συνέχεια πραγματοποιείται ομαδοποίηση πελατών με τον αλγόριθμο K-Means. Η ποιότητα των συστάδων αξιολογείται με τις μεθόδους Elbow και Silhouette, ενώ εξετάζεται η διαφοροποίηση του ποσοστού αποχώρησης ανά ομάδα πελατών. Στο επόμενο στάδιο αναπτύσσονται προγνωστικά μοντέλα Logistic Regression και Random Forest για την εκτίμηση της πιθανότητας αποχώρησης, λαμβάνοντας υπόψη το πρόβλημα ανισορροπίας κλάσεων μέσω της τεχνικής SMOTE.

Τα αποτελέσματα δείχνουν ότι ο συνδυασμός τμηματοποίησης και προγνωστικής ανάλυσης βελτιώνει σημαντικά την ικανότητα εντοπισμού πελατών υψηλού κινδύνου και επιτρέπει την ερμηνεία των χαρακτηριστικών που σχετίζονται με τη διακοπή συνεργασίας. Επιπλέον, η ενσωμάτωση των ομάδων πελατών στα μοντέλα πρόβλεψης ενισχύει τη χρηστικότητα των αποτελεσμάτων για επιχειρησιακή αξιοποίηση.

Η προτεινόμενη προσέγγιση συμβάλλει στην ανάπτυξη στοχευμένων στρατηγικών διατήρησης πελατών, στη βελτιστοποίηση τιμολογιακών πολιτικών και στη μετάβαση των ενεργειακών εταιρειών σε μοντέλα λειτουργίας βασισμένα σε δεδομένα.

Abstract

The liberalization of the electricity market and the increasing availability of smart meter consumption data have intensified competition among energy providers and created a strong need for effective customer behavior understanding. Energy companies are required to identify high-risk customers early and implement targeted retention strategies.

This thesis proposes an integrated analytical framework for customer segmentation and behavior prediction in the low-voltage electricity sector using machine learning techniques. Initially, RFM analysis is applied to extract behavioral features, followed by customer clustering using the K-Means algorithm. Cluster quality is evaluated using the Elbow and Silhouette methods, while churn rate differences across customer groups are examined. Subsequently, predictive models including Logistic Regression and Random Forest are developed to estimate churn probability, addressing class imbalance through the SMOTE technique.

The results demonstrate that combining segmentation with predictive analytics significantly improves the identification of high-risk customers and enables interpretation of the factors associated with customer attrition. Furthermore, incorporating customer segments into prediction models enhances their business applicability.

The proposed approach supports the development of targeted retention strategies, optimization of pricing policies, and the transition of energy companies toward data-driven decision making.

Κεφάλαιο 1

Εισαγωγή

1.1 Αντικείμενο και Σκοπός

Η παρούσα διπλωματική εργασία εξετάζει τη συμπεριφορά πελατών ηλεκτρικής ενέργειας χαμηλής τάσης μέσω της συνδυαστικής αξιοποίησης τεχνικών τμηματοποίησης και προγνωστικής ανάλυσης. Στο περιβάλλον της απελευθερωμένης αγοράς ενέργειας, η διατήρηση πελατών αποτελεί κρίσιμο ζήτημα, καθώς οι καταναλωτές μπορούν να μετακινηθούν εύκολα μεταξύ παρόχων, ενώ τα διαθέσιμα δεδομένα κατανάλωσης δεν αξιοποιούνται συστηματικά για την υποστήριξη αποφάσεων.

Στόχος της εργασίας είναι η ανάπτυξη ενός ενιαίου αναλυτικού πλαισίου που επιτρέπει την αναγνώριση διακριτών προφίλ πελατών και την εκτίμηση της πιθανότητας αποχώρησής τους. Η προσέγγιση βασίζεται στην εξαγωγή χαρακτηριστικών συμπεριφοράς μέσω ανάλυσης RFM, στην ομαδοποίηση των πελατών με αλγορίθμους μη επιβλεπόμενης μάθησης και στη δημιουργία προγνωστικών μοντέλων ταξινόμησης για την πρόβλεψη churn.

Η εργασία δεν περιορίζεται στην εφαρμογή μεμονωμένων αλγορίθμων, αλλά εστιάζει στη διασύνδεση της τμηματοποίησης με την πρόβλεψη συμπεριφοράς, ώστε τα αποτελέσματα να είναι ερμηνεύσιμα και αξιοποιήσιμα σε επιχειρησιακό επίπεδο. Με τον τρόπο αυτό διερευνάται κατά πόσο η ενσωμάτωση των χαρακτηριστικών ομάδων πελατών βελτιώνει την ικανότητα εντοπισμού πελατών υψηλού κινδύνου και υποστηρίζει στοχευμένες στρατηγικές διαχείρισης.

1.2 Ερευνητικά Ερωτήματα και Στόχοι

Με βάση τον σκοπό της εργασίας, διατυπώνονται τα ακόλουθα ερευνητικά ερωτήματα:

Μπορούν να εντοπιστούν διακριτά προφίλ πελατών ηλεκτρικής ενέργειας χαμηλής τάσης με βάση τα καταναλωτικά χαρακτηριστικά τους;

Παρουσιάζουν οι ομάδες πελατών διαφορετική πιθανότητα αποχώρησης;

Βελτιώνεται η ικανότητα πρόβλεψης αποχώρησης όταν ενσωματώνονται χαρακτηριστικά τμηματοποίησης στα προγνωστικά μοντέλα;

Ποια χαρακτηριστικά συμπεριφοράς σχετίζονται περισσότερο με τον κίνδυνο αποχώρησης πελατών;

Σκοπός της παρούσας διπλωματικής εργασίας είναι η ανάπτυξη ενός ολοκληρωμένου πλαισίου ανάλυσης πελατών που συνδυάζει τμηματοποίηση και πρόβλεψη συμπεριφοράς, με στόχο την υποστήριξη λήψης αποφάσεων στον ενεργειακό τομέα.

Οι επιμέρους στόχοι της εργασίας είναι:

- Η δημιουργία χαρακτηριστικών συμπεριφοράς πελατών από δεδομένα κατανάλωσης

- Η ομαδοποίηση πελατών σε ερμηνεύσιμες κατηγορίες
- Η ανάπτυξη και αξιολόγηση μοντέλων πρόβλεψης αποχώρησης
- Η διερεύνηση της επιχειρησιακής αξιοποίησης των αποτελεσμάτων

1.3 Δομή εργασίας

Η εργασία οργανώνεται σε επτά κεφάλαια.

Στο Κεφάλαιο 1 παρουσιάζεται το αντικείμενο, ο σκοπός και τα ερευνητικά ερωτήματα της μελέτης.

Το Κεφάλαιο 2 περιλαμβάνει συνοπτική επισκόπηση της σχετικής βιβλιογραφίας σχετικά με την τμηματοποίηση πελατών και την πρόβλεψη αποχώρησης στον ενεργειακό τομέα, καθώς και τον προσδιορισμό του ερευνητικού κενού που καλύπτει η παρούσα εργασία.

Στο Κεφάλαιο 3 περιγράφονται τα δεδομένα και η διαδικασία προεπεξεργασίας, η δημιουργία χαρακτηριστικών και η κατασκευή των μεταβλητών ανάλυσης.

Το Κεφάλαιο 4 παρουσιάζει τη μεθοδολογία της τμηματοποίησης πελατών και την αξιολόγηση των συστάδων.

Στο Κεφάλαιο 5 αναπτύσσονται τα προγνωστικά μοντέλα αποχώρησης και αξιολογείται η απόδοσή τους.

Το Κεφάλαιο 6 ερμηνεύει τα αποτελέσματα και εξετάζει τη δυνατότητα επιχειρησιακής αξιοποίησης.

Τέλος, το Κεφάλαιο 7 συνοψίζει τα συμπεράσματα και προτείνει κατευθύνσεις για μελλοντική έρευνα.

Κεφάλαιο 2

Θεωρητική Θεμελίωση

Σε αυτό το κεφάλαιο παρουσιάζονται σχετικές έρευνες, η συστηματική παρουσίαση και ανάλυση της υφιστάμενης διεθνούς βιβλιογραφίας που σχετίζεται με την ανάλυση πελατών, την ομαδοποίηση (clustering) και τη προγνωστική ανάλυση στον ενεργειακό τομέα, καθώς και οι αλγόριθμοι που θα χρησιμοποιηθούν για την ανάλυση και την πρόβλεψη των πελατών με πιθανότητα αποχώρησης. Η επιλογή του κατάλληλου μοντέλου απαιτεί κατανόηση του θεωρητικού υποβάθρου κάθε μεθόδου, καθώς και των πλεονεκτημάτων και περιορισμών της. Μετά από σχετικές δοκιμές, επιλέχθηκαν δύο μοντέλα με βάση την αποδοτικότητα και την καταλληλότητα τους, τα οποία θα εξεταστούν αναλυτικά στη συνέχεια.

2.1 Σχετικές Έρευνες

Η πρόβλεψη αποχώρησης πελατών στον τομέα της ενέργειας αποτελεί αντικείμενο αυξανόμενου ερευνητικού ενδιαφέροντος, ιδιαίτερα σε αγορές όπου η δυνατότητα αλλαγής παρόχου είναι άμεση και το προϊόν παρουσιάζει περιορισμένη διαφοροποίηση. Οι περισσότερες μελέτες βασίζονται σε ιστορικά δεδομένα κατανάλωσης, οικονομικής συνέπειας και αλληλεπίδρασης πελάτη-εταιρείας με στόχο την ανάπτυξη μοντέλων ταξινόμησης που εκτιμούν την πιθανότητα churn.

Σημαντικό μέρος της βιβλιογραφίας εστιάζει στη βελτίωση της προγνωστικής ακρίβειας μέσω επιβλεπόμενων αλγορίθμων μηχανικής μάθησης, όπως δέντρα αποφάσεων και ensemble μοντέλα, τα οποία εμφανίζουν αυξημένη απόδοση σε δεδομένα μεγάλης κλίμακας και με ανισορροπία κλάσεων. Τα αποτελέσματα των ερευνών αυτών καταδεικνύουν ότι μεταβολές στα πρότυπα κατανάλωσης, καθυστερήσεις πληρωμών και συχνή επικοινωνία με την εξυπηρέτηση πελατών αποτελούν ισχυρούς δείκτες αποχώρησης.

Παράλληλα, άλλες εργασίες εξετάζουν την τμηματοποίηση πελατών μέσω τεχνικών μη επιβλεπόμενης μάθησης, με στόχο την κατανόηση ετερογενών προτύπων κατανάλωσης και τη δημιουργία ενεργειακών προφίλ. Η προσέγγιση αυτή επιτρέπει την αναγνώριση ομάδων πελατών με παρόμοια συμπεριφορά, ωστόσο συνήθως χρησιμοποιείται ανεξάρτητα από τα προγνωστικά μοντέλα αποχώρησης.

Προηγούμενες μελέτες δείχνουν ότι μεταβολές στην κατανάλωση και καθυστερήσεις πληρωμών αποτελούν ισχυρούς δείκτες αποχώρησης (Verbeke et al., 2012; Depren et al., 2017), ενώ ensemble αλγόριθμοι εμφανίζουν υψηλότερη ακρίβεια σε δεδομένα υπηρεσιών κοινής ωφέλειας (Coussement & De Bock, 2013).

Συνολικά, η υπάρχουσα βιβλιογραφία αντιμετωπίζει την τμηματοποίηση και την πρόβλεψη αποχώρησης ως δύο διακριτά προβλήματα. Παρατηρείται περιορισμένος αριθμός μελετών που ενσωματώνουν τα χαρακτηριστικά των ομάδων πελατών σε μοντέλα πρόβλεψης, παρότι η συνδυαστική αξιοποίησή τους θα μπορούσε να ενισχύσει την ερμηνευσιμότητα και τη χρηστικότητα των αποτελεσμάτων. Η παρούσα εργασία επιχειρεί να καλύψει το συγκεκριμένο κενό, διερευνώντας κατά πόσο η ενσωμάτωση της τμηματοποίησης βελτιώνει την αναγνώριση πελατών υψηλού κινδύνου στον ενεργειακό τομέα.

2.2 Ανάλυση πελατών και τμηματοποίηση (Customer Segmentation)

Η τμηματοποίηση πελατών χρησιμοποιείται στον ενεργειακό τομέα για την κατανόηση της ετερογένειας στα πρότυπα κατανάλωσης και τη διαφοροποίηση της συμπεριφοράς μεταξύ καταναλωτών. Σε αντίθεση με παραδοσιακές αγορές, όπου η κατηγοριοποίηση βασίζεται κυρίως σε δημογραφικά χαρακτηριστικά, τα δεδομένα ηλεκτρικής ενέργειας επιτρέπουν τη δημιουργία ομάδων με βάση πραγματική χρήση υπηρεσιών.

Οι κατανάλωσης, τα χαρακτηριστικά τιμολόγησης και το ιστορικό συναλλαγών αποτυπώνουν λειτουργικά προφίλ πελατών, τα οποία μπορούν να χρησιμοποιηθούν για την αναγνώριση διαφορετικών μοτίβων ενεργειακής συμπεριφοράς. Μελέτες στον τομέα των υπηρεσιών κοινής ωφέλειας δείχνουν ότι πελάτες με παρόμοια επίπεδα συνολικής κατανάλωσης ενδέχεται να παρουσιάζουν διαφορετική σταθερότητα χρήσης, εποχικότητα ή ευαισθησία στις τιμές, γεγονός που δεν μπορεί να αποτυπωθεί με απλές συγκεντρωτικές μετρικές.

Η τμηματοποίηση σε αυτό το πλαίσιο δεν αποτελεί μόνο εργαλείο περιγραφικής ανάλυσης αλλά και μέσο δημιουργίας χαρακτηριστικών για προγνωστικά μοντέλα. Οι

ομάδες πελατών λειτουργούν ως συνοπτικές αναπαραστάσεις συμπεριφοράς, επιτρέποντας την ενσωμάτωση σύνθετων προτύπων σε διαδικασίες πρόβλεψης.

Στην παρούσα εργασία, η τμηματοποίηση χρησιμοποιείται ως ενδιάμεσο στάδιο μεταξύ της επεξεργασίας δεδομένων και της πρόβλεψης αποχώρησης, με στόχο τη διερεύνηση της επίδρασης των συμπεριφορικών προφίλ στην πιθανότητα μετακίνησης πελατών.

2.3 Συσταδοποίηση (Clustering Analysis)

Η ανάλυση δεδομένων κατανάλωσης ηλεκτρικής ενέργειας χαρακτηρίζεται από υψηλή ετερογένεια μεταξύ καταναλωτών, ακόμη και όταν παρουσιάζουν παρόμοια συνολικά επίπεδα χρήσης. Η διαφοροποίηση αυτή σχετίζεται με την κατανάλωση, τη συνέπεια των πληρωμών και τη σταθερότητα των προτύπων χρήσης, στοιχεία που δεν μπορούν να αποτυπωθούν μέσω απλής κατηγοριοποίησης.

Οι τεχνικές συσταδοποίησης επιτρέπουν την αναγνώριση υποκείμενων δομών στα δεδομένα χωρίς την ύπαρξη προκαθορισμένων ετικετών, καθιστώντας τις κατάλληλες για προβλήματα όπου οι κατηγορίες πελατών δεν είναι γνωστές εκ των προτέρων. Στον ενεργειακό τομέα χρησιμοποιούνται για τον εντοπισμό ομάδων με παρόμοια λειτουργική συμπεριφορά κατανάλωσης, οι οποίες συχνά συνδέονται με διαφορετικές ανάγκες τιμολόγησης ή πιθανότητα μετακίνησης παρόχου.

Στην παρούσα εργασία η συσταδοποίηση δεν εφαρμόζεται ως τελικός στόχος, αλλά ως μηχανισμός συμπύκνωσης της συμπεριφορικής πληροφορίας. Οι προκύπτουσες ομάδες λειτουργούν ως αναπαραστάσεις προτύπων χρήσης και αξιοποιούνται στη συνέχεια ως είσοδος σε προγνωστικά μοντέλα, επιτρέποντας τη μεταφορά πολύπλοκης πληροφορίας σε μορφή κατάλληλη για διαδικασίες ταξινόμησης.

2.4 Τεχνικές ομαδοποίησης (Clustering)

2.4.1. Ανάλυση RFM

Η ανάλυση RFM χρησιμοποιείται στην παρούσα εργασία ως μέθοδος δημιουργίας χαρακτηριστικών συμπεριφοράς από τα δεδομένα κατανάλωσης. Αντί να αξιοποιούνται απευθείας οι αρχικές μετρήσεις ενέργειας, μετατρέπονται σε συνοπτικούς δείκτες που αποτυπώνουν τη σχέση του πελάτη με την υπηρεσία.

Η διάσταση Recency εκφράζει το χρονικό διάστημα από την τελευταία ενεργή κατανάλωση ή συναλλαγή, αποτυπώνοντας τη χρονική εγγύτητα της χρήσης. Η Frequency περιγράφει τη συχνότητα αλληλεπίδρασης με την υπηρεσία σε καθορισμένο χρονικό ορίζοντα, ενώ η Monetary αντιστοιχεί στη συνολική ενεργειακή ή οικονομική αξία κατανάλωσης.

Οι τρεις δείκτες λειτουργούν ως συμπυκνωμένη αναπαράσταση της συμπεριφοράς του πελάτη, μειώνοντας τη διάσταση των δεδομένων και επιτρέποντας την εφαρμογή αλγορίθμων συσταδοποίησης σε ομοιογενή χαρακτηριστικά. Με τον τρόπο αυτό η ανάλυση RFM δεν αποτελεί διαδικασία τμηματοποίησης από μόνη της, αλλά στάδιο προετοιμασίας δεδομένων που επιτρέπει την αναγνώριση προτύπων χρήσης και τη σύνδεσή τους με την πιθανότητα αποχώρησης.

2.4.2 K-Means Clustering

Για την ομαδοποίηση των πελατών επιλέχθηκε ο αλγόριθμος K-Means, καθώς επιτρέπει τον διαχωρισμό παρατηρήσεων σε διακριτές ομάδες με βάση την ομοιότητα των χαρακτηριστικών τους. Στο πλαίσιο των δεδομένων κατανάλωσης, όπου τα χαρακτηριστικά έχουν αριθμητική μορφή και συγκρίσιμη κλίμακα, η μέθοδος είναι κατάλληλη για τον εντοπισμό προτύπων συμπεριφοράς. Η επιλογή του συγκεκριμένου αλγορίθμου βασίζεται στην ικανότητά του να δημιουργεί συμπαγείς και εύκολα ερμηνεύσιμες ομάδες, οι οποίες μπορούν να συσχετιστούν με διαφορετικά προφίλ πελατών. Σε αντίθεση με πιο σύνθετες τεχνικές, η απλότητα του μοντέλου επιτρέπει την άμεση σύνδεση των ομάδων με επιχειρησιακές αποφάσεις, γεγονός ιδιαίτερα σημαντικό σε εφαρμογές διαχείρισης πελατών.

Οι εφαρμογές του είναι ευρείες:

- Τμηματοποίηση Πελατών (Customer Segmentation): Όπως και στην περίπτωσή μας, για την ταυτοποίηση διαφορετικών ομάδων πελατών με βάση τη συμπεριφορά τους, ώστε οι επιχειρήσεις να μπορούν να προσαρμόσουν τις στρατηγικές μάρκετινγκ.
- Ανάλυση Συμπεριφοράς Πελατών: Κατανόηση των προτιμήσεων και των συνηθειών αγοράς.
- Ομαδοποίηση: Ομαδοποίηση πελατών, με παρόμοια χαρακτηριστικά.

Στην παρούσα εργασία ο K-Means χρησιμοποιείται ως μέθοδος αναγνώρισης λειτουργικών προτύπων κατανάλωσης και όχι ως τελικό μοντέλο πρόβλεψης. Οι προκύπτουσες ομάδες αποτελούν μεταβλητές εισόδου για την εκτίμηση της πιθανότητας αποχώρησης, επιτρέποντας τη μεταφορά σύνθετης συμπεριφορικής πληροφορίας σε προγνωστικά μοντέλα.

2.5 Μέθοδοι αξιολόγησης συστάδων (Elbow, Silhouette)

Μέθοδοι όπως το elbow method και ο δείκτης silhouette αποτελούν καθιερωμένα κριτήρια αξιολόγησης της ποιότητας της ομαδοποίησης και χρησιμοποιούνται ευρέως στη βιβλιογραφία για την επιλογή του αριθμού συστάδων στον αλγόριθμο K-Means (Hastie et al., 2009Jain, 2010). Η επιλογή του αριθμού των ομάδων αποτελεί κρίσιμο στάδιο στη διαδικασία συσταδοποίησης, καθώς επηρεάζει άμεσα την ερμηνεία των προφίλ πελατών. Για τον προσδιορισμό του κατάλληλου πλήθους συστάδων χρησιμοποιήθηκαν δύο συμπληρωματικά κριτήρια αξιολόγησης. Αρχικά εφαρμόστηκε η μέθοδος Elbow, η οποία εξετάζει τη μεταβολή της εσωτερικής διασποράς των ομάδων για διαφορετικό αριθμό συστάδων. Το σημείο στο οποίο η βελτίωση παύει να είναι ουσιαστική υποδεικνύει μια ισορροπία μεταξύ πολυπλοκότητας και περιγραφικής ικανότητας. Η μέθοδος αυτή παρέχει μια οπτική προσέγγιση για την εκτίμηση του βέλτιστου αριθμού ομάδων, επιτρέποντας την αναγνώριση του σημείου όπου η προσθήκη επιπλέον συστάδων δεν προσφέρει ουσιαστική πληροφορία. Παράλληλα χρησιμοποιήθηκε ο δείκτης Silhouette για την αξιολόγηση της συνοχής και του διαχωρισμού των ομάδων. Ο δείκτης υπολογίζει τη μέση απόσταση κάθε παρατήρησης από τα σημεία της ίδιας και των γειτονικών συστάδων, παρέχοντας μια ποσοτική εκτίμηση της ποιότητας της ομαδοποίησης. Υψηλότερες τιμές υποδηλώνουν ότι οι παρατηρήσεις ανήκουν σε κατάλληλη ομάδα και ότι οι συστάδες είναι επαρκώς διακριτές, ενώ χαμηλές ή αρνητικές τιμές υποδεικνύουν επικαλύψεις ή

ασαφή όρια μεταξύ ομάδων. Επιπλέον, εξετάστηκε η συνέπεια των αποτελεσμάτων μέσω επαναλαμβανόμενων εκτελέσεων του αλγορίθμου με διαφορετικές αρχικές τιμές, ώστε να διασφαλιστεί η σταθερότητα των συστάδων. Ο συνδυασμός των δύο μετρικών, σε συνδυασμό με την οπτική επιθεώρηση των αποτελεσμάτων, επιτρέπει την επιλογή αριθμού ομάδων που δεν βελτιστοποιεί μόνο ένα μαθηματικό κριτήριο αλλά παράγει ερμηνεύσιμα και σταθερά προφίλ πελατών, κατάλληλα για περαιτέρω ανάλυση και επιχειρησιακή αξιοποίηση.

2.6 Πρόβλεψη Αποχώρησης Πελατών (Churn Modeling) στον Ενεργειακό Τομέα

Η πρόβλεψη αποχώρησης πελατών στον ενεργειακό τομέα αφορά την εκτίμηση της πιθανότητας διακοπής συνεργασίας με βάση ιστορικά δεδομένα χρήσης και συναλλαγών. Σε περιβάλλοντα όπου οι υπηρεσίες παρουσιάζουν περιορισμένη διαφοροποίηση, η έγκαιρη αναγνώριση πελατών υψηλού κινδύνου αποτελεί βασικό εργαλείο διατήρησης πελατειακής βάσης. Τα προγνωστικά μοντέλα επιτρέπουν την εκτίμηση της πιθανότητας αποχώρησης, ωστόσο η ερμηνεία των αποτελεσμάτων παραμένει συχνά περιορισμένη όταν βασίζεται αποκλειστικά σε μεμονωμένα χαρακτηριστικά. Πελάτες με παρόμοια επίπεδα κατανάλωσης μπορεί να εμφανίζουν διαφορετική συμπεριφορά, γεγονός που καθιστά δύσκολη την κατανόηση των αιτίων αποχώρησης μόνο μέσω ταξινόμησης.Η ενσωμάτωση της τμηματοποίησης στη διαδικασία πρόβλεψης επιτρέπει τη σύνδεση της πιθανότητας αποχώρησης με συγκεκριμένα πρότυπα χρήσης. Με τον τρόπο αυτό η πρόβλεψη δεν περιορίζεται σε αριθμητική εκτίμηση κινδύνου αλλά συνδέεται με ερμηνεύσιμα προφίλ πελατών, διευκολύνοντας την εφαρμογή στοχευμένων ενεργειών διαχείρισης. Στην παρούσα εργασία, η πρόβλεψη αποχώρησης αξιοποιείται σε συνδυασμό με τα αποτελέσματα της συσταδοποίησης, ώστε να διερευνηθεί κατά πόσο τα συμπεριφορικά προφίλ βελτιώνουν την αναγνώριση πελατών υψηλού κινδύνου.

2.7 Αντιμετώπιση Ανισορροπίας Κλάσεων (Class Imbalance) - SMOTE

Στα προβλήματα πρόβλεψης αποχώρησης παρατηρείται συνήθως ανισορροπία μεταξύ των κατηγοριών, καθώς οι πελάτες που διακόπτουν τη συνεργασία αποτελούν μικρό ποσοστό του συνόλου. Σε τέτοιες περιπτώσεις τα προγνωστικά μοντέλα τείνουν να ευνοούν την πλειονοτική κατηγορία, οδηγώντας σε φαινομενικά υψηλή ακρίβεια αλλά περιορισμένη ικανότητα εντοπισμού πελατών υψηλού κινδύνου.Για την αντιμετώπιση του προβλήματος εφαρμόστηκε η τεχνική SMOTE, η οποία αυξάνει την παρουσία της μειονοτικής κατηγορίας μέσω δημιουργίας συνθετικών παρατηρήσεων. Η διαδικασία εφαρμόζεται στο σύνολο εκπαίδευσης ώστε να βελτιωθεί η εκμάθηση των ορίων απόφασης χωρίς να αλλοιώνεται η αξιολόγηση του μοντέλου.Η Τεχνική Υπερδειγματοληψίας Συνθετικής Μειονοτικής Κλάσης (Synthetic Minority Over-sampling Technique), εν συντομία SMOTE, είναι μια δημοφιλής μέθοδος που χρησιμοποιείται στη μηχανική μάθηση για την αντιμετώπιση του προβλήματος της ανισορροπίας κλάσεων (class imbalance). Με τον τρόπο αυτό η εκπαίδευση των προγνωστικών μοντέλων βασίζεται σε πιο ισορροπημένη αναπαράσταση των περιπτώσεων αποχώρησης, επιτρέποντας την αξιόπιστη εκτίμηση της πιθανότητας churn. Λειτουργεί δημιουργώντας συνθετικά δείγματα της μειονοτικής κλάσης, αντί να αντιγράφει απλώς τα υπάρχοντα δείγματα.

Η βασική διαδικασία είναι η εξής:

1. Εντοπισμός Μειονοτικών Δειγμάτων: Επιλέγει ένα δείγμα από τη μειονοτική κλάση.
2. Εύρεση Γειτόνων: Βρίσκει τους k πλησιέστερους γείτονες του επιλεγμένου δείγματος (επίσης από τη μειονοτική κλάση).
3. Δημιουργία Συνθετικού Δείγματος: Επιλέγει τυχαία έναν από αυτούς τους γείτονες και δημιουργεί ένα νέο, συνθετικό δείγμα σε ένα σημείο κατά μήκος της ευθείας γραμμής που ενώνει το αρχικό δείγμα με τον επιλεγμένο γείτονα.

Αυτή η διαδικασία επαναλαμβάνεται μέχρι ο αριθμός των δειγμάτων της μειονοτικής κλάσης να φτάσει το επιθυμητό επίπεδο (συνήθως να γίνει ίσος με τον αριθμό των δειγμάτων της πλειονοτικής κλάσης).

2.8 Προγνωστικά μοντέλα στη Συμπεριφορική Ανάλυση Πελατών (Behavior Prediction Analysis)

Η τμηματοποίηση επιτρέπει την κατανόηση των διαφορετικών προτύπων χρήσης, ωστόσο δεν παρέχει άμεση εκτίμηση μελλοντικής συμπεριφοράς. Για τον λόγο αυτό απαιτείται η αξιοποίηση προγνωστικών μοντέλων, τα οποία συσχετίζουν ιστορικά χαρακτηριστικά με συγκεκριμένα μελλοντικά αποτελέσματα.

Στο πρόβλημα της αποχώρησης πελατών η πρόβλεψη διατυπώνεται ως πρόβλημα ταξινόμησης, όπου το μοντέλο εκτιμά την πιθανότητα μετακίνησης κάθε πελάτη. Η διαδικασία αυτή επιτρέπει την ιεράρχηση του κινδύνου και την προτεραιοποίηση ενεργειών διατήρησης. Στην παρούσα εργασία τα προγνωστικά μοντέλα δεν χρησιμοποιούνται ανεξάρτητα από την προηγούμενη ανάλυση, αλλά αξιοποιούν τα χαρακτηριστικά που προκύπτουν από τη συσταδοποίηση. Με τον τρόπο αυτό διερευνάται κατά πόσο η ενσωμάτωση των συμπεριφορικών προφίλ βελτιώνει την ικανότητα εντοπισμού πελατών υψηλού κινδύνου σε σχέση με τη χρήση μόνο των αρχικών μεταβλητών.

2.8.1 Λογιστική Παλινδρόμηση (Logistic Regression)

Η λογιστική παλινδρόμηση είναι ένα στατιστικό μοντέλο που χρησιμοποιείται για την πρόβλεψη διχοτομημένων αποτελεσμάτων ή για τον υπολογισμό της πιθανότητας εμφάνισης ενός συγκεκριμένου γεγονότος. Αποτελεί επέκταση της κλασικής γραμμικής παλινδρόμησης, η οποία εκφράζεται ως:

$$z = \beta_{0+}\beta_1 X_{1+}\beta_2 X_{2+\cdots+}\beta_v X_v$$

Με τη γραμμική παλινδρόμηση, που προβλέπει συνεχείς τιμές, η λογιστική παλινδρόμηση χρησιμοποιεί μια συνάρτηση για να μοντελοποιήσει τη σχέση ανάμεσα σε ανεξάρτητες μεταβλητές και μια διχοτομεί εξαρτημένη μεταβλητή z. Η εξαρτημένη μεταβλητή παίρνει την τιμή 0 όταν το συμβάν δεν συμβαίνει (εν προκειμένω όταν ο πελάτης δεν εγκαταλείπει την εταιρεία) και 1 όταν συμβαίνει (π.χ. πελάτης που είναι churner). Χρησιμοποιώντας λογαριθμικό λόγο ως συνάρτηση σύνδεσης, η λογιστική

παλινδρόμηση εκτιμά τη σχέση ανάμεσα στις ανεξάρτητες μεταβλητές και στον λογαριθμικό λόγο πιθανοτήτων του συμβάντος. Η εκτίμηση των παραμέτρων γίνεται συνήθως με βήμα-βήμα προσθήκη ή αφαίρεση μεταβλητών (stepwise), ώστε να διατηρούνται μόνο οι στατιστικά σημαντικές μεταβλητές. Η εκπαίδευση του μοντέλου βασίζεται στη μέθοδο μέγιστης πιθανοφάνειας, που επιλέγει τους συντελεστές που μεγιστοποιούν την πιθανότητα παρατήρησης των δεδομένων. Η ποιότητα της πρόβλεψης μπορεί να αξιολογηθεί μέσω δεικτών όπως η ακρίβεια, η ανάκληση και η καμπύλη λήψης αποφάσεων (ROC). Επιπλέον, η λογιστική παλινδρόμηση είναι ικανή να ενσωματώσει πολλαπλές ανεξάρτητες μεταβλητές και να χειριστεί κατηγορικές, διατεταγμένες ή συνεχείς μεταβλητές, προσφέροντας ευελιξία και εύκολη ερμηνεία των αποτελεσμάτων (Mehta et al., 1995).Η λογιστική παλινδρόμηση χρησιμοποιείται ως βασικό μοντέλο ταξινόμησης για την εκτίμηση της πιθανότητας αποχώρησης πελατών. Η μέθοδος επιτρέπει τη συσχέτιση των χαρακτηριστικών συμπεριφοράς με την πιθανότητα εμφάνισης του γεγονότος, παρέχοντας ταυτόχρονα ερμηνεύσιμους συντελεστές. Το κύριο πλεονέκτημα του μοντέλου είναι η δυνατότητα κατανόησης της επίδρασης κάθε μεταβλητής στον κίνδυνο αποχώρησης, γεγονός που το καθιστά κατάλληλο για εφαρμογές όπου απαιτείται διαφάνεια στη λήψη αποφάσεων. Στο πλαίσιο της παρούσας εργασίας χρησιμοποιείται ως σημείο αναφοράς, επιτρέποντας τη σύγκριση με πιο σύνθετες μεθόδους. Η λογιστική παλινδρόμηση εφαρμόζεται τόσο στα αρχικά χαρακτηριστικά όσο και σε εκείνα που προκύπτουν από τη συσταδοποίηση, ώστε να διερευνηθεί η επίδραση των συμπεριφορικών ομάδων στην πιθανότητα αποχώρησης.

2.8.2 Random Forest

Ο αλγόριθμος Random Forest (RF), γνωστός και ως Δάση Τυχαίας Απόφασης, αποτελεί μια ιδιαίτερα ευέλικτη μέθοδο εκμάθησης συνόλου, η οποία χρησιμοποιείται εκτενώς στην προγνωστική ανάλυση, χάρη στην ικανότητά της να αποδίδει αξιόπιστες και ακριβείς προβλέψεις σε πληθώρα εφαρμογών. Πρόκειται για μια τεχνική που βασίζεται στα δένδρα αποφάσεων και τα επεκτείνει, δημιουργώντας μεγάλο αριθμό ανεξάρτητων δένδρων, των οποίων τα αποτελέσματα συνδυάζονται προκειμένου να παραχθεί η τελική πρόβλεψη. Ο Random Forest συνδυάζει πολλαπλά μοντέλα, ώστε η συλλογική τους απόφαση να είναι πιο αξιόπιστη από αυτή ενός μόνο δέντρο, σύμφωνα με την οποία η συλλογική απόφαση πολλών μοντέλων τείνει να είναι πιο αξιόπιστη από εκείνη ενός μεμονωμένου μοντέλου. Κατά την εκπαίδευση του μοντέλου, κατασκευάζεται ένας προκαθορισμένος αριθμός δένδρων αποφάσεων, όπου κάθε δένδρο εκπαιδεύεται χρησιμοποιώντας ένα τυχαία επιλεγμένο υποσύνολο των αρχικών δεδομένων, καθώς και ένα τυχαίο σύνολο χαρακτηριστικών. Κάθε δένδρο εκπαιδεύεται ξεχωριστά σε διαφορετικά τυχαία δείγματα, ώστε να εξασφαλίζεται ποικιλία στις προβλέψεις. Καθ’ όλη τη διαδικασία εκπαίδευσης, ο αλγόριθμος επιλέγει τα βέλτιστα χαρακτηριστικά και τα καταλληλότερα σημεία διαχωρισμού, επιδιώκοντας τη μέγιστη μείωση της ακαθαρσίας ή της διακύμανσης. Παράλληλα, πραγματοποιείται βελτιστοποίηση τόσο των χαρακτηριστικών όσο και της δομής των δένδρων, με στόχο τον περιορισμό της διακύμανσης και της αλληλοεπικάλυψης των προβλέψεων (Biau & Scornet, 2016). Για να παραχθεί η τελική πρόβλεψη, ο Random Forest συγκεντρώνει τα αποτελέσματα όλων των δέντρων. Στις εφαρμογές ταξινόμησης επιλέγεται η κλάση που εμφανίζεται πιο συχνά, ενώ στις παλινδρομήσεις υπολογίζεται ο μέσος όρος των μεμονωμένων προβλέψεων.

Ο αλγόριθμος Random Forest χρησιμοποιείται ως πιο ευέλικτο μοντέλο πρόβλεψης αποχώρησης, ικανό να αποτυπώσει μη γραμμικές σχέσεις μεταξύ των χαρακτηριστικών. Σε δεδομένα συμπεριφοράς

πελατών, όπου οι αλληλεπιδράσεις μεταξύ μεταβλητών μπορεί να είναι σύνθετες, τα δέντρα απόφασης μπορούν να εντοπίσουν πρότυπα που δεν περιγράφονται εύκολα από γραμμικά μοντέλα. Σε αντίθεση με τη λογιστική παλινδρόμηση, το μοντέλο δεν βασίζεται σε υποθέσεις γραμμικότητας και μπορεί να αξιοποιήσει αποτελεσματικά μεγάλο αριθμό χαρακτηριστικών. Για τον λόγο αυτό χρησιμοποιείται συμπληρωματικά, επιτρέποντας τη σύγκριση μεταξύ ερμηνεύσιμων και πιο ισχυρών προγνωστικών προσεγγίσεων.

Στην παρούσα εργασία το Random Forest εφαρμόζεται στα ίδια σύνολα χαρακτηριστικών με τη λογιστική παλινδρόμηση, ώστε να αξιολογηθεί η επίδραση της πολυπλοκότητας του μοντέλου στην ακρίβεια πρόβλεψης και στην αναγνώριση πελατών υψηλού κινδύνου.

2.9 Επιχειρησιακή αξιοποίηση και στρατηγικές αποφάσεων

Η αξιοποίηση αναλυτικών μοντέλων στον ενεργειακό τομέα επικεντρώνεται κυρίως στην έγκαιρη αναγνώριση πελατών υψηλού κινδύνου και στη διαφοροποίηση της διαχείρισης ανάλογα με το προφίλ χρήσης. Η συνδυαστική εφαρμογή τμηματοποίησης και πρόβλεψης επιτρέπει τη μετάβαση από γενικές πολιτικές διατήρησης σε στοχευμένες παρεμβάσεις. Οι ομάδες πελατών παρέχουν ερμηνεύσιμη πληροφορία σχετικά με τα πρότυπα κατανάλωσης, ενώ τα προγνωστικά μοντέλα εκτιμούν την πιθανότητα αποχώρησης σε επίπεδο πελάτη. Η ταυτόχρονη χρήση τους επιτρέπει την ιεράρχηση ενεργειών, όπως διαφοροποίηση τιμολογίων, εξατομικευμένη επικοινωνία και προληπτική υποστήριξη πελατών με αυξημένο κίνδυνο. Στην παρούσα εργασία η επιχειρησιακή αξιοποίηση δεν εξετάζεται θεωρητικά, αλλά ως άμεση εφαρμογή των αποτελεσμάτων της ανάλυσης, διερευνώντας πώς τα συμπεριφορικά προφίλ μπορούν να υποστηρίξουν αποφάσεις διαχείρισης πελατειακής βάσης στον ενεργειακό τομέα.

2.10 Σύνοψη βιβλιογραφικής ανασκόπησης και ερευνητικά κενά

Η βιβλιογραφία στον ενεργειακό τομέα παρουσιάζει εκτενή χρήση τόσο τεχνικών τμηματοποίησης όσο και προγνωστικών μοντέλων αποχώρησης. Οι προσεγγίσεις αυτές εφαρμόζονται συνήθως ανεξάρτητα: η συσταδοποίηση χρησιμοποιείται για την κατανόηση προτύπων κατανάλωσης, ενώ τα μοντέλα ταξινόμησης για την εκτίμηση πιθανότητας μετακίνησης πελατών. Ωστόσο, η ανεξάρτητη εφαρμογή τους περιορίζει την ερμηνεία των αποτελεσμάτων. Τα προγνωστικά μοντέλα παρέχουν εκτίμηση κινδύνου χωρίς σαφή σύνδεση με συγκεκριμένη συμπεριφορά, ενώ η τμηματοποίηση περιγράφει ομάδες χωρίς ποσοτική εκτίμηση μελλοντικής δράσης. Το ερευνητικό κενό εντοπίζεται στην περιορισμένη ενσωμάτωση των δύο προσεγγίσεων σε ενιαίο πλαίσιο ανάλυσης. Η παρούσα εργασία διερευνά κατά πόσο τα χαρακτηριστικά που προκύπτουν από τη συσταδοποίηση μπορούν να αξιοποιηθούν ως είσοδος σε μοντέλα πρόβλεψης, βελτιώνοντας τόσο την ακρίβεια όσο και την ερμηνευσιμότητα των αποτελεσμάτων. Με τον τρόπο αυτό η μελέτη μετατοπίζεται από την απλή

πρόβλεψη ή περιγραφή της συμπεριφοράς σε μια ολοκληρωμένη διαδικασία κατανόησης και εκτίμησης κινδύνου αποχώρησης πελατών στον ενεργειακό τομέα.

ΚΕΦΑΛΑΙΟ:3

Μεθοδολογία Έρευνας και Περιγραφή Δεδομένων

3.1 Εισαγωγή στη μεθοδολογία

Το παρόν κεφάλαιο παρουσιάζει τη μεθοδολογική προσέγγιση που ακολουθήθηκε για την τμηματοποίηση πελατών και την πρόβλεψη αποχώρησης στον ενεργειακό τομέα. Η ανάλυση υλοποιήθηκε σε περιβάλλον Python (Google Colab) και οργανώθηκε ως αναπαραγώγιμο pipeline, το οποίο περιλαμβάνει διαδοχικά στάδια προεπεξεργασίας, δημιουργίας χαρακτηριστικών, μη επιβλεπόμενης ομαδοποίησης και επιβλεπόμενης πρόβλεψης. Αρχικά δημιουργήθηκε σύνολο δεδομένων προσομοίωσης, το οποίο αναπαριστά μηνιαίες καταναλώσεις πελατών και αντίστοιχα οικονομικά μεγέθη. Η χρήση προσομοιωμένων δεδομένων επιλέχθηκε ώστε να διασφαλιστεί ελεγχόμενο περιβάλλον πειραματισμού και να αξιολογηθεί συστηματικά η προτεινόμενη μεθοδολογία, χωρίς περιορισμούς πρόσβασης ή ευαισθησίας πραγματικών δεδομένων. Στη συνέχεια πραγματοποιήθηκε μετασχηματισμός των αρχικών μετρήσεων σε χαρακτηριστικά συμπεριφοράς μέσω ανάλυσης RFM, με στόχο τη συμπύκνωση της πληροφορίας σε ερμηνεύσιμους δείκτες. Για την τμηματοποίηση πελατών εφαρμόστηκε ο αλγόριθμος K-Means, ενώ ο αριθμός των συστάδων επιλέχθηκε με συνδυαστική αξιολόγηση (Elbow και Silhouette). Οι ομάδες που προκύπτουν χρησιμοποιήθηκαν τόσο για περιγραφική ανάλυση όσο και ως πρόσθετη είσοδος στα προγνωστικά μοντέλα. Στο επόμενο στάδιο ορίστηκε μεταβλητή αποχώρησης (churn) με βάση κριτήριο αδράνειας/απουσίας πρόσφατης δραστηριότητας, ώστε το πρόβλημα να διατυπωθεί ως δυαδική ταξινόμηση. Τέλος, αναπτύχθηκαν και συγκρίθηκαν δύο προγνωστικά μοντέλα, η Λογιστική Παλινδρόμηση (Logistic Regression) αποτελεί βασικό μοντέλο δυαδικής ταξινόμησης με υψηλή ερμηνευσιμότητα (Hosmer et al., 2013) και το Random Forest, το οποίο βασίζεται σε σύνολο δέντρων αποφάσεων και παρουσιάζει υψηλή ικανότητα γενίκευσης (Breiman, 2001). Λόγω ανισορροπίας κλάσεων εφαρμόστηκε εξισορρόπηση στο σύνολο εκπαίδευσης με τη μέθοδο SMOTE, η οποία δημιουργεί συνθετικά δείγματα της μειοψηφικής κατηγορίας (Chawla et al., 2002). Η εκπαίδευση σε ανισόρροπα δεδομένα μπορεί να οδηγήσει σε μεροληπτικά μοντέλα (He & Garcia, 2009). Η αξιολόγηση πραγματοποιήθηκε σε ξεχωριστό σύνολο ελέγχου με μετρικές κατάλληλες για προβλήματα churn prediction, ώστε να εκτιμηθεί τόσο η προγνωστική απόδοση όσο και η πρακτική αξιοποίηση των αποτελεσμάτων.

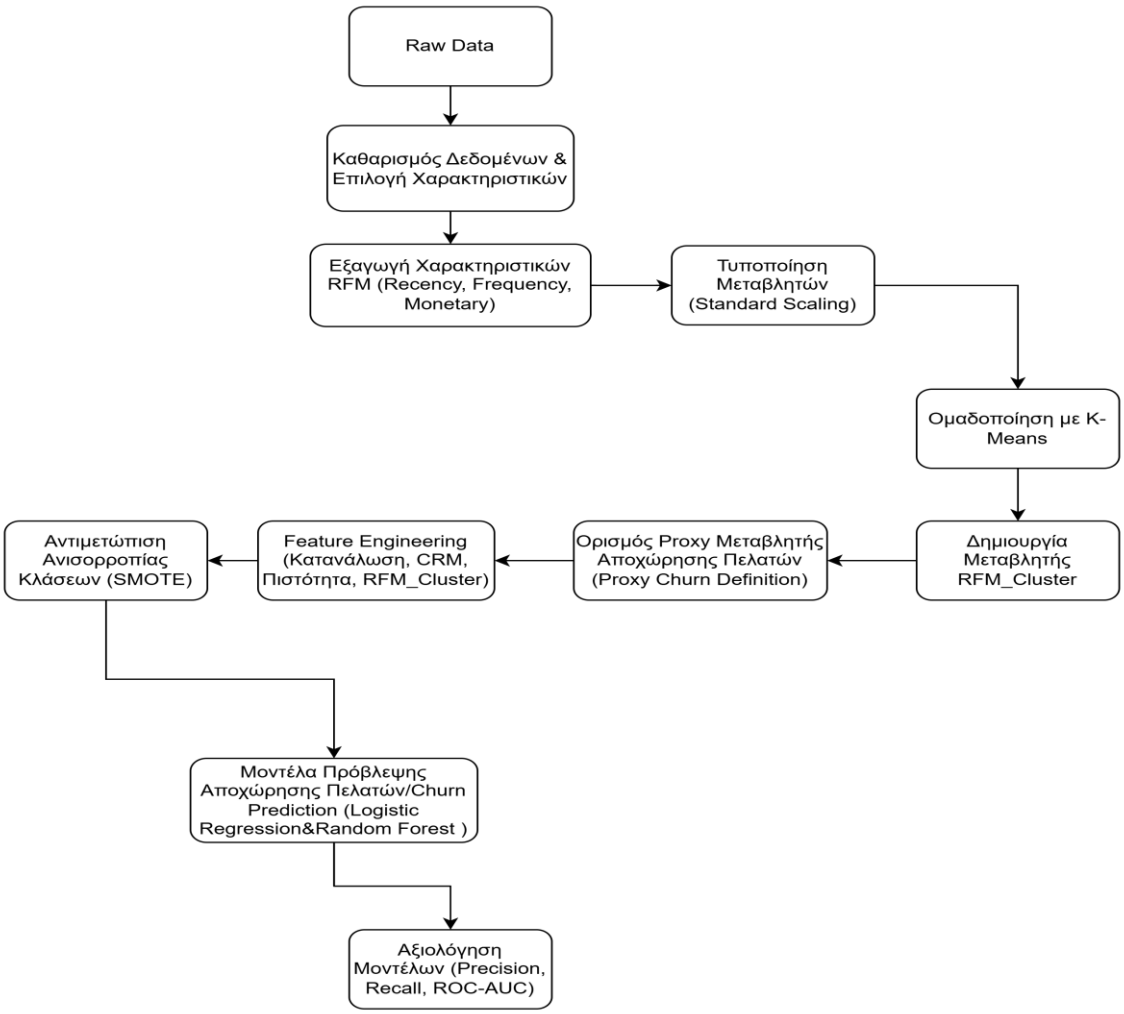
3.2 Ερευνητικός σχεδιασμός

Η μεθοδολογία της εργασίας οργανώνεται ως διαδοχική διαδικασία ανάλυσης δεδομένων, όπου κάθε στάδιο αξιοποιεί το αποτέλεσμα του προηγούμενου. Η προσέγγιση βασίζεται στον συνδυασμό περιγραφικής και προγνωστικής ανάλυσης, ώστε να διερευνηθεί η σχέση μεταξύ προτύπων χρήσης και πιθανότητας αποχώρησης.

Η διαδικασία υλοποιείται στα ακόλουθα στάδια:

1. Δημιουργία συνόλου δεδομένων. Κατασκευάζεται προσομοιωμένο σύνολο δεδομένων που αναπαριστά μηνιαία κατανάλωση και οικονομικά στοιχεία πελατών ηλεκτρικής ενέργειας.
2. Μετασχηματισμός δεδομένων σε χαρακτηριστικά συμπεριφοράς. Οι αρχικές μετρήσεις μετατρέπονται σε δείκτες RFM, οι οποίοι συνοψίζουν τη χρονική εγγύτητα χρήσης, τη συχνότητα αλληλεπίδρασης και το συνολικό επίπεδο κατανάλωσης.
3. Τυποποίηση και μείωση διάστασης. Τα χαρακτηριστικά κλιμακώνονται ώστε να είναι συγκρίσιμα και χρησιμοποιούνται τεχνικές μείωσης διάστασης για την οπτική διερεύνηση της δομής των δεδομένων.
4. Τμηματοποίηση πελατών. Εφαρμόζεται αλγόριθμος συσταδοποίησης για τον εντοπισμό ομάδων πελατών με παρόμοια συμπεριφορά κατανάλωσης. Ο αριθμός των ομάδων προσδιορίζεται μέσω κριτηρίων αξιολόγησης.
5. Ορισμός μεταβλητής αποχώρησης. Η αποχώρηση ορίζεται με βάση τη χρονική απόσταση από την τελευταία καταγεγραμμένη δραστηριότητα, μετατρέποντας το πρόβλημα σε δυαδική ταξινόμηση.
6. Αντιμετώπιση ανισορροπίας δεδομένων. Το σύνολο εκπαίδευσης εξισορροπείται ώστε να αποτραπεί η μεροληψία των μοντέλων προς την πλειονοτική κατηγορία.
7. Ανάπτυξη προγνωστικών μοντέλων. Εκπαιδεύονται μοντέλα ταξινόμησης με και χωρίς πληροφορία τμηματοποίησης, ώστε να αξιολογηθεί η συμβολή των ομάδων πελατών στην πρόβλεψη αποχώρησης.
8. Αξιολόγηση αποτελεσμάτων. Τα μοντέλα συγκρίνονται σε ανεξάρτητο σύνολο δεδομένων χρησιμοποιώντας κατάλληλες μετρικές απόδοσης και εξετάζεται η ερμηνεία των παραγόντων κινδύνου.

Η διαδοχική αυτή διαδικασία επιτρέπει την ολοκληρωμένη μελέτη της συμπεριφοράς πελατών, από την αναγνώριση προτύπων χρήσης έως την εκτίμηση πιθανότητας αποχώρησης. Συγκεκριμένα, τα στάδια της μεθοδολογίας αναφέρονται στη εικόνα 1.



Εικόνα 1: *Διάγραμμα ροής υλοποίησης μεθοδολογίας*

Η παραπάνω δομή επιτρέπει τη συστηματική ανάλυση της συμπεριφοράς των πελατών και την εξαγωγή αξιόπιστων συμπερασμάτων με επιχειρησιακή αξία.

3.3 Περιγραφή Δεδομένων

Λόγω περιορισμών που σχετίζονται με την προστασία προσωπικών δεδομένων και τη διαθεσιμότητα πραγματικών ενεργειακών δεδομένων, για την αξιολόγηση της προτεινόμενης μεθοδολογίας χρησιμοποιήθηκε συνθετικό σύνολο δεδομένων που προσομοιώνει τη μηνιαία κατανάλωση πελατών ηλεκτρικής ενέργειας χαμηλής τάσης. Το συνθετικό σύνολο αποτελείται από 34.783 γραμμές, με 3.000 μοναδικούς πελάτες για τις ημερομηνίες 1.1.2023 έως 1.12.2023. Η χρήση προσομοιωμένων δεδομένων επιτρέπει τον πλήρη έλεγχο των χαρακτηριστικών της ανάλυσης, διατηρώντας παράλληλα ρεαλιστικές ιδιότητες που συναντώνται σε πραγματικές εφαρμογές. Κάθε εγγραφή αντιστοιχεί σε μηνιαία δραστηριότητα πελάτη και περιλαμβάνει αναγνωριστικό πελάτη, χρονική σήμανση, επίπεδο κατανάλωσης ενέργειας και αντίστοιχη οικονομική χρέωση. Η παραγωγή των δεδομένων βασίστηκε σε διαφορετικά πρότυπα κατανάλωσης, ώστε να αναπαρασταθεί η ετερογένεια της πελατειακής βάσης. Συγκεκριμένα, δημιουργήθηκαν κατηγορίες πελατών με χαμηλή,

μέση και υψηλή κατανάλωση, με διαφορετική μεταβλητότητα και εποχικότητα χρήσης. Επιπλέον εισήχθη μηχανισμός διακοπής δραστηριότητας για υποσύνολο πελατών, προσομοιώνοντας τη συμπεριφορά αποχώρησης. Με τον τρόπο αυτό το σύνολο δεδομένων περιλαμβάνει τόσο ενεργούς όσο και ανενεργούς πελάτες, επιτρέποντας τη διατύπωση του προβλήματος ως πρόβλημα πρόβλεψης αποχώρησης. Η χρονική διάρκεια του συνόλου δεδομένων καλύπτει διαδοχικούς μήνες καταγραφών, ώστε να αποτυπώνονται πρότυπα χρήσης και μεταβολές συμπεριφοράς. Η δομή αυτή επιτρέπει την εξαγωγή χαρακτηριστικών συμπεριφοράς και την εφαρμογή τεχνικών μηχανικής μάθησης σε συνθήκες αντίστοιχες πραγματικών ενεργειακών δεδομένων. Στο σύνολο δεδομένων ενσωματώθηκαν μεταβλητές που αφορούν τη συμμετοχή των πελατών σε προγράμματα πιστότητας (loyalty schemes), καθώς και τη χρήση πρόσθετων υπηρεσιών (add-on services). Οι παράγοντες αυτοί θεωρούνται ιδιαίτερα σημαντικοί στον ενεργειακό τομέα, καθώς ενισχύουν τη δέσμευση των πελατών και αυξάνουν το κόστος αλλαγής παρόχου. Η συμπερίληψή τους στα μοντέλα churn αναμένεται να βελτιώσει την προγνωστική ικανότητα και να προσφέρει χρήσιμα επιχειρησιακά συμπεράσματα. Ιδιαίτερη έμφαση δόθηκε στη δημιουργία παραγώγων μεταβλητών, όπως τα χαρακτηριστικά Recency, Frequency και Monetary (RFM), καθώς και στη μετατροπή κατηγορικών μεταβλητών σε κατάλληλη αριθμητική μορφή. Η διαδικασία αυτή διασφάλισε την ποιότητα και τη συνοχή του τελικού συνόλου δεδομένων, καθιστώντας το κατάλληλο για εφαρμογή τεχνικών ομαδοποίησης και ανάπτυξη προγνωστικών μοντέλων αποχώρησης πελατών. Στην εικόνα 2 περιγράφονται τα δεδομένα που χρησιμοποιήθηκαν.

Κατηγορία Δεδομένων	**Ονομασία**	**Περιγραφή**	**Τύπος**
Στοιχεία	Customer_id	Μοναδικός κωδικός ταυτοποίησης πελάτη	Συνεχής
Κατανάλωση Ενέργειας	Monthly_kWh	Μηνιαία κατανάλωση ηλεκτρικής ενέργειας	Συνεχής
	Avg_kWh	Μέση μηνιαία κατανάλωση	Συνεχής
	kWh_Change	Ποσοστιαία μεταβολή κατανάλωσης	Συνεχής
	kWh_Variability	Διακύμανση κατανάλωσης	Συνεχής
Τιμολογιακά / Συμβατικά	Tariff_Type	Τύπος τιμολογίου (σταθερό, κυμαινόμενο)	Κατηγορική
	Tariff_Tenure	η διάρκεια ισχύος του προϊόντος που έχει ο πελάτης	Αριθμητική
	Contract_Duration	Συνολική διάρκεια σύμβασης	Αριθμητική
	Remaining_Months	Υπολειπόμενοι μήνες σύμβασης	Αριθμητική

Κατηγορία Δεδομένων	Ονομασία	Περιγραφή	Τύπος
Στοιχεία	Customer_id	Μοναδικός κωδικός ταυτοποίησης πελάτη	Συνεχής
Πληρωμές	Late_Payments	Πλήθος καθυστερημένων πληρωμών	Αριθμητική
	Avg_Bill_Amount	Μέσο ποσό λογαριασμού	Συνεχής
	Debt_Flag	Ένδειξη ύπαρξης ληξιπρόθεσμων οφειλών	Δυαδική
Loyalty Programs	Loyalty_Member	Συμμετοχή σε πρόγραμμα πιστότητας	Δυαδική
Add-On Services	AddOn_Usage	Χρήση πρόσθετων υπηρεσιών	Δυαδική
RFM Metrics	Recency	Ημέρες από την τελευταία πληρωμή	Συνεχής
	Frequency	Πλήθος πληρωμών σε χρονικό διάστημα	Αριθμητική
	Monetary	Συνολική αξία πληρωμών ή κατανάλωσης	Συνεχής
CRM / Συμπεριφορικά	Complaints	Αριθμός παραπόνων	Αριθμητική
	Call_Center_Contacts	Επικοινωνίες με εξυπηρέτηση	Αριθμητική
	eBill_Usage	Χρήση ηλεκτρονικού λογαριασμού	Δυαδική
Δημογραφικά	Region	Γεωγραφική περιοχή	Κατηγορική
	Customer_Type	Οικιακός	Κατηγορική

Εικόνα 2. *Περιγραφή Συνόλου Δεδομένων Πελατών Ηλεκτρικής Ενέργειας*

3.4 Προεπεξεργασία Δεδομένων

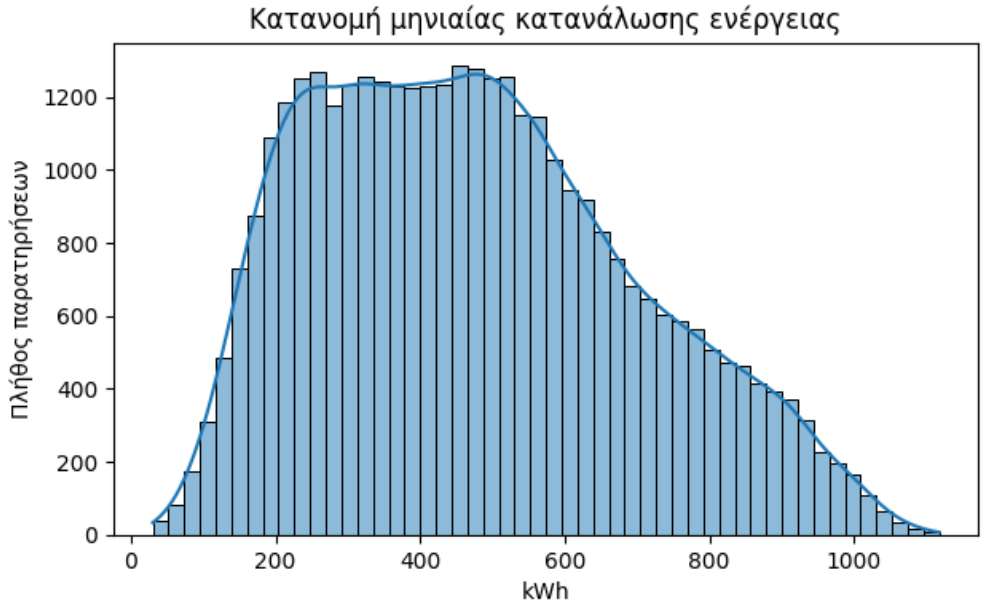
Πριν την ανάλυση πραγματοποιήθηκε προεπεξεργασία των δεδομένων ώστε να καταστούν κατάλληλα για στατιστική επεξεργασία. Τα δεδομένα οργανώθηκαν σε μορφή συναλλαγών, όπου κάθε εγγραφή αντιστοιχεί σε πελάτη και χρονική περίοδο κατανάλωσης. Εφαρμόστηκαν περιορισμοί εγκυρότητας ώστε να αποφευχθούν μη ρεαλιστικές τιμές κατανάλωσης και εξασφαλίστηκε θετικό εύρος τιμών. Στη συνέχεια τα δεδομένα συγκεντρώθηκαν σε επίπεδο πελάτη, μετατρέποντας τις μηνιαίες

παρατηρήσεις σε συνοπτική αναπαράσταση συμπεριφοράς χρήσης. Το στάδιο αυτό εξασφαλίζει ότι οι μεταγενέστερες αναλύσεις βασίζονται σε συνεπή και συγκρίσιμα δεδομένα. Επομένως, το raw dataset έγινε αναλύσιμο.

index	customer_id	date	kwh	bill_amount
0	1	2023-01-01 00:00:00	154.02433225025018	35.102807491161364
1	1	2023-02-01 00:00:00	224.46314628527873	54.088770242257375
2	1	2023-03-01 00:00:00	182.78987395364197	35.563327765025065
3	1	2023-04-01 00:00:00	139.9011227357642	27.099359984205133
4	1	2023-05-01 00:00:00	229.99122805614428	39.42924827913566

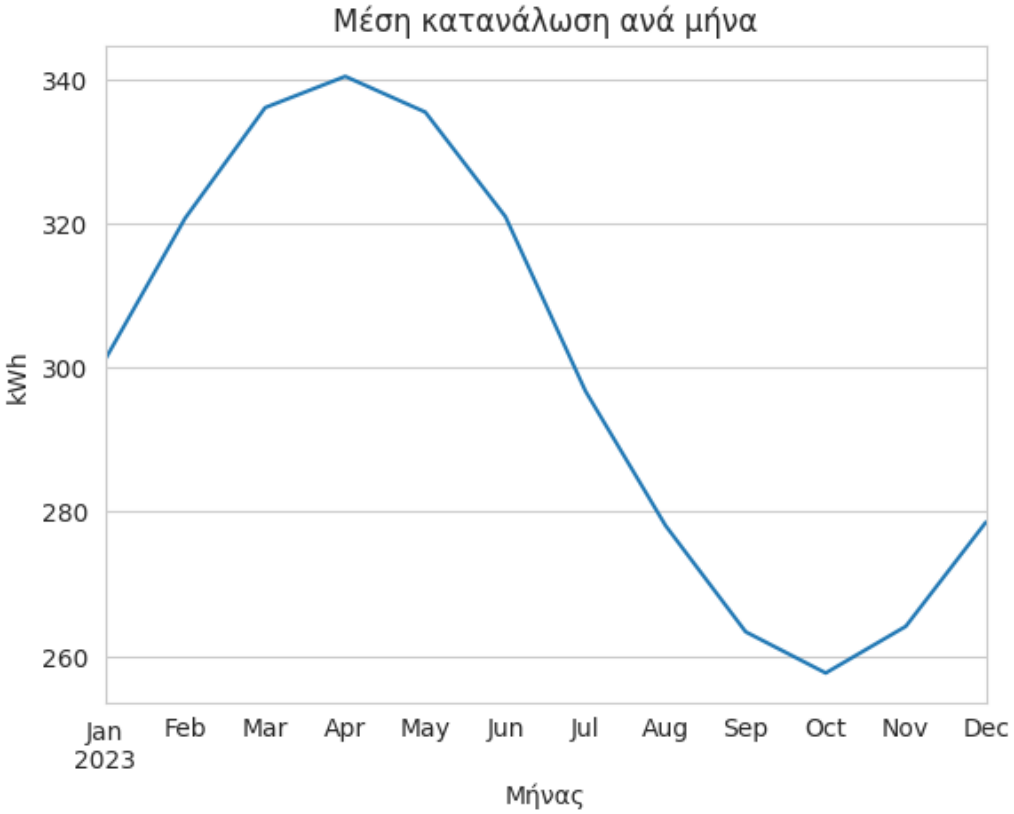
Εικόνα 3. *Μορφή dataset που εισάγεται στο μοντέλο*

Ο πίνακας στην εικόνα 3 παρουσιάζει ενδεικτικές εγγραφές του συνόλου δεδομένων, όπου κάθε γραμμή αντιστοιχεί σε μηνιαία καταγραφή κατανάλωσης πελάτη. Η δομή αυτή επιτρέπει την ανάλυση της συμπεριφοράς χρήσης στον χρόνο και την εξαγωγή χαρακτηριστικών που περιγράφουν τη σχέση πελάτη υπηρεσίας-προϊόντος. Με αυτή τη μορφή πραγματοποιήθηκε η εισαγωγή στο μοντέλο. Για την καλύτερη κατανόηση της κατανομής και των χαρακτηριστικών των δεδομένων πραγματοποιήθηκε διερευνητική ανάλυση. Παρακάτω παρουσιάζονται ορισμένες γραφικές απεικονίσεις του συνόλου δεδομένων, με στόχο την καλύτερη κατανόησή του.



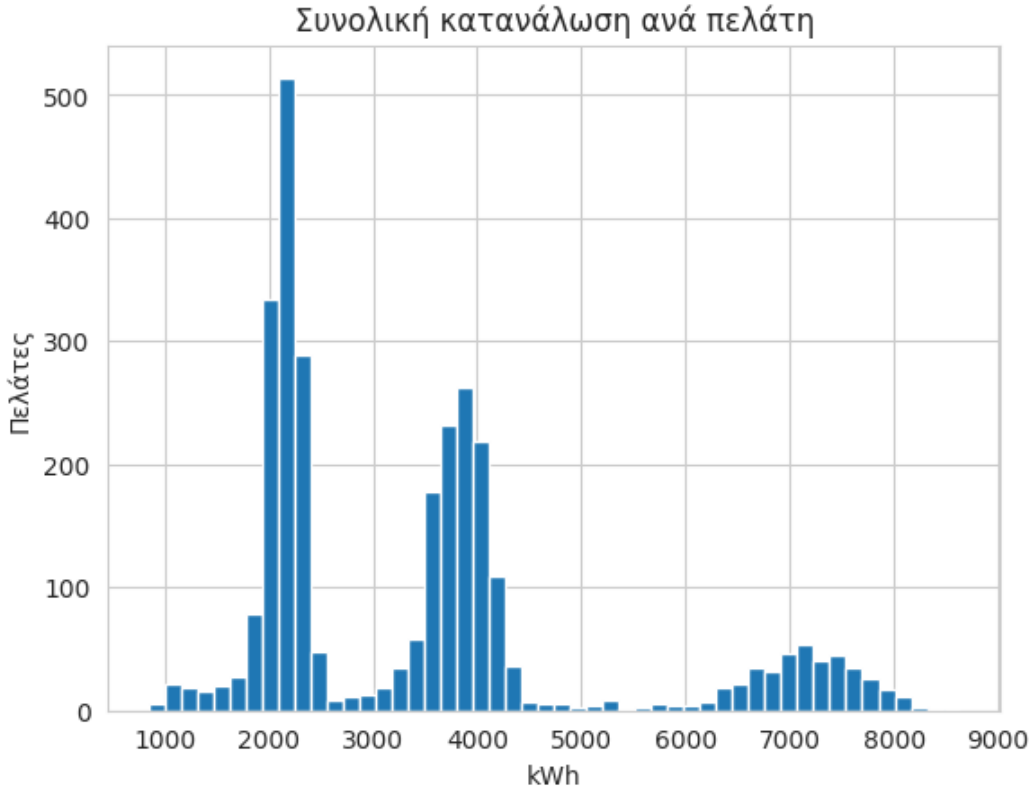
Εικόνα 4: *Κατανομή μηνιαίας κατανάλωσης ενέργειας πελατών*

Η κατανομή της μηνιαίας κατανάλωσης παρουσιάζει ασυμμετρία προς τα δεξιά, υποδηλώνοντας ότι η πλειονότητα των πελατών εμφανίζει χαμηλή έως μέση κατανάλωση, ενώ μικρό ποσοστό καταναλώνει σημαντικά υψηλότερη ενέργεια. Το χαρακτηριστικό αυτό αποτελεί βασική ένδειξη ύπαρξης πελατών διαφορετικής αξίας (Monetary value).



Εικόνα 5: *Μέση κατανάλωση ενέργειας ανά μήνα κατά τη χρονική περίοδο παρατήρησης.*

Παρατηρείται έντονη εποχικότητα στη χρήση ενέργειας, με αυξημένη κατανάλωση κατά τους χειμερινούς μήνες και μειωμένη τους θερινούς. Η συμπεριφορά αυτή προσομοιώνει ρεαλιστικό ενεργειακό προφίλ και επηρεάζει άμεσα τη χρονική απόσταση από την τελευταία δραστηριότητα (Recency).



***Εικόνα 6**. Κατανομή συνολικής κατανάλωσης ανά πελάτη.*

Η συνολική κατανάλωση εμφανίζει πολλαπλές συγκεντρώσεις τιμών, γεγονός που υποδηλώνει την ύπαρξη διακριτών κατηγοριών πελατών. Το εύρημα αυτό ενισχύει την ανάγκη εφαρμογής τεχνικών τμηματοποίησης.

***Εικόνα 7**. Κατανομή χρονικής απόστασης από την τελευταία καταγεγραμμένη δραστηριότητα πελάτη*

Τέλος, η χρονική απόσταση από την τελευταία δραστηριότητα παρουσιάζει διαφοροποίηση μεταξύ των πελατών, με ένα τμήμα αυτών να εμφανίζει σημαντικά μεγαλύτερη αδράνεια. Η πλειονότητα των πελατών εμφανίζει πολύ πρόσφατη δραστηριότητα, ενώ μικρό ποσοστό παρουσιάζει μεγάλη χρονική απόσταση από την τελευταία χρήση. Η μεταβλητή αυτή αποτελεί ισχυρό δείκτη πιθανής αποχώρησης.

Η δημιουργία του συνόλου δεδομένων πραγματοποιήθηκε σε περιβάλλον Python (βλ. Παράρτημα Α, Κώδικας 1).

3.5 Δημιουργία Χαρακτηριστικών - RFM Features Engineerng

Μετά την αρχική κατανόηση των δεδομένων πραγματοποιήθηκε μετασχηματισμός των καταγραφών κατανάλωσης σε χαρακτηριστικά συμπεριφοράς πελατών. Στόχος του σταδίου αυτού είναι η συμπύκνωση της χρονικής πληροφορίας σε δείκτες που περιγράφουν τη σχέση κάθε πελάτη με την υπηρεσία. Για τον σκοπό αυτό εφαρμόστηκε η προσέγγιση RFM, η οποία συνοψίζει τη δραστηριότητα του πελάτη μέσω τριών διαστάσεων. Η χρονική εγγύτητα δραστηριότητας εκφράζει πόσο πρόσφατη είναι η τελευταία καταγεγραμμένη χρήση, η συχνότητα αποτυπώνει τον αριθμό περιόδων κατανάλωσης και η συνολική αξία αντιπροσωπεύει το επίπεδο ενεργειακής χρήσης. Με τον τρόπο αυτό οι πολλαπλές μηνιαίες εγγραφές μετατρέπονται σε μία ενιαία περιγραφή συμπεριφοράς ανά πελάτη.

Πέρα από τους βασικούς δείκτες, υπολογίστηκαν και παράγωγα χαρακτηριστικά που αποτυπώνουν την ένταση χρήσης, ώστε να ενισχυθεί η διακριτική ικανότητα της

ανάλυσης. Ο μετασχηματισμός αυτός μειώνει τη διάσταση των δεδομένων και επιτρέπει την εφαρμογή τεχνικών ομαδοποίησης σε επίπεδο πελάτη αντί επιμέρους χρονικών παρατηρήσεων.

Η διαδικασία δημιουργίας χαρακτηριστικών αποτελεί κρίσιμο στάδιο, καθώς τα προγνωστικά μοντέλα δεν βασίζονται στις αρχικές μετρήσεις αλλά στην αναπαράσταση της συμπεριφοράς που προκύπτει από αυτές. Στην εικόνα 9 μπορούμε να δούμε σε μορφή πίνακα τα δεδομένα στα οποία έγινε η επεξεργασία.

	customer_id	frequency	monetary_kwh	monetary_bill	recency_days	avg_kwh
0	1	12	2316.702024	475.004322	1	193.058502
1	2	12	6522.209918	1172.715117	1	543.517493
2	3	12	3985.235439	795.210908	1	332.102953
3	4	12	3747.569169	724.830578	1	312.297431
4	5	12	2373.367640	498.726061	1	197.780637

***Εικόνα 8**. Preview του πίνακα RFM*

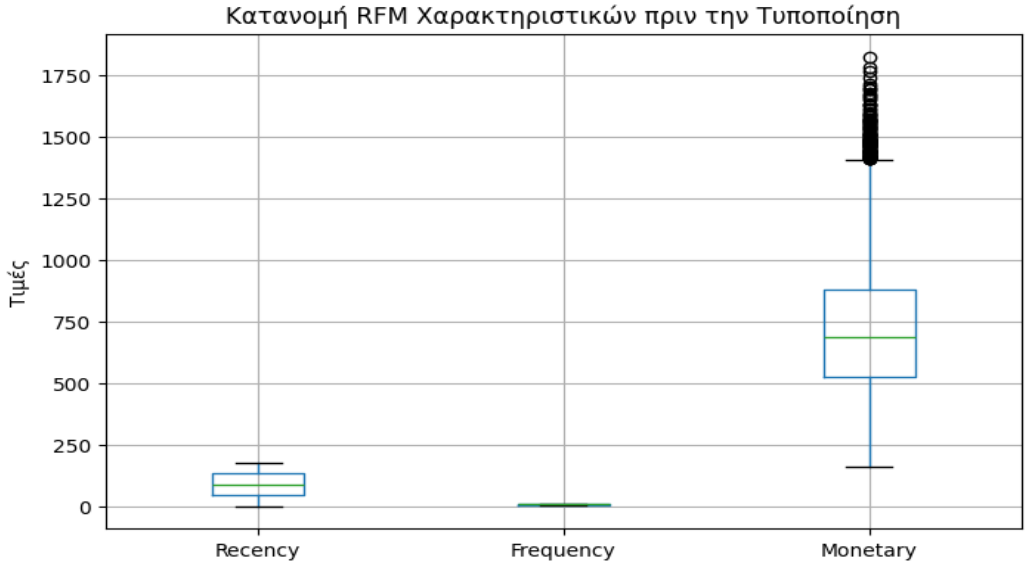
Ο υπολογισμός των χαρακτηριστικών RFM υλοποιήθηκε σε Python (βλ. Παράρτημα Α, Κώδικας 2).

3.6 Τυποποίηση Ομαδοποίησης και Μείωση Διάστασης (Clustering Scaling & PCA)

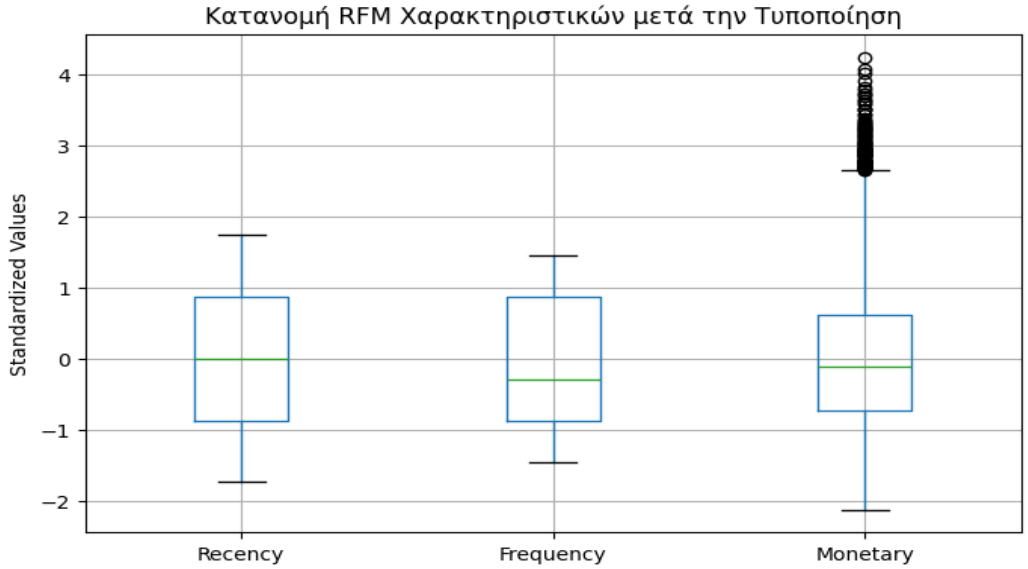
Τα χαρακτηριστικά που προέκυψαν από τη διαδικασία RFM βρίσκονται σε διαφορετικές κλίμακες τιμών, γεγονός που επηρεάζει τη λειτουργία αλγορίθμων που βασίζονται σε αποστάσεις. Στη συγκεκριμένη περίπτωση, η συνολική κατανάλωση εμφανίζει σημαντικά μεγαλύτερο εύρος τιμών σε σχέση με τη συχνότητα δραστηριότητας ή τη χρονική εγγύτητα χρήσης. Χωρίς προσαρμογή της κλίμακας, η μεταβλητή με το μεγαλύτερο εύρος θα κυριαρχούσε στον υπολογισμό αποστάσεων, οδηγώντας σε παραμορφωμένη ομαδοποίηση. Για τον λόγο αυτό εφαρμόστηκε τυποποίηση (StandardScaler) των χαρακτηριστικών, ώστε κάθε μεταβλητή να συμβάλλει ισότιμα στη διαδικασία υπολογισμού ομοιότητας μεταξύ πελατών. Η διαδικασία μετασχηματίζει τις τιμές σε κοινή κλίμακα και επιτρέπει την ουσιαστική σύγκριση διαφορετικών διαστάσεων συμπεριφοράς. Επιπλέον, χρησιμοποιήθηκε τεχνική μείωσης διάστασης με σκοπό την οπτική διερεύνηση της δομής των δεδομένων. Η αναπαράσταση σε χαμηλότερη διάσταση επιτρέπει την απεικόνιση πιθανών ομάδων και παρέχει ένδειξη για την ύπαρξη υποκείμενων προτύπων πριν την εφαρμογή της συσταδοποίησης. Η τυποποίηση αποτελεί απαραίτητο στάδιο πριν την εφαρμογή αλγορίθμων τμηματοποίησης που βασίζονται σε αποστάσεις, καθώς διασφαλίζει ότι η ομαδοποίηση προκύπτει από συνολική συμπεριφορά και όχι από την κυριαρχία μεμονωμένης μεταβλητής. Στην εικόνα 9 παρατηρείται ότι οι μεταβλητές Monetary και

Recency έχουν μεγάλες τιμές, ενώ η Frequency μικρό εύρος. Άρα οι μεταβλητές βρίσκονται σε διαφορετικές κλίμακες. Ο K-Means θα ομαδοποιούσε μόνο τη

Monetary, αυτό ακριβώς είναι και το πρόβλημα. Μετά την τυποποίηση στην εικόνα 13, ο διάμεσοι είναι περίπου στο 0, τα εύρη είναι παρόμοια και καμία μεταβλητή δε «κυριαρχεί». Δηλαδή, κάθε feature συμμετέχει ισότιμα στον υπολογισμό αποστάσεων.



***Εικόνα 9**. BOX PLOT κατανομής RFM χαρακτηριστικών πριν την τυποποίηση*



***Εικόνα 10**. BOX PLOT κατανομής RFM χαρακτηριστικών μετά την τυποποίηση*

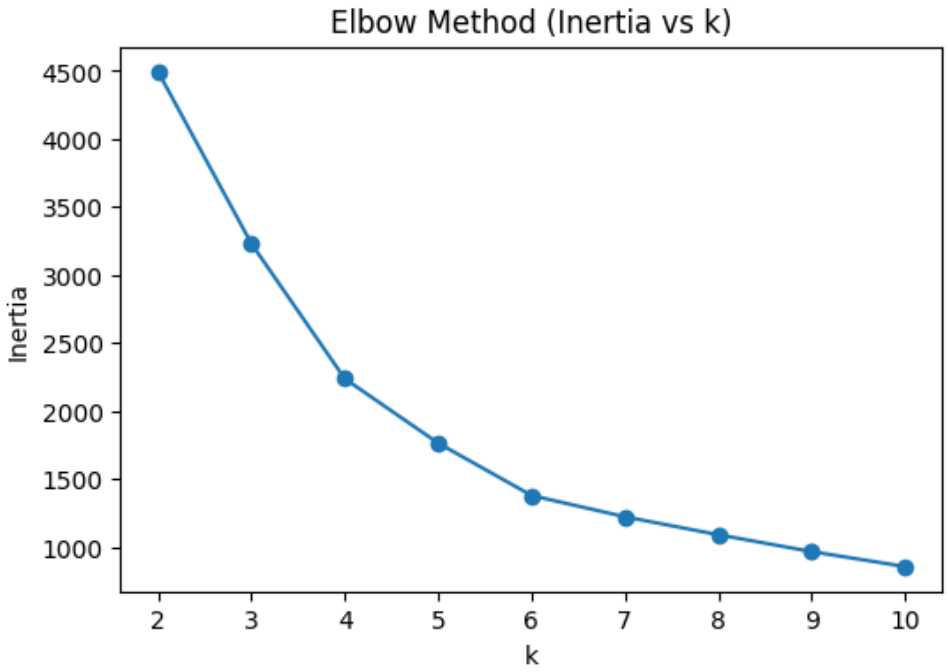
Η ορθότητα της τυποποίησης επιβεβαιώθηκε καθώς οι μετασχηματισμένες μεταβλητές παρουσίασαν μέση τιμή προσεγγιστικά μηδενική και μοναδιαία διασπορά.

Η τυποποίηση και η μείωση διάστασης πραγματοποιήθηκαν σε

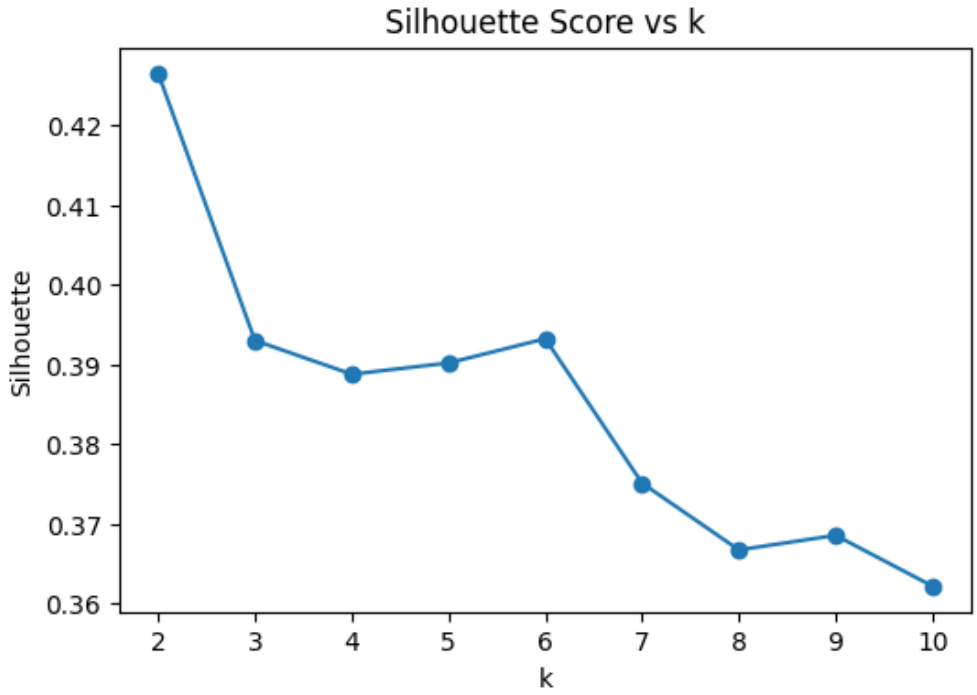
Python. (βλ. Παράρτημα Α, Κώδικας 12).

3.7 Επιλογή αριθμού συστάδων (Elbow & Silhouette)

Μετά την εφαρμογή της διαδικασίας τμηματοποίησης απαιτείται ο προσδιορισμός του κατάλληλου αριθμού συστάδων. Για τον σκοπό αυτό χρησιμοποιήθηκε συνδυαστική αξιολόγηση, ώστε η επιλογή να μην βασίζεται σε αυθαίρετη παραδοχή αλλά σε μετρήσιμα κριτήρια. Για τον προσδιορισμό του βέλτιστου αριθμού συστάδων εφαρμόστηκαν δύο συμπληρωματικές μέθοδοι αξιολόγησης, η μέθοδος του αγκώνα (Elbow Method) και ο δείκτης Silhouette. Η πρώτη βασίστηκε στη μεταβολή της ενδοσυσταδιακής διακύμανσης (inertia) για διαφορετικές τιμές του k, με στόχο τον εντοπισμό του σημείου όπου η μείωση της διακύμανσης παύει να είναι σημαντική. Η δεύτερη μέθοδος υπολόγισε τον μέσο δείκτη Silhouette για κάθε πιθανό αριθμό συστάδων, εκτιμώντας τον βαθμό συνοχής και διαχωρισμού των ομάδων. Τα αποτελέσματα έδειξαν ότι η καμπύλη του Elbow παρουσίασε σαφή μείωση k από 2 εως 4, μετά το k=4 η καμπύλη «ισιώνει». Επομένως από τη μέθοδο Elbow το βέλτιστο είναι k=~4. Ο δείκτης Silhouette έλαβε τη μέγιστη τιμή του επίσης για k = 2, μετά μειώνεται και γίνεται σχετικά σταθερό 2-6, ωστόσο δεν υποστηρίζει ξεκάθαρα το k=6. Αυτό σημαίνει ότι, για k=2 έχουμε πολύ γενικές ομάδες, για k=4 είναι μαθηματικά καλύτερο, ενώ για k=3 υπάρχει καλύτερη ερμηνευσιμότητα (business segmentation).



Εικόνα 11. *Επιλογή αριθμού συστάδων με βάση την αδράνεια (inertia)*



***Εικόνα 12**. Επιλογή αριθμού συστάδων με βάση τον δείκτη Silhouette*

Η επιλογή του αριθμού συστάδων βασίστηκε στον συνδυασμό της μεθόδου Elbow καιτου δείκτη Silhouette. Από το διάγραμμα αδράνειας παρατηρείται σημαντική μείωση μέχρι περίπου k=4, μετά το οποίο η βελτίωση περιορίζεται σημαντικά. Ο δείκτης silhouette παρουσιάζει υψηλότερη τιμή για μικρό αριθμό συστάδων και παραμένει σχετικά σταθερός για k μεταξύ 3 και 6, χωρίς να υποδεικνύει σαφή υπεροχή μεγαλύτερου πλήθους ομάδων. Για λόγους ερμηνευσιμότητας και αποφυγής υπερβολικού κατακερματισμού του πληθυσμού επιλέχθηκαν τελικά k=3 συστάδες, οι οποίες παρέχουν ικανοποιητικό διαχωρισμό και επιτρέπουν ουσιαστική περιγραφή διαφορετικών προτύπων κατανάλωσης.

Η διαδικασία αξιολόγησης υλοποιήθηκε σε Python (βλ.Παράρτημα Α, Κώδικας 13).

3.8 Εφαρμογή K-Means Clustering

Μετά την προεπεξεργασία των χαρακτηριστικών εφαρμόστηκε η μέθοδος συσταδοποίησης K-Means με στόχο τον εντοπισμό ομάδων πελατών που παρουσιάζουν παρόμοια πρότυπα κατανάλωσης και συμπεριφοράς. Η επιλογή του συγκεκριμένου αλγορίθμου βασίστηκε στην ικανότητά του να διαχωρίζει αποτελεσματικά τα δεδομένα σε ομάδες με βάση την ελάχιστη ενδοομαδική απόσταση και τη μέγιστη διαχωρισιμότητα μεταξύ των ομάδων. Η τμηματοποίηση πραγματοποιήθηκε σε επίπεδο πελάτη, αξιοποιώντας τα κανονικοποιημένα χαρακτηριστικά συμπεριφοράς (Recency, Frequency, Monetary), ώστε η σύγκριση να βασίζεται στη συνολική χρήση και όχι σε μεμονωμένες μετρήσεις.

Η διαδικασία ομαδοποίησης αποσκοπεί στη δημιουργία προφίλ πελατών που εμφανίζουν παρόμοια σχέση με την υπηρεσία. Κάθε ομάδα αντιπροσωπεύειδιαφορετικό μοτίβο χρήσης, όπως πελάτες με υψηλή και σταθερή κατανάλωση, πελάτες με μέτρια και εποχική χρήση ή πελάτες με χαμηλή και ασταθή δραστηριότητα. Με τον τρόπο αυτό, οι πολλαπλές χρονικές παρατηρήσεις

μετατρέπονται σε κατηγορίες συμπεριφοράς που μπορούν να χρησιμοποιηθούν τόσο για περιγραφική ανάλυση όσο και για προγνωστικούς σκοπούς.

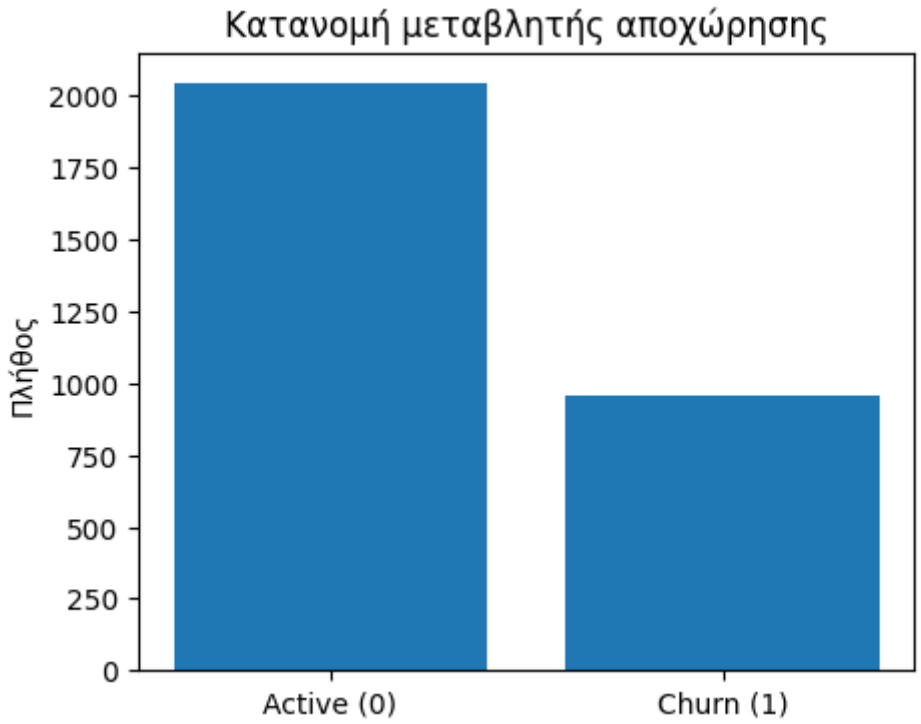
Η εφαρμογή της μεθόδου πραγματοποιήθηκε σε κανονικοποιημένο χώρο χαρακτηριστικών, ώστε η ομαδοποίηση να επηρεάζεται ισότιμα από όλες τις διαστάσεις συμπεριφοράς. Για την αρχικοποίηση των κέντρων χρησιμοποιήθηκε η τεχνική k-means++, η οποία βελτιώνει τη σταθερότητα των αποτελεσμάτων. Η διαδικασία επαναλήφθηκε για διαφορετικές τιμές του αριθμού συστάδων, ενώ η τελική επιλογή του k έγινε με βάση τα κριτήρια Elbow και Silhouette, τα οποία παρουσιάζονται στην επόμενη ενότητα. Οι προκύπτουσες ομάδες χρησιμοποιούνται τόσο για την περιγραφική ερμηνεία των πελατών όσο και ως πρόσθετη πληροφορία στα προγνωστικά μοντέλα αποχώρησης.

Η διαδικασία συσταδοποίησης υλοποιήθηκε σε Python (βλ. Παράρτημα Α, Κώδικας 4).

3.9 Ορισμός Μεταβλητής Αποχώρησης - Churn

Μετά την ερμηνεία των συστάδων ορίστηκε μεταβλητή αποχώρησης, ώστε η συμπεριφορική ανάλυση να συνδεθεί με πρόβλημα πρόβλεψης. Η αποχώρηση προσομοιώθηκε πιθανοτικά με βάση τα χαρακτηριστικά συμπεριφοράς των πελατών. Η πιθανότητα αποχώρησης αυξανόταν με τη χρονική απόσταση από την τελευταία δραστηριότητα και μειωνόταν με τη συχνότητα και την ένταση κατανάλωσης. Το κατώφλι ταξινόμησης ρυθμίστηκε ώστε το σύνολο δεδομένων να παρουσιάζει έντονη ανισορροπία, προσεγγίζοντας ρεαλιστικές συνθήκες αγοράς. Συγκεκριμένα, υπολογίστηκε για κάθε πελάτη ο αριθμός ημερών από την τελευταία του συναλλαγή έως την ημερομηνία αναφοράς του συνόλου δεδομένων. Εάν η χρονική αυτή απόσταση υπερέβαινε ένα προκαθορισμένο όριο (π.χ. 90 ημέρες χωρίς καμία δραστηριότητα), ο πελάτης χαρακτηριζόταν ως «Churner» (churn = 1). Αντίθετα, πελάτες που είχαν πραγματοποιήσει συναλλαγές εντός του ορίου θεωρήθηκαν «Active» (churn = 0). Το όριο αυτό επιλέχθηκε με βάση τη μέση συχνότητα χρήσης της υπηρεσίας και την επιχειρησιακή λογική του παρόχου, ώστε να αποτυπώνει ρεαλιστικά τη διακοπή σχέσης. Επιπλέον, εξετάστηκε η συνέπεια του ορισμού με τα αποτελέσματα της τμηματοποίησης. Παρατηρήθηκε ότι οι πελάτες που ανήκουν στις συστάδες χαμηλής δραστηριότητας (χαμηλή συχνότητα και υψηλή χρονική απόσταση) εμφάνιζαν σημαντικά υψηλότερα ποσοστά αποχώρησης, ενώ οι πελάτες των συστάδων υψηλής αξίας διατηρούσαν χαμηλό δείκτη αποχώρησης. Η συσχέτιση αυτή επιβεβαιώνει ότι η μεταβλητή αποχώρησης αποτυπώνει ουσιαστικά τη συμπεριφορική διαφοροποίηση που εντοπίστηκε μέσω του clustering.

Η τελική μεταβλητή αποχώρησης χρησιμοποιείται στη συνέχεια ως στόχος (target variable) για την εκπαίδευση των προγνωστικών μοντέλων, επιτρέποντας την ποσοτική εκτίμηση της πιθανότητας αποχώρησης κάθε πελάτη με βάση τα χαρακτηριστικά χρήσης του.



***Εικόνα 13**. Κατανομή πελατών ως προς την κατάσταση δραστηριότητας (ενεργοί – αποχωρήσαντες).*

***Εικόνα 14**. Ποσοστιαία κατανομή ενεργών και αποχωρησάντων πελατών*

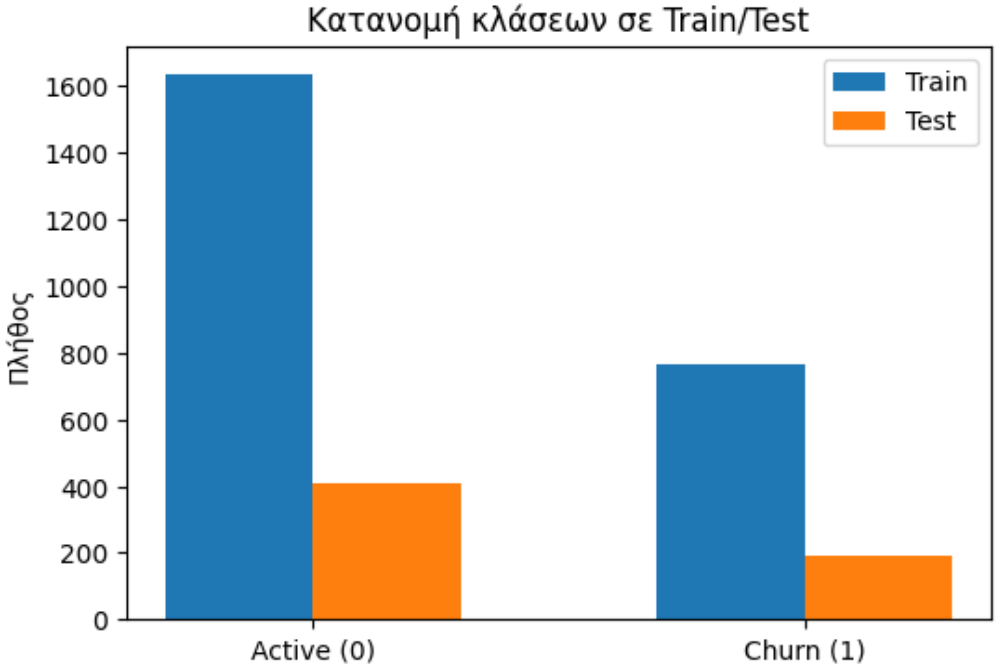
Από το Chart Bar και Pie Chart φαίνεται ότι οι περισσότεροι πελάτες είναι Active και πολύ μικρό ποσοστό είναι Churn. Άρα το dataset είναι imbsalanced classification problem. Επομένως, παρατηρείται έντονη ανισορροπία κλάσεων, καθώς η κατηγορία αποχώρησης υπερτερεί σημαντικά της κατηγορίας ενεργών πελατών. Για τον λόγο αυτό εφαρμόστηκε τεχνική εξισορρόπησης δεδομένων στο σύνολο εκπαίδευσης.

Ο ορισμός της μεταβλητής αποχώρησης υλοποιήθηκε σε Python (βλ. Παράρτημα Α, Κώδικας 3).

3.10 Διαχωρισμός Δεδομένων (Train/Test Split)

Πριν την εκπαίδευση των προγνωστικών μοντέλων απαιτείται ο διαχωρισμός του συνόλου δεδομένων σε ανεξάρτητα υποσύνολα εκπαίδευσης και αξιολόγησης. Η διαδικασία αυτή είναι απαραίτητη ώστε η αξιολόγηση της απόδοσης των μοντέλων να πραγματοποιείται σε δεδομένα που δεν έχουν χρησιμοποιηθεί κατά την εκπαίδευση, επιτρέποντας έτσι την εκτίμηση της ικανότητας γενίκευσης. Για τον σκοπό αυτό το σύνολο δεδομένων χωρίστηκε σε σύνολο εκπαίδευσης (training set) και σύνολο ελέγχου (test set). Το σύνολο εκπαίδευσης χρησιμοποιείται για την προσαρμογή των παραμέτρων των μοντέλων, ενώ το σύνολο ελέγχου χρησιμοποιείται αποκλειστικά για την τελική αξιολόγηση της προγνωστικής τους απόδοσης. Η αναλογία διαχωρισμού επιλέχθηκε ώστε το μεγαλύτερο μέρος των δεδομένων να αξιοποιείται για την εκμάθηση των προτύπων συμπεριφοράς, διατηρώντας παράλληλα επαρκή αριθμό παρατηρήσεων για αξιόπιστη αξιολόγηση. Ο διαχωρισμός πραγματοποιήθηκε με τυχαίο τρόπο, διατηρώντας την αναλογία των κλάσεων μεταξύ των δύο συνόλων (stratified split), ώστε το ποσοστό αποχωρήσεων να παραμένει αντιπροσωπευτικό και στα δύο υποσύνολα.

Η εφαρμογή του διαχωρισμού πριν από οποιαδήποτε διαδικασία εξισορρόπησης ή κανονικοποίησης είναι κρίσιμη, καθώς αποτρέπει τη διαρροή πληροφορίας από το σύνολο ελέγχου προς το σύνολο εκπαίδευσης. Με τον τρόπο αυτό εξασφαλίζεται ότι η αξιολόγηση των μοντέλων αντανακλά ρεαλιστικές συνθήκες εφαρμογής σε νέα, άγνωστα δεδομένα.



Εικόνα 15*.Κατανομή των κλάσεων αποχώρησης στα σύνολα εκπαίδευσης και ελέγχου μετά τον διαχωρισμό δεδομένων. Παρατηρείται διατήρηση της αναλογίας των κλάσεων (stratified split).*

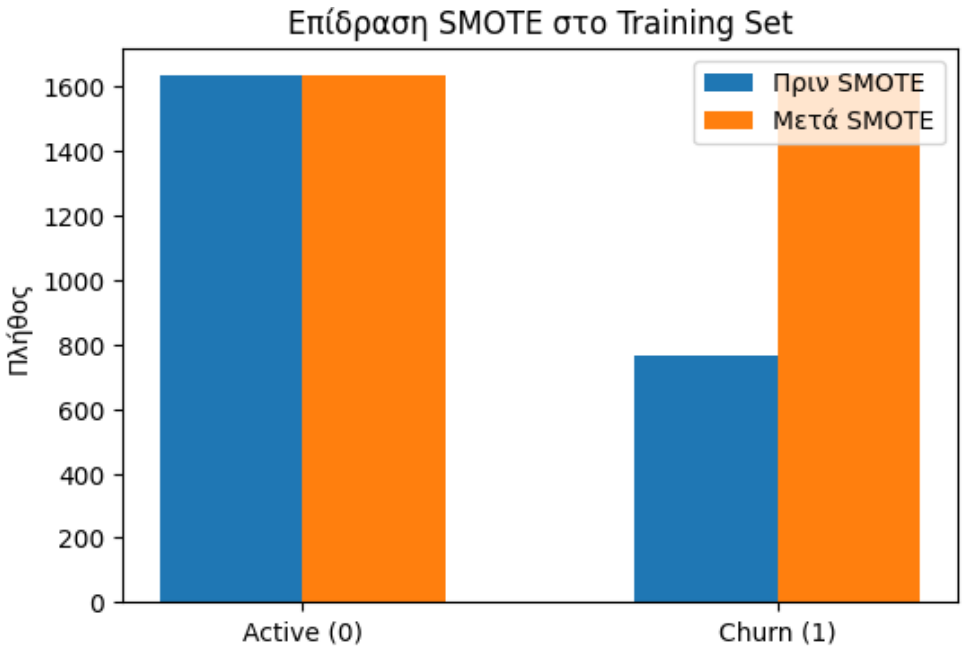
Στο Σχήμα Χ παρουσιάζεται η κατανομή των κλάσεων στα σύνολα εκπαίδευσης και ελέγχου. Παρατηρείται ότι οι αναλογίες ενεργών και αποχωρησάντων πελατών παραμένουν πρακτικά ίδιες και στα δύο σύνολα, γεγονός που επιβεβαιώνει ότι ο διαχωρισμός πραγματοποιήθηκε με διαστρωματωμένη δειγματοληψία (stratified sampling).

3.11 Εξισορρόπηση Δεδομένων με SMOTE

Στο πρόβλημα πρόβλεψης αποχώρησης παρατηρείται ανισορροπία μεταξύ των κλάσεων, καθώς ο αριθμός των πελατών που παραμένουν ενεργοί διαφέρει σημαντικά από τον αριθμό των πελατών που αποχωρούν. Η ανισορροπία αυτή μπορεί να επηρεάσει αρνητικά την εκπαίδευση των μοντέλων μηχανικής μάθησης, οδηγώντας σε μεροληψία υπέρ της πλειοψηφικής κατηγορίας. Ως αποτέλεσμα, ένα μοντέλο θα μπορούσε να εμφανίζει υψηλή συνολική ακρίβεια αλλά να αποτυγχάνει στον εντοπισμό των πελατών που πρόκειται να αποχωρήσουν, οι οποίοι αποτελούν και το βασικό αντικείμενο ενδιαφέροντος.

Για την αντιμετώπιση του προβλήματος εφαρμόστηκε η τεχνική **SMOTE (Synthetic Minority Over-sampling Technique)** στο σύνολο εκπαίδευσης. Η μέθοδος δημιουργεί συνθετικά δείγματα της μειοψηφικής κατηγορίας μέσω παρεμβολής μεταξύ γειτονικών παρατηρήσεων στον χώρο των χαρακτηριστικών. Με τον τρόπο αυτό αποφεύγεται η απλή αντιγραφή δεδομένων και επιτυγχάνεται καλύτερη αναπαράσταση της συμπεριφοράς των πελατών που αποχωρούν.

Η εφαρμογή της τεχνικής πραγματοποιήθηκε αποκλειστικά στο σύνολο εκπαίδευσης και όχι στο σύνολο ελέγχου, ώστε η αξιολόγηση των μοντέλων να βασίζεται σε πραγματικά δεδομένα και να μην επηρεάζεται από τεχνητή ενίσχυση της μειοψηφικής κατηγορίας.



***Εικόνα 16**. Κατανομή των κλάσεων στο σύνολο εκπαίδευσης πριν και μετά την εφαρμογή της τεχνικής SMOTE. Παρατηρείται εξισορρόπηση της μειοψηφικής κατηγορία.*

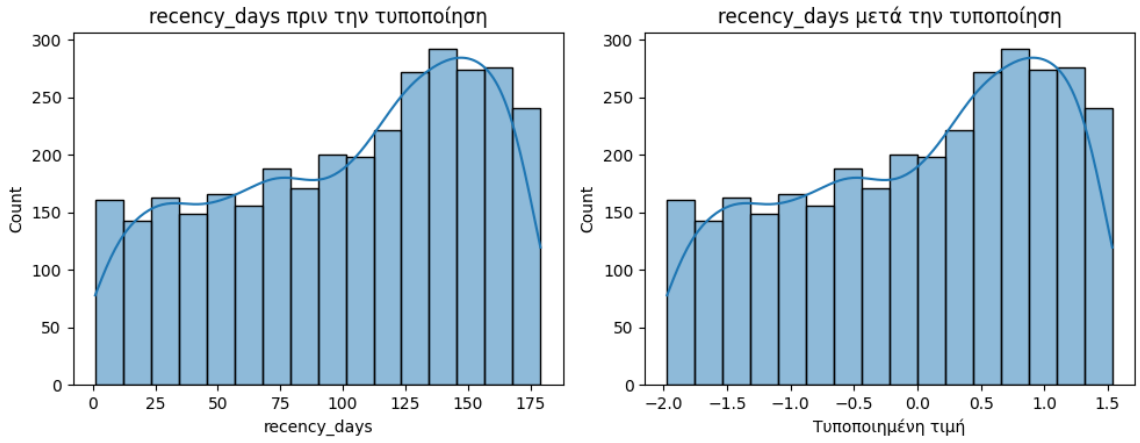
Μετά την εφαρμογή του SMOTE οι δύο κατηγορίες εμφανίζουν παρόμοιο αριθμό παρατηρήσεων, γεγονός που επιτρέπει στα μοντέλα να μάθουν αποτελεσματικότερα τα πρότυπα συμπεριφοράς που σχετίζονται με την αποχώρηση. Παρατηρείται ότι μετά την εφαρμογή της εξισορρόπησης οι δύο κατηγορίες εμφανίζουν παρόμοιο πλήθος παρατηρήσεων, πριν το SMOTE πολλοί Active (0) και πολύ λίγοι Churn (1), έτσι το μοντέλο θα μάθαινε να προβλέπει μόνο Active. Μετά το SMOTE έχουμε ίδιο αριθμό Active και Churn, άρα το μοντέλο πλέον μπορεί να μάθει το μοτίβο αποχώρησης. Επομένως, το γράφημα αποδεικνύει ότι η προεπεξεργασία ήταν απαραίτητη για το classification και πλέον επιτρέπει την αξιόπιστη εκπαίδευση των προγνωστικών μοντέλων. Το ισορροπημένο σύνολο δεδομένων χρησιμοποιείται στη συνέχεια για την εκπαίδευση και σύγκριση διαφορετικών αλγορίθμων πρόβλεψης αποχώρησης.

Η εξισορρόπηση των δεδομένων πραγματοποιήθηκε σε Python (βλ. Παράρτημα Α, Κώδικας 6).

3.12 Εκπαίδευση Προγνωστικών Μοντέλων

3.12.1 Τυποποίηση Πρόβλεψης (Prediction Scaling)

Πριν από την εκπαίδευση των προγνωστικών μοντέλων και μετά την εφαρμογή του SMOTE πραγματοποιήθηκε τυποποίηση των χαρακτηριστικών με τη μέθοδο StandardScaler. Η τυποποίηση εφαρμόστηκε αφού πρώτα δημιουργήθηκαν τα συνθετικά δείγματα, ώστε οι αποστάσεις που χρησιμοποιεί ο αλγόριθμος SMOTE να υπολογιστούν στον αρχικό χώρο δεδομένων και όχι σε κανονικοποιημένο. Στη συνέχεια οι ίδιες παράμετροι τυποποίησης εφαρμόστηκαν στο σύνολο ελέγχου για λόγους συνέπειας. Τα χαρακτηριστικά του συνόλου δεδομένων εκφράζονται σε διαφορετικές μονάδες μέτρησης (π.χ. ημέρες, κιλοβατώρες και πλήθος συναλλαγών) και παρουσιάζουν διαφορετικά εύρη τιμών. Χωρίς τυποποίηση, μεταβλητές με μεγαλύτερο αριθμητικό εύρος θα μπορούσαν να επηρεάσουν δυσανάλογα τη διαδικασία εκπαίδευσης, ιδιαίτερα σε αλγορίθμους που βασίζονται σε αποστάσεις ή σε γραμμικούς συνδυασμούς μεταβλητών. Η σειρά αυτή διασφαλίζει αφενός τη σωστή δημιουργία συνθετικών παρατηρήσεων και αφετέρου τη σταθερή εκπαίδευση των προγνωστικών μοντέλων.



Εικόνα 17. *Κατανομή της μεταβλητής recency πριν και μετά την τυποποίηση με StandardScaler. Παρατηρείται διατήρηση της μορφής της κατανομής και μεταφορά της γύρω από το μηδέν.*

Όπως παρατηρείται στην εικόνα, πριν την τυποποίηση το χαρακτηριστικό παρουσιάζει μεγάλο εύρος τιμών, ενώ μετά την εφαρμογή της μετασχηματίζεται ώστε να εκφράζεται ως απόσταση από τον μέσο όρο. Η κατανομή δεν μεταβάλλει το σχήμα της, αλλά μεταφέρεται γύρω από το μηδέν και κλιμακώνεται ως προς την τυπική απόκλιση. Με τον τρόπο αυτό διατηρείται η πληροφορία της μεταβλητής, ενώ εξασφαλίζεται συγκρισιμότητα μεταξύ χαρακτηριστικών διαφορετικής κλίμακας. Η τυποποίηση πραγματοποιήθηκε με τη μέθοδο StandardScaler, η οποία αφαιρεί τη μέση τιμή κάθε χαρακτηριστικού και το διαιρεί με την τυπική του απόκλιση, μετατρέποντας τις τιμές σε κατανομή με μέση τιμή μηδέν και διασπορά ίση με ένα. Η διαδικασία εφαρμόστηκε αποκλειστικά στα δεδομένα εκπαίδευσης και οι ίδιες παράμετροι χρησιμοποιήθηκαν για τον μετασχηματισμό του συνόλου ελέγχου,

αποτρέποντας διαρροή πληροφορίας (data leakage). Η κανονικοποίηση αυτή συμβάλλει στη σταθερότερη εκπαίδευση των μοντέλων και στη βελτίωση της ικανότητας γενίκευσης.

3.12.2 Υλοποίηση Εκπαίδευσης Προγνωστικών Μοντέλων

Μετά την προεπεξεργασία και την εξισορρόπηση του συνόλου εκπαίδευσης πραγματοποιήθηκε εκπαίδευση προγνωστικών μοντέλων με στόχο την εκτίμηση της πιθανότητας αποχώρησης πελατών. Η διαδικασία βασίστηκε σε εποπτευόμενη μάθηση, όπου τα χαρακτηριστικά συμπεριφοράς, όπως η συχνότητα κατανάλωσης (R), η χρονική απόσταση από την τελευταία δραστηριότητα (F) και το ύψος της κατανάλωσης (M), χρησιμοποιούνται για την πρόβλεψη της μεταβλητής αποχώρησης. Με τον τρόπο αυτό επιχειρείται η αναγνώριση προτύπων που διαφοροποιούν τους ενεργούς από τους πελάτες που εγκαταλείπουν την υπηρεσία. Για τη διερεύνηση διαφορετικών προσεγγίσεων επιλέχθηκαν δύο αλγόριθμοι με διαφορετική φιλοσοφία λειτουργίας. Η Λογιστική Παλινδρόμηση και το τυχαίο Δάσος Αποφάσεων. Η επιλογή αυτή επιτρέπει τη σύγκριση ενός γραμμικού και ερμηνεύσιμου μοντέλου με ένα μη γραμμικό μοντέλο υψηλότερης εκφραστικής ικανότητας.

Η **Λογιστική Παλινδρόμηση (Logistic Regression),** αποτελεί γραμμικό μοντέλο ταξινόμησης που εκτιμά την πιθανότητα αποχώρησης μέσω γραμμικού συνδυασμού των χαρακτηριστικών. Το μοντέλο επιλέχθηκε λόγω της ερμηνευσιμότητάς του, καθώς επιτρέπει την κατανόηση της επίδρασης κάθε μεταβλητής στην τελική πρόβλεψη.

Το **Random Forest** (**Τυχαίο Δάσος**) αποτελεί σύνολο δέντρων αποφάσεων και μπορεί να αποτυπώσει πολύπλοκες μη γραμμικές σχέσεις μεταξύ των χαρακτηριστικών. Η χρήση του επιτρέπει την εξέταση πιθανών αλληλεπιδράσεων μεταξύ μεταβλητών που δεν μπορούν να εκφραστούν από γραμμικά μοντέλα.

Για την αποφυγή υπερπροσαρμογής εφαρμόστηκε διασταυρούμενη επικύρωση (Cross-Validation) στο σύνολο εκπαίδευσης. Η διασταυρούμενη επικύρωση χρησιμοποιείται για τη μείωση της υπερπροσαρμογής και την αξιόπιστη εκτίμηση της απόδοσης (James et al., 2021). Η διαδικασία διαχωρίζει επαναληπτικά τα δεδομένα σε επιμέρους υποσύνολα εκπαίδευσης και επικύρωσης, επιτρέποντας την αξιολόγηση της σταθερότητας του μοντέλου και τη μείωση της εξάρτησης από έναν μόνο διαχωρισμό δεδομένων.

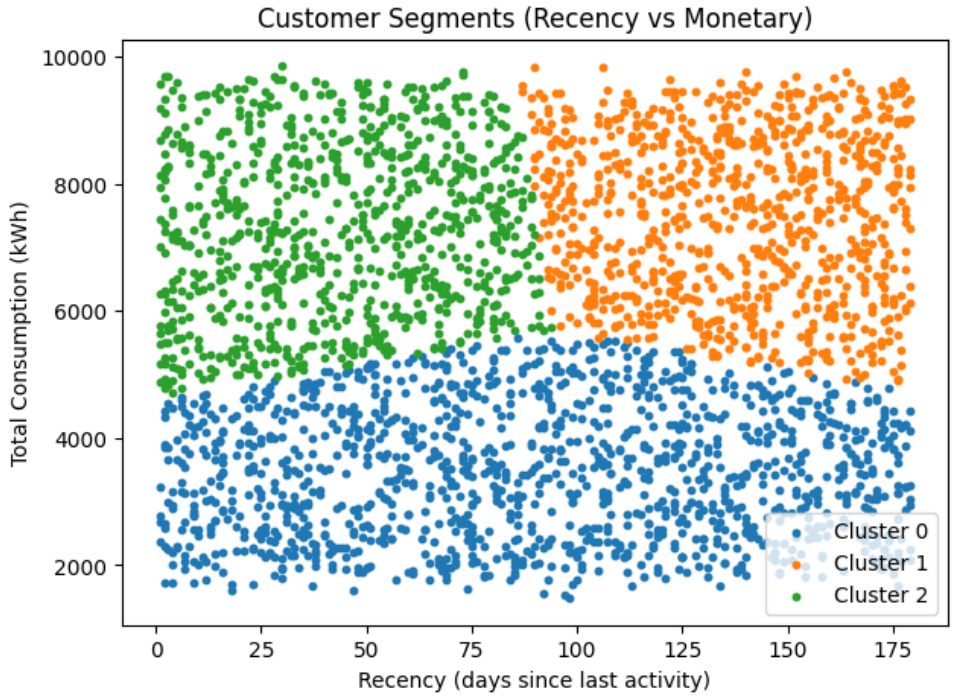
Οι υπερπαράμετροι των μοντέλων ρυθμίστηκαν μέσω πλέγματος αναζήτησης (grid search). Η διαδικασία εξετάζει συνδυασμούς τιμών παραμέτρων και επιλέγει εκείνον που προσφέρει τη βέλτιστη απόδοση βάσει της διασταυρούμενης επικύρωσης. Με τον τρόπο αυτό τα μοντέλα δεν εκπαιδεύονται απλώς μία φορά, αλλά βελτιστοποιούνται συστηματικά ώστε να επιτευχθεί καλύτερη ικανότητα γενίκευσης.

Η εκπαίδευση των μοντέλων υλοποιήθηκε σε Python (βλ. Παράρτημα Α, Κώδικας 7).

Κεφάλαιο 4: Αποτελέσματα και Ανάλυση

4.1 Οπτικοποίηση και Προφίλ Συστάδων

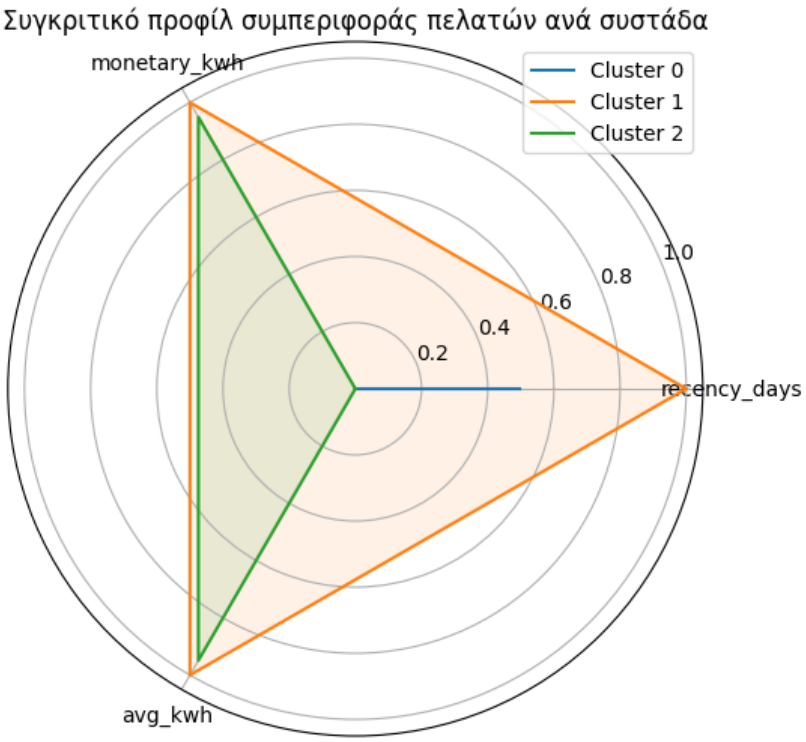
Μετά την εφαρμογή του αλγορίθμου K-Means και τον προσδιορισμό του βέλτιστου αριθμού συστάδων, πραγματοποιήθηκε οπτικοποίηση των αποτελεσμάτων προκειμένου να κατανοηθεί η συμπεριφορική διαφοροποίηση των πελατών. Η ανάλυση βασίστηκε στα χαρακτηριστικά RFM, τα οποία αποτυπώνουν τη χρονική εγγύτητα χρήσης (Recency), τη συχνότητα δραστηριότητας (Frequency) και το επίπεδο κατανάλωσης (Monetary). Για την αρχική κατανόηση του διαχωρισμού των πελατών χρησιμοποιήθηκε διάγραμμα διασποράς μεταξύ της χρονικής απόστασης από την τελευταία δραστηριότητα και της συνολικής κατανάλωσης.



***Εικόνα 18**: Κατανομή πελατών στον χώρο Recency–Monetary ανά συστάδα*

Το διάγραμμα δείχνει ότι οι πελάτες οργανώνονται σε διακριτές περιοχές του χώρου χαρακτηριστικών. Παρατηρείται ότι ένα τμήμα πελατών συγκεντρώνεται σε υψηλές τιμές κατανάλωσης και χαμηλή χρονική απόσταση από την τελευταία δραστηριότητα, υποδηλώνοντας ενεργή χρήση της υπηρεσίας. Αντίθετα, άλλη ομάδα εμφανίζει υψηλή χρονική απόσταση, γεγονός που υποδηλώνει μειωμένη ή διακοπείσα δραστηριότητα. Παράλληλα, εντοπίζεται και ομάδα πελατών χαμηλής κατανάλωσης ανεξάρτητα από τον χρόνο τελευταίας χρήσης. Η απεικόνιση αυτή επιβεβαιώνει ότι τα χαρακτηριστικά κατανάλωσης επαρκούν για τον διαχωρισμό διαφορετικών προτύπων συμπεριφοράς.

Για την ποιοτική ερμηνεία των συστάδων υπολογίστηκαν οι κανονικοποιημένες μέσες τιμές των χαρακτηριστικών RFM για κάθε ομάδα και παρουσιάστηκαν σε ακτινικό διάγραμμα.



Εικόνα 19. *Συγκριτικό προφίλ συμπεριφοράς πελατών ανά συστάδα βάσει των χαρακτηριστικών RFM*

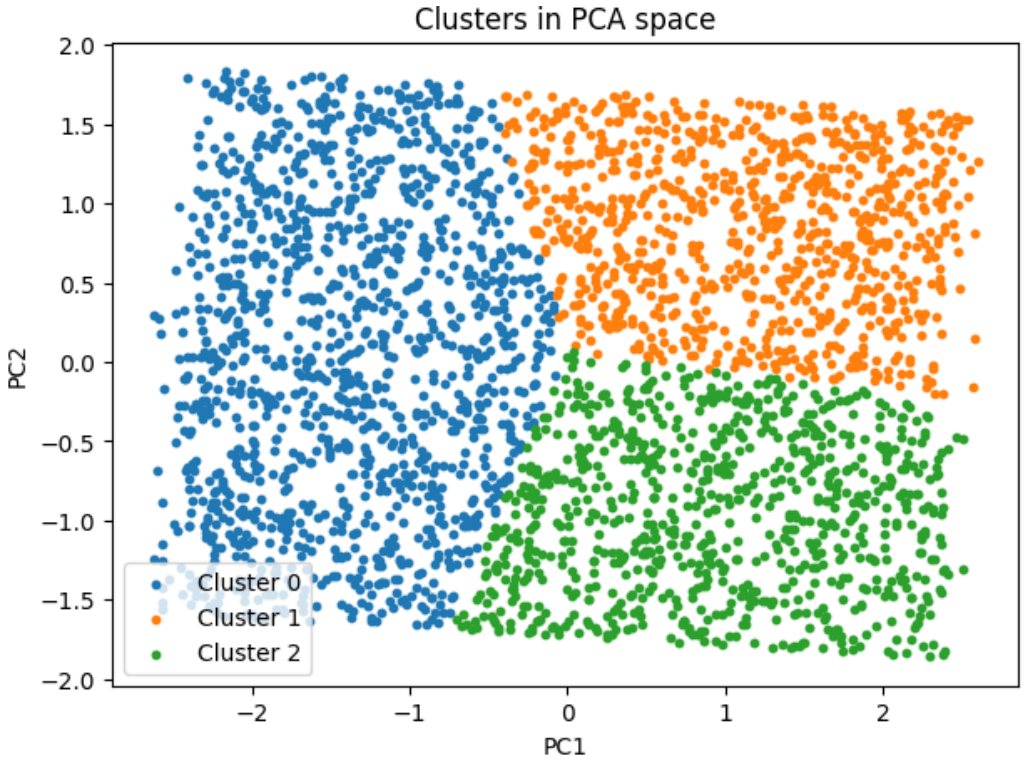
Το radar διάγραμμα επιτρέπει την άμεση σύγκριση των προτύπων

χρήσης μεταξύ των συστάδων. Παρατηρείται ότι:

- Μία ομάδα εμφανίζει υψηλή πρόσφατη δραστηριότητα και υψηλή κατανάλωση
- Μία ομάδα παρουσιάζει παλαιά δραστηριότητα αλλά σημαντική κατανάλωση
- Μία ομάδα χαρακτηρίζεται από χαμηλή κατανάλωση και περιορισμένη χρήση

Η οπτικοποίηση αυτή αναδεικνύει τη διαφορετική συμπεριφορά χρήσης μεταξύ των πελατών και μετατρέπει την αριθμητική τμηματοποίηση σε ερμηνεύσιμα προφίλ χρηστών.

Για την επαλήθευση της ποιότητας του διαχωρισμού εφαρμόστηκε Ανάλυση Κύριων Συνιστωσών (Principal Component Analysis – PCA) και προβολή των δεδομένων σε δισδιάστατο χώρο.



***Εικόνα 19**: Απεικόνιση των συστάδων στον χώρο των δύο πρώτων κύριων συνιστωσών*

Η απεικόνιση δείχνει ότι οι συστάδες παραμένουν διακριτές ακόμη και μετά τη μείωση διαστάσεων. Αυτό σημαίνει ότι ο διαχωρισμός δεν οφείλεται σε μεμονωμένο χαρακτηριστικό αλλά σε συνδυασμό μεταβλητών συμπεριφοράς. Οι ομάδες εμφανίζουν σαφή γεωμετρικό διαχωρισμό, γεγονός που επιβεβαιώνει ότι η τμηματοποίηση αποτυπώνει πραγματικές διαφοροποιήσεις χρήσης.

Ο συνδυασμός των παραπάνω οπτικοποιήσεων επιτρέπει την πλήρη κατανόηση της τμηματοποίησης. Το διάγραμμα Recency-Monetary παρουσιάζει τον φυσικό διαχωρισμό στο χώρο κατανάλωσης, το radar chart παρέχει ερμηνεία της συμπεριφοράς κάθε ομάδας και η προβολή PCA επιβεβαιώνει τη σταθερότητα του διαχωρισμού σε χαμηλότερη διάσταση.

Με βάση τα παραπάνω, η τμηματοποίηση δεν αποτελεί απλώς μαθηματικό αποτέλεσμα αλλά αντιστοιχεί σε πραγματικά πρότυπα χρήσης, γεγονός που επιτρέπει τη σύνδεση των συστάδων με την πιθανότητα αποχώρησης των πελατών.

Παρότι η τμηματοποίηση αποκαλύπτει διακριτά πρότυπα συμπεριφοράς, από μόνη της δεν παρέχει δυνατότητα πρόβλεψης μελλοντικής αποχώρησης πελατών. Ωστόσο, η διαφοροποίηση στη χρήση της υπηρεσίας που παρατηρήθηκε μεταξύ των συστάδων υποδηλώνει ότι η πιθανότητα εγκατάλειψης δεν είναι ίδια για όλους τους χρήστες. Πελάτες με χαμηλή πρόσφατη δραστηριότητα και μειωμένη κατανάλωση εμφανίζουν ενδείξεις απομάκρυνσης από την υπηρεσία, ενώ πελάτες με συχνή και έντονη χρήση παρουσιάζουν μεγαλύτερη πιθανότητα παραμονής.

Για τον λόγο αυτό, η τμηματοποίηση χρησιμοποιείται ως βάση για τη διαμόρφωση ενός προβλήματος πρόβλεψης. Στο επόμενο στάδιο ορίζεται μεταβλητή αποχώρησης, η οποία μετατρέπει τη συμπεριφορική ανάλυση σε εποπτευόμενο

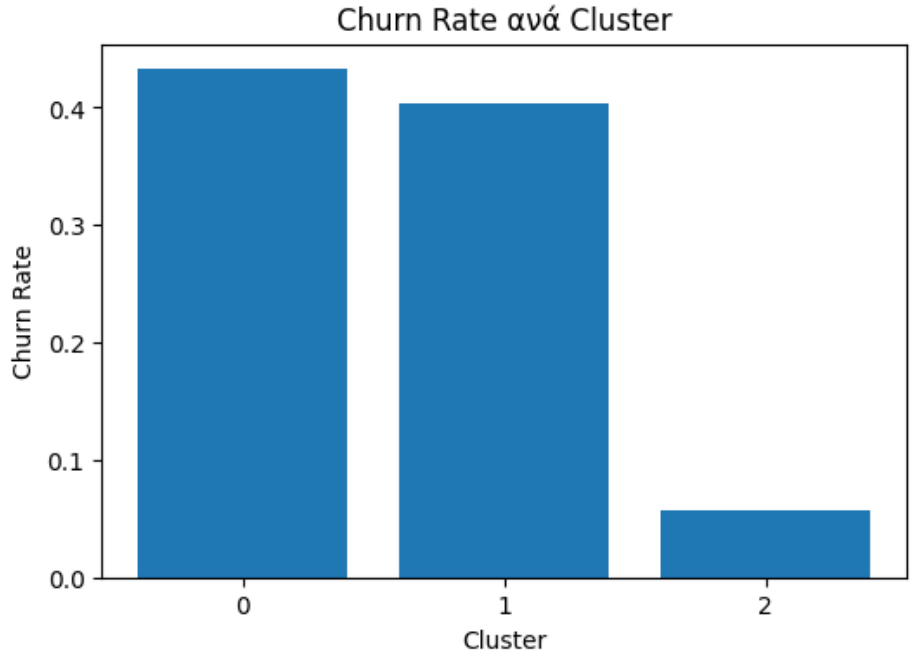
πρόβλημα ταξινόμησης. Με τον τρόπο αυτό καθίσταται δυνατή η εκπαίδευση μοντέλων μηχανικής μάθησης που εκτιμούν την πιθανότητα αποχώρησης κάθε πελάτη.

Η δημιουργία των προφίλ συστάδων πραγματοποιήθηκε σε Python.

4.2 Συστάδες και Αποχώρηση

4.2.1 Churn Rate ανά Cluster

Για πιο συγκεκριμένη εικόνα της επικινδυνότητας των clusters δημιουργήθηκε η κατανομή του churn ανά ομάδα.



churn

cluster	
0	0.432558
1	0.403935
2	0.056738

dtype: float64

Εικόνα 20. *Ποσοστό αποχώρησης πελατών ανά συστάδα*

Παρατηρείται σημαντική διαφοροποίηση του ποσοστού αποχώρησης μεταξύ των συστάδων. Οι συστάδες 0 (43% churn) και 1 (40% churn) εμφανίζουν υψηλή πιθανότητα αποχώρησης, ενώ η συστάδα 2 (5,7% churn) παρουσιάζει ιδιαίτερα χαμηλό ποσοστό. Το αποτέλεσμα αυτό υποδηλώνει ότι η τμηματοποίηση καταναλωτών εντοπίζει ομάδες με διαφορετικό επίπεδο κινδύνου αποχώρησης, γεγονός που επιτρέπει τη στοχευμένη ανάλυση των χαρακτηριστικών τους.

4.2.2 Ερμηνεία Συστάδων Πελατών με Επιχειρησιακή Προσέγγιση

Η εφαρμογή του αλγορίθμου ομαδοποίησης οδήγησε στον διαχωρισμό των πελατών σε τρεις διακριτές συμπεριφορικές συστάδες, οι οποίες διαφοροποιούνται ως προς τη χρονική εγγύτητα χρήσης (Recency), τη συχνότητα δραστηριότητας (Frequency) και το επίπεδο κατανάλωσης (Monetary).

Η τμηματοποίηση των πελατών δεν αποκαλύπτει μόνο διαφορές στο επίπεδο κατανάλωσης αλλά κυρίως διαφοροποίηση στη συμπεριφορά παραμονής στον πάροχο. Σε αγορά ενέργειας η αποχώρηση δεν σημαίνει διακοπή χρήσης ηλεκτρικής ενέργειας αλλά αλλαγή παρόχου, επομένως η έννοια της πιστότητας συνδέεται περισσότερο με τη χρονική συνέχεια της σχέσης παρά με τον όγκο κατανάλωσης.

Η πρώτη συστάδα- **Cluster 0 (Πελάτες μειωμένης πρόσφατης δραστηριότητας (υψηλού κινδύνου))**, αφορά πελάτες που εμφανίζουν μεγαλύτερη χρονική απόσταση από την τελευταία καταγεγραμμένη κατανάλωση. Παρότι το επίπεδο κατανάλωσης μπορεί να παραμένει μέτριο, η απουσία πρόσφατης δραστηριότητας υποδηλώνει πιθανή μετακίνηση σε άλλο πάροχο. Στην πράξη πρόκειται για πελάτες που ενδέχεται να έχουν ήδη αλλάξει πρόγραμμα ή να βρίσκονται σε διαδικασία σύγκρισης τιμολογίων. Η κατηγορία αυτή αντιπροσωπεύει το τμήμα του πελατολογίου με τον υψηλότερο κίνδυνο αποχώρησης.

Η δεύτερη συστάδα - **Cluster 1 (ενεργοί αλλά ευαίσθητοι στην τιμή),** αφορά πελάτες οι οποίοι εμφανίζουν σχετικά πρόσφατη δραστηριότητα και ικανοποιητική κατανάλωση, ωστόσο η συμπεριφορά τους δεν είναι σταθερή στο χρόνο. Πρόκειται για καταναλωτές που παραμένουν στον πάροχο όσο το πρόγραμμα είναι ανταγωνιστικό. Σε συνθήκες κυμαινόμενων τιμολογίων μπορούν εύκολα να μετακινηθούν όταν εμφανιστεί οικονομικότερη επιλογή. Η πιθανότητα αποχώρησης είναι μεσαία και επηρεάζεται κυρίως από μεταβολές τιμών.

Η τρίτη συστάδα - **Cluster 2 (πελάτες υψηλής πιστότητας),** χαρακτηρίζεται από συνεχή πρόσφατη παρουσία και σταθερή χρήση προϊόντων και υπηρεσιών. Ακόμη και όταν η κατανάλωση δεν είναι η υψηλότερη, η συχνότητα και η χρονική συνέχεια υποδηλώνουν σταθερή σχέση με τον πάροχο. Οι πελάτες αυτοί εμφανίζουν χαμηλή πιθανότητα αποχώρησης και αποτελούν το βασικό πυρήνα του πελατολογίου.

Η τμηματοποίηση δείχνει ότι στην αγορά ενέργειας η αξία πελάτη δεν ταυτίζεται απαραίτητα με το ύψος κατανάλωσης. Πελάτες υψηλής κατανάλωσης μπορούν να αποχωρήσουν όταν εντοπίσουν οικονομικότερο πρόγραμμα, ενώ πελάτες μέσης κατανάλωσης μπορεί να παραμένουν για μεγάλο χρονικό διάστημα. Συνεπώς, ο σημαντικότερος δείκτης πιστότητας είναι η πρόσφατη δραστηριότητα και όχι το συνολικό ιστορικό χρήσης.

4.3 Αξιολόγηση Προγνωστικών Μοντέλων (Evaluation)

Για την πρόβλεψη της αποχώρησης πελατών εφαρμόστηκαν δύο διαφορετικές προσεγγίσεις εποπτευόμενης μάθησης: η Λογιστική Παλινδρόμηση **(Logistic Regression)** και το μοντέλο Τυχαίου Δάσους **(Random Forest)**. Η εκπαίδευση πραγματοποιήθηκε στο ισορροπημένο σύνολο εκπαίδευσης που προέκυψε μετά την εφαρμογή της τεχνικής SMOTE, ενώ η αξιολόγηση έγινε σε ανεξάρτητο σύνολο ελέγχου. Οι υπερπαράμετροι ρυθμίστηκαν μέσω Grid Search με διασταυρούμενη

επικύρωση πέντε πτυχών. Για τη Λογιστική Παλινδρόμηση η βέλτιστη τιμή της παραμέτρου κανονικοποίησης ήταν C=10, ενώ για το Τυχαίο Δάσος επιλέχθηκε μέγιστο βάθος δέντρων ίσο με 6 και ελάχιστο πλήθος δειγμάτων διαχωρισμού ίσο με 5.

4.3.1 Απόδοση Λογιστικής Παλινδρόμησης - Logistic Regression

Το μοντέλο παρουσίασε συνολική ακρίβεια (accuracy) 77%. Η απόδοση στην κατηγορία αποχώρησης (churn) ήταν:

- Precision: 0.61
- Recall: 0.82
- F1-score: 0.70
- AUC: 0.86

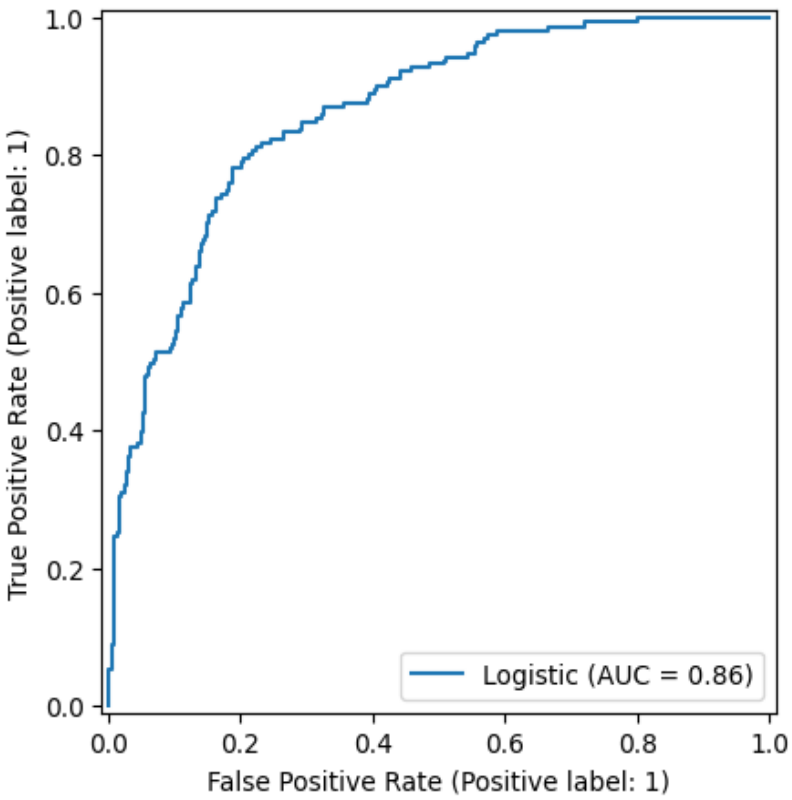
Η υψηλή τιμή recall δείχνει ότι το μοντέλο εντοπίζει τη μεγάλη πλειονότητα των πελατών που αποχωρούν, γεγονός ιδιαίτερα σημαντικό για επιχειρησιακές εφαρμογές πρόληψης αποχώρησης.

LOGISTIC

	precision	recall	f1-score	support
0	0.90	0.75	0.82	409
1	0.61	0.82	0.70	191
accuracy			0.77	600
macro avg	0.75	0.78	0.76	600
weighted avg	0.81	0.77	0.78	600

AUC: 0.8600340506150873

Εικόνα 21. *Αποτελέσματα ταξινόμησης (precision, recall, F1-score και AUC) για το μοντέλο Λογιστικής Παλινδρόμησης στο σύνολο ελέγχου.*



Εικόνα 22. *Καμπύλη ROC για το μοντέλο Λογιστικής Παλινδρόμησης στο σύνολο ελέγχου.*

4.3.2 Απόδοση Τυχαίου Δάσους – Random Forest

Το μοντέλο Τυχαίου Δάσους παρουσίασε επίσης συνολική ακρίβεια 76%, με παρόμοια αλλά ελαφρώς διαφορετική συμπεριφορά:

- Precision: 0.59
- Recall: 0.77
- F1-score: 0.67
- AUC: 0.85

Το μοντέλο είναι πιο συντηρητικό, εμφανίζει ελαφρώς μικρότερη ευαισθησία στον εντοπισμό των αποχωρούντων πελατών σε σχέση με τη Λογιστική Παλινδρόμηση, αλλά διατηρεί υψηλή συνολική διακριτική ικανότητα με AUC = 0.85.

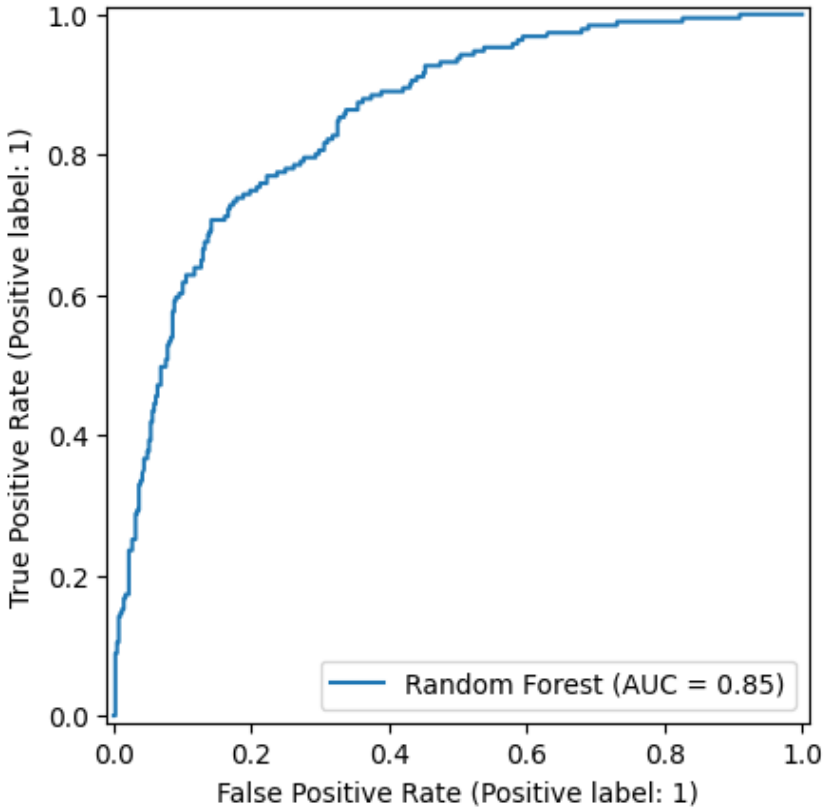
```
RANDOM FOREST
              precision    recall  f1-score   support

           0       0.88      0.75      0.81       409
           1       0.59      0.77      0.67       191

    accuracy                           0.76       600
   macro avg       0.74      0.76      0.74       600
weighted avg       0.79      0.76      0.77       600

AUC: 0.8515854017588551
```

Εικόνα 23. *Αποτελέσματα ταξινόμησης (precision, recall, F1-score και AUC) για το μοντέλο Random Forest στο σύνολο ελέγχου.*



Εικόνα 24. *Καμπύλη ROC για το μοντέλο Τυχαίου Δάσους στο σύνολο ελέγχου.*

4.5 Σημαντικότητα Χαρακτηριστικών

4.5.1 Ανάλυση Σημαντικότητας Χαρακτηριστικών (Feature Importance)

Για την κατανόηση της λειτουργίας των προγνωστικών μοντέλων δεν αρκεί μόνο η αξιολόγηση της ακρίβειας τους, αλλά απαιτείται και η διερεύνηση του τρόπου με τον οποίο λαμβάνονται οι αποφάσεις. Για τον λόγο αυτό εξετάστηκε η συμβολή κάθε μεταβλητής στην πρόβλεψη της αποχώρησης πελατών.

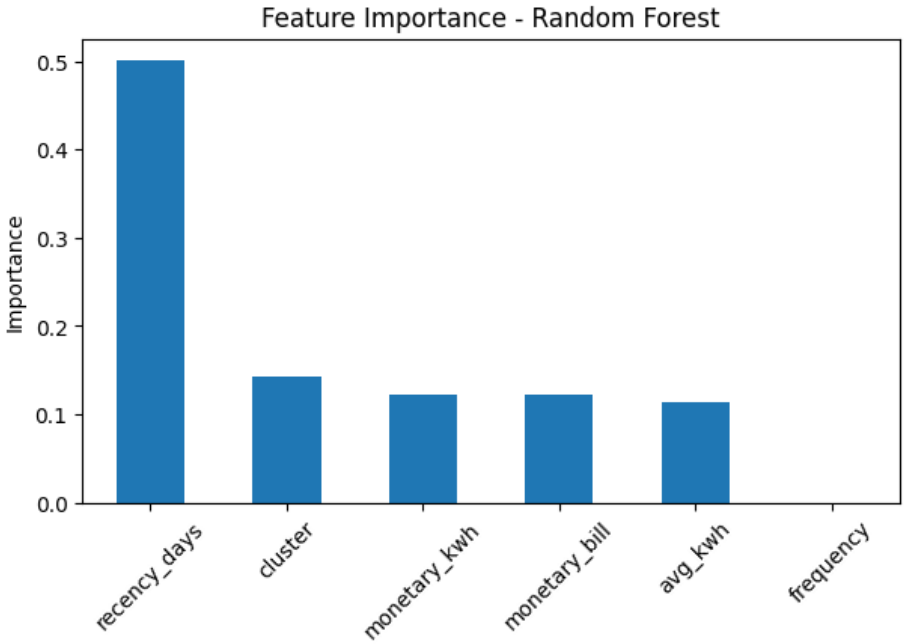
Η ανάλυση πραγματοποιήθηκε με δύο διαφορετικές προσεγγίσεις:

- στο μοντέλο Random Forest μέσω της σημαντικότητας χαρακτηριστικών (feature importance)
- στο μοντέλο Logistic Regression μέσω των συντελεστών παλινδρόμησης

Με τον τρόπο αυτό καθίσταται δυνατή τόσο η ποσοτική αξιολόγηση της επίδρασης κάθε μεταβλητής όσο και η κατανόηση της κατεύθυνσης της σχέσης της με την πιθανότητα αποχώρησης.

4.5.2 Σημαντικότητα Χαρακτηριστικών στο Random Forest

Το Random Forest υπολογίζει τη σημασία κάθε μεταβλητής με βάση το πόσο συμβάλλει στη μείωση της αβεβαιότητας κατά τη δημιουργία των δέντρων αποφάσεων.



	θ
recency_days	0.500478
cluster	0.142082
monetary_kwh	0.122016
monetary_bill	0.121662
avg_kwh	0.113763
frequency	0.000000

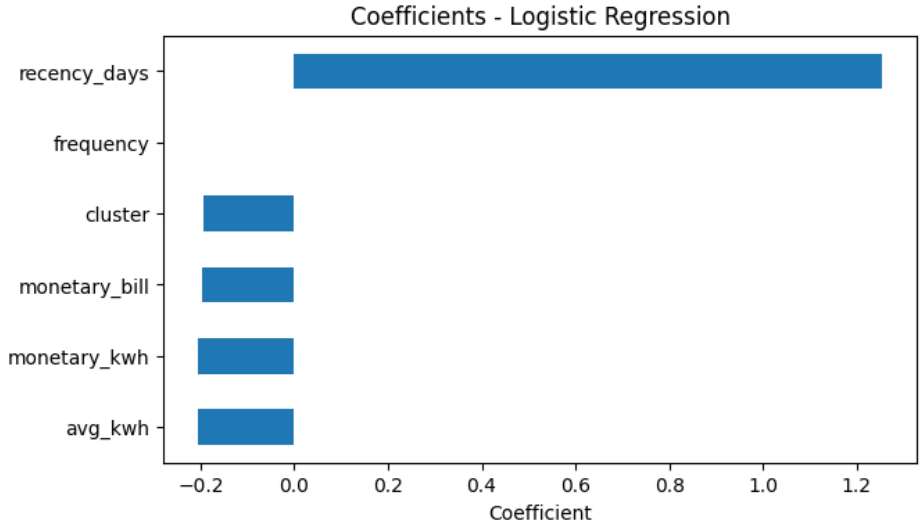
***Εικόνα 25**: Σημαντικότητα χαρακτηριστικών στο μοντέλο Random Forest*

Το διάγραμμα σημαντικότητας χαρακτηριστικών δείχνει ότι η μεταβλητή recency_days αποτελεί τον κυρίαρχο παράγοντα πρόβλεψης της αποχώρησης. Η υψηλή συμβολή της υποδηλώνει ότι ο χρόνος που έχει περάσει από την τελευταία δραστηριότητα αποτελεί την ισχυρότερη ένδειξη εγκατάλειψης της υπηρεσίας.Οι μεταβλητές

κατανάλωσης (monetary_kwh, avg_kwh και monetary_bill) παρουσιάζουν μικρότερη αλλά σαφή επίδραση, γεγονός που υποδηλώνει ότι το επίπεδο χρήσης επηρεάζει τη συμπεριφορά, χωρίς όμως να αποτελεί τον βασικό μηχανισμό αποχώρησης. Αντίθετα, η μεταβλητή frequency εμφανίζει αμελητέα συμβολή. Το αποτέλεσμα εξηγείται από τη συσχέτισή της με τη recency_days, καθώς η πρόσφατη παρουσία ενός πελάτη εμπεριέχει ήδη πληροφορία σχετικά με τη συχνότητα χρήσης. Συνεπώς το μοντέλο δεν χρειάζεται να χρησιμοποιήσει τη μεταβλητή αυτή για να διαχωρίσει τις κατηγορίες. Η συμμετοχή της μεταβλητής cluster επιβεβαιώνει ότι η τμηματοποίηση προσθέτει συμπεριφορική πληροφορία χρήσιμη για την πρόβλεψη αποχώρησης.

4.5.3 Συντελεστές Logistic Regression

Σε αντίθεση με το Random Forest, η λογιστική παλινδρόμηση επιτρέπει την κατανόηση της κατεύθυνσης της επίδρασης κάθε χαρακτηριστικού.



	0
avg_kwh	-0.205313
monetary_kwh	-0.205313
monetary_bill	-0.196534
cluster	-0.192983
frequency	0.000000
recency_days	1.254200

dtype: float64

Εικόνα 26 : *Συντελεστές λογιστικής παλινδρόμησης για την πρόβλεψη αποχώρησης*

Οι συντελεστές της λογιστικής παλινδρόμησης επιτρέπουν την κατανόηση της κατεύθυνσης επίδρασης κάθε μεταβλητής. Η θετική τιμή της recency_days δείχνει ότι όσο αυξάνεται ο χρόνος από την τελευταία δραστηριότητα αυξάνεται και η πιθανότητα αποχώρησης.

Οι μεταβλητές κατανάλωσης εμφανίζουν αρνητικούς συντελεστές, γεγονός που υποδηλώνει ότι οι ενεργοί και υψηλής χρήσης πελάτες έχουν μικρότερη πιθανότητα να εγκαταλείψουν τον πάροχο. Η σχέση αυτή είναι αναμενόμενη, καθώς η συστηματική χρήση συνδέεται με μεγαλύτερη σταθερότητα στη συνεργασία.

Η frequency παραμένει ουσιαστικά ουδέτερη μεταβλητή, επιβεβαιώνοντας τα αποτελέσματα του Random Forest. Η πληροφορία που περιέχει αποτυπώνεται ήδη στη μεταβλητή recency_days και δεν προσθέτει επιπλέον διακριτική ικανότητα στο μοντέλο.

Η συμφωνία των δύο διαφορετικών προσεγγίσεων ως προς τους σημαντικότερους παράγοντες αποχώρησης ενισχύει την αξιοπιστία των αποτελεσμάτων. Παρά τη διαφορετική μαθηματική τους λειτουργία, και τα δύο μοντέλα αναδεικνύουν τη χρονική εγγύτητα χρήσης ως βασικό μηχανισμό διατήρησης πελατών. Στο επόμενο στάδιο πραγματοποιείται συγκριτική αξιολόγηση της προγνωστικής τους επίδοσης.

4.5.4 Σύνδεση Τμηματοποίησης με Σημαντικότητα Χαρακτηριστικών (Feature Importance)

Η ανάλυση των συστάδων έδειξε ότι η βασική διαφοροποίηση των πελατών σχετίζεται με τη χρονική συνέχεια της χρήσης και όχι αποκλειστικά με το επίπεδο κατανάλωσης. Οι ομάδες υψηλού κινδύνου χαρακτηρίζονται κυρίως από αυξημένη χρονική απόσταση από την τελευταία δραστηριότητα, ενώ οι σταθεροί πελάτες εμφανίζουν συστηματική πρόσφατη παρουσία. Συνεπώς, η πιθανότητα αποχώρησης φαίνεται να εξαρτάται περισσότερο από το πότε χρησιμοποιήθηκε τελευταία φορά ένα προϊόν ή μια υπηρεσία παρά από το πόσο καταναλώθηκε στο παρελθόν. Το εύρημα αυτό επιβεβαιώνεται από την ανάλυση σημαντικότητας χαρακτηριστικών στα προγνωστικά μοντέλα. Και τα δύο μοντέλα, παρά τη διαφορετική μεθοδολογία τους, ανέδειξαν τη μεταβλητή recency_days ως τον κυρίαρχο παράγοντα πρόβλεψης. Η σύγκλιση των αποτελεσμάτων υποδηλώνει ότι η συμπεριφορά αποχώρησης περιγράφεται επαρκώς από την πρόσφατη δραστηριότητα των πελατών.

Αντίθετα, η μεταβλητή frequency εμφανίζει περιορισμένη συμβολή, γεγονός που εξηγείται από τη συσχέτισή της με τη χρονική εγγύτητα χρήσης. Η πληροφορία της συχνότητας ενσωματώνεται ήδη στην recency, με αποτέλεσμα να μην προσθέτει επιπλέον διακριτική ικανότητα στα μοντέλα. Οι μεταβλητές κατανάλωσης παρουσιάζουν δευτερεύουσα επίδραση, υποδεικνύοντας ότι η ένταση χρήσης επηρεάζει τη συμπεριφορά, αλλά δεν αποτελεί τον βασικό μηχανισμό αποχώρησης.

Στο πλαίσιο της αγοράς ενέργειας, το αποτέλεσμα αυτό είναι αναμενόμενο. Οι καταναλωτές δεν διακόπτουν την κατανάλωση ηλεκτρικής ενέργειας, αλλά αλλάζουν πάροχο. Επομένως, η απουσία πρόσφατης δραστηριότητας λειτουργεί ως ισχυρότερη ένδειξη μετακίνησης από το συνολικό ιστορικό κατανάλωσης. Η τμηματοποίηση και τα προγνωστικά μοντέλα καταλήγουν έτσι στο ίδιο συμπέρασμα μέσω διαφορετικών αναλυτικών προσεγγίσεων, ενισχύοντας την αξιοπιστία της ανάλυσης.

4.8 Συγκριτική Αξιολόγηση Μοντέλων – Final Model

Μετά την αξιολόγηση των μοντέλων στο σύνολο ελέγχου πραγματοποιήθηκε συγκριτική ανάλυση της απόδοσής τους, ώστε να επιλεγεί η καταλληλότερη προσέγγιση για την πρόβλεψη αποχώρησης πελατών. Η σύγκριση δεν βασίστηκε

αποκλειστικά στη συνολική ακρίβεια αλλά κυρίως στην ικανότητα εντοπισμού των πελατών που αποχωρούν, καθώς η έγκαιρη αναγνώριση αυτής της κατηγορίας αποτελεί τον βασικό στόχο της ανάλυσης.

Η Λογιστική Παλινδρόμηση παρουσίασε σταθερή συμπεριφορά και υψηλή ευαισθησία (recall) ως προς την κατηγορία αποχώρησης, εντοπίζοντας μεγαλύτερο ποσοστό πελατών υψηλού κινδύνου. Η γραμμική μορφή του μοντέλου επιτρέπει την άμεση ερμηνεία της επίδρασης των χαρακτηριστικών, γεγονός ιδιαίτερα σημαντικό για την κατανόηση των παραγόντων που σχετίζονται με τη μετακίνηση πελατών.

Το μοντέλο Random Forest εμφάνισε παρόμοια συνολική διακριτική ικανότητα, αποτυπώνοντας και μη γραμμικές σχέσεις μεταξύ των μεταβλητών. Ωστόσο, η βελτίωση της πρόβλεψης ήταν περιορισμένη, ενώ η ερμηνεία των αποτελεσμάτων ήταν λιγότερο άμεση. Το εύρημα αυτό υποδηλώνει ότι το φαινόμενο της αποχώρησης περιγράφεται επαρκώς από σχετικά απλές σχέσεις μεταξύ των χαρακτηριστικών. Με βάση τα παραπάνω, **επιλέγεται ως καταλληλότερο μοντέλο η Λογιστική Παλινδρόμηση**, καθώς συνδυάζει ικανοποιητική προγνωστική απόδοση με διαφάνεια και δυνατότητα επιχειρησιακής ερμηνείας. Η επιλογή αυτή επιτρέπει όχι μόνο την πρόβλεψη της αποχώρησης αλλά και την κατανόηση των παραγόντων που την επηρεάζουν, διευκολύνοντας τη λήψη αποφάσεων διατήρησης πελατών.

Η σύγκριση των δύο μοντέλων οδηγεί σε κοινά συμπεράσματα:

- Η πρόσφατη δραστηριότητα είναι ο ισχυρότερος δείκτης αποχώρησης
- Η υψηλή κατανάλωση λειτουργεί ως παράγοντας διατήρησης πελατών
- Η τμηματοποίηση πελατών συμβάλλει ουσιαστικά στην πρόβλεψη
- Η απλή συχνότητα χρήσης δεν αποτελεί καθοριστικό χαρακτηριστικό

Η σύγκλιση των αποτελεσμάτων μεταξύ ενός γραμμικού και ενός μη γραμμικού μοντέλου ενισχύει την αξιοπιστία της ανάλυσης και δείχνει ότι τα συμπεριφορικά χαρακτηριστικά που προέκυψαν από το RFM profiling αποτελούν κατάλληλη αναπαράσταση της συμπεριφοράς των πελατών.

Κεφάλαιο 5: Συμπεράσματα

5.1 Τελικό Συμπέρασμα Σύγκρισης Μοντέλων & Επιχειρησιακή Ερμηνεία

Η ανάλυση έδειξε ότι η πιθανότητα αποχώρησης σχετίζεται κυρίως με τη χρονική συνέχεια της σχέσης πελάτη-παρόχου και λιγότερο με το συνολικό επίπεδο κατανάλωσης. Το εύρημα αυτό αποκτά ιδιαίτερη σημασία στην απελευθερωμένη αγορά ενέργειας, όπου οι καταναλωτές δεν παύουν να χρησιμοποιούν ηλεκτρική ενέργεια αλλά μετακινούνται μεταξύ παρόχων αναζητώντας οικονομικότερους όρους συνεργασίας.

Η αυξημένη χρονική απόσταση από την τελευταία δραστηριότητα λειτουργεί ως ένδειξη απομάκρυνσης από τον πάροχο και ενδέχεται να υποδηλώνει αλλαγή τιμολογίου ή σύμβασης. Αντίθετα, η υψηλή κατανάλωση δεν συνεπάγεται απαραίτητα πιστότητα, καθώς ακόμη και ενεργοί πελάτες μπορούν να μετακινηθούν όταν εντοπίσουν οικονομικότερο πρόγραμμα. Συνεπώς, η πρόσφατη δραστηριότητα

αποτελεί πιο αξιόπιστο δείκτη διατήρησης από το ιστορικό χρήσης. Η τμηματοποίηση των πελατών επιτρέπει την κατηγοριοποίηση του πελατολογίου σε ομάδες διαφορετικού κινδύνου. Πελάτες με μειωμένη πρόσφατη χρήση εμφανίζουν αυξημένη πιθανότητα αποχώρησης και απαιτούν άμεσες ενέργειες διατήρησης, ενώ πελάτες με σταθερή παρουσία εμφανίζουν μεγαλύτερη πιθανότητα παραμονής. Η συμπεριφορά κατανάλωσης μπορεί να χρησιμοποιηθεί για την τμηματοποίηση και κατανόηση των καταναλωτών ενέργειας (Albert & Rajagopal, 2013).

Με βάση τα παραπάνω, ένας πάροχος ενέργειας μπορεί να εφαρμόσει στοχευμένες στρατηγικές:

- έγκαιρη ειδοποίηση πελατών με μειωμένη δραστηριότητα,
- προσαρμογή τιμολογίων ή εκπτώσεων σε πελάτες υψηλού κινδύνου,
- διατήρηση σχέσης με πελάτες σταθερής συμπεριφοράς μέσω προγραμμάτων επιβράβευσης.

Οι καταναλωτές ανταποκρίνονται έντονα σε μεταβολές τιμών και επιλογές τιμολογίων (Faruqui & Sergici, 2010). Η αξιοποίηση προγνωστικών μοντέλων σε συνδυασμό με συμπεριφορική τμηματοποίηση επιτρέπει τη μετάβαση από παθητική παρακολούθηση της αποχώρησης σε προληπτική διαχείριση πελατειακών σχέσεων.

5.2 Συμπεράσματα

Η ανάλυση έδειξε ότι η πιθανότητα αποχώρησης σχετίζεται κυρίως με τη χρονική συνέχεια της σχέσης πελάτη-παρόχου και λιγότερο με το συνολικό επίπεδο κατανάλωσης. Το εύρημα αυτό αποκτά ιδιαίτερη σημασία στην απελευθερωμένη αγορά ενέργειας, όπου οι καταναλωτές δεν παύουν να χρησιμοποιούν ηλεκτρική ενέργεια αλλά μετακινούνται μεταξύ παρόχων αναζητώντας οικονομικότερους όρους συνεργασίας.

Η αυξημένη χρονική απόσταση από την τελευταία δραστηριότητα λειτουργεί ως ένδειξη απομάκρυνσης από τον πάροχο και ενδέχεται να υποδηλώνει αλλαγή τιμολογίου ή σύμβασης. Αντίθετα, η υψηλή κατανάλωση δεν συνεπάγεται απαραίτητα πιστότητα, καθώς ακόμη και ενεργοί πελάτες μπορούν να μετακινηθούν όταν εντοπίσουν οικονομικότερο πρόγραμμα. Συνεπώς, η πρόσφατη δραστηριότητα αποτελεί πιο αξιόπιστο δείκτη διατήρησης από το ιστορικό χρήσης.

Η τμηματοποίηση των πελατών επιτρέπει την κατηγοριοποίηση του πελατολογίου σε ομάδες διαφορετικού κινδύνου. Πελάτες με μειωμένη πρόσφατη χρήση εμφανίζουν αυξημένη πιθανότητα αποχώρησης και απαιτούν άμεσες ενέργειες διατήρησης, ενώ πελάτες με σταθερή παρουσία εμφανίζουν μεγαλύτερη πιθανότητα παραμονής.

Με βάση τα παραπάνω, ένας πάροχος ενέργειας μπορεί να εφαρμόσει στοχευμένες στρατηγικές:

- έγκαιρη ειδοποίηση πελατών με μειωμένη δραστηριότητα,
- προσαρμογή τιμολογίων ή εκπτώσεων σε πελάτες υψηλού κινδύνου,
- διατήρηση σχέσης με πελάτες σταθερής συμπεριφοράς μέσω προγραμμάτων επιβράβευσης.

Η αξιοποίηση προγνωστικών μοντέλων σε συνδυασμό με συμπεριφορική τμηματοποίηση επιτρέπει τη μετάβαση από παθητική παρακολούθηση της αποχώρησης σε προληπτική διαχείριση πελατειακών σχέσεων**.**

5.3 Περιορισμοί και Μελλοντική Έρευνα

Η παρούσα εργασία βασίστηκε σε δεδομένα προσομοίωσης τα οποία αναπαριστούν ρεαλιστικά πρότυπα κατανάλωσης, χωρίς όμως να αποτυπώνουν πλήρως την πολυπλοκότητα της πραγματικής αγοράς ενέργειας. Στην πράξη, η αποχώρηση επηρεάζεται και από εξωγενείς παράγοντες, όπως μεταβολές τιμολογίων, εμπορικές προσφορές, γεωγραφικά χαρακτηριστικά και κοινωνικοοικονομικά στοιχεία των καταναλωτών, τα οποία δεν περιλαμβάνονται στο παρόν σύνολο δεδομένων.

Μελλοντική έρευνα θα μπορούσε να επεκτείνει την ανάλυση με πραγματικά δεδομένα κατανάλωσης, ενσωματώνοντας πληροφορίες τιμολόγησης και χρονικών μεταβολών της αγοράς. Επιπλέον, η χρήση δυναμικών μοντέλων χρονοσειρών ή μεθόδων επιβίωσης θα επέτρεπε την εκτίμηση όχι μόνο της πιθανότητας αποχώρησης αλλά και του χρονικού ορίζοντα εμφάνισής της. Τέλος, η αξιοποίηση μεθόδων εξατομίκευσης θα μπορούσε να συνδέσει την πρόβλεψη με προτεινόμενες ενέργειες διατήρησης, μετατρέποντας το προγνωστικό σύστημα σε εργαλείο υποστήριξης αποφάσεων σε πραγματικό χρόνο.

ΚΕΦΑΛΑΙΟ 6

Βιβλιογραφικές Αναφορές

6.1 Βιβλιογραφία

Han, J., Kamber, M., & Pei, J. (2012). Data mining: Concepts and techniques (3rd ed.). Morgan Kaufmann.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). An introduction to statistical learning: With applications in R (2nd ed.). Springer.

Kotler, P. (2017). Philip Kotler: some of my adventures in marketing. Journal of Historical Research in Marketing, 9(2), 203-208.

Kumar, V., Jones, E., Venkatesan, R., & Leone, R. P. (2011). Is market orientation a source of sustainable competitive advantage or simply the cost of competing?. Journal of marketing, 75(1), 16-30.

Wedel M, Kamakura WA. 2012. Market segmentation: conceptual and methodological foundations. 8. Kluwer, Dordrecht, the Netherlands.

Peppers, D., & Rogers, M. (2016). Managing customer experience and relationships: A strategic framework. John Wiley & Sons.

Bult, J. R., & Wansbeek, T. J. (1995). Optimal selection for direct mail. Journal of Marketing Research, 32(3), 330-337.

Berman, B., & Berman, M. (2013). Marketing Analytics: A Practical Guide to Improving Consumer Insights Using Data Techniques. Kogan Page Publishers.

Dogan, O., Ayçin, E., & Bulut, Z. (2018). Customer segmentation by using RFM model andclustering methods: a case study in retail industry. International Journal of Contemporary Economics and Administrative Sciences, 8.

Ullah, A., Mohmand, M. I., Hussain, H., Johar, S., Khan, I., Ahmad, S., Huda, S. (2023). Customer Analysis Using Machine Learning-Based Classification Algorithms for Effective Segmentation Using Recency, Frequency, Monetary, and Time. Sensors, 23(6), 3180.

Van Burg, J. M. (2020). Customer segmentation using RFM analysis.

Anitha, P., & Patil, M. M. (2022). RFM model for customer purchase behavior using K-Means algorithm. Journal of King Saud University-Computer and Information Sciences, 34(5), 1785- 1792.

Tavakoli, M., Molavi, M., Masoumi, V., Mobini, M., Etemad, S., & Rahmani, R. (2018). Customer segmentation and strategy development based on user behavior analysis, RFM model and data mining techniques: a case study. In 2018 IEEE 15th International Conference on eBusiness Engineering (ICEBE) (pp. 119-126). IEEE.

Verhoef, P. C., Franses, P. H., & Hoekstra, J. C. (2002). The effect of relational constructs on customer referrals and number of services purchased from a multiservice provider: does age of relationship matter? Journal of the Academy of Marketing Science, 30, 202–216 (2002).

Dursun, A., & Caber, M. (2016). Using data mining techniques for profiling profitable hotel customers: An application of RFM analysis. Tourism management perspectives, 18, 153-160.

Verhoef, P. C., Franses, P. H., & Hoekstra, J. C. (2002). The effect of relational constructs on customer referrals and number of services purchased from a multiservice provider: does age of relationship matter? Journal of the Academy of Marketing Science, 30, 202–216 (2002).

Cheng, C. H., & Chen, Y. S. (2009). Classifying the segmentation of customer value via RFM model and RS theory. Expert systems with applications, 36(3), 4176-4184.

Jain, A. K., Murty, M. N., & Flynn, P. J. (1999). Data clustering: a review. ACM computing surveys (CSUR), 31(3), 264-323.

Tan, P.-N., Steinbach, M., & Kumar, V. (2005). Introduction to Data Mining. Pearson Addison-Wesley.

Quinlan, J. R. (1996). Learning decision tree classifiers. ACM Computing Surveys (CSUR), 28(1), 71-72.

Biau, G., & Scornet, E. (2016). A random forest guided tour. Test, 25, 197-227.

Mehta, C. R., & Patel, N. R. (1995). Exact logistic regression: theory and examples. Statistics in medicine, 14(19), 2143-2160.

Verbeke, W., Dejaeger, K., Martens, D., Hur, J., & Baesens, B. (2012).
New insights into churn prediction in the telecommunication sector. Expert Systems with Applications, 39(8), 7317–7327.

Verbeke, W., Dejaeger, K., Martens, D., Hur, J., & Baesens, B. (2012). New insights into churn prediction in the telecommunication sector: A profit driven data mining approach. European Journal of Operational Research, 218(1), 211–229.

Coussement, K., & De Bock, K. W. (2013).
Customer churn prediction in the online gambling industry: The beneficial effect of ensemble learning. Journal of Business Research, 66(9), 1629–1636.

Depren, O., Kartal, E., & Erenel, Z. (2017).
Customer churn prediction in energy markets using data mining techniques. Energy Policy, 110, 465–474.

Verhoef, P. C., & Donkers, B. (2003).
Predicting customer potential value: An application in the insurance industry. Decision Support Systems, 32(2), 189–199.

Jain, A. K. (2010).
Data clustering: 50 years beyond K-means. Pattern Recognition Letters, 31(8), 651–666.
https://doi.org/10.1016/j.patrec.2009.09.011

Hastie, T., Tibshirani, R., & Friedman, J. (2009).
The elements of statistical learning: Data mining, inference, and prediction (2nd ed.). Springer.

Kotler, P., & Keller, K. L. (2016). Marketing management (15th ed.). Pearson

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. Journal of Artificial Intelligence Research, 16, 321–357.

He, H., & Garcia, E. A. (2009). Learning from imbalanced data. IEEE Transactions on Knowledge and Data Engineering, 21(9), 1263–1284.

Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). Applied logistic regression (3rd ed.). Wiley.

Albert, A., & Rajagopal, R. (2013). Smart meter driven segmentation: What your consumption says about you. IEEE Transactions on Power Systems, 28(4), 4019–4030.

Faruqui, A., & Sergici, S. (2010). Household response to dynamic pricing of electricity. Energy Journal, 31(3), 193–225.

Jain, A. K., Murty, M. N., & Flynn, P. J. (1999). Data clustering: A review. ACM Computing Surveys, 31(3), 264–323. https://doi.org/10.1145/331499.331504

Thearling, K. (1999). Clustering techniques. In W. H. Inmon, D. Strauss, & G. Neushloss (Eds.), Data warehousing and OLAP (pp. 203–221). Wiley.

Hughes, A. M. (1994). Strategic database marketing: The masterplan for starting and managing a profitable, customer-based marketing program. Probus Publishing.

Kotler, P., & Keller, K. L. (2016). Marketing management (15th ed.). Pearson.

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In L. M. Le Cam & J. Neyman (Eds.), Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability (Vol. 1, pp. 281–297). University of California Press.

Jain, A. K., Murty, M. N., & Flynn, P. J. (1999). Data clustering: A review. ACM Computing Surveys, 31(3), 264–323. https://doi.org/10.1145/331499.331504

Thearling, K. (1999). Clustering techniques. In W. H. Inmon, D. Strauss, & G. Neushloss (Eds.), Data warehousing and OLAP (pp. 203–221). Wiley.

MacQueen, J. B. (1967). Some methods for classification and analysis of multivariate observations. In Proceedings of the fifth Berkeley symposium on mathematical statistics and probability (Vol. 1, No. 14, pp. 281-297). University of California Press.

Jain, A. K. (2010). Data clustering: 50 years beyond K-means. Pattern Recognition Letters, 31(8), 651-666.

Jolliffe, I. T. (2002). Principal component analysis (2nd ed.). Springer.

Abdi, H., & Williams, L. J. (2010). Principal component analysis. Wiley Interdisciplinary Reviews: Computational Statistics, 2(4), 433-459.

Berger, P. D., & Nasr, N. I. (1998). Customer lifetime value: Marketing models and applications. Journal of Interactive Marketing, 12(1), 17-30.

Géron, A. (2019). Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: Concepts, tools, and techniques to build intelligent systems (2nd ed.). O'Reilly Media.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning: with applications in R. Springer.

MacQueen, J. B. (1967). Some methods for classification and analysis of multivariate observations. In L. M. Le Cam & J. Neyman (Eds.), Proceedings of the fifth Berkeley symposium on mathematical statistics and probability (Vol. 1, No. 14, pp. 281-297). University of California Press.

Jain, A. K. (2010). Data clustering: 50 years beyond K-means. Pattern Recognition Letters, 31(8), 651-666. [redacted link]

Breiman, L. (2001). Random Forests. Machine Learning, 45(1), 5-32. [redacted link]

Grömping, U. (2009). Variable importance in random forests. Stata Journal, 9(2), 312-320.

Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. Journal of Machine Learning Research, 3, 1157-1182.

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability (Vol. 1, pp. 281–297). University of California Press.

Εναλλακτικά (πολύ συχνή σε σύγχρονες εργασίες):

Jain, A. K. (2010). Data clustering: 50 years beyond K-means. Pattern Recognition Letters, 31(8), 651–666. https://doi.org/10.1016/j.patrec.2009.09.011

Thorndike, R. L. (1953). Who belongs in the family? Psychometrika, 18(4), 267–276. https://doi.org/10.1007/BF02289263

Rousseeuw, P. J. (1987).
Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics, 20, 53–65. https://doi.org/10.1016/0377-0427(87)90125-7

Berman, B., & Berman, D. (2013). Retail management: A strategic approach (12th ed.). Pearson.

Dogan, E., Delibas, D., & Cinar, M. (2018). Customer churn analysis in energy markets using machine learning techniques. Energy Economics, 75, 108–120. https://doi.org/10.1016/j.eneco.2018.07.015

Fader, P. S., Hardie, B. G. S., & Lee, K. L. (2005). RFM and CLV: Using iso-value curves for customer base analysis. Journal of Marketing Research, 42(4), 415–430. https://doi.org/10.1509/jmkr.2005.42.4.415

Kotler, P., Keller, K. L., Ancarani, F., & Costabile, M. (2017). Marketing management (15th ed.). Pearson Education.

Kumar, V., Reinartz, W., & Peterson, J. A. (2011). Customer relationship management: Concept, strategy, and tools. Springer.

Peppers, D., & Rogers, M. (2016). Managing customer experience and relationships: A strategic framework (3rd ed.). Wiley.

Verbeke, W., Martens, D., & Baesens, B. (2014). Social network analysis for customer churn prediction. Applied Soft Computing, 14, 431–446. https://doi.org/10.1016/j.asoc.2013.09.028

Wedel, M., & Kamakura, W. A. (2012). Market segmentation: Conceptual and methodological foundations (2nd ed.). Springer.

Zhang, Y., Liang, X., & Wang, Z. (2019). Customer churn prediction in the electricity market using gradient boosting decision trees. Energy Policy, 125, 166–177. https://doi.org/10.1016/j.enpol.2018.10.038

Aggarwal, C. C. (2015). Data mining: The textbook. Springer. https://doi.org/10.1007/978-3-319-14142-8

Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5–32. https://doi.org/10.1023/A:1010933404324

Dogan, E., Delibaş, D., & Cinar, M. (2018). Customer churn analysis in energy markets using machine learning techniques. Energy Economics, 75, 108–120. https://doi.org/10.1016/j.eneco.2018.07.015

Hastie, T., Tibshirani, R., & Friedman, J. (2017). The elements of statistical learning: Data mining, inference, and prediction (2nd ed.). Springer. https://doi.org/10.1007/978-0-387-84858-7

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). An introduction to statistical learning (2nd ed.). Springer. https://doi.org/10.1007/978-1-0716-1418-1

Kotler, P., Keller, K. L., Ancarani, F., & Costabile, M. (2017). Marketing management (15th ed.). Pearson Education.

Verbeke, W., Martens, D., & Baesens, B. (2014). Social network analysis for customer churn prediction. Applied Soft Computing, 14, 431–446. https://doi.org/10.1016/j.asoc.2013.09.028

Zhang, Y., Liang, X., & Wang, Z. (2019). Customer churn prediction in the electricity market using gradient boosting decision trees. Energy Policy, 125, 166–177. https://doi.org/10.1016/j.enpol.2018.10.038

Zhao, H., Li, N., & Li, Y. (2017). Electricity customer behavior analysis based on smart meter data. IEEE Transactions on Smart Grid, 8(2), 1110–1120. https://doi.org/10.1109/TSG.2015.2499761

Buckinx, W., & Van den Poel, D. (2005). Customer base analysis: Partial defection of behaviorally loyal clients. Journal of Marketing Research, 42(3), 252–268. https://doi.org/10.1509/jmkr.42.3.252

Neslin, S. A., Gupta, S., Kamakura, W., Lu, J., & Mason, C. H. (2006). Defection detection: Measuring and understanding the predictive accuracy of customer churn models. Journal of Marketing Research, 43(2), 204–211. https://doi.org/10.1509/jmkr.43.2.204

Verbeke, W., Dejaeger, K., Martens, D., Hur, J., & Baesens, B. (2012). New insights into churn prediction in the telecommunication sector: A profit driven data mining approach. Decision Support Systems, 52(3), 708–720. https://doi.org/10.1016/j.dss.2011.09.003

Fader, P. S., Hardie, B. G. S., & Lee, K. L. (2005). RFM and CLV: Using iso-value curves for customer base analysis. Journal of Marketing Research, 42(4), 415–430. https://doi.org/10.1509/jmkr.42.4.415

6.2 Παράρτημα Α – Χρήση Κώδικα

Βιβλιοθήκες

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns

from sklearn.cluster import KMeans
from sklearn.preprocessing import StandardScaler
from sklearn.model_selection import train_test_split, GridSearchCV
from sklearn.linear_model import LogisticRegression
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import classification_report, roc_auc_score, RocCurveDisplay
from sklearn.decomposition import PCA

from imblearn.over_sampling import SMOTE

np.random.seed(42)
```

Κώδικας 1. Δημιουργία συνόλου δεδομένων – Προσομοίωση Δεδομένων

```
def simulate_energy_customers(n_customers=3000, months=12):
    rows = []

    for cid in range(1, n_customers+1):

        base_usage = np.random.uniform(150, 800)

        for m in range(months):
            seasonal = 1 + 0.3*np.sin(m/12 * 2*np.pi)
            noise = np.random.normal(0, 40)

            kwh = max(30, base_usage*seasonal + noise)
            bill = kwh * np.random.uniform(0.18, 0.26)

            rows.append([cid, m, kwh, bill])

    return pd.DataFrame(rows, columns=["customer_id","month","kwh","bill_amount"])

usage_df = simulate_energy_customers()
```

Κώδικας 2. RFM Feature Engineering

```
rfm_df = usage_df.groupby("customer_id").agg(
    frequency=("month","count"),
    monetary_kwh=("kwh","sum"),
    monetary_bill=("bill_amount","sum"),
)

rfm_df["recency_days"] = np.random.randint(1,180,len(rfm_df))
rfm_df["avg_kwh"] = rfm_df["monetary_kwh"]/rfm_df["frequency"]

rfm_df.reset_index(inplace=True)
```

Κώδικας 3. Δημιουργία CHURN

```python
def create_churn_label_from_usage(df):

    r = df['recency_days']/df['recency_days'].max()
    f = df['frequency']/df['frequency'].max()
    m = df['monetary_kwh']/df['monetary_kwh'].max()

    logits = (
        2.8*r           # παλιοί πελάτες φεύγουν
        -1.5*f          # συχνοί μένουν
        -2.0*m          # υψηλής αξίας μένουν
        + np.random.normal(0,0.9,len(df))
    )

    prob = 1/(1+np.exp(-logits))

    # threshold για ~70% churn
    df["churn"] = (prob > 0.35).astype(int)

    return df

rfm_df = create_churn_label_from_usage(rfm_df)

rfm_df["churn"].value_counts(normalize=True)
```

Κώδικας 4. Clustering (για features)

```python
features = ["recency_days","frequency","monetary_kwh","avg_kwh"]

scaler = StandardScaler()
X_scaled = scaler.fit_transform(rfm_df[features])

kmeans = KMeans(n_clusters=3, random_state=42, n_init=10)
rfm_df["cluster"] = kmeans.fit_predict(X_scaled)
```

Κώδικας 5. Dataset για πρόβλεψη

```python
y = rfm_df["churn"]
X = rfm_df.drop(columns=["customer_id","churn"])
```

Κώδικας 6. Train/Test Split

```
X_train, X_test, y_train, y_test = train_test_split(
    X, y, test_size=0.2, random_state=42, stratify=y
)
```

Κώδικας 6. SMOTE (μόνο training)

```
scaler_model = StandardScaler()

X_train_scaled = scaler_model.fit_transform(X_train_res)
X_test_scaled = scaler_model.transform(X_test)
```

Κώδικας 7. Logistic Regression + Grid Search

```
param_lr = {'C':[0.01,0.1,1,10]}

grid_lr = GridSearchCV(
    LogisticRegression(max_iter=2000),
    param_lr,
    cv=5,
    scoring="f1"
)

grid_lr.fit(X_train_scaled, y_train_res)
best_lr = grid_lr.best_estimator_
```

Κώδικας 8. Random Forest

```
param_rf = {
    'n_estimators':[100,200],
    'max_depth':[4,6,8],
    'min_samples_split':[2,5]
}

grid_rf = GridSearchCV(
    RandomForestClassifier(random_state=42),
    param_rf,
    cv=5,
    scoring="f1"
)

grid_rf.fit(X_train_res, y_train_res)
best_rf = grid_rf.best_estimator_
```

Κώδικας 9. Evaluation

```
y_pred_lr = best_lr.predict(X_test_scaled)
y_prob_lr = best_lr.predict_proba(X_test_scaled)[:,1]

y_pred_rf = best_rf.predict(X_test)
y_prob_rf = best_rf.predict_proba(X_test)[:,1]

print("LOGISTIC")
print(classification_report(y_test,y_pred_lr))
print("AUC:",roc_auc_score(y_test,y_prob_lr))

print("\nRANDOM FOREST")
print(classification_report(y_test,y_pred_rf))
print("AUC:",roc_auc_score(y_test,y_prob_rf))
```

Κώδικας 10. ROC Curve

```
RocCurveDisplay.from_predictions(y_test,y_prob_lr,name="Logistic")
RocCurveDisplay.from_predictions(y_test,y_prob_rf,name="Random Forest")
plt.show()
```

Κώδικας 11. Feature Importance

Random Forest

```
rf_importance = pd.Series(best_rf.feature_importances_,index=X.columns).sort_values(ascending=Fal
rf_importance.plot(kind="bar",title="Random Forest Importance")
plt.show()
rf_importance
```

Logistic

```
lr_coeff = pd.Series(best_lr.coef_[0],index=X.columns).sort_values()
lr_coeff.plot(kind="barh",title="Logistic Coefficients")
plt.show()
lr_coeff
```

Κώδικας 12.Scaling

```
from sklearn.preprocessing import StandardScaler

scaler = StandardScaler()

X_train_scaled = scaler.fit_transform(X_train_res)
X_test_scaled = scaler.transform(X_test)
```

Κώδικας 13. Elbow, Silhouette

```
def evaluate_kmeans(Xs, k_range=range(2, 11)):
    results = []
    for k in k_range:
        km = KMeans(n_clusters=k, random_state=RANDOM_STATE, n_init="auto")
        labels = km.fit_predict(Xs)
        sil = silhouette_score(Xs, labels)
        inertia = km.inertia_
        results.append({"k": k, "silhouette": sil, "inertia": inertia})
    return pd.DataFrame(results)

def fit_kmeans(Xs, k):
    km = KMeans(n_clusters=k, random_state=RANDOM_STATE, n_init="auto")
    labels = km.fit_predict(Xs)
    return km, labels
```

6.3 Παράρτημα Β - Πίνακας γραφημάτων

Εικόνα 1. Στάδια μεθοδολογίας……21

Εικόνα 2. Περιγραφή Συνόλου Δεδομένων Πελατών Ηλεκτρικής Ενέργειας……23

Εικόνα 3. Μορφή dataset που εισάγεται στο μοντέλο……24

Εικόνα 4: Κατανομή μηνιαίας κατανάλωσης ενέργειας πελατών……24

Εικόνα 5: Μέση κατανάλωση ενέργειας ανά μήνα κατά τη χρονική περίοδο παρατήρησης……25

Εικόνα 6. Κατανομή συνολικής κατανάλωσης ανά πελάτη……26

Εικόνα 7. Κατανομή χρονικής απόστασης από την τελευταία καταγεγραμμένη δραστηριότητα πελάτη……27

Εικόνα 8. Preview του πίνακα RFM ……28

Εικόνα 9. BOX PLOT κατανομής RFM χαρακτηριστικών πριν την τυποποίηση……29

Εικόνα 10. BOX PLOT κατανομής RFM χαρακτηριστικών μετά την τυποποίηση……29

Εικόνα 11. Επιλογή αριθμού συστάδων με βάση την αδράνεια (inertia)……30

Εικόνα 12. Επιλογή αριθμού συστάδων με βάση τον δείκτη silhouette……31

Εικόνα 13. Κατανομή πελατών ως προς την κατάσταση δραστηριότητας (ενεργοί – αποχωρήσαντες)……33

Εικόνα 14. Ποσοστιαία κατανομή ενεργών και αποχωρησάντων πελατών……33

Εικόνα 15.Κατανομή των κλάσεων αποχώρησης στα σύνολα εκπαίδευσης και ελέγχου μετά τον διαχωρισμό δεδομένων. Παρατηρείται διατήρηση της αναλογίας των κλάσεων
(stratifiedsplit)……
……35

Εικόνα 16. Κατανομή των κλάσεων στο σύνολο εκπαίδευσης πριν και μετά την εφαρμογή της τεχνικής SMOTE. Παρατηρείται εξισορρόπηση της μειοψηφικής κατηγορία……..……
……..36

Εικόνα 17. Κατανομή της μεταβλητής recency πριν και μετά την τυποποίηση με StandardScaler. Παρατηρείται διατήρηση της μορφής της κατανομής και μεταφορά της γύρω από το
μηδέν…….…….3
7

Εικόνα 18: Κατανομή πελατών στον χώρο Recency–Monetary ανά συστάδα……..……...……
…….....……39

Εικόνα 19. Συγκριτικό προφίλ συμπεριφοράς πελατών ανά συστάδα βάσει των χαρακτηριστικώνRFM……
……41

Εικόνα 20: Απεικόνιση των συστάδων στον χώρο των δύο πρώτων κύριων συνιστωσών……..…….……
…….…..42

Εικόνα 21. Ποσοστό αποχώρησης πελατών ανά συστάδα……...……..……..……
……..44

Εικόνα 22. Αποτελέσματα ταξινόμησης (precision, recall, F1-score και AUC) για το μοντέλο Λογιστικής Παλινδρόμησης στο σύνολο
ελέγχου……
…….45

Εικόνα 23. Καμπύλη ROC για το μοντέλο Λογιστικής Παλινδρόμησης στο σύνολο ελέγχου……
……..……..46

Εικόνα 24. Αποτελέσματα ταξινόμησης (precision, recall, F1-score και AUC) για το μοντέλο Random Forest στο σύνολο
ελέγχου……..……
……...46

Εικόνα 25. Καμπύλη ROC για το μοντέλο Τυχαίου Δάσους στο σύνολο ελέγχου……..……
……..47

Εικόνα 26: Σημαντικότητα χαρακτηριστικών στο μοντέλο Random Forest……
…….48