



ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ
ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ
ΤΟΜΕΑΣ ΗΛΕΚΤΡΙΚΩΝ ΒΙΟΜΗΧΑΝΙΚΩΝ ΔΙΑΤΑΞΕΩΝ
ΚΑΙ ΣΥΣΤΗΜΑΤΩΝ ΑΠΟΦΑΣΕΩΝ

Τεχνικές Ομαδοποίησης με Χρήση Μηχανικής Μάθησης: Εφαρμογή σε Δεδομένα Έργων Ανακαίνισης Κτιρίων

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Ιωάννα Σλλάκου

Επιβλέπων : Ευάγγελος Μαρινάκης
Επίκουρος Καθηγητής ΕΜΠ

Αθήνα, Ιούλιος 2024



ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ
ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ
ΤΟΜΕΑΣ ΗΛΕΚΤΡΙΚΩΝ ΒΙΟΜΗΧΑΝΙΚΩΝ ΔΙΑΤΑΞΕΩΝ
ΚΑΙ ΣΥΣΤΗΜΑΤΩΝ ΑΠΟΦΑΣΕΩΝ

Τεχνικές Ομαδοποίησης με Χρήση Μηχανικής Μάθησης: Εφαρμογή σε Δεδομένα Έργων Ανακαίνισης Κτιρίων

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Ιωάννα Σλλάκου

Επιβλέπων : Ευάγγελος Μαρινάκης

Επίκουρος Καθηγητής ΕΜΠ

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 16/07 2024.

Μαρινάκης Ε.

Ψαρράς Ι.

Ασκούνης Δ.

Επίκουρος Καθηγητής ΕΜΠ

Καθηγητής ΕΜΠ

Καθηγητής ΕΜΠ

Αθήνα, Ιούλιος 2024

.....
Ιωάννα Σλλάκου

Διπλωματούχος Ηλεκτρολόγος Μηχανικός και Μηχανικός Υπολογιστών Ε.Μ.Π.

Copyright © Ιωάννα Σλλάκου, 2024

Με επιφύλαξη παντός δικαιώματος. All rights reserved.

Απαγορεύεται η αντιγραφή, αποθήκευση και διανομή της παρούσας εργασίας, εξ ολοκλήρου ή τμήματος αυτής, για εμπορικό σκοπό. Επιτρέπεται η ανατύπωση, αποθήκευση και διανομή για σκοπό μη κερδοσκοπικό, εκπαιδευτικής ή ερευνητικής φύσης, υπό την προϋπόθεση να αναφέρεται η πηγή προέλευσης και να διατηρείται το παρόν μήνυμα. Ερωτήματα που αφορούν τη χρήση της εργασίας για κερδοσκοπικό σκοπό πρέπει να απευθύνονται προς τον συγγραφέα.

Οι απόψεις και τα συμπεράσματα που περιέχονται σε αυτό το έγγραφο εκφράζουν τον συγγραφέα και δεν πρέπει να ερμηνευθεί ότι αντιπροσωπεύουν τις επίσημες θέσεις του Εθνικού Μετσόβιου Πολυτεχνείου.

Περίληψη

Οι εκπομπές ρύπων από τα κτίρια αποτελούν σημαντική πηγή περιβαλλοντικής ρύπανσης, συμβάλλοντας στην κλιματική αλλαγή και την υποβάθμιση της ποιότητας του αέρα. Τα κτίρια καταναλώνουν μεγάλες ποσότητες ενέργειας για θέρμανση, ψύξη και φωτισμό, με αποτέλεσμα την εκπομπή διοξειδίου του άνθρακα και άλλων αερίων του θερμοκηπίου. Οι ανακαινίσεις των κτιρίων παίζουν κρίσιμο ρόλο στη μείωση αυτών των εκπομπών, καθώς μπορούν να βελτιώσουν την ενεργειακή απόδοση.

Στόχος της παρούσας διπλωματικής εργασίας είναι η εφαρμογή μοντέλων τεχνητής νοημοσύνης και μηχανικής μάθησης για την υποστήριξη της λήψης αποφάσεων σχετικά με τις καλύτερες επιλογές ανακαίνισης κτιρίων. Συγκεκριμένα, υλοποιούνται αλγόριθμοι που ομαδοποιούν δεδομένα κτιρίων στα οποία έχουν γίνει ανακαινίσεις. Αυτή η ομαδοποίηση επιτρέπει στους υπεύθυνους αποφάσεων να συγκρίνουν το κτίριό τους με παρόμοια κτίρια και να επιλέξουν τις βέλτιστες στρατηγικές μείωσης του περιβαλλοντικού τους αποτυπώματος. Χρησιμοποιούνται δεδομένα από κτίρια στη Λετονία πριν και μετά τις ενεργειακές ανακαινίσεις, τα οποία εισάγονται σε τέσσερα διαφορετικά μοντέλα. Οι αλγόριθμοι μηχανικής μάθησης που συγκρίνονται είναι οι K-Means, DBSCAN, HDBSCAN και Agglomerative Clustering.

Τα εξαγόμενα συμπεράσματα βοηθούν στην αναγνώριση των διαφόρων προκλήσεων που έχει η ομαδοποίηση ενός συνόλου δεδομένων από ανακαινίσεις κτιρίων αλλά και στον εντοπισμό διάφορων σημαντικών χαρακτηριστικών με βάση τα οποία έγινε η κατηγοριοποίηση. Τα εν λόγω αποτελέσματα και συμπεράσματα έχουν ως σκοπό να ενισχύσουν σημαντικά την διαδικασία ανακαινίσεων των κτιρίων και τη μείωση εκπομπών ρύπων από αυτά.

Λέξεις Κλειδιά: Ανακαίνιση Κτιρίων, Ενεργειακή Αποδοτικότητα, Μη Εποπτευόμενη Μηχανική Μάθηση, Ομαδοποίηση, KMeans, DBSCAN, HDBSCAN.

Abstract

Emissions from buildings are a major source of environmental pollution, contributing to climate change and air quality degradation. Buildings consume large amounts of energy for heating, cooling and lighting, resulting in the emission of carbon dioxide and other greenhouse gases. Building renovations play a crucial role in reducing these emissions as they can improve energy efficiency.

The aim of this thesis is to apply artificial intelligence and machine learning models to support decision making on the best building renovation options. In particular, algorithms are implemented that cluster data of buildings in which renovations have been carried out. This clustering allows decision makers to compare their building with similar buildings and choose the best strategies to reduce their environmental footprint. Data from buildings in Latvia before and after energy renovations are used, which are input into three different models. The machine learning algorithms compared are K-Means, DBSCAN, HDBSCAN and Agglomerative Clustering.

The conclusions drawn help identify the challenges of clustering a dataset of building renovations and identify various important characteristics on which the categorization was made. These results and conclusions are intended to significantly enhance the process of building renovations and the reduction of pollutant emissions from them.

Keywords: Building Renovation, Energy Efficiency, Unsupervised Machine Learning, Clustering, Kmeans, DBSCAN, HDBSCAN.

Πίνακας Περιεχομένων

1	Εισαγωγή.....	18
1.1	Αντικείμενο-Σκοπός	18
1.2	Φάσεις Υλοποίησης.....	18
1.3	Οργάνωση Εργασίας	19
2	Εκπομπές CO ₂ από Κτιριακές Δομές και Κλιματική Αλλαγή	23
2.1	Κτιριακός Τομέας Σήμερα.....	23
2.2	Στόχοι και Δράσεις Ευρωπαϊκής Ένωσης	23
2.3	Εκπομπές CO ₂ από Κτίρια	24
2.4	Χρήση Νέων Τεχνολογιών και Εκμετάλλευση Δεδομένων.....	26
3	Σύγχρονες Προσεγγίσεις στην Ομαδοποίηση Κτιριακών Δεδομένων.....	29
3.1	Μοντέλα Μηχανικής Μάθησης	29
3.1.1	Hierarchical Clustering.....	29
3.1.2	KMeans	30
3.1.3	Model-based Clustering	30
3.1.4	PAM	30
3.2	Σχετικές Μελέτες.....	31
3.2.1	Ομαδοποίηση Κτιρίων με δεδομένα από τη Βιέννη.....	31
3.2.2	Προσδιορισμός αντιπροσωπευτικών κτιρίων και ομάδων με δεδομένα από τη Γένοβα 32	
3.2.3	Ομαδοποίηση Κτιρίων για Προσομοιώσεις	34
3.2.4	Ομαδοποίηση Fuzzy Cmeans για Σύνολα Δεδομένων με Θόρυβο, Outliers και Άνισο Αριθμό Ομάδων	34
3.2.5	Ομαδοποίηση Αστικών Hotspots	36
4	Προτεινόμενη Μεθοδολογία.....	39
4.1	Καθορισμός του Προβλήματος	41
4.2	Προτεινόμενη Μεθοδολογία	41
4.3	Στόχος Προτεινόμενης Μεθοδολογίας.....	41
5	Εφαρμογή Προτεινόμενης Μεθοδολογίας	46
5.1	Μηχανική Μάθηση	46
5.1.1	KMeans	48
5.1.2	DBSCAN.....	48
5.1.3	HDBSCAN	49
5.1.4	Spectral Clustering	49
5.1.5	GMM (Gaussian Mixture Models)	49
5.1.6	Agglomerative Hierarchical Clustering	50

5.1.7	Mean Shift Clustering.....	50
5.1.8	Affinity Propagation	50
5.1.9	BIRCH (Balanced Iterative Reducing and Clustering using Hierarchies).....	51
5.2	Δεδομένα.....	51
5.2.1	Επιλογή Σημαντικότερων Χαρακτηριστικών	58
5.2.2	Προ επεξεργασία Δεδομένων	59
5.3	Αλγόριθμοι Ομαδοποίησης	59
5.3.1	Agglomerative Clustering.....	60
5.3.2	DBSCAN.....	68
5.3.3	K-Means	70
5.3.4	HDBSCAN	74
6	Συμπεράσματα και Προοπτικές.....	84
6.1	Συμπεράσματα	84
6.2	Προοπτικές	86
7	Βιβλιογραφία	89

Ευρετήριο Πινάκων

Πίνακας 1:	Σύγκριση πιο Συχνά Χρησιμοποιούμενων Αλγορίθμων Ομαδοποίησης.....	47
Πίνακας 2:	Συνοπτική Απεικόνιση Δεδομένων από Ανακαίνισεις Κτιρίων στην Λετονία	53
Πίνακας 3:	Απεικόνιση περιγραφικών και στατιστικών στοιχείων των αριθμητικών τιμών του υποσυνόλου δεδομένων από έργα ανακαίνισης στη Λετονία.....	53
Πίνακας 4:	Πίνακας συσχέτισης των σημαντικότερων χαρακτηριστικών του συνόλου δεδομένων με έργα ανακαίνισης που πραγματοποιήθηκαν στη Λετονία	58

Ευρετήριο Σχημάτων

Σχήμα 1: Προτεινόμενο Μεθοδολογικό Πλαίσιο για την Μείωση των Εκπομπών Ρύπων Μέσω Ανακαινίσεων Κτιρίων με τη Βοήθεια της Μηχανικής Μάθησης	41
Σχήμα 2: Διαγράμματα με την Κατανομή των τιμών κάθε μεταβλητής του συνόλου δεδομένων από έργα ανακαίνισης στη Λετονία	55
Σχήμα 3: Boxplots για κάθε μεταβλητή του συνόλου δεδομένων που επιλέχτηκε από έργα ανακαίνισης στη Λετονία	57
Σχήμα 4: Εφαρμογή Agglomerative Clustering με είσοδο energy-age, co2-age, heating-age	61
Σχήμα 5: Εφαρμογή Agglomerative Clustering με είσοδο co2-heating, co2-energy, energy-heating	61
Σχήμα 6: Γραμμικά διαγράμματα χαρακτηριστικών μετά την ομαδοποίηση με Agglomerative Clustering με είσοδο age-heating-energy και 3D διάγραμμα με την απεικόνιση και των τριών χαρακτηριστικών	61
Σχήμα 7: Γραμμικά διαγράμματα χαρακτηριστικών μετά την ομαδοποίηση με Agglomerative Clustering με είσοδο age-heating-floors και 3D διάγραμμα με την απεικόνιση και των τριών χαρακτηριστικών	62
Σχήμα 8: Γραμμικά διαγράμματα χαρακτηριστικών μετά την ομαδοποίηση με Agglomerative Clustering με είσοδο age-heating-temperature και 3D διάγραμμα με την απεικόνιση και των τριών χαρακτηριστικών	62
Σχήμα 9: Γραμμικά διαγράμματα χαρακτηριστικών μετά την ομαδοποίηση με Agglomerative Clustering με είσοδο heating-energy-floors και 3D διάγραμμα με την απεικόνιση και των τριών χαρακτηριστικών	62
Σχήμα 10: Γραμμικά διαγράμματα χαρακτηριστικών μετά την ομαδοποίηση με Agglomerative Clustering με είσοδο heating-energy-temperature και 3D διάγραμμα με την απεικόνιση και των τριών χαρακτηριστικών	62
Σχήμα 11: Γραμμικά διαγράμματα χαρακτηριστικών μετά την ομαδοποίηση με Agglomerative Clustering με είσοδο heating- co2 -floors και 3D διάγραμμα με την απεικόνιση και των τριών χαρακτηριστικών	63
Σχήμα 12: Γραμμικά διαγράμματα χαρακτηριστικών μετά την ομαδοποίηση με Agglomerative Clustering με είσοδο heating- co2 -temperature και 3D διάγραμμα με την απεικόνιση και των τριών χαρακτηριστικών	63
Σχήμα 13: Γραμμικά διαγράμματα χαρακτηριστικών μετά την ομαδοποίηση με Agglomerative Clustering με είσοδο heating-floors-temperature και 3D διάγραμμα με την απεικόνιση και των τριών χαρακτηριστικών	63
Σχήμα 14: Γραμμικά διαγράμματα χαρακτηριστικών μετά την ομαδοποίηση με Agglomerative Clustering με είσοδο age-energy-floors και 3D διάγραμμα με την απεικόνιση και των τριών χαρακτηριστικών	64
Σχήμα 15: Γραμμικά διαγράμματα χαρακτηριστικών μετά την ομαδοποίηση με Agglomerative Clustering με είσοδο age-energy-temperature και 3D διάγραμμα με την απεικόνιση και των τριών χαρακτηριστικών	64
Σχήμα 16: Γραμμικά διαγράμματα χαρακτηριστικών μετά την ομαδοποίηση με Agglomerative Clustering με είσοδο energy-floors-temperature και 3D διάγραμμα με την απεικόνιση και των τριών χαρακτηριστικών	64
Σχήμα 17: Γραμμικά διαγράμματα χαρακτηριστικών μετά την ομαδοποίηση με Agglomerative Clustering με είσοδο age-co2-floors και 3D διάγραμμα με την απεικόνιση και των τριών χαρακτηριστικών	65

Σχήμα 18: Γραμμικά διαγράμματα χαρακτηριστικών μετά την ομαδοποίηση με Agglomerative Clustering με είσοδο age-co2-temperature και 3D διάγραμμα με την απεικόνιση και των τριών χαρακτηριστικών	65
Σχήμα 19: Γραμμικά διαγράμματα χαρακτηριστικών μετά την ομαδοποίηση με Agglomerative Clustering με είσοδο co2-floors-temperature και 3D διάγραμμα με την απεικόνιση και των τριών χαρακτηριστικών	65
Σχήμα 20: : Γραμμικά διαγράμματα χαρακτηριστικών μετά την ομαδοποίηση με Agglomerative Clustering με είσοδο age-floors-temperature και 3D διάγραμμα με την απεικόνιση και των τριών χαρακτηριστικών	65
Σχήμα 21: Γραμμικά διαγράμματα χαρακτηριστικών μετά την ομαδοποίηση με Agglomerative Clustering με είσοδο age-heating-co2και 3D διάγραμμα με την απεικόνιση και των τριών χαρακτηριστικών	66
Σχήμα 22: Γραμμικά διαγράμματα χαρακτηριστικών μετά την ομαδοποίηση με Agglomerative Clustering με είσοδο heating-energy-co2 και 3D διάγραμμα με την απεικόνιση και των τριών χαρακτηριστικών	66
Σχήμα 23: Γραμμικά διαγράμματα χαρακτηριστικών μετά την ομαδοποίηση με Agglomerative Clustering με είσοδο age-energy-co2 και 3D διάγραμμα με την απεικόνιση και των τριών χαρακτηριστικών	66
Σχήμα 24: Γραμμικά διαγράμματα χαρακτηριστικών μετά την ομαδοποίηση με Agglomerative Clustering με είσοδο energy-co2-floors και 3D διάγραμμα με την απεικόνιση και των τριών χαρακτηριστικών	67
Σχήμα 25: Γραμμικά διαγράμματα χαρακτηριστικών μετά την ομαδοποίηση με Agglomerative Clustering με είσοδο energy-co2-temperature και 3D διάγραμμα με την απεικόνιση και των τριών χαρακτηριστικών	67
Σχήμα 26: Γραμμικά διαγράμματα χαρακτηριστικών μετά την ομαδοποίηση με Agglomerative Clustering με είσοδο age-heating-energy-temperature-floors- co2	68
Σχήμα 27: Γραμμικό διάγραμμα age-heating μετά την ομαδοποίηση με DBSCAN με είσοδο τα χαρακτηριστικά age-heating, min_smamples=14 και eps=1528.802943482252	69
Σχήμα 28: Γραμμικό διάγραμμα μετά την ομαδοποίηση με DBSCAN με είσοδο τα χαρακτηριστικά age-energy-heating, min_samples=7 και eps=0.6	70
Σχήμα 29: Διάγραμμα Elbow για τα χαρακτηριστικά age-heating	71
Σχήμα 30: Γραμμικό διάγραμμα μετά την ομαδοποίηση με KMeans με είσοδο τα χαρακτηριστικά age-heating και k=4	71
Σχήμα 31: Διάγραμμα Elbow για τα χαρακτηριστικά age-energy	71
Σχήμα 32: Γραμμικό διάγραμμα μετά την ομαδοποίηση με KMeans με είσοδο τα χαρακτηριστικά age-energy και k=4	72
Σχήμα 33: Διάγραμμα Elbow για τα χαρακτηριστικά age- co2 και Γραμμικό διάγραμμα μετά την ομαδοποίηση με KMeans με είσοδο τα χαρακτηριστικά age- co2 και k=4	72
Σχήμα 34: Διάγραμμα Elbow για τα χαρακτηριστικά energy-heating και Γραμμικό διάγραμμα μετά την ομαδοποίηση με KMeans με είσοδο τα χαρακτηριστικά energy-heating και k=4	73
Σχήμα 35: Διάγραμμα Elbow για τα χαρακτηριστικά heating- co2 και Γραμμικό διάγραμμα μετά την ομαδοποίηση με KMeans με είσοδο τα χαρακτηριστικά heating- co2 και k=3	73
Σχήμα 36: Διάγραμμα Elbow για τα χαρακτηριστικά energy- co2 και Γραμμικό διάγραμμα μετά την ομαδοποίηση με KMeans με είσοδο τα χαρακτηριστικά energy- co2 και k=5	74
Σχήμα 37: Γραμμικά διαγράμματα χαρακτηριστικών μετά την ομαδοποίηση με HDBSCAN με είσοδο age-heating-energy και 3D διάγραμμα με την απεικόνιση και των τριών χαρακτηριστικών	75

Σχήμα 54: Γραμμικά διαγράμματα χαρακτηριστικών μετά την ομαδοποίηση με HDBSCAN με είσοδο energy-co2-temperature και 3D διάγραμμα με την απεικόνιση και των τριών χαρακτηριστικών	81
Σχήμα 55: Γραμμικά διαγράμματα χαρακτηριστικών μετά την ομαδοποίηση με HDBSCAN με είσοδο energy-floors-temperature και 3D διάγραμμα με την απεικόνιση και των τριών χαρακτηριστικών	81
Σχήμα 56: Γραμμικά διαγράμματα χαρακτηριστικών μετά την ομαδοποίηση με HDBSCAN με είσοδο co2-floors-temperature και 3D διάγραμμα με την απεικόνιση και των τριών χαρακτηριστικών	81

Πρόλογος

Η συγκεκριμένη διπλωματική εργασία εκπονήθηκε στο πλαίσιο των ερευνητικών δραστηριοτήτων του Εργαστηρίου Συστημάτων Αποφάσεων και Διοίκησης, κατά το ακαδημαϊκό έτος 2023-2024 και την περίοδο Νοεμβρίου 2023-Ιουλίου 2024. Το Εργαστήριο υπάγεται στον Τομέα Ηλεκτρικών Βιομηχανικών Διατάξεων και Συστημάτων Αποφάσεων της Σχολής Ηλεκτρολόγων Μηχανικών και Μηχανικών Ηλεκτρονικών Υπολογιστών του Εθνικού Μετσόβιου Πολυτεχνείου (ΕΜΠ). Η εργασία πραγματοποιήθηκε υπό την επίβλεψη του κ. Ευάγγελου Μαρινάκη, Επίκουρου Καθηγητή ΕΜΠ, στην Σχολή Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών, τον οποίο ευχαριστώ ιδιαίτερα για την ευκαιρία που μου έδωσε να ασχοληθώ με το συγκεκριμένο θέμα.

Στόχος της διπλωματικής εργασίας είναι η εφαρμογή μοντέλων τεχνητής νοημοσύνης και μηχανικής μάθησης, με απώτερο σκοπό την μείωση του CO₂ που εκπέμπουν οι κτιριακές δομές. Τα εν λόγω μοντέλα θα πρέπει να έχουν τη δυνατότητα λαμβάνοντας δεδομένα από προηγούμενα κτίρια στα οποία πραγματοποιήθηκαν δράσεις εξοικονόμησης ενέργειας, να βρίσκουν τη βέλτιστη στρατηγική ανακαίνισης για το εκάστοτε υποψήφιο έργο. Μέσω της ανάλυσης των δεδομένων και της εκπαίδευσης μοντέλων μη εποπτευόμενης μηχανικής μάθησης επιχειρείται η ομαδοποίηση των κτιρίων που θα βοηθήσει στην καλύτερη λήψη αποφάσεων μελλοντικών ανακαινίσεων με σκοπό την εύρεση των βέλτιστων δράσεων μείωσης βλαβερών περιβαλλοντικών ρύπων.

Θα ήθελα να ευχαριστήσω την Ντανιέλα Στόιαν και την Ιωάννα Ανδρεουλάκη, υποψήφιες Διδάκτορες του ΕΜΠ, για την ουσιαστική και καθοριστικής σημασίας συμβολή τους στην εκπόνηση της διπλωματικής μου εργασίας μέσω των εύστοχων παρατηρήσεων και των εποικοδομητικών προτάσεών τους. Η καθοδήγησή τους υπήρξε καθοριστικής σημασίας στην επιτυχημένη ολοκλήρωση της συγκεκριμένης εργασίας.

Τέλος, ιδιαίτερες ευχαριστίες στο στενό φιλικό κύκλο μου για την συνεχή συμπαράσταση και στήριξη που μου προσέφεραν, προκειμένου να μπορέσω να αντιμετωπίσω επιτυχώς τα διάφορα προβλήματα, τις δυσκολίες και τις προκλήσεις που παρουσιάστηκαν κατά τη διάρκεια των σπουδών μου.

Ιωάννα Σλλάκου
Ιούλιος, 2024

ΚΕΦΑΛΑΙΟ 1

Εισαγωγή

1 Εισαγωγή

1.1 Αντικείμενο-Σκοπός

Η κλιματική αλλαγή αποτελεί επείγουσα απειλή για τον πλανήτη μας. Η ανθρώπινη δραστηριότητα επηρεάζει σταδιακά το κλίμα της γης, προσθέτοντας τεράστιες ποσότητες αερίων του θερμοκηπίου σε εκείνες που απαντώνται φυσιολογικά στην ατμόσφαιρα. Το διοξείδιο του άνθρακα (CO_2) είναι το κυριότερο αέριο του θερμοκηπίου που παράγεται από ανθρώπινες δραστηριότητες όπως η καύση ορυκτών καυσίμων για την παραγωγή ενέργειας, η αποψίλωση των τροπικών δασών, η γεωργία, η κτηνοτροφία και η παραγωγή χημικών ουσιών. Ειδικότερα, τα κτίρια συμβάλουν σε μεγάλο βαθμό στις εκπομπές ρύπων μέσω της εκπομπής διοξειδίου του άνθρακα (CO_2) κατά τη διάρκεια της καύσης ορυκτών καυσίμων για θέρμανση, ψύξη και ηλεκτρική ενέργεια [1].

Ωστόσο, ο ρυθμός των ενεργειακών ανακαινίσεων παραμένει αρκετά χαμηλός, παρ' όλο που ο κτιριακός τομέας αποτελεί τον πιο σημαντικό ενεργειακό καταναλωτή στην Ευρωπαϊκή ένωση. Επομένως, κρίνεται σημαντική η υποστήριξη των ενεργειακών αναβαθμίσεων προκειμένου να ενισχυθεί η ενεργειακή αποδοτικότητα στα κτίρια. Σε αυτό το πλαίσιο, η ένταξη νέων τεχνολογιών στη διαδικασία λήψης αποφάσεων σχετικών με τις δράσεις που μπορούν να μειώσουν την ενεργειακή κατανάλωση στα κτίρια θα μπορούσε να οδηγήσει όχι μόνο στην αύξηση των ενεργειακών ανακαινίσεων, αλλά και στη μεγιστοποίηση της εξοικονόμησης που αναμένεται να προκύψουν από αυτές [1].

Βασικός στόχος της συγκεκριμένης διπλωματικής είναι η εφαρμογή μοντέλων τεχνητής νοημοσύνης και μηχανικής μάθησης, με απώτερο σκοπό τη σύγκριση των διαφορετικών τεχνικών που θα μπορούσαν να εφαρμοστούν για την ομαδοποίηση δεδομένων από κτίρια στα οποία έχει πραγματοποιηθεί ενεργειακή ανακαίνιση. Η ομαδοποίηση θα μπορούσε να συνεισφέρει στην υποστήριξη των αποφασίζοντων στην εύρεση των βέλτιστων δράσεων εξοικονόμησης ενέργειας σε κτιριακές δομές. Ειδικότερα, χάρη στην ομαδοποίηση των κτιρίων δίνεται η δυνατότητα στους αποφασίζοντες (για παράδειγμα ιδιοκτήτες ή διαχειριστές κτιρίων) που επιθυμούν να βελτιώσουν την ενεργειακή αποδοτικότητα ενός κτιρίου, να το συγκρίνουν με τα κτίρια που παρουσιάζουν τη μεγαλύτερη ομοιότητα με αυτό, ώστε να εξετάσουν τις δράσεις που έχουν εφαρμοστεί σε παρόμοια κτίρια προκειμένου να επιλέξουν τη βέλτιστη στρατηγική μείωσης του περιβαλλοντολογικού αποτυπώματος του δικού τους κτιρίου. Στην παρούσα εργασία χρησιμοποιούνται δεδομένα που αποτελούνται από πληροφορίες πριν και μετά από δραστηριότητες εξοικονόμησης ενέργειας οι οποίες εκτελέστηκαν σε κτίρια της Λετονίας [2]. Τα δεδομένα αυτά στη συνέχεια επεξεργάζονται για να χρησιμοποιηθούν ως είσοδος σε τέσσερα διαφορετικά μοντέλα. Οι αλγόριθμοι μηχανικής μάθησης που συγκρίνονται είναι ο αλγόριθμος Kmeans, DBSCAN, HDBSCAN και Agglomerative Clustering. Η μεθοδολογία που χρησιμοποιήθηκε και τα αποτελέσματα που προέκυψαν αναλύονται διεξοδικά στα επόμενα κεφάλαια.

1.2 Φάσεις Υλοποίησης

Η υλοποίηση της παρούσας εργασίας διακρίνεται στις παρακάτω φάσεις, κάθε μια από τις οποίες περιγράφεται συνοπτικά παρακάτω:

1. Αναζήτηση σχετικών ερευνών και εφαρμογών: Αρχικά πραγματοποιείται αναζήτηση παρόμοιων προσπαθειών σχετικών με την ομαδοποίηση κτιρίων με δεδομένα που αφορούν την ενεργειακή ανακαίνιση και την μείωση εκπομπής ρύπων.
2. Μελέτη και Ανάλυση των Δεδομένων: Τα δεδομένα που θα χρησιμοποιηθούν σε αυτή την εργασία προέρχονται από ανακαινίσεις που πραγματοποιήθηκαν στη Λετονία στα πλαίσια της αντιμετώπισης της κλιματικής αλλαγής. Είναι αναγκαίο να μελετηθούν και να αναλυθούν προσεκτικά με σκοπό την καλύτερη κατανόηση τους και τη συσχέτιση τους με το πρόβλημα της μείωσης των εκπομπών ρύπων.
3. Επιλογή μοντέλων μηχανικής μάθησης: Πραγματοποιείται συλλογή και σύγκριση διάφορων αλγορίθμων μηχανικής μάθησης που χρησιμοποιούνται σε προβλήματα ομαδοποίησης. Στη συνέχεια επιλέγονται για διεξαγωγή πειραμάτων οι αλγόριθμοι KMeans, DBSCAN, HDBSCAN και Agglomerative Clustering, καθώς ταιριάζουν καλύτερα στο διαθέσιμο σύνολο δεδομένων.
4. Επιλογή Χαρακτηριστικών: Πραγματοποιείται μία πρώτη προσέγγιση εντοπισμού των σημαντικότερων χαρακτηριστικών που παρουσιάζουν το μεγαλύτερο ενδιαφέρον από τα διαθέσιμα δεδομένα. Συγκεκριμένα επιλέγονται πληροφορίες που αφορούν την κατάσταση των κτιρίων πριν την ανακαίνιση.
5. Προ επεξεργασία Δεδομένων: Σε αυτό το σημείο πραγματοποιείται προ επεξεργασία των δεδομένων που επιλέχθηκαν για να χρησιμοποιηθούν ως είσοδος στα μοντέλα τεχνητής νοημοσύνης που θα συγκριθούν στη συνέχεια.
6. Απεικόνιση και Παρατήρηση Δεδομένων: Για την καλύτερη επιλογή μοντέλων μηχανικής μάθησης, τα επεξεργασμένα δεδομένα οπτικοποιούνται χρησιμοποιώντας διαγράμματα και εντοπίζονται χρήσιμες παρατηρήσεις σχετικά με την κατανομή τους και τη συσχέτιση των δεδομένων απεικονίζοντας τα ανά δύο.
7. Εκπαίδευση των μοντέλων: Στο συγκεκριμένο στάδιο εκπαιδεύονται τα μοντέλα, οπτικοποιούνται και αξιολογούνται τα αποτελέσματα που έδωσε το κάθε εκπαιδευμένο μοντέλο. Τα αποτελέσματα αυτά αναλύονται και συγκρίνονται διεξοδικά ώστε να βρεθεί η καταλληλότερη ομαδοποίηση που θα βοηθήσει στην λήψη αποφάσεων σχετικών με κτιριακά δεδομένα.
8. Εξαγωγή συμπερασμάτων και προτάσεων για περαιτέρω έρευνα: Στο στάδιο αυτό και μετά την αξιολόγηση των αποτελεσμάτων του προηγούμενου σταδίου, αποσαφηνίζεται η καλύτερη ομαδοποίηση, παρουσιάζονται τα συμπεράσματα και επισημαίνονται προοπτικές για περαιτέρω έρευνα.

1.3 Οργάνωση Εργασίας

Η εργασία οργανώνεται ως εξής, αρχικά στο 2ο Κεφάλαιο γίνεται αναφορά στην παρούσα κατάσταση του κτιριακού τομέα και στην συνέχεια παρουσιάζονται τόσο οι μελλοντικοί ενεργειακοί στόχοι που έχει θέσει η ΕΕ προκειμένου να μειωθούν τα υψηλά ποσοστά CO₂ μέσω ανακαινίσεων κτιρίων και να ενισχυθεί η ενεργειακή αποδοτικότητα στα κτίρια, όσο και οι δράσεις που έχουν ήδη υλοποιηθεί προς αυτήν την κατεύθυνση. Επίσης αναλύεται το φαινόμενο των υψηλών ποσοστών CO₂ και η σημασία της εύρεσης καλύτερων στρατηγικών ανακαίνισης κτιρίων. Στο 3ο Κεφάλαιο αναλύονται οι προτεινόμενες μέθοδοι για το πρόβλημα και παρόμοιες προσπάθειες που πραγματοποιήθηκαν στην ομαδοποίηση δεδομένων που αφορούν ενεργειακές ανακαίνισης κτιρίων. Επίσης, γίνεται μια σύντομη

αναφορά στα μοντέλα που χρησιμοποιούνται για την ομαδοποίηση τέτοιων δεδομένων. Στο 4ο Κεφάλαιο προσδιορίζεται το πρόβλημα που πραγματεύεται η παρούσα εργασία και παρουσιάζεται τόσο η προτεινόμενη μεθοδολογία όσο και τυχόν διαφορές και ομοιότητες που υπάρχουν σε σχέση με τις ήδη υπάρχουσες που αναλύθηκαν στο 3ο κεφάλαιο. Επίσης, αναλύεται ο στόχος της προτεινόμενης μεθοδολογίας. Στο 5ο Κεφάλαιο εφαρμόζεται η προτεινόμενη μεθοδολογία. Ειδικότερα, σε πρώτο στάδιο γίνεται μια γενική αναφορά στη μηχανική μάθηση και στους αλγόριθμους που θα μπορούσαν να χρησιμοποιηθούν για την ομαδοποίηση του διαθέσιμου συνόλου δεδομένων που περιέχει δεδομένα από ανακαινίσεις κτιρίων που πραγματοποιήθηκαν στη Λετονία. Σε δεύτερο στάδιο μελετάται το σύνολο δεδομένων και επιλέγονται οι αλγόριθμοι KMeans, DBSCAN, HDBSCAN και Agglomerative Clustering για διάφορα σενάρια πειραμάτων. Ακολουθεί η εξαγωγή των κατάλληλων αποτελεσμάτων και ο σχολιασμός τους. Τα αποτελέσματα αυτά καθώς και διάφορες ακόμη παρατηρήσεις και συμπεράσματα συμπληρώνουν το Κεφάλαιο 6, στο οποίο προτείνονται πιθανές στρατηγικές για την ορθή ομαδοποίηση κτιριακών δεδομένων καθώς και ορισμένες προτάσεις για περαιτέρω έρευνα και μελέτη.

ΚΕΦΑΛΑΙΟ 2

Εκπομπές CO₂ από Κτιριακές Δομές και Κλιματική Αλλαγή

2 Εκπομπές CO₂ από Κτιριακές Δομές και Κλιματική Αλλαγή

Στο παρόν κεφάλαιο γίνεται αναφορά στην τρέχουσα κατάσταση του κτιριακού τομέα. Στη συνέχεια, παρουσιάζονται οι μελλοντικοί ενεργειακοί στόχοι που έχει θέσει η Ευρωπαϊκή Ένωση με σκοπό τη μείωση των υψηλών επιπέδων CO₂ μέσω ανακαινίσεων κτιρίων και την ενίσχυση της ενεργειακής αποδοτικότητας στα κτίρια. Επιπλέον, παρουσιάζονται οι δράσεις που έχουν ήδη υλοποιηθεί προς αυτήν την κατεύθυνση. Αναλύεται επίσης το φαινόμενο των υψηλών επιπέδων CO₂ και η σημασία της ανάπτυξης βελτιωμένων στρατηγικών ανακαίνισης κτιρίων.

2.1 Κτιριακός Τομέας Σήμερα

Ο κτιριακός τομέας γενικά επηρεάζεται από πολλούς παράγοντες, συμπεριλαμβανομένων της οικονομικής κατάστασης, των νομικών και ρυθμιστικών πλαισίων, των δημογραφικών τάσεων και των αναγκών στέγασης, καθώς και των περιβαλλοντικών προκλήσεων[3].

Στην Ευρώπη υπάρχουν πολλές παλιές πόλεις με παλιά κτίρια. Επίσης, πολλά από αυτά δεν είναι κατασκευασμένα για αποδοτική χρήση ενέργειας αλλά και για το θερμότερο κλίμα που παρατηρείται τελευταία[4].

Ακραίες συνθήκες ζέστης και ψύχους, απαιτούν υψηλότερη κατανάλωση ενέργειας για θέρμανση για την προστασία των ανθρώπων. Ωστόσο, η μη αποτελεσματική χρήση ενέργειας, η ενεργειακή φτώχεια και η κακή κατάσταση των κτιρίων δυσχεραίνουν το πρόβλημα. Περίπου το 15% των κατοίκων της Ευρώπης διαμένουν σε σπίτια με στέγαστρα που δεν είναι αδιαβροχοποιημένα ή με υγρά τοιχώματα, δάπεδα ή θεμέλια. Ακόμα, μέχρι και το 39% των κατοίκων ζουν σε κτίρια όπου υπάρχει σήψη στα κουφώματα ή στα δάπεδα[5].

Σήμερα, οι ρυθμοί των ενεργειακών ανακαινίσεων των κτιρίων είναι πολύ αργοί. Το 75% των κτιρίων χρήζει ενεργειακής ανακαίνισης, ενώ πάνω από το 85% των σημερινών κτιρίων θα συνεχίσει πιθανότατα να χρησιμοποιείται μέχρι το 2050. Η μαζική αναβάθμιση των υφιστάμενων κτιρίων στην Ευρώπη θα συμβάλλει στην αύξηση της ενεργειακής αποδοτικότητας των κτιρίων και στην προσαρμογή τους στην κλιματική αλλαγή. Βασικός στόχος της ΕΕ ένωσης αποτελεί η επίτευξη μιας κλιματικά ουδέτερης ΕΕ έως το 2050 [6].

2.2 Στόχοι και Δράσεις Ευρωπαϊκής Ένωσης

Μια αξιοσημείωτη προσπάθεια της Ευρωπαϊκής Ένωσης για τη μείωση των εκπομπών CO₂ και τον καθορισμό ορίων στο ρυθμό αύξησης της θερμοκρασίας ήταν η Συμφωνία του Παρισιού [7]. Σύμφωνα με τη συμφωνία αυτή, η μέση αύξηση της παγκόσμιας θερμοκρασίας θα πρέπει να διατηρηθεί κάτω από τους 2 °C τις επόμενες δεκαετίες. Για την επίτευξη αυτού του στόχου, ο Διεθνής Οργανισμός Ανανεώσιμων Πηγών Ενέργειας (IRENA) προτείνει την πλήρη απεξάρτηση του ενεργειακού τομέα από τον άνθρακα έως το 2050 [8].

Επίσης, μεταξύ του 2005 και του 2020, οι υφιστάμενες πολιτικές και οι θερμότεροι χειμώνες συνέβαλαν στη μείωση κατά 25% των εκπομπών CO₂ από τα κτίρια κατά τη φάση της χρήσης τους. Ο κύριος στόχος είναι να μειωθούν οι εκπομπές αερίων του θερμοκηπίου κατά 55% έως

το 2030. Τα μέτρα ενεργειακής αποδοτικότητας στα κτίρια και οι αλλαγές που απαιτούνται για το 2021 παρέχουν σαφείς κατευθυντήριες γραμμές για την επίτευξη αυτού του στόχου, ενώ η Ευρωπαϊκή Επιτροπή εργάζεται για την προώθηση της προσαρμογής των κτιρίων στην κλιματική αλλαγή [9].

Γενικότερα, στην Ευρωπαϊκή Πράσινη Συμφωνία, η ανακαίνιση τόσο των δημόσιων όσο και των ιδιωτικών κτιρίων επισημάνθηκε ως βασική πρωτοβουλία για την προώθηση της ενεργειακής απόδοσης στον τομέα και την επίτευξη των στόχων. Για την επιδίωξη της ενεργειακής αποδοτικότητας και της οικονομικής ανάπτυξης μετά την πανδημία COVID-19, η Επιτροπή δημοσίευσε το 2020 τη στρατηγική "Ένα κύμα ανακαίνισης για την Ευρώπη - Πρασίνισμα των κτιρίων μας, δημιουργία θέσεων εργασίας, βελτίωση της ζωής" μαζί με ένα σχέδιο δράσης και ένα έγγραφο που παρουσιάζει τη διαθέσιμη χρηματοδότηση της ΕΕ. Η ενισχυμένη, προσβάσιμη και πιο στοχευμένη χρηματοδότηση αποτελεί αρχή του κύματος ανακαίνισης. Η χρηματοδότηση, επίσης για την έρευνα και την καινοτομία και για την αντιμετώπιση των εμποδίων της αγοράς και της διαθέσιμης τεχνικής βοήθειας, θα προκύψει μέσω ενός συνδυασμού άμεσων επενδύσεων και της μόχλευσης ιδιωτικών επενδύσεων. Αυτή η πρωτοβουλία βασίζεται σε διάφορες πτυχές, συμπεριλαμβανομένης:

- Εθνικές Στρατηγικές Ανακαίνισης Κτιρίων: Κάθε χώρα της ΕΕ θα πρέπει να αναπτύξει μια μακροπρόθεσμη στρατηγική για την ανακαίνιση των κτιρίων της, προκειμένου να επιτευχθούν οι στόχοι ενεργειακής απόδοσης και κλιματικής αλλαγής.
- Οδηγία για την Ενεργειακή Απόδοση των Κτιρίων: Η πρωτοβουλία βασίζεται επίσης σε υφιστάμενες οδηγίες της ΕΕ που έχουν ως στόχο τη βελτίωση της ενεργειακής απόδοσης των κτιρίων.
- Εθνικά Σχέδια για την Ενέργεια και το Κλίμα (ΕΣΕΚ): Η πρωτοβουλία αντλεί επίσης έμπνευση από τα εθνικά σχέδια ενέργειας και κλίματος των χωρών της ΕΕ, που περιλαμβάνουν στρατηγικές για την αναβάθμιση της ενεργειακής απόδοσης των κτιρίων [3].

Το κύμα ανακαίνισεων της ΕΕ έχει ως στόχο να διπλασιάσει τουλάχιστον το ετήσιο ποσοστό ενεργειακής ανακαίνισης (που σήμερα εκτιμάται στο 1%) των οικιστικών και μη οικιστικών κτιρίων έως το 2030 και να δρομολογήσει ενεργειακές ανακαίνισεις που θα μπορούσαν να μειώσουν την ενεργειακή κατανάλωση των κτιρίων κατά τουλάχιστον 60% [6].

Επιπλέον βασικό κλειδί για την απαλλαγή των κτιρίων από τις εκπομπές ρύπων αποτελούν οι καλά εκπαιδευμένοι επαγγελματίες. Ιδιαίτερα στον τομέα της ανακαίνισης κτιρίων, ο κατασκευαστικός τομέας της ΕΕ χρειάζεται περισσότερους και καλύτερα καταρτισμένους εργαζόμενους. Εκτός από τη στρατηγική Renovation Wave, η καθιερωμένη πρωτοβουλία BUILD UP Skills καθώς και το Ευρωπαϊκό Έτος Δεξιοτήτων 2023 στοχεύουν στην αντιμετώπιση αυτού του ζητήματος [10].

Οι βασικοί τομείς εστίασης είναι η αντιμετώπιση της ενεργειακής φτώχειας και η αναβάθμιση των κτιρίων με τις χειρότερες επιδόσεις, η ανακαίνιση δημοσίων κτιρίων και η αποκλιμάκωση της θέρμανσης και της ψύξης. Απώτερο σκοπό αποτελεί η μείωση των εκπομπών και η βελτίωση του συνολικού βιοτικού επιπέδου των Ευρωπαίων [11].

2.3 Εκπομπές CO₂ από Κτίρια

Τα κτίρια θεωρείται ότι είναι ένας από τους πιο καταναλωτικούς και σπάταλους τομείς της οικονομίας. Σύμφωνα με το βιβλίο 'Making the European Green Deal Work' [12], τα κτίρια

αντιπροσωπεύουν περίπου το 36% της παγκόσμιας κατανάλωσης ενέργειας. Εάν συμπεριληφθεί η ενέργεια που χρησιμοποιείται για την κατασκευή και την κατεδάφιση, το ενεργειακό μερίδιο ανέρχεται στο 50% της συνολικής κατανάλωσης ενέργειας. Παρά τα σημαντικά μέτρα εξοικονόμησης ενέργειας στις ανεπτυγμένες χώρες, η ζήτηση ενέργειας των κτιρίων αυξήθηκε περισσότερο από 20% και πάνω μεταξύ 2000 και 2017, λόγω της ταχείας αύξησης της επιφάνειας των κατοικιών, της σχετικά μικρής μείωσης της ενεργειακής έντασης και της αυξανόμενης ενεργειακής ζήτησης των ενεργειακών υπηρεσιών [13].

Η ενεργειακή τους κατανάλωση, ειδικά για θέρμανση, αερισμό, κλιματισμό (HVAC) και φωτισμό, συνιστά σημαντική πηγή εκπομπών διοξειδίου του άνθρακα (CO₂), προκαλώντας σημαντική ρύπανση στο περιβάλλον [14]. Υπολογίστηκε ότι οι άμεσες εκπομπές CO₂ που προκύπτουν από την καύση ορυκτών καυσίμων για την κάλυψη των θερμικών αναγκών των κτιρίων, τη θέρμανση χώρων και νερού και το μαγείρεμα ήταν κοντά στους 3 Gt CO₂ ενώ οι έμμεσες εκπομπές από την κατανάλωση ηλεκτρικής ενέργειας στα κτίρια, αύξανε το σύνολο σε 9,8 Gt CO₂ [15]. Ο οικιακός τομέας είναι η κύρια πηγή εκπομπών αερίων του θερμοκηπίου, αντιπροσωπεύοντας περίπου το 60% των συνολικών εκπομπών του τομέα, παρόλο που τα κτίρια κατοικιών καλύπτουν σχεδόν το 80% της συνολικής έκτασης των κτιρίων. Περίπου το 30% των συνολικών εκπομπών από κτίρια κατοικιών οφείλεται στην άμεση χρήση βιομάζας για εσωτερική καύση στις αναπτυσσόμενες χώρες. Το μερίδιο των εμπορικών κτιρίων στις έμμεσες εκπομπές CO₂ προσεγγίζει το 43%, κυρίως λόγω της υψηλής κατανάλωσης ηλεκτρικής ενέργειας σε αυτόν τον τομέα [16].

Μια στατιστική μελέτη στις Ηνωμένες Πολιτείες, μαρτυρά ότι τα κτίρια ευθύνονται για την εκπομπή 2236 εκατομμυρίων μετρικών τόνων CO₂, το οποίο αντιστοιχεί στο 39% των συνολικών εκπομπών CO₂. Αυτό δείχνει ότι τα κτίρια συνεισφέρουν στη ρύπανση περισσότερο από ό,τι οι μεταφορές ή ο βιομηχανικός τομέας [14]. Επίσης έχει αναφερθεί ότι η κατανάλωση στα κτίρια αντιπροσωπεύει περισσότερο από το ένα τρίτο της συνολικής χρήσης ενέργειας και των εκπομπών CO₂ στις αναπτυγμένες και αναπτυσσόμενες χώρες [17].

Επίσης, η συντήρηση των κτιρίων απαιτεί τη χρήση πρώτων υλών που βασίζονται κυρίως στον άνθρακα. Αυτό έχει σημαντικές επιπτώσεις στο περιβάλλον και τη βιωσιμότητα, καθώς η εξόρυξη και η χρήση του άνθρακα συνδέεται με την απελευθέρωση αερίων του θερμοκηπίου και την υπερθέρμανση του πλανήτη [13].

Παρά τις σημαντικές βελτιώσεις στην ενεργειακή ένταση ανά μονάδα επιφάνειας, οι συνολικές εκπομπές αερίων του θερμοκηπίου από τον κατασκευαστικό τομέα έχουν αυξηθεί κατά 25% από τις αρχές του 21ου αιώνα. Αυτό οφείλεται στην αντίστοιχη αύξηση της επιφάνειας κατά 65% και στη συνακόλουθη αύξηση της ζήτησης ηλεκτρικής ενέργειας των κτιρίων. Στην πραγματικότητα, από το 2000 η χρήση ηλεκτρικής ενέργειας στον τομέα των κτιρίων αυξήθηκε σχεδόν πέντε φορές ταχύτερα από τις αντίστοιχες βελτιώσεις στην απαλλαγή από τις εκπομπές στον τομέα της ηλεκτροπαραγωγής [13].

2.4 Χρήση Νέων Τεχνολογιών και Εκμετάλλευση Δεδομένων

Η ενεργειακά αποδοτική ανακαίνιση έχει γίνει ευρέως αποδεκτή ως η καλύτερη λύση για την αναβάθμιση των κτιρίων καθώς μεγιστοποιεί τα οφέλη στον ενεργειακό, περιβαλλοντικό και οικονομικό τομέα [18]. Γι' αυτό ένα βασικό ζήτημα που απασχολεί τους μηχανικούς και τους ερευνητές που ασχολούνται με τον τομέα αυτό αποτελεί η εύρεση βέλτιστων στρατηγικών που θα χρησιμοποιηθούν σε αυτήν.

Οι πρόσφατες εξελίξεις στη μηχανική μάθηση και η επιτυχία που εμφανίζουν στην πρόβλεψη και την ταξινόμηση δεδομένων μπορούν να βοηθήσουν στην ταξινόμηση δεδομένων από κτιριακά δεδομένα. Οι αλγόριθμοι και η διαθεσιμότητα και η ποιότητα των δεδομένων έχουν βελτιωθεί σημαντικά. Η αυξανόμενη υιοθέτηση τεχνολογιών και επικοινωνιών, όπως το Internet of Things και η τεχνητή νοημοσύνη, έχουν καταστήσει δυνατή τη λήψη μεγάλου όγκου δεδομένων που μπορούν να οδηγήσουν σε νέες λύσεις και αναλύσεις. Πιο συγκεκριμένα, στον τομέα της χρηματοδότησης της ΕΕ είναι πλέον δυνατή η συλλογή δεδομένων από έξυπνους μετρητές μετά την υλοποίηση ανακαινίσεων και η λήψη πληροφοριών σχετικά με την απόδοση τους. Κατά συνέπεια, οι μελλοντικές μελέτες θα πρέπει να επικεντρωθούν στην εφαρμογή μεθόδων και τεχνικών τεχνητής νοημοσύνης, καθώς και στη βέλτιστη αξιοποίηση των διαθέσιμων δεδομένων προς την κατεύθυνση της παροχής εργαλείων που θα βοηθήσουν μελλοντικά στην λήψη αποφάσεων που σχετίζονται με την ανακαίνιση κτιρίων[19].

Ειδικότερα, χρησιμοποιώντας δεδομένα που έχουν προκύψει από προηγούμενες ενεργειακές ανακαινίσεις κτιρίων μπορούν να εκπαιδευτούν διάφορες μέθοδοι ταξινόμησης για την ομαδοποίηση των δεδομένων αυτών, που αναμένεται να παράγουν αξιόπιστες προβλέψεις. Με αυτήν τη μέθοδο, θα είναι δυνατόν να εντοπιστούν οι βέλτιστες στρατηγικές ανακαίνισης που θα συμβάλουν εποικοδομητικά στη βελτίωση και ανακαίνιση άλλων κτιρίων [2].

Υπάρχουν διάφορες προσπάθειες και προσεγγίσεις που έχουν αναπτυχθεί για την ταξινόμηση συνόλων δεδομένων που αφορούν πληροφορίες σχετικά με μετρήσεις και χαρακτηριστικά από κτίρια. Στο επόμενο κεφάλαιο, θα διερευνήσουμε αυτές τις προσεγγίσεις και θα εξετάσουμε πώς μπορούν να συμβάλουν στη δημιουργία ενός πιο βιώσιμου και καθαρού αστικού περιβάλλοντος.

ΚΕΦΑΛΑΙΟ 3

Σύγχρονες Προσεγγίσεις στην Ομαδοποίηση Κτιριακών Δεδομένων

3 Σύγχρονες Προσεγγίσεις στην Ομαδοποίηση Κτιριακών Δεδομένων

Στο κεφάλαιο αυτό γίνεται μια ανασκόπηση σε μελέτες που έχουν πραγματοποιηθεί στην επεξεργασία και ομαδοποίηση δεδομένων που αφορούν ενεργειακές ανακαίνισης κτιρίων. Παρουσιάζονται τα δεδομένα που χρησιμοποιήθηκαν, η επεξεργασία που έγινε σε αυτά και οι αλγόριθμοι ομαδοποίησης που δοκιμάστηκαν. Επίσης, αναφέρονται τα αποτελέσματα στα οποία κατέληξε κάθε έρευνα.

3.1 Μοντέλα Μηχανικής Μάθησης

Τα μοντέλα μηχανικής μάθησης αποτελούν ένα ισχυρό εργαλείο που μπορεί να χρησιμοποιηθεί στη λήψη αποφάσεων που αφορά ανακαίνισης κτιρίων. Αξιοποιώντας δεδομένα από προηγούμενες ανακαίνισης, μπορούν να προβλέψουν τις βέλτιστες πρακτικές ανακαίνισης και να προτείνουν προσαρμοσμένες λύσεις, δρώντας συμβουλευτικά στους ερευνητές και ειδικούς που ασχολούνται με τον κτιριακό τομέα. Αυτή η προσέγγιση μπορεί να βοηθήσει στη μείωση του κόστους, του χρόνου και των απαιτήσεων εργασίας, ενώ ταυτόχρονα είναι ικανή να εξασφαλίσει τη μέγιστη ικανοποίηση των ανακαινιστικών αναγκών των κατοίκων.

Η πιο συνηθισμένη διαδικασία που ακολουθείται για την ομαδοποίηση δεδομένων είναι η ακόλουθη. Για να μπορέσουν να χωριστούν τα κτίρια σε συγκεκριμένες ομάδες, πρέπει πρώτα να καθοριστεί το βέλτιστο σύνολο μεταβλητών εισόδου που λαμβάνουν υπόψη οι αλγόριθμοι ομαδοποίησης. Στη συνέχεια, ακολουθεί η προ επεξεργασία των μεταβλητών εισόδου που συμβάλλει στη βελτίωση της απόδοσης των αλγορίθμων μηχανικής μάθησης. Στη συνέχεια, οι διάφορες ομάδες προσδιορίζονται από τον αλγόριθμο μηχανικής μάθησης που επιλέχτηκε. Υπάρχουν πολλοί αλγόριθμοι μηχανικής μάθησης για την ομαδοποίηση δεδομένων. Στη συνέχεια παρουσιάζονται κάποιοι σημαντικοί που έχουν χρησιμοποιηθεί για την ταξινόμηση κτιρίων. Αξίζει να σημειωθεί ότι παρ' όλο που οι αλγόριθμοι ταξινόμησης έχουν χρησιμοποιηθεί ευρέως σε διάφορα επιστημονικά πεδία, συμπεριλαμβανομένων των ιατρικών μελετών και της έρευνας αγοράς, οι δυνατότητές της για την ταξινόμηση του κτιριακού αποθέματος δεν έχουν διερευνηθεί επαρκώς [11].

3.1.1 Hierarchical Clustering

Στο Hierarchical Clustering εξετάζονται δύο τύποι ιεραρχικών αλγορίθμων: ο agglomerative και ο divisive. Στην προσέγγιση του agglomerative clustering, κάθε σημείο δεδομένων σχηματίζει μια ομάδα που αποτελείται από μόνο ένα στοιχείο. Οι πλησιέστερες ομάδες ενώνονται διαδοχικά έως ότου απομείνει μόνο μία ομάδα. Τα βήματα ανατρέχουν προς τα πίσω για να καταλήξουν στον επιθυμητό αριθμό ομάδων. Αυτή η μέθοδος απαιτεί τον μέγιστο αποδεκτό αριθμό ομάδων ως είσοδο για να παρέχει την τελική κατανομή του συνόλου δεδομένων. Στο divisive clustering, αρχικά οι παρατηρήσεις βρίσκονται όλες στην ίδια ομάδα. Σε κάθε βήμα, οι πιο διαφορετικές παρατηρήσεις διαιρούνται και κατανέμονται σε διαφορετικές ομάδες, έως ότου, τελικά, κάθε παρατήρηση βρεθεί σε διαφορετική ομάδα. Οι ομάδες συνδέονται μεταξύ τους και προσδιορίζεται μια κοινή τιμή ανομοιότητας με βάση το σχήμα σύνδεσης που υιοθετείται [20].

3.1.2 KMeans

Ο αλγόριθμος KMeans ανήκει στους αλγορίθμους μη επιβλεπόμενης μάθησης. Είναι πιο διαδεδομένος λόγω της απλότητας και της υπολογιστικής απόδοσης του. Ο αλγόριθμος ομαδοποιεί τα δεδομένα με βάση ένα κεντρικό σημείο, βελτιώνοντας τις ομάδες μέσω επαναλήψεων έως ότου συγκλίνει σε ένα αποτέλεσμα ομαδοποίησης. Συγκεκριμένα, αναθέτει N παρατηρήσεις δεδομένων σε ένα χώρο διάστασης i σε k ξεχωριστές ομάδες, με k τον προκαθορισμένο αριθμό ομάδων. Η ανάθεση των σημείων σε μια ορισμένη ομάδα k βασίζεται στον πλησιέστερο μέσο όρο χρησιμοποιώντας την ευκλείδεια απόσταση.

Η μέθοδος KMeans σε πρώτο στάδιο αρχικοποιεί τυχαία k ομάδες με k κεντρικά σημεία και κάθε σημείο ανατίθεται στην ομάδα με το πλησιέστερο κεντροειδές στο σημείο, ενώ το κεντροειδές υπολογίζεται εκ νέου σε κάθε ομάδα και ο βρόχος επαναλαμβάνεται έως ότου οι αναθέσεις δεν αλλάζουν πλέον ομάδες μεταξύ δύο διαδοχικών επαναλήψεων. Όπως έχει ήδη αναφερθεί, το κύριο πλεονέκτημα του k -means είναι η απλότητα και η υπολογιστική του αποδοτικότητα. Ωστόσο, η μέθοδος απαιτεί να καθοριστεί εκ των προτέρων ο αριθμός των ομάδων και τα αποτελέσματα της ομαδοποίησης μπορεί να εξαρτώνται σε μεγάλο βαθμό από τον επιλεγμένο αριθμό ομάδων [21].

3.1.3 Model-based Clustering

Εκτιμάται ότι το σύνολο δεδομένων μπορεί να παραχθεί από μια μίξη κατανομών πιθανότητας. Οι συνιστώσες αυτής της κατανομής (οι βασικές συναρτήσεις πυκνότητας) υποτίθεται ότι ανήκουν σε μια οικογένεια παραμετρικών συναρτήσεων πυκνότητας και συνδέονται με "βάρη" ή πιθανότητες. Στην περίπτωση του υλοποιημένου αλγορίθμου, οι συνιστώσες του μοντέλου υποτίθεται ότι είναι πολυμεταβλητές κατανομές Gauss. Κάθε υποσύνολο του συνόλου δεδομένων που αντιπροσωπεύεται από μία από αυτές τις συνιστώσες σχηματίζει μία ομάδα. Ως εκ τούτου, το πρόβλημα της κατανομής του χώρου δεδομένων αναδιατυπώνεται ως πρόβλημα επιλογής στατιστικού μοντέλου. Ο στόχος είναι η βελτιστοποίηση του αριθμού των συνιστωσών του μοντέλου (από ένα καθορισμένο εύρος), των σχετικών βαρών, καθώς και των τιμών των παραμέτρων κάθε συνιστώσας. Ο αλγόριθμος Model-based clustering επιλύει αυτό το πρόβλημα μέσω της προσαρμογής του μοντέλου στο σύνολο δεδομένων. Αφού οριστούν το στατιστικό μοντέλο και οι συνιστώσες του, τα σημεία δεδομένων που αντιπροσωπεύονται από κάθε συνιστώσα του μοντέλου ομαδοποιούνται ως ομάδα. Αυτή η μέθοδος, εξ ορισμού, αποδίδει μόνο τον βέλτιστο αριθμό κατανομών. Ως εκ τούτου, δεν απαιτείται ο καθορισμός του αριθμού των ομάδων πριν από την ανάλυση [22].

3.1.4 PAM

Ο αλγόριθμος PAM βασίζεται στην αναζήτηση k αντιπροσωπευτικών αντικειμένων (medoids) μεταξύ όλων των στοιχείων μιας δεδομένης ομάδας. Τα αντιπροσωπευτικά αντικείμενα είναι εκείνα με την ελάχιστη μέση ανομοιότητα. Τα στοιχεία αυτά ονομάζονται medoids. Ο αλγόριθμος βασίζεται στην ελαχιστοποίηση της αντικειμενικής συνάρτησης:

$$f = \min \left\{ \sum_{i=1}^n \sum_{j=1}^n d(i, j) \right\}$$

Ο αλγόριθμος PAM μπορεί να περιγραφεί σύμφωνα με τα δύο βήματά του: μια φάση δημιουργίας, όπου επιλέγονται αντικείμενα ως αντιπροσωπευτικά ενός συνόλου (αρχικά medoids), και μια φάση ανταλλαγής, όπου η ομαδοποίηση βελτιώνεται με την εναλλαγή των αντιπροσωπευτικών στοιχείων μιας ομάδας. Με άλλα λόγια, ο αλγόριθμος επισημαίνει κάθε σημείο με βάση το πλησιέστερο κέντρο και στη συνέχεια αντικαθιστά κάθε κέντρο με το medoid των σημείων της ομάδας του[23].

3.2 Σχετικές Μελέτες

Στις επόμενες παραγράφους παρουσιάζονται κάποιες σημαντικές προσπάθειες που έχουν γίνει για την ταξινόμηση κτιρίων. Βασικός σκοπός είναι τα κτίρια να χωριστούν σε ομάδες με παρόμοια ενεργειακά χαρακτηριστικά. Παρακάτω παρατίθεται ένας πίνακας σύγκρισης των προσπαθειών που έχουν πραγματοποιηθεί ενώ στα επόμενα υποκεφάλαια παρουσιάζονται πιο αναλυτικά.

3.2.1 Ομαδοποίηση Κτιρίων με δεδομένα από τη Βιέννη

Σε αυτήν την ερευνητική προσπάθεια διερευνώνται οι απαιτήσεις ενός αστικού ενεργειακού περιβάλλοντος, με στόχο την υποστήριξη της λήψης στρατηγικών αποφάσεων όσον αφορά τις φυσικές και τεχνολογικές παρεμβάσεις, καθώς και τις αλλαγές συμπεριφοράς του περιβάλλοντος. Τα δεδομένα που χρησιμοποιήθηκαν προέρχονταν από μετρήσεις που πραγματοποιήθηκαν σε μια περιοχή της Βιέννης σε 750 κτίρια με διαφορετικές λειτουργίες. Αποτελούν επίσημα δεδομένα και δεδομένα Γεωγραφικών Συστημάτων Πληροφοριών (GIS), καθώς και στατιστικές πληροφορίες και πρότυπα αξιολόγησης για την απόδοση των κτιρίων. Στη συνέχεια, συγκρίθηκαν τρεις τεχνικές ομαδοποίησης με διαφορετικά σύνολα μεταβλητών εισόδου. Ανάλογα με την εφαρμοζόμενη τεχνική ομαδοποίησης και το συγκεκριμένο σύνολο μεταβλητών εισόδου για την ομαδοποίηση, κατέληγαν σε 6 έως 21 ομάδες.

Τα χαρακτηριστικά εισόδου αφορούν μετρήσεις που σχετίζονται με τη γεωμετρία του κτιρίου όπως ο όγκος, η επιφάνεια που καταλαμβάνει το κτίριο, το ύψος του κτιρίου, η απόσταση από το έδαφος, η κατανομή των ορόφων κτλ. Άλλα σημαντικά χαρακτηριστικά είναι τα οφέλη λόγω ηλιακής ακτινοβολίας, η θερμική άνεση όπως θερμοκρασία, διάφορες παράμετροι λειτουργίας όπως η ενεργειακή απόδοση, ο ρυθμός εξαερισμού και χρονοδιαγράμματα λειτουργίας κτιρίου. Τα δεδομένα ελέγχθηκαν για ασυνέπειες ή ελλείψεις. Επίσης, κάποιες πληροφορίες συνδυάστηκαν για να δώσουν εξίσου νέα σημαντικά χαρακτηριστικά. Ο πίνακας δεδομένων που προκύπτει κάθε φορά αποθηκεύεται ως αρχείο csv και υποβάλλεται στη διαδικασία ταξινόμησης. Περιέχει 7 έως 12 χαρακτηριστικά.

Στη συνέχεια διερευνήθηκαν τρεις αλγόριθμοι ως προς την αποτελεσματικότητά τους. Οι δύο πρώτες μέθοδοι συγκатаλέγονται ανάμεσα στους πιο συχνά χρησιμοποιούμενους αλγορίθμους και η τρίτη επιλέχτηκε να δοκιμαστεί λόγω της μοναδικής προσέγγισής του για τον προσδιορισμό του βέλτιστου αριθμού ομάδων.

Ο πρώτος αλγόριθμος που διερευνήθηκε ήταν ο Hierarchical agglomerative clustering. Στο τρέχον έργο η απόσταση των σημείων δεδομένων μετράται μέσω της τετραγωνικής ευκλείδειας απόστασης. Το αποδεκτό κριτήριο σύνδεσης, το οποίο καθορίζει το ζεύγος ομάδων που πρέπει να συγχωνεύεται σε κάθε βήμα, είναι η μέθοδος Ward. Σε κάθε βήμα, ο αλγόριθμος αυτός εντοπίζει και συνδέει το ζεύγος ομάδων των οποίων η σύνδεση έχει ως αποτέλεσμα την ελάχιστη αύξηση της συνολικής διακύμανσης εντός των ομάδων μετά τη συγχώνευση. Στη συνέχεια, δοκιμάστηκε ο αλγόριθμος KMeans. Η απόσταση για την επιλογή του κοντινότερου γείτονα γίνεται με βάση την ευκλείδεια απόσταση και δίνεται ως παράμετρος στον αλγόριθμο. Τέλος, εξετάστηκε ο αλγόριθμος Model-based clustering.

Ο αριθμός των ομάδων που δόθηκε ως είσοδος στους δύο πρώτους αλγορίθμους είναι από 6 έως 30. Στη συνέχεια πραγματοποιήθηκαν πειράματα με τους συνδυασμούς των δεδομένων εισόδου και τους τρεις αλγορίθμους. Για την αξιολόγηση των αποτελεσμάτων χρησιμοποιήθηκε το σχετικό σφάλμα πρόβλεψης της ετήσιας ζήτησης θέρμανσης για ολόκληρη τη γειτονιά, ως μέτρο του πόσο καλά το δείγμα αντιπροσωπεύει τη συνολική ζήτηση θέρμανσης της αστικής περιοχής, το μέσο σχετικό σφάλμα της πρόβλεψης της ετήσιας ζήτησης θέρμανσης σε επίπεδο κτιρίου ως δείκτη των προβλεπτικών ικανοτήτων του δείγματος όσον αφορά τα μεμονωμένα κτίρια και το κλάσμα του συνολικού όγκου που σχετίζεται με κτίρια με σχετικό σφάλμα πρόβλεψης της ετήσιας ζήτησης θέρμανσης άνω του 20% για τον προσδιορισμό του συστήματος δειγματοληψίας που οδηγεί στο μικρότερο ποσό σοβαρών σφαλμάτων πρόβλεψης όσον αφορά τα επιμέρους κτίρια.

Το σχήμα ομαδοποίησης με τις καλύτερες επιδόσεις οδηγεί στον διαχωρισμό των κτιρίων σε 7 διακριτές ομάδες. Κάθε ομάδα αντιπροσωπεύεται μέσω ενός κτιρίου, με αποτέλεσμα το μέγεθος του δείγματος να είναι μικρότερο από το 1% ολόκληρου του τομέα. Παρά το μικρό μέγεθος του δείγματος, η μέθοδος αυτή οδηγεί σε σχετικά μικρά σφάλματα στην πρόβλεψη της απόδοσης, καθώς και σε χαμηλό μέσο σφάλμα στις προβλέψεις σε επίπεδο κτιρίου.

Σε αυτήν την έρευνα τα αποτελέσματα όσον αφορά την ομαδοποίηση με βάση τη ζήτηση σε θέρμανση είναι ικανοποιητικά. Ωστόσο, η αποτελεσματικότητα της μεθόδου όσον αφορά άλλες ενεργειακές πτυχές, όπως η χρονική κατανομή, απαιτεί περαιτέρω διερεύνηση με πληροφορίες υψηλής ανάλυσης. Η μέθοδος που αναπτύχθηκε βελτιώνει και αυτοματοποιεί τη διαδικασία της αναπαράστασης των κτιρίων μέσω της εφαρμογής μεθόδων μηχανικής μάθησης, διευκολύνοντας τη χρήση εργαλείων προσομοίωσης της απόδοσης των κτιρίων για ενεργειακές έρευνες σε αστικά κτίρια [24].

3.2.2 Προσδιορισμός αντιπροσωπευτικών κτιρίων και ομάδων με δεδομένα από τη Γένοβα

Σε αυτήν την προσέγγιση, χρησιμοποιήθηκε ένα σύνολο δεδομένων που περιείχε μετρήσεις από τη Γένοβα. Το σύνολο αυτό περιείχε γεωαναφερμένα δεδομένα του κτιριακού αποθέματος και των αστικών υποδομών και εδαφικές πληροφορίες.

Αρχικά, πραγματοποιήθηκε προ επεξεργασία των δεδομένων. Η προεπεξεργασία των δεδομένων είναι ένα σημαντικό βήμα που επηρεάζει την ποιότητα των αποτελεσμάτων της ομαδοποίησης. Ειδικότερα, αναλύονται τα δεδομένα και εξαλείφονται οι κενές τιμές. Με αυτόν τον τρόπο το μέγεθος του πίνακα δεδομένων μειώνεται. Ύστερα, τα δεδομένα φιλτράρονται χρησιμοποιώντας πληροφορίες όπως το είδος των κτιρίων (π.χ. κατοικίες, εμπορικά κτίρια και κτίρια μικτής χρήσης), την περίοδο κατασκευής (π.χ. 1919-1945 ή 1946-

1960) και την τελική χρήση των κτιρίων (π.χ. πολυκατοικίες ή μονοκατοικίες). Η διαδικασία αυτή δημιουργεί μικρότερα σύνολα δεδομένων, αυξάνοντας έτσι τον αριθμό των υποσυνόλων δεδομένων. Οι κατηγορικές μεταβλητές που επιλέχθηκαν επιτρέπουν τη δημιουργία ομάδων κτιρίων που χαρακτηρίζονται από κοινά χαρακτηριστικά. Τα κτίρια στο ίδιο σύνολο δεδομένων είναι πιθανό να χαρακτηρίζονται από παρόμοιες τεχνικές κατασκευής, ποιότητα κτιριακού κελύφους και γεωμετρία.

Έπειτα, πραγματοποιείται μια διαδικασία κανονικοποίησης των δεδομένων. Οι λοξές κατανομές και η παρουσία ακραίων τιμών επηρεάζουν τον αριθμό των ομάδων που δημιουργούνται από τους αλγόριθμους. Με την κανονικοποίηση των δεδομένων εγγυάται ότι δεν δίνεται ιδιαίτερη βαρύτητα σε κάποιο συγκεκριμένο χαρακτηριστικό εις βάρος κάποιου άλλου. Τα διάφορα χαρακτηριστικά του κτιρίου παρουσιάζουν τιμές που διαφέρουν κατά αρκετές τάξεις μεγέθους (106 Wh έναντι 102m²). Για το λόγο αυτό, εκτελούνται πέντε μέθοδοι στο σύνολο δεδομένων: κλιμάκωση, κεντράρισμα, z-score, PCA και Box-Cox.

Στη συνέχεια, εκτελείται μια διαδικασία προ-ομαδοποίησης σε κάθε σύνολο δεδομένων εξόδου της ταξινόμησης και της προσέγγισης κανονικοποίησης. Τέσσερις αλγόριθμοι ομαδοποίησης χρησιμοποιούνται για τη δημιουργία ομάδων: k-means, partitioning around medoids (PAM), hierarchical agglomerative και hierarchical divisive.

Τα αποτελέσματα της ομαδοποίησης μπορούν να αξιολογηθούν με δείκτες επικύρωσης. Στην παρούσα μελέτη αξιολογήθηκαν εννέα δείκτες επικύρωσης για κάθε λύση ομαδοποίησης και ένας γενικός δείκτης που συνδυάζει τους δείκτες αυτούς.

Μετά την επιλογή της καλύτερης προσέγγισης κανονικοποίησης για μια κατηγορία κτιρίων, τα αποτελέσματα των τεσσάρων αλγορίθμων ομαδοποίησης αξιολογούνται για να κατανοηθεί ποιος είναι ο καλύτερος αριθμός ομάδων για κάθε σύνολο δεδομένων. Αυτό επιτυγχάνεται με την αξιολόγηση του συνολικού δείκτη ποιότητας που ορίζεται για κάθε σύνολο δεδομένων. Έτσι, ο αριθμός των ομάδων κυμαίνεται μεταξύ 2 και 15. Ο συνολικός δείκτης ποιότητας αξιολογείται για κάθε αλγόριθμο ομαδοποίησης και για κάθε αριθμό ομάδων. Έτσι, ο δείκτης για κάθε αλγόριθμο απεικονίζεται για σύγκριση. Αυτό επιτρέπει την επιλογή του αριθμού ομάδας που μεγιστοποιεί το δείκτη σε διαφορετικά τμήματα ομαδοποίησης.

Στην ομαδοποίηση είναι επιθυμητό τα σημεία των δεδομένων που ανήκουν στην ίδια ομάδα να βρίσκονται κοντά το ένα στο άλλο αλλά και οι ομάδες να έχουν απόσταση η μία από την άλλη. Γι' αυτό άλλο ένα σημαντικό βήμα είναι η απεικόνιση των δεδομένων. Για την απεικόνιση, χρησιμοποιούνται ιστογράμματα και box-plots. Όταν εντοπιστεί ο καλύτερος αριθμός ομάδων, προσδιορίζεται το κεντροειδές (αντιπροσωπευτικό σημείο) και στη συνέχεια το μεσοειδές (αντιπροσωπευτικό στοιχείο) κάθε ομάδας. Επιπλέον, τα δεδομένα οπτικοποιούνται με τη χρήση μιας εφαρμογής γεωγραφικών συστημάτων πληροφοριών (GIS) για τον εντοπισμό των στοιχείων και της θέσης τους στο αστικό περιβάλλον. Η σύνδεση των δεδομένων με το GIS διευκολύνει την οπτικοποίηση των κτιρίων σε μια συγκεκριμένη ομάδα και τον προσδιορισμό των μεσοειδών.

Μετά την εφαρμογή της συγκεκριμένης μεθοδολογίας εντοπίστηκαν συνολικά 67 αντιπροσωπευτικά κτίρια και συναφή συγκροτήματα. Τα αποτελέσματα προέκυψαν από την ανάλυση 9 χαρακτηριστικών κτιρίων, 3 διαφορετικών τυπολογιών και 10 περιόδων κατασκευής. Οι ομάδες και τα κτίρια είναι αντιπροσωπευτικά ενός συνόλου περίπου 13600 κτιρίων. Οι καλύτερες κατατμήσεις επιτεύχθηκαν με την εξέταση ενός μικρού αριθμού ομάδων που κυμαίνεται από δύο έως πέντε. Τα καλύτερα αποτελέσματα σε διαφορετικές τυπολογίες κτιρίων επιτεύχθηκαν χρησιμοποιώντας κλιμάκωση, PCA και z-score [25].

3.2.3 Ομαδοποίηση Κτιρίων για Προσομοιώσεις

Σε αυτήν την μελέτη χρησιμοποιήθηκαν μετρήσεις από κτίρια που βρίσκονται στην περιοχή Boxbergheide. Τα χαρακτηριστικά που περιλαμβάνει είναι το εμβαδόν του κτιρίου, η θερμαινόμενη επιφάνεια δαπέδου, η συνολική επιφάνεια απωλειών, ο αριθμός των ορόφων, ο θερμαινόμενος όγκος, το συνολικό εμβαδόν του κτιρίου, η επιφάνεια απωλειών προς το δάπεδο (δηλαδή την επιφάνεια απωλειών θερμότητας του κτιρίου διαιρεμένη με τη θερμαινόμενη επιφάνεια δαπέδου), η τυπολογία (ημι-μονοκατοικίες ή μονοκατοικίες), η επιφάνεια των παραθύρων, ο λόγος παραθύρων προς τους τοίχους και η περίοδος κατασκευής. Η περίοδος κατασκευής χρησιμοποιείται στην παρούσα εργασία για την κατανομή όλων των θερμικών χαρακτηριστικών του κελύφους του κτιρίου που απαιτούνται για την ποσοτικοποίηση της ετήσιας ζήτησης θερμότητας.

Για τον εντοπισμό του βέλτιστου συνόλου χαρακτηριστικών που χρησιμοποιούνται ως μεταβλητές ομαδοποίησης, εάν ο αριθμός αυτών είναι γνωστός εκ των προτέρων, ακολουθείται μια προσέγγιση που βασίζεται στην παλινδρόμηση. Η διαδικασία περιλαμβάνει την επιλογή και κατάταξη των παραμέτρων μέσω πολλαπλής γραμμικής παλινδρόμησης (MLR) και του κριτηρίου πληροφοριών Akaike (AIC). Για τη ζήτηση θερμότητας αιχμής και την ετήσια ζήτηση θερμότητας, προσαρμόζεται ένα MLR μοντέλο χρησιμοποιώντας τη μέθοδο των ελαχίστων τετραγώνων. Η διαδικασία επιλογής προς τα εμπρός προσθέτει στο μοντέλο κάθε φορά την πιο σημαντική μεταβλητή. Το AIC χρησιμοποιείται για την επιλογή του βέλτιστου αριθμού παραμέτρων που περιλαμβάνονται στο MLR, εκτιμώντας την απώλεια πληροφορίας που προκαλεί ένα συγκεκριμένο μοντέλο σε σύγκριση με άλλα μοντέλα.

Τα δεδομένα που επιλέγονται τελικά, κανονικοποιούνται. Μετά την κανονικοποίηση, το σύνολο δεδομένων εισάγεται ως είσοδος στους αλγορίθμους k-means και agglomerative hierarchical clustering, δοκιμάζοντας διάφορους συνδυασμούς μεταβλητών.

Μετά την αξιολόγηση των αποτελεσμάτων, προκύπτει ότι για την πρόβλεψη της ζήτησης θερμότητας αιχμής, το βέλτιστο σύνολο μεταβλητών αποτελείται από το αποτύπωμα του κτιρίου, τη συνολική έκταση των απωλειών και την περίοδο κατασκευής ενώ για την ετήσια ζήτηση θερμότητας για θέρμανση χώρων, το βέλτιστο σύνολο μεταβλητών αποτελείται από την επιφάνεια του θερμαινόμενου δαπέδου, τη συνολική επιφάνεια απωλειών και την περίοδο κατασκευής. Επίσης, όσον αφορά τους αλγορίθμους τόσο για τη ζήτηση θερμότητας αιχμής όσο και για την ετήσια ζήτηση θερμότητας για θέρμανση χώρων, η τεχνική ομαδοποίησης k-means σε συνδυασμό με τη μέθοδο κανονικοποίησης των δεδομένων σε νέα κλίμακα αποδίδει καλύτερα και κατέληξαν στην ομαδοποίηση με 10 ομάδες [21].

3.2.4 Ομαδοποίηση Fuzzy Cmeans για Σύνολα Δεδομένων με Θόρυβο, Outliers και Άνισο Αριθμό Ομάδων

Η παρούσα εργασία επικεντρώνεται κυρίως σε αλγορίθμους ασαφούς ομαδοποίησης για τον χειρισμό δεδομένων που περιλαμβάνουν θόρυβο και ακραίες τιμές. Αυτοί οι αλγόριθμοι

εφαρμόζονται σε διάφορα πρακτικά προβλήματα. Ωστόσο, εκτός από το θόρυβο και τις ακραίες τιμές, μερικές φορές υπάρχουν ομάδες με διαφορετικά μεγέθη στα δεδομένα. Οι περισσότεροι αλγόριθμοι ασαφούς ομαδοποίησης αντιμετωπίζουν με το ίδιο τρόπο όλες τις ομάδες και δεν μπορούν να μετρήσουν το μέγεθος των ομάδων, γεγονός που προκαλεί μετατόπιση των κέντρων των μικρών ομάδων προς τις μεγάλες ομάδες.

Με την έννοια της αξιοποίησης των πλεονεκτημάτων και των δύο αλγορίθμων και της εξάλειψης των μειονεκτημάτων τους, οι αλγόριθμοι FCM και PCM συνδυάζονται για τη δημιουργία του αλγορίθμου Possibilistic Fuzzy C-Means (PFCM), ο οποίος αναμένεται να χειρίζεται το θόρυβο και τις ακραίες τιμές με τους πιθανοτικούς όρους του και να αποφεύγει τις συμπιπτουσες ομάδες και την ευαισθησία στην αρχικοποίηση με τους ασαφείς όρους του. Ο αλγόριθμος RFCM στοχεύει σε δύο στόχους: να αποτρέψει την προσέλκυση των κέντρων των μικρών ομάδων από τις μεγάλες ομάδες και να εξαλείψει τις επιπτώσεις του θορύβου και των ακραίων τιμών στα κέντρα των συστάδων. Για το σκοπό αυτό, χρησιμοποιούνται δύο αντικειμενικές συναρτήσεις. Η μία για τον προσδιορισμό των αρχικών κέντρων των συστάδων με μια προσέγγιση που δεν είναι ευαίσθητη στο μέγεθος και η άλλη για τον υπολογισμό των τελικών κέντρων των συστάδων με μια προσέγγιση που είναι ανθεκτική στο θόρυβο. Η ελαχιστοποίηση της πρώτης αντικειμενικής συνάρτησης παρέχει κέντρα συστάδων με τα οποία ξεκινά το δεύτερο μέρος του αλγορίθμου.

Ο θόρυβος και οι ακραίες τιμές, οι άνισες ομάδες και η ανομοιόμορφη κατανομή μάζας μειώνουν την ακρίβεια των αλγορίθμων ομαδοποίησης και οδηγούν σε ανακριβή αποτελέσματα ομαδοποίησης. Στην παρούσα εργασία παρουσιάζεται η αναθεωρημένη έκδοση του αλγορίθμου FCM (RFCM) για την ομαδοποίηση δεδομένων με τέτοια χαρακτηριστικά. Ο αλγόριθμος RFCM εξαλείφει τις συνεισφορές του θορύβου και των ακραίων τιμών εισάγοντας μια προσαρμοστική εκθετική συνάρτηση στην αντικειμενική συνάρτηση του αλγορίθμου FCM και αποτρέπει τις μεγάλες ή βαρύτερες συστάδες από το να προσελκύουν τα κέντρα των μικρών συστάδων μέσω ενός μηχανισμού που δεν είναι ευαίσθητος στο μέγεθος. Η απόδοση του προτεινόμενου αλγορίθμου συγκρίνεται με εκείνη διαφόρων γνωστών αλγορίθμων, όπως Possibilistic Fuzzy C-Means (PFCM), Possibilistic C-Means (PCM), Robust Fuzzy C-Means (FCM-σ), Noise Clustering (NC), Kernel Fuzzy C Means (KFCM), Intuitionistic Fuzzy C-Means (IFCM), Robust Kernel Fuzzy C-Mean (KFCM-σ), Robust Intuitionistic Fuzzy C-Means (IFCM-σ), Kernel Intuitionistic Fuzzy C-Means (KIFCM), Robust Kernel Intuitionistic Fuzzy C-Means (KIFCM-σ), Credibilistic Fuzzy C-Means (CFCM), Size-insensitive integrity-based Fuzzy C-Means (siibFCM), Size-insensitive Fuzzy C-Means (csiFCM), Subtractive Clustering (SC), Density Based Spatial Clustering of Applications with Noise (DBSCAN), Gaussian Mixture Models (GMM), Spectral clustering algorithm και Outlier Removal Clustering (ORC).

Οι αλγόριθμοι εφαρμόζονται αρχικά σε θορυβώδη σύνολα δεδομένων με ίσα ή άνισα μεγέθη ομάδων και ανομοιόμορφη κατανομή μάζας, αλλά με γνωστά κέντρα ομάδων. Τα πειράματα δείχνουν υπεροχή του αλγορίθμου RFCM έναντι των άλλων. Οι αλγόριθμοι εφαρμόζονται επίσης σε πραγματικά σύνολα δεδομένων, αλλά δεδομένου ότι τα κέντρα συστάδων αυτών των δεδομένων είναι άγνωστα, χρησιμοποιούνται η ποιότητα της συσταδοποίησης, ο δείκτης Xie-Beni και ο δείκτης Davies-Bouldin για την αξιολόγηση της απόδοσης των αλγορίθμων. Η μελέτη δείχνει ότι ο αλγόριθμος RFCM παρέχει τα καλύτερα αποτελέσματα όσον αφορά αυτά τα μέτρα στις περισσότερες περιπτώσεις[26].

3.2.5 Ομαδοποίηση Αστικών Hotspots

Σε αυτήν την μελέτη, αναπτύσσεται ένας νέος αλγόριθμος ομαδοποίησης KMeans που βασίζεται σε έναν αλγόριθμο θορύβου για τη σύλληψη των αστικών hotspots. Η μελέτη αποσκοπεί στον οδικό σχεδιασμό, καθώς αποτελεί ένα αποτελεσματικό μέσο για την ανακούφιση της αστικής συμφόρησης, το οποίο είναι ένα κλασικό μη ντετερμινιστικό πρόβλημα πολυωνυμικού χρόνου (NP) και έχει γίνει ένα σημαντικό ερευνητικό σημείο τα τελευταία χρόνια. Ο αλγόριθμος ομαδοποίησης KMeans είναι ένας επαναληπτικός αλγόριθμος ανάλυσης ομαδοποίησης που θεωρείται ως αποτελεσματικό μέσο για την επίλυση προβλημάτων αστικού οδικού σχεδιασμού από τους μελετητές τις τελευταίες δεκαετίες- ωστόσο, είναι πολύ δύσκολο να προσδιοριστεί ο αριθμός των ομάδων και να αρχικοποιηθεί με ευαισθησία η κεντρική ομάδα. Ο αλγόριθμος θορύβου που αναλύεται στη συνέχεια, χρησιμοποιείται για την τυχαία ενίσχυση της απόδοσης των σημείων δεδομένων και των αποτελεσμάτων εξόδου της ομαδοποίησης με την προσθήκη κρίσης θορύβου, προκειμένου να ληφθεί αυτόματα ο αριθμός των ομάδων για τα δεδομένα που δίνονται και να αρχικοποιηθεί η κεντρική συστάδα.

Για την αξιολόγηση και την ανάλυση των αποτελεσμάτων χρησιμοποιούνται τέσσερις δείκτες αξιολόγησης χωρίς επίβλεψη, οι DB, PBM, SC και SSE. Επίσης, χρησιμοποιείται μια μη παραμετρική μέθοδος στατιστικής ανάλυσης Wilcoxon για την επαλήθευση των καταστάσεων κατανομής και των διαφορών μεταξύ των αποτελεσμάτων της ομαδοποίησης.

Προκειμένου να επαληθευτεί η αποτελεσματικότητα του προτεινόμενου αλγορίθμου ομαδοποίησης θορύβου K-means, χρησιμοποιούνται πέντε σύνολα δεδομένων GPS ταξί στο εσωτερικό και στο εξωτερικό. Τα δεδομένα GPS ταξί αναφέρονται κυρίως στις πληροφορίες θέσης, κατεύθυνσης και ταχύτητας του οχήματος που καταγράφονται τακτικά από το όχημα μέσω του ενσωματωμένου παγκόσμιου συστήματος εντοπισμού θέσης κατά τη διάρκεια της διαδρομής. Οι πόλεις της παρούσας μελέτης, δηλαδή οι Aracaju (Βραζιλία), San Francisco (ΗΠΑ), Ρώμη (Ιταλία), Chongqing (Κίνα) και Πεκίνο (Κίνα), είναι αντιπροσωπευτικές πόλεις για διαφορετικές ηπείρους και χώρες και είναι πόλεις μεγάλης κλίμακας. Ο δείκτης DB, ο δείκτης PBM, ο SC (συντελεστής σιλουέτας) [49] και το SSE (άθροισμα τετραγώνων σφάλματος) χρησιμοποιούνται άμεσα για την αξιολόγηση και την ανάλυση των αποτελεσμάτων της ομαδοποίησης. Αυτές οι μέθοδοι αξιολόγησης σχετίζονται άμεσα με τον αριθμό των συστάδων.

Τα αποτελέσματα των πειραμάτων δείχνουν ότι ο αλγόριθμος θορύβου μπορεί εύλογα να λάβει τον αριθμό των ομάδων και να αρχικοποιήσει την κεντρική συστάδα, και ο προτεινόμενος αλγόριθμος ομαδοποίησης θορύβου KMeans επιδεικνύει καλύτερη απόδοση ομαδοποίησης και λαμβάνει με ακρίβεια αποτελέσματα ομαδοποίησης, καθώς και αποτελεσματική σύλληψη των αστικών hotspots. Επίσης, ο προτεινόμενος αλγόριθμος ομαδοποίησης KMeans με βάση τον θόρυβο μπορεί να εφαρμοστεί και στους τομείς της ταξινόμησης εγγράφων, της ταξινόμησης πελατών, της ανάλυσης δεδομένων διαδρομών, της ανάλυσης εγκληματικών δικτύων, της λεπτομερούς ανάλυσης αρχείων κλήσεων κ.ο.κ.

Ωστόσο, υπάρχουν ορισμένες ελλείψεις. Αφενός, η ποσότητα των δεδομένων GPS είναι πολύ μικρή για να αντικατοπτρίζει αποτελεσματικά τις κατανομές και τις σχέσεις των αστικών hotspots. Από την άλλη πλευρά, είναι δύσκολο να αποφευχθούν αποτελεσματικά κάποια κτίρια στην πόλη, ακόμη και αν τα βέλτιστα αποτελέσματα των αστικών hotspots χρησιμοποιούνται στον σχεδιασμό αστικών δρόμων[27].

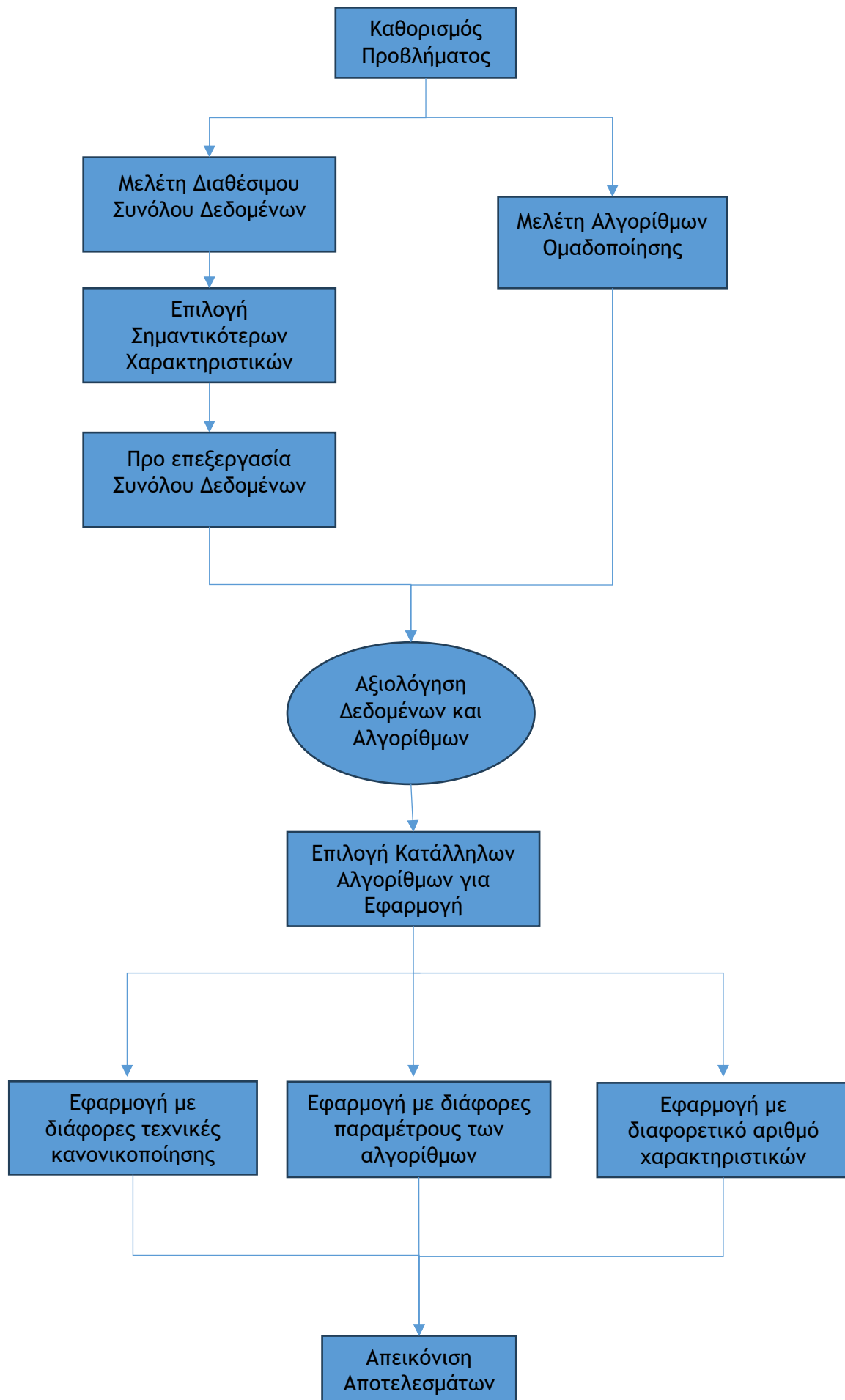
ΚΕΦΑΛΑΙΟ 4

Προτεινόμενη Μεθοδολογία

4 Προτεινόμενη Μεθοδολογία

Σε αυτό το κεφάλαιο, προσδιορίζεται το πρόβλημα που πραγματεύεται η παρούσα εργασία και παρουσιάζεται η προτεινόμενη μεθοδολογία που εφαρμόστηκε για την αντιμετώπιση αυτού του προβλήματος με τη χρήση μεθόδων επεξεργασίας δεδομένων και αλγορίθμων μηχανικής μάθησης. Στη συνέχεια, γίνεται σύγκριση με άλλες μεθόδους που έχουν ήδη εξεταστεί στο 3ο κεφάλαιο, αναδεικνύοντας τις διαφορές και τις ομοιότητές τους. Τέλος, αναλύεται ο κύριος στόχος της προτεινόμενης μεθοδολογίας.

Το παρακάτω σχήμα(Σχήμα 1), παρουσιάζει τη προτεινόμενη μεθοδολογία για την μείωση των εκπομπών ρύπων μέσω ανακαινίσεων κτιρίων με τη βοήθεια της τεχνικής νοημοσύνης



Σχήμα 1: Προτεινόμενο Μεθοδολογικό Πλαίσιο για την Μείωση των Εκπομπών Ρύπων Μέσω Ανακαίνιστων Κτιρίων με τη Βοήθεια της Μηχανικής Μάθησης

4.1 Καθορισμός του Προβλήματος

Τα κτίρια αποτελούν έναν σημαντικό παράγοντα ρύπανσης, καθώς η λειτουργία τους συχνά συνδέεται με τη χρήση ενέργειας και εκπομπές αερίων θερμοκηπίου [28]. Η ανάγκη ανακαίνισης των κτιρίων αναδύεται ως κρίσιμη για τη μείωση της περιβαλλοντικής επιβάρυνσης, με την εισαγωγή πιο βιώσιμων τεχνολογιών και υλικών. Ωστόσο, η αποτελεσματική ανακαίνιση απαιτεί συχνά δεδομένα σχετικά με την ιστορία και την κατάσταση του κτιρίου. Η ταξινόμηση παλαιών δεδομένων κτιρίων, όπως τα στοιχεία κατασκευής, οι τεχνολογικές προδιαγραφές και οι εκπομπές ρύπων, μπορεί να διευκολύνει την ανακαίνιση μέσω της καλύτερης πρόβλεψης και σχεδιασμού των αλλαγών που απαιτούνται για τη μείωση του περιβαλλοντικού αποτυπώματος των κτιρίων. Με τη συνεκτική διαχείριση δεδομένων κτιρίων από προηγούμενες ανακαίνισεις, μπορεί να επιτευχθεί μια πιο αποτελεσματική και βιώσιμη ανάπτυξη στον κτιριακό τομέα [29].

4.2 Προτεινόμενη Μεθοδολογία

Η προτεινόμενη μεθοδολογία για την αντιμετώπιση του προβλήματος περιλαμβάνει αρχικά την επιλογή συνόλων χαρακτηριστικών, η οποία πραγματοποιείται μετά από προσεκτική μελέτη και κατανόηση του προβλήματος. Ακολουθεί η προ επεξεργασία των δεδομένων, συμπεριλαμβανομένου του ελέγχου για μεταβλητές που λείπουν και της κανονικοποίησης τους για εξασφάλιση ενός σταθερού εύρους τιμών. Στη συνέχεια μελετώνται οι πιο συχνά χρησιμοποιούμενοι αλγόριθμοι για παρόμοια προβλήματα και επιλέγονται οι DBSCAN (Density-Based Spatial Clustering of Applications with Noise), KMeans, Agglomerative Clustering καθώς και ο HDBSCAN (Hierarchical Density-Based Spatial Clustering of Applications with Noise). Η εφαρμογή αυτών των μοντέλων επιτρέπει τον διαχωρισμό τους σε ομάδες και παρόμοια χαρακτηριστικά. Σε κάθε αλγόριθμο πραγματοποιούνται και αξιολογούνται πειράματα που περιλαμβάνουν τεχνικές κανονικοποίησης, αυξομείωση των χαρακτηριστικών εισόδου και δοκιμή διάφορων παραμέτρων των μοντέλων.

Για την αξιολόγηση της απόδοσης και της καταλληλότητας των μεθόδων ομαδοποίησης, αυτές συνδυάζονται και αξιολογούνται χρησιμοποιώντας οπτικοποίηση των αποτελεσμάτων. Ειδικότερα, η αξιολόγηση περιλαμβάνει την οπτική ανάλυση των ομάδων που δημιουργούνται από κάθε μέθοδο, τη σύγκριση της συνοχής και του διαχωρισμού των ομάδων, καθώς και την αξιολόγηση της εναρμόνισης με τις πραγματικές δομές και πρότυπα των δεδομένων. Με αυτόν τον τρόπο, η οπτική ανάλυση των αποτελεσμάτων παρέχει σημαντική ενδεικτική πληροφορία για την αποτελεσματικότητα και την καταλληλότητα κάθε μεθόδου ομαδοποίησης στο συγκεκριμένο πρόβλημα.

4.3 Στόχος Προτεινόμενης Μεθοδολογίας

Βασικός στόχος της μεθοδολογίας που αναλύθηκε παραπάνω είναι η αποδοτική ομαδοποίηση δεδομένων από κτίρια ώστε να δημιουργήσει ένα σαφές και οργανωμένο σύνολο δεδομένων για την καλύτερη διαχείριση και ανάλυση των πληροφοριών που σχετίζονται με κτίρια. Μια

αποτελεσματική ομαδοποίηση επιτρέπει την εύκολη πρόσβαση σε πληροφορίες σχετικά τα χαρακτηριστικά των κτιρίων. Μέσω αυτής της οργάνωσης, οι επαγγελματίες του χώρου, όπως αρχιτέκτονες, μηχανικοί, διαχειριστές κτιρίων και άλλοι, μπορούν βοηθηθούν ώστε να λαμβάνουν γρήγορες και ενημερωμένες αποφάσεις σχετικά με την ανακαίνιση κτιρίων, ενώ η ανάλυση δεδομένων μπορεί να αποκαλύψει τάσεις και πρότυπα που οδηγούν σε βελτιώσεις στην απόδοση και την ενεργειακή απόδοση των κτιρίων με απώτερο σκοπό τη μείωση των ρύπων που προέρχονται από τα κτίρια. Κατά συνέπεια, η καλύτερη ομαδοποίηση κτιριακών δεδομένων αποτελεί το θεμέλιο για μια αποτελεσματική και βιώσιμη διαχείριση των κτιρίων σε όλες τις πτυχές τους.

ΚΕΦΑΛΑΙΟ 5

Εφαρμογή Προτεινόμενης Μεθοδολογίας

5 Εφαρμογή Προτεινόμενης Μεθοδολογίας

Στο πέμπτο κεφάλαιο της εργασίας γίνεται ανάλυση της μεθοδολογίας που χρησιμοποιήθηκε και παρουσιάζονται τα αποτελέσματα που προέκυψαν. Αρχικά, γίνεται μια εισαγωγή στη μηχανική μάθηση και αναφέρονται οι αλγόριθμοι που χρησιμοποιούνται πιο συχνά σε προβλήματα ομαδοποίησης. Ειδικότερα, δίνεται μια σύντομη περιγραφή και παρουσιάζονται σύντομα κάποια πλεονεκτήματα και μειονεκτήματα της κάθε μεθόδου. Στη συνέχεια, επιλέγονται για εφαρμογή οι αλγόριθμοι KMeans, Agglomerative Clustering, DBSCAN και HDBSCAN καθώς ταιριάζουν καλύτερα στο σύνολο δεδομένων. Ακολουθεί η περιγραφή του συνόλου δεδομένων και η προ επεξεργασία του. Στο τέλος παρουσιάζεται η εφαρμογή αυτών των αλγορίθμων στο πρόβλημα που εξετάζεται, με σκοπό την εξαγωγή των κατάλληλων αποτελεσμάτων. Με αυτόν τον τρόπο, δίνεται μια πλήρης εικόνα της διαδικασίας και των αποτελεσμάτων που προέκυψαν από την εφαρμογή της μεθοδολογίας.

5.1 Μηχανική Μάθηση

Στόχος της μηχανικής μάθησης είναι η όσο το δυνατόν ακριβής πρόβλεψη. Ανάλογα με τα είδη των προβλημάτων που αντιμετωπίζονται, μπορεί να υποδιαιρεθεί στην Επιβλεπόμενη Μάθηση, στη Μη Επιβλεπόμενη Μάθηση και στην Ενισχυτική Μάθηση και σε διάφορους συνδυασμούς τους [30].

Σκοπός των προβλημάτων της Επιβλεπόμενης μάθησης είναι η πρόβλεψη της ορθούς εξόδου ή ετικέτας μέσω μιας εισόδου που δίνεται. Η πιο απλή περίπτωση είναι οι απαντήσεις-έξοδοι να έχουν τη μορφή ναι/όχι. Αυτά τα είδη προβλημάτων ονομάζονται δυαδικά προβλήματα ταξινόμησης.

Στη Μη επιβλεπόμενη μάθηση δεν υπάρχουν ετικέτες ούτε ορθές έξοδοι. Βασικός στόχος αποτελεί η ανακάλυψη της δομής των δεδομένων με σκοπό τη βέλτιστη ομαδοποίηση των δεδομένων ή τη μείωση των δεδομένων σε έναν μικρό αριθμό σημαντικών «διαστάσεων». Η οπτικοποίηση δεδομένων μπορεί επίσης να θεωρηθεί μη επιβλεπόμενη μάθηση.

Η Ενισχυτική μάθηση χρησιμοποιείται συνήθως σε καταστάσεις όπου ένα σύστημα τεχνητής νοημοσύνης, πρέπει να λειτουργήσει σε ένα περιβάλλον όπου η ανατροφοδότηση σχετικά με καλές ή κακές επιλογές είναι διαθέσιμη με κάποια καθυστέρηση. Χρησιμοποιείται επίσης σε παιχνίδια στα οποία το αποτέλεσμα μπορεί να αποφασιστεί μόνο στο τέλος του παιχνιδιού [30].

Οι κατηγορίες επικαλύπτονται και είναι σχετικά ασαφείς, οπότε ενίοτε είναι δύσκολο μια συγκεκριμένη μέθοδος να ενταχθεί σε μία κατηγορία. Για παράδειγμα, όπως υποδηλώνει η ονομασία της, η λεγόμενη ημιεπιβλεπόμενη μάθηση είναι εν μέρει επιβλεπόμενη και εν μέρει μη επιβλεπόμενη [31].

Οι αλγόριθμοι ομαδοποίησης μπορούν να παράγουν πολύ διαφορετικά αποτελέσματα και εξαρτώνται σε μεγάλο βαθμό από τα χαρακτηριστικά του συνόλου δεδομένων. Στην περίπτωση μας, που αφορά δεδομένα από κτίρια, οι επιλογές που μπορούν να αξιοποιηθούν παρουσιάζονται συνοπτικά στο παρακάτω πίνακα (Πίνακας 2). Στα επόμενα υποκεφάλαια αναλύεται περαιτέρω ο τρόπος λειτουργίας τους.

Πίνακας 1: Σύγκριση πιο Συχνά Χρησιμοποιούμενων Αλγορίθμων Ομαδοποίησης

Αλγόριθμος	Περιγραφή	Παράμετροι	Πλεονεκτήματα	Μειονεκτήματα
K-Means	Χωρίζει τα δεδομένα σε k ομάδες, όπου κάθε ομάδα αναπαρίσταται από το μέσο των δειγμάτων	number of clusters	Απλό, γρήγορο και δουλεύει καλά σε μεγάλο σύνολο δεδομένων	Υποθέτει ότι οι ομάδες είναι σφαιρικές και ίσου μεγέθους
OPTICS-DBSCAN	Σχηματίζει ομάδες ενώνοντας δείγματα με βάση την απόσταση	minimum cluster membership	Δεν χρειάζεται προκαθορισμένος ο αριθμός ομάδων	Δεν έχει καλή απόδοση σε πολυδιάστατα δεδομένα
HDBSCAN	Επιτρέπει ιεραρχία των ομάδων	minimum cluster membership, minimum point neighbors	Δεν χρειάζεται προκαθορισμένος ο αριθμός ομάδων και μπορεί να εντοπίσει ομάδες με διάφορες πυκνότητες	Μπορεί να είναι πιο αργό σε μεγάλα σύνολα δεδομένων σε σύγκριση με άλλες μεθόδους
Spectral Clustering	Χρησιμοποιεί ιδιοδιανύσματα από πίνακα ομοιότητας για να μειώσει τη διάσταση των δεδομένων πριν την ομαδοποίηση	number of clusters	Μπορεί να εντοπίσει ομάδες με αυθαίρετα σχήματα	Δεν έχει καλή απόδοση σε μεγάλα σύνολα δεδομένων ή όταν οι ομάδες έχουν διαφορετικό μέγεθος
GMM	Μοντελοποιεί τα δεδομένα σαν συλλογή από κατανομές Gauss	many	Λειτουργεί καλύτερα με κατανεμημένα δεδομένα που αποτελούν μίξη κατανομών Gauss	Απαιτεί τον καθορισμό του αριθμού των γκαουσιανών συνιστωσών
Agglomerative Hierarchical Clustering	Κατασκευάζει μια ιεραρχία δεδομένων ενώνοντας ή χωρίζοντας υπάρχουσες ομάδες	number of clusters or distance threshold, linkage type, distance	Δεν χρειάζεται προκαθορισμένος ο αριθμός ομάδων και μπορεί να παράξει ιεραρχία ομάδων	Μπορεί να είναι αργό σε μεγάλα σύνολα δεδομένων και ευαίσθητο στην επιλογή κριτηρίων σύνδεσης
Mean Shift Clustering	Εντοπίζει πυκνά σημεία σε σύνολο δεδομένων	bandwidth	Δεν χρειάζεται προκαθορισμένος ο αριθμός ομάδων	Υπολογιστικά ακριβό σε μεγάλα σύνολα δεδομένων

Affinity Propagation	Διαχωρίζει σε ομάδες μεταφέροντας μηνύματα μεταξύ ζευγών δεδομένων	damping, sample preference	Δεν χρειάζεται προκαθορισμένος αριθμός ομάδων	Υπολογιστικά ακριβό και ο αριθμός των ομάδων επηρεάζεται από τις παραμέτρους εισόδου
BIRCH	Κατασκευάζει μια δένδροειδή δομή από την οποία εξάγει τις ομάδες		Μπορεί να επεξεργαστεί μεγάλο όγκο δεδομένων χωρίς να χρειάζεται να αποθηκευτούν όλα τα δεδομένα στη μνήμη	Ευαίσθητο στη σειρά της εισόδου των δεδομένων και στην επιλογή των παραμέτρων

5.1.1 KMeans

Ο αλγόριθμος KMeans ομαδοποιεί τα δεδομένα προσπαθώντας να διαχωρίσει τα δείγματα σε k ομάδες ίσης διακύμανσης, ελαχιστοποιώντας ένα κριτήριο γνωστό ως άθροισμα τετραγώνων. Ο αριθμός των ομάδων είναι γνωστός εξ 'αρχής και δίνεται ως παράμετρος στο μοντέλο. Με βάση αυτόν διαιρεί ένα σύνολο δειγμάτων σε k διαχωρισμένες ομάδες, καθεμία από τις οποίες περιγράφεται από το μέσο όρο των δειγμάτων της ομάδας. Οι μέσοι όροι ονομάζονται συνήθως «κεντροειδή» των ομάδων. Ο αλγόριθμος K-means στοχεύει στην επιλογή κεντροειδών που ελαχιστοποιούν το άθροισμα τετραγώνων εντός της ομάδας:

Κλιμακώνεται καλά σε μεγάλους αριθμούς δειγμάτων και έχει χρησιμοποιηθεί σε μεγάλο εύρος εφαρμογών σε πολλούς διαφορετικούς τομείς. Ωστόσο, αυτό λειτουργεί με την υπόθεση ότι οι ομάδες έχουν το ίδιο μέγεθος και είναι σφαιρικές [32].

5.1.2 DBSCAN

Ο DBSCAN (Density-Based Spatial Clustering of Applications with Noise) ομαδοποιεί τα σημεία δεδομένων και ανιχνεύει το θόρυβο στο σύνολο δεδομένων. Ο DBSCAN είναι ένας αλγόριθμος ομαδοποίησης με βάση την πυκνότητα που λειτουργεί με την υπόθεση ότι οι ομάδες είναι πυκνές περιοχές στο χώρο που χωρίζονται από περιοχές χαμηλότερης πυκνότητας. Ειδικότερα, ομαδοποιεί τα "πυκνά" σημεία δεδομένων σε μια ενιαία ομάδα. Μπορεί να εντοπίσει ομάδες σε μεγάλα σύνολα χωρικών δεδομένων εξετάζοντας την τοπική πυκνότητα των σημείων δεδομένων. Το πιο συναρπαστικό χαρακτηριστικό της ομαδοποίησης DBSCAN είναι ότι είναι ανθεκτική στο θόρυβο. Επίσης, δεν απαιτεί να ειπωθεί εκ των προτέρων ο αριθμός των ομάδων, σε αντίθεση με τον KMeans, όπου πρέπει να καθοριστεί ο αριθμός των κέντρων.

Ο DBSCAN αρχικοποιείται με δύο παραμέτρους, την ϵ (epsilon) και την minPoints . Η ϵ είναι η ακτίνα του κύκλου που θα δημιουργηθεί γύρω από κάθε σημείο δεδομένων για τον

έλεγχο της πυκνότητας και η `minPoints` είναι ο ελάχιστος αριθμός σημείων δεδομένων που απαιτούνται εντός του κύκλου για να χαρακτηριστεί το σημείο αυτό ως σημείο πυρήνα.

Για τον εντοπισμό των σημείων δεδομένων στο χώρο, ο DBSCAN χρησιμοποιεί την ευκλείδεια απόσταση, αν και μπορούν να χρησιμοποιηθούν και άλλες μέθοδοι. Χρειάζεται επίσης να σαρώσει ολόκληρο το σύνολο των δεδομένων μία φορά, ενώ σε άλλους αλγορίθμους πρέπει να γίνει πολλές φορές. Ωστόσο, ένα από τα μειονεκτήματα του είναι ότι δεν έχει καλά αποτελέσματα για πολυδιάστατα σύνολα δεδομένων[33].

5.1.3 HDBSCAN

Ο αλγόριθμος HDBSCAN (Hierarchical DBSCAN) αποτελεί μία βελτίωση του DBSCAN. Συγκεκριμένα, ο HDBSCAN βελτιώνει τον DBSCAN καθιερώνοντας μια ιεραρχική αναπαράσταση των ομάδων. Η ιεραρχία που προκύπτει από την εκτέλεση του αλγορίθμου μπορεί να χρησιμοποιηθεί πολύ αποτελεσματικά για την εξαγωγή ομάδων και την ανίχνευση ακραίων τιμών. Ο HDBSCAN ξεπερνά τους περιορισμούς του DBSCAN εντοπίζοντας ομάδες οποιασδήποτε πυκνότητας. Παρόλο που υπάρχουν πολλά πλεονεκτήματα που παρέχει ο αλγόριθμος HDBSCAN, υπάρχει ένα μειονέκτημα που πρέπει να ληφθεί υπόψη. Ο αλγόριθμος έχει συνολική ασυμπτωτική πολυπλοκότητα $O(n^2)$. Αυτή η πολυπλοκότητα μπορεί να έχει σημαντικό αντίκτυπο στο χρόνο εκτέλεσης του αλγορίθμου για μεγάλα σύνολα δεδομένων[34].

5.1.4 Spectral Clustering

Το Spectral Clustering είναι μια τεχνική ομαδοποίησης που χρησιμοποιεί τις ιδιότητες του φάσματος (ιδιοτιμές) ενός γραφήματος που προκύπτει από τα δεδομένα. Αντί να αναλύει απευθείας τα δεδομένα στον αρχικό τους χώρο, μετασχηματίζει τα δεδομένα χρησιμοποιώντας το φάσμα του γράφου γειτνίασης που αναπαριστά τις σχέσεις μεταξύ των σημείων δεδομένων. Συγκεκριμένα, κατασκευάζει έναν πίνακα γειτνίασης από τις αποστάσεις μεταξύ των σημείων, και στη συνέχεια υπολογίζει τις ιδιοτιμές και τα ιδιοδιανύσματα του κανονικοποιημένου λαπλασιανού πίνακα του γράφου. Τα δεδομένα μετασχηματίζονται σε έναν χαμηλότερης διάστασης χώρο που ορίζεται από αυτά τα ιδιοδιανύσματα και εκεί εφαρμόζεται παραδοσιακή ομαδοποίηση, όπως η μέθοδος KMeans. Αυτή η προσέγγιση επιτρέπει την αποτελεσματική ανίχνευση μη γραμμικών δομών στα δεδομένα και είναι ιδιαίτερα χρήσιμη σε περιπτώσεις όπου οι κλασικές μέθοδοι ομαδοποίησης, όπως η KMeans, αποτυγχάνουν να αποδώσουν ικανοποιητικά αποτελέσματα. Ωστόσο, δεν δίνει λειτουργεί καλά με μεγάλο όγκο δεδομένων ή όταν οι ομάδες είναι διαφορετικού μεγέθους [35].

5.1.5 GMM (Gaussian Mixture Models)

Το Gaussian Mixture Models (GMM) είναι μια στατιστική μέθοδος που χρησιμοποιείται για την αναγνώριση και την περιγραφή της υποκείμενης δομής σε σύνολα δεδομένων που μπορεί να προέρχονται από πολλαπλές κανονικές κατανομές. Αντί να υποθέτει ότι όλα τα

δεδομένα προέρχονται από μία και μόνο κατανομή, το GMM υποθέτει ότι τα δεδομένα δημιουργούνται από ένα μείγμα πολλών κανονικών κατανομών με άγνωστες παραμέτρους. Κάθε κανονική κατανομή στο μείγμα χαρακτηρίζεται από τη δική της μέση τιμή και διασπορά, ενώ το μοντέλο εκτιμά επίσης τα βάρη που καθορίζουν τη συμβολή της κάθε κατανομής στο συνολικό μείγμα. Χρησιμοποιώντας τον αλγόριθμο EM (Expectation-Maximization), το GMM εκτιμά επαναληπτικά τις παραμέτρους του μείγματος, βελτιώνοντας την ακρίβεια του μοντέλου. Αυτή η ευελιξία καθιστά τα GMM κατάλληλα για εφαρμογές όπου τα δεδομένα δεν ακολουθούν απλά σχήματα. Ωστόσο, ο αριθμός των κανονικών κατανομών δίνεται αρχικά ως είσοδος στο μοντέλο και η επιλογή του κατάλληλου αριθμού είναι συχνά δύσκολη και μπορεί να απαιτεί εκτεταμένη δοκιμή και σφάλμα ή χρήση πολύπλοκων μεθόδων [36].

5.1.6 Agglomerative Hierarchical Clustering

Ο αλγόριθμος Agglomerative Hierarchical Clustering είναι μια μέθοδος ομαδοποίησης δεδομένων που ανήκει στην κατηγορία των ιεραρχικών αλγορίθμων. Λειτουργεί με την προσέγγιση "από κάτω προς τα πάνω" (bottom-up), ξεκινώντας από κάθε δεδομένο ως μία ξεχωριστή ομάδα και συγχωνεύοντας διαδοχικά τις πιο κοντινές ομάδες με βάση κάποιο κριτήριο απόστασης (π.χ. ευκλείδεια απόσταση, απόσταση Manhattan). Η διαδικασία συνεχίζεται μέχρι να δημιουργηθεί μία μοναδική ομάδα που περιέχει όλα τα δεδομένα ή μέχρι να επιτευχθεί ένα προκαθορισμένο επίπεδο ομαδοποίησης. Το αποτέλεσμα του αλγορίθμου μπορεί να αναπαρασταθεί με ένα δενδρόγραμμα (dendrogram), το οποίο απεικονίζει τη σειρά και τον τρόπο συγχώνευσης των ομάδων. Ο αλγόριθμος αυτός είναι ιδιαίτερα χρήσιμος για την εξερεύνηση της δομής των δεδομένων και την κατανόηση της φυσικής ομαδοποίησης τους. Ωστόσο, μπορεί να είναι αργό σε μεγάλα σύνολα δεδομένων και ευαίσθητο σε κριτήρια σύνδεσης των δεδομένων [37].

5.1.7 Mean Shift Clustering

Το Mean Shift Clustering είναι μια μη παραμετρική μέθοδος ομαδοποίησης που στοχεύει στον εντοπισμό των "μεγαλύτερων πυκνοτήτων" σε ένα σύνολο δεδομένων. Λειτουργεί με την επαναληπτική μετακίνηση σημείων δεδομένων προς την κατεύθυνση της μέγιστης αύξησης της πυκνότητας των δεδομένων, χρησιμοποιώντας έναν πυρήνα πυκνότητας (όπως τον πυρήνα Gaussian) για να υπολογίσει τις τοπικές κεντρικές τάσεις. Κάθε σημείο δεδομένων μετατοπίζεται προς τη μέση τιμή των δεδομένων στην περιοχή του, και αυτή η διαδικασία επαναλαμβάνεται μέχρις ότου τα σημεία συγκλίνουν σε σταθερά σημεία ή "κέντρα", τα οποία αντιπροσωπεύουν τις διάφορες ομάδες του συνόλου δεδομένων. Το Mean Shift Clustering είναι ιδιαίτερα χρήσιμο επειδή δεν απαιτεί προκαθορισμένο αριθμό ομάδων και είναι ικανό να ανιχνεύσει ομάδες οποιουδήποτε σχήματος και μεγέθους, καθιστώντας το ευέλικτο σε διάφορες εφαρμογές μηχανικής μάθησης και ανάλυσης δεδομένων. Ωστόσο, έχει υπολογιστικό κόστος σε μεγάλα σύνολα δεδομένων [38].

5.1.8 Affinity Propagation

Ο Affinity Propagation είναι ένας αλγόριθμος μηχανικής μάθησης για ομαδοποίηση δεδομένων. Σε αντίθεση με άλλους αλγόριθμους όπως ο KMeans, δεν απαιτεί τον καθορισμό του αριθμού των ομάδων εκ των προτέρων. Αντίθετα, χρησιμοποιεί μια διαδικασία ανταλλαγής μηνυμάτων μεταξύ των σημείων δεδομένων για να αναγνωρίσει παραδείγματα, τα οποία λειτουργούν ως "κεντρικά σημεία" ή "εκπρόσωποι" για τις διάφορες ομάδες. Οι ομάδες διαμορφώνονται με βάση την ομοιότητα των σημείων δεδομένων και το πόσο καλά αντιπροσωπεύονται από τα κεντρικά σημεία. Αυτή η μέθοδος είναι ιδιαίτερα χρήσιμη σε περιπτώσεις όπου η δομή των δεδομένων είναι περίπλοκη και οι κλασικές μέθοδοι ομαδοποίησης δεν είναι αποτελεσματικές. Ωστόσο, ο αλγόριθμος μπορεί να είναι υπολογιστικά απαιτητικός, ειδικά για μεγάλα σύνολα δεδομένων, λόγω της ανάγκης υπολογισμού και αποθήκευσης όλων των ζευγών ομοιοτήτων και της ανταλλαγής μηνυμάτων μεταξύ των σημείων. Άλλο ένα μειονέκτημα είναι ότι η απόδοση του αλγόριθμου μπορεί να επηρεαστεί σημαντικά από τις αρχικές παραμέτρους, όπως η προτίμηση (preference). Η επιλογή αυτών των παραμέτρων συχνά απαιτεί πειραματισμό [39].

5.1.9 BIRCH (Balanced Iterative Reducing and Clustering using Hierarchies)

Ο BIRCH (Balanced Iterative Reducing and Clustering using Hierarchies) είναι ένας αλγόριθμος ανάλυσης δεδομένων που χρησιμοποιείται για την αποδοτική και κλιμακούμενη ομαδοποίηση μεγάλων συνόλων δεδομένων. Αναπτύχθηκε με στόχο την επίτευξη υψηλής απόδοσης και χαμηλής κατανάλωσης μνήμης, καθιστώντας τον ιδανικό για εφαρμογές με μεγάλα δεδομένα. Ο αλγόριθμος λειτουργεί σε δύο βασικά στάδια: πρώτον, δημιουργεί μια συμπτυκνωμένη και ισορροπημένη αναπαράσταση των δεδομένων με τη μορφή μιας δενδροειδούς δομής, που ονομάζεται CF (Clustering Feature) tree, και δεύτερον, εφαρμόζει έναν πιο παραδοσιακό αλγόριθμο ομαδοποίησης πάνω στο CF tree για να διαχωρίσει τα δεδομένα σε ομάδες. Αυτή η προσέγγιση επιτρέπει στο BIRCH να χειρίζεται μεγάλες ποσότητες δεδομένων αποδοτικά και να παράγει καλής ποιότητας ομάδες με λιγότερους υπολογιστικούς πόρους σε σχέση με άλλους παραδοσιακούς αλγορίθμους ομαδοποίησης. Ωστόσο, είναι ευαίσθητος στην επιλογή των παραμέτρων του, όπως το μέγεθος των ομάδων και το κατώφλι που χρησιμοποιείται για τη διάσπαση των κόμβων. Αν αυτές οι παράμετροι δεν επιλεγθούν σωστά, η απόδοση του αλγορίθμου μπορεί να επηρεαστεί σημαντικά. Επίσης, επειδή ο BIRCH βασίζεται στη δημιουργία και διαχείριση ενός CF tree, μπορεί να παρουσιάσει δυσκολίες όταν τα δεδομένα έχουν υψηλή διασπορά ή δεν κατανέμονται ομοιόμορφα. Αυτό μπορεί να οδηγήσει σε υποβέλτιστες ομάδες ή σε μεγάλη χρήση μνήμης [40].

5.2 Δεδομένα

Για την αξιολόγηση της προτεινόμενης μεθοδολογίας χρησιμοποιείται ένα σύνολο 312 έργων ανακαινίσεων που ολοκληρώθηκαν στην Λετονία. Το σύνολο των δεδομένων από τα έργα περιλαμβάνει (i) το κόστος ανακαίνισης, (ii) την τρέχουσα ενεργειακή κατανάλωση του κτιρίου, (iii) τα χαρακτηριστικά του κτιρίου, καθώς και (iv) την αναμενόμενη και (v) την πραγματοποιηθείσα εξοικονόμηση ενέργειας από διάφορα έργα ΕΞΕ που ολοκληρώθηκαν στο παρελθόν. Το σύνολο δεδομένων αυτών περιλαμβάνει 3567 γραμμές και 110 στήλες.

Μια μεγάλη δοκιμασία στα προβλήματα ομαδοποίησης είναι η επιλογή των χαρακτηριστικών που θα επιλεχθούν ως είσοδος. Είναι σημαντικό το σύνολο των χαρακτηριστικών που επιλεχθούν να είναι αντιπροσωπευτικό του σημείου ενδιαφέροντος αλλά και αρκετά μικρό ώστε να ελαχιστοποιηθεί το σφάλμα μη σωστής ομαδοποίησης. Μετά από προσεκτική μελέτη του συνόλου δεδομένων και του προβλήματος επιλέγονται τα παρακάτω χαρακτηριστικά:

- 'Year of building (first year and years of renovations/additions where applicable)'
- 'Number of floors'
- 'Source of heat (Heating)'
- 'Source of heat (Hot water)'
- 'Temperature-Average'
- 'Heating period, days-Total'
- 'Heating, MWh-Total'
- 'Total heating area, m2'
- 'House functions'
- 'Energy consumption before project (MWh)'
- 'Energy consumption before project (kWh/m2)'
- 'CO₂ emission before project (tons)'
- 'CO₂ emission factor (tons CO₂/MWh) (Heating)'
- 'Total energy consumption, MWh (Heating before project)'
- 'CO₂ emission, tons (Heating before project)'

Το μέγεθος του συνόλου δεδομένων έχει, μετά την επιλογή των χαρακτηριστικών, 3567 γραμμές και 21 στήλες. Περιέχει 214 κενές τιμές (Nan values) στη στήλη 'Age of Building', 70 κενές τιμές στη στήλη 'Number of floors' και 1390 στη στήλη 'Source of heat(Hot water)'. Επίσης, περιέχει κάποιες μηδενικές τιμές. Συγκεκριμένα, περιέχει 5 στο

'Heating period, days-Total', 6 στο 'Heating, MWh-Total', 2362 στο 'CO₂ emission factor (tons CO₂/MWh) (Hot water)', 1762 στο 'CO₂ emission factor (tons CO₂/MWh) (Electricity)', 2309 στο 'Total energy consumption, MWh (Hotwater before project)', 2854 στο 'CO₂ emission, tons (Hotwater before project)', 1145 στο 'Total energy consumption, MWh (Electricity before project)' και 1819 στο 'CO₂ emission, tons (Electricity before project)'.

Ένα βασικό ζήτημα στη μηχανική μάθηση είναι η διαχείριση των τιμών που λείπουν ή οι μηδενικές τιμές καθώς έτσι το σύνολο δεδομένων περιλαμβάνει ελλιπείς πληροφορίες και επηρεάζεται η αξιοπιστία των μοντέλων που θα εκπαιδευτούν στη συνέχεια. Είναι σημαντικό να αντιμετωπιστούν αποτελεσματικά ώστε να εξασφαλιστούν τα καλύτερα αποτελέσματα. Υπάρχουν διάφοροι τρόποι για τη διαχείριση τους.

Ένας τρόπος είναι η απομάκρυνση των γραμμών με τα δεδομένα που λείπουν. Όμως έτσι το σύνολο δεδομένων γίνεται πολύ μικρό και τα δεδομένα που απομένουν δεν είναι αρκετά για τα μοντέλα ομαδοποίησης. Επομένως, σε δεύτερη φάση απομακρύνονται οι στήλες που περιέχουν πολλά μηδενικά. Τα χαρακτηριστικά που επιλέγονται για αφαίρεση είναι τα εξής:

- 'CO₂ emission factor (tons CO₂/MWh) (Hot water)'
- 'CO₂ emission factor (tons CO₂/MWh) (Electricity)'
- 'Total energy consumption, MWh (Hotwater before project)'
- 'CO₂ emission, tons (Hotwater before project)'
- 'Total energy consumption, MWh (Electricity before project)'
- 'CO₂ emission, tons (Electricity before project)'

Ένας άλλος τρόπος διαχείρισης αποτελεί η συμπλήρωση των ελλিপών τιμών. Μετά από μελέτη του συνόλου δεδομένων παρατηρούμε ότι μπορούμε να συμπληρώσουμε τις τιμές που λείπουν στο ‘Age of Building’ και το ‘Number of floors’ με τιμές από την προηγούμενη γραμμή καθώς παρατηρήθηκε ότι δεδομένα που βρίσκονται σε ίδιες γραμμές παρουσιάζουν ομοιότητες. Στο χαρακτηριστικό ‘Source of heat (Hot water)’ οι τιμές που λείπουν αντικαθίστανται με τυχαίες υπάρχουσες του συνόλου δεδομένων.

Το σύνολο δεδομένων που απομένει έχει 3567 γραμμές και 15 στήλες. Για την συνοπτική απεικόνιση των δεδομένων χρησιμοποιείται η συνάρτηση `DataFrame.info` από τη βιβλιοθήκη `pandas` της `python` [41]. Η μορφή των δεδομένων είναι η ακόλουθη (Πίνακας 3):

```
#      Column                                     Non-Null Count  Dtype
---  -
0      Age of Building                             3567 non-null     object
1      Number of floors                             3567 non-null     object
2      Source of heat (Heating)                     3567 non-null     object
3      Source of heat (Hot water)                    3567 non-null     object
4      Temperature-Average                         3567 non-null     float64
5      Heating period, days-Total                    3567 non-null     int64
6      Heating, MWh-Total                            3567 non-null     float64
7      Total heating area, m2                        3567 non-null     float64
8      House functions                              3567 non-null     object
9      Energy consumption before project (MWh)        3567 non-null     float64
10     Energy consumption before project (kWh/m2)    3567 non-null     float64
11     CO2 emission before project (tons)            3567 non-null     float64
12     CO2 emission factor (tons CO2/MWh) (Heating)  3567 non-null     float64
13     Total energy consumption, MWh (Heating before project) 3567 non-null     float64
14     CO2 emission, tons (Heating before project)    3567 non-null     float64
dtypes: float64(9), int64(1), object(5)
```

Πίνακας 2: Συνοπτική Απεικόνιση Δεδομένων από Ανακαινίσεις Κτιρίων στην Λετονία

Το χαρακτηριστικό ‘Age of Building’ περιλαμβάνει δύο ημερομηνίες σε μορφή ‘1960-1965’, επομένως αντικαθίστανται με μια τυχαία από τις δύο που δίνονται και το μετατρέπεται σε αριθμητική τιμή (int). Το ‘Number of floors’ περιέχει ορόφους σε μορφή ‘2,3,4’, οι οποίοι αντικαθίστανται με τον μεγαλύτερο αριθμό που δίνεται και μετατρέπεται σε αριθμητική τιμή. Επομένως τα δεδομένα έχουν τελικά 3 κατηγορηματικές τιμές και 12 αριθμητικές. Για τον υπολογισμό περιγραφικών και στατιστικών στοιχείων των αριθμητικών τιμών του συνόλου δεδομένων χρησιμοποιείται η συνάρτηση `Dataframe.describe` από τη βιβλιοθήκη `pandas` της `python` [42]. Ο πίνακας (Πίνακας 4) που προκύπτει παρουσιάζεται στο παρακάτω σχήμα:

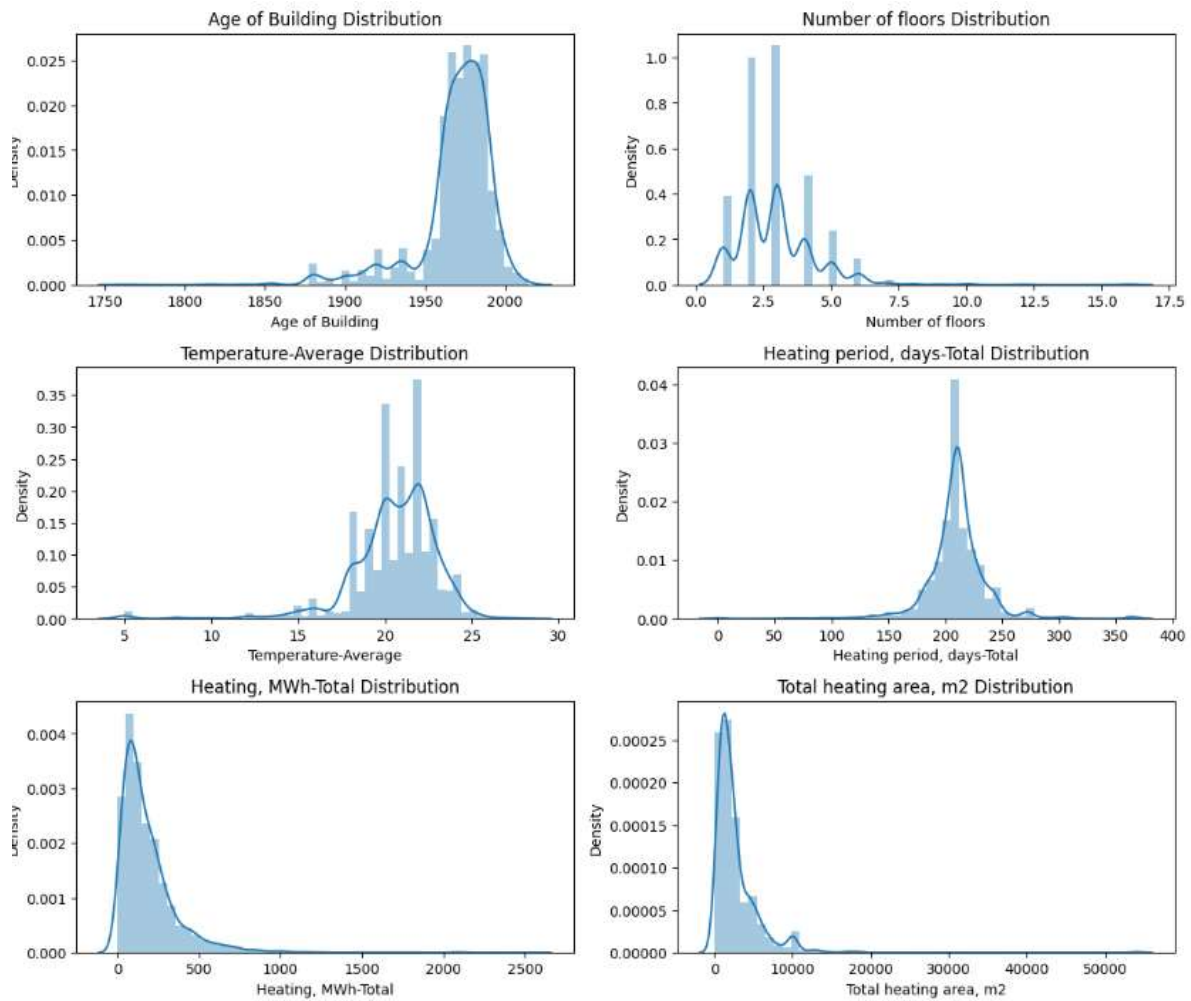
	Age of Building	Number of floors	Temperature-Average	Heating period, days-Total	Heating, MWh-Total	Total heating area, m2	Energy consumption before project (MWh)	Energy consumption before project (kWh/m2)	CO2 emission before project (tons)	CO2 emission factor (tons CO2/MWh) (Heating)	Total energy consumption, MWh (Heating before project)	CO2 emission, tons (Heating before project)
count	3567.000000	3567.000000	3567.000000	3567.000000	3567.000000	3567.000000	3567.000000	3567.000000	3567.000000	3567.000000	3567.000000	3567.000000
unique	1809.000000	2.940000	25.070000	259.980000	259.980000	2771.620000	362.706127	236.020007	229.436616	0.260000	474.941121	127.291022
idx	23.491223	1.470000	2.508934	26.722286	26.722286	3201.968377	566.980003	157.398729	258.642876	0.162000	285.642882	102.147839
mean	1760.000000	1.000000	0.000000	0.000000	0.000000	66.000000	0.500000	80.000000	2.520000	0.200000	0.000000	2.520000
std	1803.000000	2.000000	19.000000	203.000000	203.000000	3333.000000	346.000000	674.622286	65.360000	0.184000	216.700000	84.000000
min	1974.000000	3.000000	21.000000	232.000000	239.940000	3889.000000	428.110000	221.470000	219.947999	0.234000	387.970000	99.072389
25%	1369.000000	4.000000	22.000000	220.000000	241.000000	3222.700000	684.380000	282.500000	163.997999	0.204000	381.380000	163.131000
max	2223.000000	6.000000	26.100000	266.000000	262.700000	6422.100000	624.700000	1712.422286	2872.520000	0.670000	666.660000	785.194000

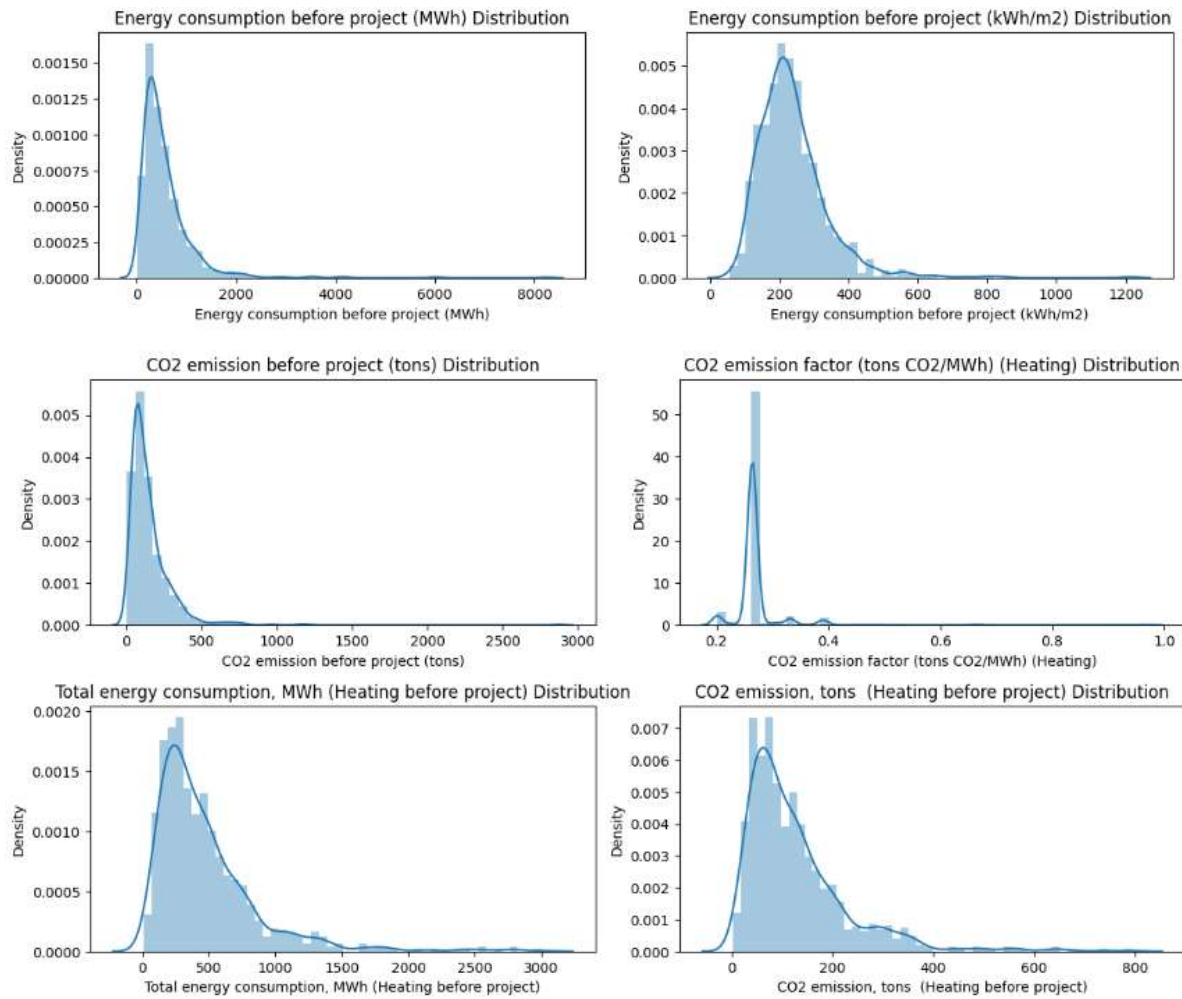
Πίνακας 3: Απεικόνιση περιγραφικών και στατιστικών στοιχείων των αριθμητικών τιμών του υποσυνόλου δεδομένων από έργα ανακαίνισης στη Λετονία

Για την εφαρμογή των δεδομένων στα μοντέλα είναι σημαντικό να προηγηθεί `encoding` για τις κατηγορηματικές μεταβλητές και `scaling` για τις αριθμητικές. Εφαρμόζεται `One-Hot-Encoding` και `StandardScaler` για τις αριθμητικές.

Για την καλύτερη ανάλυση των δεδομένων, τα δεδομένα απεικονίζονται σε διαγράμματα με την κατανομή των τιμών τους χρησιμοποιώντας τις συναρτήσεις `Dataframe.describe` [42]

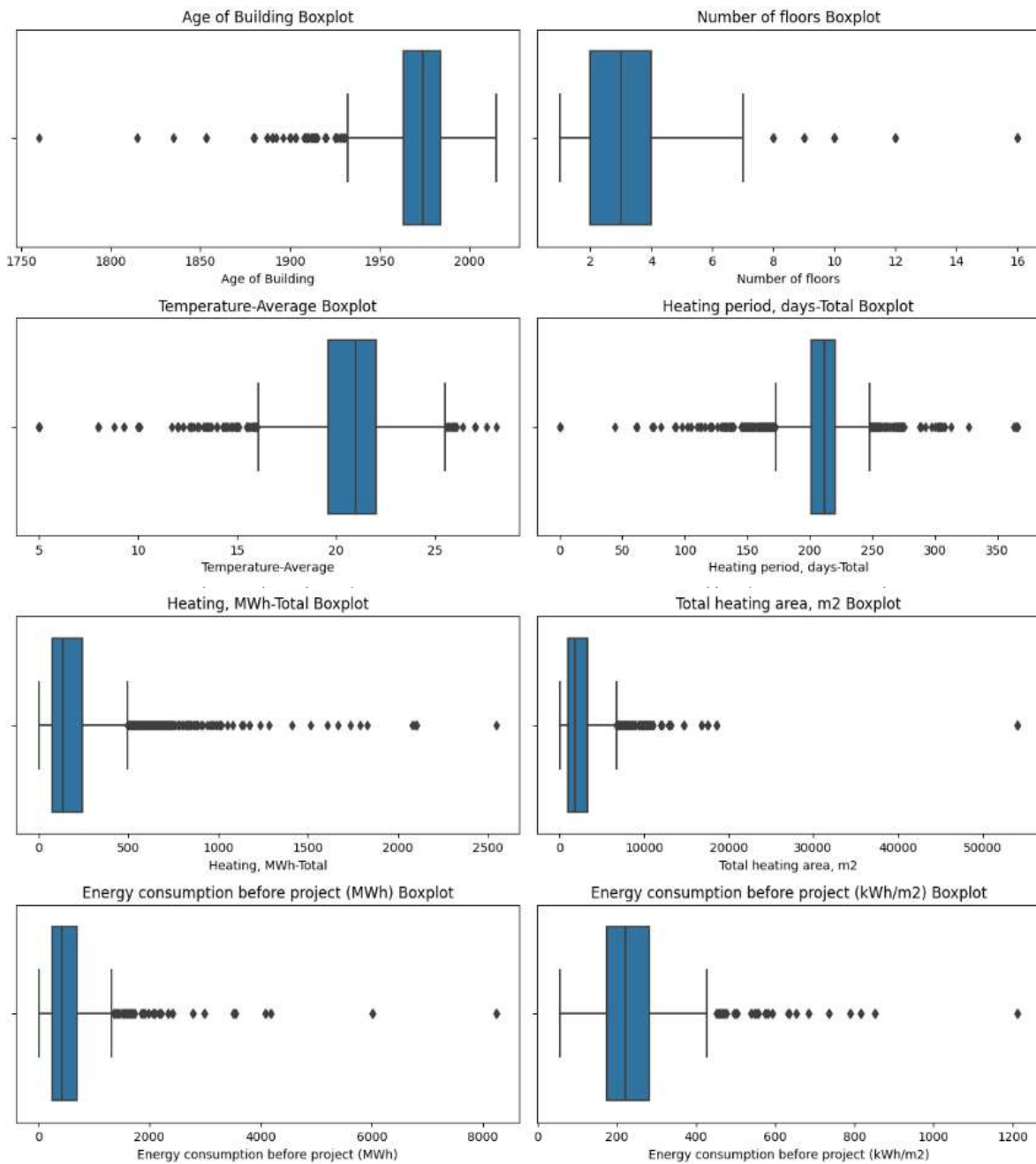
από τη βιβλιοθήκη `pandas` και `displot` [43] από τη βιβλιοθήκη `seaborn` της `python`. Τα διαγράμματα απεικονίζονται στο Σχήμα 2:

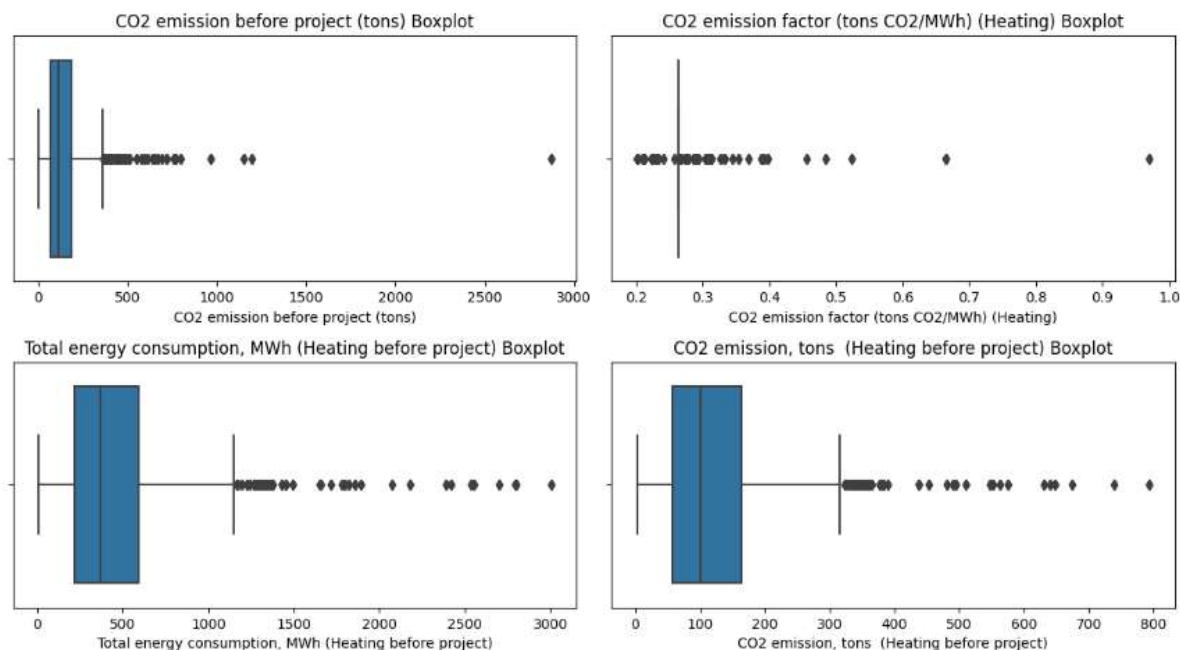




Σχήμα 2: Διαγράμματα με την Κατανομή των τιμών κάθε μεταβλητής του συνόλου δεδομένων από έργα ανακαίνισης στη Λετονία

Επίσης, απεικονίζονται τα boxplots χρησιμοποιώντας τις συναρτήσεις `Dataframe.describe` [28] από τη βιβλιοθήκη `pandas` και `boxplot` [44] από τη βιβλιοθήκη `seaborn` της `python` που παρουσιάζονται στο Σχήμα 3:





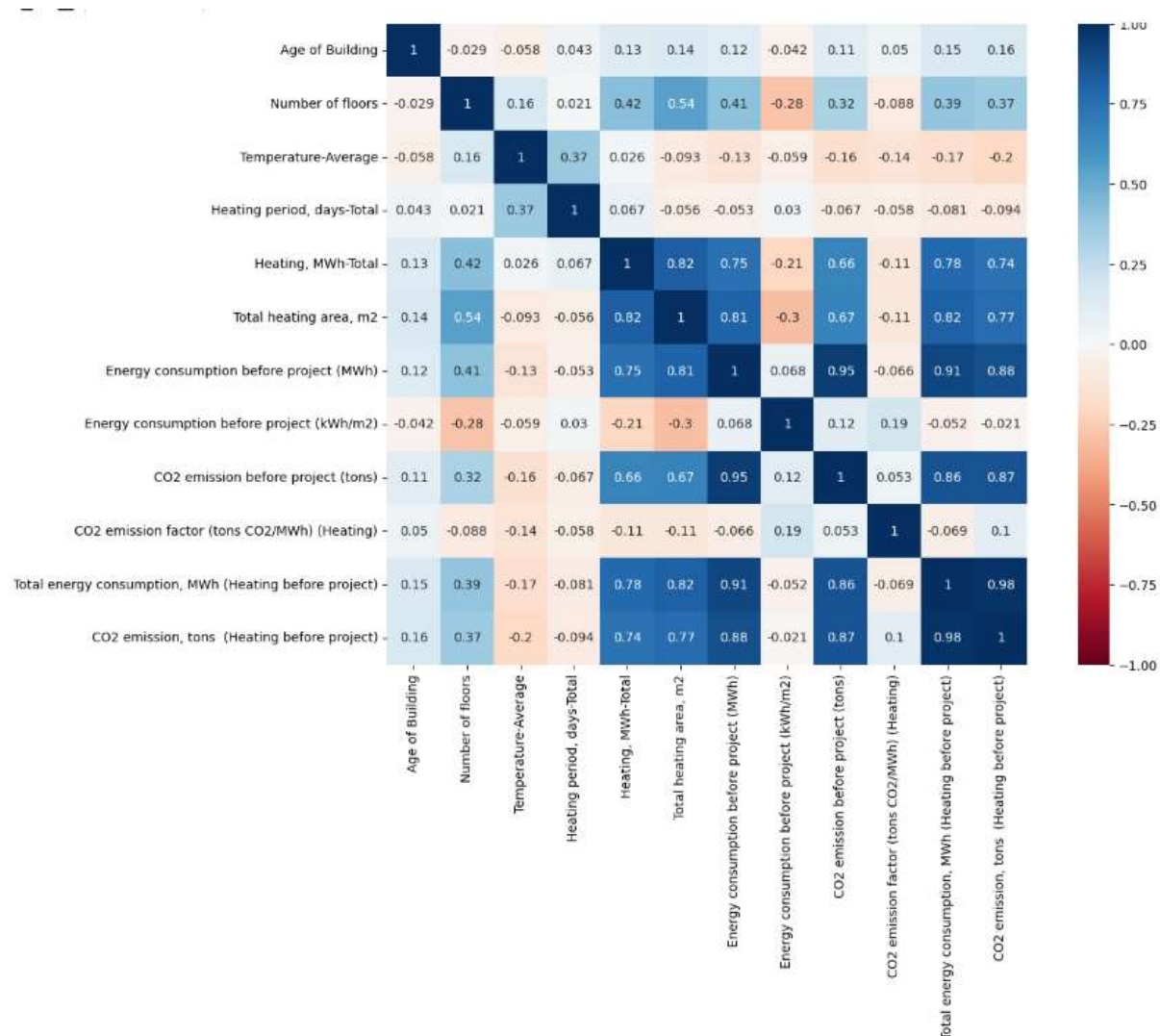
Σχήμα 3: Boxplots για κάθε μεταβλητή του συνόλου δεδομένων που επιλέχτηκε από έργα ανακαίνισης στη Λετονία

Ένα ακόμη χρήσιμο διάγραμμα για την κατανόηση και επιλογή των χαρακτηριστικών που θα δοθούν ως είσοδος στο μοντέλο είναι ο πίνακας συσχέτισης. Ο πίνακας συσχέτισης είναι μια γραφική αναπαράσταση της συσχέτισης ανά ζεύγη μεταξύ των μεταβλητών σε ένα σύνολο δεδομένων. Το διάγραμμα αποτελείται από έναν πίνακα διαγραμμάτων διασποράς και συντελεστών συσχέτισης, όπου κάθε διάγραμμα διασποράς αντιπροσωπεύει τη σχέση μεταξύ δύο μεταβλητών και ο συντελεστής συσχέτισης υποδεικνύει την ισχύ της σχέσης. Η διαγώνιος του πίνακα δείχνει συνήθως την κατανομή κάθε μεταβλητής.

Ο συντελεστής συσχέτισης είναι ένα μέτρο της γραμμικής σχέσης μεταξύ δύο μεταβλητών και κυμαίνεται από -1 έως 1. Ένας συντελεστής 1 υποδηλώνει τέλεια θετική συσχέτιση, όπου η αύξηση μιας μεταβλητής συνδέεται με αύξηση της άλλης μεταβλητής. Ο συντελεστής -1 υποδηλώνει τέλεια αρνητική συσχέτιση, όπου η αύξηση της μιας μεταβλητής συνδέεται με μείωση της άλλης μεταβλητής. Ο συντελεστής 0 υποδηλώνει ότι δεν υπάρχει συσχέτιση μεταξύ των μεταβλητών.

Η επιλογή μεταβλητών που δεν σχετίζονται μεταξύ τους μπορεί να είναι σημαντική για τη βελτίωση της απόδοσης και της αξιοπιστίας ενός μοντέλου μηχανικής μάθησης. Οι μη συσχετισμένες μεταβλητές είναι πιθανό να παρέχουν ανεξάρτητη πληροφορία, συμβάλλοντας έτσι στην αποφυγή της πολυδιεργασίας (multicollinearity), η οποία μπορεί να παραμορφώσει τα αποτελέσματα και να μειώσει την ακρίβεια του μοντέλου. Επιλέγοντας και χρησιμοποιώντας ως είσοδο στο μοντέλο μόνο τις μεταβλητές που δεν σχετίζονται έντονα μεταξύ τους, μπορούμε να δημιουργήσουμε ένα πιο σταθερό και γενικευμένο μοντέλο, το οποίο αποδίδει καλύτερα σε νέα και μη γνωστά δεδομένα [45].

Παρακάτω βλέπουμε έναν πίνακα συσχέτισης που δημιουργήθηκε με τη βοήθεια της βιβλιοθήκης `pandas.DataFrame.corr` από την `python` [46] με τις μεταβλητές του προβλήματος που αναλύουμε:



Πίνακας 4: Πίνακας συσχέτισης των σημαντικότερων χαρακτηριστικών του συνόλου δεδομένων με έργα ανακαίνισης που πραγματοποιήθηκαν στη Λετονία

5.2.1 Επιλογή Σημαντικότερων Χαρακτηριστικών

Μια μεγάλη δοκιμασία στα προβλήματα ομαδοποίησης αποτελεί η επιλογή των χαρακτηριστικών που θα επιλεχτούν ως είσοδος. Είναι σημαντικό το σύνολο των χαρακτηριστικών που επιλεχτούν να είναι αντιπροσωπευτικό του σημείου ενδιαφέροντος αλλά και αρκετά μικρό ώστε να ελαχιστοποιηθεί το σφάλμα μη σωστής ομαδοποίησης.

Μετά από προσεκτική μελέτη του συνόλου δεδομένων, την ανάλυση που προηγήθηκε στην προηγούμενη παράγραφο και του προβλήματος που καλούμαστε να αντιμετωπίσουμε επιλέγουμε τα παρακάτω χαρακτηριστικά:

- 'Year of building (first year and years of renovations/additions where applicable)'
- 'Total heating area, m2'
- 'Energy consumption before project (kWh/m2)'
- 'CO2 emission before project (tons)'
- 'Number of floors'

- 'Temperature-Average'

5.2.2 Προ επεξεργασία Δεδομένων

Η προ επεξεργασία δεδομένων αναδεικνύεται ως ο βασικός πυλώνας στη διαδικασία ανάλυσης δεδομένων και μηχανικής μάθησης. Η σημασία της είναι αναπόσπαστη, καθώς επιτρέπει την επεξεργασία των δεδομένων προτού αυτά χρησιμοποιηθούν σε μοντέλα ή αλγόριθμους μάθησης. Η διαδικασία αυτή περιλαμβάνει την ανίχνευση και αντιμετώπιση ανωμαλιών, την κανονικοποίηση των δεδομένων για τη μείωση του θορύβου και τη βελτιστοποίηση της απόδοσης των μοντέλων. Χωρίς αυτήν τη φάση, τα δεδομένα μπορεί να είναι ασυνάρτητα, παρουσιάζοντας προβλήματα και παραμένοντας αναξιοποίητα σε μεγάλο βαθμό. Ένα ανεπιθύμητο πρόβλημα μπορεί να προκύψει εάν οι τιμές των χαρακτηριστικών ομαδοποίησης διαφέρουν κατά μία ή περισσότερες τάξεις μεγέθους (π.χ. το έτος κατασκευής ποικίλλει από το 1946 έως το 2011, ενώ ο αριθμός των ορόφων είναι πάντα μεταξύ 1 και 3). Οι ομάδες ορίζονται με βάση τον υπολογισμό των αποστάσεων μεταξύ των σημείων δεδομένων σε πολλαπλές διαστάσεις. Συνεπώς, οι μεγαλύτερες μεταβλητές έχουν μεγαλύτερη βαρύτητα από τις μικρότερες μεταβλητές. Συνεπώς, η προεπεξεργασία δεδομένων διαδραματίζει καθοριστικό ρόλο στη διασφάλιση της ακρίβειας, της αξιοπιστίας και της αποτελεσματικότητας των αναλύσεων και των προβλέψεων που βασίζονται στα δεδομένα.

Στις προηγούμενες ενότητες πραγματοποιήθηκε μια πρώτη φάση επεξεργασίας των δεδομένων. Ανιχνεύτηκαν κενές τιμές, πραγματοποιήθηκε η συμπλήρωση τους κ.ά. Στη δεύτερη φάση επεξεργασίας θα πραγματοποιηθεί κανονικοποίηση (normalization) των δεδομένων.

Με την κανονικοποίηση τα δεδομένα μετασχηματίζονται έτσι ώστε να ταιριάζουν σε μια συγκεκριμένη κλίμακα, όπως 0-100 ή 0-1. Η κλιμάκωση χρησιμοποιείται όταν οι αλγόριθμοι που χρησιμοποιούνται για ομαδοποίηση βασίζονται στην απόσταση, δηλαδή στο πόσο μακριά βρίσκονται τα σημεία δεδομένων το ένα με το άλλο. Με αυτούς τους αλγόριθμους, μια αλλαγή του «1» σε οποιοδήποτε αριθμητικό χαρακτηριστικό αποκτά την ίδια σημασία. Έτσι συγκρίνονται διαφορετικές μεταβλητές με ισότιμα βάρη. Δύο τεχνικές scaling που χρησιμοποιούνται και συγκρίνονται ως προς την απόδοση των μοντέλων είναι η MinMaxScaling και η Standard Scaler [47].

5.3 Αλγόριθμοι Ομαδοποίησης

Από τους αλγόριθμους που παρουσιάστηκαν στην προηγούμενη παράγραφο, επιλέγονται για πειραματισμό και αξιολόγηση ο KMeans, ο DBSCAN, ο HDBSCAN και ο Agglomerative Clustering, καθώς ταιριάζουν καλύτερα στο σύνολο δεδομένων, για πειραματισμό και αξιολόγηση.

Κατά τη δοκιμή των μοντέλων είναι σημαντικό να ληφθούν τρία βασικά πράγματα υπόψιν. Αρχικά, τα χαρακτηριστικά και η διάστασή τους. Σε αυτό το στάδιο, εξετάζεται ποια και πόσα χαρακτηριστικά δοκιμάζονται για το μοντέλο, καθώς και το μέγεθος των δεδομένων και οι μέθοδοι για τη μείωση αυτού του μεγέθους. Δεύτερον, οι τεχνικές τροποποίησης

χαρακτηριστικών, όπου εξετάζεται πώς κωδικοποιούνται ή μετασχηματίζονται τα χαρακτηριστικά, καθώς και πώς κλιμακώνονται. Τρίτον, τα μοντέλα και η διαμόρφωσή τους, όπου αξιολογούνται ποια μοντέλα ομαδοποίησης δοκιμάζονται, ποιες παράμετροι χρησιμοποιούνται για τη ρύθμισή τους και πώς επιλέγονται αυτές οι παράμετροι.

5.3.1 Agglomerative Clustering

Πραγματοποιείται εφαρμογή του αλγορίθμου AgglomerativeClustering χρησιμοποιώντας τη συνάρτηση AgglomerativeClustering από τη βιβλιοθήκη scikit learn της python [48] χρησιμοποιώντας ως είσοδο τις παραμέτρους:

- 5 clusters
- ευκλείδεια απόσταση
- Ward

Τα features που επιλέγουμε είναι

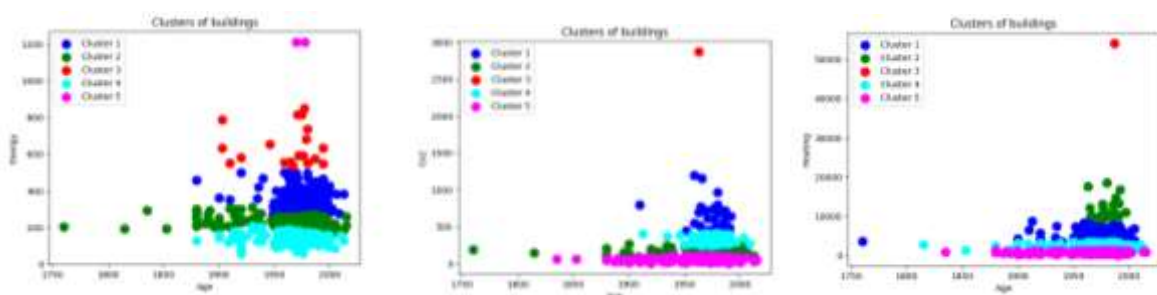
1. 'Year of building (first year and years of rennovations/additions where applicable)'
2. 'Total heating area, m2'
3. 'Energy consumption before project (kWh/m2)'
4. 'CO2 emission before project (tons)'
5. 'Number of floors'
6. 'Temperature-Average'

Για ευκολία, τα features μετονομάζονται σε age, heating, energy, co_2 , floors και temperature αντίστοιχα. Το μοντέλο εκπαιδεύεται με σενάρια εισόδου που αποτελούνται από δύο χαρακτηριστικά, τρία χαρακτηριστικά και πέντε χαρακτηριστικά. Μετά την εκπαίδευση απεικονίζονται τα γραμμικά διαγράμματα αλλά και τα 3d με την ομαδοποίηση που προέκυψε από την εκπαίδευση του αλγορίθμου σχολιάζοντας παράλληλα τα αποτελέσματα που προέκυψαν σε κάθε βήμα.

5.3.1.1 Με δύο χαρακτηριστικά

Αρχικά, εκπαιδεύεται το μοντέλο με είσοδο μόνο δύο χαρακτηριστικά, για την κατανόηση της συσχέτισης και της επιρροής που έχουν στον αλγόριθμο.

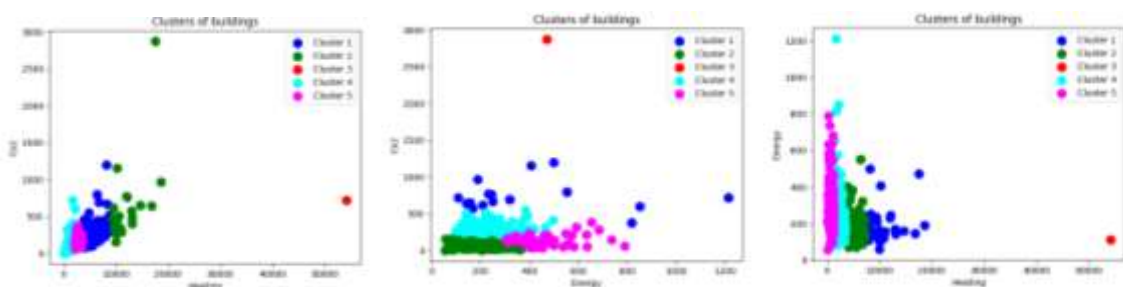
Στα σενάρια που παρουσιάζονται παρακάτω, τα χαρακτηριστικά που λαμβάνονται κυρίως υπόψιν είναι το Energy, το CO_2 και το Heating.



Σχήμα 4: Εφαρμογή Agglomerative Clustering με είσοδο energy-age, co_2 -age, heating-age

Στην περίπτωση που εκπαιδευτεί ο αλγόριθμος με τα χαρακτηριστικά που καθορίζουν την ομαδοποίηση, δηλαδή τα Energy, CO_2 και Heating παρατηρείται ότι λαμβάνει και τα δύο υπόψιν ανά εφαρμογή.

Στην περίπτωση που ο αλγόριθμος εφαρμοστεί με τα χαρακτηριστικά που καθορίζουν την ομαδοποίηση, δηλαδή τα Energy, CO_2 και Heating, ανά δύο κάθε φορά, παρατηρείται ότι λαμβάνει υπόψη και τα δύο χαρακτηριστικά που του δίνονται ως είσοδος όπως φαίνεται και στα παρακάτω διαγράμματα εκτός από την περίπτωση του energy-heating όπου ο αλγόριθμος ομαδοποιεί με βάση το heating. Ο συνδυασμός αυτών των παραμέτρων επιτρέπει στον αλγόριθμο να αναλύει και να κατανοεί καλύτερα τα δεδομένα, οδηγώντας σε πιο ακριβή και ουσιαστική ομαδοποίηση.

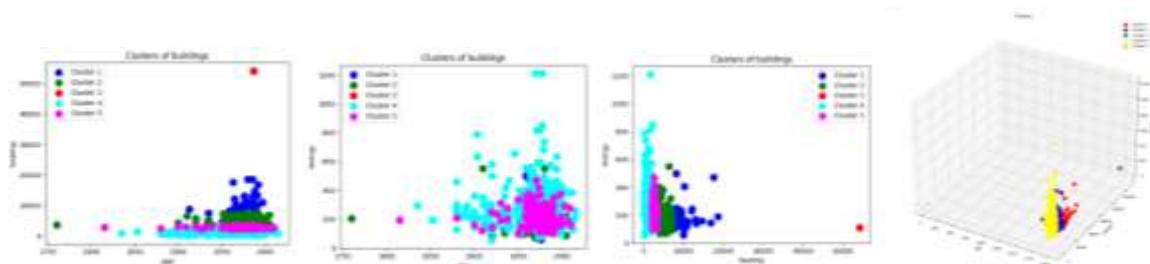


Σχήμα 5: Εφαρμογή Agglomerative Clustering με είσοδο co_2 -heating, co_2 -energy, energy-heating

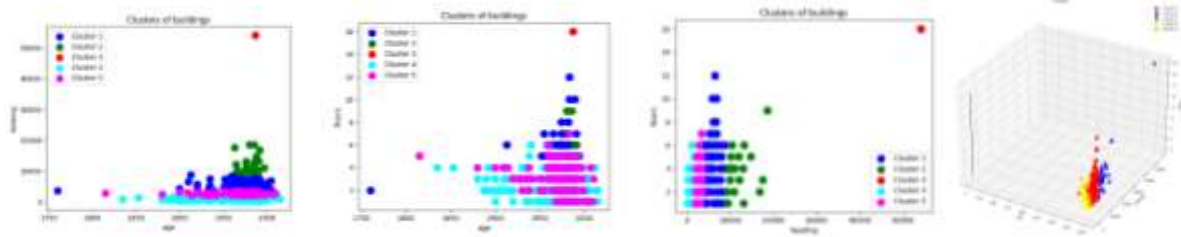
5.3.1.2 Με τρία χαρακτηριστικά

Τα πειράματα συνεχίζονται δίνοντας στο μοντέλο είσοδο τρία χαρακτηριστικά εμπλουτίζοντας την είσοδο του.

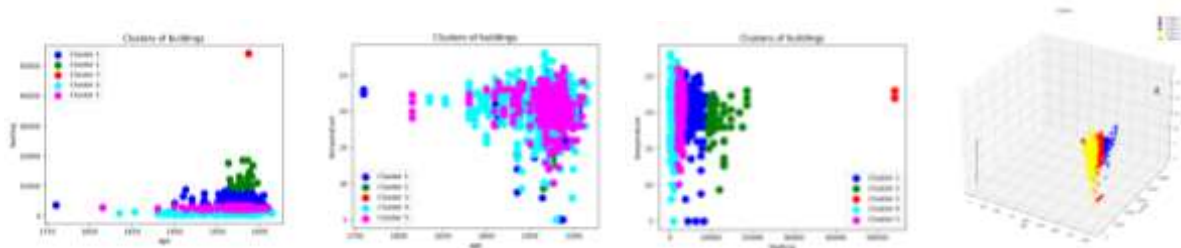
Σε ορισμένα σενάρια όπου εξετάζονταν διάφορα χαρακτηριστικά, ο αλγόριθμος χρησιμοποιούσε μόνο το χαρακτηριστικό "heating" για την ομαδοποίηση. Αυτό συνέβη επειδή το χαρακτηριστικό "heating" κρίθηκε ως το πιο σημαντικό ή αντιπροσωπευτικό για την συγκεκριμένη ανάλυση, αγνοώντας τα υπόλοιπα διαθέσιμα χαρακτηριστικά. Όπως φαίνεται και παρακάτω, το χαρακτηριστικό "heating" είχε μεγαλύτερη επιρροή και ήταν πιο καθοριστικό για την ταξινόμηση των δεδομένων σε ομάδες. Αυτό συνέβη κατά την εκπαίδευση του αλγορίθμου με τα χαρακτηριστικά:



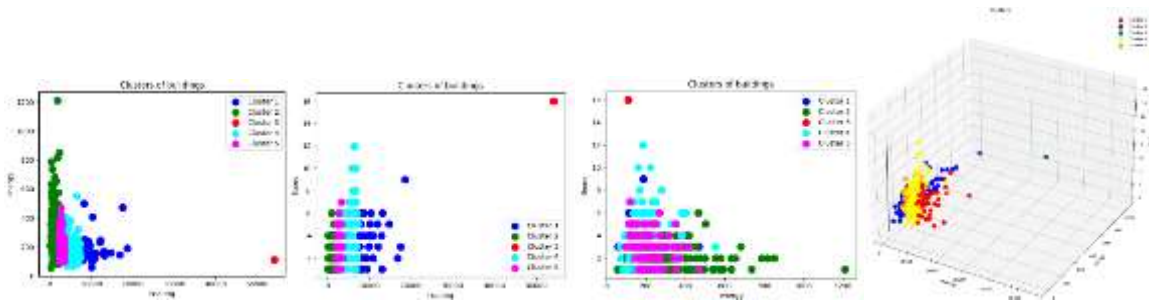
Σχήμα 6: Γραμμικά διαγράμματα χαρακτηριστικών μετά την ομαδοποίηση με Agglomerative Clustering με είσοδο age-heating-energy και 3D διάγραμμα με την απεικόνιση και των τριών χαρακτηριστικών



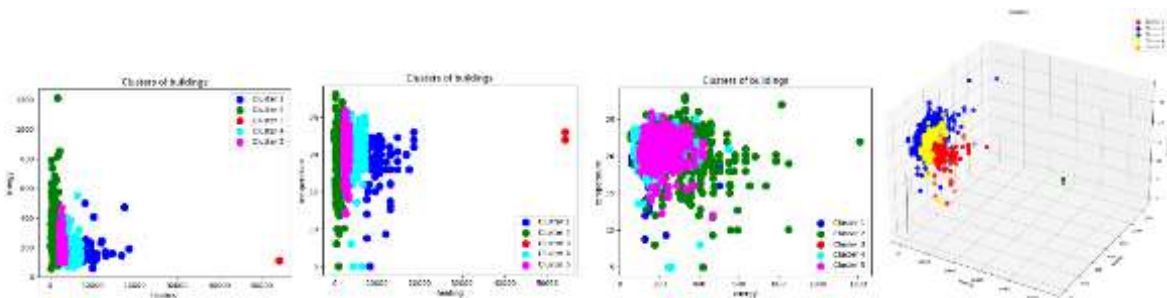
Σχήμα 7: Γραμμικά διαγράμματα χαρακτηριστικών μετά την ομαδοποίηση με *Agglomerative Clustering* με είσοδο *age-heating-floors* και 3D διάγραμμα με την απεικόνιση και των τριών χαρακτηριστικών



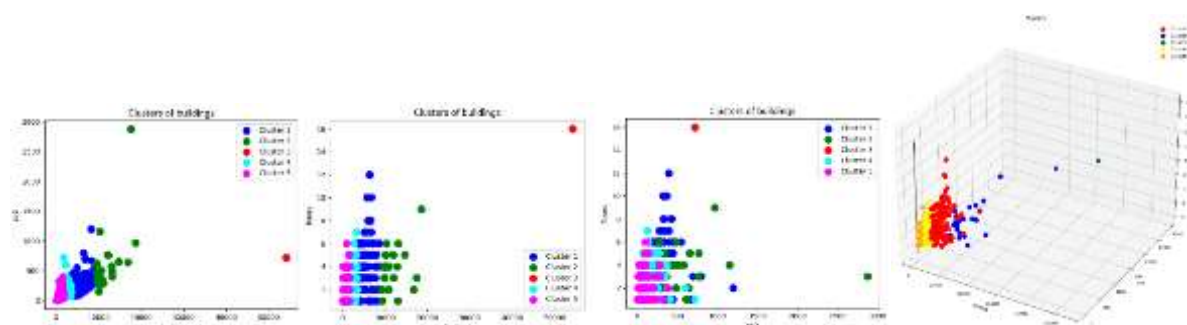
Σχήμα 8: Γραμμικά διαγράμματα χαρακτηριστικών μετά την ομαδοποίηση με *Agglomerative Clustering* με είσοδο *age-heating-temperature* και 3D διάγραμμα με την απεικόνιση και των τριών χαρακτηριστικών



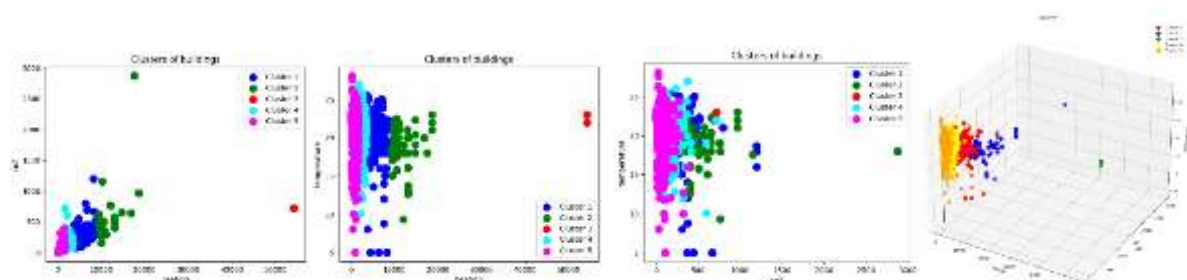
Σχήμα 9: Γραμμικά διαγράμματα χαρακτηριστικών μετά την ομαδοποίηση με *Agglomerative Clustering* με είσοδο *heating-energy-floors* και 3D διάγραμμα με την απεικόνιση και των τριών χαρακτηριστικών



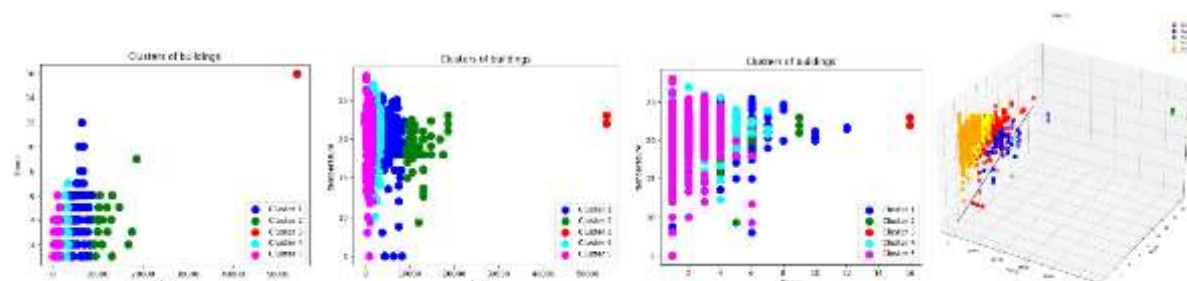
Σχήμα 10: Γραμμικά διαγράμματα χαρακτηριστικών μετά την ομαδοποίηση με *Agglomerative Clustering* με είσοδο *heating-energy-temperature* και 3D διάγραμμα με την απεικόνιση και των τριών χαρακτηριστικών



Σχήμα 11: Γραμμικά διαγράμματα χαρακτηριστικών μετά την ομαδοποίηση με Agglomerative Clustering με είσοδο heating- CO_2 -floors και 3D διάγραμμα με την απεικόνιση και των τριών χαρακτηριστικών

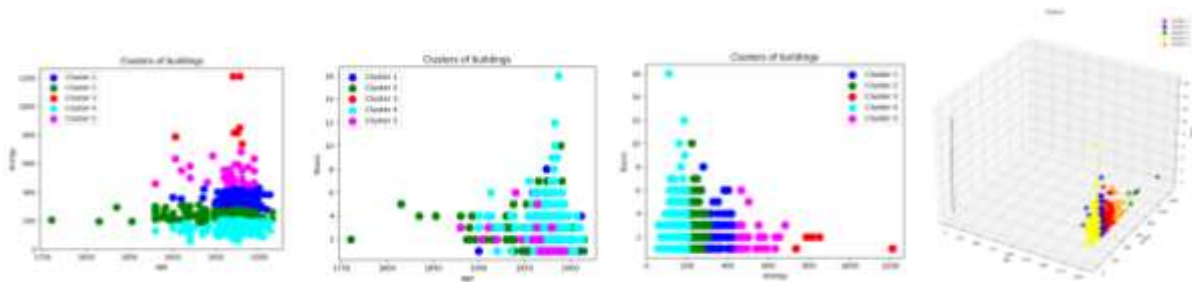


Σχήμα 12: Γραμμικά διαγράμματα χαρακτηριστικών μετά την ομαδοποίηση με Agglomerative Clustering με είσοδο heating- CO_2 -temperature και 3D διάγραμμα με την απεικόνιση και των τριών χαρακτηριστικών

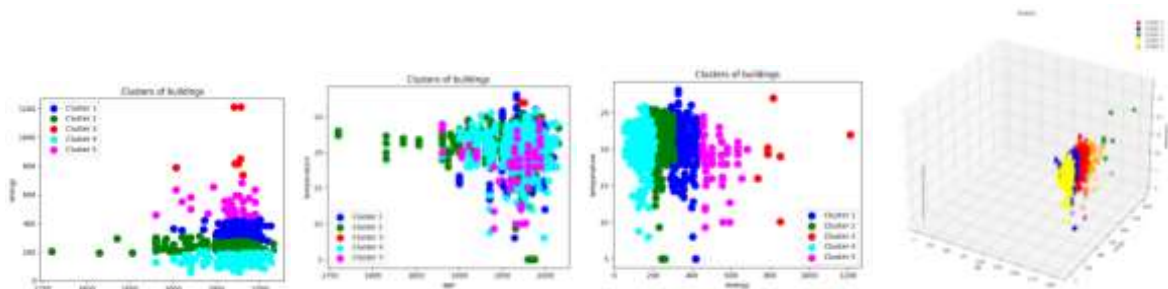


Σχήμα 13: Γραμμικά διαγράμματα χαρακτηριστικών μετά την ομαδοποίηση με Agglomerative Clustering με είσοδο heating-floors-temperature και 3D διάγραμμα με την απεικόνιση και των τριών χαρακτηριστικών

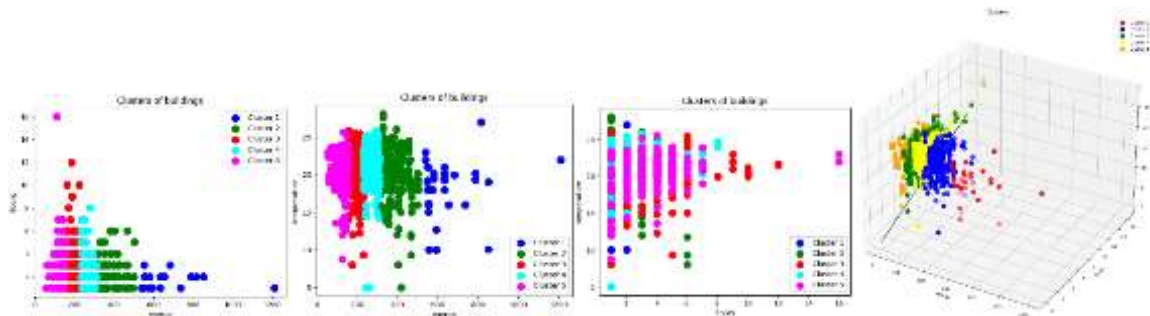
Σε άλλες δοκιμές, δίνοντας ως είσοδο άλλα χαρακτηριστικά, ο αλγόριθμος λαμβάνει υπόψη μόνο το χαρακτηριστικό "energy" για την ομαδοποίηση. Αυτή η προσέγγιση επικεντρώνεται αποκλειστικά στην ενεργειακή κατανάλωση, παρέχοντας μια καθαρή εικόνα για τις ενεργειακές απαιτήσεις και τις αποδοτικότητες των υπό μελέτη συστημάτων. Η χρήση μόνο του χαρακτηριστικού "energy" μπορεί να απλοποιήσει την ανάλυση και να διευκολύνει την αναγνώριση τάσεων και προτύπων που σχετίζονται με την κατανάλωση ενέργειας. Αυτή η στρατηγική μπορεί να είναι ιδιαίτερα χρήσιμη σε εφαρμογές όπου η ενεργειακή απόδοση είναι ο κύριος παράγοντας ενδιαφέροντος και άλλες μεταβλητές δεν έχουν τόσο μεγάλη σημασία για τον σκοπό της μελέτης.



Σχήμα 14: Γραμμικά διαγράμματα χαρακτηριστικών μετά την ομαδοποίηση με Agglomerative Clustering με είσοδο age-energy-floors και 3D διάγραμμα με την απεικόνιση και των τριών χαρακτηριστικών

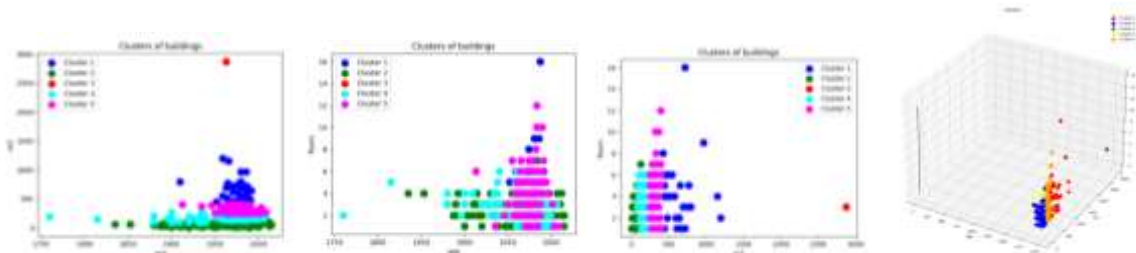


Σχήμα 15: Γραμμικά διαγράμματα χαρακτηριστικών μετά την ομαδοποίηση με Agglomerative Clustering με είσοδο age-energy-temperature και 3D διάγραμμα με την απεικόνιση και των τριών χαρακτηριστικών

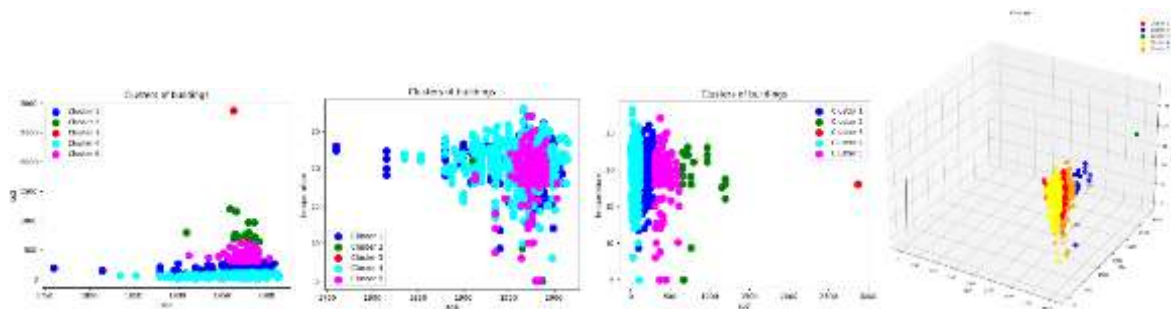


Σχήμα 16: Γραμμικά διαγράμματα χαρακτηριστικών μετά την ομαδοποίηση με Agglomerative Clustering με είσοδο energy-floors-temperature και 3D διάγραμμα με την απεικόνιση και των τριών χαρακτηριστικών

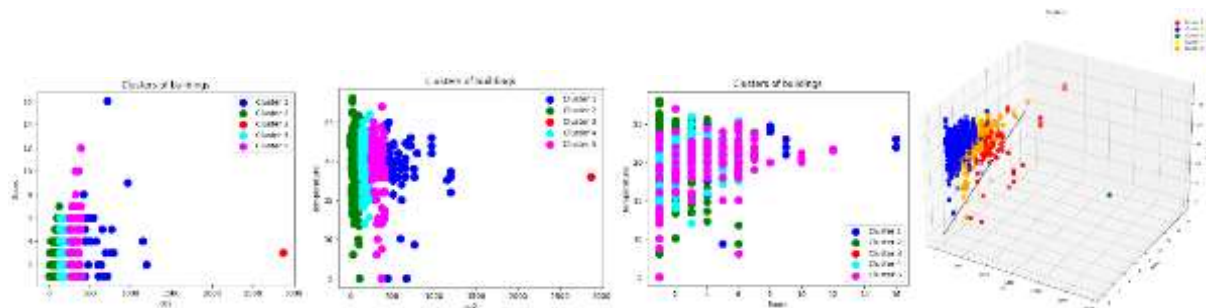
Σε άλλες δοκιμές, ο αλγόριθμος λαμβάνει υπόψη μόνο το χαρακτηριστικό " CO_2 " για την ομαδοποίηση. Αυτό μπορεί να συμβαίνει όταν η ανάλυση επικεντρώνεται κυρίως στις εκπομπές διοξειδίου του άνθρακα και τις περιβαλλοντικές επιπτώσεις τους.



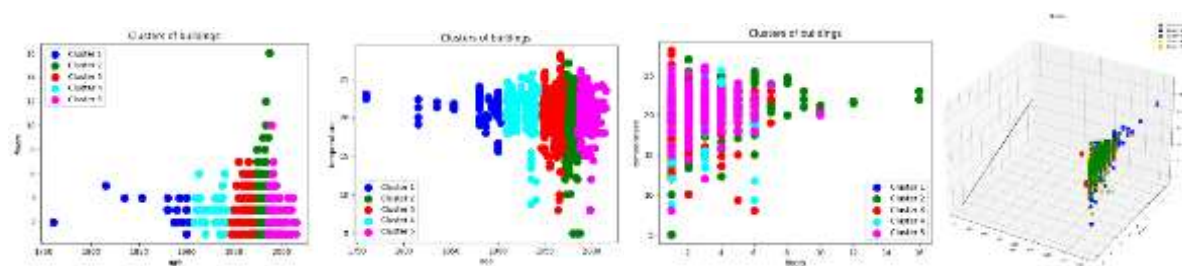
Σχήμα 17: Γραμμικά διαγράμματα χαρακτηριστικών μετά την ομαδοποίηση με Agglomerative Clustering με είσοδο age-co₂-floors και 3D διάγραμμα με την απεικόνιση και των τριών χαρακτηριστικών



Σχήμα 18: Γραμμικά διαγράμματα χαρακτηριστικών μετά την ομαδοποίηση με Agglomerative Clustering με είσοδο age-co₂-temperature και 3D διάγραμμα με την απεικόνιση και των τριών χαρακτηριστικών



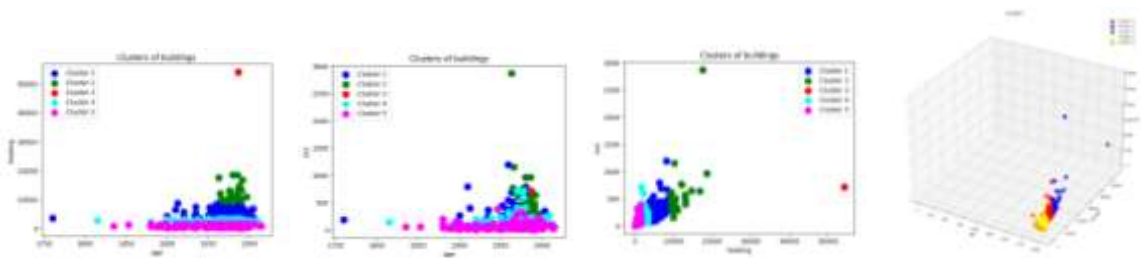
Σχήμα 19: Γραμμικά διαγράμματα χαρακτηριστικών μετά την ομαδοποίηση με Agglomerative Clustering με είσοδο co₂-floors-temperature και 3D διάγραμμα με την απεικόνιση και των τριών χαρακτηριστικών



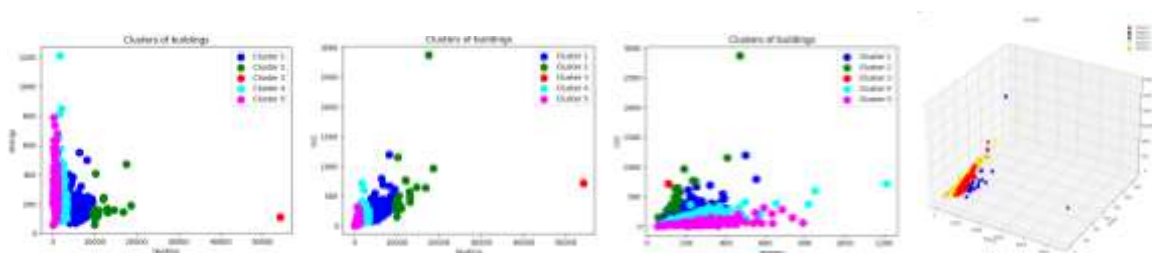
Σχήμα 20: : Γραμμικά διαγράμματα χαρακτηριστικών μετά την ομαδοποίηση με Agglomerative Clustering με είσοδο age-floors-temperature και 3D διάγραμμα με την απεικόνιση και των τριών χαρακτηριστικών

Σε άλλα σενάρια, ο αλγόριθμος λαμβάνει υπόψη τόσο το χαρακτηριστικό 'CO₂' όσο και το "heating" για την ομαδοποίηση. Σε αυτές τις περιπτώσεις, η ανάλυση γίνεται πιο πολύπλοκη καθώς συνυπολογίζονται περισσότερες παράμετροι, γεγονός που μπορεί να οδηγήσει σε πιο ακριβείς και ενημερωμένες ομάδες δεδομένων. Η επιλογή να χρησιμοποιηθούν αυτά τα δύο συγκεκριμένα χαρακτηριστικά μπορεί να βασίζεται στην υπόθεση ότι η συνδυαστική τους επιρροή προσφέρει μια πιο ολοκληρωμένη εικόνα για τα δεδομένα, επιτρέποντας την καλύτερη αναγνώριση και διάκριση των διαφορετικών ομάδων. Αυτή η προσέγγιση μπορεί

να είναι ιδιαίτερα χρήσιμη σε περιπτώσεις όπου οι εκπομπές CO_2 και η θέρμανση έχουν σημαντική αλληλεπίδραση και συνεπώς, η κοινή τους ανάλυση οδηγεί σε πιο ουσιαστικά συμπεράσματα. Τα σενάρια στα οποία παρατηρήθηκαν τα παραπάνω αποτελέσματα είναι:

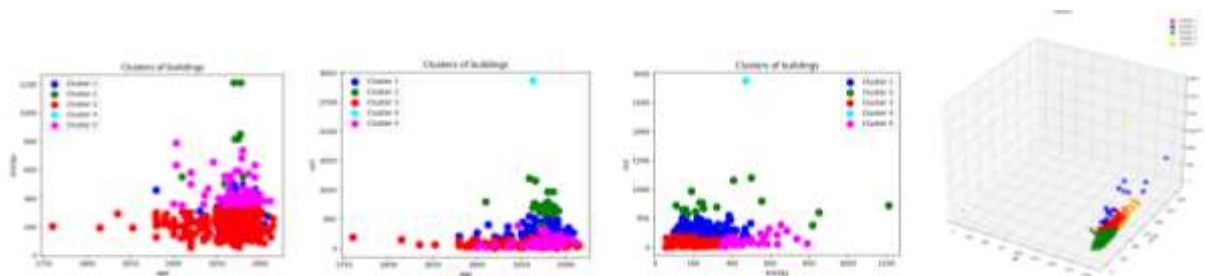


Σχήμα 21: Γραμμικά διαγράμματα χαρακτηριστικών μετά την ομαδοποίηση με Agglomerative Clustering με είσοδο age-heating-co₂ και 3D διάγραμμα με την απεικόνιση και των τριών χαρακτηριστικών

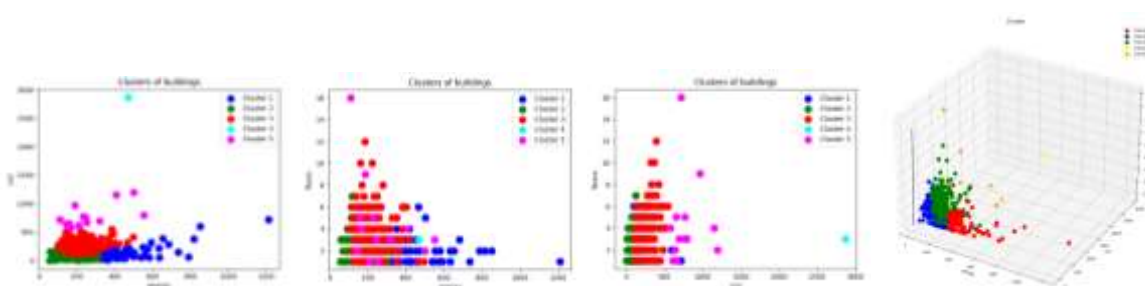


Σχήμα 22: Γραμμικά διαγράμματα χαρακτηριστικών μετά την ομαδοποίηση με Agglomerative Clustering με είσοδο heating-energy-co₂ και 3D διάγραμμα με την απεικόνιση και των τριών χαρακτηριστικών

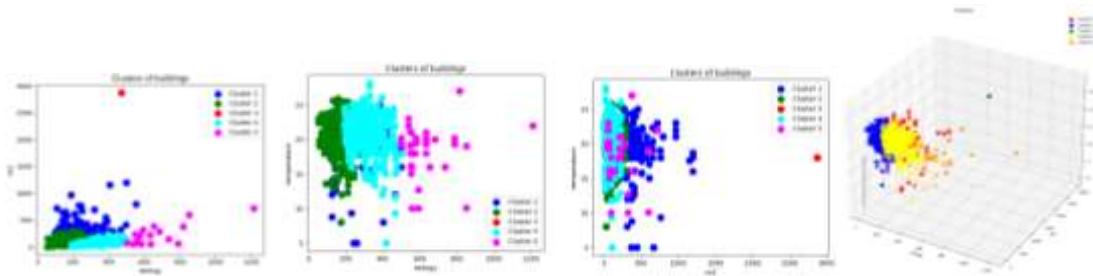
Σε άλλα σενάρια, ο αλγόριθμος εξετάζει τα χαρακτηριστικά 'CO₂' και "energy" για την ομαδοποίηση. Αυτή η προσέγγιση επιτρέπει την ανάλυση της σχέσης μεταξύ των εκπομπών διοξειδίου του άνθρακα και της ενεργειακής κατανάλωσης, προσφέροντας πολύτιμες πληροφορίες για τη δυναμική και την απόδοση των συστημάτων που μελετώνται. Αυτό παρατηρήθηκε κατά την ομαδοποίηση με τα χαρακτηριστικά:



Σχήμα 23: Γραμμικά διαγράμματα χαρακτηριστικών μετά την ομαδοποίηση με Agglomerative Clustering με είσοδο age-energy-co₂ και 3D διάγραμμα με την απεικόνιση και των τριών χαρακτηριστικών



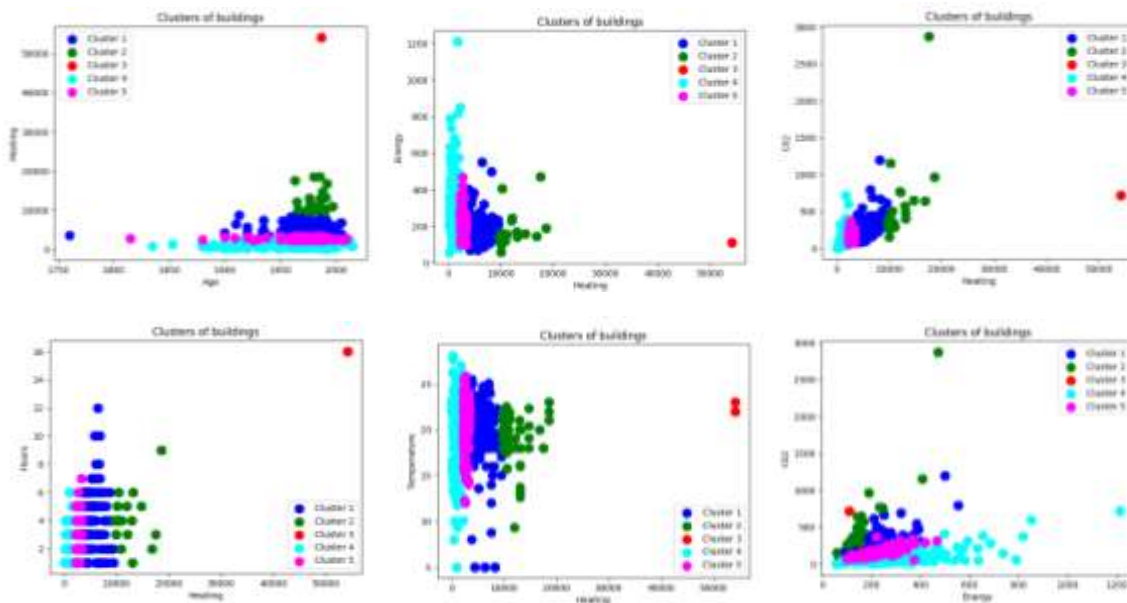
Σχήμα 24: Γραμμικά διαγράμματα χαρακτηριστικών μετά την ομαδοποίηση με Agglomerative Clustering με είσοδο *energy-co₂-floors* και 3D διάγραμμα με την απεικόνιση και των τριών χαρακτηριστικών

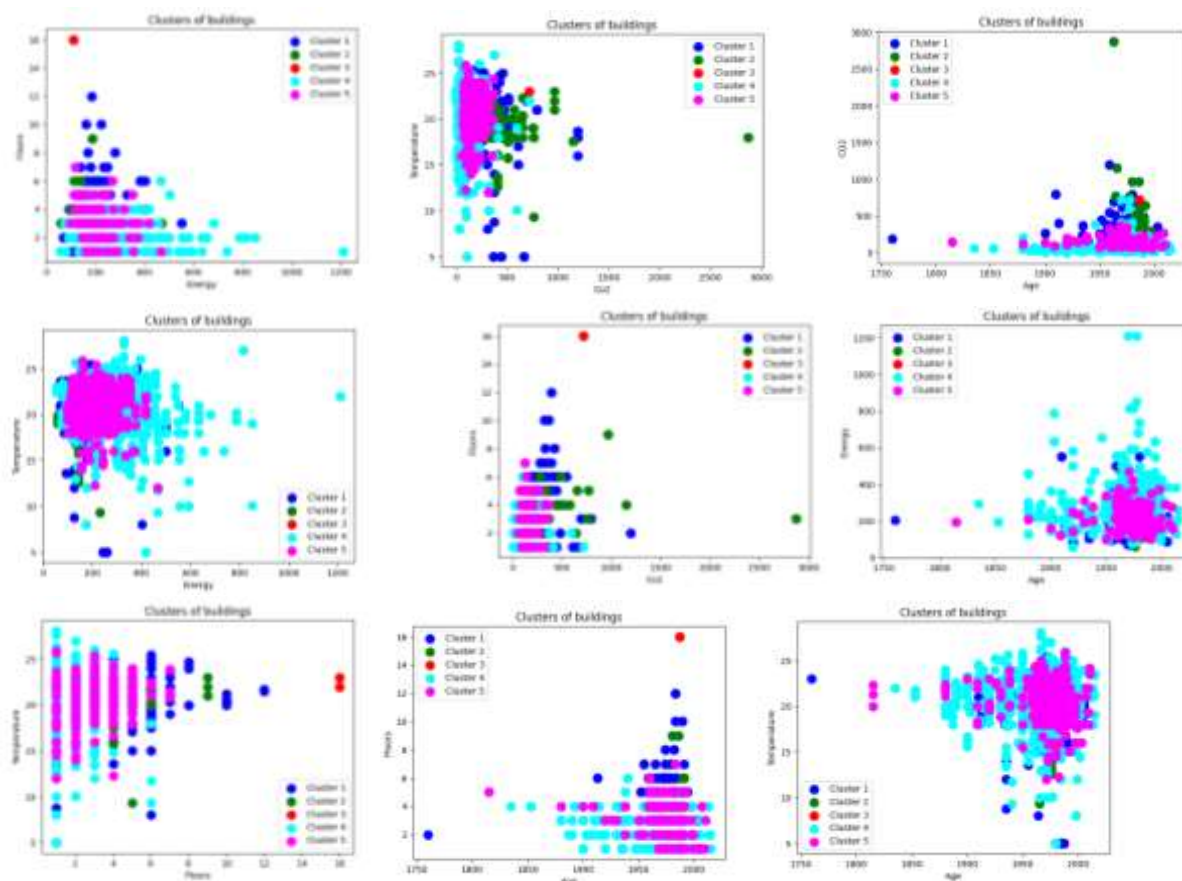


Σχήμα 25: Γραμμικά διαγράμματα χαρακτηριστικών μετά την ομαδοποίηση με Agglomerative Clustering με είσοδο *energy-co₂-temperature* και 3D διάγραμμα με την απεικόνιση και των τριών χαρακτηριστικών

5.3.1.3 Με έξι χαρακτηριστικά

Έπειτα πραγματοποιείται δοκιμή εκπαίδευσης του αλγορίθμου δίνοντας του ως είσοδο τα έξι χαρακτηριστικά *age, energy, heating, co₂, floors* και *temperature* και απεικονίζονται οι ομάδες που προέκυψαν σε γραμμικό διάγραμμα με δύο χαρακτηριστικά για καλύτερη ανάλυση και κατανόηση των αποτελεσμάτων.





Σχήμα 26: Γραμμικά διαγράμματα χαρακτηριστικών μετά την ομαδοποίηση με Agglomerative Clustering με είσοδο age-heating-energy-temperature-floors- CO_2

Από τα παραπάνω διαγράμματα παρατηρούμε ότι ο χωρισμός των ομάδων βασίστηκε κυρίως στα χαρακτηριστικά heating και CO_2 .

5.3.2 DBSCAN

5.3.2.1 Με βάση silhouette score με 2 χαρακτηριστικά

Στην εκπαίδευση του αλγορίθμου DBSCAN, δοκιμάστηκαν αρχικά διαδικασίες κωδικοποίησης και κλιμάκωσης των δεδομένων για να εξασφαλιστεί η σωστή προετοιμασία τους. Επιπλέον, οι ακραίες τιμές αποτέλεσαν πρόκληση, καθώς μπορούσαν να επηρεάσουν τα αποτελέσματα του αλγορίθμου. Για την επιλογή των παραμέτρων του μοντέλου, χρησιμοποιήθηκαν αρχικά προκαθορισμένες τιμές και στη συνέχεια η μέθοδος knee point και η ανάλυση των πλησιέστερων γειτόνων (k-nearest neighbors, KNN) για την εύρεση των βέλτιστων τιμών. Δοκιμάστηκε επίσης διαφορετικός αριθμός χαρακτηριστικών, όπου εξετάστηκαν όλοι οι δυνατοί συνδυασμοί ζευγών χαρακτηριστικών και στη συνέχεια και τα έξι χαρακτηριστικά μαζί, για να διαπιστωθεί ο καλύτερος συνδυασμός που θα οδηγούσε σε πιο ακριβή αποτελέσματα. Τα χαρακτηριστικά που επιλέχθηκαν είναι τα 'Year of building (first year and years of renovations/additions where applicable)', 'Total heating area, m²', 'Energy consumption before project (kWh/m²)', 'CO₂ emission before project (tons)', 'Number of

floors', 'Temperature-Average' τα οποία μετονομάστηκαν σε age, heating, energy, co_2 , floors και temperature αντίστοιχα.

Αρχικά, πραγματοποιούνται πειράματα με παραμέτρους min_samples στην περιοχή 2 έως 15 και neighbors στην περιοχή 2 έως 10, προκειμένου να βρεθεί το βέλτιστο silhouette score χρησιμοποιώντας ζεύγη χαρακτηριστικών. Μετά από τις δοκιμές, παρατηρείται ότι τα αποτελέσματα δεν είναι επιθυμητά. Παρακάτω, παρουσιάζεται μία δοκιμή του αλγορίθμου με βάση αυτή τη μέθοδο ως παράδειγμα στην επόμενη παράγραφο. Χρησιμοποιείται η συνάρτηση DBSCAN από τη βιβλιοθήκη scikit learn της python [49].

Για τα χαρακτηριστικά heating και age, το καλύτερο silhouette score είναι 0.8954683156100914, το οποίο επιτυγχάνεται με τις τιμές Best neigh: 5, min_samples: 14 και eps: 1528.802943482252. Όπως φαίνεται και στο παρακάτω διάγραμμα ο αλγόριθμος έχει συμπεριλάβει όλα τα σημεία σε μία ομάδα και έχει εντοπίσει κάποια ως θόρυβο.

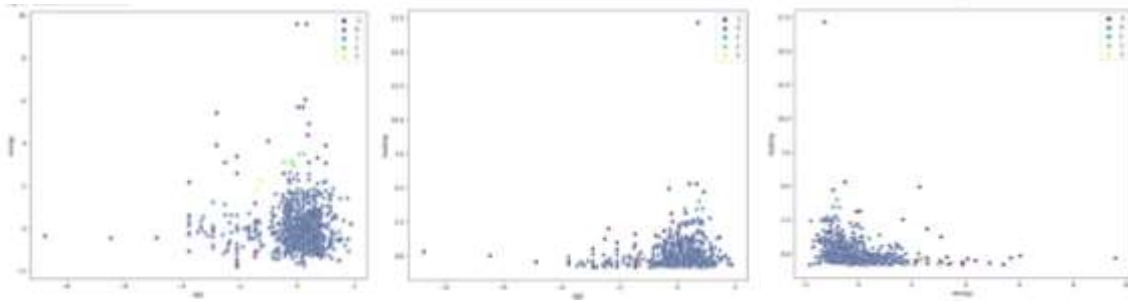


Σχήμα 27: Γραμμικό διάγραμμα age-heating μετά την ομαδοποίηση με DBSCAN με είσοδο τα χαρακτηριστικά age-heating, min_smaples=14 και eps=1528.802943482252

5.3.2.2 Χωρίς silhouette score με 3 χαρακτηριστικά

Η εκτέλεση του αλγορίθμου DBSCAN με την επιλογή παραμέτρων βάσει του silhouette score δεν απέδωσε ικανοποιητικά αποτελέσματα. Για αυτόν τον λόγο, τα πειράματα συνεχίζονται δοκιμάζοντας τις παραμέτρους min_samples = 7 και τιμές eps από 0.6 έως 0.1. Ο αριθμός των χαρακτηριστικών που δίνεται ως είσοδος στον αλγόριθμο είναι τρία τη φορά και τα αποτελέσματα οπτικοποιούνται σε γραμμικά διαγράμματα. Ωστόσο και σε αυτήν την περίπτωση τα αποτελέσματα δεν ήταν ικανοποιητικά. Στην επόμενη παράγραφο δίνεται ένα από τα πειράματα που πραγματοποιήθηκε.

Αρχικά, χρησιμοποιήθηκε η τεχνική ομαδοποίησης DBSCAN για την ανάλυση των δεδομένων με χαρακτηριστικά age, energy και heating. Με παραμέτρους min_samples: 7 και eps: 0.6, τα δεδομένα χωρίστηκαν σε πέντε ομάδες, όπου οι συχνότητες των ομάδων ήταν ({0: 2082, -1: 87, 1: 14, 2: 11, 3: 8}). Η ομάδα 0 περιλαμβάνει 2082 σημεία, υποδεικνύοντας μια μεγάλη και πυκνή περιοχή δεδομένων. Η ομάδα -1 αναφέρεται στα 87 σημεία που θεωρήθηκαν θόρυβος και δεν εντάχθηκαν σε καμία ομάδα, ενώ οι ομάδες 1, 2 και 3 είναι μικρότερες, με 14, 11 και 8 σημεία αντίστοιχα, υποδεικνύοντας μικρότερες αλλά σημαντικές συσπειρώσεις στα δεδομένα. Τα διαγράμματα που προέκυψαν από την ομαδοποίηση είναι τα εξής:



Σχήμα 28: Γραμμικά διάγραμμα μετά την ομαδοποίηση με DBSCAN με είσοδο τα χαρακτηριστικά age-energy-heating, min_samples=7 και eps=0.6

Είναι εμφανές ότι ο αλγόριθμος έχει συμπεριλάβει σχεδόν όλα τα σημεία σε μία ομάδα, κάτι το οποίο δεν είναι επιθυμητό.

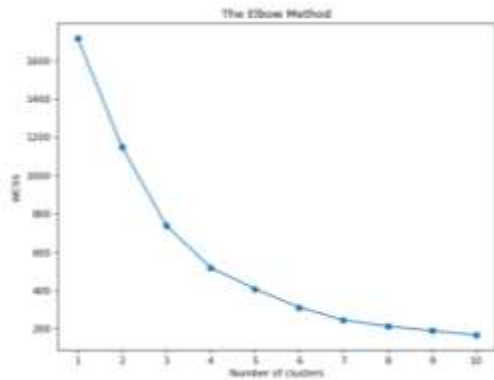
5.3.3 K-Means

Στην ανάλυση χρησιμοποιείται η μέθοδος KMeans για την ομαδοποίηση των δεδομένων. Χρησιμοποιείται η συνάρτηση KMeans από τη βιβλιοθήκη scikit learn της python [50]. Για την προεπεξεργασία των δεδομένων, εφαρμόζεται One-Hot-Encoding για τις κατηγορικές μεταβλητές και Standard Scaler για την κλιμάκωση των αριθμητικών δεδομένων. Δοκιμάστηκε διαφορετικός αριθμός ομάδων (k) και χρησιμοποιήθηκε η μέθοδος Elbow για την εύρεση του βέλτιστου αριθμού ομάδων. Επιπλέον, πραγματοποιήθηκαν πειράματα με διαφορετικά σύνολα χαρακτηριστικών, χρησιμοποιώντας και τα 6 χαρακτηριστικά μαζί, καθώς και όλα τα δυνατά ζεύγη αυτών. Τα αποτελέσματα αξιολογήθηκαν χρησιμοποιώντας μετρικές όπως το Silhouette Score. Οι οπτικοποιήσεις περιλαμβάνουν γραφήματα Elbow για την επιλογή του βέλτιστου k και διαγράμματα διασποράς των ομάδων.

5.3.3.1 Με 2 χαρακτηριστικά

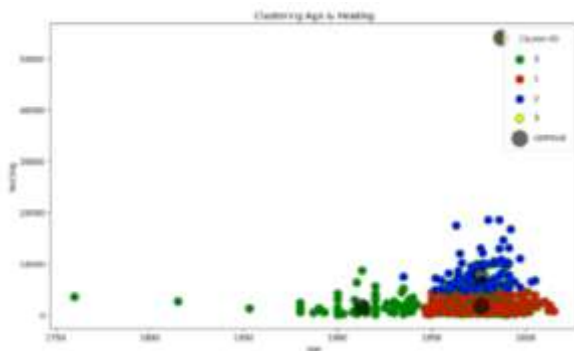
Η διαδικασία της ομαδοποίησης με τον αλγόριθμο KMeans θα πραγματοποιηθεί χρησιμοποιώντας τα χαρακτηριστικά age, heating, energy και co_2 . Αρχικά, ο αλγόριθμος εκπαιδεύεται χρησιμοποιώντας δύο χαρακτηριστικά τη φορά.

Για τα χαρακτηριστικά age και heating εφαρμόζεται η μέθοδος elbow. Το βέλτιστο σημείο εντοπίζεται εκεί όπου το διάγραμμα παρουσιάζει μία απότομη κάμψη. Για την εύρεση της κάμψης, δηλαδή του k, χρησιμοποιείται το kneed locator που επιβεβαιώνει τις παρατηρήσεις.



Σχήμα 29: Διάγραμμα Elbow για τα χαρακτηριστικά age-heating

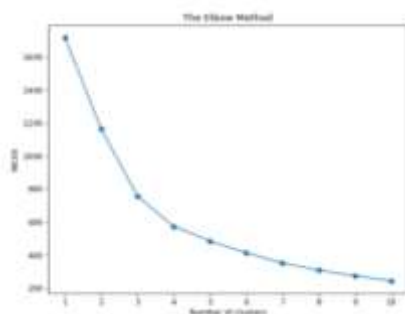
Στην αυτήν την περίπτωση, το βέλτιστο k εντοπίζεται για $k=4$ όπως φαίνεται και στο παραπάνω διάγραμμα. Στη συνέχεια, ο αλγόριθμος K-Means εφαρμόζεται με είσοδο $k=4$ και τα αποτελέσματα οπτικοποιούνται παρακάτω.



Σχήμα 30: Γραμμικό διάγραμμα μετά την ομαδοποίηση με KMeans με είσοδο τα χαρακτηριστικά age-heating και $k=4$

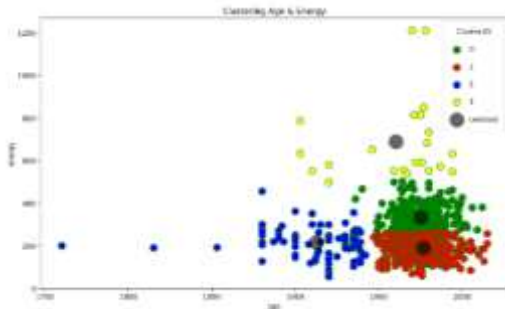
Παρατηρείται ότι ο αλγόριθμος λαμβάνει υπόψιν και τα δύο χαρακτηριστικά για τον διαχωρισμό των σημείων σε ομάδες.

Η ίδια τεχνική ακολουθείται και για τα υπόλοιπα ζεύγη χαρακτηριστικών. Με τα χαρακτηριστικά age και energy, προκύπτει το διάγραμμα Elbow με κάμψη για $k = 4$.



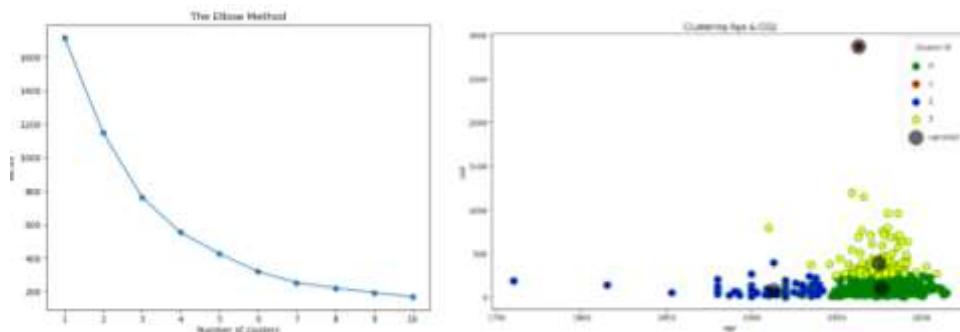
Σχήμα 31: Διάγραμμα Elbow για τα χαρακτηριστικά age-energy

Από το σχήμα γίνεται αντιληπτό ότι τα δεδομένα έχουν χωριστεί σε τέσσερις διαφορετικές ομάδες. Οι ομάδες αναπαρίστανται με διαφορετικά χρώματα: μπλε (Cluster 0), κόκκινο (Cluster 1), πράσινο (Cluster 2) και κίτρινο (Cluster 3), ενώ τα κεντρικά σημεία των clusters σημειώνονται με γκρι κύκλους. Παρατηρούμε ότι τα δεδομένα του Cluster 0 είναι κυρίως πριν το 1950 με χαμηλές τιμές ενέργειας, ενώ τα Clusters 1, 2, και 3 περιλαμβάνουν δεδομένα μετά το 1950 με το Cluster 1 να έχει χαμηλές τιμές ενέργειας, το Cluster 2 υψηλότερες και το Cluster 3 τις υψηλότερες τιμές ενέργειας. Τα κεντρικά σημεία δείχνουν τη μέση θέση των δεδομένων κάθε cluster, παρέχοντας μια οπτική αναπαράσταση της κατανομής και των χαρακτηριστικών των διαφορετικών clusters σε σχέση με το έτος κατασκευής (age) και την ενέργεια (energy).



Σχήμα 32: Γραμμικό διάγραμμα μετά την ομαδοποίηση με KMeans με είσοδο τα χαρακτηριστικά age-energy και k=4

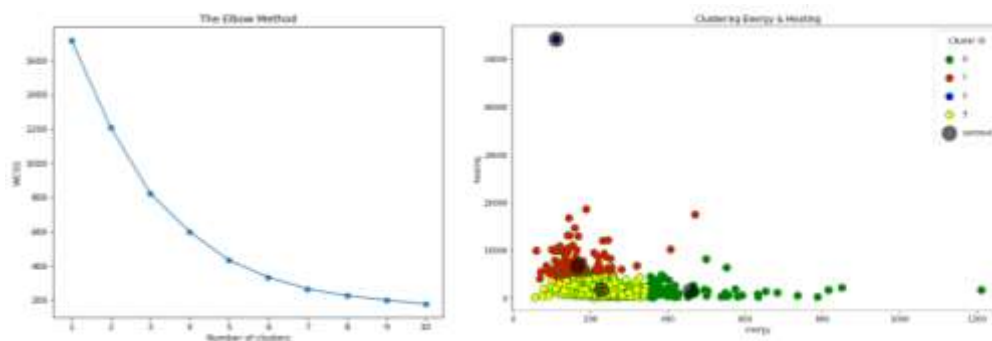
Για την εκπαίδευση με τα χαρακτηριστικά age και co_2 προκύπτει το διάγραμμα Elbow με κάμψη για k= 4 και την παρακάτω ομαδοποίηση:



Σχήμα 33: Διάγραμμα Elbow για τα χαρακτηριστικά age- co_2 και Γραμμικό διάγραμμα μετά την ομαδοποίηση με KMeans με είσοδο τα χαρακτηριστικά age- co_2 και k=4

Και σε αυτήν την περίπτωση τα δεδομένα έχουν χωριστεί σε τέσσερις διαφορετικές ομάδες. Τα δεδομένα του Cluster 0 παραμένουν κυρίως πριν το 1950 με χαμηλές εκπομπές CO_2 , ενώ τα Clusters 1, 2, και 3 περιλαμβάνουν δεδομένα μετά το 1950, με το Cluster 1 να έχει σχετικά χαμηλές εκπομπές CO_2 , το Cluster 2 υψηλότερες και το Cluster 3 τις υψηλότερες εκπομπές CO_2 . Η ανάλυση αυτή δείχνει παρόμοια τάση με το προηγούμενο διάγραμμα, υποδηλώνοντας ότι τα νεότερα δεδομένα, μετά το 1950, παρουσιάζουν υψηλότερες τιμές τόσο στην ενέργεια όσο και στις εκπομπές CO_2 , κάτι που μπορεί να υποδεικνύει αυξημένη κατανάλωση ενέργειας και περιβαλλοντική επιβάρυνση στις πιο πρόσφατες χρονικές περιόδους.

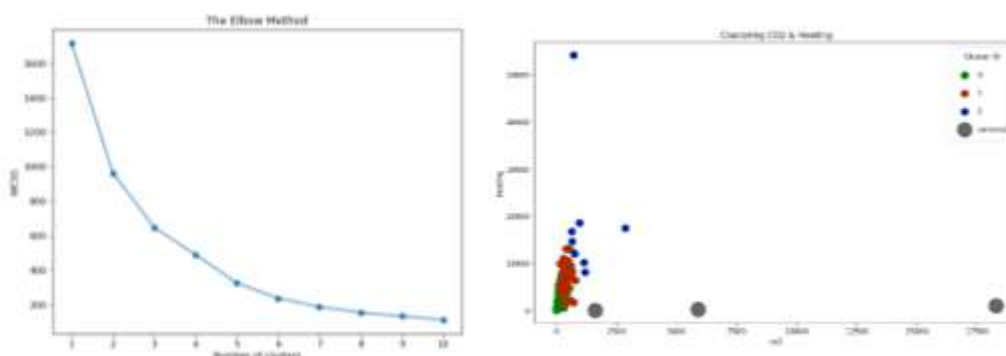
Η ίδια διαδικασία εφαρμόζεται και για το ζεύγος χαρακτηριστικών energy και heating.



Σχήμα 34: Διάγραμμα Elbow για τα χαρακτηριστικά *energy-heating* και Γραμμικό διάγραμμα μετά την ομαδοποίηση με KMeans με είσοδο τα χαρακτηριστικά *energy-heating* και $k=4$

Όπως στις προηγούμενες περιπτώσεις τα δεδομένα έχουν ταξινομηθεί σε τέσσερις ομάδες. Στο συγκεκριμένο διάγραμμα τα δεδομένα του Cluster 0 συγκεντρώνονται σε χαμηλές τιμές ενέργειας και θέρμανσης, ενώ τα Clusters 1, 2, και 3 παρουσιάζουν αυξημένες τιμές ενέργειας με διαφοροποιήσεις στις τιμές θέρμανσης. Το Cluster 1 περιλαμβάνει δεδομένα με χαμηλότερες τιμές θέρμανσης, το Cluster 2 με μέτριες τιμές, ενώ το Cluster 3 έχει τις υψηλότερες τιμές θέρμανσης. Η ανάλυση αυτή υπογραμμίζει την αυξημένη κατανάλωση ενέργειας και θέρμανσης σε συγκεκριμένα clusters, παρόμοια με την τάση που παρατηρήθηκε στα προηγούμενα διαγράμματα με την ηλικία και τις εκπομπές CO_2 , καταδεικνύοντας τη γενική αύξηση στην κατανάλωση πόρων και περιβαλλοντική επιβάρυνση στις πιο πρόσφατες χρονικές περιόδους. Αυτή η συνέπεια μεταξύ των διαγραμμάτων ενισχύει την υπόθεση ότι οι νεότερες χρονικές περίοδοι χαρακτηρίζονται από αυξημένη κατανάλωση ενέργειας και περιβαλλοντικούς δείκτες.

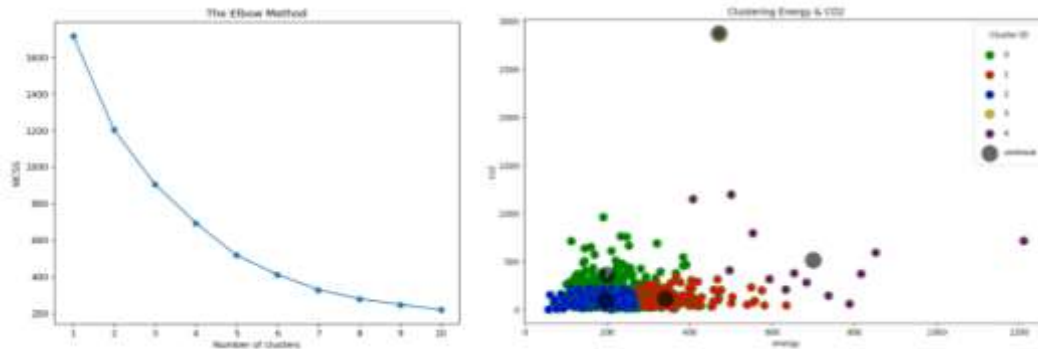
Για τα χαρακτηριστικά Heating και co_2 προκύπτουν τα διαγράμματα:



Σχήμα 35: Διάγραμμα Elbow για τα χαρακτηριστικά *heating- co_2* και Γραμμικό διάγραμμα μετά την ομαδοποίηση με KMeans με είσοδο τα χαρακτηριστικά *heating- co_2* και $k=3$

Σε αυτήν την περίπτωση τα δεδομένα χωρίζονται σε τρία clusters. Τα δεδομένα του Cluster 0 συγκεντρώνονται σε χαμηλές τιμές εκπομπών CO_2 και θέρμανσης, ενώ το Cluster 1 περιλαμβάνει δεδομένα με μέτριες τιμές εκπομπών CO_2 και θέρμανσης. Το Cluster 2, αντίθετα, έχει δεδομένα με πολύ υψηλές τιμές εκπομπών CO_2 και διαφοροποιημένες τιμές θέρμανσης. Αυτή η ανάλυση επιβεβαιώνει τη συνεχή τάση που παρατηρήθηκε στα προηγούμενα διαγράμματα, όπου οι νεότερες χρονικές περίοδοι και τα υψηλότερα επίπεδα κατανάλωσης ενέργειας συσχετίζονται με αυξημένες εκπομπές CO_2 και μεγαλύτερη περιβαλλοντική επιβάρυνση. Η σύγκριση αυτών των διαγραμμάτων αναδεικνύει τη συστηματική αύξηση στη χρήση πόρων και τις περιβαλλοντικές επιπτώσεις, ενισχύοντας την κατανόηση της σχέσης μεταξύ ηλικίας, ενέργειας, εκπομπών CO_2 και θέρμανσης.

Εκπαιδεύοντας το μοντέλο με τα χαρακτηριστικά co_2 και energy προκύπτει:



Σχήμα 36: Διάγραμμα Elbow για τα χαρακτηριστικά energy- co_2 και Γραμμικό διάγραμμα μετά την ομαδοποίηση με KMeans με είσοδο τα χαρακτηριστικά energy- co_2 και $k=5$

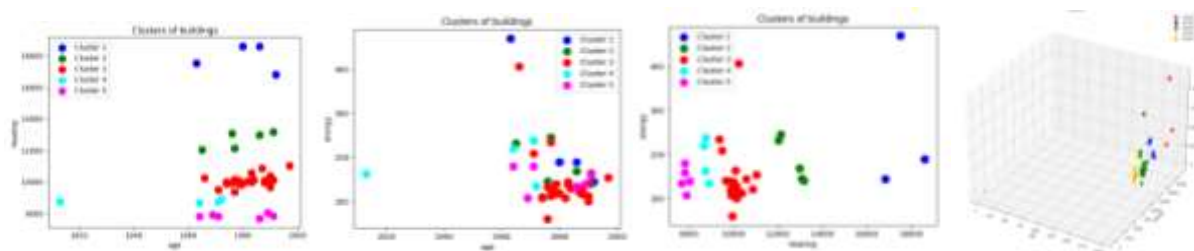
Η μεγαλύτερη κάμψη στο διάγραμμα Elbow εντοπίζεται για $k=5$, οπότε τρέχουμε τον KMeans δίνοντας του ως είσοδο $k=5$. Το Cluster 0 (πράσινο) και το Cluster 1 (μπλε) περιλαμβάνουν δεδομένα με χαμηλές τιμές ενέργειας και CO_2 , με το πράσινο να είναι πιο συγκεντρωμένο. Το Cluster 2 (κόκκινο) περιέχει δεδομένα με μέτριες τιμές, ενώ το Cluster 3 (μωβ) έχει μεγαλύτερη ποικιλία στις τιμές ενέργειας και CO_2 , υποδεικνύοντας πιθανές διαφορές στις κατηγορίες των δεδομένων. Το Cluster 4 (κίτρινο) περιλαμβάνει λίγα σημεία με πολύ υψηλές τιμές, πιθανώς outliers. Τα κεντροειδή των clusters δείχνουν τις μέσες τιμές των δεδομένων κάθε ομάδας, προσφέροντας μια γενική εικόνα των χαρακτηριστικών τους.

5.3.4 HDBSCAN

5.3.4.1 Με 3 χαρακτηριστικά

Σε πειράματα που πραγματοποιήθηκαν χρησιμοποιώντας τον αλγόριθμο ομαδοποίησης HDBSCAN, επιλέχθηκαν τα χαρακτηριστικά age, energy, heating, co_2 , floors και temperature τα οποία θα δόθηκαν ανά τρία ως είσοδο στον αλγόριθμο. Για τις ανάγκες των πειραμάτων αυτών, ορίστηκε η ελάχιστη επιτρεπτή τιμή μεγέθους ομάδας (min_cluster_size) σε 10, καθώς με μεγαλύτερες τιμές παρατηρήθηκε μεγάλη επικάλυψη των ομάδων μεταξύ τους. Οι αναλύσεις των αποτελεσμάτων παρουσιάστηκαν μέσω διαγραμμάτων που απεικονίζουν τη δομή των σχηματισμένων ομάδων και τη διασπορά των δεδομένων, προσφέροντας μια λεπτομερή εικόνα της κατανομής και της σχέσης μεταξύ των δεδομένων που εξετάστηκαν. Ειδικότερα, ο αλγόριθμος HDBSCAN εφαρμόζεται με τρία χαρακτηριστικά τη φορά και τα χαρακτηριστικά απεικονίζονται ανά ζεύγη μετά την ομαδοποίηση για την καλύτερη κατανόηση της συσχέτισης των δεδομένων μεταξύ τους.

Για τα χαρακτηριστικά age-heating-energy προκύπτουν τα παρακάτω διαγράμματα:



Σχήμα 37: Γραμμικά διαγράμματα χαρακτηριστικών μετά την ομαδοποίηση με HDBSCAN με είσοδο age-heating-energy και 3D διάγραμμα με την απεικόνιση και των τριών χαρακτηριστικών

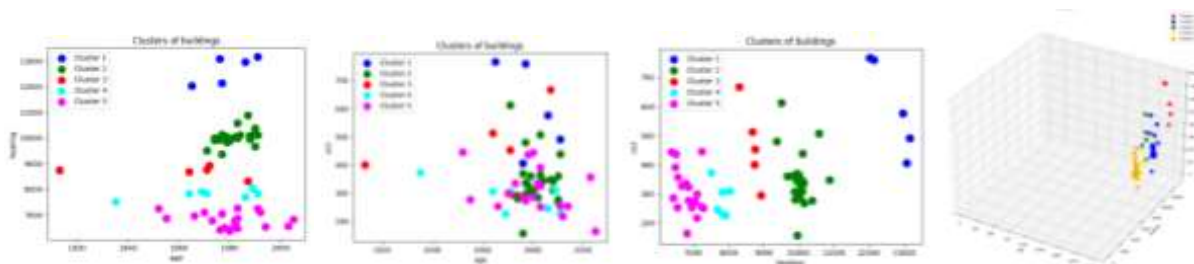
Από το παραπάνω σχήμα είναι διακριτό ότι από την ομαδοποίηση προκύπτουν πέντε διακριτές ομάδες κτηρίων με συγκεκριμένες κατανομές και συσχετίσεις. Στην πρώτη εικόνα, η ομάδα 1 (μπλε) περιλαμβάνει κτήρια με υψηλότερες απαιτήσεις θέρμανσης, ενώ οι ομάδες 2 (πράσινη) και 3 (κόκκινη) αποτελούνται από κτήρια με χαμηλότερη θέρμανση. Οι ομάδες 4 (κυανή) και 5 (κυανή) δείχνουν κτήρια με ακόμη χαμηλότερες απαιτήσεις θέρμανσης.

Στη δεύτερη εικόνα, η πρώτη ομάδα (μπλε) περιλαμβάνει ένα κτήριο με υψηλές απαιτήσεις ενέργειας. Η ομάδα δύο (πράσινη) περιλαμβάνει κτίρια διάφορων ετών με μέτρια κατανάλωση ενέργειας, ενώ η ομάδα τρία (κόκκινη) περιλαμβάνει κτίρια με διάφορες απαιτήσεις ενέργειας διάφορων ετών. Η ομάδα τέσσερα (κυανή) περιλαμβάνει πιο παλιά κτίρια με μέτρια κατανάλωση ενέργειας. Η ομάδα πέντε (φούξια) περιλαμβάνει κτίρια με χαμηλές απαιτήσεις ενέργειας διάφορων χρονολογιών.

Στην τρίτη εικόνα, η ομαδοποίηση έχει πραγματοποιηθεί λαμβάνοντας υπόψιν το χαρακτηριστικό heating, όπως και στην εικόνα 1.

Συνολικά, η ανάλυση δείχνει ότι ο αλγόριθμος λαμβάνει υπόψιν μόνο το χαρακτηριστικό heating για τον διαχωρισμό των δεδομένων σε ομάδες.

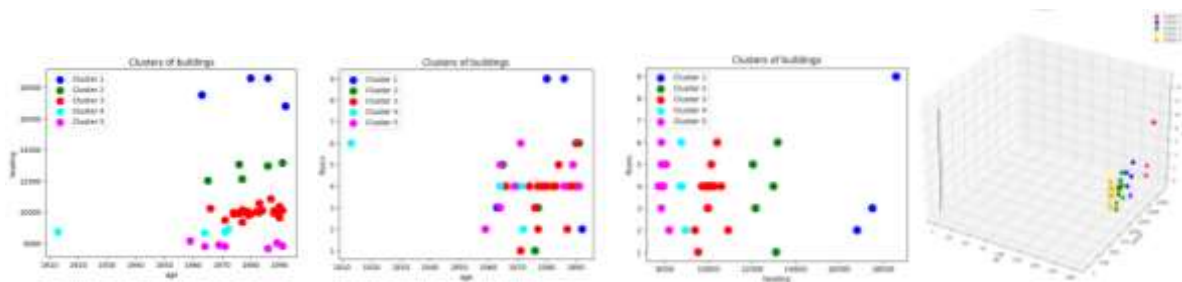
Για τα χαρακτηριστικά age-heating-co₂ προκύπτουν τα παρακάτω διαγράμματα:



Σχήμα 38: Γραμμικά διαγράμματα χαρακτηριστικών μετά την ομαδοποίηση με HDBSCAN με είσοδο age-heating-co₂ και 3D διάγραμμα με την απεικόνιση και των τριών χαρακτηριστικών

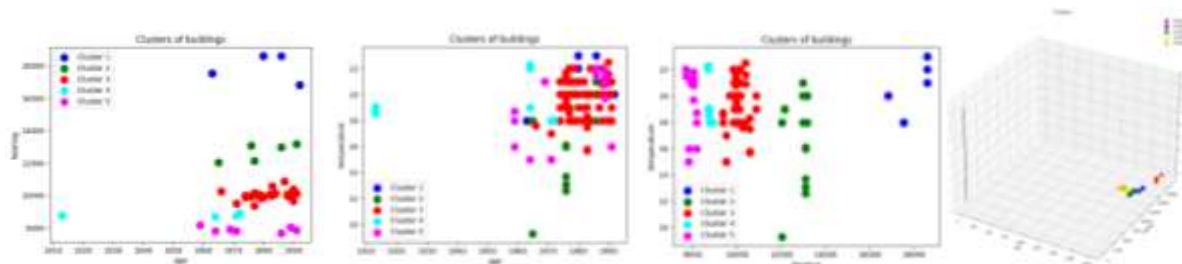
Από τα διαγράμματα μπορεί να παρατηρήσει κανείς ότι τα δεδομένα χωρίστηκαν σε πέντε διακριτές ομάδες. Ο αλγόριθμος έλαβε υπόψιν το χαρακτηριστικό heating, όπως φαίνεται και στα διαγράμματα age-heating και heating-co₂ αλλά και στο 3D διάγραμμα. Στο δεύτερο διάγραμμα (age-co₂) παρατηρούμε ότι οι ομάδες δεν είναι διακριτές άρα πιθανότατα ο αλγόριθμος να μην τα έλαβε υπόψιν κατά την ομαδοποίηση.

Για τα χαρακτηριστικά age-heating-floors οι ομάδες που προέκυψαν είναι πέντε και ο αλγόριθμος έλαβε υπόψιν και σε αυτήν την περίπτωση το χαρακτηριστικό heating.



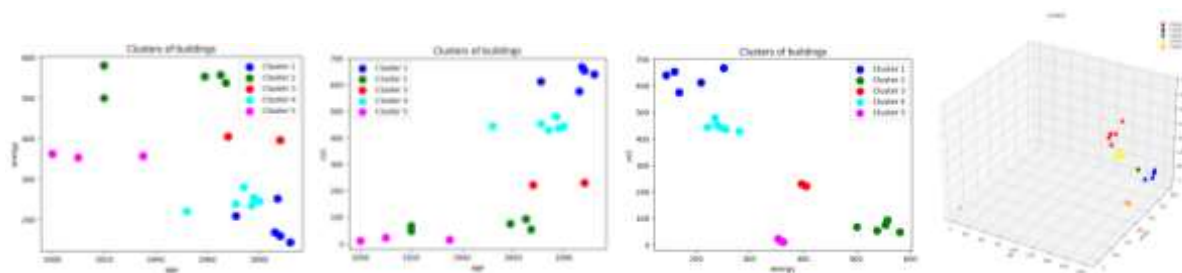
Σχήμα 39: Γραμμικά διαγράμματα χαρακτηριστικών μετά την ομαδοποίηση με HDBSCAN με είσοδο age-heating-floors και 3D διάγραμμα με την απεικόνιση και των τριών χαρακτηριστικών

Ομοίως, με τα χαρακτηριστικά age-heating-temperature:



Σχήμα 40: Γραμμικά διαγράμματα χαρακτηριστικών μετά την ομαδοποίηση με HDBSCAN με είσοδο age-heating-temperature και 3D διάγραμμα με την απεικόνιση και των τριών χαρακτηριστικών

Για τα χαρακτηριστικά age-energy-co2 έχουμε τα εξής διαγράμματα:

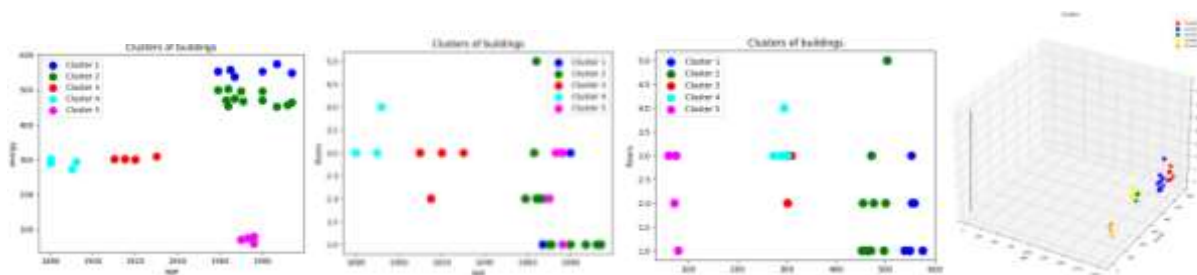


Σχήμα 41: Γραμμικά διαγράμματα χαρακτηριστικών μετά την ομαδοποίηση με HDBSCAN με είσοδο age-energy-co₂ και 3D διάγραμμα με την απεικόνιση και των τριών χαρακτηριστικών

Στην πρώτη εικόνα, τα δεδομένα έχουν χωριστεί σε πέντε ομάδες. Στην πρώτη ομάδα(μπλε) ανήκουν πιο καινούρια κτίρια με χαμηλή κατανάλωση ενέργειας. Στη δεύτερη ομάδα(πράσινο) ανήκουν κτίρια με υψηλή κατανάλωση ενέργειας τα οποία είναι αρκετά παλιά αλλά και λιγότερο σύγχρονα. Στην τρίτη(κόκκινο) και στην πέμπτη(φούξια) ανήκουν κτίρια με μεσαία κατανάλωση ενέργειας που είναι πιο σύγχρονα και πιο παλιά αντίστοιχα. Και στην τέταρτη ομάδα(κυανό) ανήκουν κτίρια με χαμηλή κατανάλωση ενέργειας.

Στη δεύτερη εικόνα παρατηρούμε ότι τα δεδομένα έχουν χωριστεί με βάση τις εκπομπές CO_2 και στην τρίτη εικόνα, βλέπουμε ότι τα δεδομένα έχουν χωριστεί με βάση την κατανάλωση ενέργειας.

Για τα χαρακτηριστικά age-energy-floors:

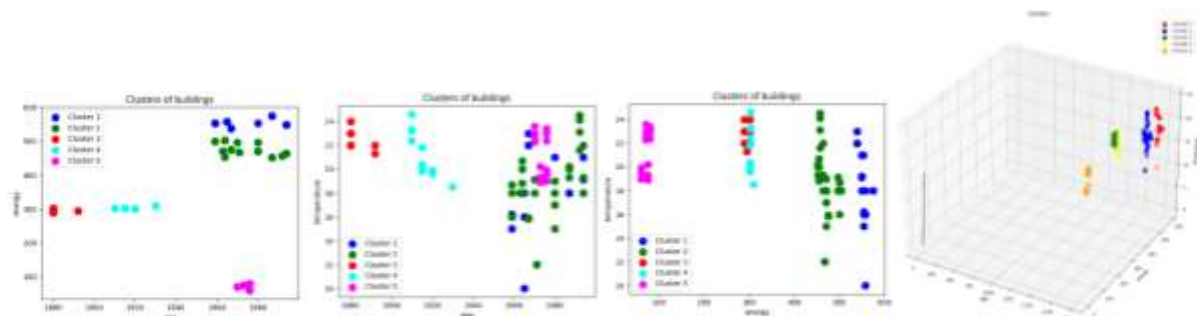


Σχήμα 42: Γραμμικά διαγράμματα χαρακτηριστικών μετά την ομαδοποίηση με HDBSCAN με είσοδο age-energy-floors και 3D διάγραμμα με την απεικόνιση και των τριών χαρακτηριστικών

Στο πρώτο διάγραμμα ο αλγόριθμος λαμβάνει υπόψιν και την κατανάλωση ενέργειας αλλά και το έτος κατασκευής για την ομαδοποίηση. Στην πρώτη(μπλε) και στην δεύτερη(πράσινη) ομάδα ανήκουν πιο σύγχρονα κτίρια με υψηλή κατανάλωση ενέργειας ενώ στην τρίτη(κόκκινη) και τέταρτη(κυανό) ανήκουν κτίρια με μέτρια κατανάλωση ενέργειας που χωρίζονται σε παλιά και πιο παλιά αντίστοιχα. Και στην τελευταία ομάδα(φούξια) ανήκουν πιο σύγχρονα κτίρια με χαμηλή κατανάλωση ενέργειας.

Στο δεύτερο διάγραμμα ο αλγόριθμος λαμβάνει υπόψιν κυρίως το έτος κατασκευής για την ομαδοποίηση, ενώ στο τρίτο την κατανάλωση ενέργειας μην λαμβάνοντας ιδιαίτερα υπόψιν τον όροφο του κτιρίου.

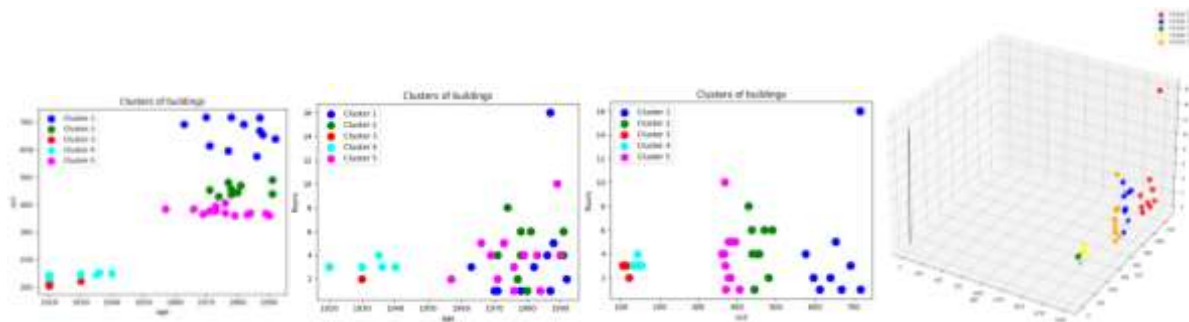
Για τα χαρακτηριστικά age-energy-temperature:



Σχήμα 43: Γραμμικά διαγράμματα χαρακτηριστικών μετά την ομαδοποίηση με HDBSCAN με είσοδο age-energy-temperature και 3D διάγραμμα με την απεικόνιση και των τριών χαρακτηριστικών

Στο πρώτο διάγραμμα η ομαδοποίηση γίνεται με βάση και τα δύο χαρακτηριστικά. Συγκεκριμένα, στην τρίτη(κόκκινο) και τέταρτη(κυανό) ομάδα ανήκουν κτίρια με μεσαία κατανάλωση ενέργειας τα οποία ανήκουν στο έτος κατασκευής 1880-1900 και 1910-1940 αντίστοιχα. Στην υπόλοιπες ομάδες ανήκουν κτίρια με έτος κατασκευής 1960 και έπειτα που χωρίζονται με βάση την κατανάλωση ενέργειας. Στο δεύτερο διάγραμμα οι ομάδες δεν είναι διακριτές ενώ στο τρίτο διάγραμμα η ομαδοποίηση πραγματοποιείται με βάση την κατανάλωση ενέργειας.

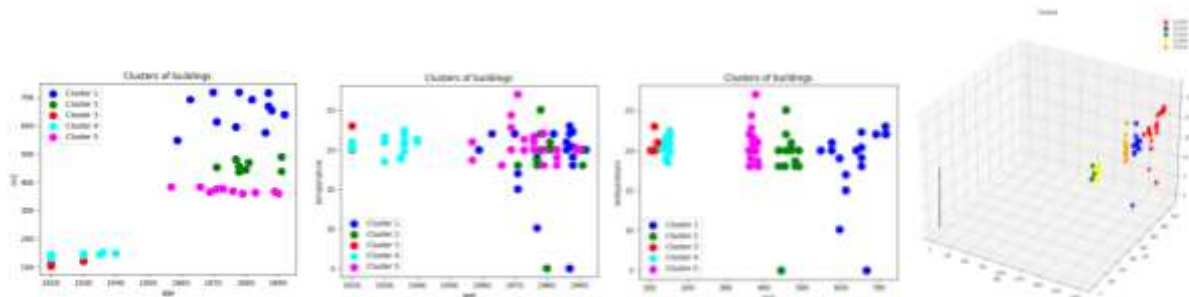
Για τα χαρακτηριστικά age-co2-floors:



Σχήμα 44: Γραμμικά διαγράμματα χαρακτηριστικών μετά την ομαδοποίηση με HDBSCAN με είσοδο age-co₂-floors και 3D διάγραμμα με την απεικόνιση και των τριών χαρακτηριστικών

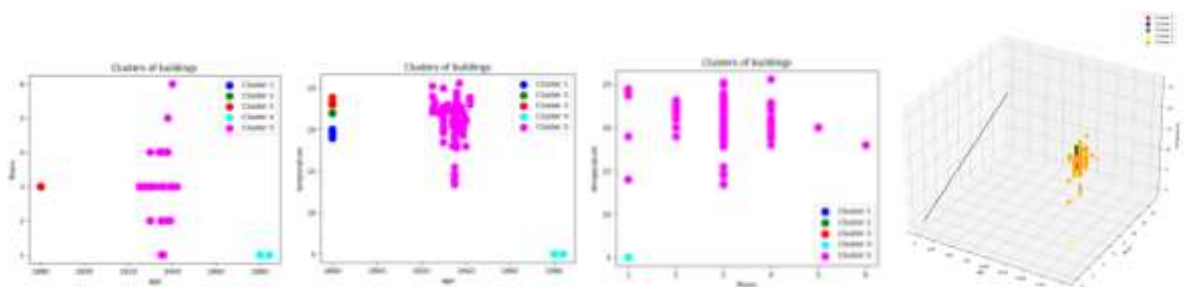
Με βάση τα παραπάνω διαγράμματα, παρατηρείται ότι η ομαδοποίηση έγινε κυρίως με βάση τις εκπομπές co₂.

Ομοίως για τα χαρακτηριστικά age-co₂-temperature, ο αλγόριθμος λαμβάνει υπόψη κυρίως τις εκπομπές co₂.



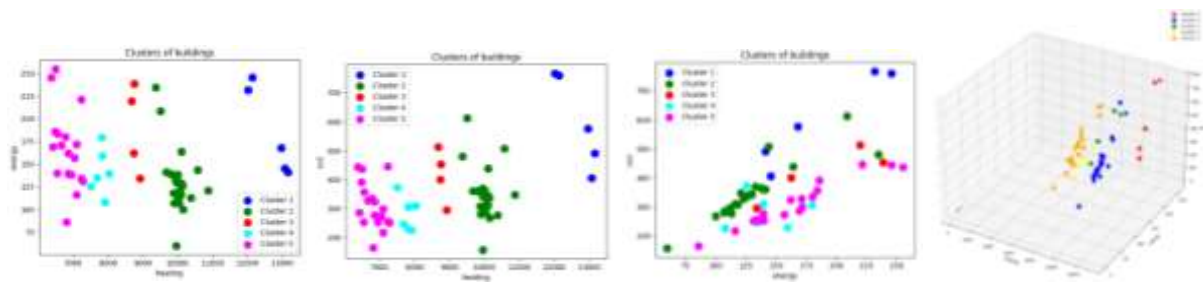
Σχήμα 45: Γραμμικά διαγράμματα χαρακτηριστικών μετά την ομαδοποίηση με HDBSCAN με είσοδο age-co₂-temperature και 3D διάγραμμα με την απεικόνιση και των τριών χαρακτηριστικών

Για τα χαρακτηριστικά age-floors-temperature δεν είναι όλες οι ομάδες ευδιάκριτες λόγω επικαλύψεων μεταξύ των ομάδων.



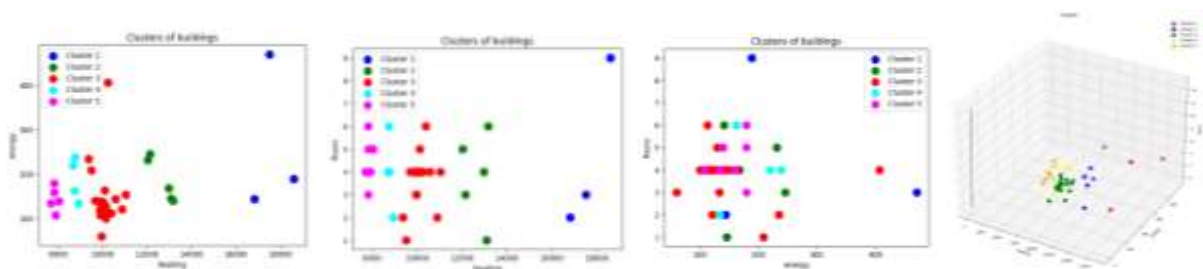
Σχήμα 46: Γραμμικά διαγράμματα χαρακτηριστικών μετά την ομαδοποίηση με HDBSCAN με είσοδο age-floors-temperature και 3D διάγραμμα με την απεικόνιση και των τριών χαρακτηριστικών

Για τα χαρακτηριστικά heating-energy-co₂, τα δεδομένα χωρίζονται με βάση τις απαιτήσεις σε θέρμανση όπως φαίνεται και στα διαγράμματα age-heating και co₂-heating.



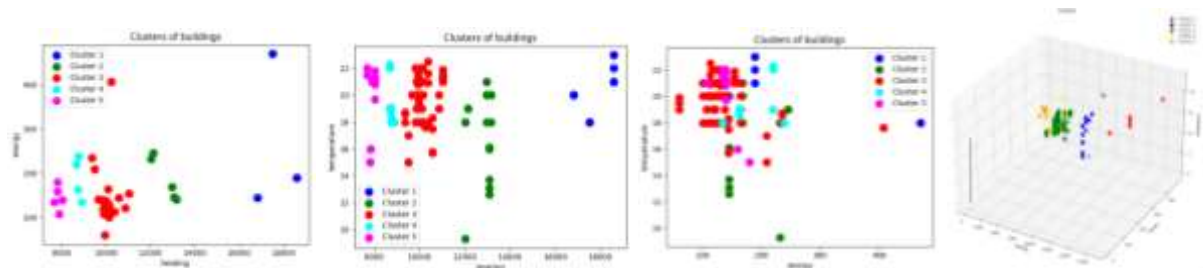
Σχήμα 47: Γραμμικά διαγράμματα χαρακτηριστικών μετά την ομαδοποίηση με HDBSCAN με είσοδο heating-energy-co₂ και 3D διάγραμμα με την απεικόνιση και των τριών χαρακτηριστικών

Για τα χαρακτηριστικά heating-energy-floors η ομαδοποίηση γίνεται με βάση το χαρακτηριστικό heating.



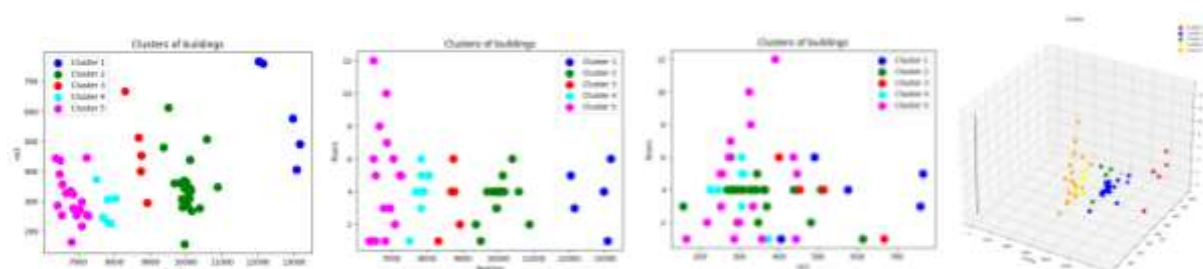
Σχήμα 48: Γραμμικά διαγράμματα χαρακτηριστικών μετά την ομαδοποίηση με HDBSCAN με είσοδο heating-energy-floors και 3D διάγραμμα με την απεικόνιση και των τριών χαρακτηριστικών

Το ίδιο ισχύει και στην περίπτωση που δίνεται ως είσοδος τα χαρακτηριστικά heating-energy-temperature:



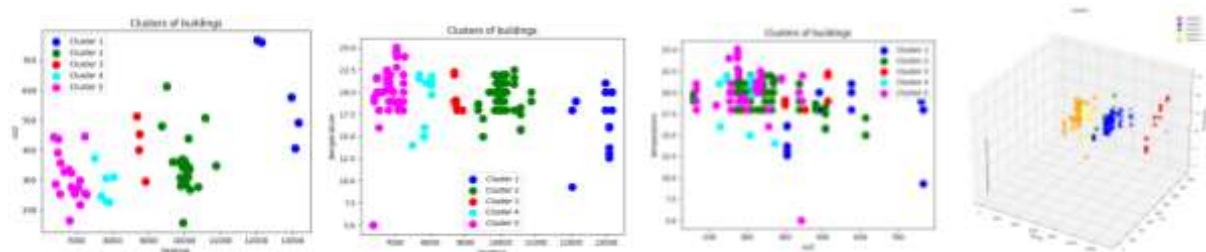
Σχήμα 49: Γραμμικά διαγράμματα χαρακτηριστικών μετά την ομαδοποίηση με HDBSCAN με είσοδο heating-energy-temperature και 3D διάγραμμα με την απεικόνιση και των τριών χαρακτηριστικών

Ομοίως, για τα χαρακτηριστικά heating-co₂-floors:



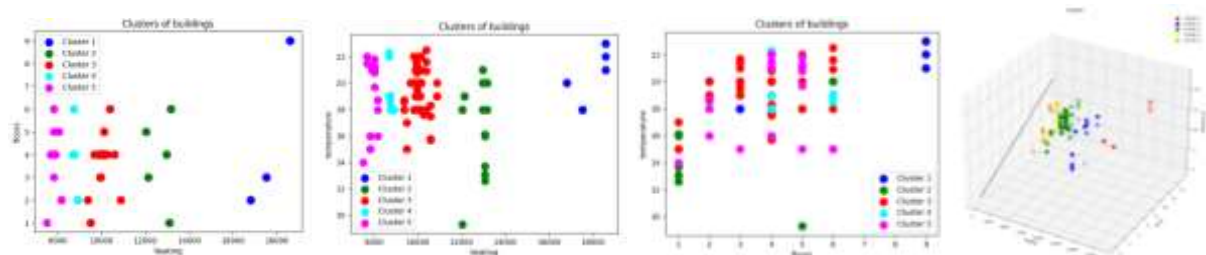
Σχήμα 50: Γραμμικά διαγράμματα χαρακτηριστικών μετά την ομαδοποίηση με HDBSCAN με είσοδο heating-co₂-floors και 3D διάγραμμα με την απεικόνιση και των τριών χαρακτηριστικών

Ομοίως για τα χαρακτηριστικά heating-co₂-temperature:



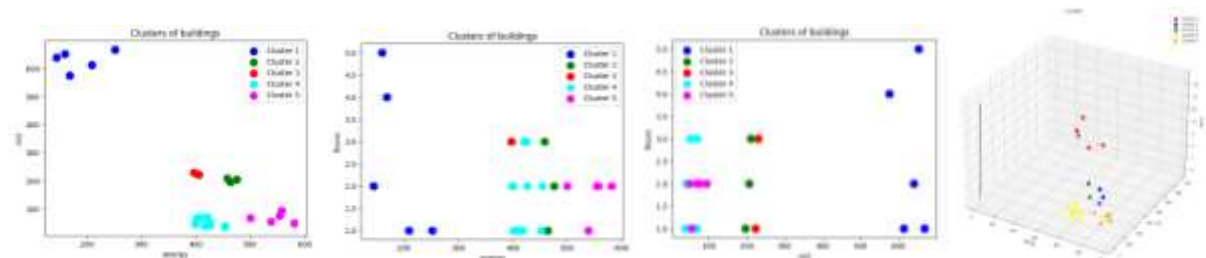
Σχήμα 51: Γραμμικά διαγράμματα χαρακτηριστικών μετά την ομαδοποίηση με HDBSCAN με είσοδο heating-co₂-temperature και 3D διάγραμμα με την απεικόνιση και των τριών χαρακτηριστικών

Ομοίως για τα χαρακτηριστικά heating-floors-temperature:



Σχήμα 52: Γραμμικά διαγράμματα χαρακτηριστικών μετά την ομαδοποίηση με HDBSCAN με είσοδο heating-floors-temperature και 3D διάγραμμα με την απεικόνιση και των τριών χαρακτηριστικών

Για τα χαρακτηριστικά energy-co₂-floors:

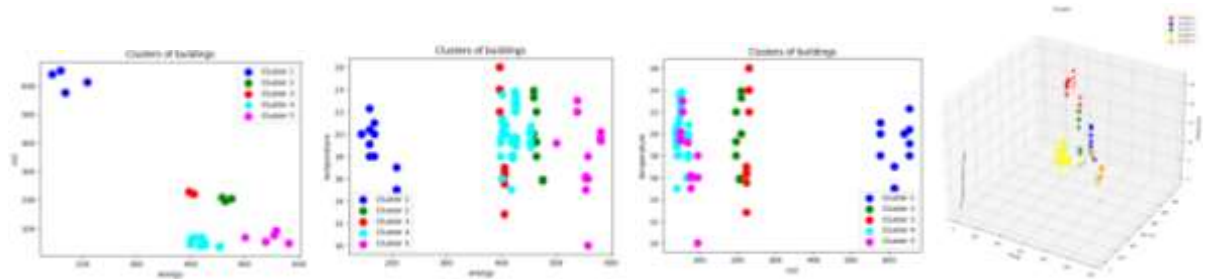


Σχήμα 53: Γραμμικά διαγράμματα χαρακτηριστικών μετά την ομαδοποίηση με HDBSCAN με είσοδο energy-co₂-floors και 3D διάγραμμα με την απεικόνιση και των τριών χαρακτηριστικών

Τα δεδομένα έχουν χωριστεί σε πέντε ομάδες. Στο πρώτο διάγραμμα βλέπουμε ότι στην πρώτη ομάδα(μπλε) ανήκουν κτίρια με υψηλές εκπομπές CO₂ ενώ στις υπόλοιπες με χαμηλότερες. Στη δεύτερη(πράσινη) και στην τρίτη(κόκκινη) ομάδα ανήκουν κτίρια με εκπομπές ρύπων 200-300 και διαχωρίζονται με βάση την κατανάλωση ενέργειας. Αντίστοιχα, στην τέταρτη(κυανό) και πέμπτη(φούξια) ομάδα ανήκουν κτίρια με χαμηλές εκπομπές ρύπων που διαχωρίζονται από την κατανάλωση ενέργειας. Στα δεύτερο και στο τρίτο διάγραμμα η ομαδοποίηση γίνεται λαμβάνοντας υπόψιν το χαρακτηριστικό energyκαι το CO₂ αντίστοιχα.

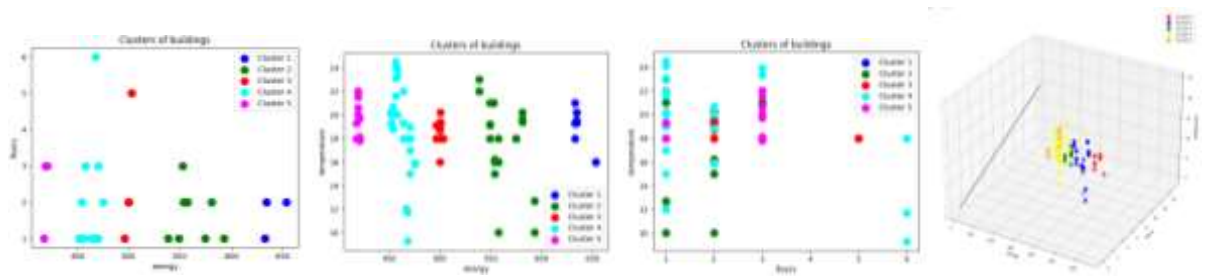
Επομένως ο αλγόριθμος λαμβάνει υπόψιν τα χαρακτηριστικά CO_2 και energy για τον διαχωρισμό των δεδομένων σε ομάδες.

Παρόμοια λογική με την προηγούμενη έχει και η ομαδοποίηση που προκύπτει από τα χαρακτηριστικά energy- co_2 -temperature:



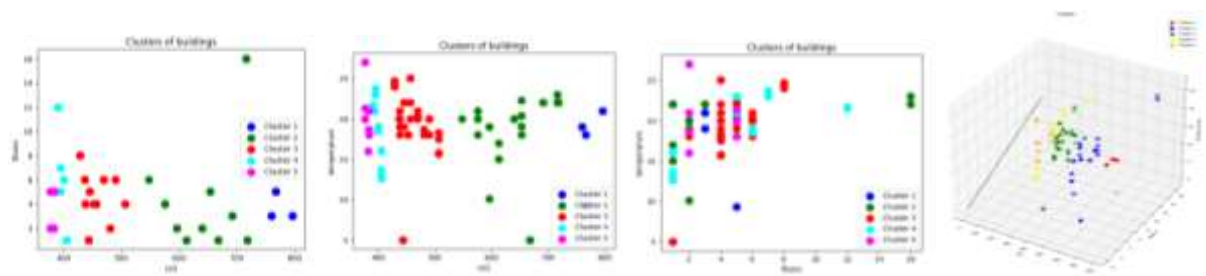
Σχήμα 54: Γραμμικά διαγράμματα χαρακτηριστικών μετά την ομαδοποίηση με HDBSCAN με είσοδο energy- co_2 -temperature και 3D διάγραμμα με την απεικόνιση και των τριών χαρακτηριστικών

Για τα χαρακτηριστικά energy-floors-temperature η ομαδοποίηση γίνεται με βάση το χαρακτηριστικό energy:



Σχήμα 55: Γραμμικά διαγράμματα χαρακτηριστικών μετά την ομαδοποίηση με HDBSCAN με είσοδο energy-floors-temperature και 3D διάγραμμα με την απεικόνιση και των τριών χαρακτηριστικών

Για τα χαρακτηριστικά co_2 -floors-temperature η ομαδοποίηση γίνεται με βάση το χαρακτηριστικό co_2 :



Σχήμα 56: Γραμμικά διαγράμματα χαρακτηριστικών μετά την ομαδοποίηση με HDBSCAN με είσοδο co_2 -floors-temperature και 3D διάγραμμα με την απεικόνιση και των τριών χαρακτηριστικών

ΚΕΦΑΛΑΙΟ 6

Συμπεράσματα και Προοπτικές

6 Συμπεράσματα και Προοπτικές

Στο πρώτο υποκεφάλαιο παρουσιάζονται και εξετάζονται τα διάφορα ευρήματα και συμπεράσματα που προέκυψαν από την εφαρμογή της μεθοδολογίας που περιγράφηκε στο Κεφάλαιο 5 και οι περιορισμοί που παρουσίαζε. Στη συνέχεια, στην επόμενη ενότητα συζητούνται πιθανές εναλλακτικές προσεγγίσεις για την εφαρμογή και την επέκταση αυτής της μεθοδολογίας.

6.1 Συμπεράσματα

Η παρούσα εργασία αναδεικνύει τη σημασία της χρήσης τεχνητής νοημοσύνης και μηχανικής μάθησης για την υποστήριξη των αποφασιζόντων στη λήψη βέλτιστων μέτρων ανακαίνισης κτιρίων με στόχο τη μείωση των εκπομπών ρύπων. Συγκεκριμένα, μέσω της ομαδοποίησης μπορεί να βρεθεί η ομάδα στην οποία ανήκει ένα κτίριο που χρήζει ανακαίνισης ώστε να μελετηθούν ανακαινίσεις από παρόμοια κτίρια που ανήκουν στην ίδια ομάδα. Οι ανακαινίσεις από κτίρια της ίδιας ομάδας συμβάλουν συμβουλευτικά στους αποφασίζοντες με αποτέλεσμα τη βελτίωση της λήψης αποφάσεων. Η ενίσχυση της ενεργειακής αποδοτικότητας των κτιρίων όχι μόνο συμβάλλει στην αντιμετώπιση της κλιματικής αλλαγής, αλλά προσφέρει και οικονομικά οφέλη, υπογραμμίζοντας την αναγκαιότητα προώθησης τέτοιων τεχνολογιών στην καθημερινή πρακτική.

Το σύνολο δεδομένων που εξετάστηκε περιλαμβάνει ένα πλήθος παρατηρήσεων που αφορούν έργα ανακαίνισης που ολοκληρώθηκαν στην Λετονία. Μετά την προ επεξεργασία και τη χρήση εργαλείων όπως πίνακας συσχέτισης, box plots και απεικόνιση των κατανομών των τιμών τους επιλέχτηκαν τα πιο ενδιαφέροντα χαρακτηριστικά δηλαδή το 'Year of building (first year and years of renovations/additions where applicable)', 'Total heating area, m²', 'Energy consumption before project (kWh/m²)', 'CO₂ emission before project (tons)', 'Number of floors' και 'Temperature-Average'. Επίσης, πραγματοποιήθηκε προεπεξεργασία στα δεδομένα ώστε να ανιχνευτούν ελλιπούς τιμές και κανονικοποιηθούν τα δεδομένα. Αυτά τα βήματα είναι απαραίτητα για τη βελτιστοποίηση της απόδοσης των μοντέλων και την εξασφάλιση ότι τα δεδομένα είναι κατάλληλα για ανάλυση.

Ακολούθησε η ανάλυση διάφορων αλγορίθμων ομαδοποίησης όπως οι k-means, DBSCAN, HDBSCAN, Spectral Clustering, Gaussian Mixture Models (GMM), Agglomerative Clustering, Mean Shift Clustering, Affinity Propagation, και BIRCH αποκαλύπτει διαφορετικά πλεονεκτήματα και μειονεκτήματα ανάλογα με τα χαρακτηριστικά των δεδομένων. Ο k-means είναι ταχύς και αποτελεσματικός για σφαιρικά clusters με ίσες διαστάσεις, αλλά είναι ευαίσθητος σε outliers. Αντίθετα, ο DBSCAN και ο HDBSCAN διαχειρίζονται καλύτερα τα μη σφαιρικά clusters και τον θόρυβο, αλλά μπορεί να είναι δύσκολο να βρεθούν οι κατάλληλες παράμετροι. Ο Spectral Clustering και ο GMM είναι χρήσιμοι για πιο περίπλοκες δομές δεδομένων αλλά απαιτούν περισσότερους υπολογιστικούς πόρους. Ωστόσο, λόγω των συγκεκριμένων δεδομένων που αναλύθηκαν, η επιλογή των αλγορίθμων περιορίζεται στους KMeans, DBSCAN, HDBSCAN, και Agglomerative Clustering, καθώς αυτοί οι αλγόριθμοι προσφέρουν την καλύτερη ισορροπία μεταξύ απόδοσης, ευκολίας υλοποίησης και διαχείρισης θορύβου και ανωμαλιών στα δεδομένα.

Ο αλγόριθμος KMeans σε συνδυασμό με τη μέθοδο του αγκώνα (elbow method) και την εφαρμογή κανονικοποίησης Standard Scaler, ομαδοποίησε τα δεδομένα σε διακριτές ομάδες όταν πραγματοποιήθηκαν πειράματα με είσοδο δύο χαρακτηριστικά. Ο αλγόριθμος ομαδοποιούσε τα δεδομένα λαμβάνοντας υπόψιν και τα δύο χαρακτηριστικά. Ωστόσο, με τα

την αύξηση της εισόδου των χαρακτηριστικών παρατηρήθηκε ο αλγόριθμος διαχώριζε τα δεδομένα σε ομάδες μόνο με βάση ένα χαρακτηριστικό ή τα διαχώριζε σε μη διακριτές ομάδες.

Η ανάλυση με HDBSCAN και Agglomerative Clustering, χρησιμοποιώντας τρία χαρακτηριστικά ως είσοδο τη φορά και την εφαρμογή κανονικοποίησης Standard Scaler, υπέδειξε ότι τα χαρακτηριστικά energy και co_2 είναι θεμελιώδη για την κατανόηση των δεδομένων, καθώς συχνά οδηγούν τις ομαδοποιήσεις ανεξαρτήτως της προσθήκης άλλων χαρακτηριστικών. Ωστόσο, παρατηρήθηκε ότι όταν το age συνδυάζεται με το χαρακτηριστικό energy, οι ομαδοποιήσεις καθορίζονται κυρίως από την ηλικία και την ενέργεια, δείχνοντας έναν ισχυρότερο συνδυασμό αυτών των δύο παραμέτρων. Επίσης όταν προστίθεται το χαρακτηριστικό heating, ο αλγόριθμος έλαβε υπόψην μόνο το χαρακτηριστικό αυτό.

Ο DBSCAN κρίθηκε ακατάλληλος για ομαδοποίηση στην περίπτωση μας, καθώς όλα τα δεδομένα κατέληγαν σε μια μόνο ομάδα είτε οι ομάδες ήταν πολύ μικρές. Αυτό συνέβη καθώς υπάρχει ένα πλήθος δεδομένων που είναι το ένα πολύ κοντά στο άλλο και κάποια πιο απόμακρα, που καθιστά δύσκολη την εύρεση των παραμέτρων του αλγόριθμου, όπως η ελάχιστη απόσταση (ϵ) και ο ελάχιστος αριθμός σημείων (minPts). Ως αποτέλεσμα, χάνεται η δυνατότητα εντοπισμού διακριτών ομάδων και η εφαρμογή του DBSCAN δεν αποδίδει τα αναμενόμενα αποτελέσματα.

Επομένως, ο αλγόριθμος KMeans, σε συνδυασμό με τη μέθοδο του αγκώνα, αποδείχθηκε αποτελεσματικός στην ομαδοποίηση δεδομένων με δύο χαρακτηριστικά, διαχωρίζοντάς τα σε διακριτές ομάδες. Ωστόσο, με την αύξηση των χαρακτηριστικών, η αποτελεσματικότητά του μειώθηκε, οδηγώντας σε διαχωρισμό βάσει ενός μόνο χαρακτηριστικού ή σε μη διακριτές ομάδες. Αντίθετα, οι αλγόριθμοι HDBSCAN και Agglomerative Clustering, χρησιμοποιώντας τρία χαρακτηριστικά, ανέδειξαν τη σημασία των χαρακτηριστικών energy και CO_2 για την κατανόηση των δεδομένων, με τον συνδυασμό των χαρακτηριστικών age και energy να παρέχε και αυτός ομαδοποιήσεις. Η προσθήκη του χαρακτηριστικού heating επηρέασε αποφασιστικά τις ομαδοποιήσεις. Ο DBSCAN κρίθηκε ακατάλληλος λόγω της αδυναμίας εντοπισμού διακριτών ομάδων σε δεδομένα με υψηλή πυκνότητα και απομακρυσμένες τιμές, καθιστώντας δύσκολη την επιλογή των κατάλληλων παραμέτρων. Ανοικτά επιστημονικά ερωτήματα παραμένουν σχετικά με τους παράγοντες που επηρεάζουν την απόδοση του KMeans με αυξανόμενα χαρακτηριστικά, τις δυνατότητες βελτίωσης της αξιοπιστίας των HDBSCAN και Agglomerative Clustering, τις προσαρμοστικές μεθόδους παραμετροποίησης του DBSCAN, και την επίδραση των χαρακτηριστικών στην ικανότητα των αλγορίθμων ομαδοποίησης να εντοπίζουν διακριτές ομάδες.

Η έρευνα αποτελεί σημείο αναφοράς για ερευνητές που θέλουν να ασχοληθούν με την ομαδοποίηση δεδομένων σε διάφορα πεδία εφαρμογής αλλά και για τους ερευνητές που ασχολούνται με την ανάλυση δεδομένων που αφορούν πληροφορίες από τον τομέα των ανακαίνισων. Παρέχει κατευθύνσεις και τεχνικές για την καλύτερη κατανόηση παρόμοιων προβλημάτων και της αντιμετώπισης τους όπως η σύγκριση των αλγορίθμων, η επιλογή τους, τα σενάρια πειραμάτων που διεξάγονται και η ανάλυση των αποτελεσμάτων καθώς και για τη συλλογή, την ερμηνεία και την επεξεργασία των δεδομένων. Επίσης, μπορεί να αποτελέσει χρήσιμο εργαλείο για τους εμπλεκόμενους σε έργα ανακαίνισης, παρέχοντας κατευθύνσεις και τεχνικές για τη βέλτιστη αξιοποίηση των δεδομένων, συμβάλλοντας έτσι στην αποτελεσματικότερη υλοποίηση των έργων.

Κάποιοι σημαντικοί περιορισμοί που παρουσίαζε η συγκεκριμένη εργασία είναι ότι το σύνολο δεδομένων που χρησιμοποιήθηκε αφορούσε δεδομένα μόνο από δημόσια κτίρια στη Λετονία. Θα μπορούσε να περιέχει και άλλα είδη κτιρίων αλλά και το είδος του όπου θα μπορούσε να αποτελέσει ένα χρήσιμο χαρακτηριστικό, επηρεάζοντας την τελική ομαδοποίηση. Το μικρό σύνολο δεδομένων που χρησιμοποιείται, είναι πιθανό να μην περιλαμβάνει όλες τις παραλλαγές και τις ιδιαιτερότητες των κτιρίων. Αυτό μπορεί να οδηγήσει σε μη ακριβείς και

παραπλανητικές αναλύσεις σε πραγματικά σύνολα δεδομένων. Άλλη μια σημαντική παρατήρηση που πρέπει να ληφθεί υπόψιν είναι ότι τα δεδομένα είναι δομημένα και δεν γίνεται ανάλυση της μεθοδολογίας σε άλλα είδη δεδομένων όπως real time ή time series δεδομένα. Τέλος, στα υπάρχοντα δεδομένα υπάρχει έλλειψη χαρακτηριστικών όπως τα δομικά χαρακτηριστικά του κτιρίου που θα μπορούσαν να αποτελέσουν σημαντικά χαρακτηριστικά για την ομαδοποίηση.

6.2 Προοπτικές

Οι προοπτικές για μελλοντική έρευνα είναι πολυάριθμες και σημαντικές. Πρώτον, η βελτιστοποίηση της απόδοσης του αλγορίθμου KMeans, του HDBSCAN και του Agglomerative Clustering με την αύξηση των χαρακτηριστικών αποτελεί ένα ανοιχτό επιστημονικό ερώτημα. Με την ανάπτυξη πιο εξελιγμένων τεχνικών κανονικοποίησης και επιλογής χαρακτηριστικών, θα μπορούσε να επιτευχθεί καλύτερη απόδοση σε πιο πολυδιάστατα δεδομένα. Επιπλέον, η αξιοπιστία των αλγορίθμων HDBSCAN και Agglomerative Clustering μπορεί να βελτιωθεί με τη διερεύνηση νέων μεθόδων για την καλύτερη κατανόηση και αξιοποίηση των ιδιοτήτων των χαρακτηριστικών, καθώς και με την προσαρμογή τους σε πιο ποικιλόμορφα σύνολα δεδομένων.

Ακόμα, η ανάπτυξη προσαρμοστικών μεθόδων παραμετροποίησης για τον DBSCAN είναι ένα σημαντικό πεδίο έρευνας. Με την εφαρμογή πιο ευέλικτων και δυναμικών τεχνικών για την επιλογή των παραμέτρων, όπως η ελάχιστη απόσταση (ϵ) και ο ελάχιστος αριθμός σημείων (minPts), ο DBSCAN θα μπορούσε να αποδώσει καλύτερα σε δεδομένα με υψηλή πυκνότητα και απομακρυσμένες τιμές. Η διερεύνηση της επίδρασης των επιμέρους χαρακτηριστικών στην ικανότητα των αλγορίθμων ομαδοποίησης να εντοπίζουν διακριτές ομάδες είναι επίσης κρίσιμη. Η κατανόηση του πώς τα διάφορα χαρακτηριστικά αλληλεπιδρούν και επηρεάζουν τις ομαδοποιήσεις μπορεί να οδηγήσει σε πιο αποτελεσματική και ακριβή ανάλυση δεδομένων.

Επίσης, θα μπορούσαν να δοκιμαστούν και άλλοι αλγόριθμοι ομαδοποίησης ή κάποιος συνδυασμός τους για ensemble methods με σκοπό τη βελτίωση των αποτελεσμάτων. Είναι χρήσιμο να ληφθούν υπόψιν και τα δεδομένα που δίνονται ως είσοδος στους αλγορίθμους. Με περισσότερα και πιο πλήρη δεδομένα αποφεύγονται προβλήματα overfitting και τα αποτελέσματα είναι πιθανότατα πιο αξιόπιστα καθώς υπάρχει ποικιλία δεδομένων για τον διαχωρισμό τους σε ομάδες. Τέλος, θα μπορούσε να δημιουργηθεί μια διεπαφή με τα αποτελέσματα για την πιο εύκολη προβολή και αξιολόγηση τους.

Με την υλοποίηση αυτών των προτάσεων, η έρευνα στον τομέα της τεχνητής νοημοσύνης και της μηχανικής μάθησης για την υποστήριξη των αποφασίζοντων στη λήψη βέλτιστων μέτρων ανακαίνισης κτιρίων θα προχωρήσει σημαντικά, προσφέροντας καλύτερα εργαλεία για τη μείωση των εκπομπών και τη βελτίωση της ενεργειακής αποδοτικότητας.

ΚΕΦΑΛΑΙΟ 7

Βιβλιογραφία

7 Βιβλιογραφία

- [1] M. A. Aktar, M. M. Alam, and A. Q. Al-Amin, “Global economic crisis, energy use, CO₂ emissions, and policy roadmap amid COVID-19,” *Sustain. Prod. Consum.*, vol. 26, pp. 770-781, 2021.
- [2] E. Sarmas, M. Kleideri, A. Zučika, V. Marinakis, and H. Doukas, “Improving energy performance of buildings: Dataset of implemented energy efficiency renovation projects in Latvia,” *Data Br.*, vol. 48, p. 109225, 2023.
- [3] “European Commission.” https://energy.ec.europa.eu/topics/energy-efficiency_en.
- [4] A. E. Stagrum, E. Andenæs, T. Kvande, and J. Lohne, “Climate change adaptation measures for buildings—A scoping review,” *Sustainability*, vol. 12, no. 5, p. 1721, 2020.
- [5] A. Kałmierczak, “Unequal exposure and unequal impacts: social vulnerability to air pollution, noise and extreme temperatures in Europe.,” 2018.
- [6] “European Environment Agency.” https://www.eea.europa.eu/en/advanced-search?q=buildings&size=n_10_n&filters%25B0%25D%25Bfield%25D=readingTime&filters%25B0%25D%25Btype%25D=any&filters%25B0%25D%25Bvalues%25D%25B0%25D%25Bname%25D=All&filters%25B0%25D%25Bvalues%25D%25B0%25D%25BrangeType%25D=fixed&filters%25B1%25D%255.
- [7] H. Doukas, A. Nikas, M. González-Eguino, I. Arto, and A. Anger-Kraavi, “From integrated to integrative: Delivering on the Paris Agreement,” *Sustainability*, vol. 10, no. 7, p. 2299, 2018.
- [8] S. Intrachooto and V. Horayangkura, “Energy efficient innovation: Overcoming financial barriers,” *Build. Environ.*, vol. 42, no. 2, pp. 599-604, 2007.
- [9] E. Commission, “A renovation wave for Europe—greening our buildings, creating jobs, improving lives,” *Off. J. Eur. Union*, p. 26, 2020.
- [10] L. Caruso and V. Buhagiar, “Educating and training for a new European Bauhaus-A proposed renovation wave for students and professional AEC curricula,” 2021.
- [11] G. Kats, *Greening our built world: costs, benefits, and strategies*. Island Press, 2013.
- [12] J. J. Schoenefeld, B. Stein, and I. Renz, “Renovating Europe: How to start and steer a wave?,” in *Making the European Green Deal Work*, Routledge, 2023, pp. 76-93.
- [13] M. Santamouris and K. Vasilakopoulou, “Present and future energy consumption of buildings: Challenges and opportunities towards decarbonisation,” *e-Prime-Advances Electr. Eng. Electron. Energy*, vol. 1, p. 100002, 2021.
- [14] M. Attia, N. Haidar, S. M. Senouci, and E.-H. Aglzim, “Towards an efficient energy management to reduce CO₂ emissions and billing cost in smart buildings,” in *2018 15th IEEE Annual Consumer Communications & Networking Conference (CCNC)*, 2018, pp. 1-6.
- [15] A. F. Pales and S. Bennett, “Energy technology perspectives 2020,” *Int. Energy Agency*, 2020.
- [16] I. E. A. (IEA), “Perspectives for the Clean Energy Transition: The Critical Role of Buildings,” 2019.

- [17] A. Ben Ayed, M. Ben Halima, and A. M. Alimi, "Adaptive fuzzy exponent cluster ensemble system based feature selection and spectral clustering," in *2017 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, 2017, pp. 1-6.
- [18] J. Ouyang and K. Hokao, "Establishment of a Methodology to Guide Energy-efficient Renovation of Existing Residential Buildings in China," in *2011 International Conference on Computer Distributed Control and Intelligent Environmental Monitoring*, 2011, pp. 1487-1491.
- [19] S. Mulero-Palencia, S. Álvarez-Díaz, and M. Andrés-Chicote, "Machine learning for the improvement of deep renovation building projects using as-built BIM models," *Sustainability*, vol. 13, no. 12, p. 6576, 2021.
- [20] F. Nielsen and F. Nielsen, "Hierarchical clustering," *Introd. to HPC with MPI Data Sci.*, pp. 195-211, 2016.
- [21] I. De Jaeger, G. Reynders, C. Callebaut, and D. Saelens, "A building clustering approach for urban energy simulations," *Energy Build.*, vol. 208, p. 109671, 2020.
- [22] P. D. McNicholas, "Model-based clustering," *J. Classif.*, vol. 33, pp. 331-373, 2016.
- [23] D. Lei, Q. Zhu, J. Chen, H. Lin, and P. Yang, "Automatic PAM Clustering Algorithm for Outlier Detection.," *J. Softw.*, vol. 7, no. 5, pp. 1045-1051, 2012.
- [24] N. Ghiassi and A. Mahdavi, "Reductive bottom-up urban energy computing supported by multivariate cluster analysis," *Energy Build.*, vol. 144, pp. 372-386, 2017.
- [25] G. Tardioli, R. Kerrigan, M. Oates, J. O'Donnell, and D. P. Finn, "Identification of representative buildings and building groups in urban datasets using a novel pre-processing, classification, clustering and predictive modelling approach," *Build. Environ.*, vol. 140, pp. 90-106, 2018.
- [26] S. Askari, "Fuzzy C-Means clustering algorithm for data with unequal cluster sizes and contaminated with noise and outliers: Review and development," *Expert Syst. Appl.*, vol. 165, p. 113856, 2021.
- [27] X. Ran, X. Zhou, M. Lei, W. Tepsan, and W. Deng, "A novel k-means clustering algorithm with a noise algorithm for capturing urban hotspots," *Appl. Sci.*, vol. 11, no. 23, p. 11202, 2021.
- [28] K. Ahmed Ali, M. I. Ahmad, and Y. Yusup, "Issues, impacts, and mitigations of carbon dioxide emissions in the building sector," *Sustainability*, vol. 12, no. 18, p. 7427, 2020.
- [29] A. Brambilla, G. Salvalai, M. Imperadori, and M. M. Sesana, "Nearly zero energy building renovation: From energy efficiency to environmental efficiency, a pilot case study," *Energy Build.*, vol. 166, pp. 271-283, 2018.
- [30] E. Alpaydin, *Introduction to machine learning*. MIT press, 2020.
- [31] M. Alloghani, D. Al-Jumeily, J. Mustafina, A. Hussain, and A. J. Aljaaf, "A systematic review on supervised and unsupervised machine learning algorithms for data science," *Supervised unsupervised Learn. data Sci.*, pp. 3-21, 2020.
- [32] K. P. Sinaga and M.-S. Yang, "Unsupervised K-means clustering algorithm," *IEEE access*, vol. 8, pp. 80716-80727, 2020.
- [33] D. Deng, "DBSCAN clustering algorithm based on density," in *2020 7th international forum on electrical engineering and automation (IFEEA)*, 2020, pp. 949-953.

- [34] M. Fuchs and W. Höpken, “Clustering: Hierarchical, k-Means, DBSCAN,” in *Applied Data Science in Tourism: Interdisciplinary Approaches, Methodologies, and Applications*, Springer, 2022, pp. 129-149.
- [35] F. Bach and M. Jordan, “Learning spectral clustering,” *Adv. Neural Inf. Process. Syst.*, vol. 16, 2003.
- [36] Y. Zhang *et al.*, “Gaussian mixture model clustering with incomplete data,” *ACM Trans. Multimed. Comput. Commun. Appl.*, vol. 17, no. 1s, pp. 1-14, 2021.
- [37] I. Davidson and S. S. Ravi, “Agglomerative hierarchical clustering with constraints: Theoretical and empirical results,” in *European Conference on Principles of Data Mining and Knowledge Discovery*, 2005, pp. 59-70.
- [38] M. A. Carreira-Perpinán, “A review of mean-shift algorithms for clustering,” *arXiv Prepr. arXiv1503.00687*, 2015.
- [39] D. Dueck, *Affinity propagation: clustering data by passing messages*. University of Toronto Toronto, ON, Canada, 2009.
- [40] K. Peng, L. Zheng, X. Xu, T. Lin, and V. C. M. Leung, “Balanced iterative reducing and clustering using hierarchies with principal component analysis (PBIRCH) for intrusion detection over big data in mobile cloud environment,” in *Security, Privacy, and Anonymity in Computation, Communication, and Storage: 11th International Conference and Satellite Workshops, SpaCCS 2018, Melbourne, NSW, Australia, December 11-13, 2018, Proceedings 11*, 2018, pp. 166-177.
- [41] “pandas.DataFrame.info.” <https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.DataFrame.info.html>.
- [42] “pandas.DataFrame.describe.” <https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.DataFrame.describe.html#pandas.DataFrame.describe>.
- [43] “seaborn.displot,” [Online]. Available: <https://seaborn.pydata.org/generated/seaborn.displot.html>.
- [44] “seaborn.boxplot,” [Online]. Available: <https://seaborn.pydata.org/generated/seaborn.boxplot.html>.
- [45] A. G. Asuero, A. Sayago, and A. G. González, “The correlation coefficient: An overview,” *Crit. Rev. Anal. Chem.*, vol. 36, no. 1, pp. 41-59, 2006.
- [46] “pandas.DataFrame.corr.” <https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.DataFrame.corr.html>.
- [47] “StandardScaler,” [Online]. Available: <https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html>.
- [48] “AgglomerativeClustering.” <https://scikit-learn.org/stable/modules/generated/sklearn.cluster.AgglomerativeClustering.html>.
- [49] “DBSCAN.” <https://scikit-learn.org/stable/modules/generated/sklearn.cluster.DBSCAN.html>.
- [50] “KMeans.” <https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html>.