ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ Μ/Υ

ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ
ΣΧΟΛΗ ΝΑΥΤΙΛΙΑΣ ΚΑΙ ΒΙΟΜΗΧΑΝΙΑΣ
ΤΜΗΜΑΤΟΣ ΒΙΟΜΗΧΑΝΙΚΗΣ ΔΙΟΙΚΗΣΗΣ & ΤΕΧΝΟΛΟΓΙΑΣ

ΔΙΑΠΑΝΕΠΙΣΤΗΜΙΑΚΟ ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ
ΣΠΟΥΔΩΝ «ΤΕΧΝΟ-ΟΙΚΟΝΟΜΙΚΑ ΣΥΣΤΗΜΑΤΑ»

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Αξιολόγηση Επενδύσεων Ενεργειακής Αποδοτικότητας με Μοντέλα Μηχανικής Μάθησης

Ελισσαίος Β. Σαρμάς

Επιβλέπων: Χάρης Δούκας

Αν. Καθηγητής Ε.Μ.Π.

Αθήνα, Φεβρουάριος 2023

ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ Μ/Υ

ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ
ΣΧΟΛΗ ΝΑΥΤΙΛΙΑΣ ΚΑΙ ΒΙΟΜΗΧΑΝΙΑΣ
ΤΜΗΜΑΤΟΣ ΒΙΟΜΗΧΑΝΙΚΗΣ ΔΙΟΙΚΗΣΗΣ & ΤΕΧΝΟΛΟΓΙΑΣ

ΔΙΑΠΑΝΕΠΙΣΤΗΜΙΑΚΟ ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ
ΣΠΟΥΔΩΝ «ΤΕΧΝΟ-ΟΙΚΟΝΟΜΙΚΑ ΣΥΣΤΗΜΑΤΑ»

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Αξιολόγηση Επενδύσεων Ενεργειακής Αποδοτικότητας με Μοντέλα Μηχανικής Μάθησης

Ελισσαίος Β. Σαρμάς

Επιβλέπων: Χάρης Δούκας

Αν. Καθηγητής Ε.Μ.Π.

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την ΧΧη ΧΧ 2022.

.....	.....	.....
Ψαρράς Ι.	Ασκούνης Δ.	Δούκας Χ.
Καθηγητής Ε.Μ.Π.	Καθηγητής Ε.Μ.Π.	Αν. Καθηγητής Ε.Μ.Π.

Αθήνα, Φεβρουάριος 2023

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.

.....

Ελισσαίος Β. Σαρμάς

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

Περίληψη

Κύριο αντικείμενο της παρούσας Διπλωματικής Εργασίας είναι η ανάπτυξη ενός μοντέλου μετα-μάθησης για την αξιολόγηση των επενδύσεων ενεργειακής αποδοτικότητας. Η ενεργειακή αποδοτικότητα των κτηρίων είναι ένας κρίσιμος παράγοντας για την επίτευξη των παγκόσμιων ενεργειακών και κλιματικών στόχων, απαιτώντας ωστόσο σημαντικές επενδύσεις. Λόγω της έλλειψης σε συστήματα υποστήριξης αποφάσεων, η οποία οδηγεί στη χρήση των παραδοσιακών επενδυτικών μηχανισμών που εστιάζουν μόνο στις οικονομικές πτυχές των έργων ενεργειακής αποδοτικότητας και παραμελούν τις περιβαλλοντικές τους επιπτώσεις, τέτοια έργα μπορεί να αντιμετωπίσουν σοβαρές δυσκολίες στη χρηματοδότηση. Στο μεταξύ, ο αντίκτυπος της εποχής της ψηφιοποίησης είναι πιο εμφανής από ποτέ, καθώς οι αλγόριθμοι τεχνητής νοημοσύνης βρίσκονται πλέον στο επίκεντρο και η διαθεσιμότητα και η ποιότητα των δεδομένων έχουν βελτιωθεί σημαντικά. Αυτή η Διπλωματική Εργασία φιλοδοξεί να γεφυρώσει το χάσμα στη χρηματοδότηση έργων ενεργειακής αποδοτικότητας με την ανάπτυξη μιας μεθοδολογίας βασισμένης σε δεδομένα που κατηγοριοποιεί τις επενδύσεις ενεργειακής αποδοτικότητας με βάση την αναμενόμενη αξία τους όσον αφορά το κόστος ανακαίνισης και την προσφερόμενη εξοικονόμηση ενέργειας. Στο πλαίσιο αυτό, αναπτύσσονται διάφορες μέθοδοι ταξινόμησης μηχανικής μάθησης οι οποίες συνδυάζονται μέσω ενός μοντέλου μετα-μάθησης με στόχο τη βελτίωση της συνολικής προβλεπτικής επίδοσης στο πρόβλημα της ταξινόμησης και τον προσδιορισμό της χρηματοδότησης που πρέπει να λάβει κάθε επένδυση σύμφωνα με τα ιδιαίτερα χαρακτηριστικά της. Η προτεινόμενη μεθοδολογία αξιολογείται χρησιμοποιώντας ένα σύνολο 312 έργων που έχουν ολοκληρωθεί στη Λετονία. Τα αποτελέσματα υποδεικνύουν ότι το μοντέλο μέτα-μάθησης παρουσιάζει καλύτερη επίδοση από όλους τους απλούς ταξινομητές, εντοπίζοντας πολύ αποτελεσματικά έργα υψηλής και μεσαίας αξίας και διακρίνοντας με απόλυτη επιτυχία έργα χαμηλής από έργα υψηλής αξίας.

Λέξεις Κλειδιά: Ενεργειακή Αποδοτικότητα, Επενδυτικός Σχεδιασμός, Κτήρια, Ανακαινίσεις, Μηχανική Μάθηση

Abstract

The main object of this Thesis is the development of a meta-learning model for assessing energy efficiency investments. Energy efficiency is critical for meeting global energy and climate targets, requiring however significant investments. Due to the lack of mature decision-support systems and the utilization of traditional investment mechanisms that focus on the economical aspects of the energy efficiency projects and neglect their environmental impact, such projects can experience difficulties in being funded. In the interim, the impact of the digitization era is more apparent than ever, as algorithms and data availability and quality have significantly improved. This Thesis aspires to bridge the gap in energy efficiency financing with the development of a data-driven methodology that labels energy efficiency investments based on their expected utility in terms of renovation cost and energy savings. Various machine learning classification methods are deployed and combined through a meta-learning model with the objective to improve overall classification performance and determine the funding that each investment should receive according to its particular characteristics. The proposed methodology is evaluated using a set of 312 projects that have been completed in Latvia. Our results indicate that the meta-learner outperforms all baseline classifiers, effectively identifying projects of high and medium potential and successfully distinguishing low from high potential ones.

Keywords: Energy Efficiency, Investment Planning, Buildings, Renovations, Machine Learning

Contents

Chapter 1

Εισαγωγή

1.1 Εισαγωγή

Η κλιματική αλλαγή επηρεάζει ήδη τις ζωές μας θέτοντας σε κίνδυνο το οικοσύστημα και απειλώντας τη βιωσιμότητα του πλανήτη. Επιστήμονες, κυβερνήσεις και ερευνητικά ιδρύματα προειδοποιούν ότι βρισκόμαστε σε μια αχαρτογράφητη περιοχή γεμάτη αβεβαιότητα. Η τελευταία μεγάλη προσπάθεια μείωσης των εκπομπών CO_2 και επιβολής ορισμένων ορίων στον ρυθμό αύξησης της θερμοκρασίας ήταν η Συμφωνία του Παρισιού [21]. Σύμφωνα με αυτή, η μέση παγκόσμια αύξηση της 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) έχει προτείνει πως απαιτείται η πλήρης απαλλαγή του ενεργειακού τομέα από τον άνθρακα έως το 2050 [32].

Ωστόσο, η αλλαγή του ενεργειακού μείγματος και η μετάβαση σε μια «πράσινη» ενεργειακή εποχή δεν είναι ομοιογενής σε όλο τον κόσμο. Διαφορετικές χώρες πληρούν διαφορετικές συνθήκες ανάπτυξης, δηλαδή διαφορετικά ιστορικά και πολιτιστικά υπόβαθρα καθώς και διαφορετικά οικονομικά, κοινωνικά, πολιτικά και περιβαλλοντικά πλαίσια [57]. Με άλλα λόγια, η επίτευξη των στόχων που έχουν τεθεί σε παγκόσμιο επίπεδο εξαρτάται από τις ενέργειες μεμονωμένων χωρών ή τομέων που δεν ενεργούν πάντα αρμονικά κατά των εκπομπών αερίων του θερμοκηπίου. Αυτά τα αδιαμφισβήτητα εμπόδια στην ενεργειακή μετάβαση ενισχύουν τη σημασία των συμπληρωματικών δράσεων με στόχο τη μείωση του ενεργειακού αποτυπώματος, όπως είναι για παράδειγμα οι δράσεις για τη μείωση της κατανάλωσης ενέργειας στον κτηριακό τομέα.

Αυτό το μονοπάτι περιλαμβάνει την προώθηση επενδύσεων ενεργειακής αποδοτικότητας, ειδικά σε κτίρια, καθώς και την προθυμία των πολιτών να συμμετάσχουν μαζικά σε έργα ανανεώσιμων πηγών ενέργειας, η οποία με τη σειρά της επηρεάζεται από πολλούς παράγοντες σύμφωνα με το [36]. Τα πιο σημαντικά θέματα είναι η περιβαλλοντική συνείδηση και η προοπτική ενεργειακής αυτάρκειας [8], αν και η επιτυχία τέτοιων έργων μπορεί επίσης να επηρεαστεί από κοινωνικούς παράγοντες, όπως η εμπιστοσύνη και οι κοινωνικοί κανόνες [63]. Από την άλλη πλευρά, η δυσκολία των παλαιότερων γενεών

να υιοθετήσουν φιλικές προς το περιβάλλον συνήθειες [2], επιβάλλει την υιοθέτηση πιο αποτελεσματικών τεχνολογιών. Από αυτή την άποψη, οι έξυπνες πόλεις και τα έργα ανακαίνισης κτιρίων είναι ζωτικής σημασίας για μια εποχή πράσινης ενέργειας.

Σύμφωνα με τον Διεθνή Οργανισμό Ενέργειας (IEA), η εφαρμογή των πολιτικών ενεργειακής αποδοτικότητας θα μπορούσε να έχει ως αποτέλεσμα τη μείωση κατά σχεδόν 36% των εκπομπών θερμοκηπίου έως το 2050 [29, 30]. Η συντριπτική πλειοψηφία αυτής της μείωσης αναμένεται να προέλθει από έργα ανακαίνισης κτιρίων [61]. Ωστόσο, εκτός από τις θεωρητικές προσδοκίες, στην πράξη οι επενδύσεις ενεργειακής αποδοτικότητας αντιμετωπίζουν ορισμένα εμπόδια. Τα έργα ενεργειακής αποδοτικότητας είναι συχνά κατακερματισμένα και το κόστος και η εξοικονόμηση ενέργειας είναι εκ των προτέρων άγνωστα ή είναι δύσκολο να εκτιμηθούν με ακρίβεια. Έτσι, οι ενδιαφερόμενοι (ιδιωτικά χρηματοπιστωτικά ιδρύματα, επενδυτικά κεφάλαια, εθνικές και περιφερειακές αρχές, καθώς και πάροχοι ενεργειακών λύσεων) στερούνται ώριμων εργαλείων λήψης αποφάσεων που θα μπορούσαν να προβλέψουν τη χρησιμότητα των μελλοντικών έργων ενεργειακής αποδοτικότητας και να τους βοηθήσουν να καθοδηγήσουν τις ενέργειές τους με μεγαλύτερη αξιοπιστία. Επιπλέον, οι αναπτυσσόμενες χώρες, οι οποίες έχουν τεράστιες δυνατότητες να αυξήσουν την αποτελεσματικότητά τους όσον αφορά την εξοικονόμηση ενέργειας, έχουν αντιμετωπίσει πολλά εμπόδια, συμπεριλαμβανομένης της έλλειψης πρόσβασης σε κατάλληλους χρηματοδοτικούς μηχανισμούς [49].

Ο όγκος των ανακαινίσεων και άλλων πρωτοβουλιών που στοχεύουν στην ενεργειακή αποδοτικότητα είναι σημαντικός όσον αφορά την κατασκευή και τους ευρύτερους στόχους πολιτικής. Οι εκτιμήσεις δείχνουν ότι οι δαπάνες για την ανακαίνιση είναι παρόμοιες σε συνολικό μέγεθος με αυτές για την κατασκευή νέων κατοικιών. Η κλίμακα αυτής της δραστηριότητας προσφέρει μια ευκαιρία για τους υπεύθυνους χάραξης πολιτικής που επιδιώκουν να παρέμβουν σε ιδιωτικές κατοικίες για τη μείωση των εκπομπών αερίων του θερμοκηπίου. Οι πολιτικές και τα προγράμματα συνήθως υποθέτουν ότι οι ιδιοκτήτες μπορούν να πειστούν να τροποποιήσουν τις δραστηριότητες βελτίωσης του σπιτιού τους με τρόπους που ενσωματώνουν τεχνικές λύσεις που οδηγούν σε μειώσεις των εκπομπών αερίων του θερμοκηπίου. Παρόλο που υπάρχει διαθέσιμη μια μεγάλη ποικιλία τεχνικών λύσεων, αυτές οι λύσεις πολύ συχνά δεν αξιοποιούνται. Έχει προταθεί ότι τα νοικοκυριά «παραβλέπουν» παρεμβάσεις ενεργειακής αποδοτικότητας και άλλες βελτιώσεις για τη μείωση των εκπομπών, παρόλο που αυτές μπορεί να είναι οικονομικά βιώσιμες. Αυτό το αποτέλεσμα υποδηλώνει ότι υπάρχει διάσταση μεταξύ της χάραξης πολιτικής για την ενεργειακή αποδοτικότητα και της καθημερινής ζωής.

Όσον αφορά τη χρηματοδότηση επενδύσεων ενεργειακής αποδοτικότητας μεγάλης κλίμακας, ένα από τα εμπόδια που υπάρχουν είναι η πιθανή έλλειψη επαρκούς όγκου δεδομένων[1] που θα μπορούσε να επιτρέψει την αξιολόγηση της χρησιμότητάς τους, αλλά,

[1]Για παράδειγμα, πολλά από τα έργα που χρηματοδοτούνται επί του παρόντος είναι μικρής κλίμακας ή σε πιλοτικό στάδιο, καθιστώντας τη συγκριτική αξιολόγηση ένα δύσκολο πρόβλημα.

το πιο σημαντικό, το γεγονός ότι ακόμη και όταν τέτοια δεδομένα υπάρχουν και είναι άμεσα διαθέσιμα προς χρήση, δεν αξιοποιούνται πάντα με συστηματικό τρόπο για την αποτελεσματική υποστήριξη της λήψης αποφάσεων. Για παράδειγμα, οι χρηματοδοτικοί φορείς χρησιμοποιούν επί του παρόντος διαφορετικές μεθόδους, πρότυπα και πηγές δεδομένων για να αξιολογήσουν τον κίνδυνο και την αποτελεσματικότητα των μελλοντικών επενδύσεων [27], καθώς και διαφορετικούς δείκτες για να λάβουν υπόψη την οικονομική δομή, το κλίμα και τη γεωγραφία τους, μεταξύ άλλων παράγοντες που επηρεάζουν. Σύμφωνα με το [48], τα εμπόδια στις ανανεώσιμες πηγές ενέργειας ενδέχεται να διαφέρουν μεταξύ τεχνολογιών και χωρών. Ειδικά η Ευρωπαική Ένωση έχει αναγνωρίσει αυτό το πρόβλημα, χρηματοδοτώντας ερευνητικά έργα που στοχεύουν στην απομάκρυνση του κινδύνου επενδύσεων σε έργα ενεργειακής αποδοτικότητας, όπως τα *EEnvest*, *Triple-A* και *Quest: 2050*, μεταξύ άλλων. Ως αποτέλεσμα, τα έργα ενεργειακής αποδοτικότητας συχνά θεωρείται ότι ενσωματώνουν υψηλότερους κινδύνους και τα χρηματοδοτικά ιδρύματα τείνουν να επικεντρώνονται σε άλλους τύπους έργων ή σε παραδοσιακές επενδύσεις [55].

1.2 Συνεισφορά της Διπλωματικής

Σε αυτή τη διπλωματική εργασία παρουσιάζεται μια μεθοδολογία βασισμένη σε δεδομένα για το πρόβλημα της αξιολόγησης των επενδύσεων ενεργειακής αποδοτικότητας. Με κίνητρο τις πρόσφατες εξελίξεις στον τομέα της μηχανικής μάθησης και την επιτυχή χρήση της στον τομέα της ταξινόμησης, προτείνουμε τη χρήση ενός μοντέλου μετα-μάθησης [2] που προβλέπει τη χρησιμότητα (utility) μελλοντικών έργων ενεργειακής αποδοτικότητας όσον αφορά το κόστος και την πραγματοποιηθείσα εξοικονόμηση ενέργειας. Χρησιμοποιώντας ένα πλούσιο σύνολο δεδομένων που αποτυπώνει (i) το κόστος της ανακαίνισης, (ii) την τρέχουσα κατανάλωση ενέργειας του κτιρίου, (iii) τα ιδιαίτερα χαρακτηριστικά του κτιρίου, καθώς και (iv) την αναμενόμενη και (v) πραγματοποιηθείσα εξοικονόμηση ενέργειας πολλαπλών έργών ενεργειακής αποδοτικότητας που έχουν ολοκληρωθεί στο παρελθόν, εκπαιδεύουμε διάφορα μοντέλα ταξινόμησης για να προσδιορίσουμε την ελκυστικότητα των μελλοντικών επενδύσεων ενεργειακής αποδοτικότητας (υψηλή, μεσαία ή χαμηλή). Στη συνέχεια, χρησιμοποιείται ένας μετα-μαθητής (μοντέλο μετα-μάθησης) για να ορίσει ποιος ταξινομητής θα πρέπει να επιλεγεί για να κάνει την πρόβλεψη σύμφωνα με τις ιδιαιτερότητες του εξεταζόμενου έργου, βελτιώνοντας έτσι περαιτέρω την ακρίβεια των αποτελεσμάτων μας. Επιδεικνύουμε τα πλεονεκτήματα της προσέγγισής μας και αναλύουμε πώς μπορεί να χρησιμοποιηθεί στην πράξη για να βοηθήσουμε τους ενδιαφερόμενους να διαφοροποιήσουν τις επενδύσεις τους με βάση την αναμενόμενη χρησιμότητα τους.

[2] Η μετα-μάθηση περιλαμβάνει τη χρήση μεθόδων μηχανικής μάθησης με στόχο τη μάθηση για το πώς να επιλέξετε ή να συνδυάσετε καλύτερα τις προβλέψεις που γίνονται με άλλες μεθόδους μηχανικής μάθησης.

Οι συνεισφορές της μελέτης μας συνοψίζονται ως εξής:

- Προσφέρουμε ένα νέο μοντέλο μηχανικής μάθησης που αξιολογεί τη χρησιμότητα των μελλοντικών έργων ενεργειακής αποδοτικότητας ως προς το κόστος και την πραγματοποιούμενη εξοικονόμηση ενέργειας με συστηματικό τρόπο, βοηθώντας έτσι τα χρηματοδοτικά ιδρύματα στη λήψη αποφάσεων. Το μοντέλο μας είναι εύκολο να αναπαραχθεί και επομένως μπορεί να ενσωματωθεί στα υπάρχοντα συστήματα των ινστιτούτων με χαμηλό κόστος.

- Αντί να ταξινομούμε μελλοντικές επενδύσεις χρησιμοποιώντας συγκεκριμένους δείκτες που εκτιμούν την αποτελεσματικότητά τους θεωρητικά, βασίζουμε τις προτάσεις μας στην υλοποιημένη χρησιμότητα πολλών, παρόμοιων έργων ενεργειακής αποδοτικότητας που έχουν ολοκληρωθεί στο παρελθόν, μαθαίνοντας έτσι εμπειρικά από την επιτυχία και τους περιορισμούς τους και προσαρμόζοντας προτάσεις σύμφωνα με τα ιδιαίτερα χαρακτηριστικά κάθε νέας επένδυσης.

- Για τον μετριασμό των υποκείμενων κινδύνων από τη χρήση ενός μόνο ταξινομητή (π.χ. όσον αφορά την ακρίβεια και την ευρωστία όταν πρόκειται για μικρά, μη ισορροπημένα ή θορυβώδη σύνολα δεδομένων) για την αξιολόγηση μελλοντικών επενδύσεων, προτείνουμε τη χρήση ενός μετα-μαθητή που είναι υπεύθυνος για τον προσδιορισμό του ταξινομητή που αναμένεται να οδηγήσει σε πιο αξιόπιστες προβλέψεις.

- Παρέχουμε τις προβλέψεις ταξινόμησής μας με τη μορφή πιθανοτήτων για να λάβουμε υπόψη την αβεβαιότητα και χρησιμοποιούμε αυτά τα αποτελέσματα για να κάνουμε δυναμικά συστάσεις σχετικά με το ποσοστό επιχορήγησης που θα πρέπει να λαμβάνουν οι μελλοντικές επενδύσεις.

1.3 Δομή της Διπλωματικής

Η παρούσα διπλωματική αποτελείται από 5 κεφάλαια.

Στο **πρώτο κεφάλαιο** πραγματοποιήθηκε μια εισαγωγή στο πρόβλημα της αξιολόγησης και κατάταξης επενδύσεων ενεργειακής αποδοτικότητας. Αξιολογήθηκε η σημασία που έχουν οι επενδύσεις ενεργειακής αποδοτικότητας στον κτηριακό τομέα και επίσης παρουσιάστηκαν επιγραμματικά η συνεισφορά και η δομή της διπλωματικής.

Στο **δεύτερο κεφάλαιο** δίνεται έμφαση στο πρόβλημα της αξιολόγησης ενεργειακά αποδοτικών επενδύσεων. Πιο συγκεκριμένα, μελετώνται τα στοιχεία του προβλήματος τα οποία είναι πολύπλευρα, δίνεται έμφαση στην παρουσίαση των μεθοδολογιών που έχουν χρησιμοποιηθεί για την αξιολόγηση τέτοιων επενδύσεων και διατυπώνεται το πρόβλημα που θα επιλύσει η συγκεκριμένη διπλωματική.

Στο **τρίτο κεφάλαιο** πραγματοποιείται η παρουσίαση της μεθοδολογίας που έχει αναπτυχθεί στην παρούσα διπλωματική εργασία. Διατυπώνεται ένας ολοκληρωμένος

ορισμός των βασικών μοντέλων για ταξινόμηση με τις αναλυτικές του εξισώσεις. Επίσης, γίνεται εισαγωγή στην έννοια της μετα-μάθησης, η οποία χρησιμοποιείται στην μεθοδολογία ταξινόμησης έργων ενεργειακής αποδοτικότητας.

Στο **τέταρτο κεφάλαιο** πραγματοποιείται η αναλυτική παρουσίαση της πειραματικής εφαρμογής της διπλωματικής. Πιο συγκεκριμένα, γίνεται λεπτομερής περιγραφή του συνόλου των δεδομένων που χρησιμοποιήθηκαν, δίνεται έμφαση στις κατανομές των δεδομένων για κάθε επιμέρους μεταβλητή και παρουσιάζονται τα αποτελέσματα της κατάταξης για κάθε βασικό μοντέλο αλλά και για το μοντέλο μετα-μάθησης.

Τέλος, στο **πέμπτο κεφάλαιο** συνοψίζονται τα συμπεράσματα της παρούσας διπλωματικής και παρουσιάζονται οι μελλοντικές προεκτάσεις για επιπλέον έρευνα.

Chapter 2

Το πρόβλημα της αξιολόγησης επενδύσεων ενεργειακής αποδοτικότητας

2.1 Εισαγωγή

Είναι γενικά αποδεκτό ότι η κατανάλωση ενέργειας στα κτίρια μπορεί να μειωθεί σημαντικά μέσω της υιοθέτησης των υφιστάμενων τεχνολογιών εξοικονόμησης ενέργειας. Ωστόσο, τα έργα ενεργειακής αποδοτικότητας ενδέχεται να αντιμετωπίζουν δυσκολίες στη χρηματοδότηση λόγω της αβεβαιότητας και της έλλειψης μιας ολοκληρωμένης μεθοδολογίας για την αξιολόγηση και τη σύγκριση των δυνατοτήτων τους [35]. Πιο συγκεκριμένα, η [31] έχει προσδιορίσει την αλληλεξάρτηση των στοιχείων ανακαίνισης, την κερδοσκοπική οικονομική απόδοση και τη νοοτροπία χαμηλότερου κόστους ως τους κύριους παράγοντες από τους οποίους προκύπτουν οικονομικά εμπόδια. Επί του παρόντος, οι ανακαινίσεις σε κτίρια αξιολογούνται με παραδοσιακούς μηχανισμούς παράδοσης επενδύσεων που έχουν αναπτυχθεί από μεγάλα χρηματοδοτικά ιδρύματα και τράπεζες. Αν και πολλά έργα έχουν χρηματοδοτηθεί μέσω αυτής της διαδικασίας, πολλές δυνατότητες μπορούν ακόμα να αξιοποιηθούν [62].

Τα τελευταία χρόνια, η αυξανόμενη υιοθέτηση τεχνολογιών πληροφοριών και επικοινωνιών, όπως το διαδίκτυο των πραγμάτων και η τεχνητή νοημοσύνη, κατέστησαν δυνατή την απόκτηση μεγάλου όγκου ετερογενών δεδομένων που μπορούν να οδηγήσουν σε νέες λύσεις [41]. Πιο συγκεκριμένα, στον τομέα της χρηματοδότησης έργων ενεργειακής αποδοτικότητας είναι πλέον δυνατή η συλλογή δεδομένων από έξυπνους μετρητές μετά την υλοποίηση των ανακαινίσεων και η λήψη πληροφοριών για την πραγματική τους απόδοση. Ωστόσο, η χρηματοδότηση έργων ενεργειακής αποδοτικότητας σε κτίρια παραμένει σχετικά υπο-ερευνημένη σύμφωνα με το [69]. Κατά συνέπεια, οι μελλοντικές μελέτες θα πρέπει να επικεντρωθούν στην εφαρμογή μεθόδων και τεχνικών τεχνητής νοημοσύνης [16], καθώς και στη βέλτιστη εκμετάλλευση των διαθέσιμων δε-

δομένων για την παροχή εργαλείων που αξιολογούν με ακρίβεια τέτοιες επενδύσεις [34].

2.2 Επισκόπηση συσχετιζόμενων μεθοδολογιών

Τα μέτρα που προσανατολίζονται στη βελτίωση της ενεργειακής αποδοτικότητας των κτιρίων έχουν αποδειχθεί αποτελεσματικά στο παρελθόν. Η πιο σημαντική βελτίωση στην ενεργειακή αποδοτικότητα των ευρωπαϊκών κτιρίων παρατηρήθηκε μετά το 1990 λόγω των αυστηρότερων οικοδομικών κανονισμών που εισήχθησαν σε πολλά κράτη μέλη στα μέσα της δεκαετίας του 1990. Ως αποτέλεσμα, τα κτίρια που κατασκευάστηκαν το 2002 καταναλώνουν 25% λιγότερη ενέργεια από αυτά που κατασκευάστηκαν το 1990. Ωστόσο, οι εκτιμήσεις των ετήσιων ποσοστών ανακαίνισης κτιρίων σε όλη την Ευρώπη είναι μόνο μεταξύ 0.5% και 2.5% του κτιριακού αποθέματος συμφωνα με το Buildings Performance Institute Europe (BPIE).

Βέβαια υπάρχει ακόμη δυνατότητα μείωσης της κατανάλωσης στον τριτογενή τομέα, συγκεκριμένα κυμαινόμενη μεταξύ 20% και 30% για το 2030 και έως 37% σε ένα τεχνικό σενάριο που λαμβάνει υπόψη προηγμένες και ακριβές τεχνολογίες. Τα τριτογενή κτίρια στην Ευρώπη αντιπροσωπεύουν το 1/4 του κτηριακού αποθέματος και η τελική τους κατανάλωση ενέργειας είναι τουλάχιστον 40% υψηλότερη από αυτή των κτιρίων κατοικιών με αποτέλεσμα, τα τριτογενή κτίρια να ευθύνονται για το 1/3 της κατανάλωσης ενέργειας. Στο πλαίσιο αυτό, θα πρέπει να πραγματοποιηθεί η συστηματική ανακαίνιση τριτογενών κτιρίων για την επίτευξη των στόχων ενεργειακής αποδοτικότητας.

Στον τομέα της ανακαίνισης, η ενεργειακή αποδοτικότητα περιλαμβάνει έναν δεσμευτικό στόχο: από το 2014, το 3% της συνολικής επιφάνειας των κτιρίων των εθνικών κυβερνήσεων θα πρέπει να ανακαινίζεται κάθε χρόνο, πράγμα που σημαίνει ότι ο δημόσιος τομέας θα γίνει πρωτοπόρος στην προώθηση των δράσεων ανακαίνισης. Αυτή η ενέργεια θα αποτελούσε την αρχή για την ανάπτυξη των εθνικών στρατηγικών ανακαίνισης που απαιτεί η ενεργειακή αποδοτικότητα. Το 2016, το Ευρωπαϊκό Κοινοβούλιο πρότεινε να επεκταθεί ο «στόχος 3%» σε όλα τα δημόσια κτίρια, δεδομένου ότι είναι ο τομέας με τις υψηλότερες δυνατότητες εξοικονόμησης ενέργειας.

Λίγες ερευνητικές εργασίες έχουν προσπαθήσει να παράσχουν μια πρόβλεψη ταξινόμησης[1] για την υποστήριξη της κοινότητας χρηματοδότησης για τον εντοπισμό μελλοντικών έργων ανακαίνισης [19]. Μια βιβλιογραφική ανασκόπηση για προσεγγίσεις αξιολόγησης έργων ενεργειακής αποδοτικότητας μπορεί να βρεθεί στο [65] όπου οι συγγραφείς κατέληξαν στο συμπέρασμα ότι οι οικονομικοί δείκτες συχνά παραβλέπονται ή καλύπτονται επιφανειακά όταν διερευνάται η ενεργειακή αποδοτικότητα. Σύμφωνα με την ίδια μελέτη, η περίοδος απόσβεσης και ο μετριασμός του κόστους είναι οι δύο

[1]Η πρόβλεψη μπορεί να είναι είτε δυαδική (επενδύστε έναντι μην επενδύσετε) είτε πολλαπλών κατηγοριών (π.χ. οι δυνατότητες της επένδυσης είναι χαμηλή, μεσαία ή υψηλή).

πιο συχνοί οικονομικοί στόχοι που εξετάζονται. Στα [59], [66] και [47] ο αναγνώστης μπορεί να βρει μελέτες που σχετίζονται με την ανάλυση βέλτιστου κόστους για την ενεργειακή ανακαίνιση κτιρίων κατοικιών στις Άλπεις, την Πορτογαλία και την Καταλονία, αντίστοιχα. Εκτός από το κόστος και την περίοδο απόσβεσης, ο οικονομικός κίνδυνος των έργων χρησιμοποιείται ως συμπληρωματικός στόχος [23]. Ωστόσο, σε αντίθεση με μια διαισθητική προσέγγιση, η οικονομική αποδοτικότητα υπάγεται στην ίδια την έννοια της ενεργειακής αποδοτικότητας, αφού η ενεργειακή αποδοτικότητα συνήθως αλληλεπιδρά με άλλες περιβαλλοντικές και κοινωνικές πτυχές [4]. Σε αυτό το πλαίσιο, στην παρούσα μελέτη συμπεριλαμβάνουμε επιπλέον πληροφορίες σχετικά με την τρέχουσα κατανάλωση ενέργειας και τα χαρακτηριστικά των κτιρίων (π.χ. έτος κατασκευής, περιοχή θέρμανσης και αριθμός διαμερισμάτων), καθώς και τις μειώσεις CO_2.

Μια άλλη καινοτομία της μελέτης μας είναι η χρήση της μηχανικής μάθησης και των μεθόδων μετα-μάθησης [40]. Οι περισσότερες ερευνητικές εργασίες αξιολόγησαν τις επενδύσεις χρησιμοποιώντας πλαίσια που βασίζονται σε ανάλυση απόφασης πολλαπλών κριτηρίων [56], αξιοποιώντας είτε διακριτές μεθόδους [50, 44] είτε συνεχή πολυαντικειμενική βελτιστοποίηση [1], - ανάλυση οφέλους (CBA) [45] και ανάλυση κόστους-αποτελεσματικότητας (CEA) [65, 22]. Ωστόσο, αυτά τα πλαίσια απαιτούν σε βάθος γνώση των οικονομικών, κοινωνικών και κλιματικών παραγόντων που επηρεάζουν κάθε επένδυση. Επιπλέον, η απόδοσή τους επηρεάζεται έντονα από τις παραδοχές που γίνονται από τις αντίστοιχες μεθόδους. Σε αυτό το πλαίσιο, οι μέθοδοι μηχανικής μάθησης θα μπορούσαν να χρησιμοποιηθούν ως εναλλακτικές προσεγγίσεις που βασίζονται σε δεδομένα στα παραδοσιακά πλαίσια αξιολόγησης έργων ενεργειακής αποδοτικότητας [68]. Μια πρόσφατη ανασκόπηση σχετικά με τις εφαρμογές μηχανικής μάθησης στον τομέα της ενεργειακής οικονομίας και χρηματοδότησης [26] έδειξε ότι οι μέθοδοι μηχανικής μάθησης χρησιμοποιούνται κυρίως στην πρόβλεψη τιμών αργού πετρελαίου και ηλεκτρικής ενέργειας. Παρά τη συνεχώς αυξανόμενη χρήση της θεωρίας μηχανικής μάθησης στον ενεργειακό τομέα, η ταξινόμηση των επενδύσεων με βάση την αναμενόμενη χρησιμότητά τους είναι ένας τομέας που θα μπορούσε να διερευνηθεί περαιτέρω.

2.3 Καινοτομία της προτεινόμενης μεθοδολογίας

Αυτή η ερευνητική εργασία επιχειρεί να αναπτύξει μεθόδους μηχανικής μάθησης και τεχνικές ανάλυσης δεδομένων στον τομέα των επενδύσεων ενεργειακής αποδοτικότητας. Στο [20], χρησιμοποιήθηκαν παραδοσιακές, στατιστικές μέθοδοι ταξινόμησης, όπως η τακτική logit, η τακτική probit και οι μέθοδοι ανάλυσης γραμμικής διάκρισης μαζί με περιορισμένο αριθμό μεθόδων μηχανικής μάθησης, όπως οι k-πλησιέστεροι γείτονες και μηχανές διανυσμάτων υποστήριξης. Από αυτή την άποψη, η παρούσα μελέτη εξετάζει επιπλέον το Gaussian Bayes, το XGBoost και τις μεθόδους τυχαίων

δασών (Random Forest) σε μια προσπάθεια να βελτιώσει την ακρίβεια των προβλέψεων που γίνονται μέσω μεθόδων μάθησης. Επιπλέον, το προτεινόμενο πλαίσιο χρησιμοποιεί ένα μετα-μαθητή με στόχο να εντοπίσει την πιο ακριβή μέθοδο ταξινόμησης για κάθε επένδυση με βάση τα ιδιαίτερα χαρακτηριστικά της, επιτρέποντας έτσι περαιτέρω βελτιστοποίηση και αύξηση της πιθανής προστιθέμενης αξίας από την προτεινόμενη διαδικασία ταξινόμησης. Η συμβολή αυτής της μελέτης μπορεί να συνοψιστεί στη χρήση ενός μοντέλου μετα-μάθησης με στόχο τη μείωση του εγγενούς κινδύνου χρήσης ενός μόνο ταξινομητή, ειδικά με μικρά ή μη ισορροπημένα σύνολα δεδομένων. Σε αντίθεση με προηγούμενες μελέτες που έχουν προτείνει προσαρμοσμένα μοντέλα για συγκεκριμένα σύνολα δεδομένων, αυτή η μελέτη παρέχει ένα γενικευμένο πλαίσιο που παράγει αξιόπιστες προβλέψεις όσον αφορά την ακρίβεια και την ευρωστία, ανεξάρτητα από τα επιλεγμένα χαρακτηριστικά και τον όγκο των διαθέσιμων δεδομένων.

Chapter 3

Προτεινόμενη Μεθοδολογία

Η μεθοδολογία που παρουσιάζεται φιλοδοξεί να παρέχει μια σύσταση χρηματοδότησης - επιχορήγησης για μελλοντικά έργα ενεργειακής αποδοτικότητας, μαθαίνοντας από μια ομάδα ήδη υλοποιημένων έργων με παρόμοια χαρακτηριστικά. Η βάση της προτεινόμενης προσέγγισης είναι ένα μοντέλο ταξινόμησης συνόλου στοίβαξης που βασίζεται σε πέντε βασικές μεθόδους ταξινόμησης μηχανικής μάθησης, και συγκεκριμένα *k-Nearest Neighbors*, *Gaussian Naive Bayes*, *Extreme Gradient Boosted Trees*, *Random Forest* και *Support Vector Machine*. Το σύνολο δεδομένων μάθησης είναι ουσιαστικά ένα σύνολο έργων που περιλαμβάνει πληροφορίες σχετικά με τις ανακαινίσεις που πραγματοποιήθηκαν σε διάφορα κτίρια στο παρελθόν. Αυτό περιλαμβάνει δεδομένα που ήταν αρχικά διαθέσιμα πριν από την υλοποίηση της ανακαίνισης (χρησιμοποιήθηκαν ως χαρακτηριστικά εισόδου) και δεδομένα μετά την ολοκλήρωση της επένδυσης. Στην πράξη, και για να επιτραπεί η αυτοματοποίηση, αυτές οι πληροφορίες θα πρέπει να αποθηκεύονται στη βάση δεδομένων του πληροφοριακού συστήματος που χρησιμοποιείται από το χρηματοδοτικό ίδρυμα για την αξιολόγηση μελλοντικών επενδύσεων.

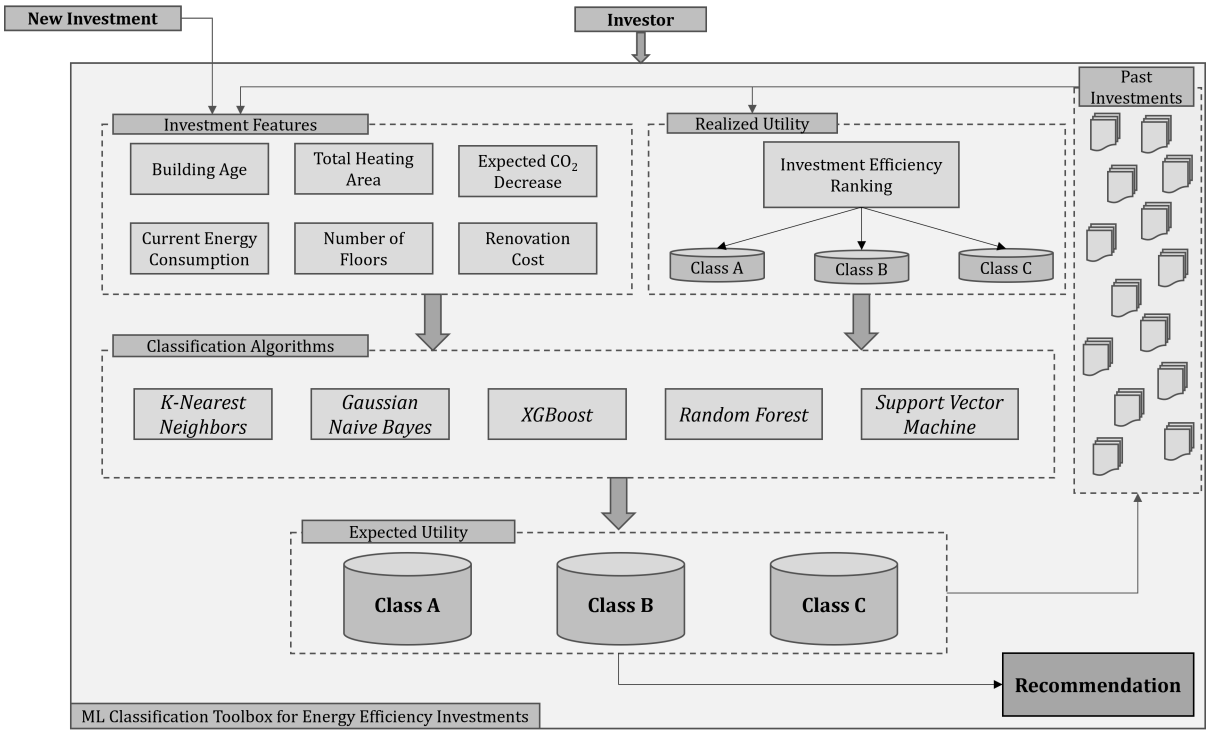
Συγκεκριμένα, τα χαρακτηριστικά της επένδυσης αφορούν συγκεκριμένα χαρακτηριστικά των έργων που, στην περίπτωσή μας, είναι η ηλικία του ανακαινισμένου κτιρίου, ο αριθμός των διαμερισμάτων, η συνολική επιφάνεια θέρμανσης και η τρέχουσα κατανάλωση ενέργειας, η αναμενόμενη μείωση CO_2, και το κόστος ανακαίνισης. Το αποτέλεσμα των ταξινομητών βάσης είναι οι πιθανότητες η επένδυση να ανήκει σε κάθε διακριτή κατηγορία μεταξύ των εξής τριών:

- *Class A*: The project should be financed.
- *Class B*: The project should be partially financed.
- *Class C*: The project should not be financed.

Προκειμένου να επισημανθούν οι προηγούμενες επενδύσεις, η πραγματοποιηθείσα χρησιμότητα κάθε έργου υπολογίζεται χρησιμοποιώντας ένα μέτρο επιλογής. Σε αυτή

τη μελέτη εξετάζουμε την κατάταξη της επενδυτικής αποδοτικότητας, όπως περιγράφεται στην ενότητα 4.1. Κατά συνέπεια, τα έργα με την καλύτερη απόδοση χαρακτηρίζονται ως *Κλάσης A*, τα έργα με τη χειρότερη απόδοση ως *Κλάσης C*, ενώ τα υπόλοιπα ως *Κλάσης B*. Έχοντας επισημάνει τις επενδύσεις, χρησιμοποιούνται οι βασικές μέθοδοι ταξινόμησης για την ταξινόμηση προηγούμενων έργων και αποθηκεύονται οι προβλέψεις τους. Χρησιμοποιώντας αυτές τις προβλέψεις ως στοιχεία εισόδου μαζί με τις πραγματικές ετικέτες των επενδύσεων και τα χαρακτηριστικά που χρησιμοποιήθηκαν αρχικά από τους ταξινομητές επιπέδου 0, ένας μετα-μαθητής, συγκεκριμένα ένας ταξινομητής λογιστικής παλινδρόμησης, εκπαιδεύεται με στόχο να προβλέψει ποια από τις πέντε βασικές μεθόδους είναι η καταλληλότερη για την πρόβλεψη της κλάσης κάθε έργου. Έτσι, ο βέλτιστος ταξινομητής χρησιμοποιείται για κάθε επένδυση για την εξαγωγή της πιθανότητας κάθε κατηγορίας. Αυτές οι πιθανότητες τελικά συνδυάζονται προκειμένου να δοθεί μια σύσταση σχετικά με το ποσοστό επιχορήγησης χρηματοδότησης για μια μελλοντική επένδυση, όπως περιγράφεται στην ενότητα 3.3. Η πλήρης διαδικασία σύστασης συνοψίζεται στο σχήμα 3.1.

Figure 3.1: Επισκόπηση της προτεινόμενης μεθοδολογίας. Η σύσταση για το ποσοστό χρηματοδότησης επιχορήγησης μελλοντικών επενδύσεων βασίζεται σε μια προσέγγιση βάσει δεδομένων που εκμεταλλεύεται χρήσιμες πληροφορίες από προηγούμενα έργα, με βάση την επιτευχθείσα μείωση της κατανάλωσης ενέργειας και το κόστος επένδυσης. Το αποτέλεσμα των βασικών μεθόδων ταξινόμησης, καθώς και του μεταμαθητή, παρέχει το ποσοστό της συνιστώμενης χρηματοδότησης επιχορήγησης.



Σημειώστε ότι το πλαίσιο της προτεινόμενης μεθοδολογίας είναι πολύ ευέλικτο

όσον αφορά τα επενδυτικά χαρακτηριστικά, τις κατηγορίες, τις μεθόδους και τα μέτρα χρησιμότητας που χρησιμοποιούνται. Για παράδειγμα, οι υπεύθυνοι λήψης αποφάσεων μπορούν να προσαρμόσουν τις μεταβλητές εισόδου που χρησιμοποιούνται από τους ταξινομητές ανάλογα με τις προτιμήσεις τους ή τη διαθεσιμότητα δεδομένων. Ομοίως, μπορούν να τροποποιήσουν το σύνολο των ταξινομητών επιπέδου 0, λαμβάνοντας υπόψη διαφορετικούς αριθμούς και τύπους μεθόδων. Τέλος, μπορούν να ορίσουν τον αριθμό των ετικετών που λαμβάνονται υπόψη από τις μεθόδους ταξινόμησης, καθώς και την προσέγγιση που χρησιμοποιείται για τον προσδιορισμό τους. Ως εκ τούτου, οι επιλογές που έγιναν στη μελέτη μας για τη δημιουργία του πλαισίου της περιγραφόμενης μεθοδολογίας και την επίδειξη της χρήσης της θα πρέπει να θεωρηθούν ενδεικτικές, αν και χρησιμοποιούν ένα σημαντικό σύνολο παραμέτρων.

Σημειώστε επίσης ότι η ευελιξία που προσφέρει το προτεινόμενο μοντέλο μπορεί γενικά να αντισταθμιστεί αποτελεσματικά από την άποψη της απόδοσης ταξινόμησης μέσω του μετα-μαθητή που χρησιμοποιείται. Ανάλογα με την αλγοριθμική τους φύση, διαφορετικές μέθοδοι ταξινόμησης μπορεί να είναι πιο κατάλληλες για το χειρισμό μικρών ή μη ισορροπημένων συνόλων δεδομένων, καθώς και για την πραγματοποίηση ακριβών προβλέψεων όταν παρέχονται πολλαπλά χαρακτηριστικά ως είσοδοι, ειδικά σε περιπτώσεις όπου δεν είναι όλα τα χαρακτηριστικά κρίσιμα ή ίδιας σημασίας. Για παράδειγμα, οι μέθοδοι που βασίζονται σε δέντρα αποφάσεων (π.χ. XGBoost και RF) μπορούν φυσικά να επιλέξουν τα πιο σχετικά χαρακτηριστικά για την πραγματοποίηση προβλέψεων με βάση τη συμβολή κάθε χαρακτηριστικού στη συνολική μείωση του σφάλματος πρόβλεψης κατά την προσαρμογή της μεθόδου. Επιπλέον, ορισμένες μέθοδοι (π.χ. RF) είναι ιδιαίτερα αποτελεσματικές στον χειρισμό πολλαπλών χαρακτηριστικών, ενώ άλλες στην αντιμετώπιση της τυχαιότητας δεδομένων. Αυτό έρχεται σε αντίθεση με μεθόδους όπως το k-NN και ο Gaussian naive Bayes που, αν και είναι πιο αποτελεσματικές με μικρά σύνολα δεδομένων, δεν μπορούν να επιλέξουν αυτόματα τα πιο σχετικά χαρακτηριστικά και, επομένως, μπορεί να επηρεαστούν αρνητικά όταν πολλές μεταβλητές παρέχονται ως είσοδος. Είναι ακριβώς ο σκοπός του μετα-μαθητή να λάβει υπόψη τέτοια ζητήματα και ιδιαιτερότητες, να μειώσει τους υποκείμενους κινδύνους της χρήσης ενός μόνο ταξινομητή για την αξιολόγηση μελλοντικών επενδύσεων και να παρέχει όσο το δυνατόν πιο αξιόπιστες προβλέψεις, δεδομένου του συνόλου δεδομένων που είναι διαθέσιμα για εκπαίδευση και των παραμέτρων του εξεταζόμενου προβλήματος ταξινόμησης.

3.1 Βασικές Μέθοδοι Ταξινόμησης

Η προτεινόμενη μεθοδολογία βασίζεται σε πέντε βασικές μεθόδους ταξινόμησης, που επιλέχθηκαν έτσι ώστε ο μετα-μαθητής να επιλέγει τον καταλληλότερο ταξινομητή

από ένα πολυποίκιλο [1] σύνολο μεθόδων, καθεμία από τις οποίες κάνει διαφορετικές υποθέσεις σχετικά με το πρόβλημα πρόβλεψης. Αυτή η προσέγγιση έχει αποδειχθεί αποτελεσματική για τη βελτίωση της ακρίβειας πρόβλεψης [60]. Οι βασικές μέθοδοι ταξινόμησης παρουσιάζονται σύντομα παρακάτω.

3.1.1 k-Nearest Neighbors

Οι k-πλησιέστεροι γείτονες (k-NN) είναι μια μη παραμετρική μέθοδος ταξινόμησης [15]. Η είσοδος της μεθόδου αποτελείται από n χαρακτηριστικά που χαρακτηρίζουν κάθε παρατήρηση (στην περίπτωσή μας επένδυση ενεργειακής αποδοτικότητας), ενώ η έξοδος είναι μια ετικέτα που καθορίζει την κατηγορία της (στην περίπτωσή μας την πραγματοποιηθείσα χρησιμότητα της επένδυσης). Ουσιαστικά, η μέθοδος ταξινομεί μελλοντικές μη επισημασμένες παρατηρήσεις προσδιορίζοντας τις k πιο παρόμοιες παρατηρήσεις με ετικέτα (πλησιέστεροι γείτονες) και λαμβάνοντας υπόψη την πληθώρα των κλάσεων τους.

Συγκεκριμένα, η μέθοδος αποτελείται από δύο φάσεις. τη φάση της εκπαίδευσης και τη φάση της αξιολόγησης. Κατά τη φάση της εκπαίδευσης, αποθηκεύονται οι σημειωμένες παρατηρήσεις, οι οποίες γενικά είναι διανύσματα σε έναν πολυδιάστατο χώρο, το καθένα αντιστοιχισμένο σε μια κλάση και ορίζεται ο αριθμός των γειτόνων k. Κατά τη φάση της αξιολόγησης, τα μη επισημασμένα διανύσματα ταξινομούνται αντιστοιχίζοντάς τα στην κλάση που εμφανίζεται συχνότερα μεταξύ των παρατηρήσεων με ετικέτα k που είναι πλησιέστερα στο δοκιμαστικό διάνυσμα. Ο όρος «πλησιέστεροι γείτονες» μπορεί να ερμηνευτεί και να υπολογιστεί χρησιμοποιώντας διάφορα μέτρα απόστασης ανάλογα με τον τύπο των χαρακτηριστικών που χρησιμοποιούνται ως είσοδοι. Για συνεχείς μεταβλητές η Ευκλείδεια απόσταση είναι το πιο συχνά χρησιμοποιούμενο μέτρο, ενώ για τις διακριτές μεταβλητές η απόσταση Hamming επικρατεί στη βιβλιογραφία [46]. Μια άλλη στρατηγική είναι η χρήση συντελεστών συσχέτισης, όπως η συσχέτιση Pearson [38]. Δεδομένου ότι η μέθοδος βασίζεται στον υπολογισμό των αποστάσεων μεταξύ των χαρακτηριστικών εισόδου, συνιστάται η κανονικοποίηση των τιμών των χαρακτηριστικών για τη βελτίωση της ακρίβειας [64].

Ο ορισμός του k, ουσιαστικά της μοναδικής υπερπαράμετρου του k-NN, είναι κρίσιμος γιατί επηρεάζει άμεσα την απόδοση της διαδικασίας επισήμανσης. Εάν το k έχει οριστεί ίσο με 1, τότε οι μη επισημασμένες παρατηρήσεις ταξινομούνται σύμφωνα με την ετικέτα του πλησιέστερου γείτονά τους. Εάν το k οριστεί ίσο με τον αριθμό των παρατηρήσεων που περιλαμβάνονται στο σύνολο εκπαίδευσης, τότε οι μη

[1] Η ποικιλομορφία αναφέρεται στη διαφορετική αλγοριθμική φύση των επιλεγμένων μεθόδων μηχανικής μάθησης και στα πλεονεκτήματα και τις αδυναμίες καθεμίας από αυτές Οι μέθοδοι αναμένεται να εμφανίζουν διαφορές όσον αφορά την ακρίβεια και την ευρωστία όταν πρόκειται για μικρά, μη ισορροπημένα ή θορυβώδη σύνολα δεδομένων, καθώς και στο χειρισμό πολλαπλών χαρακτηριστικών. Αυτές οι ιδιότητες συνοψίζονται στο εισαγωγικό μέρος της ενότητας 3 και περιγράφονται λεπτομερώς στις ακόλουθες ενότητες.

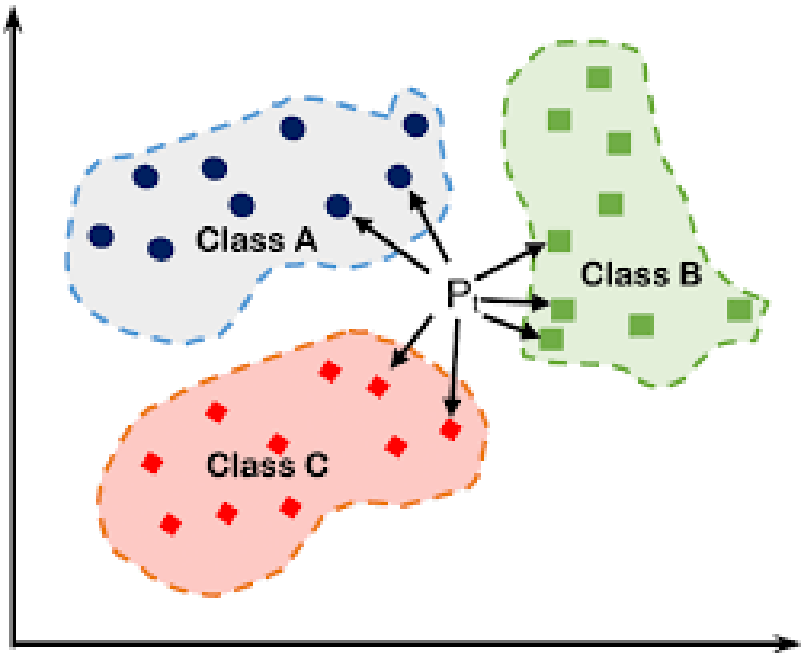
επισημασμένες παρατηρήσεις ταξινομούνται σύμφωνα με την πιο δημοφιλή ετικέτα στο πλήρες σύνολο εκπαίδευσης. Δεδομένου ότι η «βέλτιστη» τιμή του k συνήθως υπόκειται στις ιδιαιτερότητες του συνόλου εκπαίδευσης και του προβλήματος ταξινόμησης, ορίζεται συνήθως μέσω τεχνικών ευρετικής διασταυρούμενης επικύρωσης (cross validation), που χρησιμοποιούνται ευρέως στη βιβλιογραφία για την εύρεση βέλτιστων υπερπαραμέτρων.

Στο πλαίσιο αυτής της μελέτης, τα χαρακτηριστικά που εξετάζονται είναι συνεχείς μεταβλητές. Ως αποτέλεσμα, η Ευκλείδεια απόσταση χρησιμοποιήθηκε για τον προσδιορισμό των πλησιέστερων γειτόνων. Δεδομένων των καρτεσιανών συντεταγμένων δύο σημείων p και q σε έναν ευκλείδειο χώρο n-διάστατων, η απόσταση υπολογίζεται ως εξής:

$$d(p,q) = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + \cdots + (p_i - q_i)^2 + \cdots + (p_n - q_n)^2}. \quad (3.1)$$

Στο σχήμα 3.2 παρουσιάζεται ο τρόπος λειτουργίας του μοντέλου.

Figure 3.2: Ο τρόπος λειτουργίας του μοντέλου K-Nearest Neighbors [5].



3.1.2 Gaussian Naive Bayes

Η οικογένεια των πιθανοτικών ταξινομητών περιλαμβάνει ταξινομητές που είναι σε θέση να προβλέψουν την κατανομή πιθανοτήτων σε ένα διαθέσιμο σύνολο κλάσεων, αντί να προβλέπουν μία πρόβλεψη κλάσης [43]. Οι Naive Bayes ταξινομητές είναι μια υποκατηγορία πιθανοτικών ταξινομητών που βασίζονται στο θεώρημα του Bayes, υποθέτοντας ισχυρή ανεξαρτησία μεταξύ των χαρακτηριστικών. Το μοντέλο Gaussian naive Bayes έχει χρησιμοποιηθεί σε πολλούς τομείς, συμπεριλαμβανομένων προβλημάτων που

σχετίζονται με την αξιολόγηση της απόδοσης της επένδυσης [14], την πρόβλεψη της παραγωγής φωτοβολταϊκών [6] και την ανάλυση της ενεργειακής αποδοτικότητας των κτιρίων [53].

Σύμφωνα με τη μελέτη των [70], με δεδομένο ένα διάνυσμα χαρακτηριστικών x_1 έως x_n και μια μεταβλητή κλάσης y, η ακόλουθη σχέση δηλώνεται από το θεώρημα του Bayes.

$$p(y \mid x_1, \ldots, x_n) = \frac{p(y)\ p(x_1, \ldots, x_n \mid y)}{p(x_1, \ldots, x_n)}. \tag{3.2}$$

Χρησιμοποιώντας την απλή υπόθεση ανεξαρτησίας υπό όρους για όλα τα i, αυτή η σχέση μπορεί να μετατραπεί στην ακόλουθη εξίσωση:

$$p(y \mid x_1, \ldots, x_n) = \frac{p(y)\ \prod_{i=1}^{n} p(x_i \mid y)}{p(x_1, \ldots, x_n)}. \tag{3.3}$$

Ο κανόνας ταξινόμησης που δίνεται από την ακόλουθη εξίσωση προκύπτει επειδή το $p(x_1, \ldots, x_n)$ είναι σταθερό δεδομένης της εισόδου:

$$\hat{y} = \underset{y}{\operatorname{argmax}}\ p(y) \prod_{i=1}^{n} p(x_i \mid y). \tag{3.4}$$

Υπάρχουν διαφορετικοί απλοί ταξινομητές Bayes ανάλογα με τις υποθέσεις που γίνονται σχετικά με την κατανομή του $p(x_i \mid y)$. Το πρόβλημα που μελετάμαι, όμως, περιλαμβάνει συνεχή δεδομένα, επομένως μπορούμε να κάνουμε την υπόθεση ότι οι συνεχείς τιμές που σχετίζονται με κάθε Κλάση ακολουθούν την κατανομή Gauss [33]. Επομένως, χρησιμοποιείται το μοντέλο Gaussian naive Bayes, όπου η πιθανότητα των χαρακτηριστικών θεωρείται ότι ακολουθεί Gaussian κατανομή:

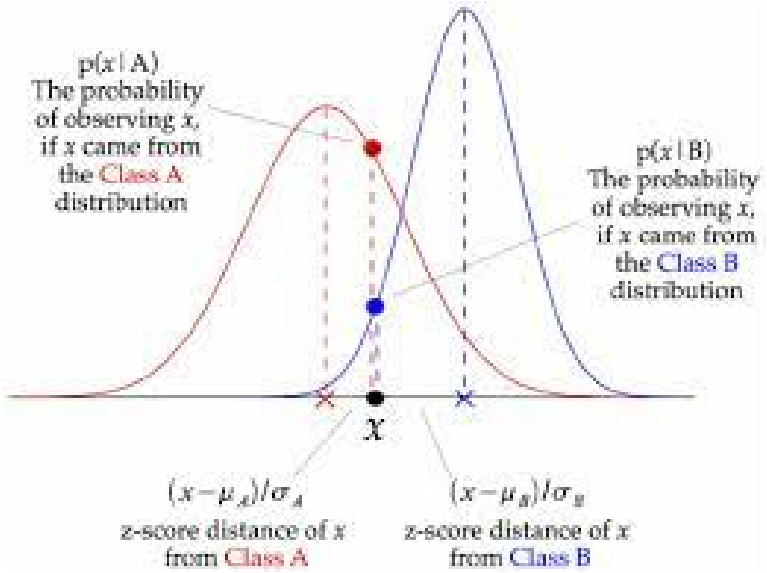
$$p(x_i \mid y) = \frac{1}{\sqrt{2\pi\sigma_y^2}}\, e^{-\frac{(x_i - \mu_y)^2}{2\sigma_y^2}}. \tag{3.5}$$

Στο σχήμα 3.3 παρουσιάζεται ο τρόπος λειτουργίας του μοντέλου.

3.1.3 Extreme Gradient Boosted Trees

Τα δέντρα ταξινόμησης είναι μια μέθοδος μηχανικής μάθησης που χωρίζει αναδρομικά τον χώρο δεδομένων μέσω κανόνων για να ταξινομήσει ένα σύνολο παρατηρήσεων [9]. Δεδομένου ότι οι κανόνες διαμερίσεων κατασκευάζονται διαδοχικά, το δέντρο ξεκινά από μια ρίζα, όπου περιλαμβάνεται ολόκληρο το σύνολο εκπαίδευσης, και χωρίζει τις διαθέσιμες παρατηρήσεις σε κλάδους, καθεμία από τις οποίες περιέχει τις παρατηρήσεις που ικανοποιούν τον πρώτο κανόνα του δέντρου. Στη συνέχεια, κάθε κλάδος μπορεί να διαχωριστεί περαιτέρω χρησιμοποιώντας πρόσθετους κανόνες. Τα κλαδιά που δεν χωρίζονται ονομάζονται φύλλα και περιλαμβάνουν τις τελικές προβλέψεις του δέντρου. Προκειμένου να καθοριστεί ποιο φύλλο πρέπει να χρησιμοποιηθεί για την πρόβλεψη κάθε παρατήρησης, πρέπει να ληφθούν υπόψη οι σύνδεσμοι κανόνων. Οι κανόνες

Figure 3.3: Ο τρόπος λειτουργίας του μοντέλου Gaussian Naive Bayes [54]



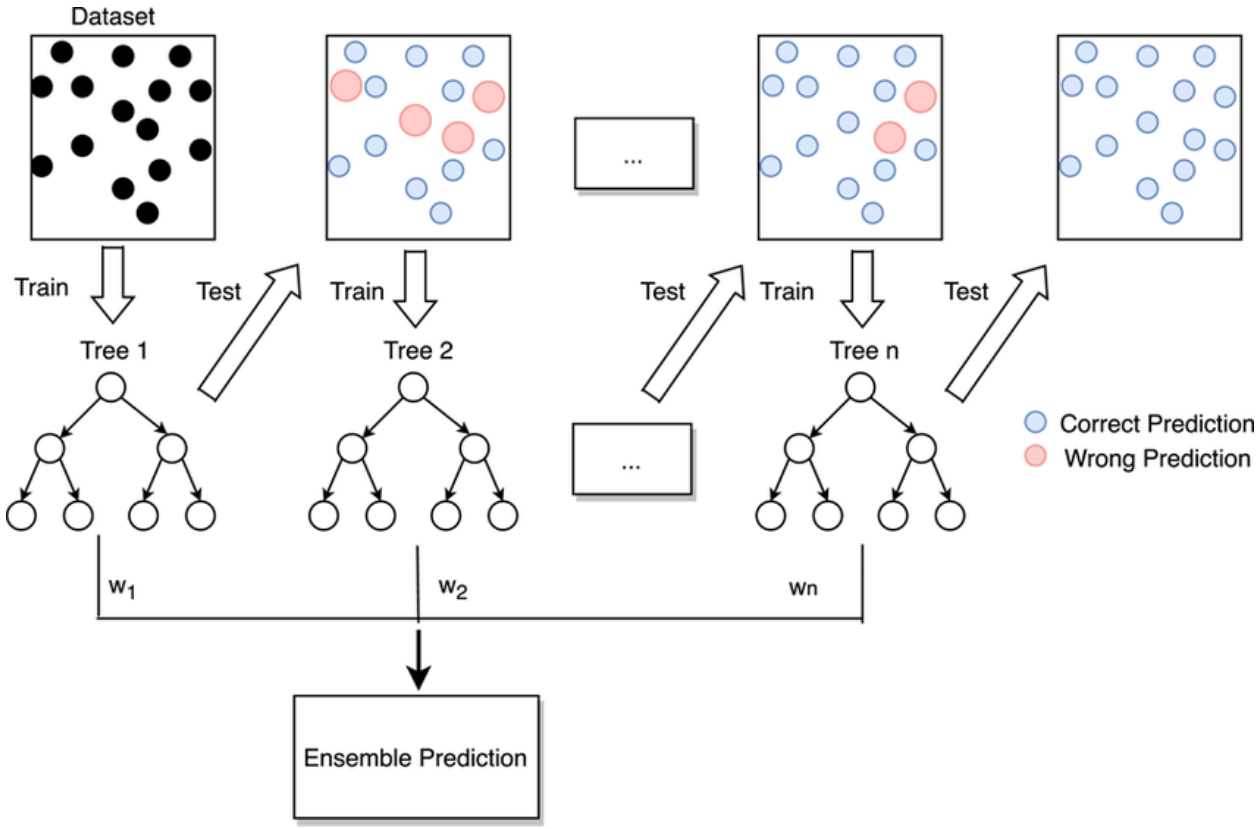
που κατασκευάζονται βασίζονται στα διαθέσιμα χαρακτηριστικά για την εκπαίδευση της μεθόδου και ορίζονται αυτόματα χρησιμοποιώντας διάφορα κριτήρια, όπως η πρόσμειξη Gini και το κέρδος πληροφοριών. Επίσης, τις περισσότερες φορές, πρόσθετα κριτήρια (π.χ. ρυθμός εκμάθησης, μέγιστος αριθμός φύλλων, μέγιστο βάθος δέντρου, ελάχιστο άθροισμα βάρους που απαιτείται σε ένα παιδί και ελάχιστη μείωση απώλειας που απαιτείται για να γίνει μια περαιτέρω κατάτμηση σε έναν κόμβο φύλλου του δέντρου) χρησιμοποιούνται με τη μορφή υπερπαραμέτρων για να καθοριστεί πότε πρέπει να τερματιστεί η ανάπτυξη του δέντρου, αποφεύγοντας έτσι την υπερπροσαρμογή και μείωση του υπολογιστικού κόστους.

Η ενίσχυση κλίσης (GB) είναι μια τεχνική που μπορεί να εξοπλιστεί με οποιαδήποτε μέθοδο μηχανικής μάθησης για να μειώσει την προκατάληψη και τη διακύμανση των προβλέψεών της και, ως εκ τούτου, να μετατρέψει τους αδύναμους μαθητές σε δυνατούς. Κατά συνέπεια, το GB συνδυάζει διαδοχικά πολλούς αδύναμους μαθητές για να δημιουργήσει έναν μόνο ισχυρό εκπαιδευόμενο με υψηλότερη ικανότητα εκμάθησης και βελτιωμένη ακρίβεια [24]. Αυτός ο συνδυασμός εκτελείται έτσι ώστε κάθε νέος αδύναμος μαθητής να εξειδικεύεται στη βελτίωση των προβλέψεων των προηγούμενων, δηλαδή στην ελαχιστοποίηση των προηγούμενων σφαλμάτων πρόβλεψης. Σε αυτό το πλαίσιο, το GB μπορεί επίσης να χρησιμοποιηθεί για να βελτιώσει την απόδοση των δέντρων ταξινόμησης και να δημιουργήσει πιο ισχυρούς ταξινομητές.

Η ακραία ενίσχυση κλίσης (XGBoost) είναι ίσως η πιο δημοφιλής εφαρμογή των δέντρων ενισχυμένων με κλίση [12]. Εκμεταλλεύεται πόρους μνήμης και υλικού για μεθόδους ενίσχυσης δέντρων, με αποτέλεσμα ανώτερη ταχύτητα και απόδοση [17]. Επίσης, το XGBoost ενσωματώνει τρεις πολύ γνωστές τεχνικές GB, τις gradient, regularized και stochastic boosting. Σε αυτή τη μελέτη, εφαρμόσαμε το XGBoost χρησιμοποιώντας την ενίσχυση κλίσης.

Στο σχήμα 3.4 παρουσιάζεται ο τρόπος λειτουργίας του μοντέλου.

Figure 3.4: Ο τρόπος λειτουργίας του μοντέλου Gradient Boosting [71]



3.1.4 Random Forest

Το Random Forest (RF) είναι μια μέθοδος ταξινόμησης συνόλου μηχανικής μάθησης που βασίζεται στην ανάπτυξη πολλαπλών δέντρων ταξινόμησης και στην εξέταση του πλήθους των ψήφων τους για να γίνει μια τελική πρόβλεψη [10]. Προκειμένου τα κατασκευασμένα δέντρα να εκμεταλλευτούν τα οφέλη του συνδυασμού, μειώνοντας τόσο την προκατάληψη όσο και τη διακύμανση των προβλέψεων, κάθε δέντρο ταξινόμησης εκπαιδεύεται σε διαφορετικές παρατηρήσεις και χρησιμοποιώντας διαφορετικό αριθμό χαρακτηριστικών, τυχαία δειγματοληψία από αυτά που ήταν αρχικά διαθέσιμα. Αυτή η τεχνική, ευρέως γνωστή ως bagging, διασφαλίζει ότι τα δέντρα που δημιουργούνται θα είναι πραγματικά διαφορετικά, θα έχουν πρόσβαση σε διαφορετικές πληροφορίες και θα είναι λιγότερο ευαίσθητα στις αλλαγές που γίνονται στο σύνολο εκπαίδευσης.

Η δυνατότητα του αλγορίθμου RF να έχει αυξημένη προβλεπτική ακρίβεια προέρχεται από το γεγονός πως αποτελεί μια προέκταση του απλού δέντρου αποφάσεων. Ειδικότερα, τα δέντρα που αναπτύσσονται πολύ βαθιά τείνουν να μαθαίνουν πολύ ακανόνιστα σχέδια: ταιριάζουν υπερβολικά στα σετ προπόνησής τους, δηλαδή έχουν

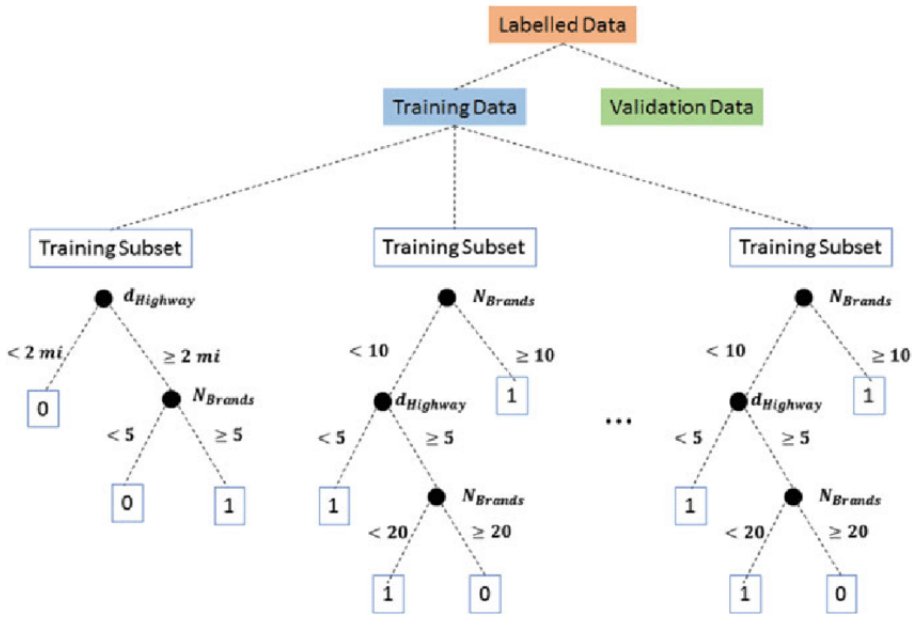
χαμηλή προκατάληψη, αλλά πολύ υψηλή διακύμανση. Τα τυχαία δάση είναι ένας τρόπος υπολογισμού του μέσου όρου πολλών δέντρων βαθιάς απόφασης, που εκπαιδεύονται σε διαφορετικά μέρη του ίδιου συνόλου εκπαίδευσης, με στόχο τη μείωση της διακύμανσης. Αυτό έρχεται σε βάρος μιας μικρής αύξησης της μεροληψίας και κάποιας απώλειας ερμηνείας, αλλά γενικά ενισχύει σημαντικά την απόδοση στο τελικό μοντέλο.

Τα δάση είναι σαν το τράβηγμα των προσπαθειών αλγορίθμου δέντρων αποφάσεων. Λαμβάνοντας την ομαδική εργασία πολλών δέντρων βελτιώνοντας έτσι την απόδοση ενός μόνο τυχαίου δέντρου. Αν και δεν είναι αρκετά παρόμοια, τα δάση δίνουν τα αποτελέσματα μιας διασταυρούμενης επικύρωσης K-fold.

Στον τομέα της ενέργειας και των κτιρίων, πολλές μελέτες έχουν χρησιμοποιήσει RF για να προβλέψουν την κατανάλωση ενέργειας σε κτίρια [3], να αναπτύξουν εξατομικευμένα συστήματα κλιματισμού στα γραφεία [39] και να εντοπίσουν σφάλματα στα δίκτυα διανομής [11], μεταξύ άλλων .

Στο σχήμα 3.5 παρουσιάζεται ο τρόπος λειτουργίας του μοντέλου.

Figure 3.5: Ο τρόπος λειτουργίας του μοντέλου Random Forest [25]



3.1.5 Support Vector Machines

Οι μηχανές υποστήριξης διανυσμάτων (SVM) βασίζονται στην κατασκευή ενός υπερ-επίπεδου ή ενός συνόλου υπερ-επίπεδων σε χώρο υψηλών ή άπειρων διαστάσεων, το οποίο, μεταξύ άλλων, μπορεί να χρησιμοποιηθεί για την επίλυση προβλημάτων ταξινόμησης δυαδικών και πολλαπλών κλάσεων [58] . Το υπερ-επίπεδο με το μεγαλύτερο λειτουργικό περιθώριο (απόσταση από τα πλησιέστερα σημεία δεδομένων του συνόλου εκπαίδευσης) επιτυγχάνει χαμηλότερα σφάλματα γενίκευσης και επομένως παρέχει καλύτερο διαχωρισμό των δεδομένων. Σε τέτοιες ρυθμίσεις, τα SVM συνήθως ονομάζονται ταξινομητές διανυσμάτων υποστήριξης (SVC) [42, 7].

Δεδομένων των διανυσμάτων $x_i \in \mathbb{R}^p,\ i = 1, \ldots, n$ και ενός διανύσματος $y \in \{1, -1\}$, το SVC υπολογίζει τις τιμές $w \in \mathbb{R}^p$ και $b \in \mathbb{R}$ προκειμένου να δημιουργηθούν σωστές προβλέψεις για οποιαδήποτε παρατήρηση χρησιμοποιώντας την τιμή του $\text{sign}(w^T\phi(x)+b)$. Το πρωταρχικό πρόβλημα που επιλύεται από τα SVC μπορεί να διατυπωθεί ως εξής:

$$\begin{aligned} \min_{w,b,z} \quad & \frac{1}{2}w^T w + C\sum_{i=1}^{N} z_i, \\ \text{s.t.} \quad & y_i(w^T\phi(x_i) + b) \geq 1 - z_i, \\ & z_i \geq 0. \end{aligned} \tag{3.6}$$

Αυτός ο μαθηματικός τύπος περιγράφει ότι τα SVC προσπαθούν να μεγιστοποιήσουν το περιθώριο, ενώ επίσης επιβάλλουν ποινή C για παρατηρήσεις που ταξινομούνται εσφαλμένα ή εμπίπτουν στο όριο του περιθωρίου. Τέλος, ο παράγοντας απόστασης z_i για κάθε διάνυσμα X_i επιτρέπει σε μια παρατήρηση να βρίσκεται σε μια ορισμένη απόσταση από το σωστό όριο περιθωρίου, επειδή η πλειονότητα των προβλημάτων δεν διαχωρίζονται από ένα υπερεπίπεδο.

Για ταξινόμηση πολλαπλών κλάσεων, η ρύθμιση προβλήματος SVC διαφέρει σημαντικά από την προσέγγιση της δυαδικής ταξινόμησης [28, 18]. Υπάρχουν δύο προσεγγίσεις που χρησιμοποιούνται για τη γενίκευση από τη δυαδική ταξινόμηση στην ταξινόμηση πολλαπλών κλάσεων. Η άμεση μέθοδος προτείνει ότι ένας δυαδικός ταξινομητής γενικεύεται προκειμένου να δημιουργηθούν προβλέψεις πολλαπλών κλάσεων, ενώ η πιο κοινή προσέγγιση προτείνει το συνδυασμό N ανεξάρτητων δυαδικών ταξινομητών για την παραγωγή ενός μόνο διανύσματος πολλαπλών κλάσεων. Στην τελευταία προσέγγιση, τα δυαδικά προβλήματα μπορούν να οριστούν με τη μορφή ενός πίνακα μεγέθους $M \times N$, όπου το M αντιπροσωπεύει τον αριθμό των κλάσεων και το N τον αριθμό των προβλημάτων. Οι τιμές κάθε στοιχείου του πίνακα $R_{i,j} \in \{-1, 0, 1\}$ που δημιουργούνται από τους δυαδικούς ταξινομητές f χρησιμοποιούνται για την πρόβλεψη της ετικέτας κλάσης σύμφωνα με την ακόλουθη εξίσωση:

$$\hat{y} = \arg\min_{y \in 1,\ldots,n} \Big\{ \sum_{k=1}^{N} R_{yk} f^k(x) \Big\}. \tag{3.7}$$

3.2 Μοντέλο Stacking Ensemble

Το σύνολο στοίβαξης (ή στοίβαξης) είναι ένα σχήμα που χρησιμοποιείται για την ελαχιστοποίηση του σφάλματος γενίκευσης πολλαπλών μεθόδων πρόβλεψης που εξετάζονται σε ένα σύνολο [67]. Συνήθως, τα σύνολα στοίβαξης χρησιμοποιούν μεταμαθητές που είναι υπεύθυνοι για το συνδυασμό των προβλέψεων από πολλαπλές βασικές μεθόδους μηχανικής μάθησης με βέλτιστο τρόπο ή απλώς για την επιλογή της πρόβλεψης που αναμένεται να είναι η πιο ακριβής για μια συγκεκριμένη περίπτωση. Κατά συνέπεια, αυτή η προσέγγιση μπορεί να εκμεταλλευτεί διάφορες μεθόδους με καλή

απόδοση που εμφανίζουν διαφορετικά χαρακτηριστικά, έχουν διαφορετικές δομές και κάνουν διαφορετικές υποθέσεις σχετικά με τα δεδομένα.

Η χρήση της μοντελοποίησης στοιβαγμένων συνόλων εμπνεύστηκε με αφορμή τη χαμηλή συσχέτιση των σφαλμάτων πρόβλεψης που συνήθως γίνονται με διαφορετικές μεθόδους βάσης. Ως αποτέλεσμα, αυτή η τεχνική είναι κατάλληλη όταν τα σφάλματα πρόβλεψης των baseline ταξινομητών δεν συσχετίζονται, γεγονός που σημαίνει ότι κάθε μέθοδος καταγράφει διαφορετικά χαρακτηριστικά του προβλήματος και, ως εκ τούτου, είναι πιο επιδέξια σε διαφορετικές προοπτικές. Όταν ικανοποιείται αυτή η υπόθεση, τα σύνολα αυτά δημιουργούν ακριβέστερες προβλέψεις από οποιαδήποτε μεμονωμένη μέθοδο, μειώνοντας τις προκαταλήψεις των επιμέρους μεθόδων.

Το ανεπτυγμένο μοντέλο στοιβαγμένων συνόλων έχει μια απλή αρχιτεκτονική, που αποτελείται από δύο επίπεδα μεθόδων ταξινόμησης. Το πρώτο επίπεδο περιλαμβάνει τους πέντε baseline ταξινομητές που περιγράφονται στην ενότητα 3.1, που ονομάζονται μέθοδοι «επιπέδου-0» ("level-0"). Η επιλεγμένη μέθοδος μετα-εκμάθησης, που θα ονομάζεται μέθοδος «επίπεδο-1» ("level-1"), είναι μια μέθοδος λογιστικής παλινδρόμησης. Η επιλογή μας βασίστηκε στην ερμηνευσημότητα που προσφέρει αυτή η μέθοδος, παρέχοντας μια απλή εξήγηση για τις προβλέψεις της, δηλαδή τους παράγοντες που οδήγησαν τον μετα-μαθητή να επιλέξει ένα baseline ταξινομητή έναντι άλλων. Αρκετές μελέτες έχουν προτείνει τη χρήση σχετικά απλών ταξινομητών ως μοντέλων επιπέδου 1 για τη βέλτιστη ενσωμάτωση των προβλέψεων των βασικών μεθόδων [13, 51]. Η συνολική αρχιτεκτονική του στοιβαγμένου μοντέλου γενίκευσης συνοψίζεται στο Σχήμα 3.6.

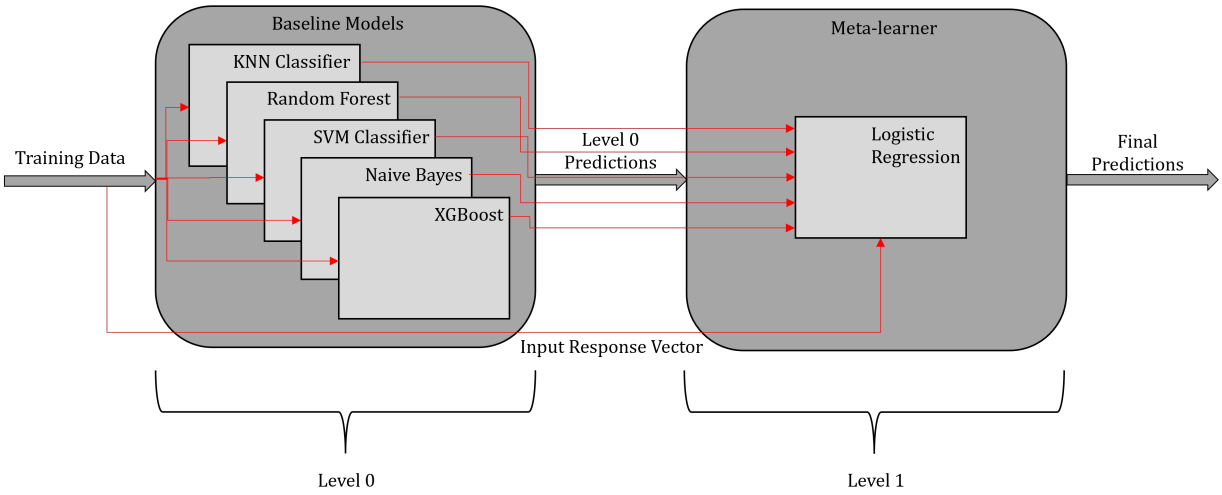
Η συνολική διαδικασία δημιουργίας μοντέλου στοίβαξης μπορεί να χωριστεί σε τρεις φάσεις. το στήσιμο του συνόλου, την εκπαίδευση και την πρόβλεψη για νέα δεδομένα. Αυτές οι φάσεις περιγράφονται παρακάτω:

Phase 1 - Ensemble Set Up: Σε αυτή τη φάση, επιλέγονται οι baseline ταξινομητές και ο ταξινομητής μετα-μάθησης. Οι βασικές μέθοδοι θα πρέπει να ενσωματώνουν διαφορετικά χαρακτηριστικά και να προέρχονται από διαφορετική κατηγορία αλγορίθμων (π.χ. δεν πρέπει να περιλαμβάνουν μόνο μεθόδους που βασίζονται σε δέντρα). Ο μετα-μαθητής θα πρέπει να είναι ένας σχετικά απλός και ερμηνεύσιμος ταξινομητής.

Phase 2 - Training: Αρχικά, καθεμία από τις βασικές μεθόδους εκπαιδεύεται χρησιμοποιώντας ένα κατάλληλο σύνολο υπερπαραμέτρων, που ορίζονται μέσω μιας διαδικασίας διασταυρούμενης επικύρωσης (π.χ. διασταυρούμενη επικύρωση k-fold). Στη συνέχεια, ο μετα-μαθητής εκπαιδεύεται χρησιμοποιώντας τις προβλέψεις και τα διανύσματα εισόδου των μεθόδων επιπέδου-0 ως είσοδο και τις πραγματικές ετικέτες των παρατηρήσεων ως έξοδο.

Phase 3 - Prediction: Το στοιβαγμένο σύνολο χρησιμοποιείται για τη δημιουργία προβλέψεων στα δεδομένα αξιολόγησης. Για να γίνει αυτό, κάθε βασικός ταξινομητής χρησιμοποιείται πρώτα για να κάνει μια μεμονωμένη πρόβλεψη. Στη συνέχεια, αυτές οι προβλέψεις τροφοδοτούνται στον μετα-μαθητή για να καθορίσει, με βάση το διάνυσμα

Figure 3.6: Επισκόπηση του προτεινόμενου μοντέλου στοιβαγμένων συνόλων. Οι βασικές μέθοδοι ταξινόμησης (επίπεδο-0) εκπαιδεύονται ακολουθώντας τη διαδικασία που περιγράφεται στο Σχήμα 3.1 και παρέχουν προβλέψεις για μελλοντικές επενδύσεις. Αυτές οι προβλέψεις στη συνέχεια τροφοδοτούνται στον μετα-μαθητή μαζί με τα διανύσματα εισόδου των baseline μεθόδων. Η έξοδος του μετα-μαθητή είναι η τελική πρόβλεψη του μοντέλου του στοιβαγμένου συνόλου, δηλαδή η πρόβλεψη του καταλληλότερου baseline ταξινομητή.



εισόδου του, ποιες βασικές προβλέψεις θα πρέπει να χρησιμοποιηθούν.

3.3 Σύσταση

Η προτεινόμενη μεθοδολογία φιλοδοξεί να παρέχει συστάσεις σχετικά με το ποσοστό επιχορήγησης που θα πρέπει να λάβουν οι μελλοντικές επενδύσεις. Εάν η προβλεπόμενη κλάση του μετα-μαθητή χρησιμοποιούταν για τον προσδιορισμό αυτού του ποσοστού, τότε οι προτάσεις του συστήματος υποστήριξης αποφάσεων θα ήταν ουσιαστικά διακριτοί αριθμοί, παίρνοντας ένα από τα τρία πιθανά ποσοστά που αντιστοιχίζονται σε κάθε κλάση. Για παράδειγμα, αν υποθέσουμε ότι η *Κλάση A* πρότεινε 100% χρηματοδότηση, η *Κλάση B* 50% χρηματοδότηση και η *Κλάση C* 0% χρηματοδότηση, τότε όλα τα έργα θα χρηματοδοτούνταν από συντελεστής $f_A = 1$, $f_B = 0,5$ ή $f_C = 0$, αντίστοιχα. Γίνεται προφανές ότι αυτή η προσέγγιση περιορίζει σημαντικά την προστιθέμενη αξία από τη μεθοδολογία, αγνοώντας επίσης την αβεβαιότητα γύρω από τις προβλέψεις. Για παράδειγμα, υποθέστε ένα σενάριο όπου μια επένδυση επισημαίνεται με πιθανότητα $0,55$ στην *Κλάση A*, $0,35$ στην *Κλάση B* και $0,10$ στην *Κλάση C*. Χρησιμοποιώντας την προαναφερθείσα προσέγγιση, το έργο θα χρηματοδοτηθεί κατά 100%, αν και υπάρχουν ενδείξεις κατά αυτής της απόφασης.

Για να μετριαστούν αυτά τα ζητήματα και να ληφθεί υπόψη η αβεβαιότητα, προτείνουμε τη δημιουργία συστάσεων χρηματοδότησης σχετικά με την πιθανότητα κάθε

κατηγορίας αντί για την ίδια την κυρίαρχη κατηγορία, συνοψίζοντας την πιθανότητα μια μελλοντική επένδυση να ανήκει σε κάθε κατηγορία. Λαμβάνοντας υπόψη τις μεταβλητές p_A, p_B και p_C, οι οποίες αντιπροσωπεύουν την πιθανότητα ότι μια επένδυση ανήκει στις κατηγορίες A, B και C, αντίστοιχα, όπως ορίζεται από το μοντέλο μετα-μάθησης καθώς και τους παράγοντες χρηματοδότησης επιχορήγησης f_A, f_B, f_C, ο τύπος χρηματοδότησης μιας μελλοντικής επένδυσης δίνεται από την ακόλουθη εξίσωση:

$$\text{Financing Percentage} = \sum_{i \in A,B,C} f_i \times p_i. \tag{3.8}$$

Σύμφωνα με την παραπάνω εξίσωση, στο προηγούμενο παράδειγμά μας το σύστημα υποστήριξης αποφάσεων θα συνιστούσε τη χρηματοδότηση του νέου έργου κατά 0,55 $\times 1,00 + 0,35 \times 0,50 + 0,10 * 0,00 = 72,5\%$, αντανακλώντας έτσι αποτελεσματικά τις γνώσεις των προτεινόμενων δεδομένων.

Chapter 4

Πειραματική Εφαρμογή και Αποτελέσματα

Αυτή η ενότητα παρέχει μια εκτενή πειραματική εφαρμογή της προτεινόμενης μεθοδολογίας σε δεδομένα που προέρχονται από πραγματικά έργα ανακαίνισης. Η διαδικασία εκπαίδευσης των πέντε baseline μεθόδων ταξινόμησης περιγράφεται λεπτομερώς, συμπεριλαμβανομένης της εύρεσης των βέλτιστων υπερπαραμέτρων τους, ακολουθούμενη από την εκπαίδευση του μοντέλου μετα-μάθησης. Τέλος, παρουσιάζονται και συζητούνται τα αποτελέσματα της εφαρμογής.

4.1 Σύνολο Δεδομένων

Η πειραματική εφαρμογή της προτεινόμενης μεθοδολογίας έχει εφαρμοστεί σε δεδομένα που συλλέχθηκαν από το Λετονικό Ταμείο Περιβαλλοντικών Επενδύσεων (LEIF), έναν οργανισμό που ανήκει στο Υπουργείο Περιβαλλοντικής Προστασίας και Περιφερειακής Ανάπτυξης της Λετονίας, που ιδρύθηκε το 1997. Από το 2009, το LEIF εποπτεύει παρακολούθηση της υλοποίησης και μετά την υλοποίηση πολλών έργων που συγχρηματοδοτούνται από χρηματοδοτικά μέσα για την κλιματική αλλαγή, αποτελώντας έτσι το μόνο ίδρυμα στη Λετονία που διαθέτει αξιόπιστα στοιχεία για την πραγματική απόδοση διαφόρων επενδύσεων όσον αφορά την εξοικονόμηση ενέργειας.

Αν και η προσέγγισή μας βασίζεται στην αξιολόγηση επενδύσεων σε ένα κτίριο, πολλά από τα αρχεία που συλλέγονται αφορούσαν επενδύσεις που εφαρμόστηκαν σε περισσότερα από ένα κτίρια. Επομένως, αυτά τα αρχεία εξαιρέθηκαν από το σύνολο δεδομένων, συμπεριλαμβανομένου ενός τελικού δείγματος 312 έργων ενεργειακής αποδοτικότητας. Οι επιλεγμένες μέθοδοι μηχανικής μάθησης εκπαιδεύτηκαν χρησιμοποιώντας ένα τυχαία επιλεγμένο 80% του συνολικού δείγματος, δηλαδή 249 επενδύσεις, ενώ οι υπόλοιπες 63 επενδύσεις χρησιμοποιήθηκαν για τον έλεγχο της απόδοσης της μεθοδολογίας μας.

Οι προαναφερθείσες επενδύσεις ταξινομήθηκαν με χρήση τριών ετικετών:

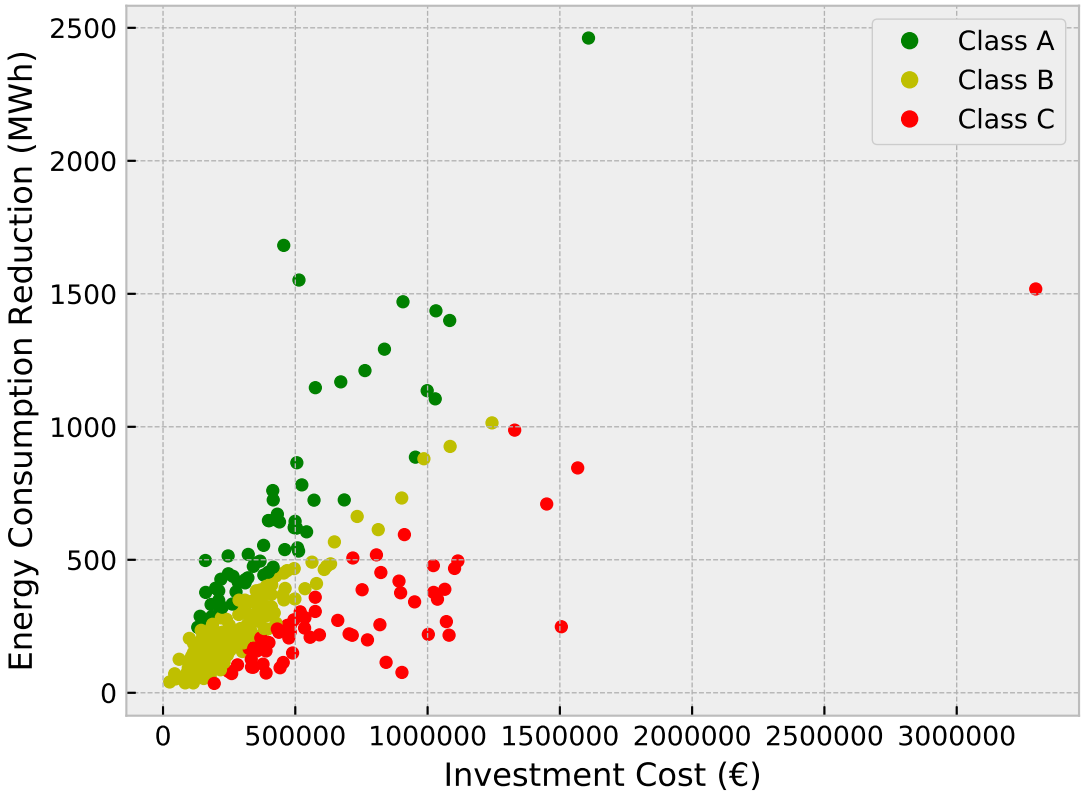
- *Κλάση A* η οποία συμπεριλαμβάνει επενδύσεις υψηλών δυνατοτήτων
- *Κλάση B* η οποία συμπεριλαμβάνει επενδύσεις μεσαίου δυναμικού
- *Κλάση C* η οποία συμπεριλαμβάνει επενδύσεις χαμηλού δυναμικού

Η επισήμανση των επενδύσεων βασίστηκε στην κατάταξη επενδυτικής αποδοτικότητας. Αυτή είναι μια απλή μέθοδος κατάταξης με δύο κριτήρια: το επενδυτικό κόστος του έργου και η μείωση της κατανάλωσης ενέργειας στο κτίριο μετά την ανακαίνιση. Αυτά τα κριτήρια συνδυάστηκαν χρησιμοποιώντας τα ακόλουθα βήματα: Πρώτον, οι τιμές για κάθε κριτήριο κανονικοποιήθηκαν σε μια κλίμακα $(0, 1)$. Δεύτερον, ο σταθμισμένος μέσος όρος των κανονικοποιημένων αξιών υπολογίστηκε για όλες τις διαθέσιμες επενδύσεις. Αυτή η διαδικασία πραγματοποιήθηκε χρησιμοποιώντας ίσα βάρη για τα επιλεγμένα κριτήρια. Ωστόσο, η επιλογή των συντελεστών στάθμισης μπορεί να προσαρμοστεί ανάλογα με τις προτιμήσεις του επενδυτή. Τέλος, οι τιμές ταξινομήθηκαν από την καλύτερη προς τη χειρότερη. Το κορυφαίο 20% των έργων επισημάνθηκαν ως επενδύσεις *Κλάσης A*, ενώ το χαμηλότερο 20% των έργων χαρακτηρίστηκαν ως επενδύσεις *Κλάσης C*. Το υπόλοιπο 60% των επενδύσεων χαρακτηρίστηκε ως *Κλάση B*. Μια αναπαράσταση του κόστους των επενδύσεων σε σχέση με την επιτευχθείσα μείωση της κατανάλωσης ενέργειας φαίνεται στο Σχήμα 4.1. Τα έργα με πράσινο χρώμα είναι επενδύσεις *Κατηγορίας A*, τα έργα με κίτρινο χρώμα είναι επενδύσεις *Κλάσης B*, ενώ τα έργα με κόκκινο χρώμα είναι επενδύσεις *Κλάσης Γ*.

Παρατηρήστε ότι, από τη σχεδίασή του, το σύστημα σήμανσης που χρησιμοποιείται έχει ως αποτέλεσμα ένα μη ισορροπημένο σύνολο εκπαίδευσης (προκατειλημμένη δειγματοληψία), με τις επενδύσεις *Κλάση B* να έχουν μεγαλύτερη πιθανότητα να αναγνωριστούν ως η σωστή κατηγορία. Παρόλο που αυτό το ζήτημα μπορεί να αντιμετωπιστεί χρησιμοποιώντας τεχνικές επαναδειγματοληψίας, όπως μεθόδους που υποδεικνύουν τυχαία την πλειοψηφική κλάση (*Κλάση B*) ή δημιουργούν συνθετικά δεδομένα για τις τάξεις μειοψηφίας (*Κλάση A* και *Κλάση C*[37]), ο μετα-μαθητής θα πρέπει ακόμα να μπορεί να αντιμετωπίσει την ανισορροπία κλάσης επιλέγοντας τη baseline μέθοδο ταξινόμησης που προβλέπει καλύτερα κάθε κατηγορία επένδυσης. Επιπλέον, δεδομένου ότι τα μη ισορροπημένα σύνολα δεδομένων είναι τυπικά στην πράξη, με ορισμένες κατηγορίες να παρατηρούνται φυσικά πιο συχνά από άλλες, αυτό το είδος ανισορροπίας χρησιμεύει ως πρόσθετο stress-test για την αξιολόγηση της προστιθέμενης αξίας από την προτεινόμενη προσέγγιση.

Τα επενδυτικά χαρακτηριστικά που χρησιμοποιήθηκαν στη μελέτη περίπτωσης μας υπόκεινται εν μέρει στη διαθεσιμότητα δεδομένων. Από αυτή την άποψη, επιλέξαμε έξι βασικά χαρακτηριστικά ως στοιχεία εισόδου στις μεθόδους ταξινόμησης, συμπεριλαμβανομένου του έτους κατασκευής του κτιρίου, της συνολικής επιφάνειας θέρμανσης (m^2), της αναμενόμενης μείωσης CO_2 λόγω δράσεις ανακαίνισης ($kgCO_2$), την τρέχουσα κατανάλωση ενέργειας του κτιρίου (MWh), τον αριθμό των ορόφων

Figure 4.1: Επισήμανση των εξεταζόμενων επενδύσεων με βάση την κατάταξη επενδυτικής αποδοτικότητας.



του κτιρίου που θα ανακαινιστεί και το ποσό της ζητούμενης επιχορήγησης χρηματοδότησης για την ανακαίνιση. Δεδομένου ότι όλες οι επενδύσεις έγιναν στην ίδια χώρα, δεν συμπεριλάβαμε γεωγραφικές μεταβλητές (π.χ. αναγνωριστικά πόλης) ως πρόσθετο χαρακτηριστικό. Οι σχέσεις ανά ζεύγη μεταξύ των επιλεγμένων επενδυτικών χαρακτηριστικών φαίνονται στο Σχήμα 4.2.

4.2 Εκπαίδευση και βελτιστοποίηση υπερπαραμέτρων

Στη μηχανική μάθηση, η βελτιστοποίηση υπερπαραμέτρων είναι το πρόβλημα της επιλογής ενός συνόλου βέλτιστων υπερπαραμέτρων για έναν αλγόριθμο εκμάθησης. Μια υπερπαράμετρος είναι μια παράμετρος της οποίας η τιμή χρησιμοποιείται για τον έλεγχο της μαθησιακής διαδικασίας. Αντίθετα, μαθαίνονται οι τιμές άλλων παραμέτρων (συνήθως βάρη κόμβων).

Το ίδιο είδος μοντέλου μηχανικής εκμάθησης μπορεί να απαιτεί διαφορετικούς περιορισμούς, βάρη ή ρυθμούς εκμάθησης για τη γενίκευση διαφορετικών μοτίβων δε-

Figure 4.2: Ζεύγη των μεταβλητών επένδυσης που λαμβάνονται υπόψη από τις προτεινόμενες μεθόδους ταξινόμησης στην εξεταζόμενη περίπτωση: Κόστος (€), Έτος κατασκευής κτιρίου, Προγραμματισμένη μείωση CO_2 ($kgCO_2$), Κατανάλωση ενέργειας πριν (MWh), Συνολική Θέρμανση (m^2), Δάπεδα

δομένων. Αυτά τα μέτρα ονομάζονται υπερπαράμετροι και πρέπει να ρυθμιστούν έτσι ώστε το μοντέλο να μπορεί να λύσει βέλτιστα το πρόβλημα μηχανικής εκμάθησης. Η βελτιστοποίηση υπερπαραμέτρων βρίσκει μια πλειάδα υπερπαραμέτρων που αποδίδει ένα βέλτιστο μοντέλο που ελαχιστοποιεί μια προκαθορισμένη συνάρτηση απώλειας σε δεδομένα ανεξάρτητα δεδομένα. Η αντικειμενική συνάρτηση παίρνει μια πλειάδα υπερπαραμέτρων και επιστρέφει τη σχετική απώλεια. Η διασταυρούμενη επικύρωση χρησιμοποιείται συχνά για την εκτίμηση αυτής της απόδοσης γενίκευσης.

Οι βέλτιστες υπερπαράμετροι για τις πέντε βασικές μεθόδους ταξινόμησης διαμορφώθηκαν μέσω μιας στρωματοποιημένης διαδικασίας διασταυρούμενης επικύρωσης k-fold. Ο όρος στρωματοποιημένος (stratified) υποδηλώνει μια παραλλαγή της παραδοσιακής διαδικασίας διασταυρούμενης επικύρωσης k-fold, όπου οι πτυχώσεις γίνονται διατηρώντας το ποσοστό των δειγμάτων για κάθε κατηγορία. Ο λόγος που το στρωματοποιημένο k-fold προτιμήθηκε έναντι του τυπικού του αντίστοιχου είναι ότι αντιμετωπίζουμε ένα πρόβλημα ταξινόμησης με ανισόρροπες κατανομές κλάσεων, που αποτελείται από 20% επενδύσεις *Κλάσης Α*, 60% επενδύσεις *Κλάσης Β*, και 20% επενδύσεις *Κλάσης Γ*. Ο αριθμός των διαχωρισμών ορίστηκε σε 10, ενώ ο αριθμός των επαναλήψεων για το στρωματοποιημένο k-fold ορίστηκε σε 3. Οι επιλεγμένες τιμές υπερπαραμέτρων για κάθε μέθοδο και οι συναρτήσεις που χρησιμοποιούνται για την εφαρμογή τους συνοψίζονται παρακάτω.

- *k-Nearest Neighbors*: Οι πιο κρίσιμες υπερπαράμετροι σε αυτή τη μέθοδο είναι ο αριθμός των γειτόνων που χρησιμοποιούνται για την εκτέλεση των ερωτημάτων, που ορίζεται ίσος με 6, και η συνάρτηση βάρους που χρησιμοποιείται για την πραγματοποίηση των προβλέψεων, ορίζεται στο ομοιόμορφο. Η συνάρτηση *neighbors* της βιβλιοθήκης *sklearn* για Python χρησιμοποιήθηκε για την υλοποίηση της μεθόδου [52].

- *Gaussian Naive Bayes*: Η μέθοδος δεν απαιτεί εύρεση των κατάλληλότερων υπερπαραμέτρων της. Η συνάρτηση *naive_bayes* της βιβλιοθήκης *sklearn* για Python χρησιμοποιήθηκε για την υλοποίηση της μεθόδου.

- *Extreme Gradient Boosted Trees*: Το XGBoost περιλαμβάνει διάφορες υπερπαραμέτρους που είναι κρίσιμες για την απόδοσή του. Τα πιο σημαντικά είναι το *ρυθμός μάθησης*, ο *αριθμός φύλλων* του δέντρου αποφάσεων, το *αριθμός φύλλων*, που επιτρέπει τη μείωση της διακύμανσης στην πρόβλεψη και το *κλάσμα χαρακτηριστικών*, επιτρέποντας την τυχαία επιλογή ενός υποσυνόλου χαρακτηριστικών σε κάθε επανάληψη. Έχοντας εφαρμόσει τη διαδικασία διασταυρούμενης επικύρωσης, επιλέχθηκαν οι ακόλουθες τιμές: Ποσοστό εκμάθησης=0,95; μέγιστο βάθος δέντρου=10; ελάχιστο άθροισμα του βάρους της περίπτωσης που απαιτείται σε ένα παιδί =1; Αριθμός δέντρων ενισχυμένων με κλίση=1000. Όλες οι άλλες υπερπαράμετροι ορίστηκαν στην προεπιλεγμένη τιμή τους. Η συνάρτηση *Classifier* της βιβλιοθήκης *XGBoost* για Python χρησιμοποιήθηκε για την υλοποίηση της μεθόδου [12].

- *Random Forest*: Ο αριθμός των δέντρων (αριθμός εκτιμητών) στη μέθοδο RF ορίστηκε σε 500, ο συντελεστής βαρών που σχετίζεται με τις κλάσεις ορίστηκε σε ισορροπημένο (χρησιμοποιώντας τις τιμές της μεταβλητής στόχου για προσαρμογή των βαρών αντιστρόφως ανάλογα με τις συχνότητες κλάσης στις μεταβλητές χαρακτηριστικών) και ο μέγιστος αριθμός χαρακτηριστικών που χρησιμοποιούνται σε κάθε επανάληψη ορίστηκε σε log_2. Η συνάρτηση *ensemble* της

βιβλιοθήκης *sklearn* για Python χρησιμοποιήθηκε για την υλοποίηση της μεθόδου.

- *Support Vector Machine*: Το SVM χρησιμοποίησε τον γραμμικό πυρήνα, ενώ η παράμετρος τακτοποίησης C ορίστηκε σε 1. Η συνάρτηση *svm* της βιβλιοθήκης *sklearn* για Python χρησιμοποιήθηκε για την υλοποίηση της μεθόδου.

Μετά τη διαμόρφωση υπερπαραμέτρων για τις baseline μεθόδους ταξινόμησης, το τελικό βήμα της εκπαιδευτικής διαδικασίας περιλαμβάνει την εκπαίδευση του μοντέλου επιπέδου-1 στο ίδιο σύνολο εκπαίδευσης. Σε αυτή τη μελέτη, ο μετα-μαθητής ήταν ένας ταξινομητής λογιστικής παλινδρόμησης λόγω της απλότητάς του, ο οποίος συνιστάται για προγνωστικούς παράγοντες επιπέδου 1. Εφόσον ο επιλεγμένος μετα-μαθητής δεν έχει υπερπαράμετρους για συντονισμό, εκπαιδεύτηκε απευθείας στο σετ εκπαίδευσης επιπέδου 1 που αποτελείται από τις προβλέψεις επιπέδου 0 από τις βασικές μεθόδους και τις πραγματικές ετικέτες. Οι προβλέψεις των βασικών μεθόδων που σχημάτισαν το σετ εκπαίδευσης επιπέδου 1 έγιναν μέσω μιας διαδικασίας διασταυρούμενης επικύρωσης 3 φορές στο σετ εκπαίδευσης.

Για τη σύγκριση της απόδοσης των ταξινομητών χρησιμοποιήθηκε ο δείκτης F1 Score. Στη στατιστική ανάλυση της δυαδικής ταξινόμησης, το F-score ή το F-measure είναι ένα μέτρο της ακρίβειας ενός τεστ. Υπολογίζεται από την ακρίβεια και την ανάκληση της δοκιμής, όπου η ακρίβεια είναι ο αριθμός των αληθινών θετικών αποτελεσμάτων διαιρεμένος με τον αριθμό όλων των θετικών αποτελεσμάτων, συμπεριλαμβανομένων αυτών που δεν προσδιορίστηκαν σωστά, και η ανάκληση είναι ο αριθμός των αληθινών θετικών αποτελεσμάτων διαιρούμενος με το αριθμός όλων των δειγμάτων που θα έπρεπε να έχουν αναγνωριστεί ως θετικά. Η ακρίβεια είναι επίσης γνωστή ως θετική προγνωστική αξία και η ανάκληση είναι επίσης γνωστή ως ευαισθησία στη διαγνωστική δυαδική ταξινόμηση.

Το σκορ F1 είναι ο αρμονικός μέσος όρος της ακρίβειας και της ανάκλησης. Η πιο γενική βαθμολογία F_β εφαρμόζει πρόσθετα βάρη, αποτιμώντας το ένα ως προς την ακρίβεια ή την ανάκληση περισσότερο από το άλλο. Το ίδιο είδος μοντέλου μηχανικής μάθησης μπορεί να απαιτεί διαφορετικούς περιορισμούς, βάρη ή ρυθμούς εκμάθησης για τη γενίκευση διαφορετικών μοτίβων δεδομένων. Αυτά τα μέτρα ονομάζονται υπερπαράμετροι και πρέπει να ρυθμιστούν έτσι ώστε το μοντέλο να μπορεί να λύσει βέλτιστα το πρόβλημα μηχανικής εκμάθησης. Η βελτιστοποίηση υπερπαραμέτρων βρίσκει μια πλειάδα υπερπαραμέτρων που αποδίδει ένα βέλτιστο μοντέλο που ελαχιστοποιεί μια προκαθορισμένη συνάρτηση απώλειας σε δεδομένα ανεξάρτητα δεδομένα. Η αντικειμενική συνάρτηση παίρνει μια πλειάδα υπερπαραμέτρων και επιστρέφει τη σχετική απώλεια. Η διασταυρούμενη επικύρωση χρησιμοποιείται συχνά για την εκτίμηση αυτής της απόδοσης γενίκευσης.

Η απόδοση των βασικών μεθόδων και του ταξινομητή στοίβαξης στη διαδικασία επικύρωσης στρωματοποιημένης k-fold (10 φορές $\times$ 3 επαναλήψεις = 30 επικυρώ-

σεις) συνοψίζεται στον Πίνακα 4.1. Από τους baseline ταξινομητές, η μέθοδος SVM αναφέρει την καλύτερη απόδοση σύμφωνα με τη βαθμολογία F1, ακολουθούμενη από τις μεθόδους RF και XGBoost. Όπως συζητήθηκε στην ενότητα 3, αυτό μπορεί να αποδοθεί στην ικανότητα των μεθόδων XGBoost και RF να επιλέγουν τα πιο σχετικά χαρακτηριστικά από ένα μεγάλο σύνολο επεξηγηματικών μεταβλητών, καθώς και στο γεγονός ότι η RF και η SVM είναι γενικά καλύτερα στην αντιμετώπιση της τυχαιότητας των δεδομένων. Επιπλέον, διαπιστώνουμε ότι το μοντέλο στοίβαξης υπερέχει όλων των βασικών μεθόδων σύμφωνα τόσο με την ακρίβεια όσο και με τα μέτρα βαθμολογίας F1, όντας ελαφρώς χειρότερο από το SVM όσον αφορά την ακρίβεια. Σημειώστε ότι αν και οι διαφορές μεταξύ του μετα-μαθητή και του ταξινομητή SVM είναι μικρές σύμφωνα με τα μέτρα ταξινόμησης που χρησιμοποιούνται, ο πρώτος είναι πιο ισχυρός, έχοντας την τάση να ταξινομεί τις περισσότερες από τις επενδύσεις με μεγαλύτερη ακρίβεια, αποφεύγοντας ακραία σφάλματα. Αυτό γίνεται εμφανές παρατηρώντας το Σχήμα 4.3, το οποίο παρέχει μια οπτική σύγκριση της απόδοσης των ταξινομητών με τη μορφή γραφικών πλαισίων. Όπως εξηγήθηκε, η διάμεσος της βαθμολογίας F1 και τα μέτρα ακρίβειας είναι υψηλότερη για το μετα-ασθενέστερο σε σύγκριση με το SVM (περισσότερες επενδύσεις ταξινομούνται σωστά), ενώ το διατεταρτημόριο του είναι σημαντικά μικρότερο (καλύτερη στιβαρότητα).

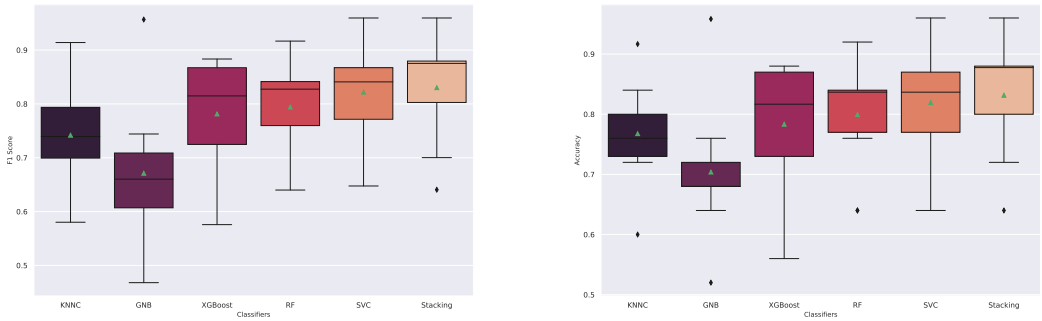
Classifier	Accuracy	Precision	F1 Score
k-Nearest Neighbors	0.77 (0.08)	0.78 (0.10)	0.74 (0.09)
Gaussian naive Bayes	0.70 (0.10)	0.66 (0.14)	0.67 (0.12)
XGBoost	0.78 (0.10)	0.79 (0.10)	0.78 (0.10)
Random Forest	0.80 (0.09)	0.82 (0.08)	0.79 (0.09)
Support Vector Machine	0.82 (0.09)	**0.85 (0.07)**	0.82 (0.09)
Stacking Model	**0.83 (0.09)**	0.84 (0.09)	**0.83 (0.09)**

Table 4.1: Classification performance (mean and standard deviation) of the baseline methods and the stacking model on the stratified 10-fold validation process. Column-wise best values are displayed in bold.

4.3 Πρόβλεψη για μελλοντικές επενδύσεις

Μετά την εκπαίδευση του μοντέλου στοίβαξης, η απόδοση της προτεινόμενης μεθοδολογίας αξιολογείται στο σύνολο δεδομένων αξιολόγησης, το οποίο, όπως εξηγήθηκε προηγουμένως, αποτελείται από 63 τυχαία επιλεγμένες επενδύσεις. Τα αποτελέσματα αυτής της εφαρμογής συνοψίζονται στον πίνακα σύγχυσης του Σχήματος 4.4, παρουσιάζοντας τον αριθμό των περιπτώσεων που ο μετα-μαθητής έχει προβλέψει κάθε κλάση σε σύγκριση με τις πραγματικές τους ετικέτες. Παρατηρούμε ότι 32 από τις 36 επενδύσεις *Κλάσης B* $(88,8\%)$ έχουν προβλεφθεί σωστά. Για την *Κλάση A*, ο αντίστοιχος

Figure 4.3: Σύγκριση των βασικών ταξινομητών και της στοίβαξης στη στρωματοποιημένη 10-πλάσια διαδικασία επικύρωσης. Η αριστερή υποεικόνα δείχνει τη βαθμολογία F1 των ταξινομητών, ενώ η δεξιά την ακρίβειά τους.



(a) F1 Score

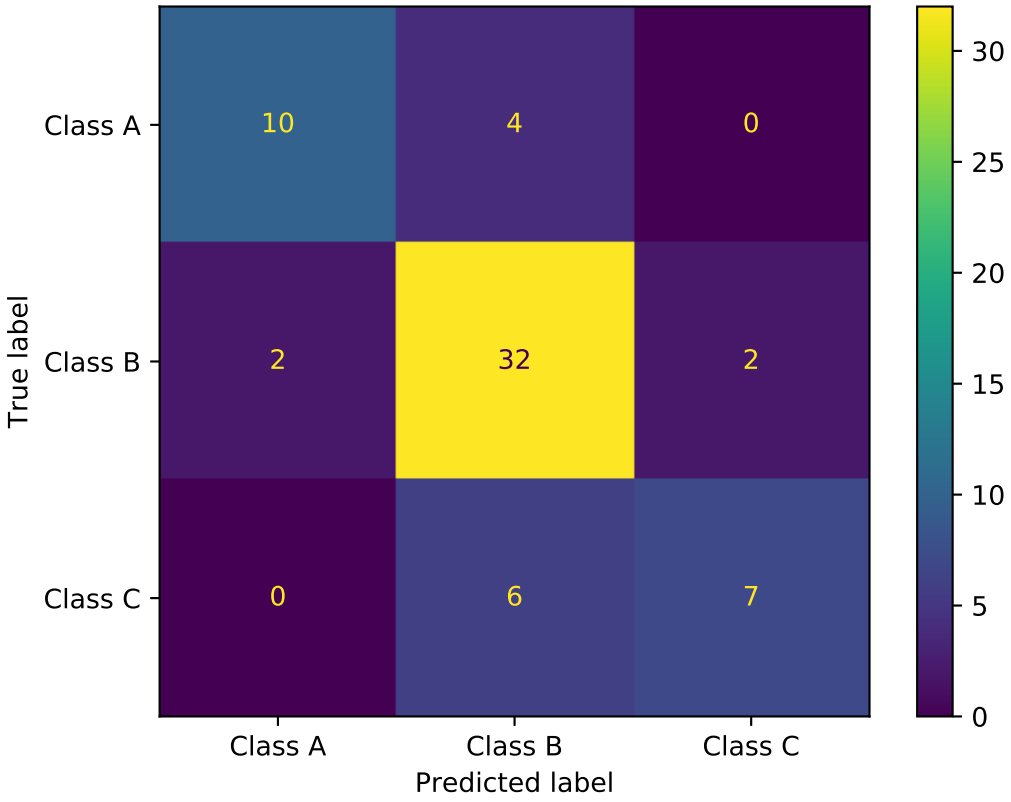
(b) Accuracy

αριθμός είναι 10 από 14 (71, 4%), ενώ για *Κλάση C*, 7 από 13 (53, 2%). Είναι ενδιαφέρον ότι ο ταξινομητής δεν έχει προβλέψει ποτέ την επένδυση *Κλάσης A* ως *Κλάσης C* ή το αντίθετο. Αυτό το εύρημα είναι ιδιαίτερα ενθαρρυντικό, υποδηλώνοντας ότι ο κίνδυνος χρήσης του προτεινόμενου μοντέλου μετα-μάθησης για τη σύσταση χρηματοδότησης (ή μη χρηματοδότησης) μιας επένδυσης χαμηλού (ή υψηλών) δυνατοτήτων είναι ασήμαντος.

Η απόδοση ταξινόμησης του μετα-μαθητή συνοψίζεται στον Πίνακα 4.2. Αυτό περιλαμβάνει την ακρίβεια, την ανάκληση και τη βαθμολογία F1 κάθε κατηγορίας, καθώς και τον μακρο και τον σταθμισμένο μέσο όρο. Ο μακρο όρος είναι ο απλός αριθμητικός μέσος όρος κάθε μέτρου σε όλες τις κατηγορίες. Δεδομένου ότι αυτό το μέτρο αποδίδει ίσα βάρη σε όλες τις τάξεις, είναι κατάλληλο για προβλήματα ισορροπημένης ταξινόμησης. Αντίθετα, ο σταθμισμένος μέσος όρος ενσωματώνει την ανισορροπία κατηγορίας, υπολογίζοντας τον μέσο όρο των δυαδικών μετρήσεων λαμβάνοντας υπόψη τον αριθμό των δειγμάτων κάθε κατηγορίας στον στόχο. Επομένως, ο σταθμισμένος μέσος όρος είναι πιο κατάλληλος για τη σύνοψη της απόδοσης του μετα-μαθητή στην παρούσα μελέτη περίπτωσης.

Είναι προφανές ότι το μοντέλο αποδίδει ελαφρώς καλύτερα με τις επενδύσεις *Κλάση B*, δηλαδή την πιο δημοφιλή κατηγορία, με αποτέλεσμα ακρίβεια 0,76 και βαθμολογία F1 0,82. Οι αντίστοιχες τιμές για την *Κλάση A* είναι 0,83 και 0,77, ενώ για την *Κλάση C* 0,78 και 0,64. Επομένως, διαπιστώνουμε ότι είναι πολύ απίθανο ο μετα-μαθητής να μην προσδιορίσει σωστά τις επενδύσεις *Κλάσης A* και *Κλάση B* όταν υπάρχουν, σε αντίθεση με τις επενδύσεις *Κλάσης C* όπου είναι πιθανό για ο μετα-μαθητής να τους ταξινομήσει ως *Κλάση B*. Ως αποτέλεσμα, μπορούμε να συμπεράνουμε ότι ο μετα-μαθητής μπορεί να χρησιμοποιηθεί αποτελεσματικά στην πράξη για τον εντοπισμό επενδύσεων υψηλού και μεσαίου δυναμικού και την υποστήριξη της χρηματοδότησης έργων ενεργειακής αποδοτικότητας μεγάλης κλίμακας. Αυτό επιβεβαιώνεται από τη

Figure 4.4: Πίνακας σύγχυσης για το μοντέλο στοίβαξης στο δοκιμαστικό σετ.



συνολική ακρίβεια του μοντέλου, που είναι 0,78 σε όλες τις κατηγορίες. Παρατηρήστε επίσης ότι, παρόλο που το σύνολο δεδομένων που χρησιμοποιήθηκε για την εκπαίδευση των baseline μεθόδων ταξινόμησης ήταν μη ισορροπημένο, οι βαθμολογίες ανάκλησης που αναφέρονται στον Πίνακα 4.2 υποδηλώνουν ότι ο μετα-απαίσιος έχει αντιμετωπίσει αποτελεσματικά αυτό το ζήτημα, παρέχοντας αμερόληπτες προβλέψεις, ειδικά όταν πρόκειται για πρόβλεψη Επενδύσεις *Κλάση A* και *Κλάση B*.

	Precision	Recall	F1 Score
Κλάση A	0.83	0.71	0.77
Κλάση B	0.76	0.89	0.82
Κλάση C	0.78	0.54	0.64
Macro Average	0.79	0.71	0.74
Weighted Average	0.78	0.78	0.77

Table 4.2: Απόδοση ταξινόμησης του μοντέλου στοίβαξης στο δοκιμαστικό σετ.

Εφόσον το μέτρο ακρίβειας δεν ενσωματώνει την ανισορροπία κλάσης, που είναι

ένας πολύ σημαντικός παράγοντας για αυτό το πρόβλημα (60% των παρατηρήσεων είναι στην ίδια κατηγορία), εκμεταλλευόμαστε τον συντελεστή κάπα του Cohen (κ). Αυτός ο συντελεστής είναι ένας στατιστικός δείκτης που χρησιμοποιείται για τη μέτρηση της αξιοπιστίας μεταξύ των αξιολογητών για κατηγορικά δεδομένα, καθώς είναι πιο ισχυρό μέτρο από τον απλό υπολογισμό της συμφωνίας ποσοστού, λόγω της παραμέτρου κ που ενσωματώνει την πιθανότητα η συμφωνία να συμβεί κατά λάθος. Ο τύπος για τον συντελεστή κάπα του Cohen δίνεται ως εξής:

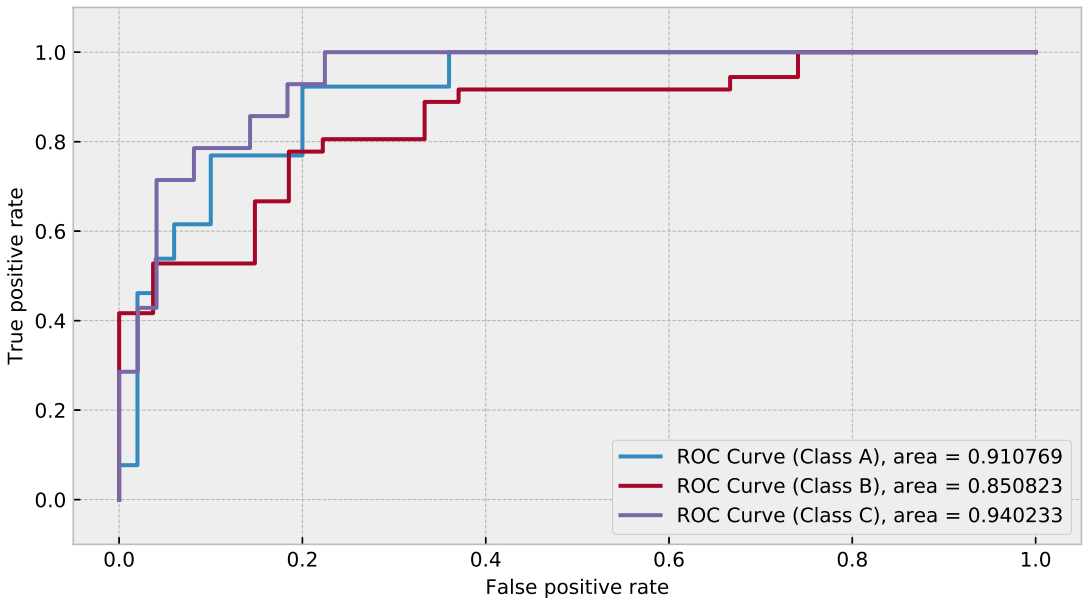
$$\kappa = \frac{p_0 - p_e}{1 - p_e}, \tag{4.1}$$

όπου το p_0 περιγράφεται ως η παρατηρούμενη αναλογική συμφωνία μεταξύ των πραγματικών και των προβλεπόμενων τιμών και το p_e είναι η αναμενόμενη συμφωνία όταν και οι δύο σχολιαστές εκχωρούν ετικέτες τυχαία. Η βαθμολογία κάπα του Cohen για το σύνολο αξιολόγησης είναι 0,594, που υποδηλώνει ουσιαστική συμφωνία για τον ταξινομητή στοίβαξης.

Ωστόσο, η προτεινόμενη μεθοδολογία βασίζεται σε ταξινομητές που μπορούν να δημιουργήσουν πιθανότητες συμμετοχής στην κλάση. Επομένως, οι προαναφερθείσες μετρήσεις αποτυγχάνουν να περιγράψουν ολόκληρη την εικόνα. Η ικανότητα του μοντέλου να διακρίνει ένα στιγμιότυπο μεταξύ κάθε κατηγορίας μπορεί να περιγραφεί από την περιοχή κάτω από τη χαρακτηριστική καμπύλη λειτουργίας του δέκτη (ROC AUC). Η καμπύλη AUC - ROC είναι μια μέτρηση απόδοσης για τα προβλήματα ταξινόμησης σε διάφορες ρυθμίσεις κατωφλίου. Το ROC είναι μια καμπύλη πιθανότητας και η AUC αντιπροσωπεύει το βαθμό ή το μέτρο της διαχωρισιμότητας. Δείχνει πόσο το μοντέλο είναι ικανό να διακρίνει μεταξύ τάξεων. Όσο υψηλότερη είναι η AUC, τόσο καλύτερα είναι το μοντέλο στην πρόβλεψη 0 τάξεων ως 0 και 1 τάξεων ως 1. Κατ' αναλογία, όσο υψηλότερη είναι η AUC, τόσο καλύτερο είναι το μοντέλο στη διάκριση μεταξύ των διαφορετικών κλάσεων.

Το ROC AUC υπολογίζεται, δημιουργώντας προβλέψεις για ένα εύρος τιμών κατωφλίου απόφασης $[0, 1]$. Για κάθε τιμή υπολογίζεται το πραγματικό θετικό ποσοστό (TPR) - γνωστό και ως ανάκληση - και το ψευδώς θετικό ποσοστό (FPR). Στην περίπτωση ταξινόμησης πολλών κατηγοριών, το μέτρο υπολογίζεται χωριστά για κάθε κλάση ακολουθώντας τη στρατηγική ταξινόμησης one-vs-rest (OVR), υποθέτοντας διαφορετικό ταξινομητή ανά κατηγορία. Η καμπύλη ROC για κάθε κλάση φαίνεται στο σχήμα 4.5. Μια μεγάλη περιοχή στο ROC AUC σημαίνει ότι υπάρχει μεγάλη διάκριση μεταξύ των κατηγοριών. Η τιμή της περιοχής κάτω από την καμπύλη ROC για τον ταξινομητή στοίβαξης συνολικά είναι 0,883, που είναι μια καλή βαθμολογία δεδομένου ότι οι τιμές AUC βρίσκονται μεταξύ 0,5 και 1 (όπου το 0,5 υποδηλώνει έναν κακό ταξινομητή και το 1 έναν εξαιρετικό ταξινομητή).

Figure 4.5: Καμπύλη ROC για τις τρεις κατηγορίες με βάση τις πιθανοτικές προβλέψεις που δημιουργούνται από τον ταξινομητή στοίβαξης.



Chapter 5

Συμπεράσματα και Μελλοντικές Προεκτάσεις

5.1 Συμπεράσματα

Αυτή η μελέτη εστιάζει στο πρόβλημα της αξιολόγησης του δυναμικού των μελλοντικών επενδύσεων στην ενεργειακή αποδοτικότητα όσον αφορά το κόστος ανακαίνισης και την πραγματοποιηθείσα εξοικονόμηση ενέργειας. Δεδομένης της έλλειψης ώριμων συστημάτων πληροφόρησης για την υποστήριξη τέτοιων αξιολογήσεων, αρκετά έργα ανακαίνισης δυσκολεύονται επί του παρόντος να λάβουν οικονομική στήριξη από ιδρύματα χρηματοδότησης, θέτοντας έτσι σε κίνδυνο τους παγκόσμιους περιβαλλοντικούς στόχους που έχουν τεθεί για τις επόμενες δεκαετίες. Από την άποψη αυτή, προτείνεται μια μεθοδολογία βασισμένη σε δεδομένα, η οποία φιλοδοξεί να ανοίξει το δρόμο για τον ακριβή εντοπισμό ελκυστικών έργων ενεργειακής αποδοτικότητας και τον προσδιορισμό των κεφαλαίων που θα πρέπει να επενδυθούν ανά περίπτωση. Για να γίνει αυτό, λαμβάνονται υπόψη κρίσιμα επενδυτικά χαρακτηριστικά έργων που έχουν ολοκληρωθεί στο παρελθόν και χρησιμοποιούνται μέθοδοι μηχανικής μάθησης για να διδαχθούν από τις επιτυχίες και τα λάθη τους. Η συνιστώμενη προσέγγιση στοχεύει στη βελτίωση των διαδικασιών χρηματοδότησης και στη διευκόλυνση διαφορετικών τύπων ενδιαφερόμενων μερών (χρηματοδότες, τραπεζίτες, επενδυτές, κ.λπ.) στη σύγκριση και την επισήμανση έργων ΕΑ με τυποποιημένο και λιγότερο αβέβαιο τρόπο.

Η προτεινόμενη μεθοδολογία περιλαμβάνει ένα μοντέλο συνόλου στοίβαξης που βασίζεται σε διάφορες baseline μεθόδους ταξινόμησης μηχανικής μάθησης με στόχο την περαιτέρω βελτίωση της απόδοσής τους. Συλλέγονται δεδομένα που περιέχουν σημαντικά χαρακτηριστικά επενδύσεων ενεργειακής αποδοτικότητας (π.χ. ηλικία κτιρίου και συνολική επιφάνεια θέρμανσης, αναμενόμενες εκπομπές CO_2 και κόστος ανακαίνισης) και χρησιμοποιούνται για την εκπαίδευση του μοντέλου ταξινόμησης. Η μεθοδολογία αξιολογείται λαμβάνοντας υπόψη μια πραγματική μελέτη περίπτωσης στη Λετονία. Τα αποτελέσματά μας υποδηλώνουν ότι το μοντέλο συνόλου στοίβαξης υπερέχει όλων

των μεθόδων ταξινόμησης μηχανικής μάθησης βάσης που εξετάζονται, επιτυγχάνοντας υψηλή ακρίβεια όταν χρησιμοποιείται για την αξιολόγηση μελλοντικών επενδύσεων. Επιπλέον, διαπιστώνουμε ότι το προτεινόμενο μοντέλο μπορεί να αναγνωρίσει αποτελεσματικά έργα υψηλών και μεσαίων επενδύσεων, όντας επίσης εξαιρετικό στη διάκριση χαμηλών από υψηλού δυναμικού. Αυτό το εύρημα δείχνει ότι τα ενδιαφερόμενα μέρη μπορούν να εκμεταλλευτούν την παρουσιαζόμενη μεθοδολογία για να μειώσουν τον κίνδυνο επενδύσεών τους και να μεγιστοποιήσουν τις αποδόσεις τους.

5.2 Μελλοντικές προεκτάσεις

Λαμβάνοντας υπόψη τους περιορισμούς της παρούσας μελέτης, η μελλοντική έρευνα στον τομέα της αξιολόγησης επενδύσεων ενεργειακής αποδοτικότητας θα πρέπει να επικεντρωθούν στην ενσωμάτωση άλλων τύπων επενδύσεων που προέρχονται από τον τομέα της μεταποίησης, των μεταφορών και του εξωτερικού φωτισμού, διευρύνοντας έτσι τα έργα που εξετάζονται από τη μεθοδολογία και αυξάνοντας τα οικονομικά και περιβαλλοντικό δυναμικό. Το αναπτυγμένο μοντέλο μετα-μάθησης μπορεί να γενικευτεί για τέτοιες εφαρμογές υπό την προϋπόθεση ότι υπάρχουν επαρκή δεδομένα για την υποστήριξη της μαθησιακής του διαδικασίας. Αυτό βέβαια θα επιφέρει την ανάγκη για εκ νέου συντονισμό των υπερπαραμέτρων των μοντέλων με βάση τα δεδομένα που θα εισαχθούν στη νέα μοντελοποίηση.

Ένα άλλο σημείο ενδιαφέροντος που θα μπορούσε να διερευνηθεί περαιτέρω είναι το πώς η γεωγραφική τοπολογία των επενδύσεων επηρεάζει την απόδοσή τους όσον αφορά την εξοικονόμηση ενέργειας. Από αυτή την άποψη, επενδύσεις που υλοποιούνται σε πολλές χώρες με διάφορα χαρακτηριστικά θα μπορούσαν να ενσωματωθούν στο μοντέλο, θέτοντας ορισμένες πρόσθετες γεωγραφικές και κλιματικές μεταβλητές για να γενικευτεί η χρήση του και να διευρυνθεί ο τύπος των ενδιαφερομένων που θα ενδιαφερόταν για τη χρησιμοποίησή του. Εναλλακτικά μέτρα σε σχέση με τα εξεταζόμενα θα μπορούσαν επίσης να εξεταστούν για την αξιολόγηση της δυνατότητας μελλοντικών έργων ενεργειακής αποδοτικότητας ώστε οι προβλέψεις της προτεινόμενης μεθόδου να αντικατοπτρίζουν καλύτερα τη βασική αλήθεια.

Το πιο σημαντικό, και προκειμένου η προτεινόμενη προσέγγιση να καταστεί πιο σχετική για εφαρμογές μεγάλης κλίμακας, θα πρέπει να εντοπιστούν εναλλακτικοί τρόποι για την ανάκτηση των δεδομένων που σχετίζονται με την ενέργεια και τις επενδύσεις που απαιτούνται για την εκπαίδευση των ταξινομητών, συμπεριλαμβανομένων των πηγών δεδομένων από τις δημόσιες αρχές και τα χρηματοπιστωτικά ιδρύματα μπορεί εύκολα να έχει πρόσβαση. Για παράδειγμα, μια εναλλακτική θα ήταν η εκμετάλλευση των πληροφοριών που παρέχονται από τα ενεργειακά πιστοποιητικά των ανακαινισμένων κτιρίων, δηλαδή η παρακολούθηση πολλών κτιρίων πριν και μετά την πραγματοποίηση συγκεκριμένων ανακαινίσεων, καθώς και το κόστος αυτών των ανακαινίσεων. Αυτός ο τύπος πληροφοριών είναι τυποποιημένος, ερμηνεύσιμος και μοιράζεται ευκολότερα

με τα ενδιαφερόμενα μέρη, διευκολύνοντας έτσι τη χρήση της προτεινόμενης μεθόδου αξιολόγησης σε ευρύτερη κλίμακα.

Bibliography

[1] N Abdou et al. "Multi-objective optimization of passive energy efficiency measures for net-zero energy building in Morocco". In: *Building and Environment* 204 (2021), p. 108141.

[2] Maria Isabel Abreu, Rui AF de Oliveira, and Jorge Lopes. "Younger vs. older homeowners in building energy-related renovations: Learning from the Portuguese case". In: *Energy Reports* 6 (2020), pp. 159–164.

[3] Muhammad Waseem Ahmad, Monjur Mourshed, and Yacine Rezgui. "Trees vs Neurons: Comparison between random forest and ANN for high-resolution prediction of building energy consumption". In: *Energy and Buildings* 147 (2017), pp. 77–89.

[4] Apostolos Arsenopoulos et al. "A Data-Driven Decision Support Tool at the service of Energy suppliers and Utilities for Tackling Energy Poverty: A case study in Greece". In: *2021 12th International Conference on Information, Intelligence, Systems & Applications (IISA)*. IEEE. 2021, pp. 1–6.

[5] Dalia M Atallah et al. "Predicting kidney transplantation outcome based on hybrid feature selection and KNN classifier". In: *Multimedia Tools and Applications* 78.14 (2019), pp. 20383–20407.

[6] Ramazan Bayindir et al. "A novel application of naive bayes classifier in photovoltaic energy prediction". In: *2017 16th IEEE international conference on machine learning and applications (ICMLA)*. IEEE. 2017, pp. 523–527.

[7] Gou Bo and Huang Xianwu. "SVM multi-class classification". In: *Journal of Data Acquisition & Processing* 21.3 (2006), pp. 334–339.

[8] Frank Pieter Boon and Carel Dieperink. "Local civil society based renewable energy organisations in the Netherlands: Exploring the factors that stimulate their emergence and development". In: *Energy Policy* 69 (2014), pp. 297–307.

[9] L Breiman. *Classification and Regression Trees*. Boca Raton, FL: Chapman & Hall, 1993.

[10] Leo Breiman. "Bagging predictors". In: *Machine learning* 24.2 (1996), pp. 123–140.

[11] Debosmita Chakraborty, Ujjal Sur, and Pradipta Kumar Banerjee. "Random Forest Based Fault Classification Technique for Active Power System Networks". In: *2019 IEEE International WIE Conference on Electrical and Computer Engineering (WIECON-ECE)*. IEEE. 2019, pp. 1–4.

[12] Tianqi Chen and Carlos Guestrin. "Xgboost: A scalable tree boosting system". In: *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*. 2016, pp. 785–794.

[13] Zhen Chen et al. "iLearn: an integrated platform and meta-learner for feature engineering, machine-learning analysis and modeling of DNA, RNA and protein sequence data". In: *Briefings in bioinformatics* 21.3 (2020), pp. 1047–1057.

[14] Chen Chia-Cheng, Yisheng Liu, and Ting-Hsin Hsu. "An analysis on investment performance of machine learning: an empirical examination on Taiwan stock market". In: *International Journal of Economics and Financial Issues* 9.4 (2019), p. 1.

[15] Thomas Cover and Peter Hart. "Nearest neighbor pattern classification". In: *IEEE transactions on information theory* 13.1 (1967), pp. 21–27.

[16] Caleb Debrah, Albert Ping Chuen Chan, and Amos Darko. "Green finance gap in green buildings: A scoping review and future research needs". In: *Building and Environment* (2021), p. 108443.

[17] Sukhpreet Singh Dhaliwal, Abdullah-Al Nahid, and Robert Abbas. "Effective intrusion detection system using XGBoost". In: *Information* 9.7 (2018), p. 149.

[18] Duleep Rathgamage Don and Ionut E Iacob. "DCSVM: fast multi-class classification using support vector machines". In: *International Journal of Machine Learning and Cybernetics* 11.2 (2020), pp. 433–447.

[19] Haris Doukas. "On the appraisal of "Triple-A" energy efficiency investments". In: *Energy Sources, Part B: Economics, Planning, and Policy* 13.7 (2018), pp. 320–327.

[20] Haris Doukas, Panos Xidonas, and Nikos Mastromichalakis. "How Successful are Energy Efficiency Investments? A Comparative Analysis for Classification & Performance Prediction". In: *Computational Economics* (2021), pp. 1–20.

[21] Haris Doukas et al. "From integrated to integrative: Delivering on the Paris Agreement". In: *Sustainability* 10.7 (2018), p. 2299.

[22] Ehab Foda, Ashraf El-Hamalawi, and Jérôme Le Dréau. "Computational analysis of energy and cost efficient retrofitting measures for the French house". In: *Building and Environment* 175 (2020), p. 106792.

[23] Aikaterini Forouli et al. “Energy efficiency promotion in Greece in light of risk: Evaluating policies as portfolio assets”. In: *Energy* 170 (2019), pp. 818–831.

[24] Jerome H Friedman. “Stochastic gradient boosting”. In: *Computational statistics & data analysis* 38.4 (2002), pp. 367–378.

[25] Eric Galin et al. “A review of digital terrain modeling”. In: *Computer Graphics Forum*. Vol. 38. 2. Wiley Online Library. 2019, pp. 553–577.

[26] Hamed Ghoddusi, Germán G Creamer, and Nima Rafizadeh. “Machine learning in energy economics and finance: A review”. In: *Energy Economics* 81 (2019), pp. 709–727.

[27] Sara Hayes et al. “What have we learned from energy efficiency financing programs”. In: *American Council for an Energy-Efficient Economy*. 2011.

[28] Sergio Herrero-Lopez, John R Williams, and Abel Sanchez. “Parallel multiclass classification using SVMs on GPUs”. In: *Proceedings of the 3rd Workshop on general-purpose computation on graphics processing units*. 2010, pp. 2–11.

[29] IEA. *Implementing Energy Efficiency Policies: are IEA Member Countries on Track?* `https://www.oecd.org/publications/implementing-energy-efficiency-policies-are-iea-member-countries-on-track-9789264075696-en.htm`. IEA Paris, France, 2009.

[30] IEA. *Net Zero by 2050: A Roadmap for the Global Energy Sector*. `https://iea.blob.core.windows.net/assets/4719e321-6d3d-41a2-bd6b-461ad2f850a8/NetZeroby2050-ARoadmapfortheGlobalEnergySector.pdf`. IEA Paris, France, 2021.

[31] Singh Intrachooto and Vimolsiddhi Horayangkura. “Energy efficient innovation: Overcoming financial barriers”. In: *Building and Environment* 42.2 (2007), pp. 599–604.

[32] IRENA. *Renewable Energy Statistics*. `https://www.irena.org/publications/2015/Jun/Renewable-Energy-Target-Setting`. 2015.

[33] Ali Haghpanah Jahromi and Mohammad Taheri. “A non-parametric mixture of Gaussian naive Bayes classifiers based on local independent features”. In: *2017 Artificial Intelligence and Signal Processing Conference (AISP)*. IEEE. 2017, pp. 209–212.

[34] Per Anker Jensen and Esmir Maslesa. “Value based building renovation–A tool for decision-making and evaluation”. In: *Building and Environment* 92 (2015), pp. 1–9.

[35] Per Anker Jensen et al. “10 questions concerning sustainable building renovation”. In: *Building and Environment* 143 (2018), pp. 130–137.

[36] Bernhard J Kalkbrenner and Jutta Roosen. "Citizens' willingness to participate in local renewable energy projects: The role of community and trust in Germany". In: *Energy Research & Social Science* 13 (2016), pp. 60–70.

[37] Bartosz Krawczyk. "Learning from imbalanced data: open challenges and future directions". In: *Progress in Artificial Intelligence* 5.4 (2016), pp. 221–232. DOI: 10.1007/s13748-016-0094-0.

[38] Justin Y Lee and Mark P Styczynski. "NS-kNN: a modified k-nearest neighbors approach for imputing metabolomics data". In: *Metabolomics* 14.12 (2018), pp. 1–12.

[39] Qing Yun Li, Jie Han, and Lin Lu. "A Random Forest Classification Algorithm Based Personal Thermal Sensation Model for Personalized Conditioning System in Office Buildings". In: *The Computer Journal* 64.3 (2021), pp. 500–508.

[40] Spyros Makridakis, Evangelos Spiliotis, and Vassilios Assimakopoulos. "Statistical and Machine Learning forecasting methods: Concerns and ways forward". In: *PLOS ONE* 13.3 (Mar. 2018), pp. 1–26. DOI: 10.1371/journal.pone.0194889. URL: https://doi.org/10.1371/journal.pone.0194889.

[41] Vangelis Marinakis. "Big data for energy management and energy-efficient buildings". In: *Energies* 13.7 (2020), p. 1555.

[42] Ajay Mathur and Giles M Foody. "Multiclass and binary SVM classification: Implications for training and classification users". In: *IEEE Geoscience and remote sensing letters* 5.2 (2008), pp. 241–245.

[43] Deiner Mena et al. "An overview of inference methods in probabilistic classifier chains for multilabel classification". In: *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 6.6 (2016), pp. 215–230.

[44] Filippos Dimitrios Mexis et al. "Leveraging Energy Efficiency Investments: An Innovative Web-based Benchmarking Tool". In: *Advances in Science, Technology and Engineering Systems Journal* 6 (2021), pp. 237–248.

[45] John Morrissey et al. "Cost-benefit assessment of energy efficiency investments: Accounting for future resources, savings and risks in the Australian residential sector". In: *Energy Policy* 54 (2013), pp. 148–159.

[46] Mohammad Norouzi, David J Fleet, and Russ R Salakhutdinov. "Hamming distance metric learning". In: *Advances in neural information processing systems*. 2012, pp. 1061–1069.

[47] Joana Ortiz et al. "Cost-effective analysis for selecting energy efficiency measures for refurbishment of residential buildings in Catalonia". In: *Energy and Buildings* 128 (2016), pp. 442–457.

[48] Jyoti P Painuly. "Barriers to renewable energy penetration; a framework for analysis". In: *Renewable energy* 24.1 (2001), pp. 73–89.

[49] Jyoti P Painuly et al. "Promoting energy efficiency financing and ESCOs in developing countries: mechanisms and barriers". In: *Journal of Cleaner Production* 11.6 (2003), pp. 659–665.

[50] Aikaterini Papapostolou et al. "Web-based Application for Screening Energy Efficiency Investments: A MCDA Approach". In: *2020 11th International Conference on Information, Intelligence, Systems and Applications (IISA*. IEEE. 2020, pp. 1–7.

[51] Bohdan Pavlyshenko. "Using stacking approaches for machine learning models". In: *2018 IEEE Second International Conference on Data Stream Mining & Processing (DSMP)*. IEEE. 2018, pp. 255–258.

[52] F. Pedregosa et al. "Scikit-learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830.

[53] Budi Prasetiyo, MA Muslim, et al. "Analysis of building energy efficiency dataset using naive bayes classification classifier". In: *Journal of Physics: Conference Series*. Vol. 1321. IOP Publishing. 2019, p. 032016.

[54] Rajeev DS Raizada and Yune-Sang Lee. "Smoothness without smoothing: why Gaussian naive Bayes is not naive for multi-subject searchlight studies". In: *PloS one* 8.7 (2013), e69566.

[55] Ashok Sarkar and Jas Singh. "Financing energy efficiency in developing countries—lessons learned and remaining challenges". In: *Energy Policy* 38.10 (2010), pp. 5560–5571.

[56] Elissaios Sarmas, Panos Xidonas, and Haris Doukas. *Multicriteria Portfolio Construction with Python*. Springer, 2020.

[57] Mauro Sarrica et al. "One, no one, one hundred thousand energy transitions in Europe: The quest for a cultural approach". In: *Energy Research & Social Science* 13 (2016), pp. 1–14.

[58] Alex J Smola and Bernhard Schölkopf. "A tutorial on support vector regression". In: *Statistics and computing* 14.3 (2004), pp. 199–222.

[59] Emanuel Stocker, Martin Tschurtschenthaler, and Lukas Schrott. "Cost-optimal renovation and energy performance: Evidence from existing school buildings in the Alps". In: *Energy and Buildings* 100 (2015), pp. 20–26.

[60] Iwan Syarif et al. "Application of bagging, boosting and stacking to intrusion detection". In: *International Workshop on Machine Learning and Data Mining in Pattern Recognition*. Springer. 2012, pp. 593–602.

[61] Peter G Taylor et al. “Final energy use in IEA countries: The role of energy efficiency”. In: *Energy Policy* 38.11 (2010), pp. 6463–6474.

[62] Robert P Taylor et al. *Financing energy efficiency: lessons from Brazil, China, India, and beyond*. World Bank Publications, 2008.

[63] John Thøgersen and Alice Grønhøj. “Electricity saving in households—A social cognitive approach”. In: *Energy policy* 38.12 (2010), pp. 7732–7743.

[64] Godfried Toussaint. “Geometric proximity graphs for improving nearest neighbor methods in instance-based learning and data mining”. In: *International Journal of Computational Geometry & Applications* 15.02 (2005), pp. 101–150.

[65] Pekka Tuominen et al. “Economic appraisal of energy efficiency in buildings using cost-effectiveness assessment”. In: *Procedia Economics and Finance* 21 (2015), pp. 422–430.

[66] Ana Brandão de Vasconcelos et al. “EPBD cost-optimal methodology: Application to the thermal rehabilitation of the building envelope of a Portuguese residential reference building”. In: *Energy and Buildings* 111 (2016), pp. 12–25.

[67] David H Wolpert. “Stacked generalization”. In: *Neural networks* 5.2 (1992), pp. 241–259.

[68] Marijana Zekić-Sušac, Saša Mitrović, and Adela Has. “Machine learning based system for managing energy efficiency of public sector as an approach towards smart cities”. In: *International journal of information management* 58 (2021), p. 102074.

[69] Dayong Zhang, Zhiwei Zhang, and Shunsuke Managi. “A bibliometric analysis on green finance: Current status, development, and future directions”. In: *Finance Research Letters* 29 (2019), pp. 425–430.

[70] Harry Zhang. “The optimality of naive Bayes”. In: *AA* 1.2 (2004), p. 3.

[71] Tao Zhang et al. “Improving convection trigger functions in deep convective parameterization schemes using machine learning”. In: *Journal of Advances in Modeling Earth Systems* 13.5 (2021), e2020MS002365.