FH MÜNSTER
University of Applied Sciences

Fachbereich
Energie · Gebäude · Umwelt

Energiebedarfsbestimmung von Bestands-Wohngebäuden anhand von maschinellen Lernmethoden

Masterarbeit

02. Mai 2024
Philipp Sommer, B.Eng.
philipp.sommer@fh-muenster.de

Erstprüfer: Prof. Dr. Elmar Brügging, FH Münster
Zweitprüfer: Amgad Agoub M.Sc., syte GmbH

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

Quellcodeverzeichnis

Abkürzungsverzeichnis

ANN	Artificial Neural Network
ALKIS	Amtliches Liegenschaftskatastar
ALS	Airborne Laser Scanning
API	Application Programmer Interface
DOM	Digitales Oberflächenmodell
EA	Energieausweis
DT	Decision Tree
EPBD	Energy Performance Building Directive
EU	Europäische Union
EVU	Energieversorgungsunternehmen
FE	Feature Engineering
FED	Fernerkundungsdaten
GEG	Gebäudeenergiegesetz
HVAC	Heating, Ventilation and Air Conditioning
KI	Künstliche Intelligenz
KNN	K-nächste Nachbarn
kWP	Kommunale Wärmeplanung
LiDAR	Light Detection and Ranging
LoD	Level of Detail
LR	Linear Regression
MAE	Mean Absolute Error
MaStR	Marktstammdatenregister
ML	Machine Learning
MSE	Mean Squared Error
NLP	Natural Language Processing
NWG	Nicht-Wohngebäude
OSM	Open Street Map
RF	Random Forest
RMSE	Root Mean Squared Error
SHAP	SHapley Additive exPlanations
SkL	Scikit-Learn

SVM	Support Vector Machine
TABULA	Typology Approach for Building Stock Energy Assessment
THG	Treibhausgasemissionen
VAE	Variational Autoencoder
WG	Wohngebäude
WPG	Wärmeplanungsgesetz

Zusammenfassung

Energieausweise informieren über den Energiebedarf und -verbrauch von Gebäuden. Für die Erstellung eines Energieausweises werden gebäudespezifische Daten benötigt, weshalb sie oft nicht für alle Gebäude vorliegen oder erst durch eine Begehung vor Ort erfasst werden können. Die vorliegende Arbeit untersucht die Möglichkeit, durch die Identifikation eines Datenschemas, basierend auf einer integrativen Analyse von Energieausweisen, Energiebedarfsvorhersagen für unbekannte Wohngebäude zu treffen. Die Zielsetzung der vorliegenden Arbeit besteht in der Identifikation wesentlicher Merkmale aus offenen Datenquellen, welche den Energiebedarf signifikant beeinflussen sowie deren Integration in ein kompaktes und effizientes Regressionsmodell. Dabei werden verschiedene maschinelle Lernmethoden, insbesondere das Extreme Gradient Boosting (XGB), eingesetzt, um Modelle zu entwickeln und zu validieren, die auf Daten aus Energieausweisen basieren. Dafür werden unter anderem Merkmale aus den beschriebenen Merkmalen zum Dach, der Außenwände, zu Fenstern und zum Boden extrahiert und in neue Merkmale überführt. Dies sind unter anderem Dach- und Wandtyp, das zugehörige Isolationsniveau und der Verglasungsgrad der Fenster. Im Anschluss werden drei Datensätze entwickelt und auf ihre Leistung untersucht. Eine Analyse der Merkmalsrelevanz hat gezeigt, dass über alle Datensätze hinweg bestimmte Merkmale wie Gebäudetyp, Baujahr, Wohnfläche, Dämmungsgrad und geografische Lage entscheidenden Einfluss auf die Vorhersage des Energiebedarfs haben. Das auf den Fernerkundungsdaten basierte Modell, wies nach Optimierung ein Bestimmtheitsmaß R^2 von 0,64 und einen mittleren absoluten Fehler (MAE) von 4,12 auf. Der Fehler bezieht sich auf eine Effizienzskala von 1-100 (Energieklasse G-A). Nach Skalierung der Pipeline und durch Ergänzung weiterer Datenpunkte, konnte der Wert auf 0,84 erhöht werden. Durch die Ergänzung von vebrauchsabhängigen Merkmalen, erreichte das XGB-Regressionsmodell ein R^2 von 0,94 und einen MAE von 1,46 über den Trainings- und Testdatensatz. Zur weiteren Validierung werden die Auswirkungen der einzelnen Merkmale über Shapley-Werte quantifiziert, um die Auswirkungen der Merkmale bei der Vorhersage zu interpretieren. Die entwickelten Modelle erreichten eine hohe Prognosegenauigkeit und demonstrierten eine signifikante Verbesserung gegenüber herkömmlichen Methoden. Die Erstellung der Datensätze erfolgte unter Verwendung der Programmiersprache Python sowie des Frameworks Scikit-learn (Version 1.4.1) zur Entwicklung der Modelle. Die erzeugten Datensätze und Modelle wurden in eine reproduzierbare Pipeline überführt und stehen nach Freigabe unter GitHub zur Verfügung.

Abstract

Energy performance certificates provide information on the energy requirements and consumption of buildings. Building-specific data is required for the creation of an energy performance certificate, which is why it is often not available for all buildings or can only be recorded through an on-site inspection. This thesis investigates the possibility of making energy demand predictions for unknown residential buildings by identifying a data schema based on an integrative analysis of energy performance certificates. The objective of this thesis is to identify key features from open data sources that significantly influence energy demand and to integrate them into a compact and efficient regression model. Various machine learning methods, in particular Extreme Gradient Boosting (XGB), are used to develop and validate models based on data from energy performance certificates. In addition, features are extracted from the descriptive features for the roof, wall, windows and floor and transferred into new features. Subsequently, three data sets are developed and analysed for their performance. An analysis of feature relevance has demonstrated that across all data sets, specific features, such as building type, year of construction, living space, degree of insulation and geographical location, exert a significant influence on the prediction of energy demand. Following optimisation, the model based on remote sensing data exhibited a coefficient of determination R^2 of 0.64 and an mean absolute error (MAE) of 4.12. The error refers to an efficiency scale of 1-100 (energy class G-A). Following scaling of the pipeline and the addition of further data points, the value increased to 0.84. The XGB regression model, which incorporates consumption-dependent characteristics, achieved an R^2 of 0.94 and an MAE of 1.46 across the training and test data set. To further validate the model, the effects of the individual characteristics are quantified using Shapley values, which enable the interpretation of the characteristics' effects on the prediction. The developed models achieved a high prediction accuracy and demonstrated a significant improvement over conventional methods. The data sets were created using the Python programming language and the Scikit-learn framework (version 1.4.1) to develop the models. The generated data sets and models were transferred to a reproducible pipeline and are available on GitHub after release.

1 Einleitung

Die wachsende Anzahl nationaler und internationaler Klimaschutzabkommen in den letzten Jahren zeigt, wie ernst politische Entscheidungsträger die Reduzierung der CO_2-Emissionen nehmen. Ein Beispiel ist das Pariser Klimaschutzabkommen aus dem Jahr 2015, dass viele Nationen in konkrete nationale Zusagen umgesetzt haben. Ebenso hat die Europäische Union (EU) ihre Ziele für 2020 auf neue, anspruchsvollere Ziele für 2030 angehoben, um ihre Verpflichtungen aus dem Pariser Abkommen in die Tat umzusetzen. Zur Verringerung des Energieverbrauchs und der CO_2-Emissionen werden verschiedene Sektoren in Betracht gezogen, darunter Landwirtschaft, Verkehr, Industrie und Gebäude. Letztere ist besonders relevant, da der Gebäudesektor 40% des Energieverbrauchs in der EU ausmacht [1]. Als Reaktion darauf hat die EU die Energy Performance Building Directive (EPBD) [2] eingeführt, die Vorgaben für neue Gebäude und die Sanierung bestehender Gebäude umfasst. Da Neubauten jährlich nur etwa 1% des Bestands ausmachen, ist die Energieeinsparung bei bestehenden Gebäuden deutlich wichtiger. Ein Schlüssel dazu ist eine Kennzeichnung, die Hausbesitzern und Entscheidungsträgern Einblicke in die aktuelle Effizienz und Verbesserungsmöglichkeiten der Gebäude bietet. In der EU ist diese Kennzeichnung als Energieausweis (EA) bekannt. Eine möglichst genaue Abschätzung des gegenwärtigen und potenziellen Energiebedarfs von (Wohn-)Gebäuden ist ein wichtiger Faktor für die Reduzierung von Treibhausgas (THG)-emissionen durch Verbesserung der Energieeffizienz. Die genaue Bestimmung des Energiebedarfs ist jedoch eine komplexe Aufgabe, die von vielen Parametern abhängt, wie z.B. dem Gebäudetyp, der Bauweise, der Lage, der Nutzung, dem Klima und dem Nutzerverhalten. Die herkömmlichen Methoden zur Energiebedarfsbestimmung basieren auf normativen Annahmen, empirischen Formeln und DIN-Normen, die ungenau, veraltet oder zeitaufwendig zu erfassen sind. Zudem erfordern sie einen hohen Aufwand an Zeit, Eingabe-Daten und Expertenwissen.

Hintergrund und Relevanz

Wie bereits in Kapitel 1 beschrieben, nehmen Bestandsgebäude einen großen Anteil des Energieverbrauchs ein. Innerhalb Deutschlands sind gegenwärtig das Gebäudeenergiegesetz (GEG) und die kommunale Wärmeplanung (KWP) die maßgeblichen Faktoren in der aktuellen Entwicklung, darunter auch im Gebäudesektor [3, 4]. Das Wärmeplanungsgesetz ist gemeinsam mit der Novelle des Gebäudeenergiegesetzes am 1. Januar 2024 in Kraft getreten. Beide Gesetze tragen dazu bei, die Klimaziele im Jahr 2045 zu

erreichen. Bis 2045 müssen dann lt. dem Wärmeplanungsgesetz (WPG) alle Wärmenetze klimaneutral sein. Die Energiequellen für das Netz müssen zu 100% regenerativ sein, um die Anforderungen zu erfüllen. Das WPG enthält Mindestziele für den Anteil von Wärme aus Erneuerbaren Energien und unvermeidbarer Abwärme. Es legt den Rahmen für die schrittweise Dekarbonisierung und den Ausbau der Fernwärme fest. Für neue Wärmenetze soll gelten: Bereits ab dem 1. Januar 2024 müssen in jedes neue Wärmenetz mindestens 65% erneuerbare Wärme eingeleitet werden. Das GEG enthält Anforderungen an die energetische Qualität von Gebäuden, die Erstellung und die Verwendung von Energieausweisen sowie an den Einsatz erneuerbarer Energien bei der Wärmeversorgung von Gebäuden. Zielsetzung des Gesetzes ist es, einen wesentlichen Beitrag zur Erreichung der nationalen Klimaschutzziele zu leisten [3, 4].

Um den Anforderungen der Gesetzgebung gerecht zu werden, bedarf es einer datengetriebenen Entscheidungsunterstützung auf Gebäudeebene und in Folge dessen die Aggregation und Unterteilung in Versorgungsgebiete. Auch wenn die Kommunen für die Erabeitung der Wärmepläne auf vorhandene Daten von Behörden, Energieversorgern und Schornsteinfegerdaten zurückgreifen können (siehe Anlage 1 zu § 15) [4], ist der öffentliche Gebäudedatenbestand im Vergleich zu anderen EU-Ländern als verbesserungswürdig einzuschätzen [5] Für eine EU-weite Vergleichbarkeit der Gebäudetypologie wurde EU TABULA (Typology Approach for Building Stock Energy Assessment) entwickelt [6]. Da es sich hier um ein Instrument der öffentlichen Politik handelt, gibt es Beschränkungen hinsichtlich der Daten und der Methode, da sich der Gebäudebestand seit Veröffentlichung im Jahr 2010 verändert hat. Daher kann es sein, dass die Typologien in TABULA heutzutage den Energieverbrauch überschätzen und das Isolationsniveau von Gebäuden unterschätzen [6].

Zielsetzung und Fragestellung

Ziel dieser Arbeit ist es, eine datengetriebene Methode zur Energiebedarfsbestimmung von Wohngebäuden (WG) auf Basis von EA zu entwickeln, die auf maschinellen Lernmethoden beruht. Anhand von öffentlichen EA-Datenbanken werden die wesentlichen Faktoren (Feature's) identifiziert, die den größten Einfluss auf den Energiebedarf bzw. -verbrauch des Gebäudes haben. Dabei ist festzustellen, welche dieser identifizierten Faktoren über (öffentliche) Daten akquiriert bzw. durch alternative Datenquellen (direkte/indirekte Proxy-Daten) oder über Fernerkundungsdaten (FED) bezogen werden können. Anhand der Untersuchung, kann ein Vorschlag für ein Datenschema zur Vorhersage des Energiebedarfs von bestehenden WG auf der Grundlage von offenen Daten getroffen werden. Darüber hinaus soll die Frage beantwortet werden, ob die Entwicklung eines Datenschemas mit relevanten Parametern, basierend auf einer integrativen Analyse (offener) Datenquellen, die Genauigkeit von Energiebedarfsprognosen von WG verbessern kann.

Aufbau der Arbeit

Die Arbeit ist in vier Kernbereiche gegliedert:

1. Theoretische Grundlagen und verwandte Arbeiten
2. Datenbeschaffung und -analyse (EDA)
3. Modellierung
4. Validierung der Modellergebnisse

Im ersten Schritt sind die Grundlagen erörtert, die zum weiteren Verständnis der verwendeten Methoden beitragen und den aktuellen Stand der Technik wiedergeben. Der zweite Bereich befasst sich mit der Datenakquise zu Datensätzen und die relevanten Parameter von WG aus verschiedenen Quellen gesammelt, wie z.B. EA, Gebäudedatenbanken, Wetterdaten etc.. Die Daten werden dann nach gängiger Praxis analysiert, aufbereitet, gefiltert und normalisiert, um sie für die Modellbildung nutzbar zu machen. Die dafür benötigten Schritte sind im Kapitel 3 erörtert und dokumentiert, um die Ergebnisse reproduzierbar zur Verfügung zu stellen. Im dritten Schritt werden verschiedene maschinelle Lernmodelle angewendet, um den Energiebedarf von WG zu schätzen. Die Modelle werden nach ihrer Leistungsfähigkeit, Robustheit und Interpretierbarkeit ausgewählt und verglichen. Im letzten Schritt werden die Modelle anhand von realen Testdaten validiert, um ihre Genauigkeit und Zuverlässigkeit zu bewerten. Die Validierung erfolgt durch den Vergleich der vorhergesagten Werte mit den tatsächlichen Werten oder mit den Werten aus anderen Methoden. Die Validierung kann auch durch Sensitivitätsanalysen oder Fehleranalysen ergänzt werden, um die Stärken und Schwächen der Modelle zu identifizieren. Zuletzt werden anhand der Ergebnisse relevante Datenquellen identifiziert, die für einen Vergleich auf deutscher Ebene benötigt werden.

2 Theoretische Grundlagen

Dieses Kapitel beschreibt den aktuellen Stand der Technik und vermittelt dem Leser die Grundlagen der Energiebedarfsbestimmung von WG.

2.1 Energiebedarf von (Wohn-)Gebäuden

Dieser Abschnitt beschreibt anhand einer systematischen Literaturrecherche aktuelle Methoden und Definitionen für die Energiebedarfsbestimmung von Gebäuden. Dabei werden die einzelnen Ansätze hinsichtlich der Vor- und Nachteile gegenübergestellt.

2.1.1 Definition und Berechnungsmethoden

Grundsätzlich lassen sich anhand der Literaturrecherche die Energiebedarfsbestimmgung von (Wohn-)Gebäuden durch vier Ansätze identifizieren [7, 8]:

- White-Box oder physikalische Modelle
- Black-Box oder statistische Modelle
- Grey-Box oder Ersatzmodelle
- Gradtagsmodelle

Physikalische (White-Box-)Modelle verwenden Energiebilanzgleichungen und eine Vielzahl von Eingabeparametern, um den Energiebedarf eines Gebäudes zu simulieren. Dieser detaillierte Ansatz ist von unschätzbarem Wert für die Optimierung der Gebäudeleistung in der Planungsphase oder bei der Nachrüstung. Die Programmierung von physikalischen Modellen ist jedoch zeit- und datenaufwändig und erfordert beträchtliche Kenntnisse des Tools [7]. Das erschwert die Nutzung dieser Modelle und ihre Skalierung über einzelne Gebäude hinaus. Es gibt immer mehr Belege dafür, dass große Diskrepanzen zwischen simuliertem und gemessenem Energiebedarf auftreten können, wenn diese Modelle ohne Kalibrierung für Vorhersagen verwendet werden [8, 9]. Die Leistung wird zudem verbessert, wenn die Ergebnisse der physikalischen Modelle mit gemessenen Energiebedarfsdaten kalibriert werden [8].

Statistische (Black-Box-)Modelle leiten rein mathematische Beziehungen zwischen Messungen des Energiebedarfs und anderen Merkmalen ab. Diese abgeleiteten Muster werden anhand der erlernten Gewichtungen für die Vorhersage unbekannter Bedarfe verwendet

[10, 11]. Diese Modelle liefern schnelle und genaue Ergebnisse, sind jedoch nur im Rahmen ihrer Trainingsdaten gültig und lassen sich daher nur schwer verallgemeinern oder an andere Kontexte anpassen. Zu den gebräuchlichsten Modellen gehören Artificial Neural Networks (ANN), Support-Vektor-Maschinen (SVM), multiple lineare Regression, Gradient Boosting (GB) und Random Forests (RF) [7].

Esatzmodelle (Surrogatmodelle) sind statistische Modelle, die anhand von Eingangs- und Ausgangsdaten physikalischer Modelle trainiert werden [12, 13]. Mit einem ausreichend großen und verallgemeinerten Trainingssatz könnte ein einzelnes Ersatzmodell theoretisch verwendet werden, um den Energiebedarf für eine umfassende Reihe von Gebäudekonstruktionen und Klimazonen schnell zu simulieren. Ein solches Modell wäre sowohl verallgemeinerbar als auch anpassbar, würde aber dennoch umfangreiche Eingabedaten und Kenntnisse erfordern, um korrekt verwendet zu werden. Surrogatmodelle würden unter denselben potenziellen Ungenauigkeiten leiden wie die physikalischen Modelle, auf denen sie beruhen, wenn sie falsch eingerichtet oder nicht kalibriert sind [12, 14].

Gradtags-Modelle bieten einen vereinfachten Ansatz zur Modellierung der Energienachfrage für Heiz- und Kühltage (Heating-Degree-Days - HDD / Cooling-Degree-Days - CDD). Diese können jedoch nur die Gesamtnachfrage für Heizung und Kühlung aufschlüsseln und nicht mehrere einzelne Endnutzungspfade. Dabei korreliert der Bedarf an Raumheizung und -kühlung stark mit der Außentemperatur. Dieser Wert ist bei hohen Auflösungen (z. B. stündlich in 0,5°-Gitterzellen) verfügbar [8]. Die Außentemperaturen werden in Gradtagen angegeben, die ausdrücken, um wie viel (in °C) und wie lange (in Tagen) die Außenlufttemperatur über oder unter einer Gleichgewichtspunkttemperatur liegt. Diese Temperatur ist der Bereich, in dem die internen Wärmegewinne eines Gebäudes die externen Verluste ausgleichen, so dass weder geheizt noch gekühlt werden muss. Die Gradtage werden genutzt, um synthetische Heiz- und Kühlbedarfsprofile abzuleiten und können als zusätzliches Merkmal für maschinelle Lernmodelle genutzt werden [7, 15]. Beispielsweise das *Hotmaps*-Projekt [16] schätzt den täglichen Heiz- und Kühlbedarf in 28 europäischen Ländern anhand der Temperatur, als einzigen meteorologischem Input ab. *When2Heat* [17] schätzt stündliche Heizprofile für 16 europäische Länder anhand von Temperatur und Windgeschwindigkeit. In Tabelle 2.1 ist ein Vergleich der methodischen Ansätze hinsichtlich der Vor- und Nachteile beschrieben, exklusive der Gradtagsmethode.

Im Rahmen der Arbeit liegt der Fokus auf den beschriebenen Black-Box-Modellen, die unter anderem auch maschinelle/statistische Lernmethoden beinhalten. Jedoch soll das Modell und die dazugehörigen Vorhersagen hinsichtlich seiner Transparenz und Interpretierbarkeit dargestellt werden.

Tabelle 2.1: Die Vor- und Nachteile der drei wichtigsten großmaßstäblichen Ansätze zur Vorhersage des Energiebedarfs von Gebäuden in Anlehnung an Gassar et al. [7]. *Hinweis: Der Begriff „makroökonomische Effekte“ wird verwendet, um auf die potenziellen Veränderungen in Faktoren bezogen zu sein, die mit Einkommen, Gebäudepreisen, der Bevölkerung und dem Wohnungszensus zusammenhängen.*

Item	Black-Box	White-Box	Grey-Box
Vorteile	• Eingangsdaten sind historische Daten • Hohe Ausführungsgeschwindigkeit, außer bei Support Vector Machine • Einbeziehung von makroökonomischen Effekten • Bewältigung von Eingangsdaten für lineare und nichtlineare Probleme außer bei Regression	• Eingangsdaten sind physische Informationen • Ergebnisse können in physischen Begriffen interpretiert werden • Keine Trainingsdaten erforderlich • Explizite Darstellung von Endenergieverbräuchen • Hohe Genauigkeit	• Eingangsdaten sind physische Informationen und historische Daten • Ergebnisse können in physischen Begriffen interpretiert werden • Einbeziehung von makroökonomischen Effekten • Explizite Darstellung von Endenergieverbräuchen • Sehr hohe Genauigkeit
Nachteile	• Abhängigkeit von historischen Daten • Mehrere Schwierigkeiten bei der Interpretation der Ergebnisse in physischen Begriffen • Große Menge an Trainingsdaten erforderlich • Niedrige Ausführungsgeschwindigkeit mit Support Vector Machine • Niedrige Genauigkeit bei Regression • Keine explizite Darstellung von Endverbrauchen	• Detaillierte physische Eingangsinformationen erforderlich • Repräsentative Gebäude • Annahme des Bewohnerverhaltens • Keine wirtschaftlichen Faktoren • Nicht einfach zu verwenden, benötigt Erfahrung • Geschwindigkeit ist mittel	• Eine ungefähre Beschreibung des Gebäudes ist erforderlich • Nicht einfach zu verwenden, benötigt Erfahrung • Geschwindigkeit niedrig

2.1.2 Einflussfaktoren und Kennwerte

Verwandte Arbeiten haben sich bereits mit dem Thema der Einflussfaktoren beschäftigt [7, 18, 19]. Gassar et al. [7] hat in seinem Review die Faktoren zusammengefasst, die innerhalb verwandter Arbeiten als signifikant und relevant eingestuft wurden.

Klimafaktoren

Temperatur: Die Temperatur beeinflusst die Leistung der Gebäudehülle aufgrund des Wärmeübergangs. Wenn die Außentemperatur aufgrund des Klimawandels steigt, nimmt der Kühlbedarf der Gebäude zu. **Relative Feuchtigkeit:** Die relative Feuchtigkeit beeinflusst das Niveau des thermischen Komforts in Innenräumen, insbesondere in Regionen mit hoher Luftfeuchtigkeit. **Solareinstrahlung:** Die Variation der solaren Einstrahlung führt zu Schwankungen sowohl in der Lufttemperatur, als auch in der relativen Feuchtigkeit, was wiederum Änderungen im Energiebedarf von Gebäuden bewirkt.

Gebäudefaktoren

Wohnraumfläche: Der Effekt einer größeren Fläche ist mit einer Steigerung des Energiebedarf durch Zunahme von Heiz- und Kühlbereichen in Gebäuden verbunden, wobei eine 1% Zunahme der Gebäudegröße zu einer 0,61 %-igen Steigerung des Energieverbrauchs führt. **Gebäudealter:** Die Literatur berichtet, dass das Gebäudealter einen Einfluss auf den Energieverbrauch hat und dass ältere Gebäude die 6,1% mehr Heizenergie pro Woche benötigen verglichen mit neuen oder sanierten Gebäuden. An der Stelle sollte auch der Sanierungsstand mit einbezogen werden. **Gebäudetyp:** Es wurde beobachtet, dass Haushalte in Doppel- und Reihenhäusern 6,9% weniger Energie pro Woche verbrauchen, als solche in freistehenden Häusern. Walter und Sohn [20] verwendeten Gebäudemerkmale wie Grundfläche, Gebäudetyp, Gebäudealter, Wandtyp, Fenstertyp für eine Stichprobe von 870.000 Gebäuden in den USA als Eingangsdaten für Modelle zur Entwicklung eines Prognosemodells, dass Gebäudeeigentümern und politischen Entscheidungsträgern Schätzungen der erwarteten Energieeinsparungen liefert, wenn bestimmte Anlagen in Wohn- und Geschäftsgebäuden geändert werden. Die Schlussfolgerung war, dass die Entwicklung eines Modells, das sowohl Daten über die Gebäudeeigenschaften als auch über den empirischen Nachrüstungsprozess kombiniert, die Möglichkeiten für zukünftige Verbesserungen in Gebäuden erhöhen kann.

Sozioökonomische Faktoren

Anzahl der Bewohner: Es besteht eine signifikant positive Beziehung zwischen der Anzahl der Bewohner und dem Verbrauchsniveau in Gebäuden, wobei ein zusätzlicher Bewohner in Familien den Energieverbrauch um etwa 21% erhöht. **Einkommen:** Der Effekt des Einkommens auf den Energieverbrauch ist mit größeren, wärmeren Gebäuden

verbunden, die mehr Geräte enthalten. **Mietverhältnis:** In verwandten Arbeiten wird der Einfluss der Eigentumsverhältnisse auf den Energiebedarf dem Gebäudeeigentum zugeschrieben, während andere Faktoren auf gemietete Gebäude zurückgeführt werden.

Geräte und sonstige Ausrüstung (HVAC)

Verschiedene Haushaltsgeräte, wie Computer, Ventilatoren, Waschmaschinen, Kühlschränke usw. beeinflussen den Energieverbrauch abhängig von der Anzahl und Art der Geräte, der Leistungsnachfrage und der Nutzungshäufigkeit.

Geografische Faktoren

Geografische Lage: Der Einfluss der geografischen Lage auf den Energiebedarf wird den Unterschieden in der Anzahl der Heiz- und Kühltage in der entsprechenden Region zugeschrieben. **Urbanisierung:** Weiterhin wurde beobachtet, dass der Grad der Verstädterung den Energiebedarf hinsichtlich Kühlung erhöht und städtische Wärmeinseln aufgrund der Bebauungsdichte vergrößert.

2.1.3 Energieausweise als Informationsquelle

Die Verfügbarkeit und der Umfang öffentlich zugänglicher Daten in der EA-Datenbank variieren signifikant zwischen den Mitgliedstaaten. Eine Übersicht zu den öffentlichen Zugängen von EA-Datenbanken ist in Abbildung 2.1.3 dargestellt. In einigen Ländern wie Dänemark, Estland, Ungarn, Irland, Litauen, den Niederlanden, Portugal, Schweden, Teilen des Vereinigten Königreichs (England und Wales) und Norwegen werden spezifische EA-Informationen direkt aus der Datenbank bereitgestellt. Im Gegensatz dazu beschränken sich andere Länder, darunter Belgien-Flandern, Griechenland, Kroatien, Ungarn und Rumänien, auf die Veröffentlichung von aggregierten Daten. Zudem wird in einigen Ländern der Zugriff auf EA-Informationen auf Antrag von Drittparteien, vorwiegend zu Forschungs- und teilweise zu kommerziellen Zwecken, gewährt. Länder wie Bulgarien, Deutschland, Finnland, Malta und Zypern bieten keinen öffentlichen Zugang zur EA-Datenbank. Bis auf Deutschland und Rumänien verfügen alle Länder über eine EA-Datenbank und generieren daraus Daten, um Aussagen über den Gebäudebestand treffen zu können. Hier werden im Rahmen einer Stichprobenprüfung die ausgestellten EA beim *Deutschen Institut für Bautechnik* (DIBt) auf Plausibiliät geprüft [21].

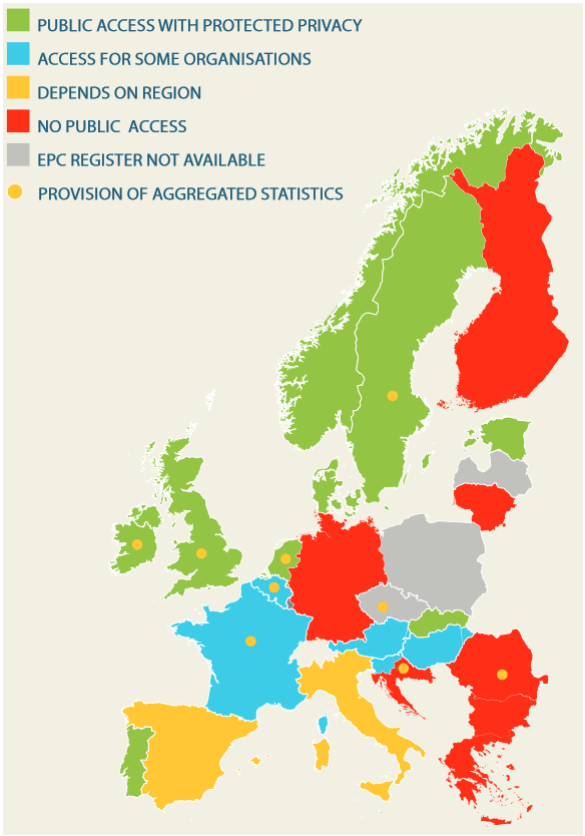

Abbildung 2.1: Öffentlicher Zugang zu den EA-Datenbanken in der EU

Ein Auszug der verfügbaren EA-Daten, die während dieser Arbeit abgerufen werden, finden sich in Tabelle 2.1.3 wieder. Teilweise sind die Zugänge bzw. der Download der Daten erst mit einer Registrierung möglich, andere widerum haben eine integrierte Schnittstelle (API), um die Daten direkt maschinenlesbar oder als .csv-Datei abzurufen.

Tabelle 2.2: Auszug zu den Zugängen von EA-Datenbanken innerhalb der EU

Land	Verfügbarkeit	URL-Link
England	Nach Registrierung/Anmeldung stehen mehrere Optionen zur Verfügung	`https://epc.opendatacommunities.org/docs/index`
Niederlande	Für Forschungseinrichtungen	`https://www.ep-online.nl/`
Dänemark	Einzelansicht der Gebäude möglich	`https://old.sparenergi.dk/demo/addresses/map`

Es gibt aber auch alternative Ansätze, wie z.B. das Projekt *Forschungsdatenbank Nichtwohngebäude* (ENOB:dataNWG) [22] oder der *IWU Datenbestand Wohngebäude 2016* [23], die mit Stichprobenansätzen viele Merkmale, insbesondere im energetischen Bereich, abdecken. Dies gilt auch für empirische Ansätze mit Datenspenden (z.B. CO_2-Online) [24]. Allerdings ist hier die Nachhaltigkeit aufgrund der abnehmenden Datenspenden fraglich, so dass die ENOB:dataNWG-Methode den empirischen Ansätzen (basierend auf Datenspenden) vorzuziehen ist [5].

2.1.4 Quantifizierungsmethoden anhand von Energieausweisen

In der Literatur und Praxis werden verschiedene Aspekte von EA diskutiert [25–27]. Neben der Untersuchung, inwieweit EA den Immobilienmarkt beeinflussen sowie der Auswirkung und Relevanz von EA auf Sanierungs- und Kaufentscheidungen, stellt die Energieeffizienzlücke, also der Unterschied zwischen Bedarf und Verbrauch eine große Herausforderung dar [26]. Die Energieeffizienlücke beschreibt das Phänomen, dass der tatsächlich gemessene Endenergieverbrauch erheblich vom vorhergesagten Endenergieverbrauch abweicht. Weitere Arbeiten beschreiben Abweichungen von bis zu 287% [9]. Wenninger und Wiethe [15] beschreiben eine Möglichkeit, diese Lücke zu schließen. Diese besteht darin, datengetriebene Quantifizierungsmethoden, anstelle von berechneten Bedarfen zu verwenden. Ein Vorteil bei der Anwendung datengetriebener Methoden ist, dass kein Domänenwissen über die physikalische Eigenschaften der Gebäude benötigt wird [28]. Eine weitere Möglichkeit, diese Lücke zu minimieren, ist ein Bedarfs-Verbrauchs-Vergleich, der in Beiblatt 1 der DIN V 18599 [29] für die Sanierungsberatung geregelt ist, jedoch kein Bestandteil der offiziellen EA ist [15]. Die Norm definiert Kennzahlen und Zusammenhänge, um den berechneten Bedarf schrittweise an den gemessenen Verbrauch anzunähern und so die Leistungslücke durch verbesserte Sanierungsentscheidungen zu minimieren [30]. Der datengetriebene Ansatz mit den gemessenen Verbrauchsdaten hatte einen um 50% geringeren Fehler, als die Vorhersagen bei dem berechneten Bedarfswert

[15]. In Deutschland bildete die Energieeinsparverordnung (EnEV) den rechtlichen Rahmen für EA mit der Endenergieeffizienz als Zielgröße. Diese Verordnung wurde mit dem GEG 2020 abgelöst. Grundsätzlich wird ein EA entweder durch Messung (Verbrauchsausweis) oder durch Berechnung (Energiebedarfsasuweis) erstellt. Verbrauchsausweise spiegeln den tatsächlich gemessenen Jahresverbrauch aller Energiequellen wider, die in den letzten drei aufeinanderfolgenden Jahren zur Heizung, Belüftung und Kühlung eines Hauses beigetragen haben, und berücksichtigen damit auch direkt das Verhalten der Bewohner. Die Energiebedarfsausweise spiegeln den Energiebedarf wider und bestimmen den Endenergiebedarf durch eine technische Analyse einer Vielzahl von Gebäudeparametern [15]. Um die erforderlichen Informationen für die Durchführung von rechnerischen EA zu sammeln, sind Vor-Ort-Begehungen durch qualifizierte Auditoren erforderlich. Die deutsche Normung DIN V 18599 ist das Standardberechnungsschema zur Ermittlung des Endenergiebedarfs von Gebäuden [29]. Für WG konnte bis Ende des Jahres 2023 auch die Norm DIN V 4108-6 in Kombination mit DIN V 4701-10 oder DIN V 4701-12 angewendet werden. Nach den aktuellen Richtlinien müssen für fast zwei Drittel aller WG in Deutschland EA berechnet werden. Ein weiteres Beispiel beschreibt Hettinga et al. [31], die ein maschinelles Lernmodell mit offenen Daten aus den Niederlanden kombinierte. Die Validierung erfolgte mit dem TABULA-Modell [6]. Ergebnis der Untersuchung war, dass der RF-Klassifizierer, der auf offenen Daten trainiert wurde, das TABULA-Modell mit einer Genauigkeit von 71% gegenüber 26% bei der korrekten Vorhersage der Energiekennzeichnung der Gebäude übertraf. Zu den offenen Daten gehören das Baujahr, der Gebäudetyp, die Grundfläche und Daten aus dem niederländischen Adress- und Gebäuderegister (BAG) und dem niederländischen Statistikamt (CBS). Wie ebenfalls von Ahmad [32] vorgeschlagen, können datengetriebene Methoden neue Optionen für Energiezertifizierungsmaßnahmen bieten. Eine zunehmend beliebte datenwissenschaftliche Technik ist der Einsatz von künstlicher Intelligenz (KI), die von Khayatian et al. [33] erfolgreich angewendet wurde. Dabei wurden neuronale Netze mit einem sehr detaillierten Datensatz von Gebäudeeigenschaften kombiniert, um das Energielabel einzelner Gebäude vorherzusagen. Das Modell weist eine Genauigkeit von 95% auf, die innerhalb von ±3 Konfidenzintervallen des gemessenen Wärmebedarfsindikators in [kWh/m^2] liegt. Für das Training des Algorithmus wurden jedoch zumindest einige der detaillierten Merkmale verwendet, die für die direkte Bestimmung des Energielabels von Gebäuden verwendet werden. Diese Merkmale sind meist nur für Gebäude verfügbar, die bereits einen EA besitzen.

2.1.5 Erfassung energetischer Merkmale über Fernerkundungsverfahren

Zu den Fernerkundungsverfahren zählen zum einen satellitengestützte Verfahren, zum anderen aber auch alle Informationen zu Gebäude- und Siedlungscharakteristika, die auf verschiedenen Ebenen bei Bund, Ländern und Kommunen vorliegen und ausgewertet werden. Dies umfasst alle Geobasisdatenbestände des Bundesamtes für Karto-

graphie und Geodäsie in all ihren Detaillierungsgraden und Ausbildungen, Daten der ESA und der NASA aus unterschiedlichen Satellitenaufnahmen sowie die Produkte des Copernicus-Programms (www.copernicus.eu), welche zur Überwachung von Land, Meer, Atmosphäre und Klima vorgehalten werden. Darüber hinaus werden auf kommunaler Ebene vielfach zusätzliche Geodatenbestände geführt, beispielsweise Informationen zur tatsächlichen Nutzung von Gebäuden, wie sie bspw. für öffentlich zugängliche Gebäude in Stadtplänen verzeichnet werden [5]. Die Verwendung von Fernerkundungsdaten, insbesondere hochauflösenden Luft- und Satellitenbildern, hat sich als wertvolle Quelle für die Extraktion von Gebäudemerkmalen und die Anreicherung von Daten erwiesen, die für die Modellierung des Energiebedarfs städtischer Gebäude erforderlich sind. Die Fernerkundungstechnologie kann, insbesondere bei der Anwendung im Bausektor, eine Vielzahl wertvoller Merkmale und Daten liefern. In mehreren Arbeiten [34–37] konnte der Einsatz u.a. von Deep-Learning-Techniken, insbesondere von Faltungsneuronalen Netzen (engl. Convolutional neural networks (CNNs)), für die genaue Extraktion von Gebäudegrundrissen, Höhen und 3D-Geometrien aus Fernerkundungsbildern nachgewiesen werden [35]. Diese geometrischen Informationen bilden die Grundlage für die Bottom-up-Modellierung des Wärmebedarfs von Gebäuden auf der Grundlage von Bauarten und Bauzeiten. Die geometrischen Faktoren stellen wie bereits in Kap. 2.1.2 vorgestellt einen wesentlichen Aspekt bei der Modellierung des Energiebedarfs dar. Die Schätzung von Gebäudetypen und -alter ist jedoch ebenfalls von entscheidender Bedeutung [38, 39]. Deep-Learning-Methoden, als auch die Anwendung von Random Forest haben vielversprechende Ergebnisse bei der Klassifizierung von Gebäudetypen aus Luftbildern gezeigt [35, 40]. Maschinelle Lernverfahren hingegen sind in der Lage, das Alter von Gebäuden durch die zusätzliche Analyse von Bildmerkmalen zu bestimmen [34, 37]. Weiterhin kann durch die Nutzung von FED die Bewertung des Potenzials für Photovoltaikanlagen auf Dächern erfolgen. Dieses Potenzial ist eng mit dem externen Energiebedarf von Gebäuden verbunden. Weitere Techniken, wie die Erstellung von Orthofotos aus Drohnenbildern und die Verarbeitung von LiDAR-Daten ermöglichen dabei die genaue Vermessung geeigneter Dachflächen oder die Rekonstruktion unter Berücksichtigung von Neigung, Azimut[1] und Abschattungseffekten [41–43]. Die aus den FED extrahierten Gebäudedaten werden anschließend in die Arbeitsabläufe der Energiemodellierung integriert [44]. So können beispielsweise durch Verwendung des Bautyp und Baualtersklasse und Hinzunahme der TABULA-Methodik [45] die Modellierung des Wärmebedarfs von WG erfolgen [34], während die Landoberflächentemperatur auf Basis von FED mit lokalen Klimazonen und Gradtagszahlen kombiniert wurde, um die räumliche Verteilung des Energiebedarfs zu schätzen. Zu den wesentlichen Herausforderungen bei der Verwendung von FED zählen der Bedarf an umfangreichen, gelabelten Trainingsdatensätzen sowie die Extraktion

[1] Azimut ist der Winkel zwischen der Nordrichtung und der Projektion des beobachteten Objekts auf die horizontale Ebene des Beobachtungspunktes, gemessen im Uhrzeigersinn von Norden aus. Dieser Winkel wird in Grad angegeben, wobei Norden 0° ist, Osten 90°, Süden 180° und Westen 270°

relevanter Merkmaler für den Anwendungsfall [43]. Krapf et al. haben beispielsweise auf Basis der EA-Daten von England und Wales die Merkmale mit LiDAR-Daten kombiniert [46]. Um die EA mit den Gebäudepunktwolken zu verknüpfen, haben Krapf et al. die Unique Property Reference Numbers (UPRN) verwendet, die für die meisten Gebäude in England existieren. Das Deutsches Zentrum für Luft- und Raumfahrt e.V. (DLR) gibt in einer Konzeptentwicklung an, dass durch die Kombination von Geobasisdaten und Fernerkundungsaufnahmen neue oder verbesserte Merkmale zum Gebäudebestand für Einzelgebäude bereitgestellt werden können [47]. Folgende Gebäudemerkmale können aus der Kombination gewonnen werden. Bauvolumen und Geschossflächenzahl: Das Bauvolumen kann aus der Grundfläche und zusätzlichen Höheninformationen der Gebäude berechnet werden. Die Höheninformationen können entweder aus einem Gebäudemodell mit CityGML Level of Detail 1 (LoD) oder aus einem Digitalen Oberflächenmodell (DOM), das durch Laserscanning erzeugt wurde, entnommen werden. Aus Gebäudehöhe, Grundfläche, mittlerer Geschosshöhe und Grundstücksfläche kann die Geschossfläche berechnet werden. Daraus kann auch die Bebauungsdichte abgeleitet werden [5]. Gebäudenutzung: Zur Bestimmung der Gebäudenutzung werden die Bilddaten mit anderen Modellen verknüpft, z.B. mit dem CityGML - LoD1-DE. Hier hat jedes einzelne Haus Attribute, aus denen auf die Funktion des Gebäudes geschlossen werden kann. Funktionen, die im ALKIS-Objektkatalog geführt werden, sind z.B. Wohngebäude, Parkhäuser und öffentliche Gebäude. Auf Basis der LoD2-Daten und der Gebäudehöhe kann die Gebäudenutzung (Wohnheim, Altenheim, Forsthaus etc.) den Unterklassen EFH/ZFH und MFH zugeordnet werden [47]. Daraus könnte im nächsten Schritt eine Aussage über die mögliche Bewohnerzahl abgeleitet werden. Eine weitere mögliche Quelle für Nutzungsdaten sind auch Open Street Map-Karten (OSM), allerdings sind die Nutzungsklassen dort nicht standardisiert [5]. Dachform: Die Dachform kann teilweise aus den Daten der Hausumringe entnommen werden [5]. Wo dies nicht möglich ist, können Informationen zur Dachform auch aus digitalen Modellen mit LoD2 entnommen werden. Darüber hinaus könnte die Dachform auch aus dem noch höher aufgelösten Digitalen DOM abgeleitet werden, was aber im Vergleich zu den Hausumringen oder dem LoD2-Modell vermutlich mit einem höheren Aufwand verbunden ist. Dachaufbauten, Solaranlagen und Solarflächenpotenzial: Mit hochauflösenden Fernerkundungsdaten, wie Luftbildern, aber auch Laserscan- (LiDAR) und Hyperspektraldaten besteht in begrenztem Umfang die Möglichkeit, Dachaufbauten und Solaranlagen zu identifizieren. Mit der Methode der spektralen Signatur können Solaranlagen und deren Ausdehnung auf Dächern bereits automatisiert (aber nur lokal) erkannt werden. Um diese Methode auf ganz Deutschland auszuweiten, kann eine KI eingesetzt werden. Als Eingangsdaten dienen hier DOP20 Luftbilder und die Hausumringe. Der Aufwand für die Erstellung eines solchen KI-Modells ist noch hoch und es müssen ausreichend Trainingsdaten zur Verfügung stehen. Zur Ermittlung des Solarflächenpotenzials auf Dächern werden LiDAR und hier insbesondere Airborne Laser Scanning (ALS) Daten ausgewertet. Dachflächen werden von Vegetation und an-

deren Gebäudeflächen unterschieden und hinsichtlich Neigung und Ausrichtung bewertet [47]. Fassadenauswertung: Die Erfassung von Fassaden ist mit herkömmlichen Fernerkundungsdaten nicht möglich. Hier können neue Methoden wie Google Streetview oder Schrägluftbilder eingesetzt werden. Mit diesen Bildern kann eine trainierte KI den Fensterflächenanteil der Fassade aus den CityGML-Daten von Berlin identifizieren [48, 49]. Dachmaterial: Zur Bestimmung des Dachmaterials von Gebäuden wird auf eine spektrale Bibliothek mit verschiedenen Materialien zugegriffen. Da die Bestimmung dieser Materialien nur mit hyperspektralen Daten möglich ist, werden nur bestimmte Gebiete dargestellt. Eine Möglichkeit, diese Methode auf ganz Deutschland auszuweiten, wäre der Einsatz von Satelliten wie EnMAP [47].

2.2 Maschinellen Lernmethoden zur Energiebedarfsbestimmung

Maschinelles Lernen ist ein Teilgebiet der künstlichen Intelligenz, das es ermöglicht, aus Daten, Muster und Zusammenhänge zu erkennen und Vorhersagen für ungesehene Daten zu treffen. Der Ansatz zielt darauf ab, die Leistung von Maschinen (Computer) zu steigern, indem sie anhand von Daten und Mustern, die als Input dienen, lernen, ohne sich an eine definierte Regelbindung zu halten. Die Regeln werden aus der Verteilung der Daten generiert, in einem Modell gebündelt und auf neue, unbekannte Daten angewendet. [50] Eine vereinfachte Darstellung zur Unterteilung von maschinellen Lernmethoden kann der Abbildung 2.2 entnommen werden. Die einzelnen Ansätze sind gesondert im Abschnitt 2.2.1 erörtert. Die Vorteile dieser Methode sind, dass sie flexibel, adaptiv und selbstlernend ist, und dass sie weniger Eingangsdaten und Expertenwissen benötigt, als die herkömmlichen Methoden, die im Abschnitt 2.1.1 beschrieben sind.

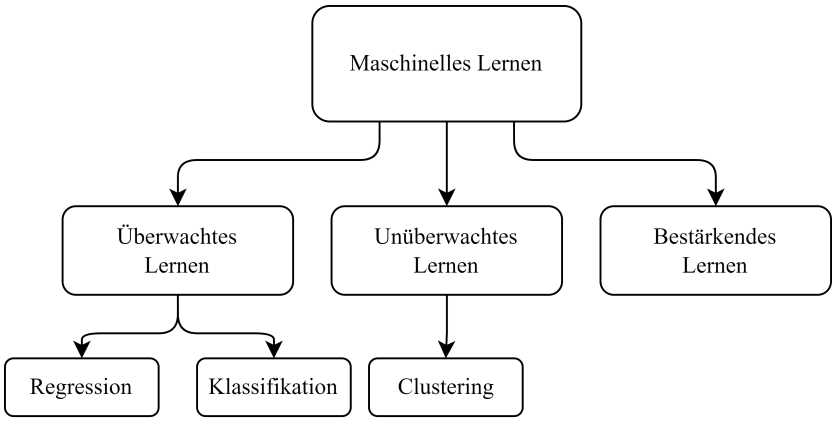

Abbildung 2.2: Methoden des Maschinellen Lernens

2.2.1 Klassifikation von maschinellen Lernmethoden

Es existieren viele verschiedene Arten von maschinellen Lernsytemen. Diese Systeme werden kurz genannt und erklärt, um die Wahl des Systems zu begründen. Grundsätzlich lassen sich ML-Systeme in drei Bereiche kategorisieren [50]. Da es sich bei der vorliegenden Arbeit um ein überwachtes Modell handelt, wird auf eine Definition der anderen Systeme verzichtet.

- Trainingsüberwachung
- Batch- und Online-Learning
- Instanz- und modellbasiertes Lernen

ML-Systeme lassen sich je nach Art und Umfang der Überwachung beim Training in verschiedene Kategorien einteilen. Die wichtigsten sind überwachtes Lernen, unüberwachtes Lernen, selbstüberwachtes Lernen, teilüberwachtes Lernen und verstärkendes Lernen (Reinforcement Learning) [28, 50]. Überwachtes Lernen kann weiter in Regressions- und Klassifizierungsalgorithmen unterteilt werden [51]. Ein Klassifizierungsalgorithmus wird verwendet, wenn die Ausgangsvariable ein Label darstellt, wie bspw. die Energiebewertung und der Gebäudetyp [52]. Regressionsalgorithmen werden verwendet, wenn die Ausgangsvariable ein realer, kontinuierlicher Wert ist, wie beispielsweise der Energieverbrauch [53]. Zu den gängigen überwachten Lernalgorithmen gehören k-Nearest-Neighbor (kNN), Naive Bayes (NB), Support Vector Machinen (SVM) und neuronale Netze (NN). Unüberwachtes Lernen wird angewendet, wenn keine entsprechenden Ausgangsvariablen für die Eingangsdaten vorhanden sind [51]. Einige gängige Algorithmen für unüberwachtes Lernen sind k-means-Clustering und Assoziationsregeln. Beim unüberwachten Lernen sind die Trainingsdaten nicht gelabelt. Das System versucht, ohne Anleitung zu lernen und eine Struktur in den Daten zu erkennen und bspw. Cluster in den Daten herauszustellen. Eine verwandte Aufgabe ist die Dimensionsreduktion, um die Dimensionen der Daten zu vereinfachen, ohne dabei allzu viele Information zu verlieren [50]. Diese Methodik kann auch als Extraktion von Merkmalen (engl. Feature Extraction) verstanden werden, indem mehrere Merkmale miteinander kombiniert werden. Die Wahl eines ML-Systems schließt die Einbindung eines anderen Systems nicht aus, wodurch eine Kombination innerhalb der Systeme möglich ist [50]. Bei der vorliegenden Datenanalyse werden ebenfalls unüberwachte Systeme mit überwachten Systemen kombiniert. Dies wird innerhalb des Kapitels 3 weiter erörtert.

2.2.2 Auswahl und Bewertung von Modellen

Bei der Auswahl eines Modells für maschinelles Lernen sollten folgende Faktoren berücksichtigt werden: Das zu lösende Problem und das gewünschte Ergebnis sollten klar und

deutlich definiert werden, um ein Modell auszuwählen, das für den speziellen Anwendungsfall geeignet ist. Die Kenntnis der Merkmale innerhalb der Daten ist von essenzieller Bedeutung, um das Verständnis zu gewährleisten. Die verfügbaren Daten sollten geprüft werden, ob sie für das zu lösende Problem geeignet sind. Hierbei ist es wichtig, die Verteilung der Daten sowie fehlende Werte und Ausreißer zu berücksichtigen. Im Anschluss sollten drei bis fünf Modelle gewählt werden, das für den entsprechenden Anwendungsfall geeignet ist. Jedes Modell hat dabei Vor- und Nachteile. Daher ist es wichtig, das Modell auszuwählen, das am besten zum Anwendungsfall passt. Für das Training und die Auswertung sollten die Daten in Trainings- und Testsätze aufgeteilt werden. Die Trainingsmenge dient dazu, um das Modell zu trainieren, und die Testmenge, um die Verallgemeinerungsleistung des Modells zu bewerten. Bei der Bewertung werden geeignete Metriken gewählt, um die Leistung des Modells zu bewerten. Anhand der Bewertungsergebnisse, können Bereiche identifiziert werden in denen das Modell verbessert werden kann. Dies kann die Anpassung von Hyperparametern, das Sammeln von weiteren Daten oder das Testen eines anderen Modells beinhalten. Sobald die Leistung des Modells zufriedenstellend ist, kann das Modell genutzt werden, um das ursprünglich definierte Problem zu lösen. Zusammenfassend stellt die Auswahl des richtigen Modells einen iterativen Prozess dar. Um die beste Lösung für das spezifische Problem zu finden, sollten mehrere Modelle ausprobiert und über den Ansatz neu iterieren [50]. Dabei sind Ensemble-Methoden wie RF und XGB, ANN und SVM die am häufigsten verwendeten maschinellen Lernalgorithmen für die Bestimmung Energieeffizienz von Gebäuden [15, 28, 54].

3 Daten und Methodik

Im Folgenden wird die grundlegende Vorgehensweise dargelegt, die in dieser Untersuchung Anwendung findet. Die Abb. 3.1 verweist dabei auf die durchgeführten Schritte und die verwendete Methodik, angelehnt an Ali et al. [51]. Die methodische Vorgehensweise beginnt mit der Datensammlung, gefolgt von der Datenanalyse und -vorverarbeitung, der Datenpartitionierung[1] und der Erstellung von Modellen unter Verwendung verschiedener Algorithmen für maschinelles Lernen und Ensemble-Methoden[2]. Schließlich werden die finalen Modelle anhand verschiedener, modellinterner Parameter optimiert, bewertet und validiert. Dabei ist die Reproduzierbarkeit in Form einer *Pipeline* der Arbeit sichergestellt und in einem GitHub-Repository hinterlegt (siehe Anhang D). Die verwendeten Methoden orientieren sich an der gängigen Praxis, die in der Literatur gut dokumentiert sind und den aktuellen Stand wiedergeben.

3.1 Datenbeschreibung und- aufbereitung

Gemäß der EPB-Richtlinie muss für WG ein EA ausgestellt werden, wenn sie gebaut, verkauft oder vermietet werden. EA können auch im Rahmen einer Green-Deal-Bewertung erstellt werden. Die Gesamteffizienz wird durch das „Energy Efficiency Rating“ (EER) erfasst, das von 1 bis 100 reicht. Abbildung 3.2 zeigt die Aufschlüsselung des EER nach den Energielabeln. Eine Bewertung von 100 steht für höchste Effizienz und damit für niedrigere Energiekosten. EA dienen der Messung der Energieeffizienz von WG und Nicht-Wohngebäude (NWG). Üblicherweise wird der EER-Wert auch in Energieeffizienzklassen von A bis G eingeteilt. Die Berechnung des Energieeffizienz-Verhältnisses basiert auf den Kosten für die Beheizung der Immobilie, die Warmwasserbereitung und die Beleuchtung. Dabei wird die Wohnfläche berücksichtigt und es werden standardisierte Annahmen für die Belegung und das Verhalten der Bewohner zugrunde gelegt, um potenziellen Hauskäufern oder Mietern eine einheitliche Vergleichsmöglichkeit zu bieten. Ein entsprechendes Gutachten ist nach Fertigstellung 10 Jahre lang gültig. Hierbei ist bereits zu erkennen, dass die Verteilung der einzelnen Label unausgewogen ist. Innerhalb des Datensatzes gibt es sehr wenige effiziente (A) sowie auch ineffiziente WG mit dem Label G. Der Großteil der zertifizierten Gebäuden ist mit dem Energielabel C ausgewiesen.

[1] Aufteilung in Trainings- und Testdaten

[2] Verwendung einer endlichen Menge von verschiedenen Lernalgorithmen, um bessere Ergebnisse zu erhalten als mit einem einzelnen Lernalgorithmus

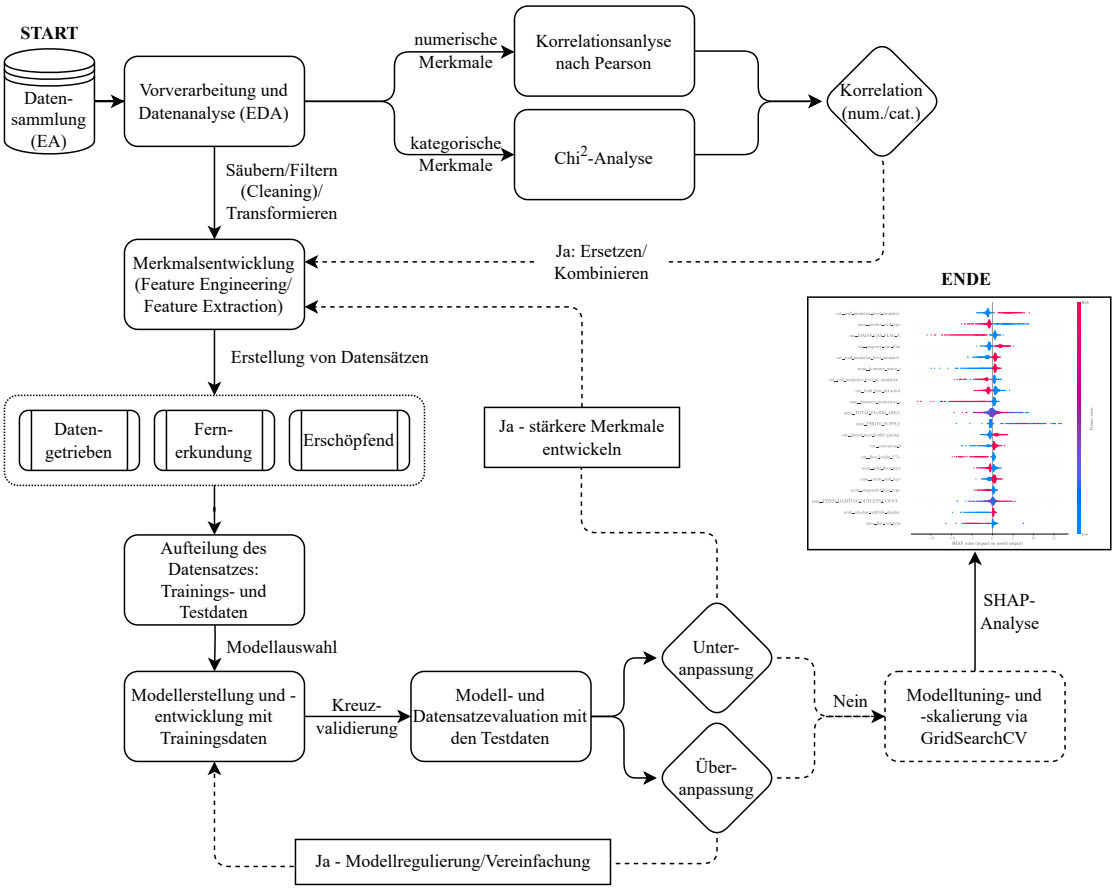

Abbildung 3.1: Ablaufdiagramm der verwendeten Methodik

Die Verteilung der Bewertungen zeigt (siehe Abb. 3.3), dass die meisten WG in den EA-Klassen C und D zu finden sind, die in der Regel als mäßig energieeffizient gelten. Die Aufschlüsselung der Immobilientypen ergibt folgendes Bild. Häuser sind der häufigste Gebäudetyp bei den meisten EA-Einstufungen, insbesondere bei den Einstufungen D und E. Hinsichtlich der Energieeffizienz lässt sich feststellen, dass Wohnungen im Vergleich zu anderen Gebäudetypen tendenziell eine höhere Energieeffizienz aufweisen. Dies kann einerseits dadurch begründet werden, dass Wohnungen in der Regel nach neueren Standards gebaut werden, andererseits dadurch, dass Wohnungen in der Regel neuer sind als andere Gebäudetypen. Bungalows hingegen weisen eine hohe Präsenz in der Bewertungskategorie D auf, jedoch nur eine geringe Präsenz in den höheren und niedrigeren Effizienzklassen. Dies lässt auf eine moderate Energieeffizienz schließen. Bungalows sind insgesamt am wenigsten vertreten, weisen jedoch eine beachtliche Präsenz in den Kategorien C und D auf. Daten für „Park Homes“ sind am wenigsten vertreten, mit einer geringen Ausprägungen in den Bewertungen A und C. Ebenso ist die Anzahl der Immobilien mit den Bewertungen F und G sehr gering, was darauf hindeutet, dass extrem ineffiziente Immobilien ebenfalls selten sind. Die meisten Gebäude liegen nicht an den Extremen der

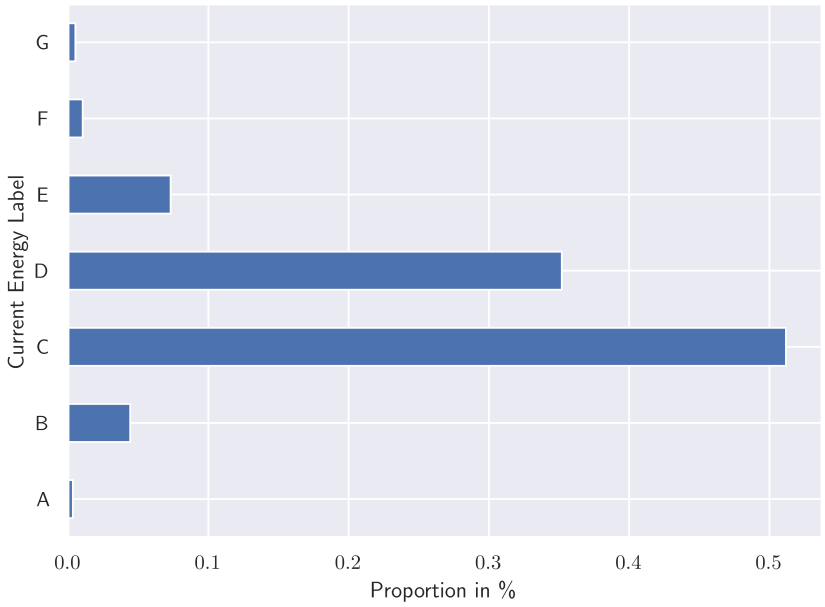

Abbildung 3.2: Rating der Energieeffizienzklasse

Energieeffizienz, was darauf hindeutet, dass es Raum für Verbesserungen gibt, insbesondere bei Häusern, die die Bewertung D dominieren. Die hohe Konzentration von Häusern in den Kategorien C und D lässt darauf schließen, dass eine signifikante Anzahl von Häusern von Energieeffizienzverbesserungen profitieren und in höhere Effizienzkategorien aufsteigen könnten. Energieeffizienzmaßnahmen müssen demnach möglicherweise auf bestimmte Gebäudetypen zugeschnitten werden, da die Verteilung auf die EA-Klassen, je nach Typ variiert.

3.1.1 Datenquellen- und erhebung

Die vorliegende Arbeit verwendet offene bzw. berechnete Daten, die aus EA abgeleitet werden können. Die verwendeten Daten sind für 5.000 WG zufällig ausgewählt worden. Es werden keine direkten Eingabedaten verwendet, die normalerweise für die Berechnung des offiziellen Energielabels erforderlich sind, beispielsweise bei Khayatian et al. [33] Das Modell wurde entwickelt, um das Energielabel bzw. die Energieffizienzklasse für WG zu ermitteln. Dabei sind auch detaillierte Informationen, wie z. B. Heizungssysteme oder Fensterisolierungswerte enthalten, die im Allgemeinen nicht für alle WG verfügbar sind. Die in der Arbeit getesteten Algorithmen sind mit Hilfe von Scikit-Learn (SkL) [55] (Version 1.4.1) als reproduzierbare *Pipeline* implementiert, wobei im Allgemeinen die Programmiersprache Python mit ausgewählten Bibliotheken für die Datenenanalyse verwendet wird. Zur Konsolidierung von Energieeffizienzbewertungen bzw. Energiebedarfsbestimmung von WG, als Ansatz für die Datenerhebung, werden in dieser Arbeit offene Daten der „Abteilung für die Gesamtenergieeffizienz von Gebäuden: England und

Abbildung 3.3: Anzahl der Ausweise anhand des Gebäudetyps

Wales“ verwendet. Es wurden ausschließlich EA für WG verwendet, die seit 2008 gebaut, verkauft oder vermietet wurden. Diese Daten enthalten Informationen über die Energieeffizienzwerte von WG während des Energiebewertungsprozesses. Genauer gesagt werden für diese Arbeit die aktuellen Energieeffizienzbewertungen von A bis G (wobei A sehr effizient und G am wenigsten effizient ist), die zwischen April 2018 und April 2021 eingereicht wurden, verwendet.

Der Tabellenauszug 3.1.1 beschreibt die Zielmerkmale aus der EA-Datenbank, die in dieser Arbeit verwendet werden. Die EA-Daten enthalten insgesamt 92 Merkmale mit umfangreichen Ausprägungen und detaillierten Informationen über Gebäudemerkmale und Energieverbrauch. Eine Beschreibung der einzelnen Merkmale sowie die Verwendung für die Analyse befindet sich im Anhang C. Zusätzlich sind auch potenzielle Abschätzungen sowie Maßnahmen zur Verbesserung des energetischen Zustands enthalten. Die ausführliche Auflistung sowie der verwendeten Merkmale ist dem C zu entnehmen. Für weitere Informationen wird auf folgende Seite verwiesen, wo die EA gelistet sind: EA-Datenbank England. Der Datensatz der EA wurde am 29. Februar 2024 aktualisiert und enthält Ausweise, die bis einschließlich 31. Januar 2024 ausgestellt wurden.

Der Rohdatensatz wurde extrahiert und als Komma-getrennte .csv-Datei heruntergeladen. Zur Erstellung eines adäquaten Modells wurde der vorliegende Rohdatensatz einer Vorverarbeitung unterzogen, welche die Durchführung von Datenimputation und Ausreißererkennung umfasste. Es wurden Techniken zur Skalierung und Kodierung von Merkmalen implementiert, um die Auswahl von Merkmalen oder unabhängigen Variablen zu ermöglichen. Dadurch wird die Vorhersagekraft der ML-Algorithmen durch Hyperpa-

rameteroptimierung erhöht. Der bereinigte und vorverarbeitete Datensatz wird für das Training zufällig in einen Trainings- und einen Testdatensatz im Verhältnis von 70/30% aufgeteilt (vgl. Ausschnitt 3.1).

```
# from sklearn.model_selection import train_test_split
X_train, y_train, X_test, y_test = train_test_split(X, y, random_state=0)
```

Listing 3.1: Partitionierung von Datensätzen: Training- und Testdaten

Mehrere ML-Algorithmen wurden unter Verwendung von scikit-learn in eine laufende Instanz eines *Jupyter Notebook* importiert. Die für die Erstellung der Energieeffizienzausweise erhobenen Daten sind die primäre Datenquelle für diese Analyse. Sekundäre Quellen als Ergänzung zu den bestehende Datensätzen werden beschrieben.

Tabelle 3.1: Zielmerkmale für die Energiebedarfsbestimmung

Merkmal	Beschreibung	Ausprägung
Energielabel	Aktuelle Energiebewertung, umgerechnet in eine lineare Bewertung aus der Energieeffizienz.	A-G
Energieeffizienz	Basierend auf den Energiekosten, d. h. der für Raumheizung, Warmwasserbereitung und Beleuchtung benötigten Energie [kWh/Jahr] multipliziert mit den Brennstoffkosten	1-100

3.1.2 Datenqualität und -bereinigung

EA werden im Allgemeinen durch akkreditierte Gutachter erstellt. Obwohl die Erstellung dieser Gutachten durch strenge Richtlinien und den Einsatz spezifischer Software unterstützt wird, erfordern Situationen, in denen Gutachter nicht auf alle Bereiche des WG Zutritt haben, eine Schätzung bestimmter Werte basierend auf empirischen Werten oder vereinfachten Abschätzungen. Die Genauigkeit der erfassten Daten ist zudem abhängig von der präzisen Übermittlung der Informationen durch die Gutachter selbst. Einige der Vorverarbeitungsmethoden wurden unter Zuhilfenahme des Github-Repositorys[1] durchgeführt. Zur Förderung der Konsistenz zwischen den Bewertungen und zur Steigerung der Datenqualität werden mehrere Datenbereinigungen durchgeführt, darunter:

- Entfernung von Variablen mit einem signifikanten Anteil an fehlenden Daten
- Ausschluss von Immobilien mit ungültigen Postleitzahlen aus der Datenbank
- Standardisierung der Erfassungsmethoden für „fehlende Daten“
- Festlegung eines Mindestwerts = 0 für begrenzte Datenwerte bspw. der Energieverbrauch
- Ausreißerkontrolle

[1] `https://github.com/datasciencecampus/energy-efficiency`

Trotz dieser Maßnahmen zur Qualitätssicherung bestehen weitere Herausforderungen hinsichtlich der Datenqualität, einschließlich:

- Auftreten fehlender Werte (NaN)
- Inkonsistenzen im Vergleich zu anderen Erhebungen bzw. „Samples“
- Mangel an Standardisierung bei den Beschreibungsmerkmalen
- Vorhandensein nicht plausibler Antworten
- Die Zertifikate eine Gültigkeit von 10 Jahren aufweisen und somit veraltet sein können
- Schätzungen von Werten in Fällen, in denen Gutachter nicht in der Lage waren, Teile der Immobilie zu betreten

Durch die Filterung der Merkmale im weiteren Modellierungsprozess, werden diese nicht weiter analysiert oder auf andere potenzielle Fragestellungen hin untersucht. Diese Einschränkungen unterstreichen bereits die Notwendigkeit kontinuierlicher Bemühungen zur Verbesserung der Datenqualität und -verarbeitung im Kontext von Energiezertifizierungsmaßnahmen von Gebäuden.

3.1.3 Korrelationsanalyse und Merkmalsextraktion

Einige der insgesamt 92 Merkmale in den EA-Daten werden aus dem Datensatz entfernt, da sie miteinander stark korrelieren oder eine große Anzahl fehlender Werte aufweisen. Die Abschätzung der Korrelation der numerischen Werte erfolgt dabei über die Korrelationsmethode nach *Pearson*. *Pandas*, eine Python-Bibliothek zur Manipulation von tabellarischen Daten, besitzt eine mitgelieferte Methode zur Überprüfung der Korrelationen innerhalb des *Dataframes*. Dies ist auch im Ausschnitt 3.2 dokumentiert.

```
#House descriptions numerical values
corr_house_desc = epc_clean[['CURRENT_ENERGY_EFFICIENCY',
                             'TOTAL_FLOOR_AREA',
                             'MULTI_GLAZE_PROPORTION',
                             'EXTENSION_COUNT',
                             'NUMBER_HABITABLE_ROOMS',
                             'NUMBER_HEATED_ROOMS',
                             'LOW_ENERGY_LIGHTING',
                             'NUMBER_OPEN_FIREPLACES',
                             'FLOOR_HEIGHT']].corr(method='pearson')
```

Listing 3.2: Korrelationsanalyse numerischer Merkmale

Die Korrelationen zwischen den Merkmalen kann über eine Matrix dargestellt werden (vgl. 3.1.3). Die Beibehaltung korrelierter Variablen im Modell könnte die Schätzung der Beziehung zwischen den Variablen und dem vorhergesagten Zielwert erschweren und

Multikollinearität hervorrufen. Fehlende Werte hingegen können die statistische Aussagekraft des Modells verringern und die daraus gezogenen Schlussfolgerungen verfälschen. Überschreitet die Anzahl der fehlenden Werte nicht die Mindestgrenze von 50%, werden die numerischen Werte innerhalb der *Pipeline* durch den Mittelwert und bei den kategorialen Werten durch den zumeist auftretenden Wert (Modus) innerhalb der Kategorie ersetzt. Das Entfernen von korrelierten Merkmalen reduziert den Datensatz und verkürzt die Laufzeit bei der Modellerstellung. Zur weiteren Verbesserung des Datensatzes wurde die Anzahl der bewohnbaren Räume beibehalten, während die Anzahl der beheizten Räume entfernt wurde. Es wurden auch Merkmale, wie die Aufschlüsselung nach Umwelt und Energieeffizienz (Wände, Böden, Fenster, Dächer) sowie von anderen Merkmalen abgeleitete Merkmale aus dem Roh-Datensatz entfernt. Die einzigartigen Variablen, bei denen es unwahrscheinlich ist, dass diese durch eine andere Quelle (Proxy-Daten[1]) akquiriert werden können, wurden beibehalten.

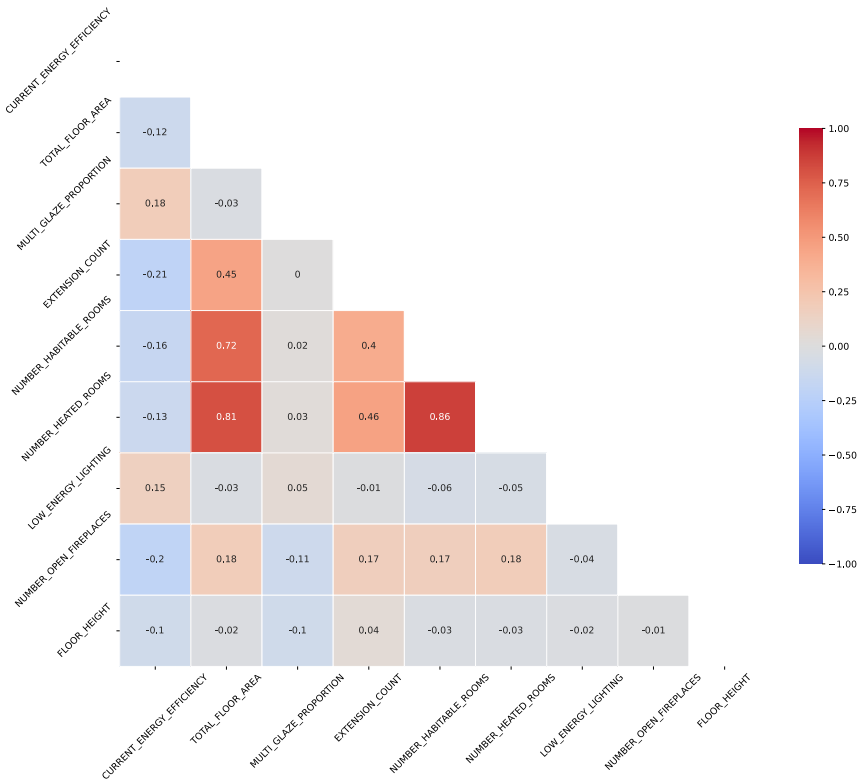

Abbildung 3.4: Korrelationsanalyse der numerischen Gebäudemerkmale

Es lassen sich drei wesentliche Methoden zur Merkmalsauswahl unterscheiden, nämlich Filter, Wrapper und eingebettete Methoden [56, 57]. Die Filter- und Wrapper-Methoden finden bereits in verwandten Arbeiten zur Energievorhersage Anwendung. Kusiak et al.

[1] „Proxy-Daten“ dienen als Ersatz oder Stellvertreter für direkte Daten, die für die Lösung eines bestimmten Problems relevanter oder präziser sind, jedoch schwieriger zu beschaffen, zu messen oder direkt in einem Modell zu verwenden sind.

wenden beispielsweise einen Boosting-Baum an, um die Eingangsmerkmale für die Energielastvorhersage zu bewerten [58]. Die Anwendung der Embedded-Methode ist nachweislich mit einem hohen Rechenaufwand verbunden. Zudem liefert sie mitunter Ergebnisse, die hinter denen der Filter- und Wrapper-Methode zurückbleiben und wurde im Bereich der Energiebedarfsvorhersage nur selten eingesetzt [56]. Aus diesem Grund findet sie in dieser Arbeit keine Berücksichtigung und die Filtermethode des Chi-Quadrat (Chi^2)-Test wird angewendet. Die Filtermethode zielt darauf ab, irrelevante und redundante Variablen, die sich auf Statistiken konzentrieren, zu entfernen. Im Gegensatz dazu zielt die Wrapper-Methode darauf ab, den besten Satz von Variablen für spezifische ML-Algorithmen zu bestimmen [57].

Ein Chi^2-Test auf Unabhängigkeit wurde durchgeführt, um kategoriale Merkmale zu finden, die miteinander und zum Zielwert in Beziehung stehen. Die Merkmale, die als signifikant mit einem p-Wert $< 0{,}05$ zum Zielwert identifiert werden, bleiben erhalten. Sollte zwischen den Merkmalen eine Abhängigkeit festgestellt werden, wird eines der beiden Merkmale auch entfernt. Abhängige und unabhängige Merkmale werden anhand des p-Wertes des Chi^2-Tests unterschieden. Wenn der p-Wert kleiner als das gewählte Signifikanzniveau (in der Regel $Alpha = 0,05$) ist, besteht ein statistisch signifikanter Zusammenhang zwischen den Merkmalen. Das bedeutet, dass die Verteilung des einen Merkmals von dem anderen abhängt. Wenn der p-Wert größer oder gleich *Alpha* ist, sind die Merkmale unabhängig voneinander. Wenn der p-Wert größer oder gleich dem Signifikanzniveau ist, deutet dies darauf hin, dass es keinen statistisch, signifikanten Zusammenhang zwischen den Merkmalen gibt. Das bedeutet, dass die Beweise nicht auf eine Beziehung zwischen den Verteilungen der beiden Merkmale hindeuten und sie somit unabhängig voneinander sind. Dies wird aus zwei Gründen angewendet: **1. Vermeidung von Abhängigkeiten zwischen Merkmalen:** Wenn sich zwei Merkmale als abhängig erweisen, ist es möglicherweise nicht sinnvoll, beide Merkmale in das Modell zu integrieren, da sie redundante Informationen liefern. Dies kann zu Problemen wie Multikollinearität in Regressionsmodellen führen. Stark korrelierte Prädiktoren können das Modell instabil machen und Interpretationen erschweren. In diesem Fall sollte nur das abhängige Merkmal einbezogen werden, das eine stärkere Beziehung zur Zielvariable aufweist oder besser interpretierbar ist. **2. Relevanz für das Zielmerkmal:** Der Chi^2-Test bewertet die Beziehung zwischen zwei kategorialen Merkmalen, sagt jedoch nichts über die Beziehung zwischen einem Merkmal und der Zielvariablen aus, es sei denn, eines der Merkmale im Test ist die Zielvariable. Die Merkmalsauswahl sollte dabei nur diejenigen Merkmale einbeziehen, die eine starke und statistisch signifikante Beziehung zum Zielmerkmal aufweisen. Dabei gibt es aber auch Ausnahmen. Neben der statistischen Signifikanz kann auch die Bereichsrelevanz und die praktische Bedeutung der Merkmale berücksichtigt werden. Ein Merkmal kann eine schwache, statistische Assoziation mit dem Zielmerkmal aufweisen, kann aber dennoch aufgrund von Fachwissen wichtig sein. Weiterhin reagieren verschiedene Modelle unterschiedlich auf Merkmalsabhängigkeiten. Baumbasierte Mo-

delle wie RF und GB können effizienter mit korrelierten Merkmalen umgehen als lineare Modelle [31]. Scikit-learn besitzt geeignete, statistische Methoden, um diese Beziehungen darzustellen (vgl. Liste 3.3). Hierfür werden entsprechende Funktionen geschrieben und auf den vorhandenen Datensatz angewendet. Die Auswertung und Selektion führt zu einer tabellarischen Aufstellung, welche die Ergebnisse der Filterung der Merkmale präsentiert. Eine weitere Möglichkeit wäre, das Modell selbst entscheiden zu lassen, welche Merkmale relevant sind. Dazu könnte die *SelectFromModel*-Klasse verwendet werden, um diese Merkmale im Nachhinein zu filtern und zu selektieren [55]. Diese Vorgehensweise findet an der Stelle keine Verwendung, da dies eher dem datengetriebenen Ansatz entspricht.

```
from sklearn.feature_selection import chi2
from sklearn.preprocessing import LabelEncoder

def encode_features(epc_clean, categorical_features):
  """
  Encode categorical features using LabelEncoder.
  Returns a DataFrame with encoded features.
  """
  le = LabelEncoder()
  encoded_df = pd.DataFrame()
  for feature in categorical_features:
      encoded_df[feature] = le.fit_transform(epc_clean[feature].astype(
    str))
  return encoded_df

def calculate_chi_square(encoded_df, features, target_feature):
  """
  Calculate Chi-square stats between each feature and the target feature.
  Returns a DataFrame with Chi-square scores and p-values.
  """
  chi_scores = []
  for feature in features:
      if feature != target_feature:
          chi_score, p_value = chi2(encoded_df[[feature]], encoded_df[
    target_feature])
          chi_scores.append((feature, chi_score[0], p_value[0]))
  chi_scores_df = pd.DataFrame(chi_scores, columns=['Feature', 'Chi2
    Score', 'p-value'])
  return chi_scores_df.sort_values(by='Chi2 Score', ascending=False)

def analyze_feature_relationships(encoded_df, features):
  """
  Analyze relationships between all pairs of features using Chi-square
    test.
  Identifies potentially redundant features.
  """
```

```
33  chi_scores_pairs = []
34  for feature1, feature2 in combinations(features, 2):
35      chi_score, p_value = chi2(encoded_df[[feature1]], encoded_df[
      feature2])
36      chi_scores_pairs.append((feature1, feature2, chi_score[0], p_value
      [0]))
37  chi_scores_pairs_df = pd.DataFrame(chi_scores_pairs, columns=['Feature
      1', 'Feature 2', 'Chi2 Score', 'p-value'])
38  return chi_scores_pairs_df.sort_values(by='Chi2 Score', ascending=False
      )
```

Listing 3.3: Chi-Quadrat-Unabhängigkeitstest der kategorialen Merkmale

Die Auswertungen und weitere Details zu den Korrelationen der kategorischen Merkmale befinden sich im Anhang A. Weiterhin sind auch die Gründe für den Ausschluss der einzelnen Merkmale in Anhang C beschrieben. Weitere Informationen finden sich auch im GitHub-Repository unter: EPC-Modelling

Erstellung der relevanten Datensätze

Die EA-Daten enthalten mehrere kategoriale Variablen mit einer Vielzahl von Werten (siehe Abschnitt 3.1.3). Um geeignete Merkmale zu finden, die die meisten Informationen enthalten, werden drei Datensätze erstellt und untersucht. Der Hauptfokus liegt hier auf den FED-Datensatz, der die Basis für die Identifizierung geeigneter Daten zur Übertragung darstellt.

- Data-driven - Datengetriebener Ansatz
- Remotesensing - Fernerkundung oder domänengetriebener Ansatz
- Exhaustive - Erschöpfender/Bereinigter Datensatz

Der erste Ansatz, der als *datengetrieben* bezeichnet wird, verwendet statistische Methoden, um die Anzahl der Variablen zu reduzieren und in Gruppen aufzuteilen oder durch andere Merkmal zu ersetzen. Da einige Merkmale, die textliche Beschreibungen der Eigenschaft enthalten, freihändig erstellt wurden, enthalten diese zum Teil eine große Anzahl an einmaligen Ausprägungen. Obwohl die erzeugten Gruppen aus dem daten-getriebenen Ansatz aus mathematischer Sicht korrekt erscheinen, sind diese schlecht zu interpretieren. Um dies zu beheben, gruppiert der Fernerkundungsdatensatz neue Kategorien unter Verwendung von baulichen Eigenschaften und Beschreibungen auf Basis der bestehenden Informationen. Beispiele hierfür sind die Extrahierung von Schlüsselmerkmalen wie „geneigtes Dach“ (Dachtyp), „isolierter Boden“ oder die „Gebäudealtersklasse“ aus den Beschreibungen der einzelnen Instanzen. Dadurch wird die Dimensionalität des Datensatzes erhöht, da weitere Merkmale hinzugefügt werden und die Leistungsfähigkeit verringert sich. Jedoch ist dieser Ansatz leichter zu interpretieren, als der datengesteuerte Ansatz.

Weiterhin kann eine Aussage über die Relevanz und den Beitrag der einzelnen Merkmale entschieden werden. Der dritte Ansatz liefert dem Modell die größtmögliche Information, ohne die Daten weiter zu manipulieren. Dadurch können fortgeschrittene, maschinelle Lernverfahren selbst entscheiden, wie die Daten während des Trainings aufgeteilt werden sollen, da die kategorischen Werte nicht in weitere Gruppen unterteilt werden. Dieser Ansatz besitzt grundsätzlich den größten Informationsgehalt, jedoch ist die Rechenzeit für das Training der Modelle länger, da alle Einträge kodiert werden müssen. Das Ergebnis dieser Herangehensweise sind vier neue Sub-Datensätze als .csv-Datei, die auf dem ursprünglichem Datensatz aufbauen. Bei dem dritten Ansatz wurde das Merkmal *Energieverbrauch* ergänzt, um die Auswirkungen der Korrelation auf die Leistung des Modells zu untersuchen.

1. Datengetrieben → *epc_ dd.csv*
2. Fernerkundung → *epc_ rs.csv*
3. Fernerkundung+ → *epc_ rs_ plus.csv*
4. Erschöpfend → *epc_ ex.csv*

3.1.4 Merkmalsentwicklung und Dimensionalitätsreduktion

Die Entwicklung von Merkmalen (engl. Feature Engineering (FE)) und die Reduzierung der Dimensionalität von Daten sind wichtige Faktoren zur Verbesserung der Leistung von maschinellen Lernmodellen, insbesondere bei der Verarbeitung von tabellarischen Daten [59]. Beim Feature Engineering werden neue Merkmale aus vorhandenen erstellt, um die Genauigkeit des Modells zu verbessern. Dieser Prozess kann verborgene Muster in den Daten aufdecken, die nicht sofort ersichtlich sind, und dem Modell aufschlussreichere Eingangsmerkmale für die Vorhersagen zu liefern [60]. FE kann die Lernfähigkeit eines Modells erheblich verbessern, was zu genaueren und robusteren Vorhersagen führt durch Vermeidung einer Unteranpassung des Modells. Ein Beispiel ist die Baualtersklasse innerhalb des Datensatzes. Teilweise sind dabei Präfixe wie „England and Wales: “ mit der verbundenen Baultersklasse und fest definierte Baujahre angegeben. Um ein einheitliches Schema sowie ein verwertbares Merkmal zu generieren, werden die Präfixe entfernt und die Baujahre den entsprechenden Baualtsklassen gemäß Mikrozensus zugeordnet. Hierfür wird eine Funktion geschrieben, die auf den vorhanden Datensatz angewendet werden kann (vgl. 3.4).

```
1  def map_age_band(age_band):
2      """
3      Cleaning and mapping the 'CONSTRUCTION_AGE_BAND' feature to a new
     schema.
4      """
5      if not isinstance(age_band, str):
```

```
        return None

    # Removing the "England and Wales:" prefix and strip any leading/trailing spaces
    age_band_clean = age_band.replace("England and Wales:", "").strip()

    # Special case for 'before 1900'
    if 'before' in age_band_clean:
        return "before 1919"

    # Extracting the year(s) from the string
    years = [int(s) for s in age_band_clean.split('-') if s.strip().isdigit()]

    # If no years are found, return None
    if not years:
        return np.nan

    # Using the earliest year in the range for mapping
    year = years[0]

    # Mapping to the new schema (Microcensus-Building Classes)
    if year < 1919:
        return "before 1919"
    elif year <= 1948:
        return "1919 - 1948"
    elif year <= 1978:
        return "1949 - 1978"
    elif year <= 1986:
        return "1979 - 1986"
    elif year <= 1990:
        return "1987 - 1990"
    elif year <= 1995:
        return "1991 - 1995"
    elif year <= 2000:
        return "1996 - 2000"
    elif year <= 2004:
        return "2001 - 2004"
    elif year <= 2008:
        return "2005 - 2008"
    else:
        return "2009 and later"

# Apply this function to the 'CONSTRUCTION_AGE_BAND' column
epc_rs['construction_age_band'] = epc_rs['CONSTRUCTION_AGE_BAND'].apply(map_age_band)
```

Listing 3.4: Merkmalsentwicklung für die Baualtersklasse

Ein weiteres Beispiel ist die Merkmalsentwicklung für den „Dachtyp“. Der Dachtyp der

WG ist im Merkmal der Dachbeschreibung („Roof Description“) in Kombination mit dem Isolierungsgrad enthalten. Daraus können zum einen das Merkmal *Dachtyp* sowie auch *Isolierungsgrad* entwickelt werden. Die exemplarische Implementierung kann dem Ausschnitt 3.5 entnommen werden.

```
1 def roof_types(df):
2
3   df['pitched_roof_type'] = df.apply(lambda row: 1 if 'pitched' in str(
      row['ROOF_DESCRIPTION']).lower() else 0, axis = 1)
4   df['flat_roof_type'] = df.apply(lambda row: 1 if 'flat' in str(row['
      ROOF_DESCRIPTION']).lower() else 0, axis = 1)
5   df['thatched_roof_type'] = df.apply(lambda row: 1 if 'thatched' in str(
      row['ROOF_DESCRIPTION']).lower() else 0, axis = 1)
6
7   return df
```

Listing 3.5: Merkmalsentwicklung für den Dachtyp

Die Dimensionalitätsreduzierung von Daten hingegen zielt darauf ab, den Datensatz durch Verringerung der Anzahl der Eingabevariablen zu vereinfachen. Techniken wie die Hauptkomponentenanalyse (PCA) und Methoden zur Merkmalsauswahl helfen dabei, die informativsten Merkmale zu ermitteln und zu behalten, während überflüssige oder irrelevante Merkmale eliminiert werden. Die Merkmalsauswahl resultiert aus einem vorangestellten Unabhängigkeitstest (siehe 3.1.3), jedoch gibt es auch andere Methoden für die Merkmalsauswahl bspw. die rekursive Auswahl von Merkmalen auf Basis des erstellten Modells (siehe Dokumentation [55] und [50]). Diese Methoden werden nicht im Detail beschrieben. Durch die Reduzierung der Dimensionalität verbessert sich nicht nur die Leistung des Modells, indem es sich auf die relevantesten Daten konzentriert, sondern es verringert auch die Komplexität der Berechnungen und das Risiko einer Überanpassung des Modells. Die Modelle, die auf Datensätzen mit reduzierter Dimensionalität trainiert werden, können besser auf neue, ungesehene Daten angewendet werden. Dadurch sind sie in praktischen Anwendungen effektiver. *Feature-Engineering* und Dimensionalitätsreduktion haben in der Vorverarbeitungsphase eine zentrale Bedeutung. Nicht nur die Modellgenauigkeit und -effizienz werden verbessert, sondern auch Probleme wie *Overfitting*[1] und Rechenkomplexität können bewältigt werden. Durch sorgfältige Anwendung dieser Techniken können Datenwissenschaftler leistungsfähigere und effizientere Modelle für maschinelles Lernen erstellen, die aussagekräftige Erkenntnisse aus tabellarischen Daten gewinnen können. Für den datengetriebenen Ansatz wird PCA verwendet, um die Information aus den beschreibenden Merkmalen zu extrahieren und in neue Merkmale umzuwandeln. Eine beispielhafte Anwendung für den datengetriebenen Ansatz ist im Folgenden dokumentiert (siehe Ausschnitt 3.6)

[1] „Überanpassung“: Das Modell hat eine gute Leistung auf den Trainingsdaten, aber nicht auf den Testdaten.

```python
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.decomposition import PCA

# Step 1: Preprocessing (this example uses TF-IDF)
epc_dd['combined_description'] = epc_dd[descriptive_cols].apply(lambda x:
    ' '.join(x.dropna()), axis=1)

# Step 2: Vectorization
vectorizer = TfidfVectorizer(stop_words='english')
X_tfidf = vectorizer.fit_transform(epc_dd['combined_description']).
    toarray()

# Step 3: Scaling before PCA
scaler = StandardScaler(with_mean=False)
X_scaled = scaler.fit_transform(X_tfidf)

# Apply PCA with all components to determine the variance explained
pca_full = PCA()
pca_full.fit(X_scaled)
explained_variance_ratio = pca_full.explained_variance_ratio_

# Determine n_components: choosing a value that retains 90% or  of
    variance
cumulative_variance = explained_variance_ratio.cumsum()
n_components = (cumulative_variance < 0.9).sum() + 1

# Step 3: Apply PCA with chosen n_components
pca = PCA(n_components=n_components)
X_pca = pca.fit_transform(X_scaled)

# Step 4: Integrate PCA Components Back to Original Dataset
for i in range(n_components):
    epc_dd[f'pca_comp_desc{i+1}'] = X_pca[:, i]
```

Listing 3.6: Dimensionalitätsreduktion der beschreibenden Merkmale

3.2 Modellentwicklung und -optimierung

Wie bereits in Kapitel 2 erwähnt, werden für den hier dargestellten Anwendungsfall überwachte, maschinelle Lernmethoden verwendet, da der vorhandene Datensatz bereits gelabelt bzw. das Ziel-Merkmal (Energieeffizienz/Rating) bereits vorhanden ist.

3.2.1 Erstellung eines Basismodells

Ein Basismodell ist ein erster Modellierungsversuch, der eine Basismessung liefert und während der gesamten Entwicklung als Referenzpunkt dient. Es handelt sich hierbei oft um ein heuristisches Modell oder ein einfaches maschinelles Lernmodell [61]. Zu Beginn

eines Projekts ist es vorteilhaft, ein Basismodell zu erstellen, um das Verhältnis von Nutzen und Kosten zu verstehen. Die Entwicklung von Modellen für maschinelles Lernen ist teuer, sowohl in Bezug auf die Zeit, als auch auf die benötigten Ressourcen. Daher ist es wichtig zu wissen, ob das Basismodell, dass bspw. nur 5% weniger genau ist als das finale Modell die Kosten wert ist. Das Basismodell dient als Referenz und liefert wertvollen Kontext für die Bewertung des vollständigen Modells [61, 62]. Ein weiterer Vorteil der Erstellung eines Basismodells ist die Möglichkeit, Leistungsverbesserungen zuzuordnen. Es dient damit als Ausgangspunkt für die weitere Erstellung und Optimierung von leistungsfähigeren Modellen. Durch das Wissen darüber, welche technischen oder parameterbedingten Änderungen zu einer Leistungsverbesserung geführt haben, erhält man ein besseres Verständnis und kann seine Bemühungen gezielter darauf konzentrieren. Das Basismodell ist somit ein wichtiger Bestandteil des iterativen Modellierungsvorgangs. Da ein regelbasiertes Modell aufgrund der Vielzahl an vorhandenen Merkmalen nicht möglich ist, sind für den hier dargestellten Anwendungsfall zwei maschinelle Lernmethoden als Basismodell bereitgestellt:

- Regression: Lineares Regressionsmodell (LR)
- Klassifikation: Entscheidungsbaum (DT)
- Klassifikation: Random Forest (RF)

Die Basismodelle werden mit dem bereinigten und unmanipulierten Datensatz trainiert und über *Kreuzvalidierung* getestet, um festzustellen, ob eine Über- oder Unteranpassung vorliegt. Im Ausschnitt 3.7 ist der prinzipielle Ablauf für das Basis-Regressionsmodell zur Vorhersage der Energieeffizienz dargestellt. Das gleiche Verfahren kann für das Entscheidungsbaummodell durch den Austausch des Zielmerkmals, den Import der benötigten Bibliotheken adaptiert und angewendet werden. Die *Pipeline* umfasst dabei verschiedene Schritte zur Datenimputation, Skalierung, Kodierung sowie zur Modellanwendung. Die Pipeline für numerische Daten besteht dabei aus zwei Schritten. 1. Imputation: Fehlende Werte werden durch den Median ersetzt, was eine robuste Wahl ist, da der Median weniger anfällig für Ausreißer ist als der Mittelwert. 2. Skalierung: Die *MinMaxScaler*-Transformation skaliert die Daten auf ein Intervall zwischen 0 und 1, um die Leistung zu verbessern und numerische Merkmale mit einem hohen Skalenniveau nicht überwiegen. Die Pipeline für kategoriale Daten beinhaltet ebenfalls zwei Komponenten. 1. Imputation: Hierbei werden fehlende Werte durch den häufigsten Wert (Modus) in jeder Spalte ersetzt. 2. Kodierung: Der *OneHotEncoder* wandelt die kategorialen Merkmale in eine numerische Form um, die für maschinelle Lernmodelle geeignet ist, wobei unbekannte Kategorien während der Transformation ignoriert werden. Der *ColumnTransformer* integriert die zuvor definierten numerischen und kategorischen Pipelines. Dies ermöglicht die simultane Anwendung der jeweiligen Transformationen auf die entsprechenden Spaltengruppen. Schließlich wird eine Gesamtpipeline erstellt, die den *preprocessor* und eine

Modellstruktur enthält. Daraufhin wird die Pipeline auf die Trainingsdaten angewendet, um das Modell zu trainieren.

```
# Importing the necessary libraries
from sklearn.linear_model import LinearRegression
from sklearn.impute import SimpleImputer
from sklearn.preprocessing import OneHotEncoder, MinMaxScaler
from sklearn.model_selection import train_test_split
from sklearn.pipeline import Pipeline
from sklearn.compose import ColumnTransformer

# Replace 'target_column' with actual target column name
target = "CURRENT_ENERGY_EFFICIENCY"

# Separate features and target and drop leading features
X_lr = epc_regress.drop(columns=[target, "CURRENT_ENERGY_RATING"])
y_lr = epc_regress[target]

# Preprocessing pipeline
X_train_lr, X_test_lr, y_train_lr, y_test_lr = train_test_split(X_lr,
    y_lr, test_size=0.2, random_state=42)

numeric_features = X_lr.select_dtypes(include=['int64', 'float64']).
    columns
numeric_transformer = Pipeline(
    steps=[
        ("imputer", SimpleImputer(strategy="median")),
        ("scaler", MinMaxScaler())
        ]
)

categorical_features = X_lr.select_dtypes(include=["object"]).columns
categorical_transformer = Pipeline(
    steps=[
        ("imputer", SimpleImputer(strategy="most_frequent")),
        ("encoder", OneHotEncoder(handle_unknown="ignore"))
        ]
    )

# Resulting preprocessor which combines numerical and categorical
    transformations
preprocessor = ColumnTransformer(
    transformers=[
        ("num", numeric_transformer, numeric_features),
        ("cat", categorical_transformer, categorical_features)
        ]
)

lr = Pipeline(
```

```
    steps=[("preprocessor", preprocessor),
           ("lr", LinearRegression())
           ]
       )

lr.fit(X_train_lr, y_train_lr)

# For R-squared scores
scores = cross_val_score(lr, X_train_lr, y_train_lr, cv=10, scoring="r2")
```

Listing 3.7: Erstellung des Basismodells (Lineare Regression)

3.2.2 Auswahl von Lernverfahren und Modellarchitekturen

Wie bereits beschrieben hängt die Auswahl des Modells von der Struktur der Eingangsdaten sowie von der Aufgabe (Regression/Klassifikation) ab. Eine gängige Praxis bei der Konzeptionierung von Modellen ist der parallele Vergleich mehrerer Modelle mit einem Basisdatensatz. Nach Aufbereitung und Bereinigung des Datensatzes wird mit Hilfe von *PyCaret* [63] das beste Modell identifiziert. *PyCaret* ist eine Open-Source, *Low-Code*-ML-Bibliothek, die den Prozess der Durchführung von End-to-End-ML-Experimenten vereinfachen kann. Weiterhin beinhaltet PyCaret weitere Modelle, die nicht standardgemäß in SkLearn enthalten sind, u.a. XGB, ET und CB. Es automatisiert viele Aspekte des maschinellen Lernprozesses, wie Datenvorbereitung, FE, Modellauswahl, Hyperparameter-Tuning und Modellbewertung. Folgende fünf Modelle können über eine 10-fache Kreuzvalidierung identifiziert werden:

1. CatBoost - Categorical Boosting (CB)
2. Extra Trees - Extremely Randomized Trees (ET)
3. Extreme Gradient Boosting (XGB)
4. Gradient Boosting (GB)
5. Random Forest (RF)

Die identifizierten Modelle basieren auf dem Prinzip von Entscheidungsbäumen. Der Entscheidungsbaum-Algorithmus generiert ein logisches Modell, welches sich ausgehend von einem Wurzelknoten über mehrere Verzweigungen bis zu einem finalen Klassifizierungsblattknoten verzweigt (vgl. Abb. 3.5). Die Anzahl der Verzweigungen bzw. die Tiefe (engl. Max Depth) eines Entscheidungsbaums wird im Algorithmus festgelegt, wobei sich dies sowohl auf die Berechnungszeit, als auch auf die Genauigkeit der Ergebnisse auswirkt. Um ein Optimum zu finden, müssen verschiedene Tiefen des Baums erkundet werden. Es gibt verschiedene Strategien, um zu bestimmen, welche Entscheidungen im Baum ausgewählt werden sollen. Die Gängigste ist, den Algorithmus so einzustellen, dass er eine Kostenfunktion verwendet, um zu analysieren, welche der Aufteilungen den geringsten Genauigkeitsverlust darstellt. Dies geschieht über Validierungskurven (engl. Validation Curves) anhand verschiedener modellinterner Parameterbereiche. Wenn ein Gebäude

dem Baum bis zur letzten Ebene folgt, wird ihm schließlich ein bestimmtes Energielabel/Energieeffizienzbewertung zugewiesen. So gelangt es zur Teilmenge 1.2, wandert ein paar Ebenen weiter nach unten und erhält schließlich das Energielabel D bei Klassifikation oder den numerischen Wert bei dem hier gezeigten Anwendungsfall. [31] Dieser Prozess kann mit einer nahezu unendlichen Anzahl von „Zweigen“ oder Entscheidungen fortgesetzt werden.

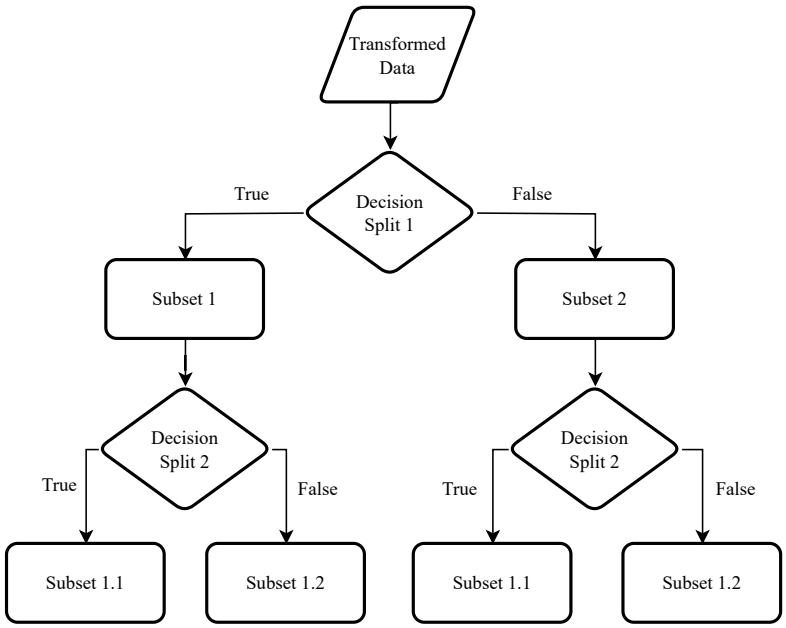

Abbildung 3.5: Ein Datenpunkt beginnt an der Spitze des Baums und wird nach Entscheidung 1 in zwei Teilmengen aufgeteilt (eigene Darstellung in Anlehnung an Hettinga et al. [31]).

Der RF-Algorithmus stellt eine weiterentwickelte Version des Entscheidungsbaum dar. Es handelt sich um eine Gruppe von Bäumen (vgl. Abb. 3.5), die jeweils auf einer Teilmenge der Trainingsdaten trainiert werden und jeweils ihre eigenen Vorhersagen erstellen. Die von einem Baum getroffenen Entscheidungen (unter Verwendung der Kostenfunktionen) unterscheiden sich häufig von den Entscheidungen eines anderen Baumes. Dies kann letztlich zu einer identischen Entscheidung oder zu einem divergierenden Ergebnis führen [31]. Das Resultat wird gemäß dem Mehrheitsprinzip[2] selektiert, welches das Ergebnis priorisiert, was von den meisten Bäumen generiert wird. Die Modifikation des RF-Algorithmus, ET, zeichnet sich durch eine zufällige Wahl der Aufteilungen der Entscheidungsbäume aus. Dies resultiert in einer verkürzten Trainingszeit sowie in einigen Fällen in einer verbesserten Modellgenauigkeit. Zudem ist ET gegenüber Überanpassungen resistent und eignet sich für Klassifikations- und Regressionsproblemen [54]. Das GB basiert zwar ebenfalls auf der Erstellung eines Entscheidungsbaums, jedoch wird hier-

[2]engl. Majority Vote: Untergliederung in „Soft“-Voting: Berücksichtigung der Wahrscheinlichkeiten bei dem Mehrheitsentscheid der Klasse und „Hard“-Voting: Mehrheitsentscheid ohne Berücksichtigung der Wahrscheinlichkeit der vorhergesagten Klasse

bei davon ausgegangen, dass nicht alle Entscheidungen, die im Baum getroffen werden, eine ähnliche Genauigkeit aufweisen. Aus diesem Grund beginnt der Algorithmus beim GB mit der Erstellung des ersten Baums. Für jede Vorhersage wird die Fehlerspanne betrachtet, die durch diese Vorhersage entsteht. Bei Vorhersagen mit geringem Fehler lassen sich die Daten in der Teilmenge leicht anpassen bzw. klassifizieren und müssen nicht weiter berücksichtigt werden. Teilmengen mit höherer Fehlerspanne sind hingegen schwieriger anzupassen bzw. zu klassifizieren und müssen weiter verfeinert werden, um eine bessere Anpassung bzw. Klassifizierung zu finden. Die Unterscheidung erfolgt durch den Algorithmus, indem er den Vorhersagen mit geringem Fehler ein niedriges Gewicht und denjenigen mit hohem Fehler ein höheres Gewicht zuweist. Diejenigen mit der höchsten Gewichtung werden dann in einem nachfolgenden Baum neu bewertet, um eine bessere Übereinstimmung zu erzielen. Der zweite Baum wird zum ersten Baum hinzugefügt, wodurch der erste Baum verbessert wird. Dies geschieht in so vielen Iterationen, wie im Algorithmus festgelegt ist. Die Hyperparameter dieses Algorithmus sind identisch mit denen der RF-Algorithmen [31]. CB verwendet ebenfalls das Gradienten-Boosting auf Entscheidungsbäumen und ist speziell für den Umgang mit kategorischen Daten optimiert, die auch im Datensatz stark vertreten sind. CatBoost bietet eine hohe Vorhersagegenauigkeit und ist robust gegenüber typischen Problemen, wie der Überanpassung und dem Umgang mit fehlenden Daten [64].

Die identifizierten Modelle besitzen noch die Standardwerte für die Hyperparameter und sind somit nicht reguliert bzw. neigen zur Überanpassung. In Tabelle 3.2.2 sind die Leistungsmetriken ersichtlich, die absteigend nach dem Bestimmtheitsmaß (R^2) aufgelistet sind. Für die weitere Analyse wird das XGB-Regressor verwendet, da dieses Modell die kürzeste Trainigszeit aufweist und eine Vielzahl von Hyperparametern besitzt, die für das Tuning sowie die weitere Optimierung verwendet werden können. Dieses Modell ist neben den Modellen CB, ET und RF das vielversprechenste und auch laut der Literatur die meist verwendete Modellarchitektur. Eine Erklärung zu den Evaluationsmetriken findet sich in Tablle 3.2.2.

Tabelle 3.2: Identifizierung geeigneter Modelle anhand des Basisdatensatzes mit MAE - Mean Absoulte Error, MSE - Mean Squared Error, RMSE - Root Mean Squared Error und TT - Training Time in Seconds

Modell	R^2	MAE	MSE	RMSE	TT (Sec)
CatBoost Regressor	0,747	3,279	26,662	5,138	2,330
Extra Trees Regressor	0,727	3,364	28,707	5,329	1,095
Gradient Boosting Regressor	0,707	3,671	31,010	5,533	0,370
Extreme Gradient Boosting	0,696	3,463	32,198	5,636	0,246
Random Forest Regressor	0,659	3,655	36,281	5,976	0,948

Leistungsindikatoren (Metriken)

Ein wesentlicher Bestandteil bei der Erstellung von maschinellen Lernmodellen ist die Bewertung und Generalisierbarkeit des Modells auf neue Daten. Die Bewertungskriterien unterscheiden sich auch hier hinsichtlich des Zielmerkmals bzw. der abhängigen Variable (numerisch/kategorisch). Da die Zielvariable eine numerische Ausprägung beinhaltet, kann ein Regressionsmodell verwendet werden. F_i und A_i sind die vorhergesagten und wahren Werte für eine Instanz i, N ist die Größe der Stichprobe (5.000) und $\bar{A}$ der Mittelwert der wahren Werten A_i. In Tabelle 3.2.2 sind die meist verwendeten Evaluationsmetriken für Regressionsmodelle aufgelistet. Da der Datensatz auch das übersetzte Energielabel A bis G beinhaltet, sind wegen der Vollständigkeit die Evaluationsmetriken von Klassifikationsmodellen in Tabelle 3.2.2 dargestellt.

Tabelle 3.3: Übersicht der meist verwendeten Evaluationsmetriken in Anlehnung an Amasyali et al. (2018) [28] und Wenninger et al. [15] für Regressionsmodelle

Bewertungsmaß	**Gleichung**	**Einheit, Wertebereich**	**Bester Wert**
Bestimmtheitsmaß (R^2)	$R^2 = \frac{\sum(F_i-\bar{A})^2}{\sum(A_i-\bar{A})^2}$	$1, [0, 1]$	1
Variationskoeffizient (CV)	$CV = \frac{1}{\bar{A}}\sqrt{\frac{1}{N}\sum_{i=1}^{N}(F_i - A_i)^2}$	$-1, [0, \infty)$	0
Mittlerer absoluter prozentualer Fehler (MAPE)	$MAPE = \frac{1}{N}\sum_{i=1}^{N}\left\lvert\frac{F_i-A_i}{A_i}\right\rvert \cdot 100$	%, $[0, \infty)$	0
Wurzel des mittleren quadratischen Fehlers (RMSE)	$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(F_i - A_i)^2}$	$\frac{kWh}{m^2a}$, $[0, \infty)$	0
Mittlerer absoluter Fehler (MAE)	$MAE = \frac{1}{N}\sum_{i=1}^{N}\lvert F_i - A_i\rvert$	$\frac{kWh}{m^2a}$, $[0, \infty)$	0
Mittlerer quadratischer Fehler (MSE)	$MSE = \frac{1}{N}\sum_{i=1}^{N}(F_i - A_i)^2$	$\left(\frac{kWh}{m^2a}\right)^2$, $[0, \infty)$	0

Bei der Verwendung von Klassifikationsmodellen, empfiehlt es sich, den *F-Score* als Evaluation zu verwenden, da diese Metrik *Precision* und *Recall* kombiniert und eine Übersicht über die Gesamtleistung bietet. Im weiteren Verlauf befindet sich das XGB-Regressionsmodell im Fokus.

3.2.3 Kreuzvalidierung und Hyperparameteroptimierung

Eine *Kreuzvalidierung* oder Kreuzüberprüfung (engl. cross-validation) gibt Aufschluss darüber, wie gut das Modell generalisiert und ob es auch eine gute Leistung bei neuen Daten besitzt. Dabei wird der vorhandene Datensatz in k gleich große Teilmengen, sog. *Folds* aufgeteilt. Die Aufteilung hängt dabei von der Variante der Kreuzvalidierung ab. Hier wird die K-fache Kreuzvalidierung verwendet. Dabei ist k eine festgelegte Zahl, die angibt, in wie viele Folds der Datensatz aufgeteilt wird. Ein häufig gewählter Wert

Tabelle 3.4: Übersicht zu den meist verwendeten Evaluationsmetriken bei mehrklassigen Klassifikationsproblemen in Anlehnung an Sokolova [65].

Maßnahme	**Formel**	**Bewertungsfokus**
Average Accuracy	$\frac{1}{l}\sum_{i=1}^{l}\frac{tp_i+tn_i}{tp_i+fn_i+fp_i+tn_i}$	Die durchschnittliche Klasseneffektivität eines Klassifizierers
Error Rate	$\frac{1}{l}\sum_{i=1}^{l}\frac{fp_i+fn_i}{tp_i+fn_i+fp_i+tn_i}$	Der durchschnittliche Klassenfehler bei der Klassifizierung
Precision_{μ}	$\frac{1}{l}\sum_{i=1}^{l}\frac{tp_i}{tp_i+fp_i}$	Übereinstimmung der Klassenbezeichnungen der Daten mit denen eines Klassifizierers bei Berechnung aus der Summe von Entscheidungen pro Instanz
Recall_{μ}	$\frac{1}{l}\sum_{i=1}^{l}\frac{tp_i}{tp_i+fn_i}$	Effektivität eines Klassifizierers zur Identifizierung von Klassenbezeichnungen, wenn aus der Summe von Entscheidungen pro Instanz berechnet
Fscore_{μ}	$\left(\beta^2+1\right)\frac{Precision_{\mu}\cdot Recall_{\mu}}{\beta^2\cdot Precision_{\mu}+Recall_{\mu}}$	Beziehungen zwischen positiven Datenbezeichnungen und solchen, die von einem Klassifizierer basierend auf der Summe von Entscheidungen pro Instanz gegeben werden
Precision_{M}	$\frac{1}{l}\sum_{i=1}^{l}\frac{tp_i}{tp_i+fp_i}$	Eine durchschnittliche Klassenübereinstimmung der Datenklassenbezeichnungen mit denen eines Klassifizierers
Recall_{M}	$\frac{1}{l}\sum_{i=1}^{l}\frac{tp_i}{tp_i+fn_i}$	Eine durchschnittliche Klasseneffektivität eines Klassifizierers zur Identifizierung von Klassenbezeichnungen
Fscore_{M}	$\left(\beta^2+1\right)\frac{Precision_{M}\cdot Recall_{M}}{\beta^2\cdot Precision_{M}+Recall_{M}}$	Beziehungen zwischen positiven Datenbezeichnungen und solchen, die von einem Klassifizierer basierend auf einem Klassendurchschnitt gegeben werden

für k ist 10 wenn der Datensatz ausgewogen und groß genug ist, um Teilmengen zu bilden [50, 66]. Das Modell wird daraufhin k-Mal über die Teilmengen trainiert und validiert. Bei jedem Durchlauf wird ein unterschiedlicher Fold als Validierungsdatensatz verwendet, während die restlichen $k-1$ Folds zum Trainieren des Modells verwendet. Dies garantiert, dass jeder Fold genau einmal als Validierungsdatensatz verwendet wird. Aus den k-Durchläufen werden die Evaluationsmetriken aggregiert, oft durch Bildung des Mittelwerts, um eine Gesamteinschätzung der Modellleistung zu erhalten. Dieser Prozess kann nach erfolgreicher Identifizierung des geeigneten Modells ebenfalls in die Pipeline integriert werden (siehe Ausschnitt 3.8 unter „scores“).

```
from sklearn.model_selection import cross_val_score

rs_pipeline = Pipeline(steps=[('preprocessor_rs', preprocessor_rs),
                              ('model_rs', XGBRegressor(random_state=0))
    ])

scores = cross_val_score(rs_pipeline, X_train_rs, y_train_rs, cv=10,
    scoring='r2')

rs_pipeline.fit(X_train_rs, y_train_rs)
```

Listing 3.8: Implemtierung der Kreuzvalidierung

Die Verbesserung bzw. Optimierung des Modells erfolgt durch Anpassung der modellinternen Hyperparameter, die zur Regulierung des Modells beitragen, um eine Überanpassung zu vermeiden [50, 67]. Die bereits in *Sklearn* implementierten Klassen, wie *GridSearchCV* oder *RandomizedSearchCV* können dazu verwendet werden [55]. Für das finale Modell wird *GridSearchCV* verwendet, um die besten Parameter für das Tuning zu finden. Die verwendeten Parameterbereiche sowie die Implementierung sind der Abbildung 3.9 zu entnehmen.

```
from sklearn.model_selection import GridSearchCV

parameters = {
      "model__n_estimators": [100, 200, 300, 500]
      "model__learning_rate": [0.05, 0.10, 0.15, 0.3, 0.5],
      "model__max_depth": [3, 4, 5, 6, 8]
      }

grid_search = GridSearchCV(
        model_pipeline,
        param_grid=parameters,
        cv=10,
        scoring='neg_mean_squared_error'
        )

grid_search.fit(X_train, y_train)

best_params = grid_search.best_params_
```

Listing 3.9: Hyperparameter Tuning

Mit den identifizierten Parameter kann ein finales Modell trainiert werden und zielt darauf ab, die Hyperparameter zu finden, um die Leistung des Modells zu maximieren und eine Überanpassung zu vermeiden (vgl. Ausschnitt 3.10). Ist die Leistung des Modells zufriedenstellend, kann anstatt der partitionierten Trainingsdaten der komplette Datensatz verwendet werden, falls mehr Daten das Modell verbessern.

```
pipeline.named_steps['model_rs'].set_params(**best_params)
final_model = pipeline.fit(X_train_rs, y_train_rs)
```

Listing 3.10: Training des Modells mit den besten Parametern

Ein letzter Schritt bei der Modellerstellung ist die Speicherung des Modells sowie der vorgelagerten Vorverarbeitungsschritte. Auf diese Weise kann sichergestellt werden, dass auf das Modell zugegriffen werden kann, um beispielsweise Verbesserungsmöglichkeiten vorzunehmen, es mit anderen Modellen zu vergleichen oder das Modell in die Produktion zu schicken (engl. Deploying), um Vorhersagen für den Anwendungsfall zu treffen. Für die Speicherung und das Laden, kann die Bibliothek *joblib* verwendet werden. Darüber kann das Modell als *.pkl*-Datei (Pickle) gespeichert und geladen werden (vgl. Ausschnitt 3.11).

```
import joblib

joblib.dump(final_model, 'tuned_model.pkl')
model = joblib.load('tuned_model.pkl')
model.predict(X)
```

Listing 3.11: Speichern und Laden des finalen Modells

3.2.4 Sensitivitätsanalyse und Feature Importance

Die Bedeutung der Merkmale in einem Modell wird durch den relativen Beitrag jedes Merkmals zum Entscheidungsprozess des Modells dargestellt. Diese Bedeutung oder Relevanz (engl. *Feature Importance*) wird auf der Grundlage berechnet, wie stark jedes Merkmal die gewichtete Unreinheit in einem Baum verringert, gemittelt über alle Bäume im Wald. Eine hohe Relevanz zeigt an, dass ein Merkmal einen erheblichen Einfluss auf die Vorhersagen des Modells hat [68]. Änderungen an den Werten dieses Merkmals werden voraussichtlich signifikante Auswirkungen auf die Ausgabe haben. Eine niedrige Relevanz deutet darauf hin, dass Änderungen an den Werten dieses Merkmals keinen signifikanten Einfluss auf das Modellergebnis haben. Wenn ein Merkmal einen Bedeutungswert von Null hat, trägt es nicht zum Entscheidungsprozess des Modells bei. Bei der Interpretation dieser Bedeutungen ist es wichtig, die Art des Modells und der Daten zu berücksichtigen. Dabei ist zu beachten, dass eine hohe Relevanz von Merkmalen keine Kausalität impliziert und sich von Modell und Eingangsdaten unterscheiden kann. Insofern ist bei Entscheidungen, die auf diesen Werten basieren, Vorsicht geboten, insbesondere bei Vorhandensein korrelierter Merkmale, welche die wahrgenommene Wichtigkeit verwandter Merkmale beeinflussen können. Bereits korrelierte Merkmale sind im Vorfeld aus dem Datensatz isoliert, ersetzt oder mit anderen kombiniert.

Interpretation mit SHAP

SHAP (SHapley Additive exPlanations) ist ein Erklärungsansatz im ML, der auf der Spieltheorie basiert [69]. Es bewertet den Beitrag einzelner Merkmale zu den Vorhersagen eines Modells. SHAP ermöglicht eine detaillierte und faire Zuschreibung von Einflüssen einzelner Features auf das Ergebnis eines Modells [70]. Dieser modellagnostische Ansatz kann auf eine Vielzahl von Modellen angewendet werden, um die Transparenz und Nachvollziehbarkeit komplexer Modelle zu verbessern. SHAP hilft bei der Erklärung von Einzelvorhersagen und bietet Einblicke in die globale Wichtigkeit von Merkmalen und interpretiert den geleisteten Beitrag [71]. Dabei wird ein Vektor v mit den Beiträgen jedes Merkmals zur Vorhersage für jedes Eingabeobjekt und dem Erwartungswert der Modellvorhersage für das Objekt geschaffen.

- v_i ist der Beitrag des i_{th} features.
- $v_{feature\_count}$ ist der Erwartungswert für die Vorhersage des Modells.

Für ein gegebenes Objekt ist die Summe $\sum_{i=0}^{feature\_count} v[i]$ gleichwertig zur Vorhersage des Objekts [69]. Die *feature importance* $ShapValues_i$ wird wie folgt über jedes feature i berechnet:

$$ShapValues_i = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!\,(M - |S| - 1)!}{M!} [f_x(S \cup \{i\}) - f_x(S)], wo \tag{3.1}$$

- M ist die Anzahl der Eingangsmerkmale.
- N ist die Menge von allen Eingangsmerkmalen.
- S ist die Menge der Merkmalsindizes, die nicht Null sind (die Merkmale, die beobachtet werden und nicht unbekannt sind).
- $f_x(S) = E[f(x)|x_s]$ ist die Vorhersage des Modells für die Eingabe x, wobei $E[f(x)|x_s]$ der Erwartungswert der Funktion unter der Bedingung einer Teilmenge S der Eingabemerkmale ist.

SHAP stellt eine Erweiterung der allgemeinen *Feature Importance* von maschinellen Lernmodellen dar, indem es zusätzlich zur Relevanz den spezifischen Beitrag für einzelne Vorhersagen ausgibt. Dadurch können Merkmale, die eine großen Einfluss auf die Vorhersagen haben stärker priorisiert werden als Merkmale mit geringer Relevanz.

4 Ergebnisse

Es werden drei maschinelle Lernalgorithmen getestet, u.a LR, DT sowie zwei Ensemble-Methoden, nämlich RF und XGBoost. Die Ergebnisse der ML- und Ensemble-Methoden sind in den Tabellen 4.1.1 und 4.1.2 beschrieben. Die Ergebnisse dieser Arbeit zeigen, dass anhand von maschinellen Lernmethoden (eine vielversprechende Alternative) zur Energiebedarfsbestimmung von WG ggü. herkömmlichen Methoden herausgestellt werden kann. Die Modelle können die Energieeffizienz mit hoher Genauigkeit bestimmen. Weiterhin können die Modelle auch verschiedene Einflussfaktoren erkennen und quantifizieren, wie z.B. das Gebäudealter, die Gebäudefläche oder die Gebäudeform. Im weiteren Verlauf werden die Modellergebnisse über die erstellten Datensätze beschrieben.

4.1 Modellergebnisse und Validierung

Ein wichtiger Aspekt der Modellleistung ist die Über- und Unteranpassung von statistischen Modellen. Unteranpassung tritt auf, wenn das Modell bei den Trainingsdaten eine sehr geringe Leistung aufweist und daher die Daten nicht korrekt darstellen kann. Bei Unteranpassung kann das Modell zu „einfach“ sein, um genaue Ergebnisse zu erzeugen, da es nicht in der Lage ist, Muster oder Beziehungen zwischen Eingabe- und Ausgabemerkmalen für Daten zu extrahieren. Um das Problem der unzureichenden Anpassung zu lösen, gibt es verschiedene Techniken: [1] Der Datensatz kann durch Hinzufügen effektiverer Merkmale neu formuliert oder ergänzt werden. [2] Es können geeignete Vorverarbeitungsmethoden ausgewählt werden, um die Nachteile fehlender Daten oder Ausreißerwerte zu beheben. [3] Die Menge oder Art von Modellregularisierungstechniken kann verringert oder geändert werden. Das Problem der Überanpassung tritt auf, wenn die Modellleistung bei den Trainingsdaten sehr hoch ist, bei den Validierungs- oder Testdaten jedoch niedrig ausfällt. Der Grund für das Problem der Überanpassung liegt in der Unfähigkeit des Modells, die globale Lösung des Problems für den gesamten Datensatz zu erreichen und die Verallgemeinerung auf neue Daten gering ausfällt. Um eine geringere Leistung für ungesehene Daten zu vermeiden, empfiehlt es sich, weniger Merkmalskombinationen zu verwenden und den Umfang der Regularisierung zu erhöhen [72].

4.1.1 Basismodelle

In Tabelle 4.1.1 sind die Ergebnisse der betrachteten Basismodelle dargestellt. Hierbei findet eine Modellunterscheidung zwischen Regression und Klassifikation statt, um Un-

terschiede in der Leistungsbetrachtung anzugeben. Für kontinuierliche, numerische Zielmerkmale wird lineare Regression verwendet. Das Modell der linearen Regression zeigt einen R^2-Wert von 0,65, was bedeutet, dass es 65% der Varianz in der Zielvariable mit den gegebenen Merkmalen erklären kann. Dies kann als moderat effektiv angesehen werden, deutet jedoch auf eine Unteranpassung hin.

Der DT ist ein Modell für Klassifikation- als auch für Regressionaufgaben und erreicht eine durchschnittliche Genauigkeit von 66%. Diese Genauigkeit gibt an, wie oft das Modell die korrekte Klassifikation der Energielabel vorhersagt. Entscheidungsbäume sind typischerweise bekannt für ihre Fähigkeit komplexe Datenstrukturen abzubilden. Allerdings können DT anfällig für Überanpassung sein, da diese solange „wachsen“ und Blätterknoten (engl. leaf nodes) ausbilden, bis die resultieren Blätter rein sind [73].

Der RF hingegen, eine Erweiterung des DT-Modells, kann durch die Integration mehrerer Bäume (Ensemble) die Varianz reduzieren und kann eine Überanpassung vermeiden. Das Basismodell zeigt ggü. den anderen Modellen eine höhere, mittlere Genauigkeit von 73%. Dies stellt im Rahmen der maschinellen Lernmethoden eine signifikante Verbesserung gegenüber dem einfachen DT-Classifier dar und bestätigt die Effektivität des RF-Ansatzes für komplexere Datenstrukturen, sowohl bei Regressions- als auch Klassifikationsaufgaben.

Tabelle 4.1: Evaluation der Basismodelle mit k=10 Kreuzvalidierung; (a) vor Regulierung und (b) nach Regulierung der Hyperparameter

Modell	Mean Accuracy (a)	Mean R^2(a)	Mean Accuracy (b)	Mean R^2 (b)
LR	-	0,65	-	-
DT	0,66	-	0,63	-
RF	0,73	-	0,68	-

Während LR die Regularisierung nicht direkt unterstützt, können verwandte Modelle, wie Ridge (L2 Regularisierung) oder Lasso (L1 Regularisierung) verwenden, um einen Regularisierungsterm hinzuzufügen. Diese Modelle sind mit Hyperparametern (bspw. *Alpha*) ausgestattet, die die Stärke der Regularisierung steuern [55].

4.1.2 Extreme Gradient Boosting (XGB)

Die Evaluierungsergebnisse vom XGBoost-Model sind in Tabelle 4.1.2 über alle Datensätze dargestellt. Das Modell kann auf Basis des *Fernerkundung*-Datensatzes die geringste Varianz erklären und besitzt den höchsten Fehler. Dies deutet darauf hin, dass das Modell hinsichtlich der Hyperparameter optimiert werden muss, um effektiv das Muster in den Daten zu erkennen. Der angepasste Datensatz *Fernerkundung+*, der zusätzlich mit dem Merkmal des aktuellen Energieverbrauchs trainiert wurde, besitzt mit dem *datengetriebenen* Ansatz die höchste Leistung ohne weitere Optimierung. Um die Leistung des

Modells zu verbessern, müssen die modellinternen Parameter angepasst werden, um ein finales Abbild der Modellleistung zu erhalten.

Tabelle 4.2: Evaluation des XGB-Modells über alle Datensätze mit k=10 Kreuzvalidierung

Datensatz	Mean R^2	Mean MSE	Mean MAE
Fernerkundung	0,61	39,62	4,12
Fernerkundung+	0,82	18,18	2,18
Datengetrieben	0,84	13,52	2,11
Erschöpfend	0,74	26,41	3,3

Die Abbildungen 4.1 und 4.4 zeigen zwei verschiedene Kurven: Zum einen zwei Validierungskurven für das XGB-Regressor Modell bezogen auf unterschiedliche Hyperparameter (Lernrate und maximale Tiefe) und zwei Lernkurven, die die Modellleistung in Bezug auf die Anzahl der Trainingsbeispiele zeigen. Bei der Validierungskurve 4.1 bleibt die rote Kurve (Trainings-Score) relativ konstant und hoch über den Bereich der Lernraten, was zeigt, dass das Modell sich den Trainingsdaten gut anpasst. Die grüne Kurve (Cross-Validation-Score) erreicht ein Plateau bzw. einen Höchstwert bei niedrigeren Lernraten und fällt daraufhin ab, wenn die Lernrate zunimmt. Dies deutet darauf hin, dass das Modell bei einer niedrigen Lernrate am besten verallgemeinert und dass eine höhere Lernrate zu einer schlechteren Generalisierung führt. Die schattierte Fläche um die grüne Kurve (Variabilität der Cross-Validation-Scores) ist relativ schmal, was darauf hindeutet, dass die Modellleistung über verschiedene Validierungssets konsistent ist. In Abb. 4.2 steigt der Trainings-Score mit zunehmender maximaler Tiefe an, was den Erwartungen entspricht, da tiefe Bäume die Trainingsdaten genauer anpassen können. Der *Cross-Validation-Score* steigt ebenfalls zunächst an, fällt aber nach einer bestimmten Tiefe ab. Dies zeigt, dass die Erhöhung der Baumtiefe eine Überanpassung zur Folge hat, was zu einer schlechten Generalisierung für neue, unbekannte Daten führt. Die optimale, maximale Tiefe liegt im mittleren Bereich der getesteten Werte, wo der Cross-Validation-Score am höchsten ist.

In den Abb. 4.3 und 4.4 sind die Lernkurven für den FED und FED+ Datensatz dargestellt. Während bei den Validierungskurven R^2-Scores verwendet wird, werden bei den Lernkurven der Mean Squared Error (MSE) verwendet, bei dem niedrigere Werte besser sind. Auf beiden Kurven bleibt der Trainings-Score (rot) sehr niedrig und konstant, was darauf hindeutet, dass das Modell mit mehr Daten nicht notwendigerweise besser wird, möglicherweise wegen eines hohen Bias. Der Cross-Validation-Score (grün) verbessert sich mit mehr Daten, nähert sich aber einem Plateau an, was darauf hindeutet, dass das Hinzufügen weiterer Daten allein die Modellleistung nicht wesentlich verbessern wird. Der Unterschied zwischen den Trainings- und Cross-Validation-Scores ist besonders im Fall des MSE sehr groß, was auf eine mögliche Überanpassung oder hohe Varianz im Modell hinweisen könnte.

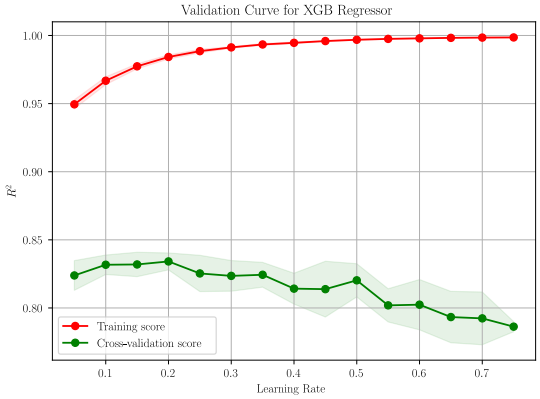

Abbildung 4.1: Validierung der Lernrate

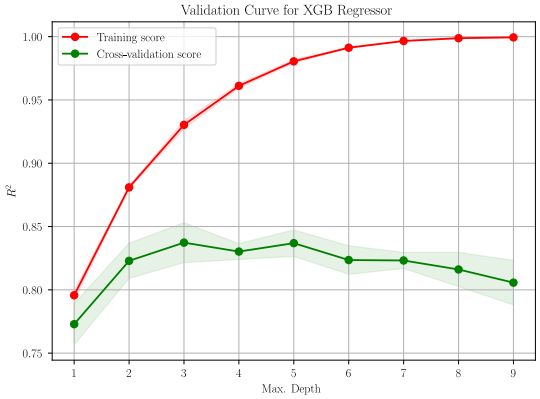

Abbildung 4.2: Validierung der Baumtiefe

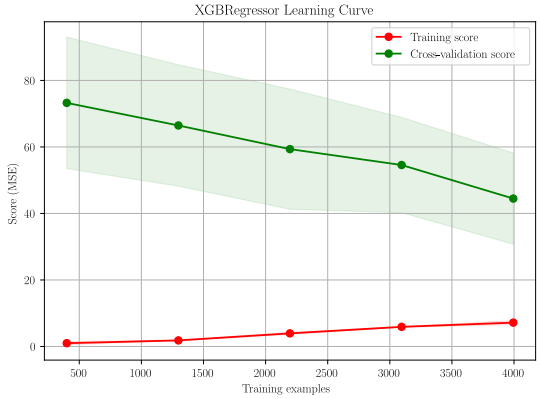

Abbildung 4.3: Lernkurve für FED

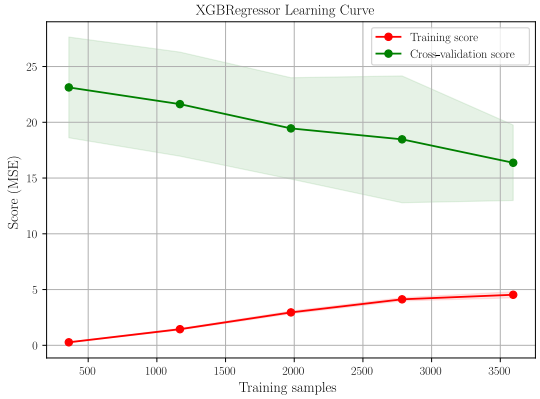

Abbildung 4.4: Lernkurve für FED+

Basierend auf diesen Kurven könnten folgende Maßnahmen ergriffen werden: Die Modellkomplexität kann geprüft und gegebenenfalls eine weitere Anpassung der Hyperparameter erfolgen, um eine Überanpassung zu verhindern (z.B. durch Einführung von Regularisierung). Weiterhin kann versucht werden eine größere Datenmenge zu verwenden, um zu sehen, ob sich die Lernkurven weiterhin verbessern. Eine weitere Möglichkeit zur Optimierung des Modells stellt die Feinabstimmung der Lernrate und der maximalen Tiefe dar. Dabei ist es zielführend, das Gleichgewicht zwischen Bias und Varianz[1] zu finden, das zu einer optimalen Leistung des Modells führt.

[1] Bekannt als *Bias-Variance-Tradeoff*: Der Kompromiss zwischen der Fähigkeit eines Modells, komplexe Muster in den Daten zu erkennen (Varianz) und seiner Genauigkeit bei der Vorhersage von Daten, die es nicht während des Trainings gesehen hat (Bias) [67].

4.2 Analyse der Schlüsselfaktoren und Interpretation der Modellentscheidungen

Über die *Feature Importance* (FI) der Merkmale kann die Relevanz herausgestellt werden, die für das Modell besonders entscheidend sind. Da die Ausprägungen der kategorialen Merkmale kodiert sind (0/1), ist es schwierig, den Beitrag der einzelnen Ausgangsmerkmale zum endgültigen Modell zu bestimmen. Als Alternative zeigt die Abbildung 4.5 eine gewichtete Sicht der Merkmale, die sich aus der Summe jedes kodierten Merkmals in der XGBoost-Bedeutung ergibt. Nur kodierte Merkmale mit einer Bedeutung größer als 0 werden berücksichtigt. Daraus geht hervor, dass bspw. der Isolierungsgrad für Dach und Wände sowie die Baualtersklasse die höchste Relevanz für das Modell haben. Die aggregierte Merkmalsrelevanz für den datengetriebenen, erschöpfenden und FED+-Ansatz sind dem Anhang E beigefügt.

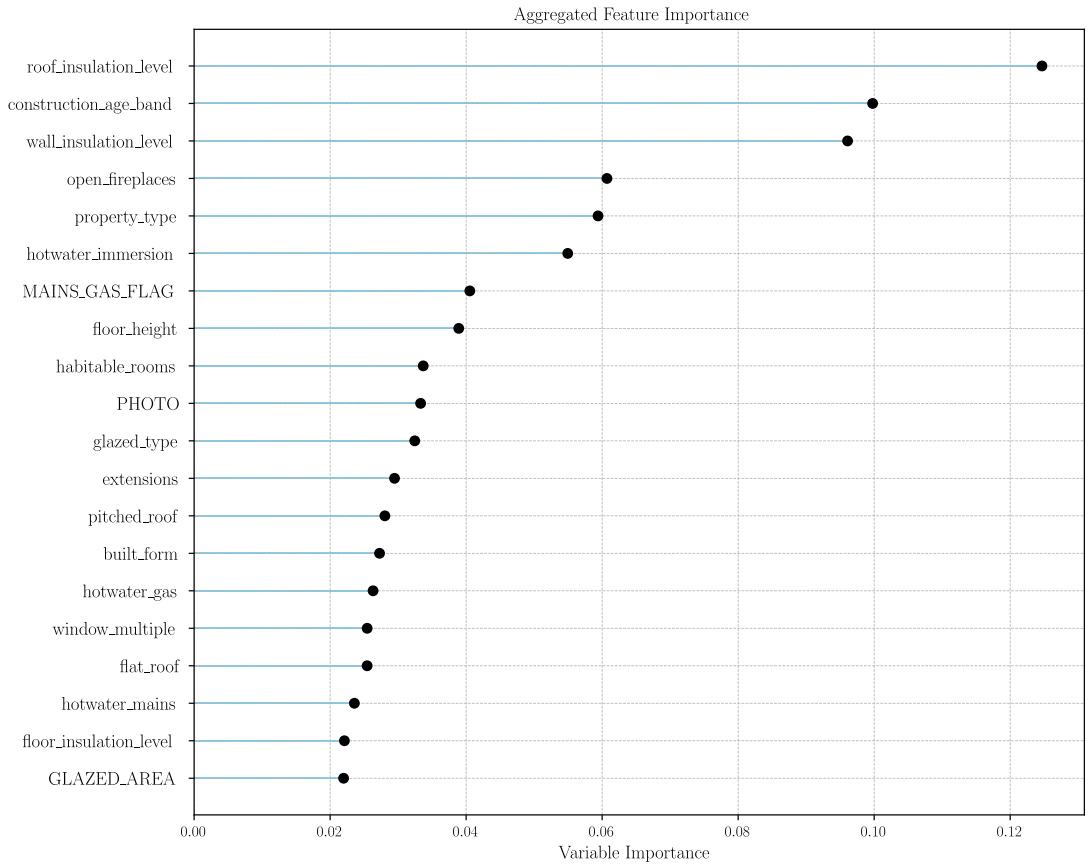

Abbildung 4.5: Aggregierte Relevanz der Merkmale für Fernerkundungsdaten

In Abbildung 4.6 ist die SHAP-Zusammenfassung abgebildet. Dabei basiert die Interpretation darauf, dass die Merkmale in absteigender Reihenfolge ihrer Bedeutung von oben nach unten geordnet sind. Ein einflussreiches Merkmal ist das *wall_ insulation_ level*, was

darauf hindeutet, dass das Isolationsniveau der Wände eine Schlüsseldeterminante für die Energieeffizienz eines Gebäudes ist. Die SHAP-Werte geben die Richtung der Auswirkung eines Merkmals auf die Modellausgabe an. Ein positiver SHAP-Wert für ein Merkmal deutet darauf hin, dass das Merkmal die Vorhersage der Energieeffizienz erhöht, während ein negativer SHAP-Wert auf einen Rückgang der vorhergesagten Energieeffizienz hindeutet. Eine positive Auswirkung (rot) lässt sich bei Merkmalen wie *wall_ insulation_ level* und *MAINS_ GAS_ FLAG* feststellen, die in der Regel positive SHAP-Werte aufweisen. Dies deutet darauf hin, dass ein höheres Niveau der Wandisolierung und das Vorhandensein von Gasanschlusses mit einer höheren Energieeffizienzvorhersage verbunden sind. Demgegenüber stehen negative Auswirkungen (blau), welche durch die Merkmale *roof_ insulation_ level* und *solid_ floor_ type* repräsentiert werden. Hierbei ist zu beachten, dass das Fehlen einer Dachisolierung sowie das Vorhandensein von Massivböden mit niedrigeren Energieeffizienzprognosen assoziiert sind. Die Farbe des Werts gibt Aufschluss über den Wert des Merkmals (hoch oder niedrig), wobei *Rot* für höhere und *Blau* für niedrigere Werte steht. Dies lässt sich beispielsweise anhand des Merkmals "Wandisolierung" veranschaulichen. Gebäude mit einer besseren Wandisolierung weisen demnach tendenziell eine höhere, prognostizierte Energieeffizienz auf.
Eine alternative Darstellungsform stellt der *Waterfall-Plot* in Abb. 4.7 für eine einzelne Vorhersage dar. Die x-Achse repräsentiert den SHAP-Wert, welcher den Einfluss der einzelnen Merkmale auf die Vorhersage des Modells quantifiziert. Die y-Achse zeigt die Merkmale, die den größten Einfluss auf die Vorhersage für diesen speziellen Fall haben. Jedes Merkmal ist mit einem Wert von 0 oder 1 versehen, welcher den binären Status dieses Merkmals für den analysierten Fall darstellt. Die Balken repräsentieren den Beitrag jedes Merkmals zur finalen Vorhersage. Rote Balken zeigen einen Anstieg der Vorhersage des Energieeffizienzwertes an, während blaue Balken einen Rückgang anzeigen. Das Diagramm beginnt mit einem Basiswert $E[f(X)]$, der die durchschnittliche Modellleistung über den zum Training des Modells verwendeten Datensatz darstellt. In diesem Fall beträgt er ungefähr 67,892. Der endgültige Wert $f(x)$ auf der rechten Seite stellt den tatsächlich vorhergesagten Energieeffizienzwert für diesen speziellen Fall dar, der ungefähr 75,893 beträgt. Die einflussreichsten Merkmale scheinen mit dem Dachtyp und der Art des Gebäudes zusammenzuhängen. Die Isolierung stellt einen entscheidenden Faktor für die Energieeffizienz dar, wie die negativen Auswirkungen einer fehlenden Isolierung zeigen. Die Summe der Auswirkungen anderer Merkmale deutet darauf hin, dass sich viele kleine Beiträge zu einer signifikanten Auswirkung auf die Vorhersage des Modells summieren können. Es besteht die Möglichkeit, dass Wechselwirkungen oder Multikollinearität zwischen den Merkmalen bestehen (beispielsweise verschiedene Dämmungsattribute), die einer weiteren Untersuchung bedürfen, um ihre individuellen Auswirkungen vollständig zu verstehen. Für das vorliegende Beispiel wird ein höherer Energieeffizienzwert als der Durchschnitt vorhergesagt, was vor allem auf die Art des Gebäudes (eine Wohnung) und das Fehlen bestimmter Merkmale zurückzuführen ist, die normalerweise die Ener-

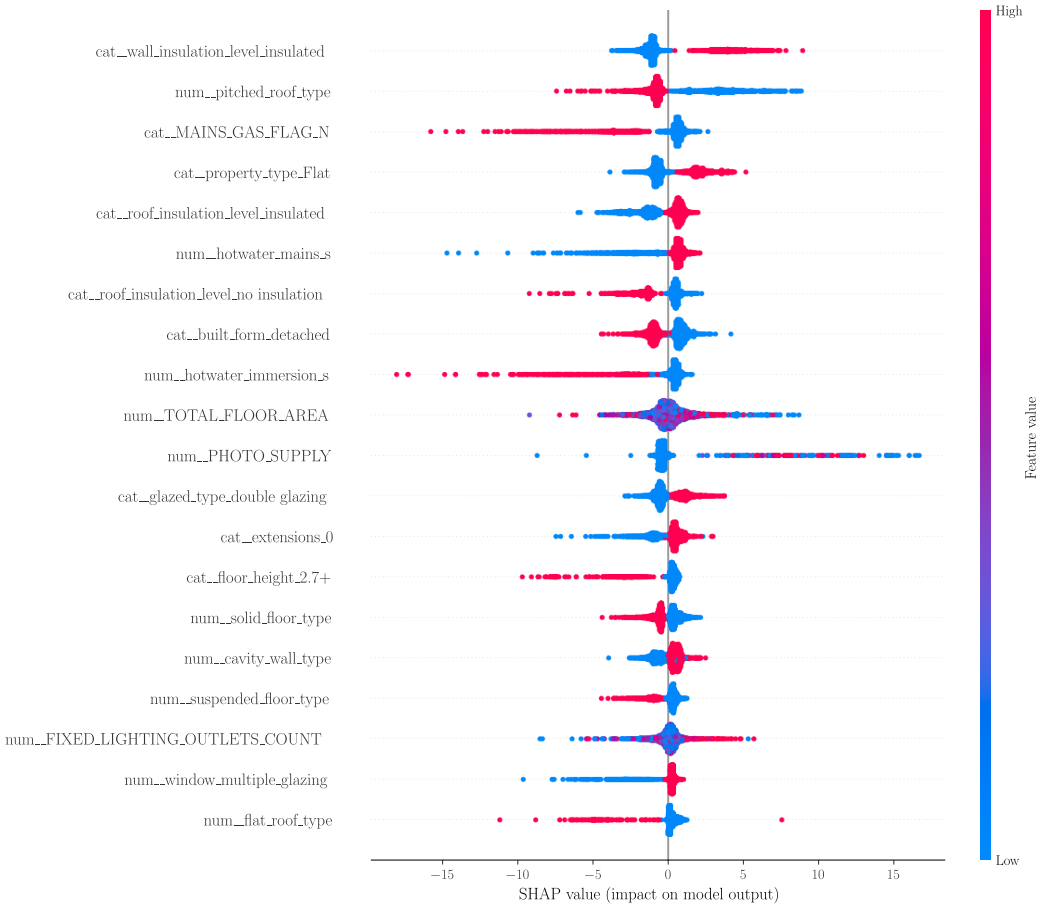

Abbildung 4.6: SHAP Zusammenfassung auf Basis des Fernerkundungsdatensatzes

gieeffizienz verringern (z. B. ein geneigtes Dach). Es sei darauf hingewiesen, dass binäre Werte von 0 und 1 im Kontext der Kodierung der Daten interpretiert werden müssen. Zusammenfassend kann mit dieser Darstellungsweise der Einfluss auf die Vorhersagen quantifiziert werden, um das Model und deren Merkmale zu interpretrieren.

Zusätzlich zum Regressionsmodell für die Vorhersage der Energieeffizienz, ist in Abbildung 4.8 die Konfusionsmatrix für das experimentelle XGB-Modell zur Klassifizierung der aktuellen Energielabel dargestellt. Das Klassifikationsmodell erreicht mit Hyperparametertuning eine Genauigkeit von 0,71, um das richtige Energielabel vorherzusagen. Es ist aber auch ersichtlich, dass obere und untere Abweichungen hinsichtlich der Energieklasse vorliegen. Bezieht man diese Information als Bewertungsgrundlage mit ein, erreicht das Klassifikationsmodell eine Genauigkeit (Accuracy) von 0,97 bei ± 1 Abweichung zum wahren Energielabel.

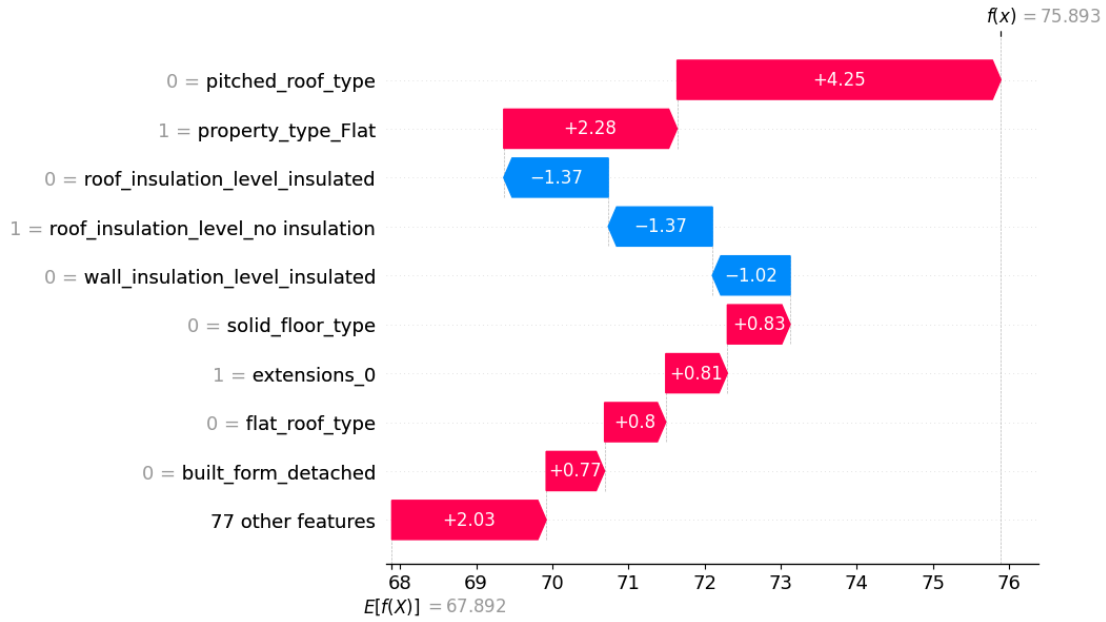

Abbildung 4.7: SHAP Zusammenfassung: Beitrag der Merkmale zur ersten Vorhersage

4.3 Robustheit und Generalisierbarkeit der Modelle

Die Modellkonfiguration sowie das Hyperparameter-Tuning für das XGB-Modell in Kombination mit Verbrauchsdaten (FED+) ist in Tabelle 4.3 dokumentiert. Die Validierung erfolgt anhand des Trainings- und Testdatensatzes. Auf beiden Datensätzen erzielt das Modell ein Bestimmtheitsmaß R^2 von 0,94 und MAE von 1,46, d.h. mittlere Abweichung um 1,46 auf einer Ratingskala von 1-100 für unbekannte Daten. Dies zeigt, dass das Modell die Varianz in den Daten effektiv erfassen kann. Im Rahmen einer Sensitivitätsanalyse könnte die Reaktion des Modells auf Veränderungen in den Eingabedaten getestet werden bspw. durch die Reduktion von Eingangsmerkmalen. Bezüglich der Generalisierbarkeit bleibt offen, inwiefern eine Übertragbarkeit möglich, da die spezifischen Daten hinsichtlich der regionalen Gebäudetypologie variieren. Grundsätzlich kann aber die Aussage getroffen werden, dass durch Ergänzung des aktuellen Energieverbrauchs eine gute Verallgemeinerungsleitung mit geringem Fehler auf dem Testdatensatz besteht.

Tabelle 4.3: Ergebnisse der Kreuzvalidierung und Hyperparametersuche für FED+

Parameter	Hyperparameterbereich	Wert
Anzahl der Bäume (n_estimators)	100 - 500	300
Lernrate (learning_rate)	0,1 - 0,4	0,1
Baumtiefe (max_depth)	3 - 8	4

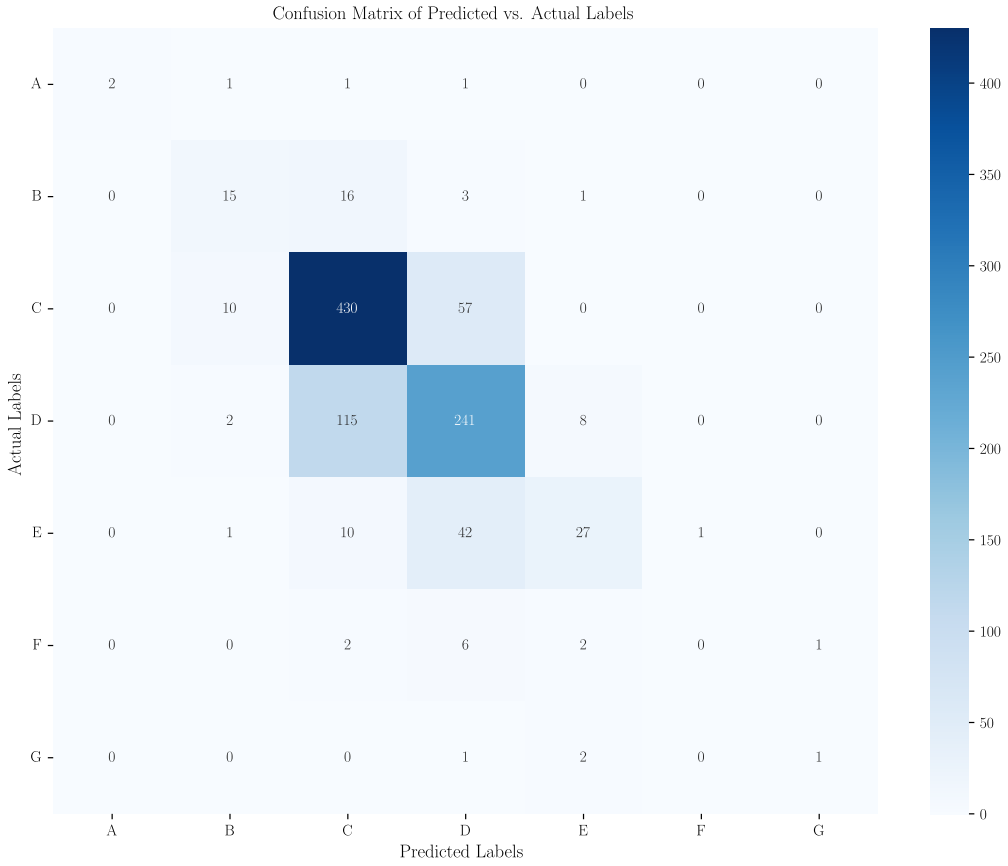

Abbildung 4.8: Konfusionsmatrix der vorhergesagten und wahren Energielabel

4.4 Skalierung der Pipeline

Da der bisher untersuchte Datensatz nahezu keine Datenpunkte für Gebäude der Energieklasse A und G aufweist, werden im weiteren 967.937 Datenpunkte bezogen. Dies dient unter anderem dazu, ein repräsentatives Abbild des Gebäudebestands zu bekommen und weitere Muster aufzudecken. Weiterhin kann auf Basis der Lernkurve 4.3 die Aussage getroffen werden, dass mehr Datenpunkte die Modellleistung verbessert.

Dabei werden die Daten zufällig unter der Bedingung einer ausgewogenen Anzahl von Energielabeln A-G ausgewählt. Insgesamt weist der neue Datensatz 150.000 Datenpunkte für jedes Energielabel auf. Eine Ausnahme stellt das Energielabel A mit 67.937 Datenpunkten dar, da die Anzahl der Einträge nicht mehr ausreichend ist. Im Rahmen der Skalierung werden die gleichen Transformationsschritte wie beim FED-Datensatz vorgenommen, um ein Benchmark zum Vorgängermodell vorzunehmen. Das Modell wird mit den gleichen Parametern, wie in Tabelle 4.3 trainiert. Die erhöhte Anzahl an Instanzen hat auch zur Folge, dass neue, bislang nicht beobachtete Ausprägungen innerhalb des Datensatzes aufgetreten sind, die bei der Merkmalsextraktion berücksichtigt werden müssen. Dies betrifft insbesondere die Merkmalsbeschreibung für Dach und Wände. Die

Ergebnisse können der Tabelle 4.4 entnommen werden.

Tabelle 4.4: Evaluation des XGB-Modells über den skalierten FED-Datensatz mit k=10 Kreuzvalidierung

Datensatz	Mean R^2 Train	Mean R^2 Test	Mean MSE Test	Mean MAE Test
FED	0,83	0,83	122,18	8,16

Im Vergleich zum Vorgängermodell (vgl. 4.1.2) mit 5.000 Datenpunkten, hat sich das Bestimmtheitsmaß R^2 von 0,61 auf 0,83 erhöht, was einer wesentlichen Steigerung entspricht. Eine Erhöhung des MAE und MSE bei gleichzeitiger Verbesserung von R^2 lässt den Schluss zu, dass das Modell zwar besser darin geworden ist, die Muster in den Daten zu erfassen, jedoch Schwierigkeiten hat, präzise Vorhersagen bei einzelnen Datenpunkten zu treffen. Als mögliche Gründe für die Veränderungen können die höhere Datenkomplexität, das Auftreten von mehr Ausreißern oder eine größere Varianz in den neuen Daten angenommen werden. An den Stellen zeigt die Analyse einer weiteren Lernkurve, ob mehr Daten diesen Fehler verringern oder sogar erhöhen (vgl. 4.9)

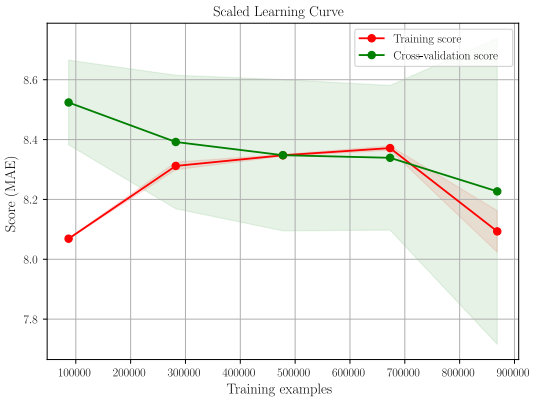

Abbildung 4.9: FED-Lernkurve skaliert

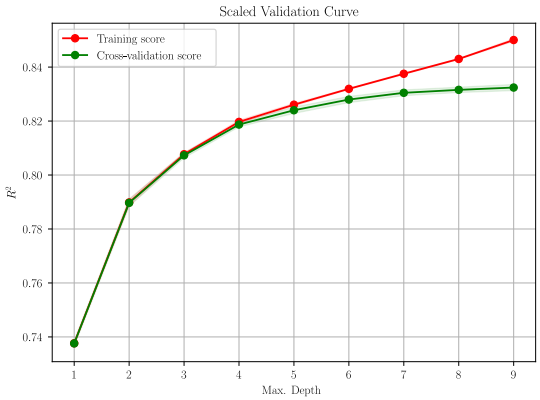

Abbildung 4.10: FED-Validierungskurve

Die Analyse der Lernkurve zeigt, dass der Trainings-Score mit zunehmender Datenmenge sinkt, während der Cross-Validation-Score (CVS) sich verbessert. Die Ergebnisse deuten auf eine wirksame Generalisierbarkeit des Modells hin, da die Werte beider Scores eine Annäherung zeigen und ein Plateau erreichen. Dies lässt darauf schließen, dass das Modell gut ausbalanciert ist, jedoch möglicherweise keine signifikanten Verbesserungen durch zusätzliche Trainingsdaten mehr erzielt werden können. Weiterhin ist die Unsicherheit (grüner Bereich) vom CVS sehr breit, was auf eine erhöhte Inkonsistenz innerhalb der Validierungsergebnisse hinweist. Dies lässt den Schluss zu, dass Anpassungen am Modell selbst oder durch weitere Merkmalsentwicklung erzielt werden müssen. Die aggregierte Merkmalsrelevanz des Modells wird in der Abbildung 4.11 anhand der Relevanz der einzelnen Merkmale verdeutlicht. Hierbei zeigt sich, dass das Merkmal *wall_ type* die höchste

Relevanz aufweist. Dies lässt den Schluss zu, dass die Wandkonstruktion einen primären Einflussfaktor auf die Modellergebnisse darstellt. Die signifikante Rolle dieses Merkmals lässt sich möglicherweise auf die Vielfalt seiner Ausprägungen zurückführen, die eine eine hohe Varianz innerhalb des Datensatzes aufweisen. Des Weiteren ist es auch naheliegend, dass die Wandkonstruktion einen direkten Einfluss auf die thermischen Eigenschaften eines Gebäudes ausübt, insbesondere in Kontexten, in denen die Energieeffizienz bewertet wird. Die direkte Korrelation mit dem Zielparameter könnte eine Erklärung dafür sein, warum der Wandtyp als dominant in der Merkmalsrelevanz hervortritt. Weiterhin ist zu berücksichtigen, dass dieses Merkmal in Wechselwirkung mit weiteren relevanten Merkmalen, wie der Fensterstruktur und der Dachisolierung, steht, was seine Bedeutung im Modell weiter erhöht. Dies lässt den Schluss zu, dass zukünftige Optimierungen oder detailliertere Untersuchungen bezüglich der Wandkonstruktionen einen merklichen Einfluss auf die Modellgenauigkeit haben könnten.

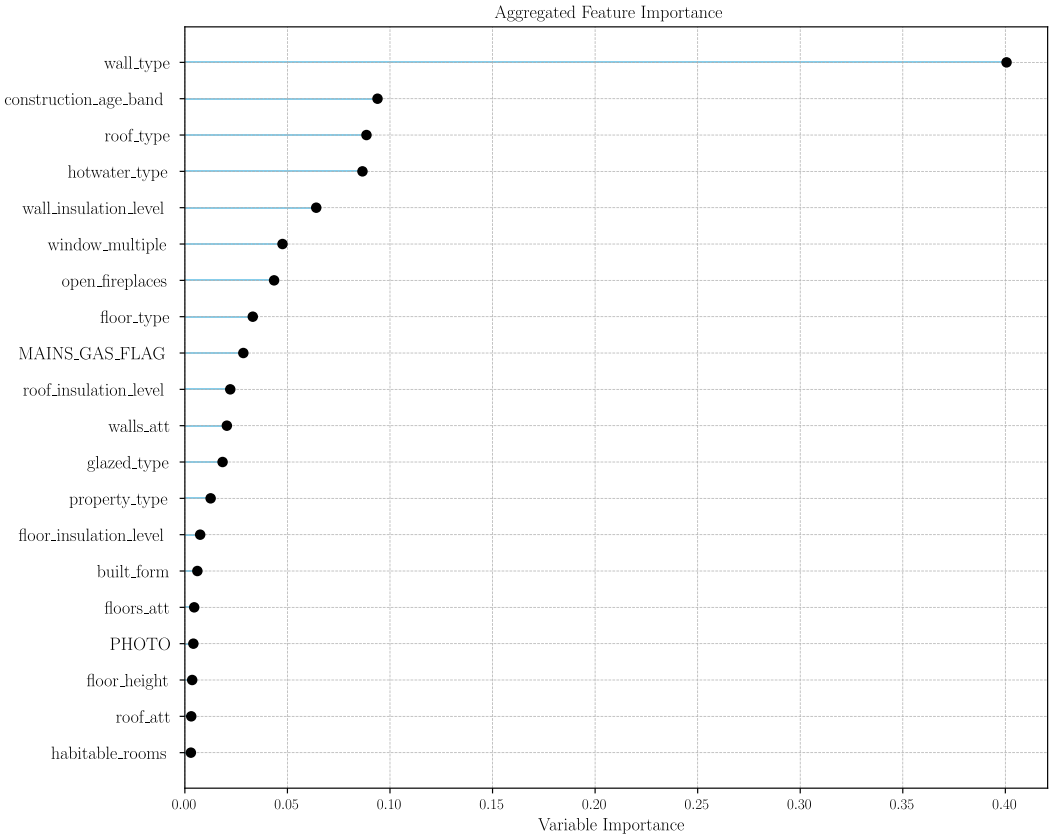

Abbildung 4.11: Aggregierte Merkmalsrelevanz für den skalierten FED-Datensatz

4.5 Resultierendes Datenschema

Aus der Analyse der aggregierten Merkmalsrelevanz in Abb. 4.5, der SHAP-Analyse und Literaturrecherche in Kap. 2.1.2, kann eine Empfehlung getroffen werden, welche Merkmale zukünftig bezüglich der Energiebedarfsbestimmung von WG priorisiert werden sollten. In Tabelle 4.5 sind die Schlüsselfaktoren aus der Modellanalyse zusammengefasst, die den größten Einfluss haben. Das resultierende Datenschema ergibt sich aus der Analyse der bestehenden Datensätze.

4.6 Erweiterung des Datenschemas für die Übertragbarkeit

Zur Übertragung der verwendeten Methodik nach Deutschland müssen geeignete Datenquellen akquiriert werden. Wie bereits angemerkt, existiert für Deutschland keine öffentliche Datenbank für EA. Daher müssen alternative Quellen untersucht werden. Der veröffentlichte „KWW-Datenkompass zur Kommunalen Wärmeplanung“ [75] kann dabei als umfassender Leitfaden zur Datenerhebung dienen. Der Fokus dieses Leitfadens liegt bei der Beschaffung und Nutzung verschiedener Datenquellen zur Unterstützung effizienter Planungsprozesse innerhalb der kWP. Weiterhin soll dieser Datenkompass zu einer Metadatenbank ausgebaut werden, um die Daten in Zukunft direkt abrufen zu können [75]. Der Datenkompass deckt dabei eine breite Palette von Themenbereichen ab, die u.a. für die Wärmeplanung von Bedeutung sind. Dabei gibt die Tabelle F im Anhang die Themengruppen und verknüpften Datenquellen wieder. Die Datenerhebung stützt sich auf eine Vielzahl von öffentlich zugänglichen und auf Anfrage verfügbaren Quellen u.a. Marktstammdatenregister (MaStR), Energieversorgungsunternehmen (EVU), Amtliches Liegenschaftskatasterinformationssystem (ALKIS), Zensus-Daten, Energieatlas NRW und Geoportal NRW, die für die Wärmeplanung relevant sind. Von besonderer Relevanz sind die Themengruppen für *Gas- und Wärmeverbräuche*, *Gebäudedaten* sowie *Wärmeerzeuger*. Dies impliziert jedoch nicht, dass die anderen Themengruppen eine geringere Relevanz für zukünftige Datenanalyse haben. Daher sind in Tabelle 4.6 geeignete (offene) Datenquellen/Datenlieferanten aufgelistet aus denen die identifizierten Merkmale gewonnen werden können.

Tabelle 4.5: Merkmale zur Charakterisierung der Gebäude für die automatisierten Energiebedarfsbestimmung (in Anlehnung an Wiethe et al. [74]).

Kategorie	**Merkmale**	**Werte**
Verschiedenes	Keller vorhanden	Ja, nein
	Baujahr des Gebäudes	Jahr
	Gebäudealterklasse*	Sieben Gebäudealtersklassen
	Gebäudetyp	Einzelstehend, angebaut etc.
	Wohnfläche	m^2
	Regionale Zugehörigkeit	Postleitzahl
	Gasanschluss vorhanden	Ja, nein
	Bewohner	Personenzahl
Gebäudegeomtrie	Orientierung	degree
	Fenster-zu-Wand Verhältnis	%
	Relative Kompaktheit	-
	A/V-Verhältnis	-
	Gebäudehöhe	m
Wandisolierung	Baujahr der Außenwand	Jahr
	Dicke der Außenwandisolierung	cm
	Vorhandensein einer Außenwandisolierung	Ja, nein
Heizungssystem	Art des Energieträgers	Öl, Gas, Fernwärme, Wärmepumpe etc.
	Baujahr des Kessels	Jahr
	Heizungseffizienzwert	%
Dach	Vorhandensein einer Dachisolierung	Unbekannt/keine, teilweise, voll
	Material der Dachdeckung	Material und U-Wert
Fenster	Art der Fensterverglasung	Einfach, Doppel, Dreifachverglasung, etc.
	Baujahr des Fensters	Jahr
Energieverbrauch	Gas- und Stromverbrauch über ein Jahr	kWh
Energieeffizienz des Gebäudes	Witterungsbereinigte Energieeffizienz des Gebäudes	kWh/m^2a

*Lt. dem deutschen Mikrozensus 2011, einer statistischen Erhebung, die den Haushalts- und Gebäudebestand des Statistischen Bundesamtes Deutschland (2011) repräsentiert.

Tabelle 4.6: Angepasste Auflistung von geeigneten Datenquellen zu Gebäuden, Verbräuchen sowie Wärmeerzeuger [75]

Datenquelle	**Zugang**	**Link/Lieferant**	**Anmerkung**
Zensus2022	öffentlich	Zensus2022 Datenbank	Quelle für Gebäudealter, Energieträger durch synthetische VAE Profile [76]
ALKIS	auf Anfrage	Liegenschaftskataster	Gebäudegrundrisse und Nutzung
Geoportal NRW	öffentlich	Geoportal	LoD1/2 Daten: Auswertung zur Dachform. Alternativ aus dem DOM
Energieatlas NRW	öffentlich	Wärmekataster	
OpenGeodata.NRW	öffentlich	Wärmebedarfsmodell	aktualisiertes Raumwärmebedarfsmodell NRW
Schornsteinfeger	auf Anfrage	Schornsteinfeger	Technische Angaben zur Heizung. Anfragen unter `kehrdaten@energy4climate.nrw`
MaStR	öffentlich	Markstammdaten	Angaben zu Erzeugungs- und Speicheranlagen (Gas/Strom)
EVU	auf Anfrage	auf Kommunalebene	adressbezogene Gas- und Stromverbräuche
Energieausweise WG/NWG	auf Anfrage	dena	Hausverwaltungen, Vermieter, (kommunale) Wohnungsgenossenschaften
Energieausweise (kommunale)	auf Anfrage	dena	Amt für Gebäudemanagement
OSM	öffentlich	OpenStreetMap	Gebäudefunktion und Höhe
GoogleStreetView	öffentlich	GoogleStreetView	Quelle für Fassadenauswertung oder alternativ mit Schrägluftbildern [37]
Erdbeobachtung	öffentlich	EnMAP Satellit	Auswertung des Dachmaterials durch Hyperspektraldaten [47]

5 Diskussion

Die erarbeitete Methodik basiert auf dem Einsatz und Entwicklung eines XGB-Modells zur Vorhersage des Energiebedarfs von WG. Die Wahl dieser Methode begründet sich durch ihre Effizienz in der Verarbeitung großer Datenmengen sowie ihre Fähigkeit, komplexe nichtlineare Beziehungen zu modellieren. Diese Eigenschaften sind entscheidend, da die Genauigkeit der Energiebedarfsprognose stark von der Qualität und Komplexität der verarbeiteten Daten abhängt.

Der extrahierte Gebäudedatensatz weist eine ungleiche Verteilung der Energielabel auf, wie in Abbildung 3.3 dargestellt. Gebäude mit dem Energielabel C und D stellen dabei die Mehrheitsklassen dar und repräsentieren mehr als 80% der Gebäude. Bei der Anwendung der Klassifizierung strebt beispielsweise der RF eine Minimierung der Gesamtfehlerrate an und neigt bei unausgewogenen Lernproblemen dazu, die Vorhersagegenauigkeit der Mehrheitsklasse überzubewerten [77]. Um dieser Tendenz entgegenzuwirken, wird in der Praxis häufig eine Kombination aus Unter- und Überauswahl der Mehrheitsklasse mit der *Synthetic Minority Oversampling* (SMOTE)-Methode verwendet [78]. Unter Verwendung dieser Methode könnte die Sensitivität um 10% erhöht werden, ohne dass es zu weiteren Einbußen bei der Klassifizierungsleistung kommt [19]. Wie in verschiedenen Arbeiten empfohlen, wurde das Resampling innerhalb der Kreuzvalidierungsschleife [79] unter Verwendung der SMOTE-Implementierung aus dem Python-Paket *imbalanced-learn* [80] durchgeführt.

Der Einsatz von maschinellen Lernmethoden erfordert eine hinreichende Menge an relevanten Daten sowie geeignete, alternative Datenquellen, falls keine direkten Daten verfügbar sind. Hinsichtlich der Komplexität, der Übertragbarkeit und der in EA enthaltenen Informationen, stellt dies eine besondere Herausforderung dar. Daher werden im Folgenden Vorschläge für die Akquise alternativer Datenquellen aufgezählt. Dachtyp und -material: Eine Möglichkeit, dies zu tun, ist die Verwendung von Luftbildern, um die Form und das Material von Dächern zu klassifizieren. Es gibt bereits Untersuchungen den Dachtyp zu klassifizieren. Dadurch kann das Dachmaterial einen Teil der Informationen ersetzen, die in den Beschreibungen der Gebäude enthalten sind [37, 81]. Infrarotbilder: Bei Vorhandensein einer ausreichenden Auflösung könnten thermische Infrarotsignale, die in Satellitenbildern erfasst werden, als Maß für die vom Dach entweichende Wärme verwendet werden [82]. Wandtyp und -material: Wenn eine geeignete Quelle von Bildern auf Straßenebene zur Verfügung steht, kann das Material der Wände und möglicherweise auch Gebäude mit einem höheren Anteil an Fenstern (eine Quelle für Wärmeverluste)

identifiziert werden [83, 84]. Mit ähnlichen Luftbildern können auch Gebäude mit Sonnenkollektoren klassifiziert werden [42]. Es wäre interessant, sowohl Luftbilder als auch Bilder von Immobilien auf Straßenebene in ein neuronales Netz einzuspeisen, das anhand der EA-Daten trainiert wurde, und zu testen, ob es möglich ist, die Energieeffizienzwerte direkt aus den Bildern zu lernen oder die EA mit LiDAR-Daten zu kombinieren [46]. Weiterhin werden drei Methoden zur Vorverarbeitung der Daten und zur Erstellung von Merkmalssätzen getestet. Ein weiterer Ansatz wäre die Anwendung von Natural Language Processing (NLP) auf die kategorialen Felder, die Individualtext zur Beschreibung der Gebäude enthalten. Die beschreibenden Merkmale zum Dach, Wand, Boden und Fenster der Gebäude ("..._DESCRIPTION") können mit Hilfe von *CatBoost* [64] über Tokens beschrieben werden. Zukünftige Datensätze könnten die Fähigkeit des Modells, die Energieeffizienz oder das Label vorherzusagen, erheblich verbessern. Insbesondere ein Trainingsatensatz, der Informationen über das Alter des Gebäudes, Sanierungszyklus, den Energieverbrauch oder den Isolierungsstandard enthält, wäre ein erheblicher Beitrag.

6 Schlussfolgerungen und Ausblick

Eine Implementierung von zentralen EA-Registern ist in den EU-Mitgliedstaaten noch nicht flächendeckend erfolgt 2.1.3. Daraus resultieren signifikante Unterschiede in Qualität, Glaubwürdigkeit und Nutzen dieser EA zwischen den Staaten. Es ist daher wichtig, die Einführung der EA-Systeme auf nationaler Ebene weiter zu fördern und Richtlinien dafür zu etablieren. Auf dieser Ebene finden bereits in Deutschland Bestrebungen statt, um einen besseren Datenaustausch für den öffentlichen und wissenschaftlichen Bereich zu gewährleisten [85]. Eine Chance zur Verbesserung besteht durch die effektive Implementierung der überarbeiteten EPBD (Europäische Richtlinie über die Gesamtenergieeffizienz von Gebäuden, 2010/91/EU). Dies könnte beispielsweise durch die Einrichtung eines unabhängigen Kontrollsystems für EA und die Einführung von Strafen bei Nichtbefolgung erreicht werden. Zusätzlich ist es notwendig, weitere Maßnahmen zur Qualitätssicherung, vor allem in der Anfangsphase des Zertifizierungsprozesses, zu ergreifen. Mehrere Länder haben bereits Verfahren zur Überprüfung der Plausibilität der EA-Daten in der Berechnungssoftware und/oder in den EA-Registern eingeführt. Ein häufiges Problem, das die Qualität der EA beeinträchtigt, sind Fehler in den Eingabedaten. In Ungarn und Irland beispielsweise wird eine erste Überprüfung der EA-Daten vorgenommen, bevor das Zertifikat offiziell ausgestellt wird. Es gibt einen klaren Bedarf an Richtlinien zur Schaffung zentralisierter Register für EA, welche nicht nur das unabhängige Kontrollsystem unterstützen, sondern auch als wichtiges Werkzeug zur Erfassung und Überwachung des nationalen Gebäudebestands dienen sollen. In diesem Zusammenhang sollte die Europäische Kommission zusätzliche Empfehlungen abgeben und den Austausch von Best Practices für die Entwicklung funktionaler, vernetzter und automatisierter Datenbanken fördern. Es ist ebenfalls wichtig, die effiziente Nutzung von EA-Daten zu verstärken. Ein leistungsfähiges Monitoring oder Screening, ergänzt durch eine EA-Datenbank, kann eine sofort verfügbare Informationsquelle über den Gebäudebestand bieten. In ganz Europa zeigen immer mehr Praxisansätze den zusätzlichen Nutzen von EA-Daten auf, sowohl für die Politikgestaltung beispielsweise zur Information über relevante Renovierungsstrategien, als auch für Überwachungszwecke und für Markt- sowie Forschungsanalysen. Ein Beispiel hierfür ist Bulgarien, wo das EA-Register als Grundlage für die nationale Renovierungsstrategie gemäß Artikel 4 der Energieeffizienzrichtlinie (EED) genutzt wurde [86]. Ein anderes Beispiel ist Dänemark wo EA für die nationale Wärmewendestrategie genutzt wird [87]. Die technische Integration der Gebäudedaten aus bspw. ENOB:dataNWG mit einer EA-Datenbank ist grundsätzlich möglich. Die Nutzung spezifischer Daten aus EA

könnte die Zuverlässigkeit der durch Stichprobenhochrechnungen gewonnenen Gebäudedaten verbessern. Langfristig könnte ein umfassendes Gebäuderegister, das energetische Merkmale aller Gebäude berücksichtigt, andere Methoden überflüssig machen. Diese Vorgehensweise findet sich auch in der Praxis anderer europäischer Staaten wieder, wo EA-Daten standardmäßig bei Neubau, Kauf, Verkauf und größeren Sanierungen erhoben werden. Die Sammlung einer ausreichend großen und repräsentativen Datenmenge erfordert jedoch Zeit und zusätzliche Daten, um die Repräsentativität gewährleisten zu können. Die Anwendung von Fernerkundungsmethoden ermöglicht bereits heute die flächendeckende Bestimmung von Lage, Adresse, Anzahl und Größe von Gebäuden. Dies kann insbesondere durch eine Kombination mit den Geoinformationen der Vermessungsbehörden erfolgen. Die Ermittlung von Nutzungsklassen auf Einzelgebäudeebene kann in naher Zukunft mit einer ausreichenden Datenqualität für Hochrechnungen automatisiert erfolgen. Des Weiteren könnten FED-Daten zur Ermittlung repräsentativer Stichproben von Gebäudenutzungen und Gebäudegrößen herangezogen werden, was Verfahren wie das Screening im ENOB:dataNWG vereinfachen könnte. Die Zuweisung von Baujahren sowie die automatisierte Auswertung von Gebäudemerkmalen durch die optische Analyse von Fassadenaufnahmen befinden sich noch im experimentellen Stadium, sind jedoch für eine feinere Clusterung der Gebäude hinsichtlich Sanierungsansätzen von essentieller Bedeutung. In der Zukunft könnte auch der Gebäudetomograph *Gtom* [49] des DLR zum Einsatz kommen, um eine berührungslose energetische Analyse von Gebäudehüllen durchzuführen, Schwachstellen zu identifizieren und hohe Energieverbräuche zu ermitteln.

Zusammenfassung der wichtigsten Erkenntnisse

Diese Arbeit leistet einen Beitrag zur Forschung im Bereich der Energiebedarfsbestimmung von WG durch die Anwendung von maschinellen Lernmethoden zur Identifizierung relevanter Merkmale, die über FED akquiriert werden können. Die Methode bietet eine Reihe von Vorteilen gegenüber den herkömmlichen Methoden, wie z.B. eine höhere Genauigkeit, eine geringere Daten- und Expertenanforderung, eine größere Flexibilität und eine bessere Anpassungsfähigkeit. Arbeit demonstriert das Potential datengetriebener Ansätze, nicht nur die Genauigkeit von Energiebedarfsprognosen zu verbessern, sondern auch Entscheidungsträgern in der kommunalen Wärmeplanung praktisch anwendbare Werkzeuge zur Verfügung zu stellen. Die Ergebnisse tragen dazu bei, zielgerichtete Energieeffizienzmaßnahmen zu entwickeln, die auf einer soliden und datenbasierten Grundlage basieren. Darüber hinaus bietet die erfolgreiche Integration von maschinellem Lernen in die Analyse von Energieeffizienz einen vielversprechenden Ansatz für zukünftige Forschungen und praktische Anwendungen im Bereich der Gebäudeenergiebewertung. Durch die Nutzung weiterer Datenpunkte und Einbindung relevanter Parameter, wie Gebäudetyp, Baujahr, Wohnfläche und geografische Lage konnte ein robustes Modell mit einem

Bestimmtheitsmaß R^2 von 0,84 erstellt werden. Gegenüber den Basismodellen stellt dies eine signifikante Steigerung dar. Durch Einbindung des Energieverbrauchs, konnte eine weitere Steigerung von R_2 0,94 erreicht werden, wodurch das Modell den Großteil der Varianz bei geringem Fehler erklären kann. Die Methode hat jedoch auch einige Einschränkungen und Herausforderungen, wie z.B. die Datenqualität, die Datenverfügbarkeit, die Modellauswahl oder die Modellinterpretation. Die Generalisierbarkeit der Modelle auf andere Regionen (Deutschland) oder Gebäudetypen, die Abhängigkeit von der Qualität und Verfügbarkeit der Daten sowie die Komplexität der Modelle stellen Herausforderungen dar, die in zukünftigen Arbeiten adressiert werden müssen.

Potenziale für aktuelle/zukünftige Anwendungsfelder

Grundsätzlich kann die Methodik aus zwei Sichtweisen betrachtet werden. Aus der wissenschaftlichen Perspektive bietet die energetische Bedarfsabschätzung von WG mit Hilfe von datengetriebener Methoden ein großes Potenzial für mehrere Handlungsfelder. Im Rahmen der kWP sind die Kommunen gesetzlich dazu verpflichtet die erforderlichen Daten für die Bestands -und Potenzialanalyse abzurufen [4]. Auf dieser Grundlage könnte eine Datenbank mit den fehlenden sowie relevanten Gebäudemerkmalen angereichert werden, die den Transformationsprozess erheblich beschleunigen kann. Nicht nur die Kommune, auch die Privatperson kann durch diese Datenabfrage einen Einblick in die aktuelle und potenzielle, energetische Situation bekommen. Eine ähnliche Methodik verfolgte auch Wederhake et al. [88], indem Bewohner relevante Merkmale über einen Abfragebogen bereitstellen könnten. Dies kann individuelle Sanierungsanreize auslösen, indem entsprechende Empfehlungen in Verbindung mit Förderporgrammen auf Basis des aktuellen Gebäudezustands gegeben werden. Das resultierende Modell ermöglicht die Ermittlung der Merkmale, die am meisten zur Energieeffizienz beitragen, und erlaubt somit gezielte Eingriffe. So können beispielsweise Bemühungen zur Verbesserung der Isolierung priorisiert werden, wenn Wand- und Dachisolierung starke, positive Auswirkungen haben. Die Erkenntnisse können den politischen Entscheidungsträgern Aufschluss darüber geben, wo Vorschriften oder Anreize zur Verbesserung der Energieeffizienz ansetzen sollten. So könnten zum Beispiel Anreize für die Verbesserung der Isolierung oder die Installation effizienterer Heizsysteme in Betracht gezogen werden. Aus der technisch, wirtschaftlichen Perspektive kann das Wissen um die Merkmale, die zu einer höheren Energieeffizienz führen, die Grundlage für Planungsentscheidungen bilden, um von Anfang an energieeffizientere Gebäude zu planen oder den Sanierungsfahrplan vorab einzuschätzen. Dies gilt bspw. für Architekten, Hauseigentümer, Planer und Energieberater. Auf dieser Grundlage könnten großräumig automatisierte EA ausgewiesen werden und einen Sanierungsanreiz bei Hauseigentümern auslösen. Diese Aussage wird auch durch Hettinga et al. [31] und Wenninger et al. [74] gestützt. Ein weiterer Punkt ist die Kombination von ML mit Energiesystemmodellierung bzw. Optimierungsalgorithmen. ML ermöglicht die

Extraktion von Mustern sowie die Prognose zukünftiger Systemzustände aus historischen Datensätzen, was für die Planung und Steuerung von Energiesystemen bspw. Wärmenetze von essentieller Bedeutung ist. Gleichzeitig bieten Optimierungsalgorithmen eine Schnittstelle für die Lösung von linearen, nichtlinearen und gemischt-ganzzahligen Programmieraufgaben, die für die strategische Planung von Energieerzeugung, -verteilung und -verbrauch erforderlich sind.

Limitationen und offene Fragestellungen

Aufgrund eines fehlenden Trainingsdatensatzes für Deutschland, konnte keine direkte Übertragung erfolgen. Die identifizierten Merkmale aus dem Datensatz für England stimmen jedoch mit den relevanten Merkmalen aus verwandten Arbeiten überein [15, 54]. Des Weiteren wurden nur Ausweise von WG betrachtet. Die Datenbank beinhaltet noch weitere Daten zu NWG sowie entsprechende Maßnahmenempfehlungen, die den energetischen Standard verbessern. Eine Kombination oder separate Betrachtung dieser Ausweisdaten, kann zu weiteren Erkenntnissen führen. Weiterhin sollte ein weiterer Datensatz mit empfohlenen Gebäudemerkmalen getestet werden, die auf Basis von FED akquiriert werden können. In dieser Arbeit wurden anfänglich mehrere Modelle verglichen. Aus diesem Vergleich wurde ein Modell (XGB) genauer betrachtet und validert. Das untersuchte Modell deckte sich zwar mit den Empfehlungen aus der Literatur, jedoch würde ein paralleler Vergleich aus mehreren Modellarchitekturen Aufschluss über die Relevanz der Merkmale geben. Weiterhin kann durch den Vergleich eine Abschätzung erfolgen, welches Modell hinsichtlich Genauigkeit und rechnerischem Aufwand das geeignete Model darstellt. Die Hinzunahme von NWG würde die Dimensionalität des Datensatzes erhöhen und Interpretierbarkeit erschweren. Daher hätte die gesonderte Betrachtung eine praktische Relevanz, um weitere Einblicke zu erhalten. Die Anzahl der Bewohner sowie das Verhalten wäre auch noch ein interessanter Aspekt, da dies nicht innerhalb des Datensatzes aufgeführt ist. Ein weiterer Punkt ist die Auswirkung der Hyperparameteroptimierung auf die Qualtität der Vorhersage. Durch *GridSearchCV* und ähnlicher Validierungsmethoden können geeignete Parameter für die Regulierung des Modells herausgefunden werden. Duch die Parametrierung kann die Überanpassung und damit die Verallgemeinerungsleistung verbessert werden. Bei genauer Betrachtung dieser Parameter könnte die Auswirkung ebenfalls mit *ShapValues* analysiert werden. Dadurch können die Parameter explizit identifiziert werden, die einen großen Einfluss auf den Vorhersagewert haben. Für zukünftige Arbeiten mit Fokus auf Labelklassifikation wäre die Betrachtung eines ausbalancierten Datensatzes interessant, da nur ein geringer Anteil von A- und G-Label vorhanden waren. Hierfür kann die in Python implementierte Bibliothek *Imbalanced-learn* verwendet werden. Dies ist eine quelloffene, MIT-lizensierte Bibliothek, die auf scikit-learn aufbaut und Methoden für die Klassifizierung von unausgewogenen Klassen bereitstellt [89].

Literaturverzeichnis

[1] H. Visscher u. a., „Improved governance for energy efficiency in housing," *Building Research and Information*, 2016. DOI: 10.1080/09613218.2016.1180808.

[2] EPBD. „Energy performance of buildings directive." (2023), Adresse: https://energy.ec.europa.eu/topics/energy-efficiency/energy-efficient-buildings/energy-performance-buildings-directive_en.

[3] *GEG - Gesetz zur Einsparung von Energie und zur Nutzung erneuerbarer Energien zur Wärme- und Kälteerzeugung in Gebäuden*. Adresse: https://www.gesetze-im-internet.de/geg/ (besucht am 03. 04. 2024).

[4] *WPG - Gesetz für die Wärmeplanung und zur Dekarbonisierung der Wärmenetze*. Adresse: https://www.gesetze-im-internet.de/wpg/BJNR18A0B0023.html (besucht am 03. 04. 2024).

[5] S. Becker. „Metastudie zur Verbesserung der Datengrundlage im Gebäudebereich," Die Bundesregierung informiert | Startseite. (6. Apr. 2022), Adresse: //www.bundesregierung.de/breg-de/service/publikationen/metastudie-zur-verbesserung-der-datengrundlage-im-gebaeudebereich-2024482 (besucht am 14. 12. 2023).

[6] „IEE Project TABULA." (2012), Adresse: https://episcope.eu/iee-project/tabula/ (besucht am 05. 03. 2024).

[7] A. A. A. Gassar und S. H. Cha, „Energy prediction techniques for large-scale buildings towards a sustainable built environment: A review," *Energy and Buildings*, Jg. 224, S. 110 238, 1. Okt. 2020, ISSN: 0378-7788. DOI: 10.1016/j.enbuild.2020.110238. Adresse: https://www.sciencedirect.com/science/article/pii/S0378778820301237 (besucht am 13. 11. 2023).

[8] I. Staffell, S. Pfenninger und N. Johnson, „A global model of hourly space heating and cooling demand at multiple spatial scales," *Nat Energy*, S. 1–17, 14. Sep. 2023, Publisher: Nature Publishing Group, ISSN: 2058-7546. DOI: 10.1038/s41560-023-01341-5. Adresse: https://www.nature.com/articles/s41560-023-01341-5 (besucht am 14. 12. 2023).

[9] P. d. Wilde, „The gap between predicted and measured energy performance of buildings: A framework for investigation," *Automation in Construction*, 2014. DOI: 10.1016/j.autcon.2014.02.009.

[10] C. Robinson u. a., „Machine learning approaches for estimating commercial building energy consumption," *Applied Energy*, 2017. DOI: 10.1016/j.apenergy.2017.09.060.

[11] N. Fumo und M. A. Rafe Biswas, „Regression analysis for prediction of residential energy consumption," *Renewable and Sustainable Energy Reviews*, Jg. 47, S. 332–343, 1. Juli 2015, ISSN: 1364-0321. DOI: 10.1016/j.rser.2015.03.035. Adresse: https://www.sciencedirect.com/science/article/pii/S1364032115001884 (besucht am 16. 12. 2023).

[12] P. W. Westermann, „Advancing surrogate modelling for sustainable building design.," Diss., University of Victoria, 2020. Adresse: http://oatd.org/oatd/record?record=%22handle%5C%3A1828%2F12127%22 (besucht am 14. 12. 2023).

[13] H. Zhang, H. Feng, K. Hewage und M. Arashpour, „Artificial neural network for predicting building energy performance: A surrogate energy retrofits decision support framework," *Buildings*, Jg. 12, Nr. 6, S. 829, Juni 2022, Number: 6 Publisher: Multidisciplinary Digital Publishing Institute, ISSN: 2075-5309. DOI: 10.3390/buildings12060829. Adresse: https://www.mdpi.com/2075-5309/12/6/829 (besucht am 14. 12. 2023).

[14] T. Catalina, V. Iordache und B. Caracaleanu, „Multiple regression model for fast prediction of the heating energy demand," *Energy and Buildings*, Jg. 57, S. 302–312, 1. Feb. 2013, ISSN: 0378-7788. DOI: 10.1016/j.enbuild.2012.11.010. Adresse: https://www.sciencedirect.com/science/article/pii/S0378778812005993 (besucht am 16. 12. 2023).

[15] S. Wenninger und C. Wiethe, „Benchmarking energy quantification methods to predict heating energy performance of residential buildings in germany," *Business & Information Systems Engineering*, Jg. 63, Nr. 3, S. 223–242, 1. Juni 2021, ISSN: 1867-0202. DOI: 10.1007/s12599-021-00691-2. Adresse: https://doi.org/10.1007/s12599-021-00691-2 (besucht am 16. 11. 2023).

[16] *Hotmaps Project - The open source mapping and planning tool for heating and cooling*. Adresse: https://www.hotmaps-project.eu/ (besucht am 04. 04. 2024).

[17] O. Ruhnau, L. Hirth und A. Praktiknjo, „Time series of heat demand and heat pump efficiency for energy system modeling," *Sci Data*, Jg. 6, Nr. 1, S. 189, 1. Okt. 2019, ISSN: 2052-4463. DOI: 10.1038/s41597-019-0199-y.

[18] R. Nouvel, M. Zirak, V. Coors und U. Eicker, „The influence of data quality on urban heating demand modeling using 3D city models," *Computers, Environment and Urban Systems*, Jg. 64, S. 68–80, 1. Juli 2017, ISSN: 0198-9715. DOI: 10.1016/j.compenvurbsys.2016.12.005. Adresse: https://www.sciencedirect.com/science/article/pii/S0198971516304306 (besucht am 14. 12. 2023).

[19] O. M. Garbasevschi u. a., „Spatial factors influencing building age prediction and implications for urban residential energy modelling,“ *Computers, Environment and Urban Systems*, Jg. 88, S. 101 637, 1. Juli 2021, ISSN: 0198-9715. DOI: 10.1016/j.compenvurbsys.2021.101637. Adresse: https://www.sciencedirect.com/science/article/pii/S0198971521000442 (besucht am 14. 12. 2023).

[20] T. Walter und M. D. Sohn, „A regression-based approach to estimating retrofit savings using the Building Performance Database,“ *Applied Energy*, Jg. 179, S. 996–1005, 1. Okt. 2016, ISSN: 0306-2619. DOI: 10.1016/j.apenergy.2016.07.087. Adresse: https://www.sciencedirect.com/science/article/pii/S0306261916310297 (besucht am 29. 04. 2024).

[21] *GEG-Registrierstelle*, DIBt. Adresse: https://www.dibt.de/de/wir-bieten/geg-registrierstelle (besucht am 21. 04. 2024).

[22] „ENOB DataNWG: Forschungsdatenbank.“ (2020), Adresse: https://www.datanwg.de/forschungsdatenbank/ (besucht am 30. 04. 2024).

[23] „IWU wohngebaeudedaten2016.“ (2016), Adresse: https://wohngebaeudedaten2016.iwu.de/ (besucht am 30. 04. 2024).

[24] E. Örtl, „Hintergrundbericht: Wohnen und Sanieren,“ Umweltbundesamt, 23. Mai 2019. Adresse: https://www.umweltbundesamt.de/publikationen/hintergrundbericht-wohnen-sanieren (besucht am 16. 11. 2023).

[25] Y. Li, S. Kubicki, A. Guerriero und Y. Rezgui, „Review of building energy performance certification schemes towards future improvement,“ *Renewable and Sustainable Energy Reviews*, Jg. 113, S. 109 244, 1. Okt. 2019, ISSN: 1364-0321. DOI: 10.1016/j.rser.2019.109244. Adresse: https://www.sciencedirect.com/science/article/pii/S1364032119304447 (besucht am 27. 11. 2023).

[26] O. Pasichnyi, J. Wallin, F. Levihn, H. Shahrokni und O. Kordas, „Energy performance certificates — New opportunities for data-enabled urban energy policy instruments?“ *Energy Policy*, Jg. 127, S. 486–499, 1. Apr. 2019, ISSN: 0301-4215. DOI: 10.1016/j.enpol.2018.11.051. Adresse: https://www.sciencedirect.com/science/article/pii/S0301421518307894 (besucht am 08. 01. 2024).

[27] U. Ali u. a., „Urban building energy performance prediction and retrofit analysis using data-driven machine learning approach,“ *Energy and Buildings*, Jg. 303, S. 113 768, 15. Jan. 2024, ISSN: 0378-7788. DOI: 10.1016/j.enbuild.2023.113768. Adresse: https://www.sciencedirect.com/science/article/pii/S0378778823009982 (besucht am 17. 12. 2023).

[28] K. Amasyali und N. M. El-Gohary, „A review of data-driven building energy consumption prediction studies,“ *Renewable and Sustainable Energy Reviews*, Jg. 81, S. 1192–1205, 1. Jan. 2018, ISSN: 1364-0321. DOI: 10.1016/j.rser.2017.04.095. (besucht am 26. 01. 2024).

[29] *DIN V 18599: Energetische Gebäudebilanzierung.*

[30] U. Bigalke, H. Marcinek, M. Grafe, M. Großklos, T. Loga und R. Born, „Auswertung von Verbrauchskennwerten energieeffizienter Wohngebaeude," Studie, Juni 2016. Adresse: https://www.kompetenzzentrum-contracting.de/fileadmin/dena/Dokumente/Pdf/9164_dena-Studie._Auswertung_von_Verbrauchskennwerten_energieeffizienter_Wohngebaeude.pdf (besucht am 01. 05. 2024).

[31] S. Hettinga, R. van 't Veer und J. Boter, „Large scale energy labelling with models: The EU TABULA model versus machine learning with open data," *Energy*, Jg. 264, S. 126 175, 1. Feb. 2023, ISSN: 0360-5442. DOI: 10.1016/j.energy.2022.126175. Adresse: https://www.sciencedirect.com/science/article/pii/S0360544222030614 (besucht am 14. 12. 2023).

[32] T. Ahmad, H. Chen, Y. Guo und J. Wang, „A comprehensive overview on the data driven and large scale based approaches for forecasting of building energy demand: A review," *Energy and Buildings*, Jg. 165, S. 301–320, 15. Apr. 2018, ISSN: 0378-7788. DOI: 10.1016/j.enbuild.2018.01.017. Adresse: https://www.sciencedirect.com/science/article/pii/S0378778817329225 (besucht am 16. 12. 2023).

[33] F. Khayatian, L. Sarto und G. Dall'O', „Application of neural networks for evaluating energy performance certificates of residential buildings," *Energy and Buildings*, Jg. 125, S. 45–54, 1. Aug. 2016, ISSN: 0378-7788. DOI: 10.1016/j.enbuild.2016.04.067. Adresse: https://www.sciencedirect.com/science/article/pii/S0378778816303322 (besucht am 18. 12. 2023).

[34] P. Gorzalka u. a., „Collecting data for urban building energy modelling by remote sensing and machine learning," Jg. 17, S. 1139–1146, 2021. DOI: 10.26868/25222708.2021.30184. Adresse: https://publications.ibpsa.org/conference/paper/?id=bs2021_30184 (besucht am 07. 03. 2024).

[35] M. Wurm, A. Droin, T. Stark, C. Geiß, W. Sulzer und H. Taubenböck, „Deep Learning-Based Generation of Building Stock Data from Remote Sensing for Urban Heat Demand Modeling," *International Journal of Geo-Information*, Jg. 10, 12. Jan. 2021. DOI: 10.3390/ijgi10010023.

[36] T. Tooke und N. Coops, „A review of remote sensing for urban energy system management and planning," Journal Abbreviation: Joint Urban Remote Sensing Event 2013, JURSE 2013 Pages: 170 Publication Title: Joint Urban Remote Sensing Event 2013, JURSE 2013, 1. Apr. 2013, ISBN: 978-1-4799-0213-2. DOI: 10.1109/JURSE#.2013.6550692.

[37] K. Mayer, L. Haas, T. Huang, J. Bernabé-Moreno, R. Rajagopal und M. Fischer, „Estimating building energy efficiency from street view imagery, aerial imagery, and land surface temperature data," *Applied Energy*, Jg. 333, S. 120 542, 1. März

2023, ISSN: 0306-2619. DOI: 10.1016/j.apenergy.2022.120542. Adresse: https://www.sciencedirect.com/science/article/pii/S0306261922017998 (besucht am 26. 02. 2024).

[38] T. R. Tooke, N. C. Coops und J. Webster, „Predicting building ages from LiDAR data with random forests for building energy modeling," *Energy and Buildings*, Jg. 68, S. 603–610, 1. Jan. 2014, ISSN: 0378-7788. DOI: 10.1016/j.enbuild.2013.10.004. Adresse: https://www.sciencedirect.com/science/article/pii/S0378778813006506 (besucht am 07. 03. 2024).

[39] F. Nachtigall, N. Milojevic-Dupont, F. Wagner und F. Creutzig, „Predicting building age from urban form at large scale," *Computers, Environment and Urban Systems*, Jg. 105, S. 102 010, 1. Okt. 2023, ISSN: 0198-9715. DOI: 10.1016/j.compenvurbsys.2023.102010. Adresse: https://www.sciencedirect.com/science/article/pii/S019897152300073X (besucht am 10. 11. 2023).

[40] L. A. Blanco Bohorquez, M. Aditya, B. Schiricke und B. Hoffschmidt, „Classification of Building Properties from the German Census Data for Energy Analysis Purposes," Juni 2023. Adresse: https://elib.dlr.de/199041/ (besucht am 07. 03. 2024).

[41] P. S. Prakash und P. R. Vyas, „Remote sensing using drone and machine learning for computation of rooftop solar energy potential," in *2023 IEEE Applied Sensing Conference (APSCON)*, Jan. 2023, S. 1–3. DOI: 10.1109/APSCON56343.2023.10101182. Adresse: https://ieeexplore.ieee.org/document/10101182 (besucht am 17. 04. 2024).

[42] X. Huang, K. Hayashi, T. Matsumoto, L. Tao, Y. Huang und Y. Tomino, „Estimation of rooftop solar power potential by comparing solar radiation data and remote sensing data—a case study in aichi, japan," *Remote Sensing*, Jg. 14, Nr. 7, S. 1742, Jan. 2022, Number: 7 Publisher: Multidisciplinary Digital Publishing Institute, ISSN: 2072-4292. DOI: 10.3390/rs14071742. Adresse: https://www.mdpi.com/2072-4292/14/7/1742 (besucht am 17. 04. 2024).

[43] A. Wichmann, A. Agoub und M. Kada, „RoofFN3D: Deep Learning Training Data for 3D Building Reconstruction," *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, Jg. XLII-2, S. 1191–1198, 2018. DOI: 10.5194/isprs-archives-XLII-2-1191-2018. Adresse: https://isprs-archives.copernicus.org/articles/XLII-2/1191/2018/.

[44] I. Dochev u. a., „Calculating urban heat demands: An analysis of two modelling approaches and remote sensing for input data and validation," *Energy and Buildings*, Jg. 226, S. 110 378, 1. Aug. 2020. DOI: 10.1016/j.enbuild.2020.110378.

[45] T. Loga, B. Stein und N. Diefenbach, „TABULA building typologies in 20 European countries—Making energy-related features of residential building stocks comparable," *Energy and Buildings*, Towards an energy efficient European housing stock: monitoring, mapping and modelling retrofitting processes, Jg. 132, S. 4–12, 15. Nov. 2016, ISSN: 0378-7788. DOI: 10.1016/j.enbuild.2016.06.094. Adresse: https://www.sciencedirect.com/science/article/pii/S0378778816305837 (besucht am 16. 11. 2023).

[46] S. Krapf, K. Mayer und M. Fischer, „Points for energy renovation (PointER): A point cloud dataset of a million buildings linked to energy features," *Sci Data*, Jg. 10, Nr. 1, S. 639, 20. Sep. 2023, Number: 1 Publisher: Nature Publishing Group, ISSN: 2052-4463. DOI: 10.1038/s41597-023-02544-x. Adresse: https://www.nature.com/articles/s41597-023-02544-x (besucht am 21. 02. 2024).

[47] T. Esch, S. Starmans, W. Heldens, B. Leutner, J. Zeidler und C. Ji. „Konzeptentwicklung für die Informationsgewinnung zum Gebäudebestand in Deutschland aus Fernerkundungsdaten - G-DAT DE." Num Pages: 120. (30. Juni 2021), Adresse: https://www.bbsr.bund.de/BBSR/DE/forschung/programme/zb/Auftragsforschung/5EnergieKlimaBauen/2018/fernerkundungsdaten/01-start.html (besucht am 30. 04. 2024).

[48] M. Moenks, „Building facade segmentation of oblique aerial images using convolutional neural networks for urban climate modeling," Diss., 1. Feb. 2019.

[49] J. Estevam Schmiedt, M. Peichl, S. Plattner, S. Pless und J. Goettsche, „Kurzzeitmessung: Gtom - energetische analyse von gebäudehüllen," in *Energiewende Bauen: Forschungserkenntnisse von der Komponente bis zum Quartier*, Fraunhofer IRB Verlag, 2020, S. 99–105, ISBN: 978-3-948234-88-1. Adresse: http://www.energiewendebauen.de (besucht am 30. 04. 2024).

[50] „Hands-on machine learning with scikit-learn, keras, and TensorFlow, 2nd edition [book]." ISBN: 9781492032649. (2019), Adresse: https://www.oreilly.com/library/view/hands-on-machine-learning/9781492032632/ (besucht am 03. 12. 2023).

[51] U. Ali, M. Shamsi, F. Alshehri, E. Mangina und J. O'Donnell, „Application Of Intelligent Algorithms For Residential Building Energy Performance Rating," 2020. DOI: 10.26868/25222708.2019.210232. Adresse: https://www.researchgate.net/publication/336349588_Application_Of_Intelligent_Algorithms_For_Residential_Building_Energy_Performance_Rating_Prediction (besucht am 01. 12. 2023).

[52] B. Abhilash, E. Busari, C. Syranidou, J. Linssen und D. Stolten, „Classification of Building Types in Germany: A Data-Driven Modeling Approach," *Data*, Jg. 7, S. 45, 9. Apr. 2022. DOI: 10.3390/data7040045.

[53] B. Yildiz, J. I. Bilbao, B. Yildiz, A. B. Sproul, J. I. Bilbao und A. B. Sproul, „A review and analysis of regression and machine learning models on commercial building electricity load forecasting," *Renewable & Sustainable Energy Reviews*, 2017. DOI: 10.1016/j.rser.2017.02.023.

[54] C. N. Egwim, O. O. Egunjobi, A. Gomes und H. Alaka, „A comparative study on machine learning algorithms for assessing energy efficiency of buildings," in *Machine Learning and Principles and Practice of Knowledge Discovery in Databases*, M. Kamp u. a., Hrsg., Ser. Communications in Computer and Information Science, Cham: Springer International Publishing, 2021, S. 546–566, ISBN: 978-3-030-93733-1. DOI: 10.1007/978-3-030-93733-1_41.

[55] F. Pedregosa u. a., „Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, Jg. 12, S. 2825–2830, 2011.

[56] R. Olu-Ajayi, H. Alaka, I. Sulaimon, F. Sunmola und S. Ajayi, „Machine learning for energy performance prediction at the design stage of buildings," *Energy for Sustainable Development*, Jg. 66, S. 12–25, 1. Feb. 2022, ISSN: 0973-0826. DOI: 10.1016/j.esd.2021.11.002. Adresse: https://www.sciencedirect.com/science/article/pii/S0973082621001307 (besucht am 04. 03. 2024).

[57] L. Zhang und J. Wen, „A systematic feature selection procedure for short-term data-driven building energy forecasting model development," *Energy and Buildings*, Jg. 183, S. 428–442, 15. Jan. 2019, ISSN: 0378-7788. DOI: 10.1016/j.enbuild.2018.11.010. Adresse: https://www.sciencedirect.com/science/article/pii/S0378778818321625 (besucht am 18. 04. 2024).

[58] A. Kusiak, M. Li und Z. Zhang, „A data-driven approach for steam load prediction in buildings," *Applied Energy*, Jg. 87, Nr. 3, S. 925–933, 1. März 2010, ISSN: 0306-2619. DOI: 10.1016/j.apenergy.2009.09.004. Adresse: https://www.sciencedirect.com/science/article/pii/S0306261909003808 (besucht am 18. 04. 2024).

[59] R. Zebari, A. Abdulazeez, D. Zeebaree, D. Zebari und J. Saeed, „A comprehensive review of dimensionality reduction techniques for feature selection and feature extraction," *Journal of Applied Science and Technology Trends*, Jg. 1, Nr. 1, S. 56–70, 15. Mai 2020, Number: 1, ISSN: 2708-0757. DOI: 10.38094/jastt1224. Adresse: https://www.jastt.org/index.php/jasttpath/article/view/24 (besucht am 28. 03. 2024).

[60] A. L. Blum und P. Langley, „Selection of relevant features and examples in machine learning," *Artificial Intelligence*, Relevance, Jg. 97, Nr. 1, S. 245–271, 1. Dez. 1997, ISSN: 0004-3702. DOI: 10.1016/S0004-3702(97)00063-5. Adresse: https://www.sciencedirect.com/science/article/pii/S0004370297000635 (besucht am 14. 12. 2023).

[61] D. Dzyabura und J. Hauser, „Active Machine Learning for Consideration Heuristics," *Mark. Sci.*, Jg. 30, S. 801–819, 2011. DOI: 10.1287/mksc.1110.0660.

[62] D. A. Waterman, „Generalization Learning Techniques for Automating the Learning of Heuristics," *Artif. Intell.*, Jg. 1, S. 121–170, 1970. DOI: 10.1016/0004-3702(70)90004-4.

[63] A. Moez, *PyCaret: An open source, low-code machine learning library in Python*, Version PyCaret version 3.0, Apr. 2020. Adresse: https://pycaret.org (besucht am 20. 11. 2023).

[64] A. V. Dorogush, V. Ershov und A. Gulin, *CatBoost: gradient boosting with categorical features support*, 24. Okt. 2018. DOI: 10.48550/arXiv.1810.11363. arXiv: 1810.11363[cs,stat]. Adresse: http://arxiv.org/abs/1810.11363 (besucht am 26. 02. 2024).

[65] M. Sokolova und G. Lapalme, „A systematic analysis of performance measures for classification tasks," *Information Processing & Management*, Jg. 45, Nr. 4, S. 427–437, 1. Juli 2009, ISSN: 0306-4573. DOI: 10.1016/j.ipm.2009.03.002. Adresse: https://www.sciencedirect.com/science/article/pii/S0306457309000259 (besucht am 30. 03. 2024).

[66] *The Elements of Statistical Learning*. Adresse: https://link.springer.com/book/10.1007/978-0-387-84858-7 (besucht am 08. 04. 2024).

[67] C. M. Bishop, *Pattern recognition and machine learning* (Information science and statistics). New York: Springer, 2006, 738 S., ISBN: 978-0-387-31073-2.

[68] Y. Ding, L. Fan und X. Liu, „Analysis of feature matrix in machine learning algorithms to predict energy consumption of public buildings," *Energy and Buildings*, Jg. 249, S. 111 208, 15. Okt. 2021, ISSN: 0378-7788. DOI: 10.1016/j.enbuild.2021.111208. Adresse: https://www.sciencedirect.com/science/article/pii/S0378778821004928 (besucht am 16. 11. 2023).

[69] S. M. Lundberg, G. G. Erion und S. Lee, „Consistent Individualized Feature Attribution for Tree Ensembles," *CoRR*, Jg. abs/1802.03888, 2018. arXiv: 1802.03888. Adresse: http://arxiv.org/abs/1802.03888.

[70] S. M. Lundberg und S.-I. Lee, „A Unified Approach to Interpreting Model Predictions," in *Advances in Neural Information Processing Systems 30*, I. Guyon u. a., Hrsg., Curran Associates, Inc., 2017, S. 4765–4774. Adresse: http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf.

[71] S. M. Lundberg u. a., „From local explanations to global understanding with explainable AI for trees," *Nature Machine Intelligence*, Jg. 2, Nr. 1, S. 2522–5839, 2020.

[72] M. A. B. Abbass und M. Hamdy, „A generic pipeline for machine learning users in energy and buildings domain," *Energies*, Jg. 14, Nr. 17, S. 5410, Jan. 2021, Number: 17 Publisher: Multidisciplinary Digital Publishing Institute, ISSN: 1996-1073. DOI: 10.3390/en14175410. Adresse: https://www.mdpi.com/1996-1073/14/17/5410 (besucht am 14. 12. 2023).

[73] Z. Yu, F. Haghighat, B. C. M. Fung und H. Yoshino, „A decision tree method for building energy demand modeling," *Energy and Buildings*, Jg. 42, Nr. 10, S. 1637–1646, 1. Okt. 2010, ISSN: 0378-7788. DOI: 10.1016/j.enbuild.2010.04.006. Adresse: https://www.sciencedirect.com/science/article/pii/S0378778810001350 (besucht am 04. 03. 2024).

[74] C. Wiethe und S. Wenninger, „The influence of building energy performance prediction accuracy on retrofit rates," *Energy Policy*, Jg. 177, S. 113 542, 1. Juni 2023, ISSN: 0301-4215. DOI: 10.1016/j.enpol.2023.113542. Adresse: https://www.sciencedirect.com/science/article/pii/S0301421523001271 (besucht am 16. 11. 2023).

[75] Kompetenzzentrum Kommunale Wärmewende (KWW). „KWW-Datenkompass." (5. März 2024), Adresse: https://www.kww-halle.de.

[76] S. S. Borysov, J. Rich und F. C. Pereira, „Scalable population synthesis with deep generative modeling," *Transportation Research Part C: Emerging Technologies*, Jg. 106, S. 73–97, Sep. 2019, ISSN: 0968090X. DOI: 10.1016/j.trc.2019.07.006. arXiv: 1808.06910[cs,stat]. Adresse: http://arxiv.org/abs/1808.06910 (besucht am 01. 05. 2024).

[77] C. Chen, A. Liaw und L. Breiman, „Using random forest to learn imbalanced data," Adresse: https://statistics.berkeley.edu/sites/default/files/tech-reports/666.pdf.

[78] N. V. Chawla, K. W. Bowyer, L. O. Hall und W. P. Kegelmeyer, „SMOTE: Synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research*, Jg. 16, S. 321–357, 1. Juni 2002, ISSN: 1076-9757. DOI: 10.1613/jair.953. Adresse: https://www.jair.org/index.php/jair/article/view/10302 (besucht am 20. 04. 2024).

[79] M. S. Santos, J. P. Soares, P. H. Abreu, H. Araujo und J. Santos, „Cross-Validation for Imbalanced Datasets: Avoiding Overoptimistic and Overfitting Approaches," *IEEE Computational Intelligence Magazine*, Jg. 13, Nr. 4, S. 59–76, Nov. 2018, Conference Name: IEEE Computational Intelligence Magazine, ISSN: 1556-6048. DOI: 10.1109/MCI.2018.2866730. Adresse: https://ieeexplore.ieee.org/document/8492368 (besucht am 20. 04. 2024).

[80] G. Lemaître, F. Nogueira und C. K. Aridas, „Imbalanced-learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning," *Journal of Machine Learning Research*, Jg. 18, Nr. 17, S. 1–5, 2017, ISSN: 1533-7928. Adresse: http://jmlr.org/papers/v18/16-365.html (besucht am 20. 04. 2024).

[81] T. Tooke, M. van der Laan, N. Coops, R. Kellett und A. Christen, *Classification of Residential Building Architectural Typologies using LiDAR*. 11. Apr. 2011, Journal Abbreviation: 2011 Joint Urban Remote Sensing Event, JURSE 2011 - Proceedings Publication Title: 2011 Joint Urban Remote Sensing Event, JURSE 2011 - Proceedings. DOI: 10.1109/JURSE.2011.5764760.

[82] A. Anand und C. Deb, „The potential of remote sensing and GIS in urban building energy modelling," *Energy and Built Environment*, 23. Juli 2023, ISSN: 2666-1233. DOI: 10.1016/j.enbenv.2023.07.008. Adresse: https://www.sciencedirect.com/science/article/pii/S2666123323000685 (besucht am 26. 02. 2024).

[83] T. H. Meles, N. Farrell und J. Curtis, „How well do building energy performance certificates predict heat loss?" *Energy Efficiency*, Jg. 16, Nr. 7, S. 74, 4. Sep. 2023, ISSN: 1570-6478. DOI: 10.1007/s12053-023-10146-0. Adresse: https://doi.org/10.1007/s12053-023-10146-0 (besucht am 21. 12. 2023).

[84] K. S. Atwal, T. Anderson, D. Pfoser und A. Züfle, „Predicting building types using OpenStreetMap," *Sci Rep*, Jg. 12, Nr. 1, S. 19 976, 20. Nov. 2022, Number: 1 Publisher: Nature Publishing Group, ISSN: 2045-2322. DOI: 10.1038/s41598-022-24263-w. Adresse: https://www.nature.com/articles/s41598-022-24263-w (besucht am 30. 12. 2023).

[85] „Fortschritt durch Datennutzung," Bundesregierung, Aug. 2023. Adresse: https://www.bmi.bund.de/SharedDocs/downloads/DE/veroeffentlichungen/2023/datenstrategie.pdf;jsessionid=484F7BDC422138E2984808635342651C.live871?__blob=publicationFile&v=3 (besucht am 14. 12. 2023).

[86] BPIE. „Energy performance certificates across the EU," BPIE - Buildings Performance Institute Europe. (2014), Adresse: https://www.bpie.eu/publication/energy-performance-certificates-across-the-eu/ (besucht am 16. 11. 2023).

[87] „Die dänische Wärmewende," Wärmewende. Section: International. (1. Aug. 2021), Adresse: https://www.waermewende.de/daenischewaermewende/ (besucht am 21. 04. 2024).

[88] L. Wederhake, S. Wenninger, C. Wiethe, G. Fridgen und D. Stirnweiß, „Benchmarking building energy performance: Accuracy by involving occupants in collecting data - A case study in Germany," *Journal of Cleaner Production*, Jg. 379, S. 134 762, 15. Dez. 2022, ISSN: 0959-6526. DOI: 10.1016/j.jclepro.2022.134762. Adresse: https://www.sciencedirect.com/science/article/pii/S0959652622043347 (besucht am 16. 11. 2023).

[89] G. Lemaître, F. Nogueira und C. K. Aridas, „Imbalanced-learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning,“ *Journal of Machine Learning Research*, Jg. 18, Nr. 17, S. 1–5, 2017. Adresse: `http://jmlr.org/papers/v18/16-365.html`.

ANHANG

A Korrelationsanalysen

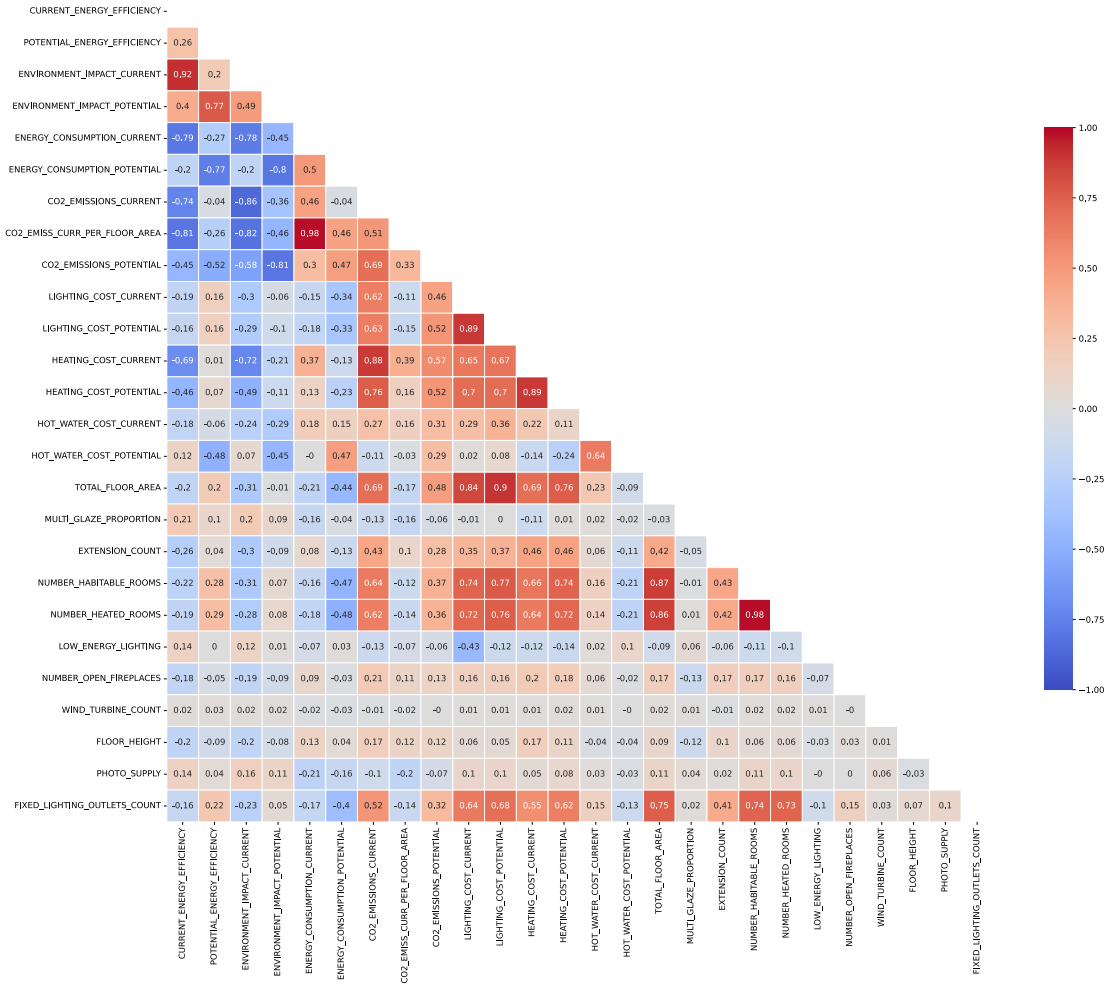

Abbildung A.1: Korrelationsanalyse der numerischen Merkmale

Tabelle A.1: Chi^2-Analyse der kategorischen Merkmale zum Zielmerkmal

ID	Feature	Chi2 Score	p-value
19	ROOF DESCRIPTION	2073.436482	0.000000
0	CURRENT ENERGY RATING	1259.052336	0.000000
12	WINDOWS DESCRIPTION	1155.849145	0.000000
22	MAINHEAT DESCRIPTION	755.832676	0.000000
28	LIGHTING DESCRIPTION	490.549354	0.000000
17	WALLS ENV EFF	422.659187	0.000000
16	WALLS ENERGY EFF	422.659187	0.000000
32	SOLAR WATER HEATING FLAG	400.703137	0.000000
15	WALLS DESCRIPTION	346.710221	0.000000
11	FLOOR DESCRIPTION	271.854618	0.000000
9	HOT WATER ENERGY EFF	249.767781	0.000000
18	SECONDHEAT DESCRIPTION	223.014369	0.000000
3	TRANSACTION TYPE	213.705501	0.000000
23	MAINHEAT ENERGY EFF	203.361528	0.000000
34	CONSTRUCTION AGE BAND	192.414369	0.000000
31	MAIN FUEL	124.961869	0.000000
20	ROOF ENERGY EFF	88.966841	0.000000
21	ROOF ENV EFF	88.966841	0.000000
24	MAINHEAT ENV EFF	84.399353	0.000000
27	MAINHEATC ENV EFF	79.227915	0.000000
26	MAINHEATC ENERGY EFF	79.227915	0.000000
13	WINDOWS ENERGY EFF	75.508281	0.000000
14	WINDOWS ENV EFF	75.508281	0.000000
25	MAINHEATCONT DESCRIPTION	61.744209	0.000000
10	HOT WATER ENV EFF	61.585880	0.000000
35	TENURE	49.339901	0.000000
5	MAINS GAS FLAG	40.042863	0.000000
2	BUILT FORM	35.410391	0.000004
6	GLAZED TYPE	33.857832	0.000007
4	ENERGY TARIFF	30.493627	0.000032
8	HOTWATER DESCRIPTION	26.598368	0.000172
1	PROPERTY TYPE	26.276871	0.000198
29	LIGHTING ENERGY EFF	18.060460	0.006083
30	LIGHTING ENV EFF	18.060460	0.006083
33	MECHANICAL VENTILATION	0.940513	0.987767
7	GLAZED AREA	0.755909	0.993204

B Statistik der EA-Stichprobe

Tabelle B.2: Beschreibende Statistik der EA-Stichprobe der kategorialen Merkmale

feature	count	unique	top	freq
ADDRESS1	4993	4313	Flat 2	59
ADDRESS2	2498	1527	Wombwell	94
ADDRESS3	787	440	Browns Green	110
CURRENT ENERGY RATING	4993	7	C	2556
POTENTIAL ENERGY RATING	4993	7	B	2598
PROPERTY TYPE	4993	5	House	2605
BUILT FORM	4993	6	Mid-Terrace	1806
LOCAL AUTHORITY	4993	315	E08000025	204
CONSTITUENCY	4993	564	E14000542	155
COUNTY	1257	21	Kent	120
TRANSACTION TYPE	4993	10	rental	2630
ENERGY TARIFF	4158	6	Single	3603
MAINS GAS FLAG	4936	2	Y	3862
FLAT TOP STOREY	2051	2	N	1275
GLAZED TYPE	4639	6	double glazing, unknown install date	2261
GLAZED AREA	4936	5	Normal	4634
HOTWATER DESCRIPTION	4993	21	From main system	3751
HOT WATER ENERGY EFF	4993	5	Good	3863
HOT WATER ENV EFF	4993	5	Good	3898
FLOOR DESCRIPTION	4993	31	Solid, no insulation (assumed)	2006
FLOOR ENERGY EFF	38	3	Very Good	31
FLOOR ENV EFF	38	3	Very Good	31

Fortsetzung auf nächster Seite

Tabelle B.2 – *Fortsetzung von vorheriger Seite*

feature	count	unique	top	freq
WINDOWS DESCRIPTION	4993	12	Fully double glazed	4523
WINDOWS ENERGY EFF	4993	5	Average	2813
WINDOWS ENV EFF	4993	5	Average	2813
WALLS DESCRIPTION	4993	58	Cavity wall, filled cavity	1553
WALLS ENERGY EFF	4993	5	Average	1394
WALLS ENV EFF	4993	5	Average	1394
SECONDHEAT DESCRIPTION	1371	10	Room heaters, electric	603
ROOF DESCRIPTION	4993	44	(another dwelling above)	1212
ROOF ENERGY EFF	3769	5	Good	1643
ROOF ENV EFF	3769	5	Good	1643
MAINHEAT DESCRIPTION	4993	41	Boiler and radiators, mains gas	3672
MAINHEAT ENERGY EFF	4993	5	Good	4120
MAINHEAT ENV EFF	4993	5	Good	4125
MAINHEATCONT DESCRIPTION	4984	28	Programmer, room thermostat and TRVs	3026
MAINHEATC ENERGY EFF	4993	5	Good	3544
MAINHEATC ENV EFF	4993	5	Good	3544
LIGHTING DESCRIPTION	4993	83	Low energy lighting in all fixed outlets	3486
LIGHTING ENERGY EFF	4993	5	Very Good	4284
LIGHTING ENV EFF	4993	5	Very Good	4284
MAIN FUEL	4993	19	mains gas (not community)	3735
HEAT LOSS CORRIDOR	1994	3	no corridor	805
SOLAR WATER HEATING FLAG	4936	2	N	4904
MECHANICAL VENTILATION	4936	3	natural	4908
ADDRESS	4993	4988	25b Melbourne Road	2
LOCAL AUTHORITY LABEL	4993	315	Birmingham	204

Fortsetzung auf nächster Seite

Tabelle B.2 – *Fortsetzung von vorheriger Seite*

feature	count	unique	top	freq
CONSTITUENCY LABEL	4993	564	Barnsley East	155
POSTTOWN	4971	765	LONDON	771
CONSTRUCTION AGE BAND	4993	17	England and Wales: 1950-1966	1047
LODGEMENT DATETIME	4993	4957	2023-12-30 12:36:54	3
TENURE	4993	4	Owner-occupied	2003
UPRN SOURCE	4727	2	Energy Assessor	4639

Tabelle B.1: Beschreibende Statistik der EA-Stichprobe der numerische Merkmale

	count	mean	std	min	25%	50%	75%	max
CURRENT_ENERGY_EFFICIENCY	5000.0	68.57	12.58	1.0	63.00	69.0	75.00	107.00
POTENTIAL_ENERGY_EFFICIENCY	5000.0	85.45	6.97	27.0	82.00	85.0	89.00	123.00
ENVIRONMENT_IMPACT_CURRENT	5000.0	65.71	14.59	1.0	57.00	66.0	74.00	106.00
ENVIRONMENT_IMPACT_POTENTIAL	5000.0	83.33	9.18	24.0	78.00	83.0	88.00	121.00
ENERGY_CONSUMPTION_CURRENT	5000.0	211.63	107.13	-63.0	150.00	205.0	257.00	1236.00
ENERGY_CONSUMPTION_POTENTIAL	5000.0	90.39	54.62	-198.0	63.00	93.0	119.00	523.00
CO2_EMISSIONS_CURRENT	5000.0	3.74	2.64	-0.9	2.30	3.2	4.50	48.00
CO2_EMISS_CURR_PER_FLOOR_AREA	5000.0	37.89	19.93	-11.0	27.00	36.0	46.00	354.00
CO2_EMISSIONS_POTENTIAL	5000.0	1.76	1.66	-2.0	0.90	1.5	2.20	47.00
LIGHTING_COST_CURRENT	5000.0	164.68	61.90	37.0	127.00	152.0	187.00	1452.00
LIGHTING_COST_POTENTIAL	5000.0	150.92	46.93	39.0	123.00	143.0	170.00	730.00
HEATING_COST_CURRENT	5000.0	1483.09	1086.09	134.0	880.75	1241.0	1791.25	20614.00
HEATING_COST_POTENTIAL	5000.0	1107.74	677.96	151.0	749.00	988.0	1313.25	17410.00
HOT_WATER_COST_CURRENT	5000.0	293.99	174.62	65.0	223.00	254.0	315.00	2627.00
HOT_WATER_COST_POTENTIAL	5000.0	176.52	61.84	41.0	143.00	169.0	195.00	770.00
TOTAL_FLOOR_AREA	5000.0	102.46	55.76	25.0	74.00	88.0	113.00	1321.00
MULTI_GLAZE_PROPORTION	5000.0	96.59	15.71	0.0	100.00	100.0	100.00	100.00
EXTENSION_COUNT	4252.0	0.74	0.89	0.0	0.00	1.0	1.00	4.00
NUMBER_HABITABLE_ROOMS	4252.0	4.95	1.58	1.0	4.00	5.0	6.00	22.00
NUMBER_HEATED_ROOMS	4252.0	4.88	1.62	0.0	4.00	5.0	6.00	22.00
LOW_ENERGY_LIGHTING	5000.0	88.15	22.66	0.0	85.00	100.0	100.00	100.00
NUMBER_OPEN_FIREPLACES	5000.0	0.085	0.36	0.0	0.00	0.0	0.00	6.00
WIND_TURBINE_COUNT	5000.0	0.0002	0.01	0.0	0.00	0.0	0.00	1.00
FLOOR_HEIGHT	5000.0	2.415	0.25	0.0	2.33	2.4	2.50	3.81
PHOTO_SUPPLY	4142.0	1.93	9.31	0.0	0.00	0.0	0.00	100.00
FIXED_LIGHTING_OUTLETS_COUNT	5000.0	13.13	8.26	1.0	9.00	11.0	15.00	100.00

C Merkmalsbeschreibung

Tabelle C.1: Merkmalsbeschreibung aus der EA-Datenbank mit *Feldname* - Originalname aus der Datenbank, *Typ* - Datentyp des Merkmals und *Exkl.* - Exkludiertes Merkmal bei der Modellerstellung "1" = Ja und "0" = Nein

ID	Feldname	Typ	Beschreibung	Exkl.	Begründung
0	LMK KEY	str	Individual lodgement identifier	1	Kein Bezug zum Zielmerkmal
1	ADDRESS1	str	Addresszeile 1	1	Kein Bezug zum Zielmerkmal
2	ADDRESS2	str	Addresszeile 2	1	Kein Bezug zum Zielmerkmal
3	ADDRESS3	str	Addresszeile 3	1	Kein Bezug zum Zielmerkmal
4	POSTCODE	str	Postleitzahl vom Objekt	1	Kein Bezug zum Zielmerkmal
5	BUILDING REFERENCE NUMBER	str	Eindeutiger Identifikator für das Grundstück	1	Kein Bezug zum Zielmerkmal aber verknüpft mit den Empfehlungen
6	CURRENT ENERGY RATING	str	Aktuelle Energiebewertung in eine lineare Bewertung von "A bis G" umgewandelt (wobei A die höchste Energieeffizienz und G die geringste Energieeffizienz bedeutet)	0	Abgleitet bzw. übersetzt von der Energieeffizienz

Fortsetzung auf nächste Seite

Tabelle C.1 – *Fortsetzung von vorheriger Seite*

ID	Feldname	Typ	Beschreibung	Exkl.	Begründung
7	POTENTIAL ENERGY RATING	str	Geschätzte potenzielle Energiebewertung, umgewandelt in eine lineare Bewertung von A bis G	1	Theoretisches Rating basierend auf den Empfehlungen
8	CURRENT ENERGY EFFICIENCY	int	Basierend auf den Energiekosten, d. h. der für Raumheizung, Warmwasserbereitung und Beleuchtung benötigten Energie [in kWh/Jahr] multipliziert mit den Brennstoffkosten. (£/m2/Jahr, wobei die Kosten aus kWh abgeleitet werden)	0	
9	POTENTIAL ENERGY EFFICIENCY	int	Potenzielle Energieeffizienzklasse	1	Theoretische Effizienz auf Basis der Empfehlungen
10	PROPERTY TYPE	str	Beschreibt den Gebäudetyp (Haus, Wohnung, Bungalow, Maisonette)	0	
11	BUILT FORM	str	Zusammen mit dem Gebäudetyp eine erhält man eine strukturierte Beschreibung des Objekts: freistehend, Doppelhaushälfte...	0	
12	INSPECTION DATE	date	Datum der tatsächlichen Insprektion durch den Gutachter	1	Zuordung ohne Ausweis nicht möglich
13	LOCAL AUTHORITY	str	Gemeindegebiet wo sich das Objekt befindet	1	Kein Bezug zum Zielmerkmal
14	COUNTY	str	Landkreis in dem das Gebäude steht	1	Kein Bezug zum Zielmerkmal

Fortsetzung auf nächste Seite

Tabelle C.1 – *Fortsetzung von vorheriger Seite*

ID	Feldname	Typ	Beschreibung	Exkl.	Begründung
15	LODGEMENT DATE	date	Datum des Eintrags in das Register	1	Zuordung ohne Ausweis nicht möglich
16	TRANSACTION TYPE	str	Grund für die Erstellung des Ausweises (Verkauf, Vermietung...)	1	Kein Bezug zum Zielmerkmal
17	ENVIRONMENT IMPACT CURRENT	int	Ein Rating für die aktuellen Auswirkungen der Immobilie auf die Umwelt in Bezug auf Kohlendioxid (CO_2)-Emissionen.	1	Einzigartiges (Ziel-)Merkmal ohne geeignete Alternativquelle
18	ENVIRONMENT IMPACT POTENTIAL	int	Geschätztes Umwelt-Rating	1	Theoretisches Rating basierend auf den Empfehlungen
19	TOTAL FLOOR AREA	float	Die gesamte nutzbare Grundfläche ist die Summe aller umschlossenen Räume, gemessen bis zur Innenseite der Außenwände.	0	
20	ENERGY TARIFF	str	Art des Stromtarifs für die Immobilie, z.B. Singletarif	0	
21	MAINS GAS FLAG	Y/N	Ob ein Anschluss zum Gasnetz besteht	0	
22	FLOOR LEVEL	str	Nur Wohnungen und Maisonetten	1	59 % fehlende Einträge
23	FLAT TOP STOREY	Y/N	Ob die Wohnung im obersten Stock ist	1	58 % fehlende Einträge
24	FLAT STOREY COUNT	int	Anzahl der Stockwerke	1	99 % fehlende Einträge
25	MAIN HEATING CONTROLS	str	Art der Heizungssteuerung	1	100 % fehlende Einträge

Fortsetzung auf nächste Seite

Tabelle C.1 – *Fortsetzung von vorheriger Seite*

ID	Feldname	Typ	Beschreibung	Exkl.	Begründung
26	MULTI GLAZE PROPORTION	int	Prozentsatz des verglaste Bereiche. Mehrfachverglasungsanteil anhand der Fläche und Art des jeweiligen Fensters bzw. Dachfensters berechnet	0	
27	GLAZED TYPE	int	Art der Verglasung: einfach, doppel, dreifach	0	
28	GLAZED AREA	str	Flächenschätzung der gesamten verglasten Wohnfläche	0	
29	EXTENSION COUNT	int	Anzahl der Anbauten zwischen 0-4	0	
30	NUMBER HABITABLE ROOMS	int	Anzahl der bewohnbaren Räume	0	
31	NUMBER HEATED ROOMS	int	Anzahl der beheizten Räume mit einem Wärmetauscher	1	0.86 Korrelation mit Merkmal ID29
32	CONSTRUCTION AGE BAND	str	Angabe zur Baualtersklasse sowie Angabe des Baujahres	0	
32	LOW ENERGY LIGHTING	int	Prozentsatz der vorhandenen Niedrigenergiebeleuchtung	1	Gleiche Information in Merkmal ID42
33	NUMBER OPEN FIREPLACES	int	Anzahl der offenen Feuerplätze z.B. Kamin	0	
34	HOTWATER DESCRIPTION	str	Beschreibung der Warmwasserbereitung	0	
35	FLOOR DESCRIPTION	str	Beschreibung des Bodens	0	

Fortsetzung auf nächste Seite

Tabelle C.1 – *Fortsetzung von vorheriger Seite*

ID	Feldname	Typ	Beschreibung	Exkl.	Begründung
36	WINDOWS DESCRIPTION	str	Beschreibung der Fenster	0	
37	WALLS DESCRIPTION	str	Beschreibung der Wände	0	
38	SECONDHEAT DESCRIPTION	str	Beschreibung der Sekundärheizung	1	Chi^2 Signifikanz mit ID25 und ID21
39	ROOF DESCRIPTION	str	Beschreibung des Dachs	0	
40	MAINHEAT DESCRIPTION	str	Beschreibung der Raumwärmebereitung	0	
41	MAINHEATCONT DESCRIPTION	str	Beschreibung zur Heizungssteuerung	1	Gleiche Information in ID25
42	LIGHTING DESCRIPTION	str	Beschreibung zur Beleuchtung	0	
43	MAIN FUEL	str	Art des Brenstoffes zur Wärmeversorgung	1	Chi^2 Signifikanz mit ID21 und ID25 aber Alternativquelle für ID21
44	WIND TURBINE COUNT	int	Anzahl installierter Windräder	0	
45	HEAT LOSS CORRIDOOR	str	Indikator, ob ein nicht beheizter Flur vorhanden ist	1	60 % fehlende Einträge
46	UNHEATED CORRIDOR LENGTH	float	Angabe wenn ein ID45 vorhanden ist	1	82 % fehlende Einträge
47	FLOOR HEIGHT	float	Geschosshöhe	0	
48	SOLAR WATER HEATING FLAG	Y/N	Indikator ob Solarthermie vorhanden ist	0	

Fortsetzung auf nächste Seite

Tabelle C.1 – *Fortsetzung von vorheriger Seite*

ID	Feldname	Typ	Beschreibung	Exkl.	Begründung
49	MECHANICAL VENTILATION	str	Art der Belüftung	1	Einmaliges Merkmal ohne Alternativquelle
50	..._ENERGY EFF Merkmale	str		1	Abgeleitet von anderen Merkmalen im Datensatz
51	..._ENV EFF Merkmale	str		1	Abgeleitet von anderen Merkmalen im Datensatz
52	LIGHTING COST CURRENT, HEATING COST CURRENT	float	Energiekosten für Heizen, Warmwasser und Beleuchtung	1	Einmaliges Merkmal ohne Alternativquelle. Proxy für den Verbrauch.

D Digitaler Anhang

Die in dieser Masterarbeit verwendeten Parameter sind als digitaler Anhang unter folgender Adresse verfügbar: `https://github.com/psomm/EPC-Modelling` In dem Repository sind die

- Vorverarbeitungsmethoden,
- Sub-Datensätze,
- Modelle sowie dazugehörige Pipeline und
- Validierungen

von allen Untersuchungen enthalten. Die Untersuchungen wurden im Rahmen der Arbeiten in einem privaten Repository durchgeführt. An der Stelle wird nochmal darauf hingewiesen, dass im Falle einer Veröffentlichung („public“) des Repositorys keine Stichproben zu den Energieausweisen enthalten sind und entsprechend entfernt wurden, um den Datenschutz zu gewährleisten. Zur Reproduktion der Ergebnisse wird empfohlen eine Stichprobe unter folgender Adresse zu beziehen: `https://epc.opendatacommunities.org/`

E Aggregierte Merkmalsrelevanzen

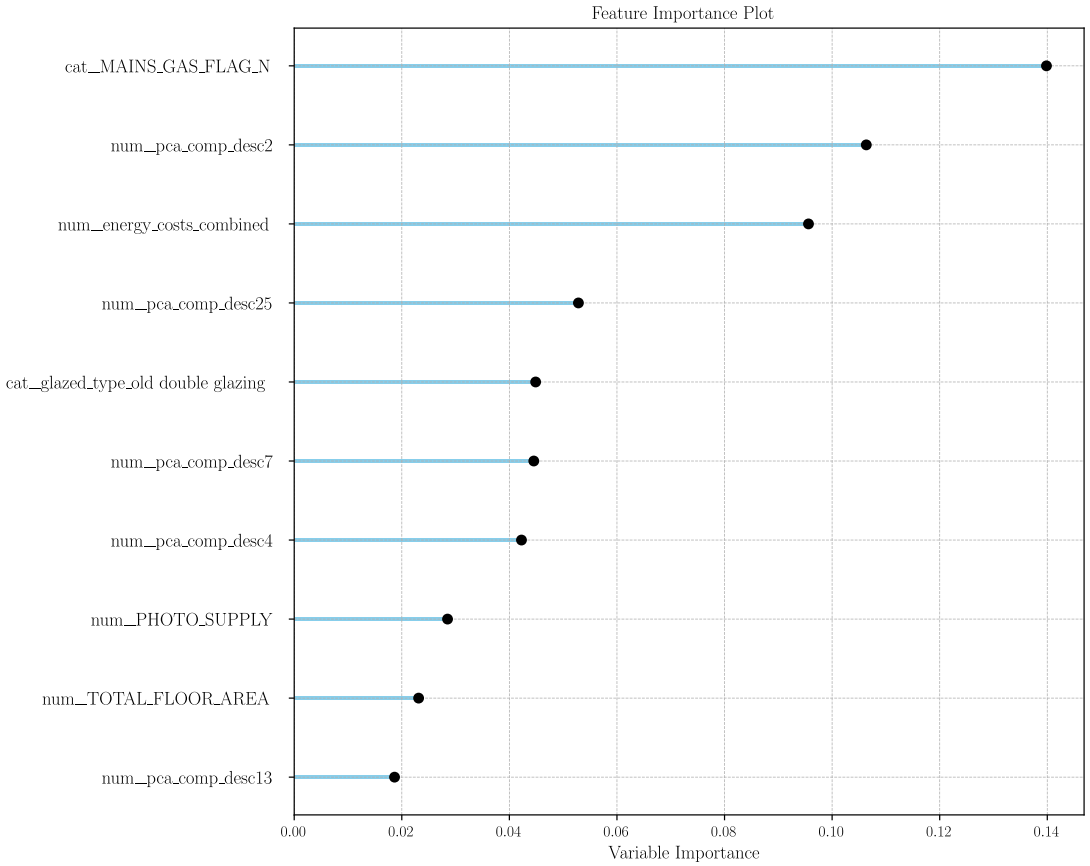

Abbildung E.1: Aggregierte Relevanz der Merkmale für den datengetriebenen Ansatz

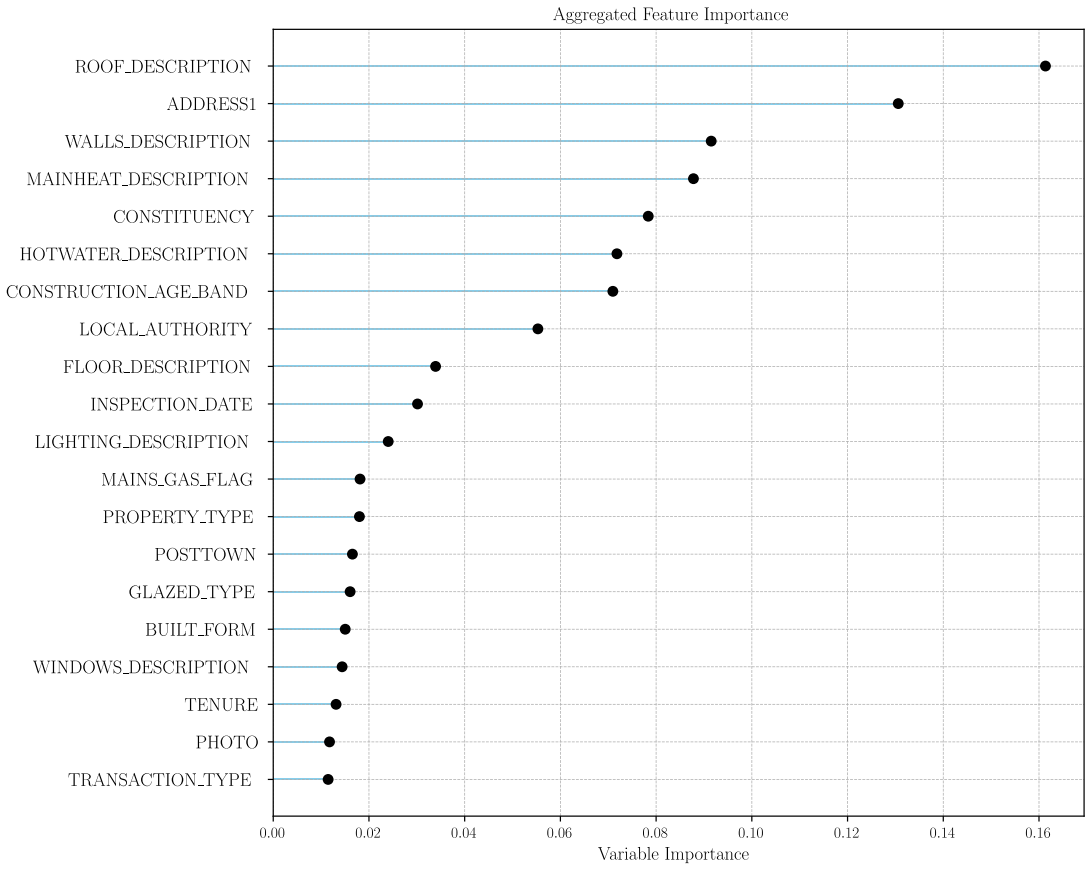

Abbildung E.2: Aggregierte Relevanz der Merkmale für den erschöpfenden Ansatz

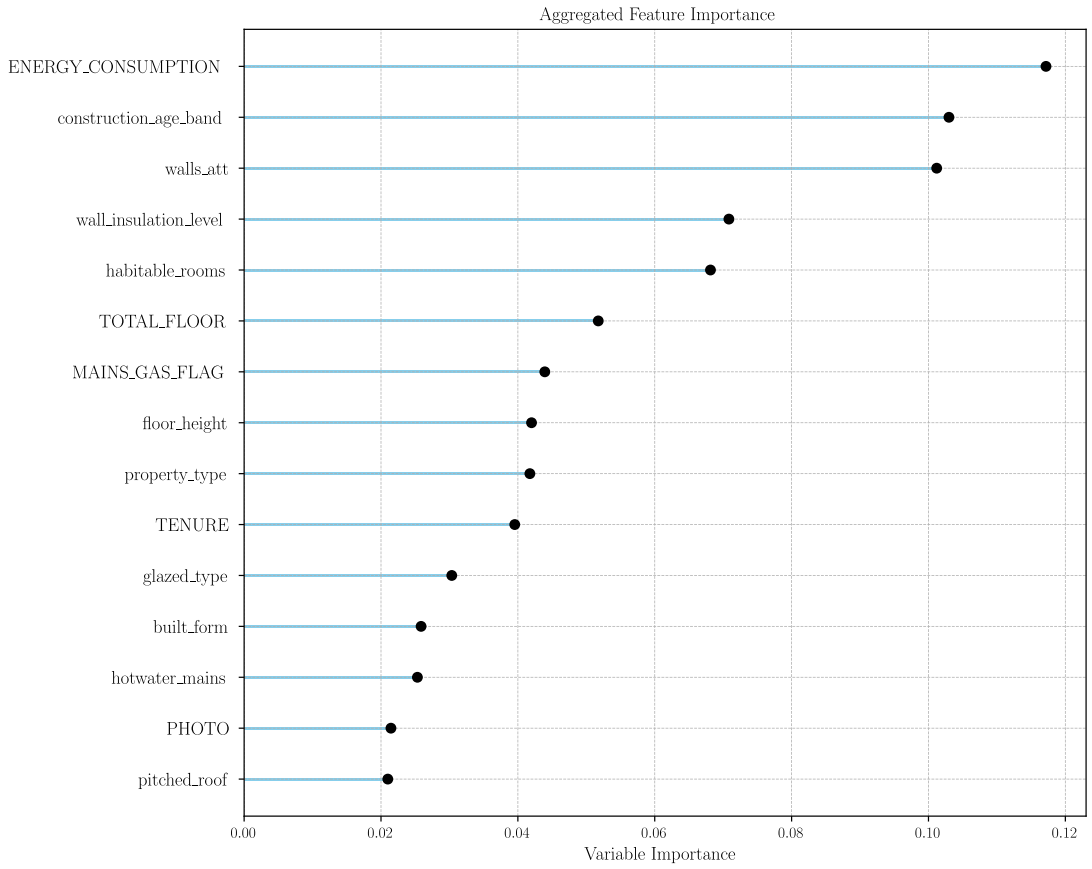

Abbildung E.3: Aggregierte Relevanz der Merkmale für den Fernerkundung+ Ansatz

F Datenquellen geordnet nach Themengruppen

Tabelle F.1: Angepasste Übersicht der (offenen) Datenquellen und Zuständigkeiten mit den Abk.: MaStR - Marktstammdatenregister, EVU - Energieversorgungsunternehmen, ALKIS - Amtliches Liegenschaftskatasterinformationssystem, LEP - Landesenwticklungsplan, FNP - Flächennutzungsplan, B-Plan - Bebauungsplan [75]

Themengruppen	**MaStR**	**Kehrbuchdaten**	**EVU**	**ALKIS**	**Geoportal NRW**	**Zensus2022**
Gas- und Wärmeverbräuche	X		X			
Dezentrale Wärmeerzeugungsanlagen Verbrennungstechnik		X				
Gebäudedaten				X	X	X
Industrie, Gewerbe und Unternehmen (Prozess- u. Abwärme)	X					
Wärmenetze						
Wärmeerzeuger	X	X	X			
Gasnetze						
Stromnetze (Hoch- und Mittelspannung)	X	X				
Niederspannungsnetze	X	X				
Kläranlagen		X	X			
Abwassernetze		X				
Bauleitpläne (LEP, FNP, B-Plan)				X		