POLITECNICO DI TORINO

Corso di Laurea Magistrale
in Ingegneria Energetica e Nucleare

Tesi di Laurea Magistrale

Analisi di attestati di prestazione energetica di edifici attraverso processi di intelligenza artificiale

Relatori
Prof. Alfonso Capozzoli
Dr. Marco Savino Piscitelli

Candidato
Giacomo Buscemi

Anno Accademico 2021-2022

Sommario

Gli Attestati di Prestazione Energetica (*APE*) sono un importante strumento per la valutazione e il miglioramento dell'efficienza energetica negli edifici. Per tale motivo la stima del fabbisogno energetico di un edificio ricopre un ruolo fondamentale per i professionisti del settore che vogliono, ad esempio, confrontare le prestazioni di un gruppo di edifici o redigere un progetto di riqualificazione energetica, ancor più se ciò avviene in modo veloce, affidabile e scalabile su più configurazioni diverse. Questo lavoro di tesi propone diverse metodologie data-driven per la valutazione dell'Indice di prestazione energetica non rinnovabile di un edificio, sviluppate su un database di circa 50000 *APE* relativi alla Regione Piemonte, che possono avere finalità di verifica del lavoro di redazione svolto o di rapido calcolo dell'andamento energetico dell'immobile. L'idea alla base è quella di avere la possibilità di confrontare l'edificio in esame con la molteplicità di immobili simili le cui informazioni sono catalogate nei certificati di prestazione, mettendo a disposizione una baseline di riferimento con la quale paragonare il livello di efficienza energetica. Algoritmi non supervisionati di clustering sono stati utilizzati per rilevare gli edifici di riferimento all'interno del database, i quali presentavano attributi caratteristici rappresentativi di un determinato gruppo di fabbricati. In seguito si è proseguito con l'utilizzo di modelli di regressione, per stimare il valore degli indici di consumo identificanti le prestazioni energetiche di un immobile, sfruttando degli attributi di input specifici precedentemente filtrati e selezionati. Cinque algoritmi di machine learning (KNN, Regression Tree, Random Forest, Boosting and Artificial Neural Networks) sono stati utilizzati e comparati, al fine di ottenere la predizione con il minore tasso di errore. Ai modelli di predizione black-box, sono state applicate le metodologie di *eXplainable Artificial Intelligence (XAI)* al fine di rendere maggiormente interpretabili i risultati prodotti. Comprendere le motivazioni dietro la formulazione del risultato fornito dal modello, diventa di cruciale importanza qualora si investa su tali risultati con opere di ristrutturazione o di nuova costruzione. L'obiettivo della tesi è dunque quello di far emergere il potenziale degli attestati di prestazione energetica, sfruttando la loro finalità di raccolta e catalogazione delle informazioni del parco edilizio italiano al fine di: fornire una base statistica di supporto alla compilazione del certificato; generare una linea guida di riferimento attraverso la quale confrontare le prestazioni di un edificio esistente o di nuova costruzione; stimare i consumi futuri di un edificio; fare emergere diverse possibilità di riqualificazione per un edificio esistente.

Ringraziamenti

Parte di questa tesi è stata sviluppata nell'ambito di un progetto, finanziato nel contesto del bando regionale VIR-Piemonte, dal titolo "Definizione di processi innovativi di intelligenza artificiale per il supporto alle decisioni in ambito energetico attraverso l'analisi di attestati di prestazione energetica di edifici – APE4IA" sviluppato in collaborazione tra il Politecnico di Torino - Dipartimento Energia e la società EDILCLIMA s.r.l.
Ringrazio il BAEDA Lab, in particolare il Prof. Alfonso Capozzoli e il Dr. Marco Savino Piscitelli, per avermi concesso l'opportunità di iniziare un percorso di crescita personale e professionale e soprattutto per la fiducia accordatami.
Grazie ai miei genitori, ai quali ho sempre cercato di esprimere la mia gratitudine più con i fatti che con le parole, ma mi concedo un' eccezione.
E a Vera, grazie per essere stata sempre lì al mio fianco in questa avventura chiamata vita.

Indice

Elenco delle tabelle VI

Elenco delle figure VII

Introduzione 1

1 L'Attestato di Prestazione Energetica 6

1.1 Quadro normativo di riferimento 6

1.2 Dataset utilizzati 9

1.3 Le Classi Energetiche 9

1.4 Sezioni Attestato di Prestazione Energetica 13

1.4.1 Sezione Dati Generali 13

1.4.2 Sezione Prestazione Energetica 15

1.4.3 Sezione Raccomandazioni 15

1.4.4 Sezione Dati di dettaglio 16

2 Metodi e Tecniche di Analisi 19

2.1 Tecniche di statistica descrittiva 19

2.2 Analisi di Clustering 23
2.2.1 Concetto di distanza nell'analisi di clustering 23
2.2.2 Tecniche di clustering partitivo 26
2.2.3 Tecniche di clustering gerarchico 28
2.2.4 Tecniche di identificazione del numero ottimale di cluster . . 31
2.3 Analisi di Regressione e Classificazione 33
2.3.1 K-Nearest Neighbors (KNN) 34
2.3.2 Classification And Regression Tree (CART) 36
2.3.3 Metodi Ensamble 40
2.3.4 Bayesian Additive Regression Tree 42
2.3.5 Artificial Neural Network 43
2.4 Tecniche di eXplainable Artificial Intelligence 46
2.4.1 Partial Dependence Plot (PDP) e Accumulated Local Effects (ALE) Plot 47
2.4.2 Permutation Feature Importance 50
2.4.3 Break-down Plot 51

3 Framework Metodologico 53

3.1 Obiettivo 1 ~ Identificazione del dataset di riferimento e pre-processamento dei dati 54
3.2 Obiettivo 2 ~ Statistica descrittiva 55
3.3 Obiettivo 3 ~ Clustering 56
3.4 Obiettivo 4 ~ Modelli di previsione 58
3.5 Obiettivo 5 ~ Layer di XAI 60

4 Risultati 61

4.1 Obiettivo 1 61

4.1.1 Identificazione e pre-elaborazione del dataset di attestati di riferimento 61

4.1.2 Rimozione valori anomali 66

4.2 Obiettivo 2 70

4.3 Obiettivo 3 73

4.4 Obiettivo 4 78

4.4.1 Approccio a un livello di analisi per lo sviluppo di un modello previsionale regressivo 79

4.4.2 Approccio a due livelli di analisi per lo sviluppo di un modello previsionale regressivo 84

4.5 Obiettivo 5 90

4.5.1 Global Methods 90

4.5.2 Local Methods 92

5 Conclusioni e Discussione dei Risultati 97

5.1 Prospettive future di implementazione 99

Bibliografia 103

Elenco delle tabelle

1.1 Dataset disponibili nei cataloghi data.gov e dati.piemonte 10

1.2 Scala di classificazione degli edifici sulla base dell'indice di prestazione energetica globale non rinnovabile, secondo l'allegato 1 DM requisiti minimi n.162/2015 12

1.3 Categorie destinazione d'uso secondo il D.P.R 412/93 [1] 14

3.1 Variabili analisi Clustering 56

3.2 Variabili modello di stima di $Ep_{glo,nren}$ 58

4.1 Dominio di esistenza degli attributi per la rimozione dei valori anomali 68

4.2 Decremento cardinalità del dataset 69

4.3 metriche di validazione del clustering 73

4.4 Centroidi *k-means* 75

4.5 Risultati approccio ad un livello di analisi 81

4.6 Risultati classificazione 87

4.7 Risultati Regressione algoritmo BART sui range reali 88

4.8 Risultati Regressione algoritmo BART sui range predetti dal classificatore 88

4.9 Tempi computazionali relativi al KNN 89

Elenco delle figure

1.1 Fac-simile prestazione energetica globale dell'edificio APE 11

2.1 Esempio di un boxplot 20

2.2 Probabilità definita come l'area sottesa alla funzione densità di probabilità 21

2.3 Funzione di probabilità cumulata in relazione a una funzione di densità gaussiana 22

2.4 Rappresentazione della distanza in uno spazio bi-dimensionale . . . 24

2.5 Rappresentazione distanza Euclidea e Manhattan 25

2.6 Passaggi algoritmo k-means 27

2.7 Dendogramma clustering gerarchico 29

2.8 Linkage Methods 30

2.9 Metodo di stima del punto di gomito 32

2.10 Metodo della silhouette 33

2.11 KNN per la Classificazione 34

2.12 KNN per la Regressione 35

2.13 Stima del numero ottimale di vicini che ottimizza l'accuratezza del modello 36

2.14 Esempio di albero di Regressione 38

2.15 Schema del Random Forest: Bagging 41
2.16 Schema del Boosting 41
2.17 Deep Learning 44
2.18 Schema Neurone 44
2.19 Schema ANN 45
2.20 Metodologia XAI 47
2.21 Distribuzione marginale e condizionata 48
2.22 Valutazione dell'ALE per l'attributo x_1 correlato a x_2 49
2.23 Permutation Feature Importance plot, per un modello KNN 50
2.24 Esempio *Break-down Plot* 51
3.1 Framework Metodologico 53
3.2 Metodologia clustering 57
3.3 Approcci per la predizione 59
4.1 Distribuzione delle province piemontesi nel dataset 62
4.2 Distribuzione delle destinazioni d'uso all'interno del dataset 63
4.3 Distribuzione tipologie di immobile oggetto dell'attestato 64
4.4 Distribuzione servizi energetici 65
4.5 Distribuzione edifici in base al numero di impianti installati e al servizio energetico fornito 65
4.6 Distribuzione del combustibile utilizzato dagli impianti 66
4.7 Scatter plot dei valori prima dell'imposizione dei filtri 67
4.8 Decremento cardinalità del dataset 70
4.9 Analisi sensitività rendimenti sottosistema 70

4.10 Distribuzione delle trasmittanza evidenziando i tre intervalli di valori 71

4.11 Distribuzione U_{op} per tipologia edilizia e periodo di costruzione . . . 72

4.12 Distribuzione APE per periodo di costruzione 72

4.13 Metodo del gomito e indice di Silhouette 74

4.14 Coordinate parallele centroidi *k-means* 75

4.15 Distribuzione $Ep_{glo,nren}$ per cluster 76

4.16 Confronto cluster 1 e 16 77

4.17 Assegnazione di un cluster a un nuovo edificio 78

4.18 *k-fold cross validation* applicata all'iperparametro K dell'algoritmo *KNN* 80

4.19 Scatter plot dei valori dell'$Ep_{glo,nren}$ reale e predetto con il modello regressivo BART 82

4.20 Grafico dei residui comparati alla stima dell'$Ep_{glo,nren}$ 83

4.21 Suddivisione $Ep_{glo,nren}$ in tre range di consumo *Low*, *Medium* e *High* 84

4.22 Concettualizzazione metriche accuratezza 86

4.23 Scatter plot dei valori dell'$Ep_{glo,nren}$ reale e predetto con il modello regressivo BART, applicato a tre intervalli di consumo 89

4.24 ALE plot BART 90

4.25 Feature importance plot BART 91

4.26 BreakDown plot per un edificio con alto consumo stimato, con ordinamento delle variabili predeterminato 93

4.27 BreakDown plot per un edificio con alto consumo stimato, con ordinamento delle variabili euristico 94

5.1 Rappresentazione valore stimato $Ep_{glo,nren}$ nel suo cluster di riferimento 98

5.2 Attributi dell'edificio i cui consumi sono rappresentati in Figura 5.1 99

Introduzione

L'efficienza energetica è un tema di crescente interesse politico e sociale per molti paesi sparsi per il mondo, per ragioni sia economiche che ambientali. La Commissione europea ha reso noto che gli edifici dei paesi membri l'UE consumano circa il 40% dell'energia e rilasciano il 36% delle emissioni di gas serra associate alla produzione energetica, ma annualmente solamente l'1% di essi viene sottoposto a lavori di ristrutturazione e retrofit a fini di efficientamento energetico [2]. Si è ritenuto dunque necessario implementare dei piani d'azione nazionali di regolamentazione dei consumi energetici, al fine di ridurre gli sprechi di energia, dovuti a un utilizzo poco efficiente delle risorse disponibili, e ad incentivare la transizione verso l'utilizzo di fonti energetiche rinnovabili e dal basso impatto ambientale.

In questo scenario gli *Attestati di Prestazione Energetica (APE)* sono stati sviluppati come uno strumento per migliorare l'efficienza energetica, diminuire il consumo e fornire maggiore trasparenza sull'uso dell'energia negli edifici. Essi ricoprono un ruolo fondamentale per la stima e il miglioramento dell'efficienza energetica degli edifici, fornendo informazioni riguardo le loro performance e caratteristiche termofisiche e geometriche, che racchiudono un elevato potenziale di conoscenza del patrimonio edilizio. Tale conoscenza per essere utilizzata, deve essere estratta da un numero significativo di dati; le tecniche di intelligenza artificiale sono lo strumento chiave per adempiere a questo scopo [3]. Ciononostante, la differenza fra le prestazioni energetiche di un edificio valutate con modelli di stima e quelle reali effettive, risulta ancora significativa. Come avvalorato in [4], i valori di consumo inseriti nel dataset svedese dei certificati di prestazione energetica presentavano un divario di circa il 20% rispetto a quelli stimati da modelli regressivi. Recenti studi hanno individuato fra le cause determinanti tali divergenze la poca accuratezza delle variabili di input relative alla fisica dell'edificio, le condizioni meteo mutevoli e la difficoltà nel determinare un profilo energetico di utilizzo delle risorse da parte dell'occupante [5]. Altri fattori che influenzano il calcolo delle prestazioni dell'edificio sono le assunzioni effettuate nei modelli di calcolo, riguardanti le infiltrazioni

e la tenuta dell'involucro edilizio, il sistema di riscaldamento degli ambienti, l'inerzia termica delle pareti massive e i parametri per modellare i guadagni solari e l'effetto dell'ombreggiamento [6]. Inoltre, possono verificarsi dei malfunzionamenti o regolazioni errate al sistema di climatizzazione o altre carenze impiantistiche dell'edificio, che si traducono in un consumo di energia variabile nel tempo non ponderabile a priori [7]. Un'*APE* non è dunque perfettamente rappresentativo delle prestazioni energetiche dell'edificio durante il suo effettivo funzionamento, ma rimane uno strumento valido per effettuare confronti ed analisi di benchmark tra edifici.

Pasichnyi [8] nel 2019 ha analizzato gli utilizzi nella letteratura scientifica dei certificati di prestazione energetica e quali tematiche trattassero, valutandone le potenzialità e i molteplici campi di applicazione. Lo studio è stato condotto su un campione di 79 pubblicazioni europee, ed ha presentato le opportunità future che gli *APE* possono offrire nel campo dell'efficienza energetica degli edifici, evidenziando anche le criticità che li affliggono, riguardanti principalmente la qualità e l'interoperabilità dei dati forniti. Nella maggior parte degli articoli esaminati i certificati energetici sono stati utilizzati per analizzare tematiche riguardanti la classificazione delle prestazione energetiche edilizie, operazioni di riqualificazione energetica, valutazione degli effetti degli *APE* sul processo decisionale nel settore del mercato immobiliare, validazione della qualità dei dati e valutazione del divario fra consumi reali e riportati negli attestai. É stato riscontrato anche un notevole aumento della complessità degli studi condotti, evidenziato dal crescente numero di analisi svolte sulla qualità e validazione dei certificati energetici, sull'impiego sempre più frequente di dataset ausiliari e l'applicazione di metodologie di analisi avanzate, basate su processi di intelligenza artificiale.

In accordo con la letteratura scientifica, gli open data relativi agli *APE* rappresentano al giorno d'oggi una grande fonte di informazioni, e un numero sempre più significativo di ricercatori li sta utilizzando per affrontare diversi compiti nell'ambito della gestione energetica degli edifici [9]. In questo contesto, il sistema di benchmarking energetico assume un ruolo fondamentale nella valutazione delle performance di un edificio, coadiuvando il lavoro di diverse figure professionali e protagonisti del settore. Lo scopo principale è valutare, in maniera sistematica, le divergenze fra le prestazioni energetiche di un edificio o di un raggruppamento di essi, rispetto a un edificio di riferimento che funge da linea guida. Esistono due metodologie principali di benchmarking: il confronto delle prestazione di un edificio rispetto a se stesso *(benchmark interno)*, valutando ad esempio i cambiamenti che esso ha assunto a monte e a valle di un intervento di riqualificazione, ed il confronto delle prestazioni di un edificio rispetto ad altri edifici simili *(benchmark*

esterno) [10]. L'identificazione dell'edificio di riferimento può avvenire sia attraverso metodi basati su calcoli analitici sia su metodi data-driven. Con i metodi analitici i consumi vengono confrontati con quelli di un edificio di riferimento simulato da software o strumenti di calcolo. Questo approccio pur avendo un alto grado di affidabilità dei risultati forniti, presenta delle forti limitazioni qualora si volesse analizzare un numero elevato di edifici, dovute sia al tempo computazione esoso che alla dettagliata richiesta di informazioni specifiche, non sempre facilmente reperibili all'interno di un vasto dataset [10], [11]. In alternativa vi sono i metodi data-driven basati su approcci black-box, che sfruttano i valori derivati dai dati di consumo effettivi degli edifici inseriti nel dataset. I processi più comuni vengono sviluppati attraverso modelli statistici o tecniche di data analytics [11]. Su queste ultime in particolare si focalizzerà il lavoro di tesi, nel quale verranno utilizzati algoritmi di clustering per individuare raggruppamenti di edifici simili, da cui estrarre dei profili di riferimento con cui confrontare i nuovi edifici.

Oltre al bechmark energetico, un altro aspetto rilevante trattato nel lavoro di tesi è la stima del valore degli indici di consumo caratterizzanti l'edificio e conseguenzialmente il certificato energetico. Le tecniche di data analytics trovano larga applicazione anche in questo campo e lo dimostrano i molteplici studi condotti sull'argomento. Tso e Yau [12] hanno comparato le accuratezze ottenute da modelli di regressione lineare, reti neurali artificiali *(Artificial Neural Network: ANN)*, e da alberi decisionali nella predizione della media settimanale dei consumi elettrici ad Hong Kong durante l'inverno e l'estate. Yu [13] ha utilizzato un albero decisionale per modellare i consumi reali degli edifici residenziali, al fine di predire il consumo energetico di edifici di nuova costruzione. Melo et al. [14] hanno sviluppato una rete neurale per migliorare l'accuratezza di modelli surrogati utilizzati per fini di classificazione. Attanasio, Piscitelli et al. [9] hanno utilizzato una metodologia a due livelli di analisi per la stima dell'indice di energia primaria disponendo di un dataset di circa 90 000 certificati energetici. La distribuzione dei valori degli indici di prestazione è stata suddivisa in tre range di consumo, basso, alto e molto alto. Il primo livello di analisi consta nel predire il corretto range di consumo precedentemente definito, per un nuovo edificio, attraverso diversi algoritmi di classificazioni basati su alberi decisionali, reti neurali e support vector machine (SVM). Si passa dunque al secondo livello di analisi, nel quale viene sviluppato un modello regressivo di stima del valore numerico dell'indice di prestazione, per ogni range di consumo. In questo modo il primo livello ha lo scopo di raggruppare edifici simili dal punto di vista delle prestazioni, aumentando così l'efficacia dell'addestramento del modello di regressione finale. Un approccio analogo verrà affrontato all'interno di questa tesi.

I risultati ottenuti dai modelli data-driven seppur affidabili, sono tipicamente

complicati da interpretare e comprendere in maniera approfondita. Per sopperire a tale mancanza sono state utilizzate delle metodologie di interpretazione dei modelli di intelligenza artificiale *(eXplainable Artificial Intelligence: XAI)*, le quali mirano a fornire strumenti per rompere il tipico compromesso fra la complessità e l'interpretabilità del modello black box, sfruttandone appieno le potenzialità [10].

Obiettivi e contributo della tesi

Il lavoro di tesi si articola su cinque diversi obiettivi che convogliano nella fase di interpretazione dei risultati.

L'**Obiettivo 1** sarà volto a identificare il dataset di *APE* di riferimento ed applicare un framework di pre-processamento dei dati al fine di irrobustire la base di dati considerata e garantire un adeguato livello qualitativo dei processi analitici successivamente implementati. Ad esso faranno riferimento tutti gli obiettivi successivi.

L'**Obiettivo 2** sarà volto ad individuare valori e trend di riferimento per il fabbisogno di energia primaria e di caratteristiche termofisiche ed impiantistiche degli edifici che costituiscono il campione di *APE* disponibile. A tal fine verranno impiegate tecniche di statistica descrittiva e di apprendimento non supervisionato, in grado di caratterizzare dettagliatamente un campione di dimensioni significativamente rappresentative. L'adempimento di questo obiettivo consentirà di visualizzare le caratteristiche termofisiche o impiantistiche di un edificio rispetto alle tendenze statisticamente più ricorrenti nel campione di riferimento.

L'**Obiettivo 3** ha lo scopo di identificare all'interno del campione in analisi, dei raggruppamenti di edifici rappresentativi da cui sia possibile estrarre un riferimento rispetto alla configurazione delle variabili termofisiche, impiantistiche e di prestazione energetica. Da tali raggruppamenti sarà possibile estrarre dei benchmark multivariati rispetto ai quali confrontare i risultati di prestazione energetica ottenuti per l'edificio in considerazione. A tal fine saranno valutate diverse tecniche di analisi clustering partizionale e gerarchico.

L'**Obiettivo 4** è finalizzato a fornire uno strumento semplice per stimare il fabbisogno energetico di un edificio e che possa essere utilizzato anche per la la stima di scenari di retrofit. Per questo scopo verranno comparati diversi algoritmi di machine learning di apprendimento supervisionato, al fine di selezionare il miglior algoritmo in termini di accuratezza e precisione considerando un set minimo di variabili di input. Tale modello, se allenato su un campione significativamente

rappresentativo consentirà di condurre, in maniera veloce e semplificata, anche analisi di scenari energetici volti a quantificare l'impatto di una potenziale azione di retrofit energetico sull'edificio in considerazione.

Infine, l'**Obiettivo 5** riguarda la definizione di un layer di analisi che fornisca una spiegazione ed interpretazione della stima ottenuta dal modello data-driven precedentemente sviluppato. A tale scopo verranno utilizzate tecniche agnostiche di *XAI (eXplainable Artificial Intelligence)* che consentiranno di rendere il processo di stima trasparente, indipendentemente dall'approccio modellistico perseguito.

A valle di ogni obiettivo, ognuno dei quali verrà descritto nello specifico nel Capitolo 3 relativo al framework metodologico, vi sarà una fase di interpretazione dei risultati ottenuti.

Il lavoro di tesi si sviluppa nei seguenti capitoli: **Capitolo 1**, che funge da presentazione del caso studio relativo all'Attestato di Prestazione Energetica; si ripercorre la storia normativa della certificazione energetica che ha condotto all'attestato odierno, analizzando le sezioni di cui è composto e il Sistema Informativo con cui è gestito. Sono inoltre descritti i dataset utilizzati e gli attributi che li caratterizzano. Nel **Capitolo 2** vengono presentati i metodi analitici utilizzati per condurre il lavoro di tesi. Nel **Capitolo 3** è esposto il framework metodologico con il quale è stata sviluppata l'analisi, descrivendo le fasi che lo compongono. Nel **Capitolo 4** vengono riassunti tutti i risultati ottenuti applicando il caso studio alla metodologia sviluppata. Infine nel **Capitolo 5** si effettua un'analisi critica dei risultati ottenuti in cui si traggono le conclusioni del lavoro svolto e delle possibili prospettive future.

Capitolo 1

L'Attestato di Prestazione Energetica

La certificazione energetica è un metodo di valutazione delle prestazioni di un edificio, attraverso l'assegnazione di una classe energetica, definita in funzione dei consumi annui. L'obiettivo è quello di incentivare le operazioni di miglioramento dell'efficienza energetica dell'edificio, attraverso un sistema informativo che erudisca i proprietari e gli acquirenti di immobili sulle caratteristiche e sui costi energetici effettivi dell'immobile stesso. In questo modo viene fornita una conoscenza approfondita del sistema edificio che consente di effettuare un efficientamento mirato ed efficace sull'abitazione di interesse.

Una moltitudine di studi condotti sull'analisi dei consumi energetici, ha rivelato che gli edifici sono responsabili del 30 - 40% del consumo totale di energia primaria dei paesi maggiormente sviluppati. La consapevolezza del peso del settore edilizio ha dunque portato l'Unione Europea a sviluppare politiche inerenti la prestazione energetica degli edifici, per le quali la certificazione energetica assume un ruolo determinante per studiare le misure di intervento più efficaci per la riduzione dei costi e dell'uso ottimale delle risorse.

1.1 Quadro normativo di riferimento

Il primo passo verso una maggiore attenzione da parte delle politiche nazionali riguardo la tecnologia costruttiva edilizia ed impiantistica avvenne intorno agli anni

'70 del Novecento, in seguito alla pubblicazione dell'Energy Building Conscious Design da parte della Commissione per l'Ambiente della Comunità Economica Europea[15]. Esso si prestava ad essere una raccolta di tutti i vari errori, riguardanti le tematiche energetiche, effettuati fino a quel momento, in modo che gli Stati ne prendessero coscienza ed intervenissero con l'emanazione di leggi opportune. Tutto ciò si sviluppò nel contesto della crisi petrolifera di quegli anni, che inevitabilmente mise in moto quel processo di ricerca, attivo ancora oggi, di un utilizzo sempre più efficiente e mirato delle risorse energetiche disponibili, e soprattutto di fonti primarie alternative.

La prima legge relativa al contenimento dei consumi energetici per usi termici negli edifici fu la L. 373/76, emanata nel 1976, riguardante generalmente gli impianti di produzione di calore e i terminali di regolazione annessi e l'isolamento termico degli edifici. Inoltre furono introdotti per la prima volta i concetti di zone climatiche e di gradi giorno.

Il punto di svolta riguardo le tematiche energetiche nell'edilizia si ebbe con la pubblicazione sulla gazzetta ufficiale della L. 10/1991, che fu la prima legge a regolamentare la progettazione e la gestione del *sistema edificio*, ponendo come obiettivi fondamentali l'uso consapevole dell'energia e il comfort degli individui all'interno dei locali climatizzati. Oltre a ciò la legge proponeva un metodo di valutazione del bilancio energetico invernale, valutato in funzione degli apporti e delle dispersioni di calore. A tal proposito la legge imponeva anche la verifica dell'isolamento dell'involucro edilizio al fine di contenere le dispersioni termiche. Un altro vincolo riguardava il rendimento del sistema di riscaldamento; al di sotto di certi valori non era possibile effettuare il risparmio energetico prefissato. Inoltre la norma indicava di effettuare una relazione tecnica attinente l'edificio, ad opera di un professionista, che ne attestasse l'adeguatezza alle disposizioni della legge stessa.

L'attuazione della *Legge 10/1991* avvenne tramite il D.P.R n. 412/1993, il quale aggiungeva: i criteri di progettazione energetica, la destinazione d'uso con cui classificare gli edifici e la suddivisione del territorio nazionale in una delle sei zone climatiche, determinate in base al periodo convenzionale di riscaldamento annuo, funzione dei gradi giorno[16].

Nel 2002, in accordo con il Protocollo di Kyoto, venne emanata la direttiva 2002/91/CE, detta anche *Energy Performance of Building Directive - EPBD* dal Parlamento Europeo ed il Consiglio dell'Unione, al fine di sensibilizzare i paesi membri a ridurre i consumi energetici in ottemperanza agli obiettivi di riduzione dell'impatto ambientale e dell'inquinamento. In tale direttiva, recepita in Italia

attraverso il dlgs 19 agosto 2005 n. 192, viene introdotto il concetto di *Certificato energetico*, nel quale indicare la prestazione energetica dell'edificio, al fine di facilitare al cittadino la conoscenza dell'efficienza energetica dell'immobile.

Nel 2007 con il dlgs 29 dicembre 2006 n. 311 venne introdotto in via transitoria l'attestato di qualificazione energetica *(AQE)*, sostituito in seguito dall' *ACE*, avente come parametro principale l'*indice di prestazione energetica per la climatizzazione invernale (EPi)* [16].

Due anni dopo, con il D.P.R n.59/2009 si definirono le metodologie e i requisiti minimi impiantistici riguardanti la climatizzazione invernale ed estiva, la produzione di acqua calda sanitaria e l'illuminazione artificiale negli edifici non residenziali. Il D.P.R individuava nell'UNI/TS 11300 (parte uno e due), le norme tecniche nazionali per il calcolo delle prestazioni energetiche degli edifici. Successivamente con il DM 26 giugno 2009 vennero definite le linee guida nazionali per la redazione dell'*Attestato di Certificazione Energetica (ACE)*

Nel 2013 con il D.L n. 63/2013 (Decreto Eco-bonus/Energia) l'*Attestato di Certificazione Energetica (ACE)* fu sostituito dall'*Attestato di Prestazione Energetica (APE)*, il quale recepì la Direttiva 2010/31/UE, più aggiornata rispetto alla 2002/91/CE. Il quadro normativo fu aggiornato in seguito con ulteriori tre decreti interministeriali del 26 giugno 2015: decreto requisiti minimi, il decreto relazione tecnica di progetto e le nuove linee guida del nuovo APE. Queste ultime in particolare sancirono la nascita del *Sistema Informativo sugli Attestati di Prestazione Energetica (SIAPE)*, realizzato e gestito da *ENEA* a livello nazionale, al fine di raccogliere in un'unica banca dati centrale tutti i vari APE di edifici e unità immobiliari presenti nei Catasti Regionali [17].

In tempi più recenti sono stati emanati il D.L. 10 giugno 2020 n. 48 [18], che costituisce l'attuazione della direttiva europea 2018/844/UE la quale andava a modificare la direttiva precedente 2010/31/UE sui temi relativi alla prestazione energetica nell'edilizia e la direttiva 2012/27/UE per quel che concerne l'efficienza energetica. I principali obiettivi indicati dalla direttiva sono: rendere più efficaci le strategie di ristrutturazione degli immobili a lungo termine, al fine di ottenere un parco edilizio decarbonizzato entro il 2050; promuovere gli investimenti privati per il recupero del patrimonio edilizio esistente; definire le modalità di esercizio, conduzione, controllo, ispezione e manutenzione degli impianti termici per la climatizzazione invernale ed estiva e per la preparazione dell'acqua calda sanitaria; migliorare la trasparenza delle metodologie di calcolo della prestazione energetica definite dagli stati membri.

Il piano normativo italiano riguardante la certificazione energetica si conclude ad

oggi con il D.L. 8 novembre 2021 n. 199, attuazione della direttiva *RED II*, ovvero la n. 2001 del 2018, sulla promozione dell'uso dell'energia da fonti rinnovabili, con l'obiettivo di accelerare il percorso di crescita sostenibile del paese recando disposizioni in materia di energia da fonti rinnovabili. [19]

1.2 Dataset utilizzati

Gli attestati energetici analizzati nel lavoro di tesi sono relativi alla Regione Piemonte. Essi sono stati rilasciati dal *CSI-Piemonte (Consorzio Sistema Informativo)* in formato open data, con la possibilità di essere liberamente consultati e fruibili da qualsiasi individuo o ente informativo. In questo modo è possibile disporre delle stesse informazioni richieste al certificatore durante la compilazione online dell'attestato.
La disponibilità di tali dati è stata resa possibile grazie al lavoro congiunto della Regione con il CSI-Piemonte, che nel 2015 hanno realizzato il *Sistema Informativo per la Prestazione Energetica degli Edifici (SIPEE)* relativo agli edifici e unità immobiliare del Piemonte. I dati raccolti a livello regionale riguardanti l'ultimo anno trascorso vengono poi inviati entro il 31 marzo di ogni anno ad *ENEA* per alimentare il *SIAPE*.

Gli attestati analizzati sono stati reperiti dal portale nazionale della pubblica amministrazione[1] e sul sito web della Regione Piemonte[2]. I dati sono suddivisi in più dataset scaricabili in formato *csv*, ognuno caratterizzante un determinato dettaglio informativo dell'APE. Nella Tabella 1.1 a pagina 10 sono indicati i dataset utilizzati con una breve descrizione del loro contenuto.

1.3 Le Classi Energetiche

L'*APE* è un documento redatto da un professionista abilitato che descrive le caratteristiche energetiche di un immobile, sintetizzando le informazioni relative al metodo di costruzione e ai consumi energetici attraverso una scala di performance composta da dieci valori alfanumerici (A4, A3, ecc. fino alla G), chiamata *classe energetica*. La classe A4 rappresenta gli edifici aventi le prestazioni energetiche

[1]Dati pubblica amministrazione: https://www.dati.gov.it/

[2]Dati Regione Piemonte: https://www.dati.piemonte.it/

Dettaglio attestato di prestazione energetica (APE)	**Descrizione contenuto**
sezione dati generali	• Informazioni relative alle motivazioni di richiesta dell'APE • Geolocalizzazione dell'edifici • Dati catastali • Caratteristiche geometriche dell'edifici
sezione consumi	• Informazioni contenute nei dati generali • Informazioni relative alla sezione consumi
sezione dati energetici	• Informazioni contenute nei dati generali • Classe Energetica • Indici di Energia Primaria
sezione impianti	• Informazioni contenute nei dati generali • Informazioni relative agli impianti presenti negli edifici certificati
sezione raccomandazioni	• Informazioni contenute nei dati generali • Codice REN relativo alla tipologia di intervento
dati tecnici aggiuntivi	• Proprietà termofisiche dell'edificio • Gradi giorno
dati tecnici edificio reale	• Fabbisogno riscaldamento involucro edilizio • Rendimenti dei sottosistemi
dati tecnici edificio di riferimento	• Fabbisogni di riscaldamento e raffrescamento dell'involucro edilizio • Energie Primarie di riferimento • Rendimenti di riferimento

Tabella 1.1: Dataset disponibili nei cataloghi data.gov e dati.piemonte

migliori, indice energetico e consumi bassi, mentre la classe G identifica gli edifici caratterizzati da un indice energetico e consumi elevati, con performance peggiori. Per questi motivi la classe energetica rappresenta uno dei parametri maggiormente significativi nella determinazione del valore commerciale di un immobile, sia per quanto riguarda il valore di vendita sia per i consumi annui relativi ai servizi energetici utilizzati. In questo panorama l'*APE* costituisce uno strumento fondamentale per una rapida e trasparente valutazione della convenienza economica all'acquisto e alla locazione di un immobile, o all'attuazione di interventi di riqualificazione energetica dello stesso [20].

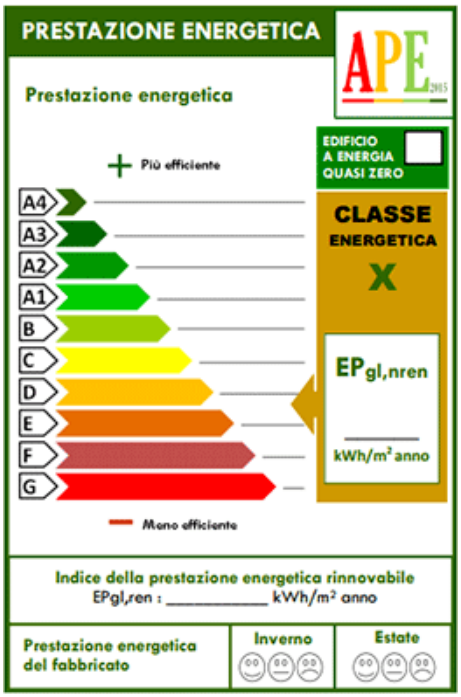


Figura 1.1: Fac-simile prestazione energetica globale dell'edificio APE

Dal Decreto requisiti minimi del 2015 con l'introduzione del certificato energetico *APE*, il metodo di calcolo per l'attribuzione della classe energetica è basato sulla definizione dell'*Indice di prestazione energetica globale non rinnovabile* ($Ep_{gl,nren}$), parametro che esprime la quota non rinnovabile del consumo totale di energia primaria per la climatizzazione, in regime continuo degli impianti, in relazione all'unità di superficie utile climatizzata. Esso si determina dalla somma degli indici di prestazione non rinnovabile dei singoli servizi energetici dell'edificio in esame, esprimibile con l'equazione:

$$Ep_{gl,nren} = Ep_{H,nren} + Ep_{C,nren} + Ep_{W,nren} + Ep_{V,nren} + Ep_{L,nren} + Ep_{T,nren} \quad (1.1)$$

Gli elementi presenti nell'equazione 1.1 sono rispettivamente: il fabbisogno di energia primaria non rinnovabile per la climatizzazione invernale ed estiva ($Ep_{H,nren}$ ed $Ep_{C,nren}$), per la produzione di acqua calda sanitaria ($Ep_{W,nren}$), per la ventilazione ($Ep_{V,nren}$) e, nel caso del settore non residenziale, per l'illuminazione artificiale ($Ep_{L,nren}$) e il trasporto di persone o cose ($Ep_{T,nren}$).
L'$Ep_{gl,nren}$ tiene dunque conto del rapporto tra l'energia necessaria per soddisfare i vari bisogni connessi a un uso standard dell'edificio, e la superficie netta calpestabile dell'ambiente stesso. Viene dunque espresso attraverso l'unità di misura del $\frac{kWh}{m^2 anno}$.

Per l'assegnazione della classe di prestazione energetica l'indice di prestazione ottenuto con il metodo di calcolo viene confrontato con l'*Indice di prestazione energetica globale non rinnovabile dell'edificio di riferimento* ($Ep_{gl,nren,rif,std}$), calcolato ipotizzando che l'edificio sia dotato di elementi caratterizzanti l'involucro edilizio e impianti standard ottemperanti i requisiti minimi previsti dalla legge. In base ai vincoli rappresentati in Tabella 1.2 viene determinata la classe energetica relativa all'edificio certificato.

Vincoli $Ep_{gl,nren}$	**Classe Energetica**
$Ep_{gl,nren} \leq 0.40 Ep_{gl,nren,rif,std}$	A4
$0.40 Ep_{gl,nren,rif,std} < Ep_{gl,nren} \leq 0.60 Ep_{gl,nren,rif,std}$	A3
$0.60 Ep_{gl,nren,rif,std} < Ep_{gl,nren} \leq 0.80 Ep_{gl,nren,rif,std}$	A2
$0.80 Ep_{gl,nren,rif,std} < Ep_{gl,nren} \leq 1.00 Ep_{gl,nren,rif,std}$	A1
$1.00 Ep_{gl,nren,rif,std} < Ep_{gl,nren} \leq 1.20 Ep_{gl,nren,rif,std}$	B
$1.20 Ep_{gl,nren,rif,std} < Ep_{gl,nren} \leq 1.50 Ep_{gl,nren,rif,std}$	C
$1.50 Ep_{gl,nren,rif,std} < Ep_{gl,nren} \leq 2.00 Ep_{gl,nren,rif,std}$	D
$2.00 Ep_{gl,nren,rif,std} < Ep_{gl,nren} \leq 2.60 Ep_{gl,nren,rif,std}$	E
$2.60 Ep_{gl,nren,rif,std} < Ep_{gl,nren} \leq 3.50 Ep_{gl,nren,rif,std}$	F
$3.50 Ep_{gl,nren,rif,std} < Ep_{gl,nren}$	G

Tabella 1.2: Scala di classificazione degli edifici sulla base dell'indice di prestazione energetica globale non rinnovabile, secondo l'allegato 1 DM requisiti minimi n.162/2015

1.4 Sezioni Attestato di Prestazione Energetica

L'*APE* presenta un formato standard, valido su tutto il territorio italiano, articolato per fornire informazioni semplici e chiare sull'efficienza, le prestazioni e il fabbisogno energetico dell'immobile e dei servizi energetici. L'uniformità degli attributi costituenti il certificato permette di poter confrontare facilmente e in maniera diretta edifici localizzati in tutta la nazione. Il certificato è composto da diverse sezioni fondamentali contenenti i dati generali dell'immobile, la prestazione energetica globale del fabbricato e degli impianti con i relativi consumi stimati, le raccomandazioni per eventuali riqualificazioni energetiche, dati di dettaglio del fabbricato e degli impianti e infine informazioni riguardanti il certificatore e il software utilizzato per la redazione dell'attestato. Tutti i dati inseribili nell'*APE* sono riportati nei dataset forniti dal *Sistema Informativo*.

1.4.1 Sezione Dati Generali

Corrispondono alla prima sezione presente nell'*APE*, e forniscono informazioni riguardanti:

- *Destinazione d'uso*: indica la finalità di utilizzo dell'edificio, viene effettuata una distinzione a monte fra edificio residenziale e non residenziale e successivamente viene inserita la classificazione in base al D.P.R. 412/93 rappresentata in Tabella 1.3 a pagina 14;
- *Oggetto dell'attestato*: definisce se l'edificio in questione è un'unità immobiliare, un intero edificio o gruppo di unità immobiliari;
- *Motivazione del rilascio*: si identifica la motivazione per cui è stato redatto l'*APE*, le più frequenti sono per passaggi di proprietà o contratti di locazione;
- *Dati identificativi*: sono contenuti gli attributi che danno informazioni riguardo la locazione geografica dell'edificio: Regione, Comune, indirizzo, piano, interno e coordinate GIS dell'immobile in questione. A queste vengono aggiunti i dati catastali (sezione, foglio, particella e subalterni), in modo da poter individuare univocamente l'edificio.
  Infine vi sono indicate la zona climatica del comune dove sorge l'edificio e ulteriori dati del fabbricato (anno di costruzione, superficie utile riscaldata $[m^2]$, superficie utile raffrescata $[m^2]$, volume lordo riscaldato $[m^3]$ e volume lordo raffrescato $[m^3]$);

- *Servizi energetici presenti*: in questa parte vengono indicati i servizi energetici presenti nel calcolo della prestazione energetica nell'equazione 1.1.

Categoria destinazione d'uso	**Descrizione**
E.1	Edifici adibiti a residenza e assimilabili:
E.1 (1)	abitazioni adibite a residenza con carattere continuativo, quali abitazioni civili e rurali, collegi, conventi, case di pena, caserme;
E.1 (2)	abitazioni adibite a residenza con occupazione saltuaria, quali case per vacanze, fine settimana e simili;
E.1 (3)	edifici adibiti ad albergo, pensione ed attività similari;
E.2	Edifici adibiti a uffici e assimilabili: pubblici o privati, indipendenti o contigui a costruzioni adibite anche ad attività industriali o artigianali, purché siano da tali costruzioni scorporabili agli effetti dell'isolamento termico;
E.3	Edifici adibiti a ospedali, cliniche o case di cura e assimilabili ivi compresi quelli adibiti a ricovero o cura di minori o anziani nonché le strutture protette per l'assistenza ed il recupero dei tossico-dipendenti e di altri soggetti affidati a servizi sociali pubblici;
E.4	Edifici adibiti ad attività ricreative, associative o di culto e assimilabili:
E.4 (1)	quali cinema e teatri, sale di riunione per congressi;
E.4 (2)	quali mostre, musei e biblioteche, luoghi di culto;
E.4 (3)	quali bar, ristoranti, sale da ballo;
E.5	Edifici adibiti ad attività commerciali e assimilabili: quali negozi, magazzini di vendita all'ingrosso o al minuto, supermercati, esposizioni;
E.6	Edifici adibiti ad attività sportive:
E.6 (1)	piscine, saune e assimilabili;
E.6 (2)	palestre e assimilabili;
E.6 (3)	servizi di supporto alle attività sportive;
E.7	Edifici adibiti ad attività scolastiche a tutti i livelli e assimilabili;
E.8	Edifici adibiti ad attività industriali ed artigianali e assimilabili.

Tabella 1.3: Categorie destinazione d'uso secondo il D.P.R 412/93 [1]

1.4.2 Sezione Prestazione Energetica

In questa sezione sono indicate:

- *Prestazione energetica del fabbricato*: viene indicata la prestazione energetica dell'involucro al netto del rendimento degli impianti presenti.
  Per quanto riguarda la prestazione energetica *invernale* dell'involucro, l'indicatore è definito a partire dal valore dell'indice di prestazione termica utile per il riscaldamento dell'edificio di riferimento ($EP_{H,nd,limite}$), calcolato secondo quanto previsto dal decreto requisiti minimi, ipotizzando, che in esso siano installati elementi edilizi dotati dei requisiti minimi di legge in vigore. Tramite questo valore vengono valutati tre fasce di qualità dell'involucro edilizio, bassa, media ed alta.
  Per la prestazione energetica *estiva* dell'involucro per definire l'indicatore sono necessari la trasmittanza termica periodica Y_{IE} e all'area solare equivalente estiva per unità di superficie utile $A_{sol,est}/A_{sol_utile}$ definite anch'esse nel decreto requisiti minimi.

- *Prestazione energetica globale*: in questa sezione (Figura 1.1) devono essere riportati l'indice di prestazione energetica globale non rinnovabile (equazione 1.1) e la classe energetica; inoltre sarà necessario indicare se l'edificio e ad energia quasi zero, cioè se rispetta tutti i requisiti minimi vigenti per legge e se utlizza una determinata quantità di fonti rinnovabili;

- *Prestazione energetica degli impianti e consumi stimati*: vengono indicate le tipologie e le quantità annue consumate in uso standard dei vettori energetici utilizzati. Inoltre vengono indicate le prestazioni energetiche dell'edificio espresse attraverso gli indici di prestazione energetica globale rinnovabile e non rinnovabile ($\frac{kWh}{m^2 anno}$) e le emissioni di CO_2 ($\frac{kg}{m^2 anno}$). Tutti gli indici presentano al denominatore la superficie utile, definita come la superficie netta calpestabile dei volumi interessati dalla climatizzazione; altro non è che l'*unione* delle superfici riscaldate e raffrescate dell'edificio.

1.4.3 Sezione Raccomandazioni

Sono riportati gli interventi consigliati e la stima dei risultati conseguibili, classe energetica e indice di prestazione, con l'applicazione del singolo intervento o con la totalità di essi. Con ciò è possibile valutare un ipotetico potenziale di miglioramento dell'edificio oggetto dell'attestato di prestazione.

1.4.4 Sezione Dati di dettaglio

In questa parte sono riportati altri dati energetici, come l'*energia esportata*, che da normativa può essere soltanto elettrica, definita come l'eccedenza di energia prodotta in loco rispetto al fabbisogno mensile, che non concorre alla prestazione energetica dell'edificio. Seguono altri dati relativi al fabbricato:

- *Volume riscaldato* $[m^3]$: volume lordo delle zone climatizzate dell'edificio, definito dalle superfici delimitanti;
- *Superficie disperdente* $[m^2]$: superficie delimitante il volume riscaldato rispetto all'ambiente esterno, al terreno, e ad altri locali non climatizzati o a diversa temperatura;
- *Rapporto S/V* $[m^{-1}]$: rapporto di forma tra la superficie disperdente e il volume riscaldato precedentemente definiti;
- *Indice di prestazione termica utile per il riscaldamento*$[\frac{kWh}{m^2 anno}]$: espresso come $Ep_{H,nd}$, definisce il fabbisogno termico invernale dell'involucro edilizio per mantenere costante la temperatura di setpoint dell'ambiente riscaldato;

Infine la sezione si conclude con i dati di dettaglio degli impianti installati, suddivisi per servizio energetico fornito. Le informazioni più rilevanti riguardano gli indici di prestazione rinnovabili e non e soprattutto i rendimenti di tali sistemi, poiché da essi si identifica il loro livello di efficienza. I rendimenti indicati riguardano i servizi energetici inerenti alla produzione di acqua calda sanitaria, climatizzazione estiva e invernale. Su quest'ultima viene posto particolare interesse, scorporando il rendimento medio stagionale per il riscaldamento in quattro quote:

- *Rendimento del sottosistema di distribuzione*: tiene conto dell'energia termica dissipata dal fluido termovettore durante il transito nel sistema di distribuzione;
- *Rendimento del sottosistema di emissione*: relativo a tutte le tipologie di terminali installati in ambiente, viene valutato come il rapporto fra il calore fornito dal sistema di emissione reale rispetto al calore richiesto dai locali con un sistema di emissione ideale in grado di mantenere una temperatura in ambiente uniforme;
- *Rendimento del sottosistema di generazione*: tipico del generatore di calore utilizzato, considera il rapporto fra l'energia effettivamente fornita in ambiente rispetto al calore utile prodotto;

- *Rendimento del sottosistema di regolazione*: riguarda il sistema dedicato al controllo della temperatura degli ambienti e serve a valutare le perdite energetiche relative a una non perfetta regolazione del calore richiesto in ambiente.

Infine le ultime pagine dell'*APE* sono dedicate alla raccolta di informazione relative soggetto certificatore, ai sopralluoghi effettuati e ai software utilizzati per la redazione dell'attestato.

Capitolo 2

Metodi e Tecniche di Analisi

Gli open data contengono una significativa quantità di informazioni, la cui comprensione esula spesso dalle abilità umane. Si dimostrano dunque necessari metodi e algoritmi afferenti al campo della statistica e del machine learnig, in grado di estrapolare la conoscenza da questi *big data*. In questo capitolo sono presentati i metodi relativi alla visualizzazione, l'analisi cluster e l'analisi predittiva utilizzati nel framewok metodologico adottato.

2.1 Tecniche di statistica descrittiva

L'estrazione della conoscenza dagli attestati energetici avviene anche e soprattutto attraverso una buona visualizzazione grafica, in grado di attribuire un significato alle informazioni in esso racchiuse. La visualizzazione oltre a facilitare la comprensione dei contenuti, organizzando i dati in maniera più comprensibile, permette di evidenziare i valori anomali, *outliers*, che pervadono i datasets, agevolandone l'identificazione e conseguente rimozione, al fine di ottenere dei dati utili e fruibili.

Boxplots

Il diagramma a scatola e baffi, meglio noto come *boxplot*, è un metodo grafico standard per la visualizzazione della distribuzione di un insieme di dati. Concettualmente viene rappresentato con un rettangolo nel quale vengono indicati il primo quartile Q_1, il secondo quartile Q_2 o mediana e il terzo quartile Q_3 della

distribuzione dei dati. Per quartile si intende l'unità statistica che suddivide la popolazione di dati in quattro parti contenenti lo stesso numero di elementi. Nel primo quartile Q_1 dunque, sono contenuti il 25% dei valori costituenti l'intera popolazione in esame. La mediana, come i quartili, rappresenta un indice di posizione della distribuzione ed individua il punto di mezzeria del numero di elementi; il 50% di essi si troverà a sinistra della mediana e il restante 50% alla sua destra.

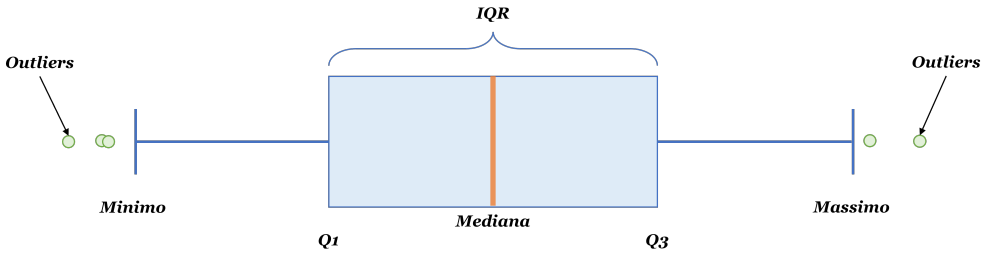


Figura 2.1: Esempio di un boxplot

Un altro indice rappresentativo dei boxplot è il range interquartile $IQR = Q_3 - Q_1$, all'interno del quale sono contenuti i valori più frequenti della popolazione, pari al 50 % degli elementi totali. Inoltre viene utilizzato come parametro per definire i valori soglia di massimo e minimo per l'accettabilità di un elemento nella sua distribuzione. All'interno del boxplot tali limiti di massimo e minimo vengono generalmente valutati rispettivamente come $Q_3 + 1.5 \cdot IQR$ e $Q_1 - 1.5 \cdot IQR$ e corrispondono ai baffi del diagramma. I valori all'esterno di tale intervallo vengono considerati come *outliers*, valori poco probabili che si presentano con poca frequenza all'interno del dataset e dunque da considerarsi anomali o poco significativi(Figura 2.1).

Funzione di densità di probabilità

Il boxplot appena descritto consente una rappresentazione semplice del concetto di distribuzione della probabilità, la quale viene espressa in maniera più completa dalla funzione di densità di probabilità (*Probability Density Function: PDF*). Molti attributi analizzati in seguito verranno visualizzati attraverso questo strumento, ed è opportuno evidenziarne le caratteristiche fondamentali. La *PDF* rappresenta la probabilità che una determinata variabile continua X assuma un determinato valore x, all'interno di un determinato intervallo.

Matematicamente, data la variabile casuale continua X che assume valori nell'intervallo (a,b) compresi fra più e meno infinito, la funzione di densità di probabilità è la funzione $f_X : \Re \to \Re$ che ad ogni elemento reale associa il limite per dx che

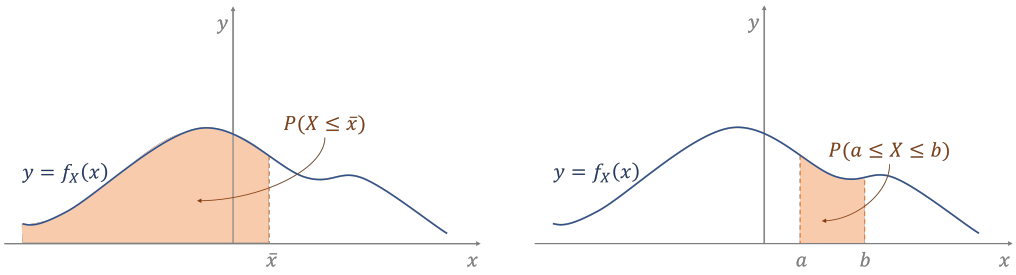
tende a 0, del rapporto tra la probabilità che la variabile casuale assuma valori nell'intervallo $[x, x + dx]$ e l'ampiezza dx [21].

$$f_X : \Re \to \Re : x \to \lim_{dx \to 0} \left[\frac{P(x < X \leq x + dx)}{dx} \right] \tag{2.1}$$

La funzione di densità in x, allora, rappresenta quanto vale la probabilità nell'intorno di x, rapportata all'ampiezza dell'intorno stesso, per questo viene evocato il concetto di *densità*. La probabilità che una variabile aleatoria X assuma valori minori o uguali di un valore reale $\bar{x}$, o compresi fra un determinato intervallo [a,b] è rispettivamente:

$$P(X \leq \bar{x}) = \int_{-\infty}^{\bar{x}} f_X(x)dx \tag{2.2}$$

$$P(a \leq X \leq b) = \int_{a}^{b} f_X(x)dx \tag{2.3}$$



(a) *equazione 2.2* (b) *equazione 2.3*

Figura 2.2: Probabilità definita come l'area sottesa alla funzione densità di probabilità

La *PDF* rispetta le seguenti proprietà:

- Poiché la probabilità non può avere un valore negativo, la funzione $f_X(x) \geq 0$ assumerà sempre valori positivi.
- L'area totale sottesa alla funzione è pari a 1 ovvero alla probabilità legata all'accadimento dell'evento certo.

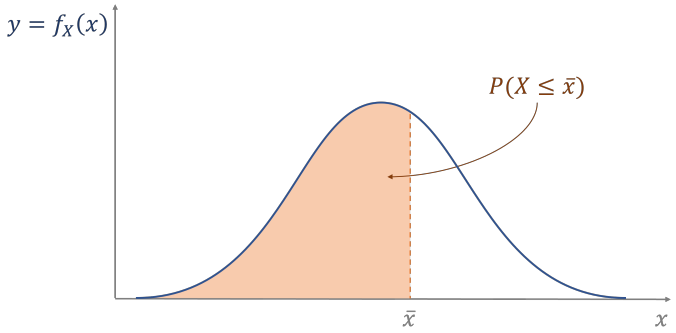
$$\int_{-\infty}^{\infty} f_X(x)dx = 1$$

- La probabilità che la variabile casuale continua X assuma un determinato valore puntuale del dominio di esistenza è nulla. Ciò è dovuto al fatto che un singolo valore corrisponde ad un intervallo di ampiezza zero, quindi la corrispondente area è anch'essa zero. Ciò implica che non ha influenza l'inclusione, nel calcolo della probabilità, degli estremi dell'intervallo.

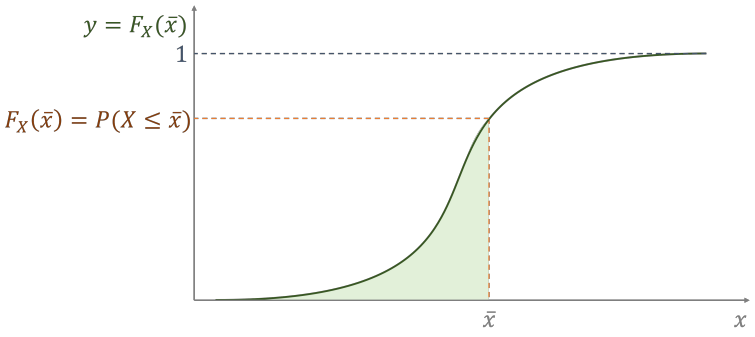
$$P(a \leq X \leq b) = P(a < X \leq b) = P(a \leq X < b) = P(a < X < b)$$

Funzione di probabilità cumulata

Nelle analisi successive in alcuni casi l'attenzione non verterà sulla probabilità che un determinato attributo assuma uno specifico valore, ma sulla probabilità che esso sia di entità minore o uguale a un dato valore, caso rappresentato dall'equazione 2.2 e dalla Figura 2.2a.



(a) *PDF gaussiana*



(b) *CDF gaussiana*

Figura 2.3: Funzione di probabilità cumulata in relazione a una funzione di densità gaussiana

Data una variabile casuale X, la funzione che fa corrispondere ai valori di x, le probabilità cumulate $P(X \leq \bar{x})$ viene detta funzione di probabilità cumulata (*Cumulative Density Function: CDF*) definita come:

$$F_X(\bar{x}) = P(X \leq \bar{x}) = \int_{-\infty}^{\bar{x}} f_X(x)dx \tag{2.4}$$

Essendo una rappresentazione diretta della probabilità $F_X(\bar{x}) \in [0,1]$ ed è monotona non decrescente.

In Figura 2.3 si osserva l'andamento di una funzione densità di probabilità gaussiana o normale, che descrive una distribuzione di probabilità continua spesso usata come prima approssimazione per esprimere l'andamento di variabili casuali a valori reali a concentrarsi nell'intorno di uno specifico valore medio, e la sua funzione di probabilità cumulata relativa. Il vantaggio della *CDF* è quello di esprimere in maniera diretta sull'asse delle ordinate la probabilità che la variabile casuale X sia minore o uguale a un determinato valore $\bar{x}$, rappresentando di fatto l'andamento della funzione integrale della *PDF*.

2.2 Analisi di Clustering

L'analisi cluster, o analisi dei raggruppamenti, è uno dei metodi di data mining più importanti per la scoperta della conoscenza in un dataset multivariato. Lo scopo fondamentale del clustering è identificare dei pattern, dei profili, o dei raggruppamenti di elementi simili all'interno del dataset analizzato. In letteratura fa riferimento al concetto di *algoritmo di apprendimento non supervisionato*, così definito perché non viene "guidato" nel raggiungere il suo scopo, non viene definita una relazione aprioristica fra input ed output al quale l'algoritmo deve fare riferimento. Al modello viene fornito solamente un insieme di valori di input, che verranno classificati e organizzati in base alle loro caratteristiche comuni, al fine di effettuare previsioni sulle osservazioni future.

2.2.1 Concetto di distanza nell'analisi di clustering

La similarità fra gli elementi dell'insieme viene determinata attraverso il concetto di distanza multidimensionale: un elemento è tanto più simile ad un altro, quanto più esso è vicino all'altro nello spazio n-dimensionale. L'idealizzazione del concetto di spazio per l'essere umano è fortemente correlata al numero delle dimensioni a

cui esso fa riferimento. Se nell'immagine in Figura 2.4, rappresentante uno spazio a due dimensioni, è facile stabilire che il cerchio sia più vicino al quadrato che al triangolo, quest'astrazione diventa molto difficile se fossero stati considerati degli spazi con dimensioni superiori a tre, ma il concetto di distanza rimane immutato.

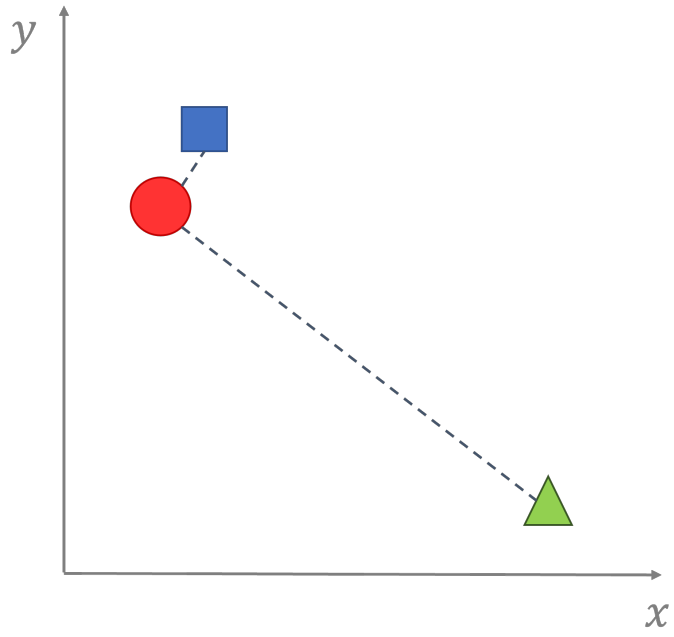


Figura 2.4: Rappresentazione della distanza in uno spazio bi-dimensionale

Esistono diversi metodi per stabilire la distanza fra due elementi in uno spazio n-dimensionale, nell'analisi cluster le tipologie più utilizzate sono la *distanza Euclidea* e la *distanza di Manhattan*

1. Distanza Euclidea: è il concetto di distanza più semplice e più utilizzato, definisce la lunghezza del segmento congiungente due punti nello spazio euclideo ed è espressa dalla formula

$$d_{euc}(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

2. Distanza Manhattan: la distanza fra due punti nello spazio viene valutata come la somma del valore assoluto delle differenze delle loro coordinate

$$d_{man}(x, y) = \sum_{i=1}^{n} |(x_i - y_i)|$$

Dove x e y sono due vettori di lunghezza n. In Figura 2.5 sono visualizzate graficamente le due distanze.

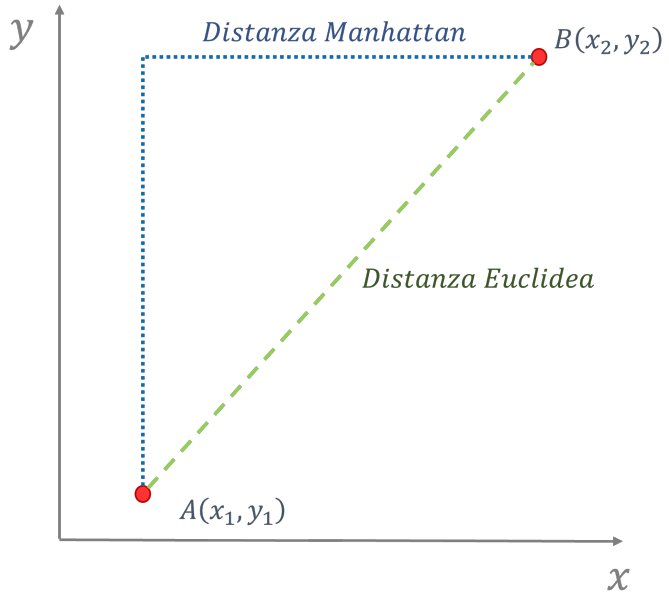


Figura 2.5: Rappresentazione distanza Euclidea e Manhattan

Tecniche di normalizzazione

I dataset utilizzati presentano diverse tipologie di informazioni, espresse in scale diverse fra loro; basti pensare ai valori dei rendimenti in forma decimale e all'anno di costruzione che assume valori superiori alle migliaia. Poiché la distanza ricopre un ruolo fondamentale nell'analisi cluster, al fine di evitare delle interpretazioni errate delle informazioni racchiuse nei dati, dovute all'ampiezza del dato analizzato piuttosto che al suo peso nella caratterizzazione dell'insieme, è opportuno effettuare una normalizzazione. Esistono diverse tipologie di normalizzazione:

- *Normalizzazione max-min*: è la tipologia più semplice, noti i valori di massimo e minimo del dataset, consente di scalare e normalizzare i dati in un intervallo compreso fra 0 e 1

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

- *Z-score*: ridimensiona la distribuzione degli attributi in modo che essi assumano una media nulla e deviazione standard unitaria

$$z - score = \frac{x - \mu}{\sigma}$$

dove μ e σ sono rispettivamente la media e la deviazione standard degli attributi. Questo tipo di normalizzazione non comprende dei limiti di esistenza superiori o inferiori, ma permette di visualizzare di quante deviazioni standard il valore x si discosta dal suo valor medio. Non è dunque fortemente influenzata dalla presenza degli outlier come la normalizzazione max-min, ed anzi può essere utilizzata per evidenziare eventuali valori anomali. Poiché in una funzione normale il 99.7% degli elementi di un campione si concentra nel range $[-3\sigma, +3\sigma]$, tutti i valori all'infuori di questo intervallo possono essere considerati degli outliers.

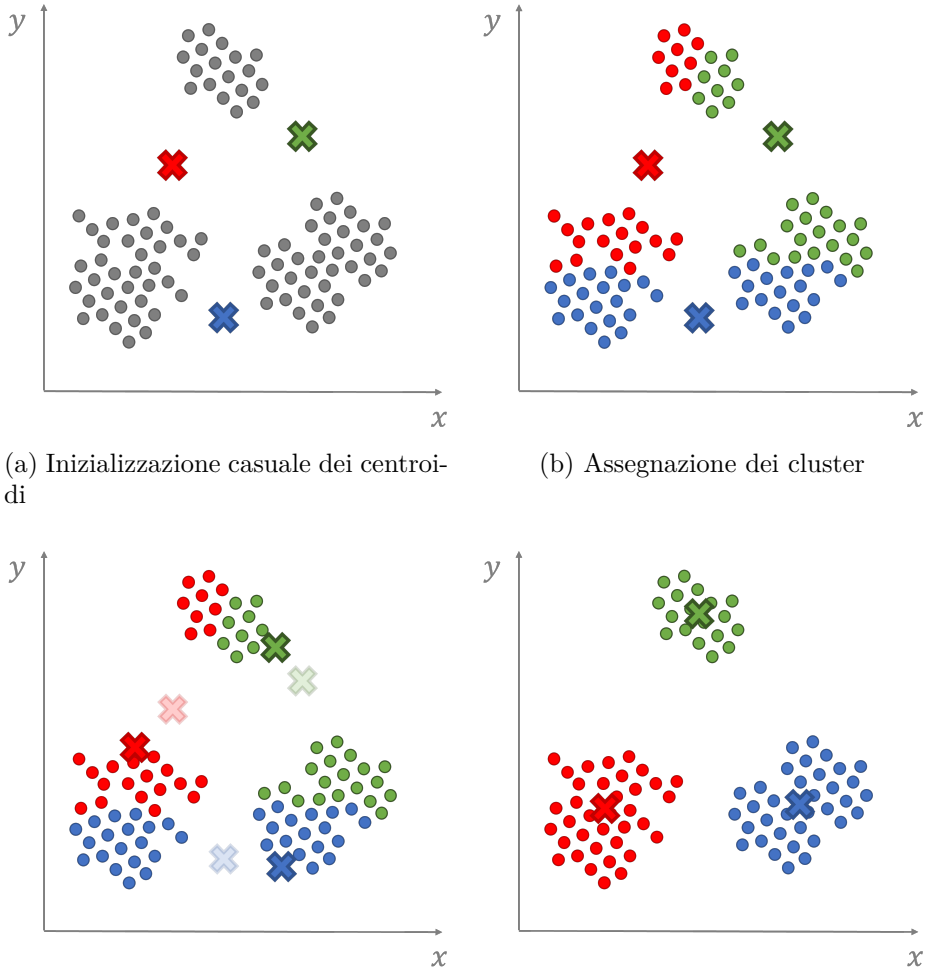
2.2.2 Tecniche di clustering partitivo

Questi algoritmi suddividono gli elementi presenti in un insieme di dati in un numero prestabilito di k ripartizioni, ciascuna di esse rappresentante un cluster. Ogni cluster è caratterizzato da un centroide, definito ad esempio come la media aritmetica dei valori assunti dagli elementi del cluster nello spazio n-dimensionale. La partizione dell'insieme di elementi viene svolta al fine di ottimizzare un determinato criterio, come ad esempio la similarità degli oggetti in funzione della loro distanza; gli oggetti presenti all'interno della stessa partizione sono simili fra loro (alta similarità intra-cluster) e diversi rispetto agli oggetti contenuti negli altri raggruppamenti (bassa similarità inter-cluster) [22], [23].

K-means

Il K-means rappresenta l'algoritmo di clustering partizionale più semplice e utilizzato. Il primo passaggio consiste nell'imporre un determinato valore al parametro k, stabilendo così a priori quante partizioni del dataset si vogliono ottenere. Vengono così selezionati k elementi casuali all'interno del dataset come elementi rappresentativi del *k-esimo* cluster. I restanti elementi vengono dunque assegnati al centroide a loro più vicino all'interno dello spazio n-dimensionale, utilizzando il concetto di distanza (2.2.1) fra l'oggetto e la media aritmetica del cluster . Dopo questo step l'algoritmo valuta la nuova distanza media fra gli elementi del cluster definendo così dei nuovi centroidi e si rieseguono i passaggi sopra esposti. L'algoritmo termina quando la somma delle distanze intra-cluster è ridotta la minimo,

ovvero quando nessuno oggetto cambia più cluster e i centroidi non cambiano più la loro posizione (condizione di convergenza), oppure quando viene raggiunto un numero massimo di iterazioni.



(a) Inizializzazione casuale dei centroidi

(b) Assegnazione dei cluster

(c) Ricalcolo e spostamendo dei centroidi

(d) Riassegnazione dei cluster e convergenza

Figura 2.6: Passaggi algoritmo k-means

Partitioning Around Medioids (PAM)

L'algoritmo *PAM (Partitioning Around Medioids)* è un algoritmo di clustering partizionale che introduce il concetto di medioide, definito come l'oggetto più significativo della partizione. La differenza rispetto al centroide, definito come la

media fra i vari elementi del cluster, consiste nel fatto che il medioide coincide con un oggetto realmente presente nell'insieme. Il procedimento operativo del *PAM* è poi analogo a quello del *k-means*, utilizzando il medioide come elemento di riferimento invece del centroide. Questo semplice cambiamento irrobustisce l'algoritmo, rendendolo meno sensibile alla presenza degli outliers. Il problema principale del *PAM*, come anche del *k-means* è la scelta predeterminata del numero di cluster.

Clustering Large Applications (CLARA)

CLARA, acronimo di *Clustering Large Applications*, è un algoritmo del tipo k-medioide come il *PAM*, utilizzato tipicamente in dataset contenenti un elevato numero di istanze. La differenza principale rispetto agli altri due algoritmi precedentemente esposti è che invece di fissare a priori il numero k di cluster in cui suddividere il dataset, *CLARA* campiona un sottogruppo di elementi, dalla grandezza definita, a cui applica l'algoritmo *PAM*, per generare un set ideale di medioidi relativi a quel sottogruppo. Per valutare la qualità dei mediodi ottenuti viene definita una funzione obiettivo in grado di calcolare la dissimilarità media fra ogni oggetto nell'intero dataset e il medioide del suo cluster. L'algoritmo ripete il campionamento un determinato numero di volte fino ad ottenere la configurazione di medioidi che minimizza la funzione obiettivo.

2.2.3 Tecniche di clustering gerarchico

Le metodologie di clustering gerarchico costituiscono un approccio alternativo al clustering partizionale per raggruppare oggetti con caratteristiche simili,con il vantaggio di non necessitare di un numero prefissato di raggruppamenti per essere utilizzato. Il funzionamento alla base degli algoritmi gerarchici consiste nel dividere o aggregare gli elementi del dataset in una sequenza di partizioni innestate [23]. La gerarchia delle partizioni può essere *agglomerativa (bottom-up)* o *divisiva (top-down)*:

- *Clustering Agglomerativo*: ogni oggetto del dataset viene considerato appartenente a un cluster costituito inizialmente solo da se stesso *(singleton o foglia)*. Successivamente i cluster più simili vengono agglomerati in un solo cluster in maniera iterativa, fino a formare un unico grande cluster contenente tutti gli elementi *(radice)*.

- *Clustering Divisivo*: inversamente al caso precedente, inizialmente si considera un solo grande cluster in cui sono accorpati tutti gli elementi del dataset. In seeguito i cluster più eterogeni vengono scorporati fino a quando non si ottiene la configurazione dove esistono solamente cluster foglia contenenti un solo elemento.

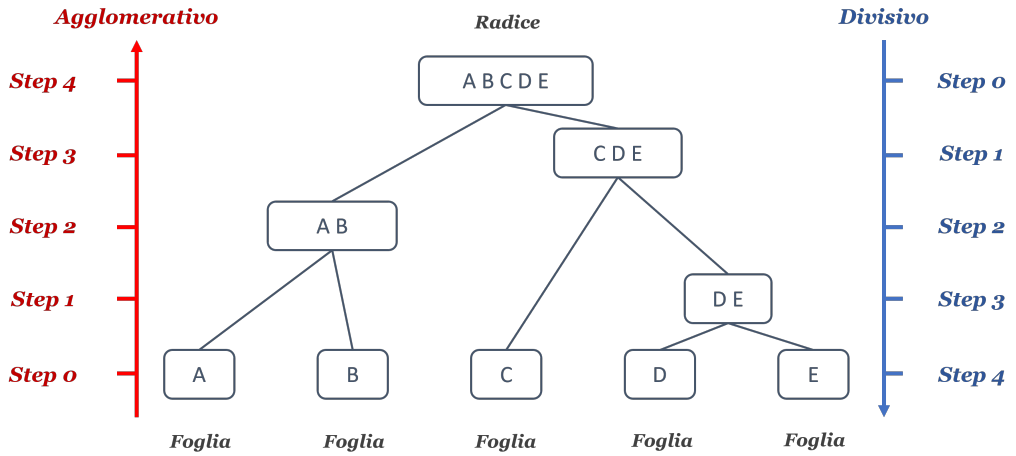
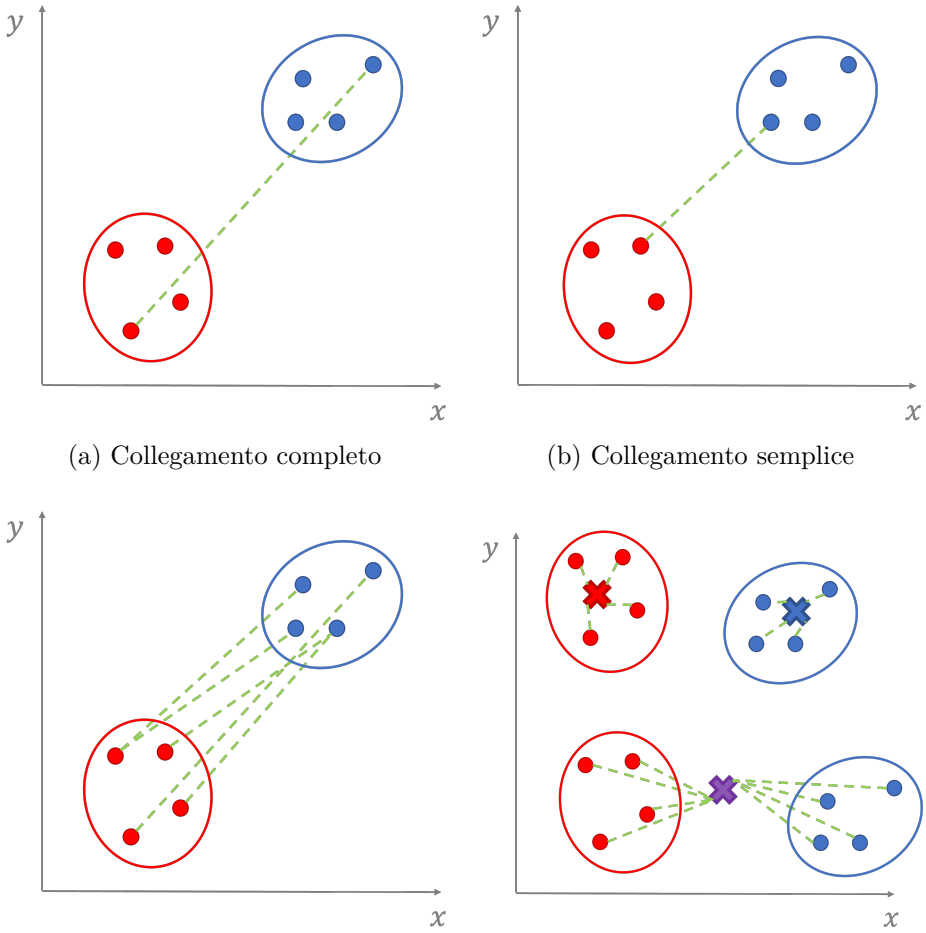


Figura 2.7: Dendogramma clustering gerarchico

Il clustering gerarchico assume di fatto una configurazione ad albero binario chiamata *dendogramma*, Figura 2.7. Seppur tale configurazione consenta di poter tagliare l'albero al livello di profondità desiderato e analizzare i cluster che vi si sono formati, l'algoritmo gerarchico non dice a priori quanti cluster significativi ci sono, o qual è il punto corretto per effettuare il taglio. L'unione o la divisione dei cluster viene eseguita in base a determinati criteri di similarità, che rappresentano di fatto metodi diversi per misurare le distanze e vengono chiamati *linkage methods*. Per semplicità verrà considerato il caso di algoritmi agglomerativi per la trattazione, maggiormente utilizzati rispetto agli algoritmi divisivi poiché presentano una complessità minore.

Metodi di linkage negli algoritmi di clustering gerarchico

Le funzioni di collegamento utilizzano la distanza, tipicamente Euclidea, fra gli oggetti dell'insieme per raggruppare coppie di oggetti in base alla loro similarità, in maniera iterativa, fino a formare la struttura gerarchica ad albero. Dunque, se la distanza è il metro di giudizio con il quale vengono accorpati gli oggetti, i *linkage methods* stabiliscono dove e fra quali elementi tale distanza deve essere valutata. I metodi più comuni sono:



(a) Collegamento completo

(b) Collegamento semplice

(c) Collegamento medio

(d) Collegamento Ward

Figura 2.8: Linkage Methods

a) *Collegamento completo o massimo*: la distanza fra due cluster è valutata come la massima distanza fra un elemento di un cluster rispetto a quello appartenente a un altro cluster e tende a generare cluster compatti;

b) *Collegamento singolo o minimo*: la distanza fra due cluster è valutata come la minima distanza fra un elemento di un cluster rispetto a quello appartenente a un altro cluster e tende a generare lunghe catene di cluster;

c) *Collegamento medio*: La distanza fra due cluster è valutata come la distanza media fra tutti gli elementi di un cluster rispetto a tutti gli elementi di un altro cluster, questo metodo è meno sensibile alla presenza degli outliers;

d) *Collegamento Ward*: vengono accorpati i cluster per i quali l'aumento della varianza della distanza intra-cluster è la minore possibile, valutando come distanza la somma dei quadrati delle distanze fra gli elementi.

2.2.4 Tecniche di identificazione del numero ottimale di cluster

Dalla descrizione dei metodi di cluster descritti nella sezione precedente è emerso come il punto più critico dell'analisi sia proprio la determinazione del numero corretto di raggruppamenti. Se alla base del concetto di cluster vi è la generazione di un raggruppamento di elementi che presentano una forte similarità intra-cluster e bassa similarità inter-cluster, purtroppo non vi è un metodo inequivocabile per poter determinare il numero ottimale di cluster che soddisfi questi requisiti. Ciò deriva dal fatto che vi è una forte dipendenza con il metodo utilizzato per stabilire la similarità e con i parametri sui quali è stata valutata la distanza. Esistono dunque diversi indici con cui valutare il clustering, in questa tesi sono stati adoperati due metodi diretti, chiamati così perché giungono al risultato ottimizzando una determinata funzione obiettivo, il *metodo del gomito* e della *silhouette* [22].

Una delle funzioni più utilizzate per stabilire la bontà di un analisi cluster è la *Total Sum of Square error: SST*, definita come:

$$SST = \sum_{k=1}^{k} W(C_k) = \sum_{k=1}^{k} \sum_{x_i \in C_k} (x_i - \mu_k)^2 \tag{2.5}$$

dove $W(C_k)$ rappresenta la varianza intra-cluster, x_i è un oggetto appartenente al *k-esimo* cluster C_k, μ_k è la media degli elementi assegnati al cluster C_k. Il numero ottimale di cluster che minimizza l'equazione 2.5 può essere determinato con il *metodo del gomito*. Tendenzialmente maggiore è il numero di cluster considerato e minore sarà *SST*, ma ciò comporta una perdita di significato dell'analisi stessa. Dunque, è opportuno effettuare un'analisi di sensitività al fine di individuare il numero di cluster per il quale l'aggiunta di un ulteriore raggruppamento non provochi una riduzione sostanziale dell'errore. Il *gomito* della curva iperbolica, viene determinato geometricamente come il punto per cui la distanza rispetto al segmento congiungente i due punti estremali della curva è massima (Figura 2.9). Il *metodo della silhouette* valuta la qualità di un cluster, determinando quanto un elemento sia simile rispetto al cluster in cui è stato assegnato e diverso rispetto agli altri. Il coefficiente di silhouette comprende valori fra $+1$ e -1, dove i valori positivi più alti indicano che l'oggetto è ben rappresentato dal suo cluster e poco

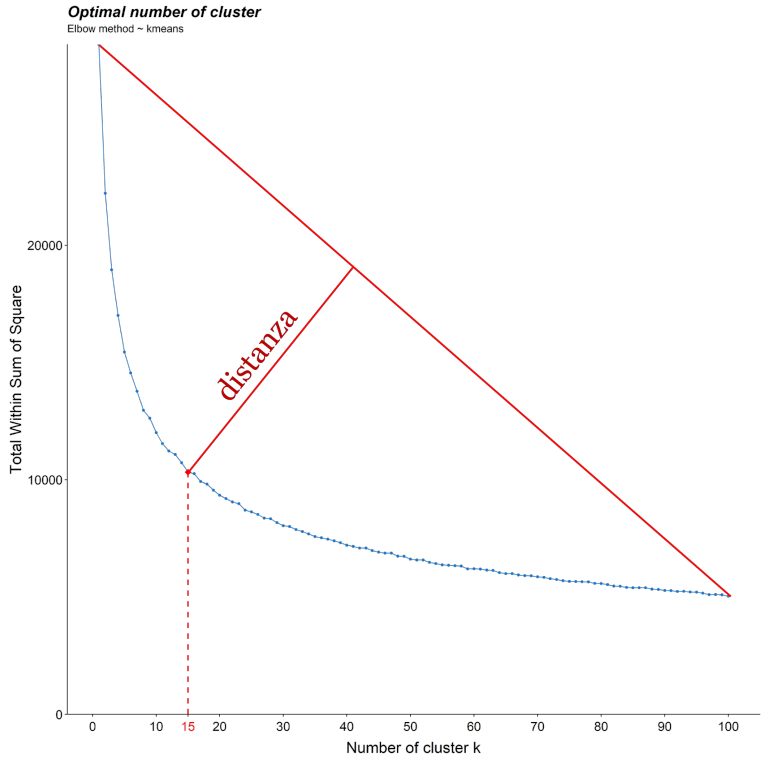


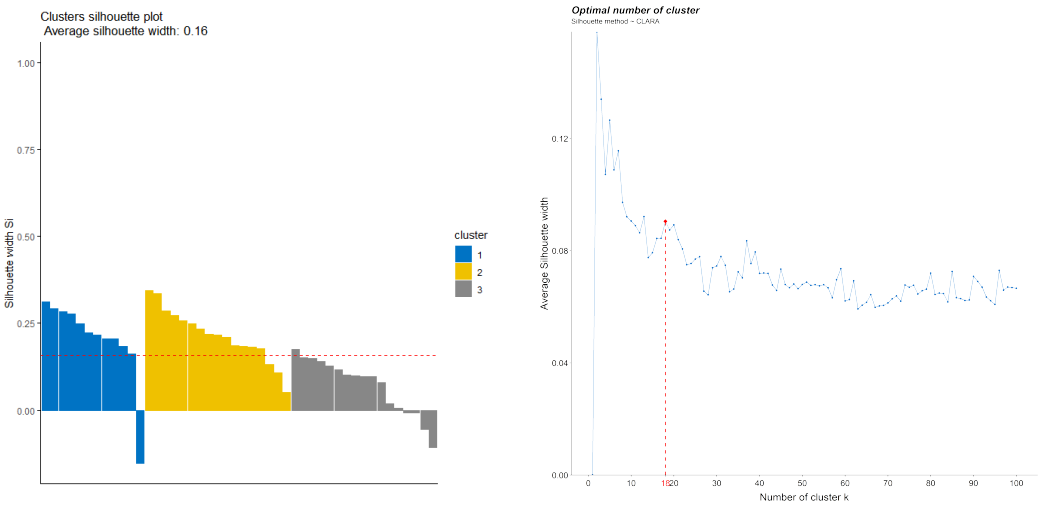
Figura 2.9: Metodo di stima del punto di gomito

coerente con i cluster vicini. Il coefficiente di silhouette viene definito come:

$$S_i = (b_i - a_i) / \max(a_i, b_i) \tag{2.6}$$

- a_i rappresenta la dissimilarità fra l'$i - esimo$ elemento del cluster rispetto a tutti gli altri oggetti del cluster a cui appartiene;
- b_i corrisponde alla minima dissimilarità fra l'$i - esimo$ elemento del cluster e tutti gli altri cluster vicini.

Se in un cluster emergono molti coefficienti di shilouette negativi è molto probabile che esso sia un raggruppamento poco rappresentativo. La Figura 2.10a rappresenta un esempio di ripartizione degli elementi di una distribuzione in tre cluster. Il valore del coefficiente di silhouette riportato corrisponde alla media dei coefficienti di ciascun elemento. Un oggetto caratterizzato da un valore negativo evidenzia una disomogeneità fra esso e il suo cluster. In Figura 2.10b, viene visualizzato l'andamento del valore medio del coefficiente di silhouette in funzione del numero di cluster, in analogia alla SST per il metodo del gomito.



(a) (b)

Figura 2.10: Metodo della silhouette

2.3 Analisi di Regressione e Classificazione

In questa sezione verranno presentati gli *algoritmi di apprendimento supervisionato*, metodi di machine learning in grado di stimare il risultato di osservazioni future. Vengono definiti supervisionati poiché al modello sono fornite delle coppie di dati di ingresso e uscita dalle quali esso estrae le correlazioni e le dipendenze in modo da comprendere quali output sono associati a determinati input, e come essi influenzano il risultato. Il modello sarà poi in grado di stimare il risultato di nuove istanze future sulla base delle conoscenze apprese.

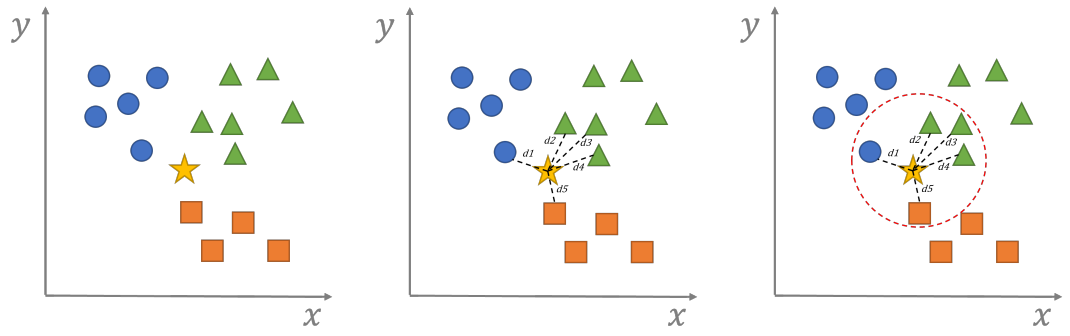
I modelli predittivi si distinguono in due gruppi principali:

- **Modelli di Regressione** per la stima di variabili numeriche continue, come ad esempio i consumi energetici di un edificio;

- **Modelli di Classificazione** in grado di stimare la classe o il gruppo di appartenenza di una determinata istanza. In questo caso la variabile stimata è categorica, come ad esempio la classe energetica di un edificio.

2.3.1 K-Nearest Neighbors (KNN)

L'algoritmo *K-Nearest Neighbors: KNN* è un algoritmo supervisionato di machine learning, semplice da implementare e dal costo computazione non elevato, utilizzato per risolvere problemi di classificazione e regressione. Il KNN si basa sul concetto che gli elementi vicini nello spazio n-dimensionale sono simili fra loro, ed assumono andamenti e caratteristiche analoghe. Per una nuova istanza dunque, il suo output verrà stimato comparandola con i k casi simili ad essa contenuti nel dataset, dove k è un parametro definito a priori che identifica il numero di oggetti più vicini considerati dall'algoritmo. In maniera analoga a quanto visto per il *k-means*, la somiglianza viene calcolata tramite la distanza euclidea. Minore sarà la distanza, maggiore sarà la somiglianza tra gli oggetti del dataset e l'istanza da prevedere. Per la classificazione l'assegnazione dell'oggetto più vicino viene eseguita secondo il voto maggioritario, mentre per la regressione trattando valori numerici continui, il vicino viene individuato valutando la minima distanza media fra gli oggetti dell'insieme. In Figura 2.11 viene rappresentato un esempio di classificazione per il quale la nuova istanza verrà assegnata alla classe "triangolo" poiché per l'intorno scelto essa risulta essere vicina a 3 oggetti "triangolo", a 1 oggetto "cerchio" e a 1 oggetto "quadrato". La Figura 2.12 esemplifica invece un caso regressivo dove l'output numerico dell'istanza n.16 verrà determinato come $Output_{16} = (Output_5 + Output_8 + Output_9 + Output_{11} + Output_{12})/5$



(a) classificazione di una nuova istanza (stella) nelle tre classi presenti(cerchio, quadrato,triangolo)

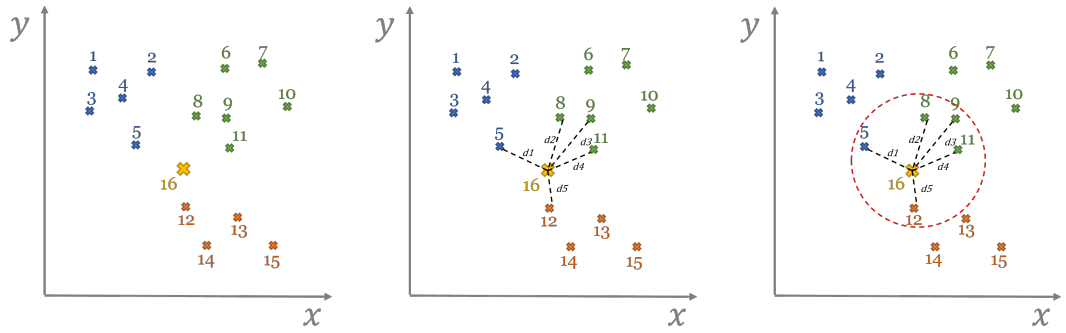
(b) valutazione della distanza euclidea rispetto agli elementi dell'insieme

(c) selezione dei k elementi più vicini

Figura 2.11: KNN per la Classificazione

L'algoritmo KNN si compone dei seguenti passaggi:

1. inizializzazione del parametro k per scegliere il numero di elementi vicini da



(a) individuazione del valore dell'output della nuova istanza n.16

(b) valutazione della distanza euclidea rispetto agli elementi dell'insieme

(c) selezione dei k=5 elementi più vicini

Figura 2.12: KNN per la Regressione

considerare;

2. per ogni elemento viene calcolata la distanza rispetto a tutti agli altri elementi componendo una matrice delle distanze

3. riordinamento crescente degli elementi in funzione del valore della distanza, dalla più piccola alla più grande;

4. selezione dei primi k elementi che presentano la distanza minore

5. **Classificazione**: per il nuovo oggetto viene selezionata la classe che compare più volte fra le etichette dei k elementi selezionati;

6. **Regressione**: per il nuovo oggetto ritorna come valore la media aritmetica della distanze dei k elementi selezionati.

Per scegliere il valore ottimale di k l'algoritmo viene eseguito diverse volte, ognuna di esse con un parametro k differente. Dopo un determinato numero di iterazioni viene assunto il valore di k che massimizza l'accuratezza della predizione per il dataset fornito, con l'obiettivo di ridurre l'errore compiuto dall'algoritmo sui nuovi dati che non ha mai processato.

È buona norma impostare un numero dispari di vicini da considerare, cosicché per la classificazione si ovvia il problema della parità di voti, caso in cui la scelta viene di fatto effettuata casualmente. Il *KNN* è un algoritmo versatile e semplice da implementare, presenta pochi parametri da impostare e si adatta sia a problemi

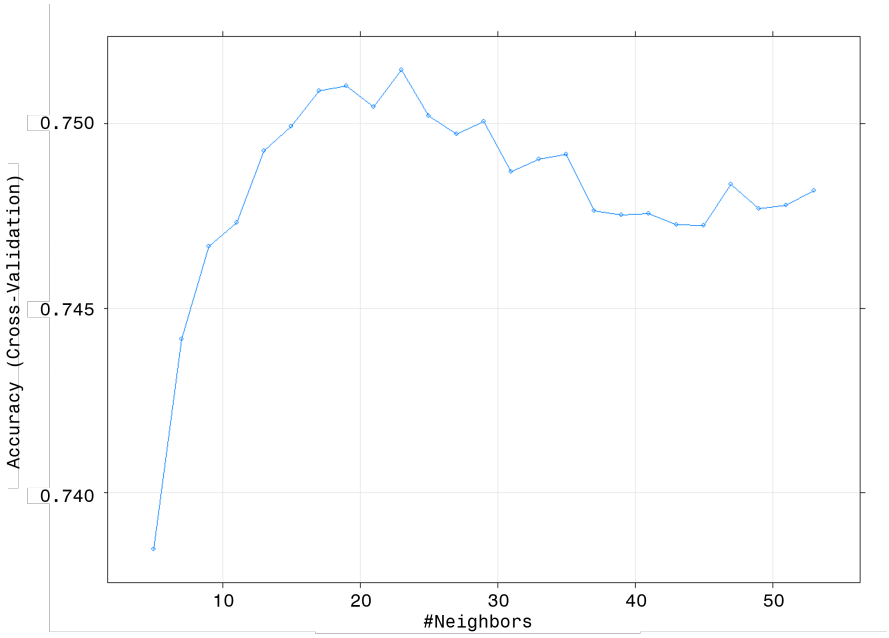


Figura 2.13: Stima del numero ottimale di vicini che ottimizza l'accuratezza del modello

di regressione che di classificazione. Di contro, la sua complessità aumenta notevolmente all'aumentare del numero di attributi caratteristici dell'insieme dei dati; ogni attributo corrisponde a una dimensione nello spazio, per la quale deve essere valutata la distanza fra gli oggetti.

2.3.2 Classification And Regression Tree (CART)

Il *Classification And Regression Tree: CART* è un algoritmo supervisionato di machine learning in grado di generare un modello predittivo data-driven. Questo modello viene costruito effettuando dei partizionamenti ricorsivi dell'insieme dei dati, in due sottoinsieme discendenti al fine di classificare gli elementi in gruppi omogenei all'interno di una partizione, e quando più differenziati rispetto all'altra.

Graficamente esso assume la tipica forma di un albero decisionale, partendo dal nodo radice, in cui l'insieme non è stato ancora suddiviso, si diramano una serie di nodi e rami. Ogni nodo dell'albero corrisponde a un sottoinsieme dell'insieme iniziale e in ognuno di essi avviene una ripartizione binaria dei dati, chiamata *split*. Ad ogni *split* il modello identifica quali attributi del sottoinsieme, e quali livelli di soglia sono i migliori discriminanti della successiva ripartizione, in accordo

con determinati requisiti di purezza prestabiliti. La ripartizione binaria ricorsiva segue quindi la logica degli algoritmi *greedy*, nei quali viene perpetuata la scelta più "golosa", più appetibile, che presenta un costo, definito in questo caso come impurezza del nodo, minore.

I nodi che non subiscono ulteriori *split* sono chiamati nodi terminali dell'albero decisionale, o nodi foglia. Ognuno di essi costituisce una ripartizione dell'insieme di partenza ben caratterizzata da determinati attributi e condizioni. Uno dei passaggi più critici dell'algoritmo è la determinazione delle condizioni di *split*: per ogni nodo si esegue una serie di *split* e viene scelto quello che rende più omogenei i dati all'interno dei due nodi discendenti.

La selezione del miglior *split* segue due approcci diversi in funzione del tipo di variabile da stimare. Nel caso della classificazione, per ogni nodo si definisce una funzione di impurità in grado di determinare la diversità di classi presenti nel sottoinsieme; l'indice di impurità più famoso e utilizzato è il coefficiente di Gini. Per la regressione invece l'obiettivo è minimizzare la funzione d'errore spesso definita dalla somma quadratica dei residui.

Coefficiente di Gini

Si consideri l'insieme $X = x_1, x_2, ..., x_n$ degli attributi di ingresso di un modello di classificazione, dove la variabile x_i può essere discreta, e quindi presentare un numero K di valori che definisce le modalità con le quali si presenta la variabile categorica, o continua. L'output discreto Y si presenta con J modalità. Per determinare lo *split* più efficacie, l'algoritmo svolge i seguenti passaggi:

1. Per ogni attributo dell'insieme X, viene calcolato il coefficiente di Gini per la $k - esima$ modalità assunta dalla variabile categorica x_i in relazione al verificarsi della variabile di output Y:

$$G(x_i = k) = \sum_{j=1}^{J} p(j|k)(1 - p(j|k)) = 1 - \sum_{j=1}^{J} p(j|k)^2$$

   dove $p(j|k)$ esprime la probabilità condizionata della modalità j della variabile di output Y al verificarsi della $k - esima$ modalità dell'attributo di ingresso x_i. Se l'attributo x_i è una variabile continua, il coefficiente di Gini viene valutato nel seguente modo:

$$G(x_i < c) = \sum_{j=1}^{J} p(j|(x_i < c))(1 - p(j|(x_i < c))) = 1 - \sum_{j=1}^{J} p(j|(x_i < c))^2$$

dove questa volta $p(j|(x_i < c))$ determina la probabilità che, dato per certo che x_i sia minore di un determinato valore c, Y assuma una delle sue j modalità. Il valore c viene valutato come la media del valore dell'attributo di un'istanza e quella successiva,

2. Calcolo del coefficiente di Gini ponderato per ciascun attributo x_i:

$$G(x_i) = \sum_{k}^{K} G(x_i = k) p_k$$

con p_k uguale al rapporto fra il numero di volte in cui l'attributo x_i assume la modalità $k - esima$ e il numero totale di istanze.

3. L'attributo che presenta il coefficiente di Gini ponderato più basso viene scelto come attributo sul quale effettuare lo *split*.

4. Il coefficiente di Gini più basso dell'attributo selezionato determina il punto di *split*.

Questa sequenza viene ripetuta per ogni nodo dell'albero decisionale. Il coefficiente di Gini assume valor compresi fra 0 e 1. Minore è il valore di G più l'insieme si presenta omogeneo.

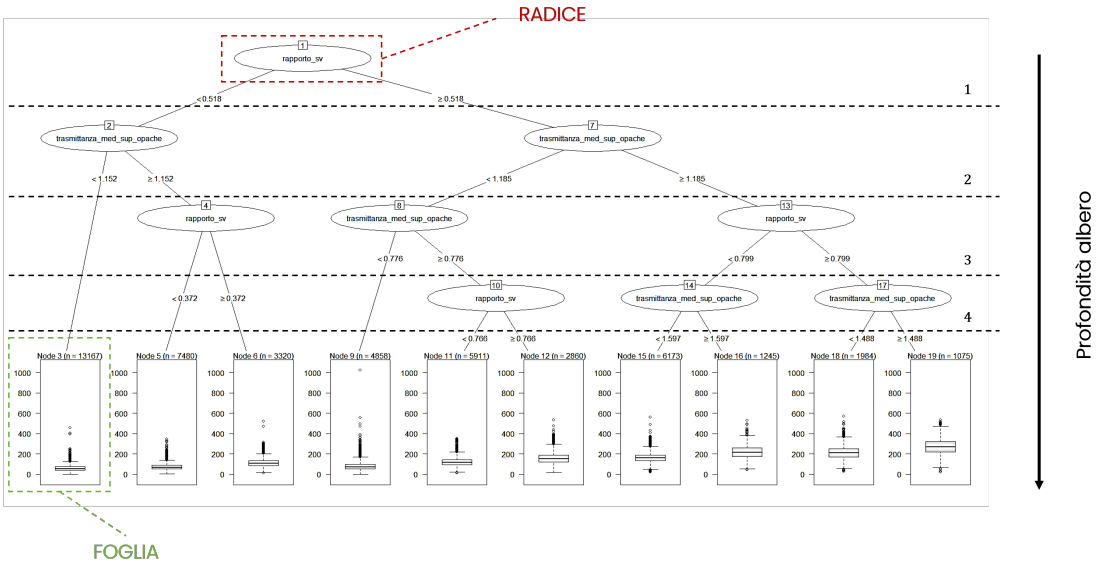


Figura 2.14: Esempio di albero di Regressione

Senza nessuna limitazione il *CART* potrebbe effettuare un numero elevato di ripartizioni fino ad arrivare a considerare un numero massimo di nodi foglia pari

al numero di istanze presenti nel dataset. Un modello siffatto presenta un sovradattamento ai dati forniti per addestrarlo, overfitting, diventando inaccurato e di difficile interpretazione per le previsioni successive. Per evitare tale problema, è necessario imporre delle limitazioni alla crescita dell'albero, una di queste rappresentata dal parametro di complessità *cp*. Ogniqualvolta l'albero effettua uno *split*, si genera un errore determinato dalla somma quadratica dei residui (*SSE: Sum of Square Error*), al quale viene sommato il *cp*.

$$minimize(SSE + cp)$$

L'algoritmo tende a minimizzare questa somma per ogni *split*, dunque quando si imposta un *cp* alto, verranno accettati solamente le ripartizioni che generano un SSE limitato, dalle quali si formerà un albero compatto è poco sviluppato. Un valore del *cp* basso invece, penalizzerà in quota minore la crescita dell'albero, che tenderà ad effettuare molte più ripartizioni. Altre tipologie di limitazioni da imporre alla crescita dell'albero sono:

- *Maxdepth*: si impone il valore della massima profondità che può assumere l'albero. Essendo l'albero binario, in questo modo è possibile determinare il numero massimo di nodi foglia che si genereranno;

$$maxdepth = 4 => n_{max} foglie = 2^4 = 16$$

- *Minsplit*: minimo numero di elementi che devono essere osservati all'interno di un nodo per poter effettuare uno *split*;

- *Minbucket*: numero minimo di elementi che deve essere presente in un nodo foglia per poter essere accettato.

Fra i vantaggi principali del *CART* vi è la sua semplicità sia in termini di implementazione, che di interpretazione. Infatti l'albero decisione è in grado di gestire variabili continue e categoriche, senza la necessità di generare variabili ausiliarie, *dummy variables*, ed è in grado di fornire un modello discretamente accurato anche con un numero ridotto di dati con cui allenarsi. Inoltre la rappresentazione grafica della struttura ad albero consente un intuitiva comprensione del funzionamento dell'algoritmo da parte dell'essere umano, cosa non scontata quando si parla di modelli black-box. Di contro però i modelli basati su un singolo albero risultano essere poco robusti e poco accurati rispetto ad altri algoritmi di machine learning comunemente usati.

2.3.3 Metodi Ensamble

In statistica e nel campo dell'intelligenza artificiale i metodi ensemble usano più algoritmi di apprendimento per incrementare le performance predittive del modello globale, il quale risulta più accurato rispetto ad ogni suo singolo modello costituente [24].
Tra le metodologie di ensambling più diffuse si pone l'attenzione su:

- *Bagging*: allena indipendentemente e parallelamente fra loro più algoritmi di base e successivamente media i risultati ottenuti secondo un determinato criterio;

- *Boosting*: allena sequenzialmente gli algoritmi di base, in modo che l'algoritmo successivo sia influenzato dal precedente.

Bagging e Random Forest

Il *bagging*, conosciuto anche come *bootstrap aggregation*, è una tecnica statistica che consiste nel generare dei campioni (*bag*) di dimensione B, estratti in maniera casuale dal dataset iniziale di dimensione N. Si generano così dei sottoinsiemi quanto più indipendenti fra loro e ben rappresentativi della distribuzione del dataset iniziale, con cui allenare i vari singoli modelli, al fine di minimizzare la varianza del modello finale. Le *Random Forest* sono degli algoritmi supervisionati di machine learnig, composti da un ampio numero di singoli alberi decisionali, da qui il concetto di "foresta" espresso nel nome, che operano insieme attraverso il *bagging*. L'approccio che si segue nel *Random Forest* è di aggiungere alla metodologia *bagging* anche la scelta casuale del numero di attributi con cui allenare i singoli modelli. In questo modo si limita ulteriormente la correlazione fra i singoli alberi decisionali, rendendo il modello più robusto alle diverse distribuzioni del dataset e meno influenzato dalla presenza di eventuali valori mancanti. In funzione del tipo di problema trattato esistono diversi metodi per aggregare i risultati dei singoli modelli allenati in parallelo. Per i problemi di regressione l'output finale del modello aggregato si ottiene effettuando una media dei risultati ottenuti dai singoli modelli. Per i problemi di classificazione uno dei criteri con il quale decidere il risultato finale è il voto maggioritario; si tiene conto di tutti i risultati predetti dai singoli modelli, e la classe che è stata stimata il maggior numero di volte viene assunta come risultato del modello aggregato.

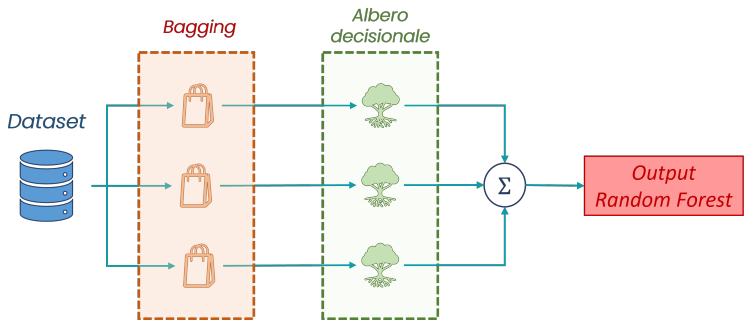


Figura 2.15: Schema del Random Forest: Bagging

Boosting

Nel *boosting* i singoli modelli vengono allenati sequenzialmente in maniera indipendente fra loro. L'idea alla base è quella di allenare iterativamente i singoli modelli sui risultati ottenuti dai modelli precedenti. In questo modo i modelli successivi tendono ad apprendere ciò che ha causato gli errori o le imprecisioni nei suoi predecessori, fino ad ottenere un modello finale aggregato avente un'accuratezza maggiore rispetto ai singoli modelli. La robustezza del modello finale aggregato deriva dal fatto che ogni singolo modello viene allenato partendo dai punti deboli del modello precedente, dando, ad esempio, maggiore importanza alle osservazioni nel dataset che hanno prodotto i risultati più divergenti. Come per il

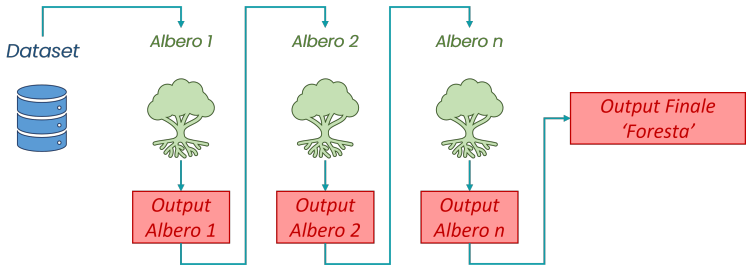


Figura 2.16: Schema del Boosting

bagging, anche il *boosting* può essere utilizzato sia per problemi di regressione che di classificazione, ricavando il risultato finale dalla media delle singole predizioni ponderate sui pesi associati a ciascun modello o da tecniche di voto maggioritario.

2.3.4 Bayesian Additive Regression Tree

Il *Bayesian Additive Regression Tree: BART* è un metodo di ensambling simile al *boosting*, che sfrutta più alberi decisionali per ottenere un modello più robusto. L'idea alla base del *BART* consiste nell'imporre una probabilità a priori che regolarizzi l'adattamento del modello *sum-of-tree*, mantenendo gli effetti dei singoli alberi limitati, evitando un'assunzione troppo stringente dei parametri. In questo modo ognuno di essi sarà in grado di spiegare una porzione specifica dei risultati ottenuti [25]. Per generare il modello aggregato di alberi decisionali, il *BART* impiega una versione adattata del modello bayesiano di backfitting basato sul metodo Monte Carlo e catene di Markov (MCMC).

La regressione ha lo scopo di evidenziare come la variazione degli attributi di input del modello X, utilizzati come predittori, influenzino la risposta Y, attraverso una funzione $f(X) = E(Y|X)$, esprimibile in termini di probabilità condizionata.

$$Y = f(X) + \epsilon \qquad \epsilon \sim N(0, \sigma^2)$$

Per approssimare la funzione $E(Y|X)$ il *BART* utilizza il modello *sum-of-tree* $E(Y|X) \approx \sum_{j=1}^{m} g_j(x)$, di m alberi decisionali g_j, che rappresenta un modello di regressione additivo con componenti multivariati, in grado di incorporare gli effetti di interazione in maniera molto più naturale [25]. In particolare ogni albero viene identificato nella sommatoria con i termini T, che denota l'albero binario caratterizzato da un insieme specifico di nodi interni, regole di *split* e nodi terminali, e da $M = \mu_1, \mu_2, ..., \mu_b$ che esprime l'insieme dei valori dei parametri associati al $b - esimo$ nodo foglia terminale. Ad ogni variabile x viene assegnato il valore μ_i relativo al nodo foglia di T, ad essa associabile tramite la sequenza delle decisioni prese per ogni *split* a partire dal nodo radice. Esprimendo questa notazione al modello si ottiene la relazione

$$Y = \sum_{j=1}^{m} g(x; T_j, M_j) + \epsilon \qquad \epsilon \sim N(0, \sigma^2)$$

dove per ogni albero di regressione T_j e i parametri M_j associati ai suoi nodi terminali, $g(x; T_j, M_j)$ è la funzione che assegna $\mu_{ij} \in M_j$ a x. Ogni μ_{ij} può dunque rappresentare un effetto primario della variabile indipendente su quella dipendente quando $g(x; T_j, M_j)$ dipende da una sola variabile x, o un effetto di interazione quando $g(x; T_j, M_j)$ dipende da più di un componente di x [25].

In un modello siffatto, un albero più sviluppato risulterebbe più influente rispetto a quelli più piccoli, compromettendo l'essenza del metodo additivo. É necessario imporre dei vincoli alla crescita dei singoli alberi. Se i metodi *boosting* limitavano la crescita dell'albero moltiplicando il risultato di ogni albero per una piccola

costante scelta tramite una validazione incrociata, il *BART* impone una funzione di probabilità a priori sui parametri caratterizzanti il modello *sum-of-tree*, come la profondità massima e i vincoli di crescita (*shrinkage)*, per regolarizzare l'apprendimento. In questo modo ogni singolo albero sarà in grado di rappresentare solo una porzione specifica dei risultati ottenuti, e presi singolarmente ognuno di essi presenta dunque un contributo limitato al modello finale (*weak learner)*. La generazione del modello additivo ad albero viene affidata all'algoritmo *Markov Chain Monte Carlo: MCMC* tramite un metodo di backfitting. L'idea alla basa è di simulare la distribuzione di probabilità a posteriori della variabile dipendente, generatasi dall'unione della distribuzione a priori e i dati, mediando i risultati di stima ottenuti dai vari campionamenti effettuati dal $MCMC$, invece di stimare in maniera diretta tutta la distribuzione una volta sola. Al metodo del campionamento casuale *Monte Carlo* si aggiunge il concetto delle catene Markoviane, in questo modo il singolo campionamento viene generato in funzione del precedente. L'algoritmo viene inizializzato con un numero m prestabilito di singoli alberi (m potrebbe anche essere valutato come un parametro a priori, ma si evita di farlo per non inficiare i tempi computazionali), ad ogni iterazione per ogni albero può aumentare o diminuire il numero di nodi foglia o cambiare qualche regola decisionale di *split*, facendo variare μ e la varianza σ. L'algoritmo prosegue fino a un criterio di convergenza delle iterazioni.

2.3.5 Artificial Neural Network

Le reti neurali (*Artificial Neural Network: ANN*) sono delle tecniche di machine learning facenti parte della sottocategoria del deep learning (Figura 2.17).

Il termine *deep* viene utilizzato poiché tali algoritmi utilizzano più livelli (*layers*) di apprendimento, per ricreare le relazioni fra input e output con maggiore significatività. Il loro nome deriva dall'analogia di funzionamento con le reti neurali biologiche. Il concetto alla base è quelli di collegare fra loro singole unità in grado di svolgere funzioni elementari, neuroni, al fine di elaborare compiti complessi. Le relazioni input-output vengono ricavate attraverso delle semplici trasformazioni sequenziali delle informazioni ad ogni passaggio nei *layer*. Tecnicamente queste trasformazioni avvengono a seguito di una parametrizzazione dei pesi w attribuiti alle variabili di ingresso x. In questo contesto, l'apprendimento consiste nel trovare una configurazione di valori per i pesi in tutti i *layer* della rete neurale, rendendola in grado di interpretare le corrette relazioni esistenti tra i dati in ingresso e i relativi output Y [26]. Generalmente i neuroni sono dislocati in un layer o vettore, in modo che l'output di uno di essi funga da input a quello successivo. Le connessioni

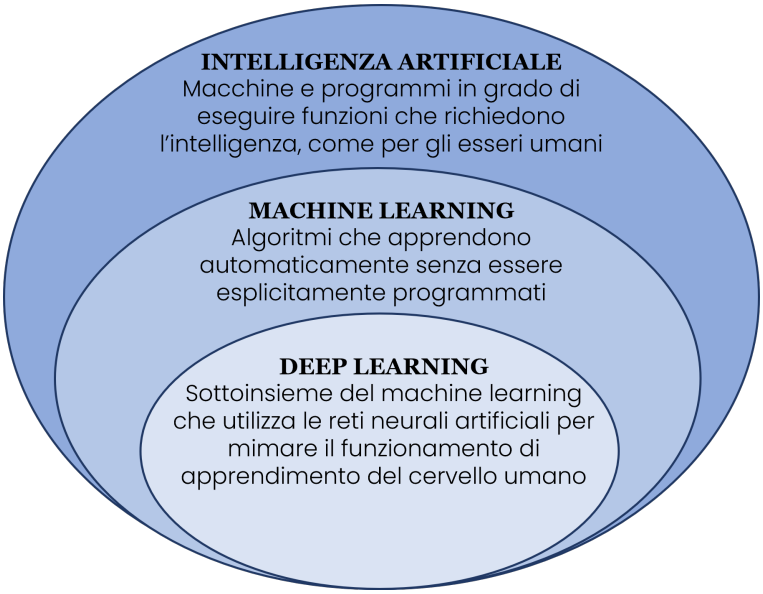


Figura 2.17: Deep Learning

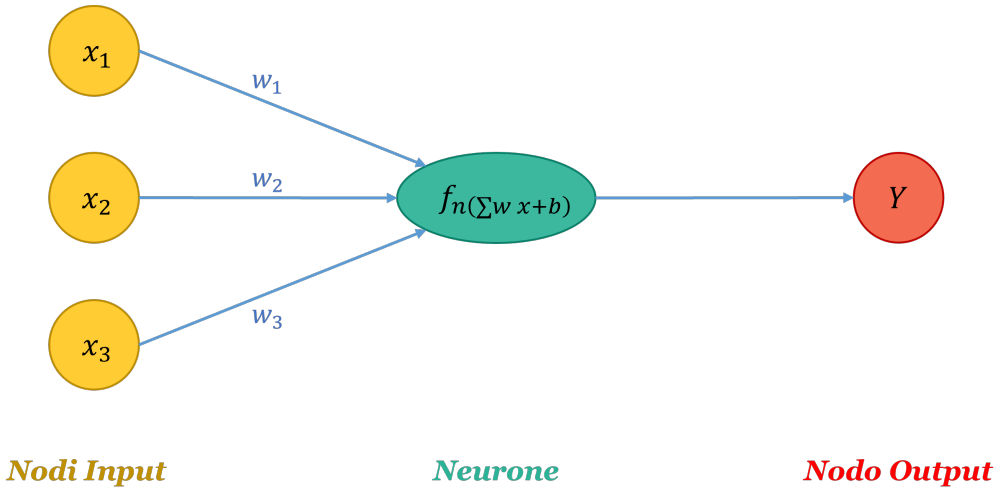


Figura 2.18: Schema Neurone

che ogni neurone può avere con gli altri neuroni del layer seguente, simulano le sinapsi cerebrali del cervello biologico. La struttura fondamentale delle reti neurali consta di tre livelli principali:

- Un *input layer* costituito dai dati da cui estrarre le informazioni per la modellazione dell'output;

- Uno o più *hidden layers* composti dalle unità neuronali programmate per analizzare i dati in ingresso valutandone il peso nella stima finale, attraverso l'utilizzo di funzioni di attivazione specifiche;

- Un *output layer* composto da uno o più elementi in funzione del risultato che si vuole ottenere.

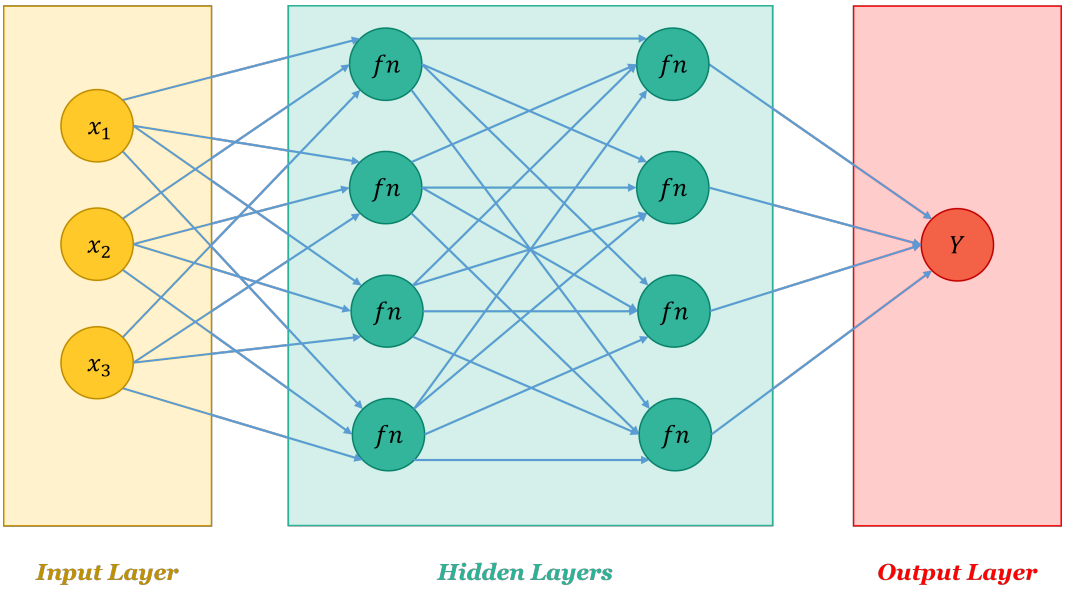


Figura 2.19: Schema ANN

In una rete neurale però, vi sono migliaia di parametri da tenere in considerazione, generati dalla concatenazione fra neuroni. Prima di tutto è necessario dunque poterli osservare, per stabilire come essi influiscono sull'output e successivamente misurare la differenza fra il valore del risultato atteso e quello ricavato dalla rete, e regolare i parametri di conseguenza. La *loss function* è in grado di adempiere a questo scopo: considera le previsioni effettuata della ANN e le confronta con il valore atteso per quell'output, computando un parametro di scostamento che determina quanto bene sta lavorando la rete su quello specifico esempio considerato. Il vantaggio consiste nell'utilizzare questo scostamento come feedback, per correggere poi il valore del peso assegnato a un determinato input al fine di indirizzare la predizione verso il valore desiderato. L'aggiustamento dei pesi è un compito demandato all'*optimizer*, il quale implementa l'algoritmo di *backpropagation* che consente alle reti di riaggiornarsi iterativamente in funzione dei dati e degli scostamenti osservati. Ad ogni iterazione l'algoritmo corregge i vari pesi

cercando di minimizzare la *loss function*, secondo determinate metriche di valutazione, ottenendo così una previsione che si avvicini il più possibile al valore atteso [26].

2.4 Tecniche di eXplainable Artificial Intelligence

Gli algoritmi di machine learning sono uno strumento molto efficace per estrarre le informazioni da un dataset, ma molto spesso, come è anche emerso dai paragrafi precedenti, il loro funzionamento non è sempre trasparente e comprensibile per l'essere umano. Questo problema è tipico dei modelli *black-box*, per i quali l'utente fornisce un set di dati di input e ne riceve uno di output, ma la funzione di trasferimento sviluppata dal modello, come esso ha interpretato e riarrangiato i dati per poter stimare il risultato, rimane un concetto sconosciuto. In alcuni casi sapere solamente **cosa** il modello ha predetto non è più sufficiente, mentre assume sempre di più un ruolo fondamentale conoscere il **perché** il modello ha effettuato una determinata stima, o ha assegnato un edificio ad una classe piuttosto che ad un'altra. La conoscenza del perché facilità la comprensione e l'interpretabilità del problema stesso. Secondo Miller [27], l'interpretabilità è il grado in cui un essere umano può comprendere la causa di una decisione. Per Kim [28] l'interpretazione è il grado in cui un essere umano può prevedere in modo coerente il risultato di un modello. Più alto è il livello di interpretabilità di un modello, più facile sarà comprendere il perché sono state effettuate determinate decisioni. Se un utente non può fidarsi di un modello, molto probabilmente non lo userà. Oltre alla fiducia riposta nella predizione generata, è necessario avere fiducia anche verso il modello che produce tale risultato. Gli studi svolti nel campo dello *XAI (eXplainable Artificial Intelligence)* mirano a trovare un modo per poter rendere l'utente finale padrone del modello data-driven che sta utilizzando, e non viceversa.

Esistono due tipologie principali di interpretazione modellistica:

- *Model-specific*: livello di interpretazione limitata a una specifica classe di modelli, da spiegazioni riguardo i parametri caratteristici di un modello, come ad esempio l'interpretazione dei pesi in un modello di regressione lineare;
- *Model-agnostic*: metodo più generale e applicabile a qualsiasi modello, non ha accesso alle strutture interne dello specifico modello, e tende a spiegare come gli output sono influenzati dagli attributi di ingresso a posteriori del processo di addestramento.

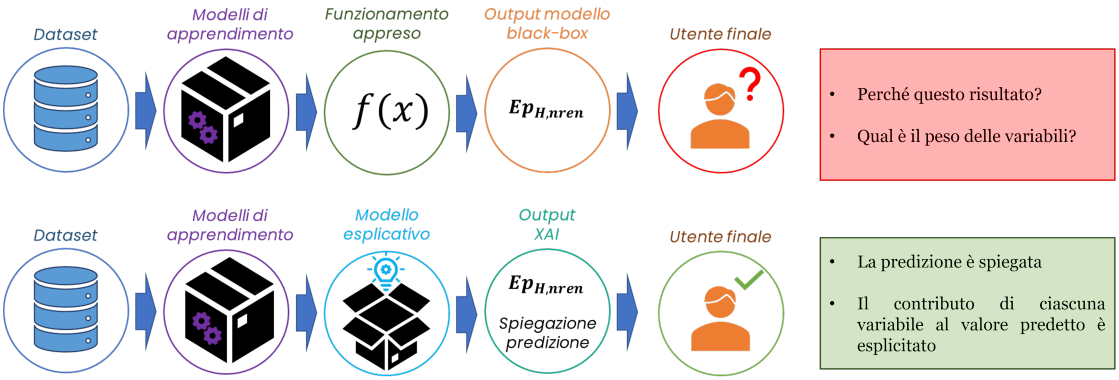


Figura 2.20: Metodologia XAI

Ognuna di esse può inoltre dare informazioni riguardo l'interpretabilità globale dell'intero modello, di come gli attributi influenzano mediamente la predizione, spesso molto difficile da ottenere, o ad un livello di interpretabilità locale di una singola o di un insieme ristretto di istanze, per le quali si cerca di analizzare il perché il modello ha stimato uno specifico risultato in funzione degli specifici input forniti [29].

L'obiettivo dei vari algoritmi di *XAI* è reinterpretare i risultati ottenuti dal modello black-box ed esprimerli attraverso un modello più semplice, più interpretabile ad alto livello. Ideali per questo scopo sono i modelli di regressione lineare, in grado di esprimere l'output desiderato attraverso una funzione polinomiale nella quale ad ogni attributo è associato un determinato peso e segno, che permette di capire come e quanto esso influisce sul risultato finale. Anche gli alberi decisionali assolvono bene questo compito, percorrendo l'albero dal nodo foglia fino alla radice è possibile capire il perché di una determinata scelta osservando le regole decisionale compiute dall'algoritmo ad ogni *split*.

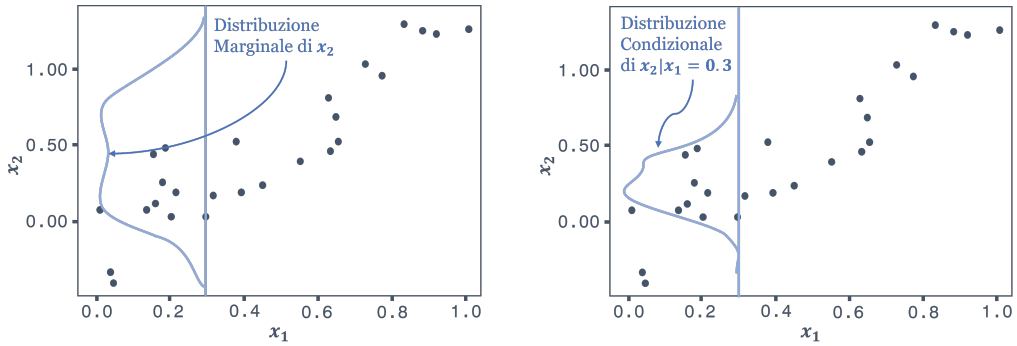
2.4.1 Partial Dependence Plot (PDP) e Accumulated Local Effects (ALE) Plot

Il *PDP* è un metodo di interpretabilità globale del modello in grado di visualizzare l'effetto marginale di uno o più attributi sul risultato di output di un modello di machine learning. La visualizzazione grafica è uno strumento molto potente che facilità la comprensione della relazione tra input ed output [30]. La funzione di

dipendenza parziale per la regressione è definita come:

$$\hat{f}_S(x_s) = E_{X_C}[\hat{f}(x_S, X_C)] = \int \hat{f}(x_S, X_C) dP(X_C) \tag{2.7}$$

dove x_S è l'attributo di interesse che si vuole visualizzare e X_C sono tutti gli altri attributi utilizzati dal modello di machine learning. La funzione 2.7 effettua un'analisi di sensitività sulla distribuzione dell'output in funzione della sola variabile x_S, escludendo il contributo alla predizione delle altre variabili X_C. Idealmente per rendere significativa la rappresentazione, ogni variabile x_S dovrebbe non essere correlata a nessuna variabile dell'insieme C, in questo modo la *PDP* rappresenterebbe il mutamento della predizione associato solamente alla variazione di x_S. Se così non fosse non sarebbe possibile visualizzare questa relazione diretta a causa dei contributi relativi agli effetti inferenziali degli attributi. Un altro problema relativo all'utilizzo del *PDP* è la generazione di istanze insensate ed inusuali, che non dovrebbero essere considerate nella spiegazione dell'output. Come detto l'algoritmo fissa i valore degli attributi non di interesse e modifica man mano i valori di quello che si vuole attenzionare. Ciò comporta che se si sta ad esempio valutando il contributo della superficie disperdente totale di un edificio, nel tracciare la curva verrà valutato un valore di 15 m^2, quando poi magari risulta per una determinata istanza una superficie dell'involucro opaco di 18 m^2, il che è evidentemente insensato. Questo è ciò che avviene quando viene applicata la probabilità marginale sul valore dell'attributo x_S.



(a) Stimatori PDP

(b) Stimatori ALE

Figura 2.21: Distribuzione marginale e condizionata

L'*ALE* plot è uno dei metodi alternativi utilizzati per sopperire a questo problema. Esso sfrutta la probabilità condizionata dell'attributo in esame per diminuire l'effetto di correlazione con le altre variabili, e traccia il grafico mediando le

differenza nella predizione [29].

$$\hat{f}_{j,ALE}(x) = \sum_{k=1}^{k_j(x)} \frac{1}{n_j(k)} \sum_{i:x_j^{(i)} \in N_j(k)} [\hat{f}(z_{k,j}, x_j^{(i)}) - \hat{f}(z_{k-1,j}, x_j^{(i)})] \tag{2.8}$$

Dove $z_0, z_1, etc.$ sono i valori dell'attributo x_1, tipicamente identificati dai quantili, in modo da creare degli $N(k)$ intervalli $[z_{k-1} - z_k)$ con un numero simile di elementi, $n(k)$ definisce il numero di elementi presenti all'interno dell'intervallo $N(k)$, e $k(x_1)$ definisce l'indice dell'intervallo a cui appartiene x_1. Il termine $\hat{f}(z_{k,j}, x_j^{(i)})$ indica che per l'istanza i il valore di x_j viene sostituito con il valore z_k del limite destro dell'intervallo (ciò avviene in maniera analoga per il secondo membro della sottrazione con z_{k-1} identificante il limite sinistro dell'intervallo), lasciando immutato il resto degli attributi e valutando la differenza nelle predizioni in questi due punti.

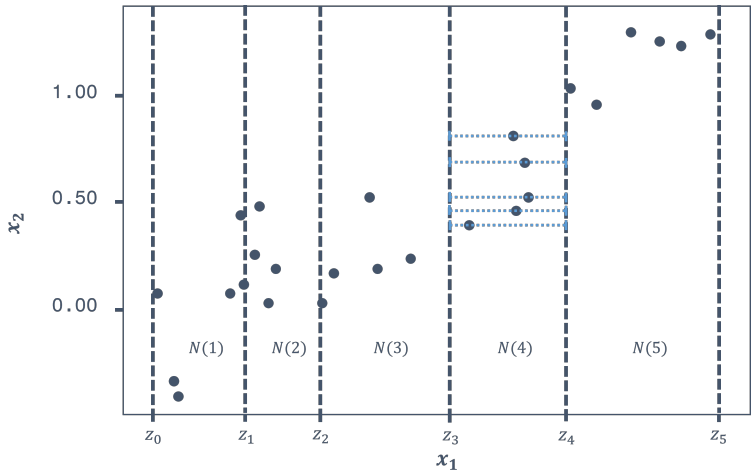


Figura 2.22: Valutazione dell'ALE per l'attributo x_1 correlato a x_2

In Figura 2.22 la distribuzione dei valori di x_1 in funzione di x_2 è stata suddivisa in 5 intervalli con verosimilmente egual numero di elementi. Concentrandosi sull'intervallo $N(4)$, per ogni punto al suo interno, il valore di x_1 viene sostituito con z_3 e con z_4, viene calcolata la differenza della predizione per questi punti e successivamente valutata la media, dividendola per il numero $n(4)$ di elementi appartenenti a $N(4)$. Si effettua questa operazione per tutti gli intervalli e alla fine si sommano i risultati [31].

2.4.2 Permutation Feature Importance

La *Permutation Feature Importance* (Figura 2.23)è una tecnica di visualizzazione che mira a mostrare il peso di ogni variabile predittiva nella predizione finale dell'output, misurando l'aumento dell'errore (espresso come RMSE) nella predizione del modello, dopo aver permutato i valori del singolo attributo. Una variabile risulta quindi più importante delle altre se in seguito alla permutazione dei suoi valori, l'errore verificatosi nella predizione è massimo. Ciò significa che il modello considera rilevante la variabile ai fini della stima del risultato, e una sua variazione è direttamente correlata all'output. Al contrario, se il cambiamento dei valori di una variabili non produce un errore significativo nella predizione, tale variabile non è considerata importante dal modello.

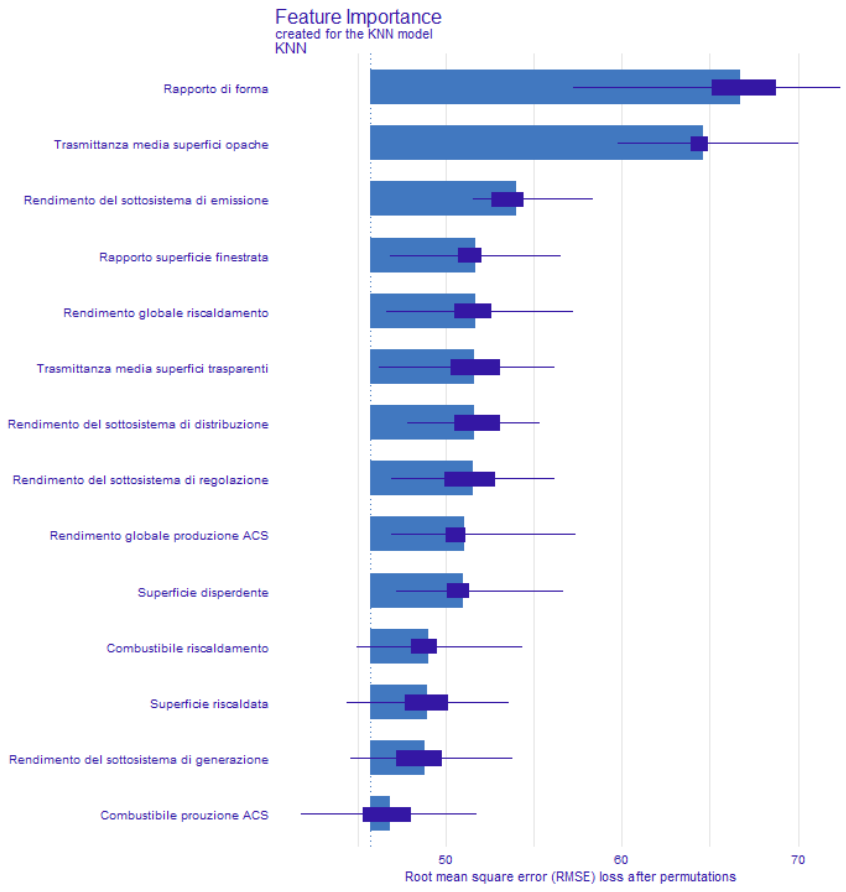


Figura 2.23: Permutation Feature Importance plot, per un modello KNN

2.4.3 Break-down Plot

Il *breakDown Plot (BD)* è un modello di interpretazione locale della predizione che cerca di spiegare come i vari attributi contribuiscono al raggiungimento del risultato predetto per un'istanza. L'idea alla base del *BD* è di identificare i contributi di ogni attributo valutando il cambiamento dell'output, mantenendo fissi i valori delle altre variabili similmente a quanto avviene per la *Permutation Feature Importance*.

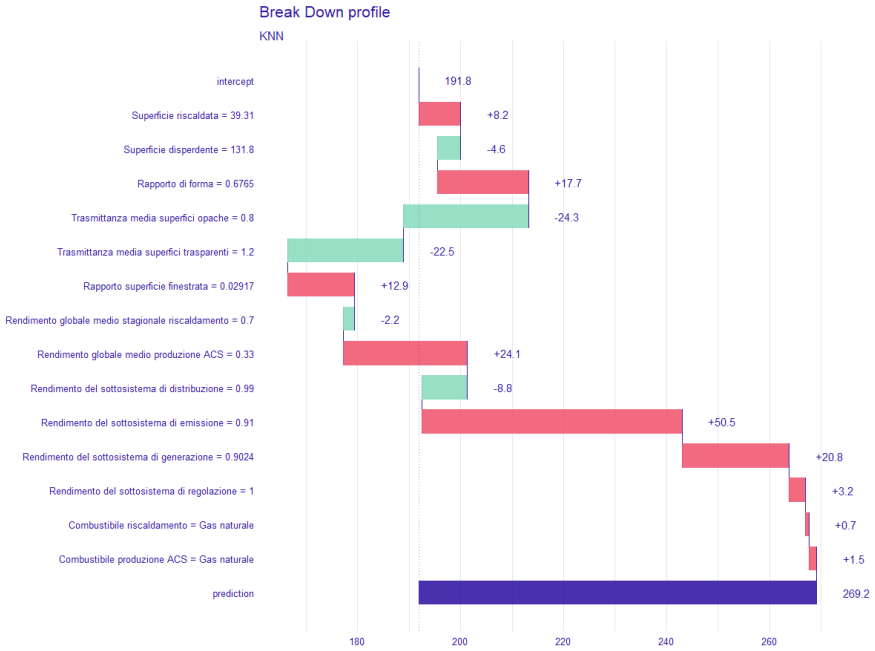


Figura 2.24: Esempio *Break-down Plot*

I passaggi dell'algoritmo di *BD* possono essere riassunti nei seguenti punti [32]:

1. Calcolo della media dell'output predetto, che fungerà poi da punto di riferimento e da intercetta nel grafico;

2. Valutazione dell'importanza dell'attributo, del peso di ciascun predittore. Per ogni variabile, viene sostituito e fissato all'interno del dataset il valore che essa assume nell'istanza definita come nuova osservazione, e successivamente viene ricalcolata la media delle predizioni con questa nuova configurazione, e valutata la differenza con la media originale. Il valore di tale differenza Δj determina il peso della variabile nella predizione.

3. Dopo aver effettuato il punto precedente per tutti ogni attribuo dell'istanza, essi vengono ordinati in maniera decrescente in funzione del valore calcolato di Δj. Questo approccio euristico determina l'ordine con cui saranno poi visualizzati i contributi di ogni variabile nella predizione finale, ed è fortemente dipendente dai valori contenuti nell'istanza. Cambiando istanza cambia la sequenza di variabili rappresentate nel grafico e i contributi di ciascuna di esse possono assumere significati talvolta anche opposti. In alternativa è possibile impostare un ordine aprioristico di visualizzazione per ogni istanza, in modo da poter evidenziare con maggiore facilità il cambiamento relativo di ciascuna variabile sulla predizione.

4. Si procede dunque nella valutazione della predizione fissando man mano i valori degli attribui dell'istanza secondo l'ordine determinato nel punto precedente. A differenza del punto 2, nel quale la variabile veniva modificata solo quando doveva essere valutato il suo contributo, in questo step le modifiche rimangono anche per le valutazioni delle variabili successive.

5. Convergenza di tutti gli attributi ai valori contenuti nell'istanza, e valutazione di come gli attributi contribuiscono alla predizione e di come essa si discosta dal valor medio delle predizioni del dataset (intercetta valutata nel punto 1).

Capitolo 3

Framework Metodologico

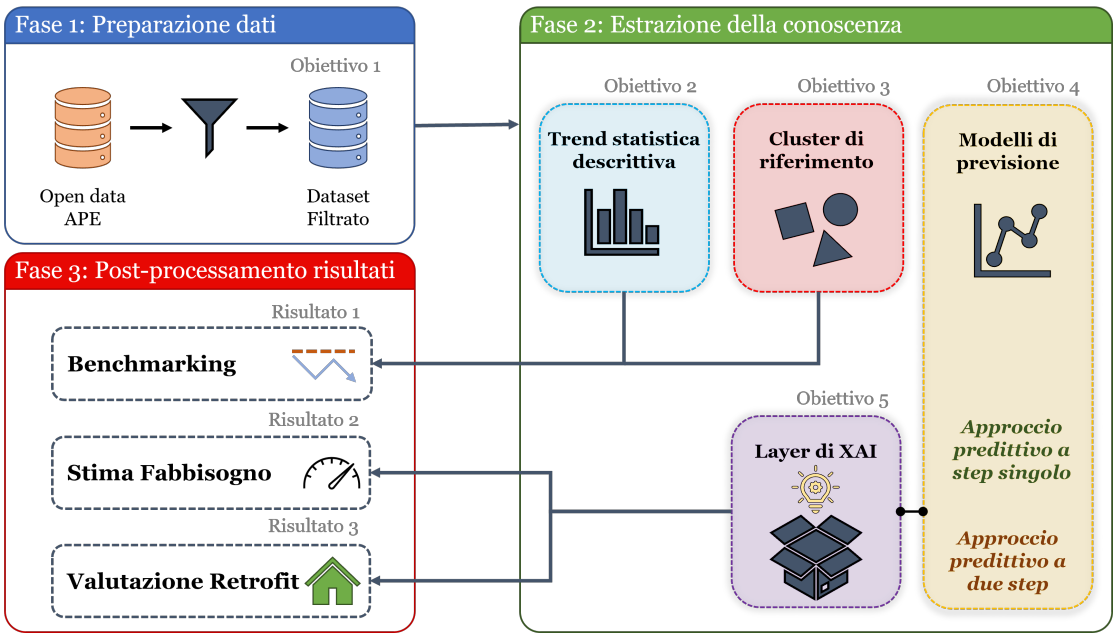


Figura 3.1: Framework Metodologico

Il lavoro di tesi è stato sviluppato applicando un approccio metodologico induttivo, attraverso l'implementazione di un framework in grado di stimare i consumi futuri di un nuovo edificio, sfruttando le informazioni contenute negli open data dei certificati energetici relativi alla Regione Piemonte. La stima del fabbisogno, fortemente correlata all'interpretazione del modello data-driven, può essere sfruttata per analizzare i punti deboli di un edificio e ipotizzare eventuali azioni di retrofit

per migliorarne prestazioni. Oltre ai modelli di previsione dell'indice di prestazione energetica globale dell'edificio, perpetuata con due approcci diversi analizzati in seguito, il framework propone anche un modello di analisi di benchmarking attraverso il clustering degli edifici accomunati da caratteristiche simili. In questo modo è possibile stimare le performance energetiche dell'edificio in relazione a una baseline di riferimento, in maniera diretta e intuitiva.

3.1 Obiettivo 1 ~ Identificazione del dataset di riferimento e pre-processamento dei dati

Al fine di utilizzare un dataset statisticamente significativo sono stati apportati dei filtri su alcune variabili ritenute maggiormente caratterizzanti i certificati energetici.

- *Filtro temporale*: sono stati selezionati gli attestati redatti fra il 2019 e il 2021;
- *Destinazione d'uso e oggetto attestato*: è stata considerata la destinazione d'uso residenziale relativa alla categoria edilizia E1(1). In questo contesto l'analisi è stata ristretta ulteriormente alle unità immobiliari.
- *Servizi energetici*: l'attenzione è stata rivolta agli edifici che usufruivano del riscaldamento invernale e della produzione di acqua calda sanitaria, in cui è installato un singolo impianto per servizio energetico.

Tali restrizioni sono state effettuate per concentrare l'attenzione sulla tipologia di attestati che si ripresentavano nel dataset con maggior frequenza, in modo da poter disporre di un valido campione statistico su cui effettuare le analisi successive. Come tutti i dataset adibiti alla raccolta di un numero massiccio e decentralizzato di informazioni, anche negli open data prelevati dal sito della Regione Piemonte sono presenti anomalie, che devono essere gestite per non inficiare la qualità dei modelli elaborati nelle fasi successive. Le motivazioni principali per le quali è necessario effettuare un pre-processamento e una pulizia (*data cleaning*) dei dati sono principalmente la presenza di valori mancanti, inconsistenze e outliers all'interno dei database. I valori mancanti vengono spesso gestiti attraverso metodi di interpolazione o rimpiazzo con costanti globali e locali. In questo lavoro però è stato ritenuto opportuno non recuperare gli attestati che presentavano valori mancanti attraverso tecniche di machine learning, per evitare una possibile "corruzione" dei dati e una loro scorretta interpretazione. Gli unici valori mancanti rimaneggiati riguardano l'attributo dei *Gradi Giorno*, fondamentali per la caratterizzazione

ambientale dell'edificio, che sono stati recuperati per ogni Comune del Piemonte da un database esterno, e inseriti del dataset usando come chiave di associazione l'attributo *Comune* presente negli open data. Le inconsistenze riscontrate negli open data sono da associarsi principalmente ai rendimenti. Il problema è da imputare ad un'errata digitazione o trasposizione del valore del rendimento, che invece di essere scritto nella sua forma decimale 0.951, in alcuni attestati viene espresso come 95.1, risultando completamente fuori scala rispetto agli altri. I valori che presentano questo problema evidente sono stati individuati e riconvertiti nel formato corretto per poter essere utilizzati, prima della fase di individuazione dei valori anomali. Infine gli outliers sono stati rimossi imponendo dei vincoli al dominio di esistenza delle variabili di interesse, in modo da escludere a priori le istanze che presentavano valori fuori scala.

3.2 Obiettivo 2 ~ Statistica descrittiva

Dopo aver filtrato e ripulito i dati, la seconda parte del framework consiste nell'estrarre la conoscenza intrinseca nei dati. Tramite gli strumenti della statistica descrittiva (boxplot, diagrammi a barre etc.) è possibile evidenziare i trend di riferimento delle variabili che compongono il dataset, in modo da individuare degli schemi topologici comuni e ricorrenti. In questa fase sono stati evidenziati gli attributi maggiormente significativi per la redazione dell'*APE*, come ad esempio le trasmittanze di involucro o i rendimenti di impianto che non sono di immediata determinazione. Sono stati delineati tre range di valori (basso, medio e alto) attraverso un processo di segmentazione monovariata della distribuzione, che segue un approccio derivante dal metodo *k-means*. Esso si basa sulla determinazione adattiva dei limiti dell'intervallo secondo il seguente procedimento [33]:

1. impostato il numero k di intervalli, l'algoritmo posiziona i $k-1$ limiti in modo da suddividere la distribuzione in intervalli equiprobabili;

2. per ogni *k-esimo* intervallo viene calcolato il centroide $r_k = \frac{1}{N_k} \sum_{t_k \in [\beta_i, \beta_{i+1})} t_k$ dove N_k sono il numero di elementi presenti nel $k - esimo$ intervallo, ciascuno con un determinato valore t_k. β_i e β_{i+1} identificano i limiti iniziali dell'intervallo, con $i = [1, ..., k-1]$ e $\beta_0 = -\infty$ e $\beta_k = +\infty$.

3. calcolo del nuovo limite $\beta_i = \frac{r_{i-1} - r_i}{2}$;

4. valutazione dell'errore totale di rappresentazione come:

$$\Delta' = \sum_{i=1}^{k} \sum_{t_k \in [\beta_i, \beta_{i+1})} (t_n - r_i)^2$$

5. se $\frac{\Delta - \Delta'}{\Delta} < \gamma$ con $\Delta \approx \infty$ e $\gamma > 0$, l'algoritmo è arrivato a convergenza e si ferma, altrimenti $\Delta = \Delta'$ e si ritorna al punto 1.

In questo modo è possibile visualizzare che range di valori assume l'attributo in esame, ad esempio in funzione della tipologia edilizia e dell'intervallo temporale di costruzione dell'edificio.

3.3 Obiettivo 3 ~ Clustering

Successivamente sono stati testati tre algoritmi non supervisionati di clustering, il *k-means*, il *CLARA* e il *gerarchico* per l'individuazione di una baseline di riferimento con la quale poter confrontare gli edifici dal punto di vista prestazionale. Dal dataset filtrato, ottenuto dall'obiettivo 1, sono stati selezionati nove attributi maggiormente significativi dell'attestato energetico. Sono stati esclusi dall'analisi

Nome	**Simbolo**
Superficie riscaldata	S_{heat}
Superficie disperdente	S_{disp}
Rapporto superficie finestrata	W_R
Rapporto di forma	S/V
Trasmittanza media superfici opache	U_{op}
Trasmittanza media superfici trasparenti	U_{tr}
Gradi Giorno	DD
Rendimento globale medio stagionale riscaldamento	η_H
Rendimento globale medio stagionale ACS	η_W

Tabella 3.1: Variabili analisi Clustering

gli attributi con una forte correlazione. Il modello tende a carpire tutte le caratteristiche utili dagli input forniti, nel caso di variabili correlate una determinata relazione viene ripetuta e spiegata più volte. Le viene attribuito un peso maggiore dal modello, che tralascia così altri aspetti e relazioni che sarebbero potute essere rilevanti. Per poter valutare la *distanza euclidea* nello spazio $n - dimensionale$ dove $n = 9$ pari al numero di attribui considerati, è stata effettuata una normalizzazione $z - score$ per rendere il dataset omogeneo. Per ogni algoritmo testato è stato valutato il numero ottimale di cluster attraverso il metodo del *gomito* (paragrafo 2.2.4). Infine è stato scelto l'algoritmo che mantenesse la migliore qualità del clustering, attraverso tecniche di validazione statistica basate sul coefficiente di silhouette e sulla SST (equazione 2.5). Ad ogni istanza del dataset è stata as-

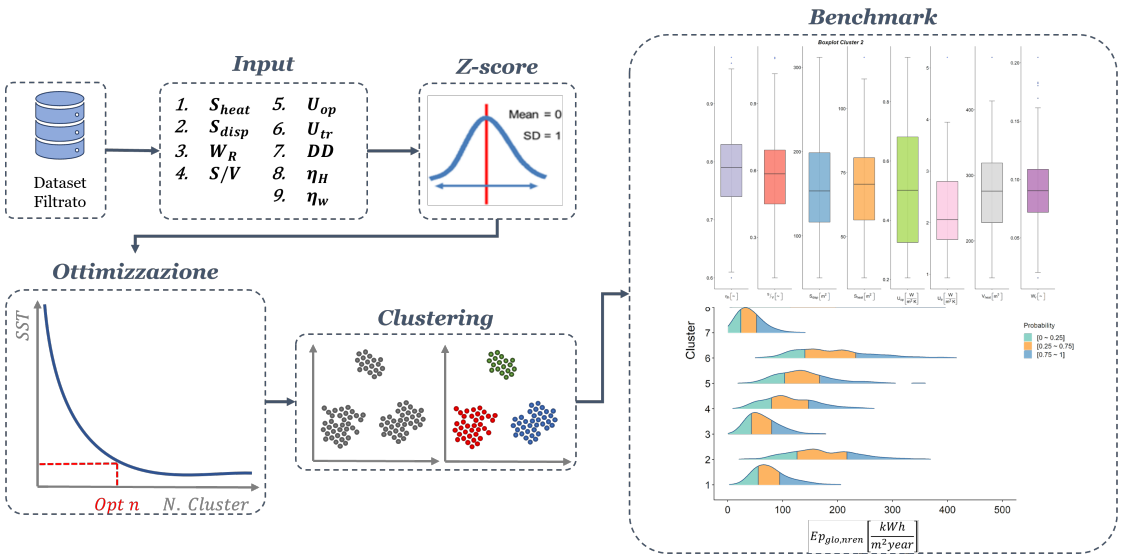


Figura 3.2: Metodologia clustering

segnata un etichetta di cluster, valutata in funzione della minore distanza euclidea fra il centroide dell'$i - esimo$ cluster e gli attributi dell'istanza stessa. Attraverso questa etichetta è possibile generare dei raggruppamenti di attestati energetici con caratteristiche simili, per i quali è possibile analizzare la distribuzione delle variabile termofisiche e geometriche come gli indici di prestazione energetica, valutando anche le relazioni fra essi. Ad esempio, un cluster che presenta una distribuzione con valori elevati dell'$Ep_{glo,nren}$, sarà caratterizzato da bassi valori del rendimento, o da alti valori delle trasmittanze di involucro. L'obiettivo 3 fornisce dunque una visualizzazione delle distribuzioni degli attributi caratteristici di ogni cluster, che può essere sfruttata per confrontare edifici diversi, o valutare come si rapporta un nuovo edificio rispetto a un campione di edifici a lui simili.

3.4 Obiettivo 4 ~ Modelli di previsione

L'obiettivo 4 consta nello sviluppo di algoritmi supervisionati di stima regressiva dell'indice di prestazione energetica globale non rinnovabile, su cui si basa l'attribuzione all'edificio di una determinata classe energetica che ne accerti le prestazioni per ragioni sia energetico-ambientali, sia legate al mercato immobiliare. Gli *APE* racchiudono diversi attributi influenzanti le performance energetiche dell'edificio, identificabili attraverso l'$Ep_{glo,nren}$. Il processo di selezione degli attributi è stato guidato dalle precedenti esperienze riguardanti gli attestati di prestazione, maturate negli studi condotti da [9], [10], [34] e [35], con lo scopo di utilizzare un numero discreto di variabili di input, che siano anche di facile reperibilità.

Categoria	**Nome**	**Simbolo**	**Valore medio**
Geometriche	Superficie riscaldata	S_{heat}	$72.6m^2$
	Superficie disperdente	S_{disp}	$157.3m^2$
	Rapporto superficie finestrata	W_R	0.11
	Rapporto di forma	S/V	0.54
Involucro	Trasmittanza media superfici opache	U_{op}	$1.10\frac{W}{m^2k}$
	Trasmittanza media superfici opache	U_{tr}	$3.31\frac{W}{m^2k}$
Ambientali	Gradi Giorno	DD	$2693DD$
Impianto	Rendimento globale medio stagionale riscaldamento	η_H	0.70
	Rendimento globale medio stagionale ACS	η_W	0.56
	Rendimento del sottosistema di distribuzione	η_d	0.96
	Rendimento del sottosistema di emissione	η_e	0.95
	Rendimento del sottosistema di generazione	η_g	0.90
	Rendimento del sottosistema di regolazione	η_r	0.97
	Combustibile riscaldamento	$Fuel_H$	Gas naturale, Energia elettrica, Biomasse, etc...
	Combustibile ACS	$Fuel_W$	Gas naturale, Energia elettrica, Biomasse, etc...

Tabella 3.2: Variabili modello di stima di $Ep_{glo,nren}$

Tali attributi sono stati identificati in quattro diverse categorie:

- *Geometriche*: le variabili in questa categoria descrivono le diverse caratteristiche geometriche dell'edificio che influenzano le sue prestazioni energetiche;
- *Involucro*: sono contenute le variabili che caratterizzano l'edificio dal punto di vista termofisico;
- *Ambientali*: a questa categoria appartengono i gradi giorno, utilizzati per attribuire un peso anche alla zona climatica in cui sorge l'edificio. L'analisi è stata condotta solamente su edifici appartenenti alla regione Piemonte, di conseguenza i gradi giorno si distribuivano con una scarsa varianza, ma qualora lo studio si estendesse anche a edifici appartenenti ad altre Regioni con condizioni climatiche differenti, i gradi giorno costituirebbero un determinante fondamentale.
- *Impianto*: sono contenuti i rendimenti globali medi stagionali relativi alla produzione di acqua calda sanitaria e al riscaldamento invernale, per il quale sono presenti anche i rendimenti di sottosistema.

La Tabella 3.2 riporta le variabili ritenute più rilevanti per ogni categoria.

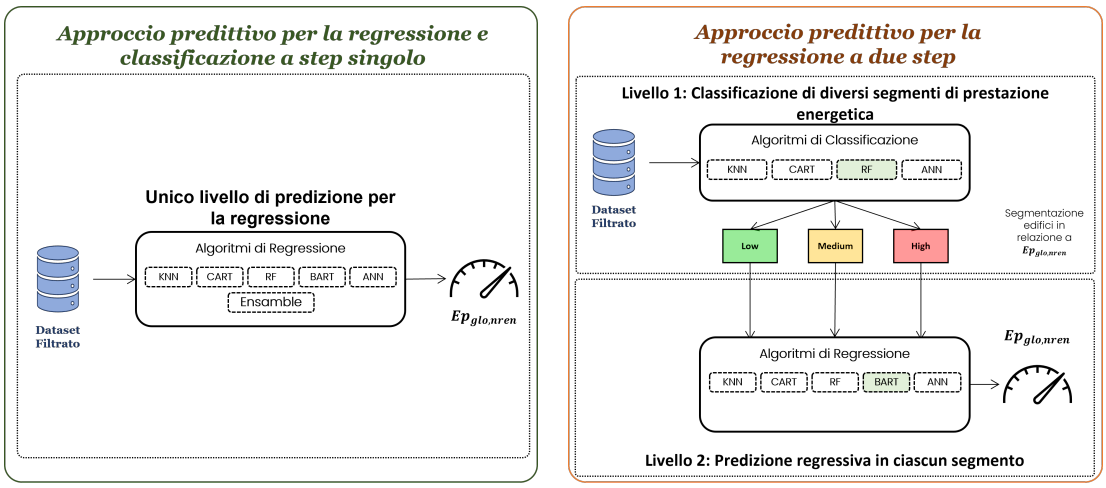


Figura 3.3: Approcci per la predizione

In particolare per il modello di stima sono state valutate due differenti metodologie (Figura 3.3):

- *Approccio a un livello di analisi* nel quale l'output è stimato in maniera diretta dal modello. Il dataset di riferimento viene suddiviso in un sottoinsieme

di *train*, con il quale allenare il modello, e in un sottoinsieme di *test*, necessario alla validazione del modello sviluppato. *Train* e *test* corrispondono relativamente all'80% e al 20% del dataset iniziale. L'obiettivo è fornire la giusta quantità di informazioni al modello, affinché sia in grado di carpire tutte le relazioni esistenti fra input ed output, senza però cadere nel problema dell'*overfitting*, che si presenta quando esso si adatta troppo ai dati del *train*, tanto da non riuscire ad essere efficiente nella predizione di nuove configurazioni di dati del *test*, che non ha mai processato.

- *Approccio a due livelli di analisi*, ispirandosi al modello HEDEBAR di [9], ogni istanza del dataset viene contrassegnata con un'etichetta di consumo, ricavata dalla suddivisione della distribuzione dell'$Ep_{glo,nren}$ in tre segmenti di consumo *Low, Medium e High*. Successivamente viene allenato un modello classificatore in grado di attribuire ad ogni nuova istanza la giusta etichetta, che servirà infine per ridistribuire il dataset in tre gruppi distinti, per i quali verrà predisposto un relativo modello di regressione. Il raggruppamento degli edifici sulla base dei consumi simili, permette di diminuire la varianza del sottoinsieme di dati generando un modello più accurato.

3.5 Obiettivo 5 ~ Layer di XAI

Strettamente correlato all'obiettivo 4, il layer di XAI ha lo scopo di rendere interpretabili le relazioni fra input ed output rilevate dal modello black-box, sfruttando tali informazioni per valutare le performance future di un nuovo edificio ed eventuali azioni di retrofit per poterle migliorare. Sono stati implementati dei modelli di interpretabilità globali, in grado di valutare come gli attributi di input influenzino mediamente l'output, e l'importanza che essi assumono nell'effettuare la predizione. Per l'interpretazione di un'istanza specifica, sono stati sviluppati metodi di interpretabilità locale basati sul *breakDown plot* in grado di valutare la quota parte di ciascun attributo nella determinazione del valore del risultato finale. In questo modo è possibile evidenziare quali attributi, contribuiscano ad un aumento dell'indice di prestazione stimato, agendo in maniera mirata sui loro valori per ottenere un risultato più performante.

Capitolo 4

Risultati

In questo capitolo verranno presentati i risultati degli obiettivi costituenti il framework metodologico discusso nel capitolo precedente. Le analisi metodologiche, sono state sviluppate con il software statistico R [36] nella loro totalità.

4.1 Obiettivo 1

Questa sezione riassume le tecniche di pre-elaborazione dei dataset, al fine di ottenere un campione di *APE* statisticamente valido su cui condurre le analisi successive. Unendo le informazione contenute negli open data, descritti in Tabella 1.1, sono stati ottenuti circa 500 000 certificati energetici, redatti dal 2015 al 2021 e riferiti alla Regione Piemonte, ognuno di essi contrassegnato da 134 attributi univoci caratterizzanti l'edificio dal punto di vista geografico, geometrico e termofisico.

4.1.1 Identificazione e pre-elaborazione del dataset di attestati di riferimento

L'obiettivo dell'analisi degli attestati è di riuscire a stimare con un tasso di errore relativamente basso le prestazioni energetiche di un edificio, sfruttando le informazioni intrinsecamente contenute negli attestati stessi. Per raggiungere tale scopo e per poter usufruire di dati con un valore qualitativo significativo, è stato ritenuto necessario effettuare delle selezioni preliminari degli attestati disponibili, secondo

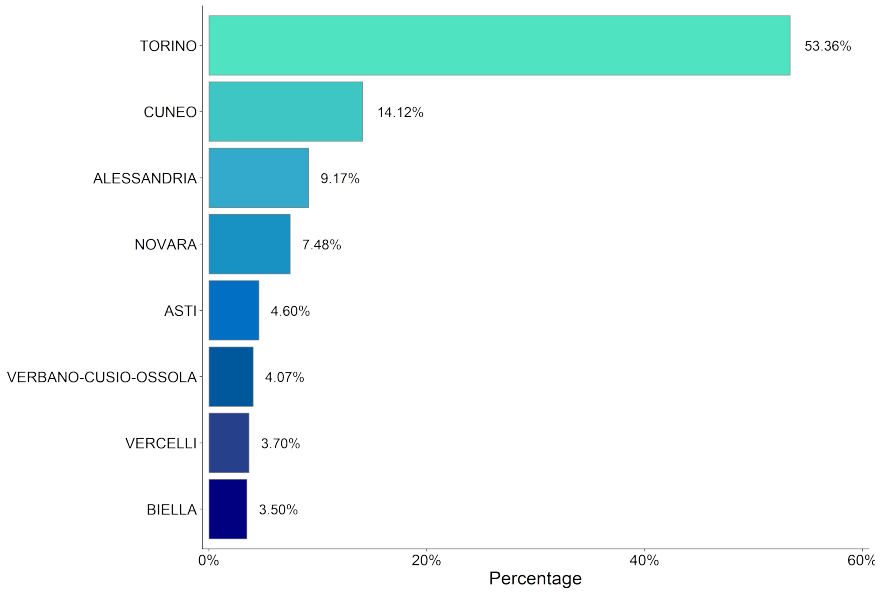


Figura 4.1: Distribuzione delle province piemontesi nel dataset

diversi criteri statistici e logici basati sulla conoscenza della materia, al fine di delineare un dominio di esistenza degli edifici.

Filtro temporale

Il parametro scelto per la prima scrematura del dataset è la data di invio dell'attestato energetico al sistema informativo. In base alle ultime modifiche apportate alle leggi di bilancio si è ritenuto opportuno considerare solamente gli attestati inviati nell'ultimo triennio, ovvero dal 2019 fino al 2021. Questo perché negli ultimi anni vi è stata una larga diffusione di incentivi e bonus relativi alla riqualificazione edilizia, sostenuti dal Decreto Legge 19 maggio 2020, n. 34, meglio noto come *Decreto Rilancio* [37]. La scelta di questo determinato intervallo temporale consente di disporre di un insieme di certificati più omogeneo rispetto a quelli redatti attualmente, in modo da poter sfruttare una base statistica più significativa.
In seguito a tale campionamento il dataset è stato ridotto del 50%, contando circa 250 000 attestati.

Filtro sulla Destinazione d'uso e oggetto attestato

Se il range temporale è stata una scelta aprioristica, le successive selezioni del dataset sono state effettuate andando a valutare la distribuzione della tipologia di attributi categorici presente in esso. Il primo attributo ritenuto maggiormente significativo è la *destinazione d'uso* dell'edificio certificato, che permette di stabilire quale categoria edilizia, Tabella 1.3 a pagina 14, sia maggiormente caratteristica del parco edilizio piemontese censito dalla certificazione energetica.

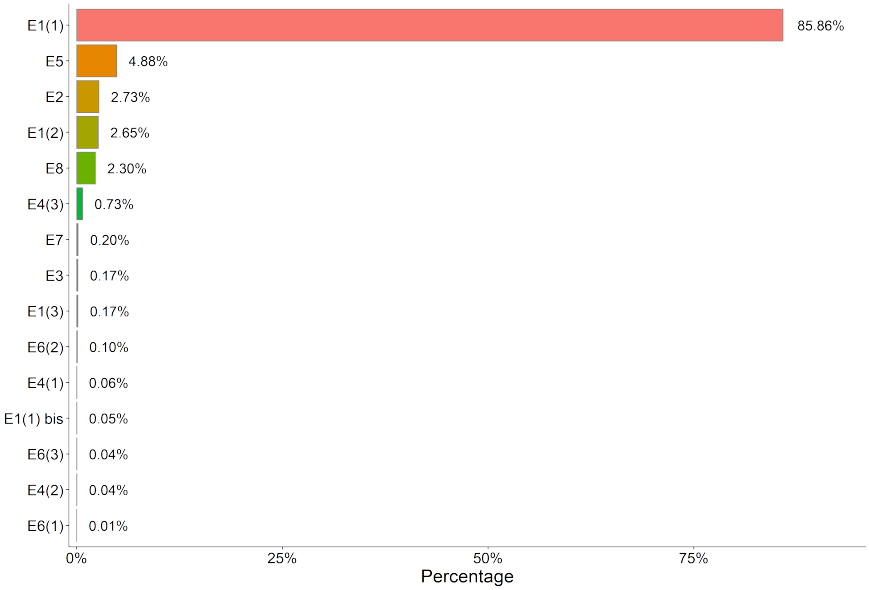


Figura 4.2: Distribuzione delle destinazioni d'uso all'interno del dataset

Dalla Figura 4.2 emerge come la categoria edilizia E1(1), riferita agli edifici adibiti a residenza a carattere continuativo, siano nettamente preponderanti rispetto alle altre. Per questo motivo si è scelto di focalizzare l'analisi solamente sugli edifici residenziali appartenenti alla suddetta destinazione d'uso. Su tale campione è stata effettuata un'ulteriore cernita relativa all'oggetto del certificato.

La Figura 4.3 mostra come circa l'86% degli immobili corrisponda a un'unità immobiliare appartenente a un edificio pluriunità, generalmente appartamenti di un condominio. Dunque analogamente a quanto deciso per le destinazioni d'uso, è stata considerata per l'analisi solamente questa tipologia di immobile.

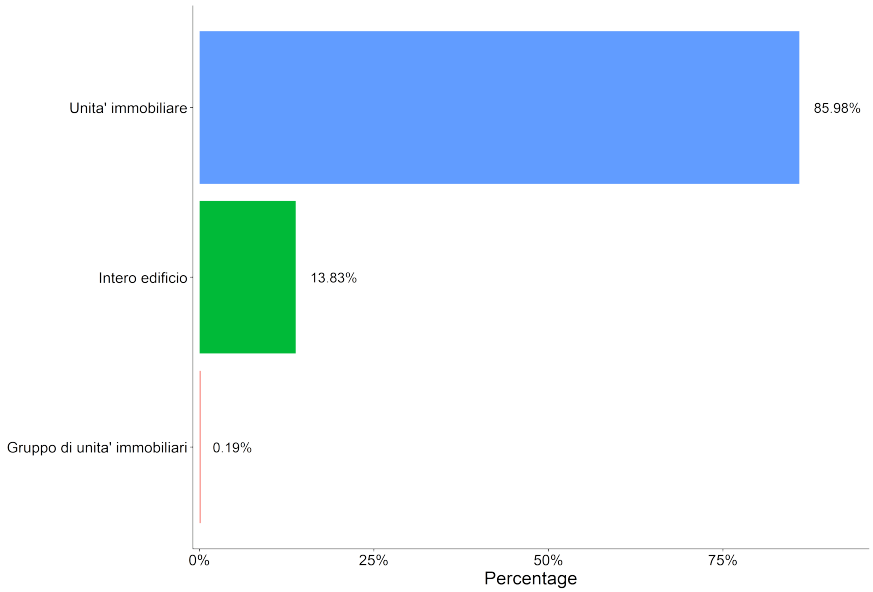


Figura 4.3: Distribuzione tipologie di immobile oggetto dell'attestato

Filtro sui Servizi energetici

Dopo aver ristretto il campo di analisi ai soli edifici residenziali è stata effettuata una valutazione sui servizi energetici utilizzati. Gli open data contenevano i dati relativi agli impianti installati, ognuno dei quali associato a un determinato edificio. Si presenta una situazione nel quale generalmente ogni edificio possiede un singolo impianto dedito al riscaldamento invernale e un singolo impianto destinato alla produzione di acqua calda sanitaria. I casi in cui un immobile disponga di più impianti dedicati allo stesso servizio, o a servizi diversi dai due precedentemente elencati sono molto limitati, e per tali ragioni trascurati dall'analisi (Figure 4.4 e 4.5).

Con questa configurazione, ad ogni singolo impianto installato nell'edificio è possibile associare un determinato combustibile. In Figura 4.6 sono rappresentate le prime sette tipologie di combustibile maggiormente utilizzato dagli impianti. Si osserva come il gas naturale sia la risorsa più utilizzata per il soddisfacimento dei servizi selezionati, quasi in egual misura per il riscaldamento e per la produzione di acqua calda sanitaria *(Domestic Hot Water: DHW)*. L'energia elettrica invece viene utilizzata prevalentemente per la DHW, ipotizzando una presenza discreta di boiler elettrici nel parco edilizio.

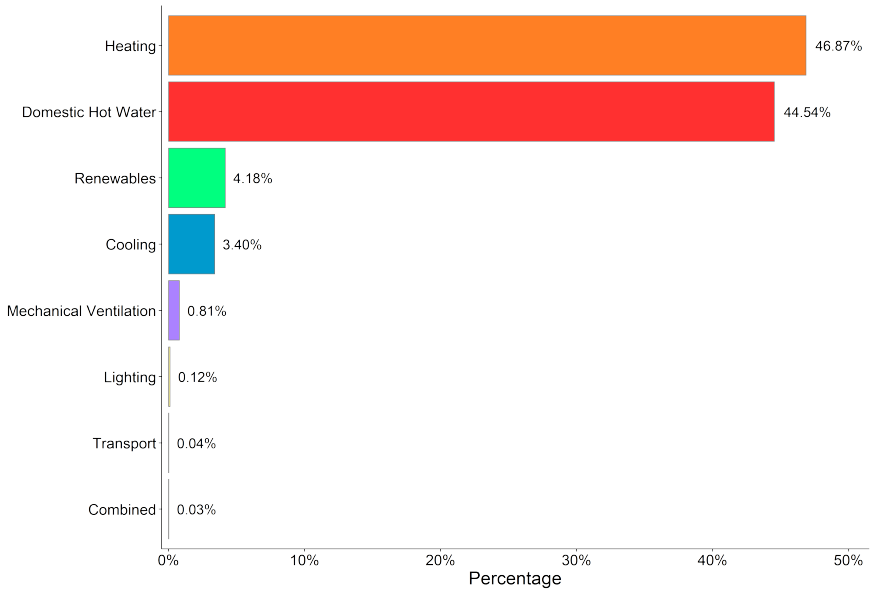


Figura 4.4: Distribuzione servizi energetici

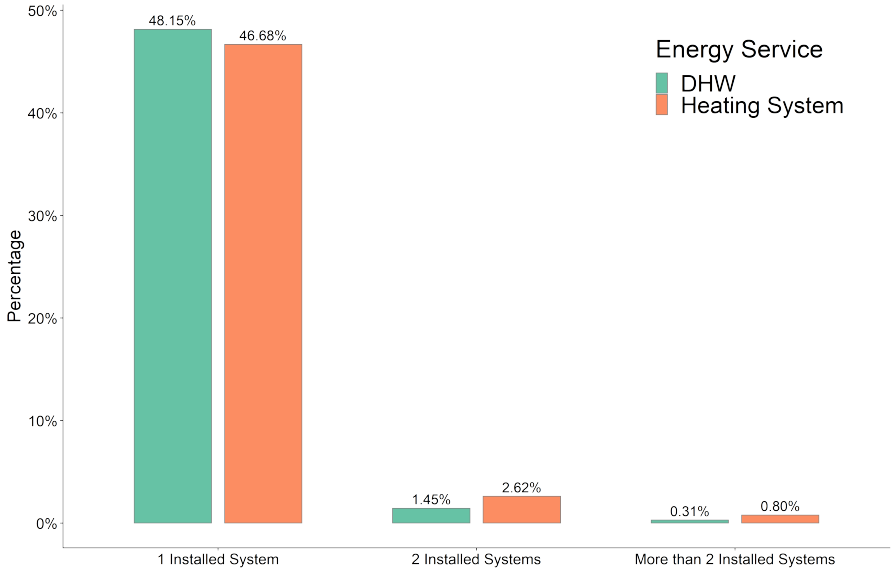


Figura 4.5: Distribuzione edifici in base al numero di impianti installati e al servizio energetico fornito

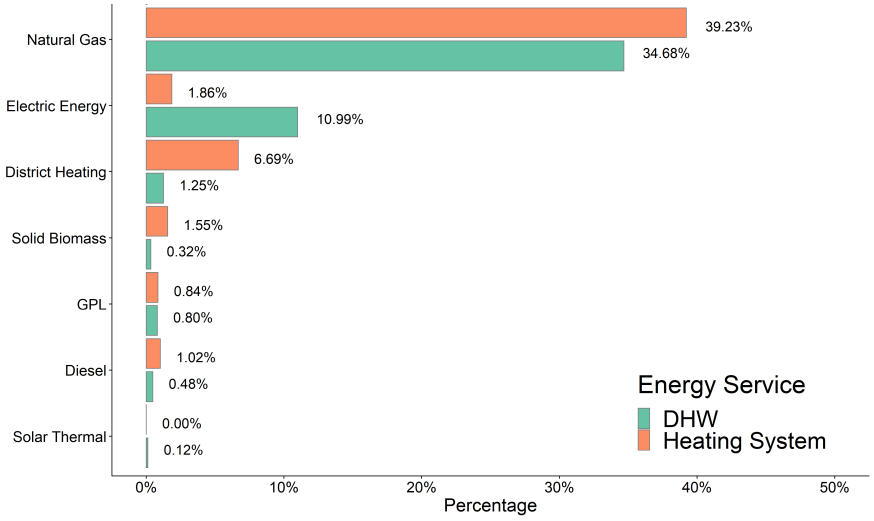
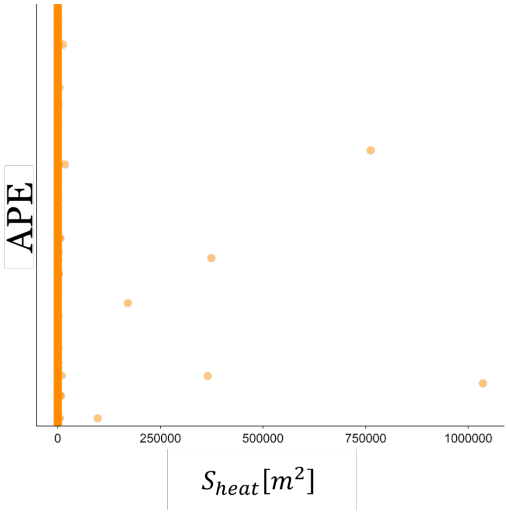


Figura 4.6: Distribuzione del combustibile utilizzato dagli impianti

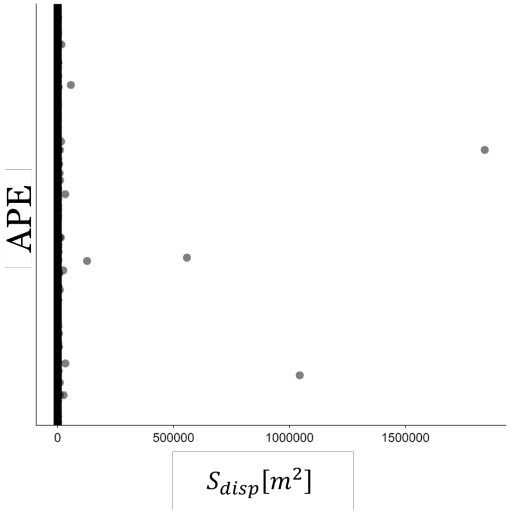
4.1.2 Rimozione valori anomali

Conseguentemente alla fase preliminare di filtraggio, la base di dati di cui si dispone è composta da circa 120 000 *APE*, relativi ad unità immobiliari residenziali di categoria E1(1), con mono impianto di acqua calda sanitaria e di riscaldamento invernale, certificate nel triennio 2019-2021.

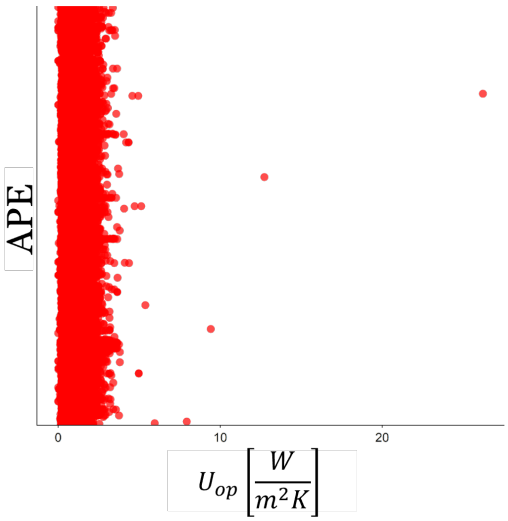
I dati dei certificati sono inseriti manualmente da esseri umani, e come tutti i dati reali possono essere soggetti ad imprecisioni, inconsistenze e anomalie che comportano una perdita inevitabile della qualità dell'informazione raccolta. Le risposte ottenute dalle metodologie data-driven sviluppate in seguito sono fortemente dipendenti dalla bontà del dato in ingresso, rispecchiando il concetto ben noto in ambito informatico *"Garbage in, garbage out"* [38]. Tutte le anomalie, i rumori e le incongruenze presenti in input, si riversano nell'output. Diventa dunque necessario individuare ed eliminare tali valori, al fine di fornire risultati attendibili e veritieri nelle analisi successive. L'approccio con il quale è stata effettuata la pulizia degli attestati energetici si basa sulle conoscenze dell'esperto del dominio nell'individuare le variabili maggiormente significative e ad esse attribuire un range di esistenza. In particolare in seguito allo studio monovariato di tali attributi selezionati, sono stati applicati dei filtri (riassunti in Tabella 4.1) a variabili geometriche del fabbricato, superfici e volumi, a variabili termofisiche, come le trasmittanze medie delle superfici di involucro, e sulle variabili inerenti al rendimento dei sistema di riscaldamento.



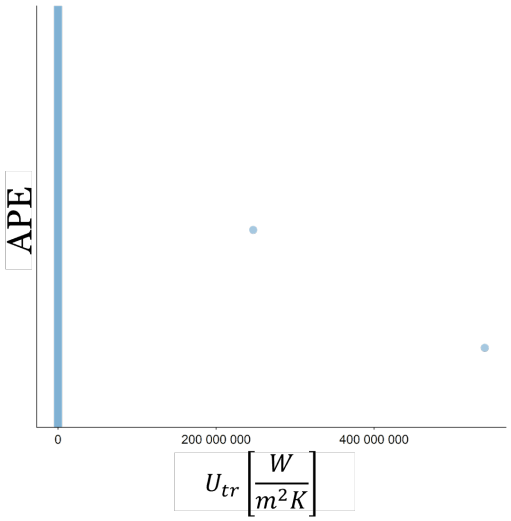
(a) Anomalie dati Superficie Riscaldata



(b) Anomalie dati Superficie Disperdente



(c) Anomalie dati Trasmittanza media superfici opache



(d) Anomalie dati Trasmittanza media superfici trasparenti

Figura 4.7: Scatter plot dei valori prima dell'imposizione dei filtri

Variabile	**Simbolo**	**Dominio**
Superficie riscaldata	$S_{heat}[m^2]$	$\in$ [20,250]
Superficie disperdente	$S_{disp}[m^2]$	$\in [5, \infty)$
Rapporto di forma	$S/V[m^{-1}]$	$\in$ [0.1,1.5]
Volume lordo riscaldato	$V_{heat}[m^3]$	$\in$ [25,1000]
Trasmittanza media superfici opache	$U_{op}[\frac{W}{m^2K}]$	$\in$ [0.1,3]
Trasmittanza media superfici trasparenti	$U_{tr}[\frac{W}{m^2K}]$	$\in$ [0.9,7]
Rendimento globale medio stagionale riscaldamento	$\eta_H[-]$	$\in$ [0,1]
Rendimento globale medio stagionale ACS	$\eta_W[-]$	$\in$ [0,1]
Rendimento sottosistema distribuzione	$\eta_d[-]$	$\in$ [0.75,1]
Rendimento sottosistema emissione	$\eta_e[-]$	$\in$ [0.85,1]
Rendimento sottosistema generazione	$\eta_g[-]$	$\in$ [0.5,1.2]
Rendimento sottosistema regolazione	$\eta_r[-]$	$\in$ [0.6,1]
Anno di costruzione	$y_c[year]$	$\in$ [1700,2021]
Fabbisogno di energia primaria non rinnovabile per la climatizzazione invernale	$Ep_{H,nren}$	$\in$ [0,700]
Fabbisogno di energia primaria non rinnovabile per ACS	$Ep_{W,nren}$	$\in$ [0,200]

Tabella 4.1: Dominio di esistenza degli attributi per la rimozione dei valori anomali

Come è possibile osservare dalla Figura 4.7 alcuni certificati presentano valori completamente fuori scala per gli attributi selezionati, tanto che non è possibile apprezzare la loro distribuzione effettiva. La scelta del dominio per ogni attributo mira ad eliminare tali anomalie e ad inquadrare le informazioni anche in funzione delle scelte di analisi adottate. Ad esempio, nel caso del rendimento del sottosistema di generazione, valori superiori all'unità non sono delle anomalie fisiche, poiché sovente viene inserito il COP delle pompe di calore che assume valori tipici compresi fra 3-6. Con la scelta dell'intervallo di $\eta_g \in [0.5,1.2]$ si sono di fatto escluse tutte le macchine termiche a ciclo inverso, poiché ritenute statisticamente irrilevanti per l'analisi dal momento in cui essa si concentra sugli appartamenti condominiali, in cui sono rari gli impianti installati di questo tipo. Inoltre il limite superiore non è stato imposto pari all'unità per considerare anche le caldaie a condensazione, il cui rendimento supera il 100% a causa della mancata valutazione del calore latente di condensazione nel rapporto fra energia utile ed energia spesa. Dopo aver filtrato gli attributi caratteristici dell'edificio è stato ritenuto opportuno

verificare le distribuzioni dei valori degli indici di prestazione inseriti nei certificati, rimuovendo gli attestati caratterizzati da valori anomali, rimasti come refuso dai filtri precedenti. Essi sono stati limitati solamente superiormente con i valori visualizzabili in Tabella 4.1.

Attributo filtrato	**Cardinalità**	**Δ Percentuale**
	119 776	0%
Superficie riscaldata	118 125	1.38%
Superficie disperdente	118 003	0.10%
Rapporto di forma	117 487	0.44%
Volume lordo riscaldato	117 210	0.24%
Trasmittanza media superfici opache	116 972	0.20%
Trasmittanza media superfici trasparenti	116 186	0.67%
Rendimento globale medio stagionale riscaldamento	112 065	3.55%
Rendimento globale medio stagionale ACS	111 055	0.90%
Rendimento sottosistema distribuzione	69 095	37.78%
Rendimento sottosistema emissione	68 862	0.34%
Rendimento sottosistema generazione	47 815	30.56%
Rendimento sottosistema regolazione	47 431	0.80%
Anno di costruzione	47 288	0.30%
Fabbisogno di energia primaria non rinnovabile per la climatizzazione invernale	46 662	1.32%
Fabbisogno di energia primaria non rinnovabile per ACS	46 464	0.42%

Tabella 4.2: Decremento cardinalità del dataset

La Tabella 4.2 illustra come si sia ridotta la cardinalità del dataset in seguito ad ogni filtro adottato. É possibile osservare come il troncamento più significativo sia dovuto ai rendimenti. Per evitare di imporre un dominio troppo restringente, che avrebbe escluso un numero eccessivo di attestati, la scelta dei limiti di esistenza dei rendimenti di distribuzione e generazione è stata coadiuvata da un'analisi di sensitività (Figura 4.9), attraverso la quale è stato individuato il valore del rendimento che non escludesse dall'analisi un numero eccessivo di attestati validi.

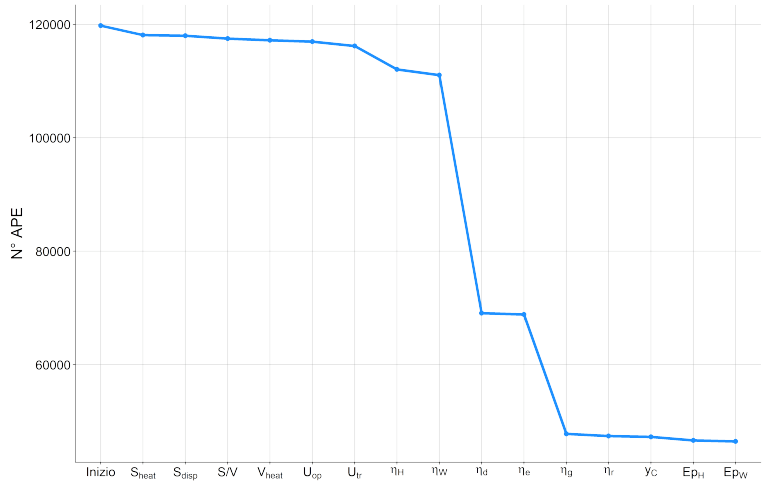
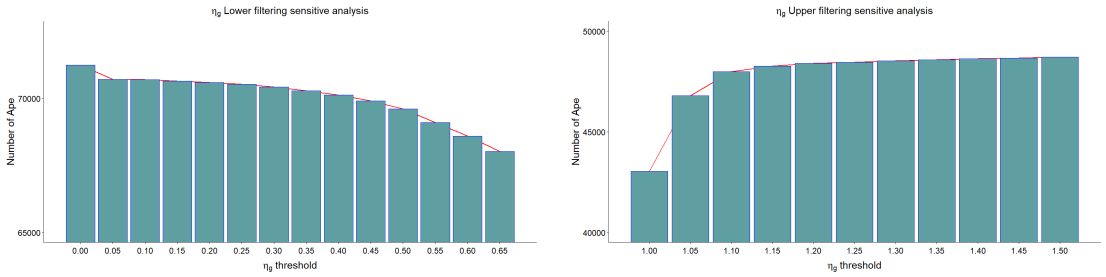


Figura 4.8: Decremento cardinalità del dataset



(a) Limite inferiore η_g

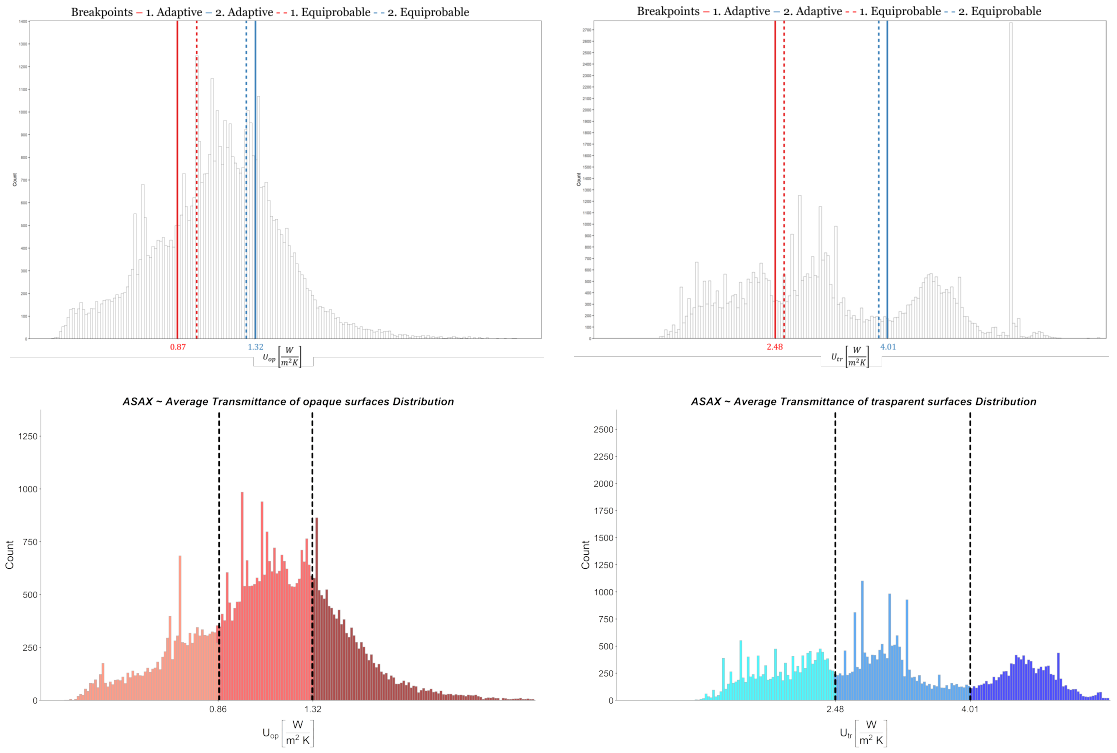
(b) Limite superiore η_g

Figura 4.9: Analisi sensitività rendimenti sottosistema

4.2 Obiettivo 2

L'obiettivo 2 è incentrato nel fornire delle visualizzazioni efficaci nel caratterizzare il dataset di attestati. Successivamente all'operazione di pulizia e filtraggio dei dati, sono state analizzate le distribuzioni delle variabili di maggior interesse segmentandole in tre range, in modo da individuare a primo impatto quali siano i punti di confine fra gli intervalli contenenti i valori ritenuti bassi, medi o alti.

Proprio per concentrare l'attenzione sul valore di soglia, la procedura di segmentazione è stata effettuata sfruttando l'algoritmo di determinazione adattiva basato sul *k-means* presentato nel capitolo 3. Il vantaggio di questo metodo consiste nell'utilizzo di intervalli di ampiezza variabile per raggruppare gli elementi



(a) Trasmittanza media superfici opache (b) Trasmittanza media superfici trasparenti

Figura 4.10: Distribuzione delle trasmittanza evidenziando i tre intervalli di valori

più omogenei fra loro. La Figura 4.10 mostra la distribuzione delle conducibilità termiche medie delle superfici di involucro, sezionate con il metodo enunciato, e di come siano cambiati i limiti passando da una distribuzione equiprobabile ad una adattiva.

Nel caso in cui si disponga già di tutte le informazioni necessarie per la redazione di un nuovo attestato, lo strumento di visualizzazione può rivelarsi utile per confrontare i valori dei vari attributi in possesso rispetto alla loro distribuzione nel dataset. La collocazione del valore all'interno di uno dei range evidenziati, determina l'entità dell'attributo stesso. Sfruttando la Figura 4.10 come esempio, se per un edificio è stato valutato un valore di $U_{op} = 3.5 \frac{W}{m^2 K}$, è evidente come esso si collochi nel range di valori elevati per quell'attributo, rispetto alla media. Estendendo questa analisi a più input si riesce ad avere un quadro preliminare su quelli che potrebbero essere i consumi dell'edificio in esame.

Gli stessi intervalli sono stati successivamente utilizzati per condurre delle analisi in frequenza che riassumessero i pattern principali e la loro diversificazione, ad esempio rispetto all'anno di costruzione e alla tipologia edilizia (Figura 4.11). La rappresentazione degli attributi rispetto ad altri fattori, e non soltanto tramite un'analisi monovariata, consente una caratterizzazione più efficace dell'informazione.

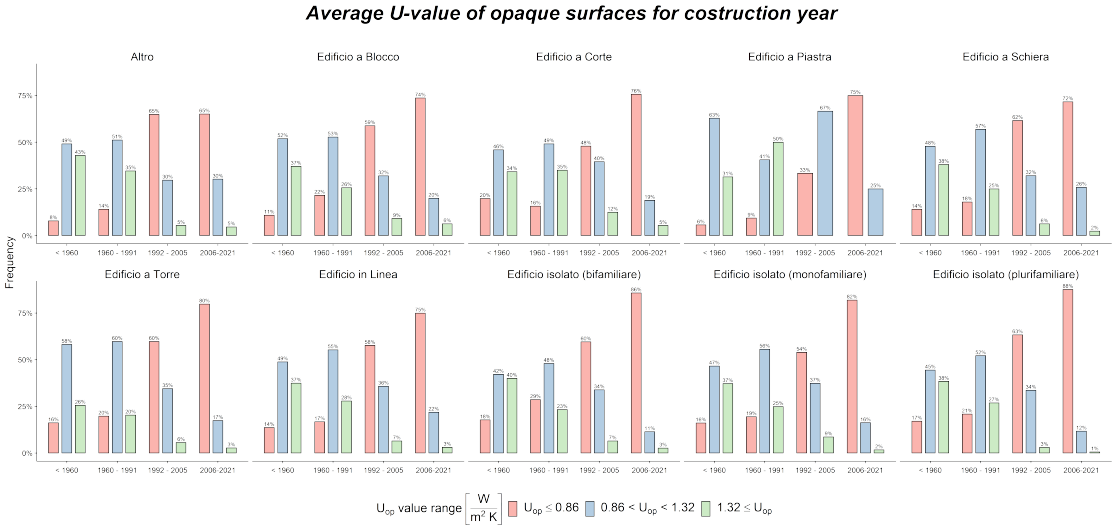


Figura 4.11: Distribuzione U_{op} per tipologia edilizia e periodo di costruzione

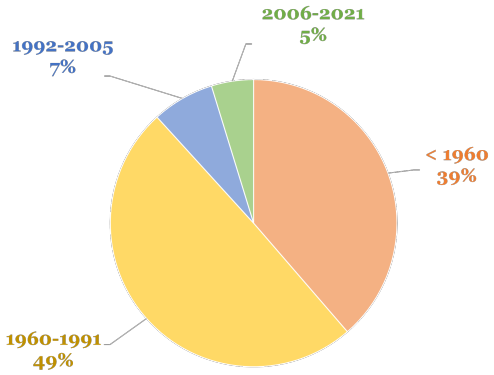


Figura 4.12: Distribuzione *APE* per periodo di costruzione

Sono stati evidenziati quattro differenti raggruppamenti per il periodo di costruzione rappresentati in Figura 4.12. Il primo include il 39% del dataset ed è composto dagli edifici costruiti prima del 1960, caratterizzati generalmente da

carenti proprietà termofisiche, com'è possibile osservare dai valori della conducibilità termica, per i quali è consigliabile attuare una ristrutturazione energetica. Il secondo intervallo è costituito dagli edifici costruiti tra il 1960 e il 1991, anno in cui è entrata in vigore la Legge 10, una fra le prime normative riguardanti la relazione energetica di un edificio, e ad esso afferisce il 49% del dataset. Il terzo e il quarto intervallo, caratterizzanti relativamente il 7% e il 5%, sono relativi agli edifici di nuova costruzione, suddivisi a monte e a valle del 2005, anno in cui è stata emanata in Italia la direttiva Europea *EPBD*, che ha sancito la nascita del "Certificato Energetico". Seguendo la direzione tracciata dalle normative è possibile osservare il miglioramento della qualità dell'involucro edilizio nel corso degli anni; se soltanto il 10%-20% degli edifici costruiti prima degli anni Sessanta presentava valori ottimali per la trasmittanza media delle superfici opache, $U_{op} \leq 0.86 \frac{W}{m^2 K}$, per gli edifici moderni tale percentuale si assesta fra il 65% e l' 88%.

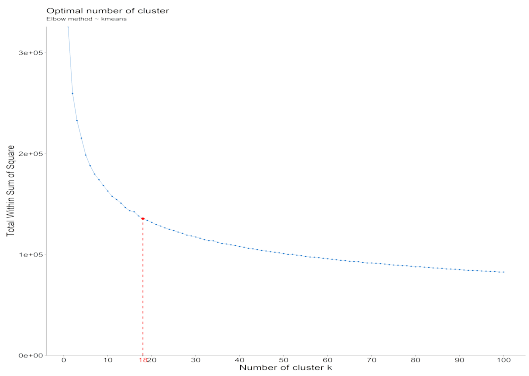
4.3 Obiettivo 3

In questa parte dello studio per ottenere un modello in grado di partizionare il dataset in sottogruppi di certificati simili, sono state valutate tre differenti metodologie di clustering: il *k-means* e il *CLARA*, afferenti alla tipologia di clustering partitivi, e il clustering *gerarchico divisivo*. Per ognuno di essi è stato stimato il numero ottimale di cluster attraverso il *metodo del gomito*. Come è possibile osservare in Figura 4.13, vengono restituiti 18 cluster per il *k-means*, 17 per il *CLARA* e 21 per il *gerarchico*. Seppur non coincidenti, tali risultati si concentrano in un intorno specifico di valori, che con buona approssimazione contiene il numero di raggruppamenti ideale con cui identificare il dataset di certificati energetici.

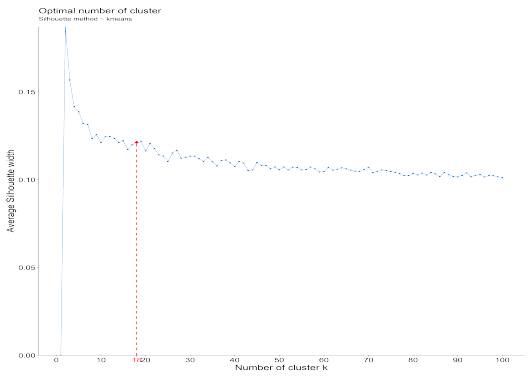
Metodo di Clustering	*SST*	*Shilouette*
k-means	14 3750	0.13
CLARA	16 2500	0.09
gerarchico	15 6250	0.05

Tabella 4.3: metriche di validazione del clustering

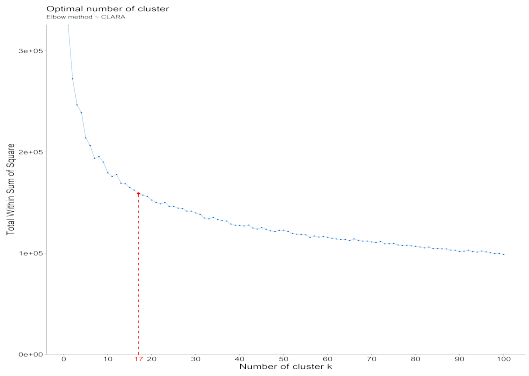
Per stabilire quale metodo sia più efficace nel ripartizionare il dataset, ci si è avvalsi della stima della *SST* e dell'indice di *Silhouette* (Figura 4.13), valutati per ciascun metodo in funzione del loro numero ottimale di cluster.



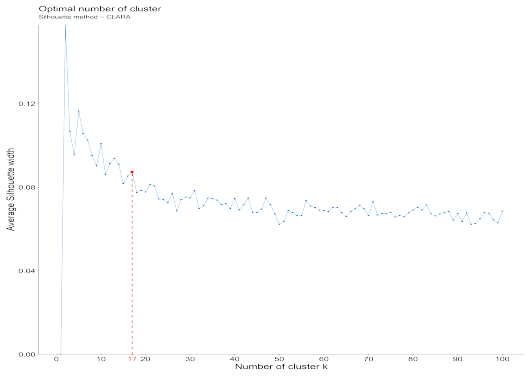
(a) Cluster ottimali *k-means*



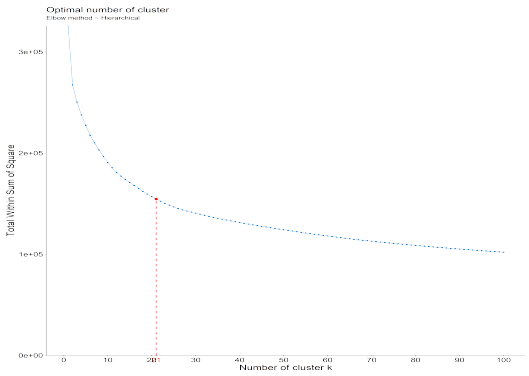
(b) Indice di Silhouette *k-means*



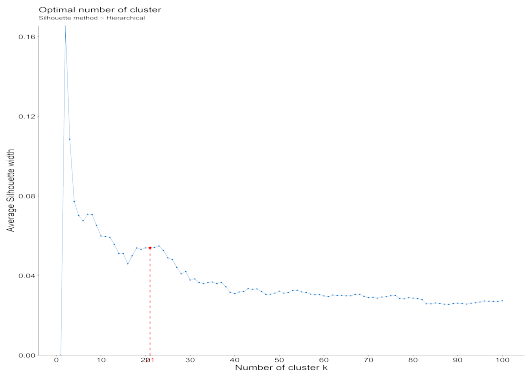
(c) Cluster ottimali *CLARA*



(d) Indice di Silhouette *CLARA*



(e) Cluster ottimali *gerarchico*



(f) Indice di Silhouette *gerarchico*

Figura 4.13: Metodo del gomito e indice di Silhouette

La Tabella 4.3 riassume i risultati ottenuti, evidenziando come l'algoritmo *k-means* presenti i valori migliori per entrambe le metriche di validazione adottate.

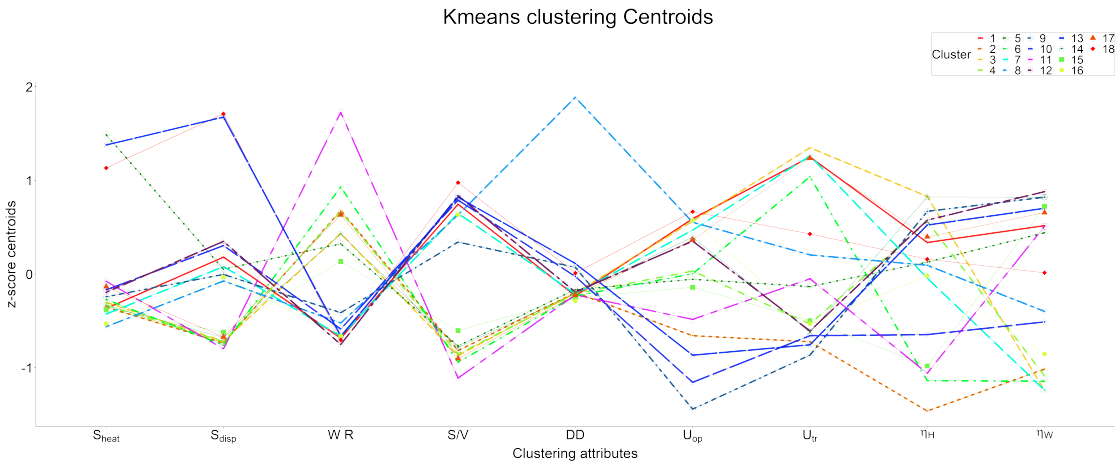


Figura 4.14: Coordinate parallele centroidi *k-means*

Cluster	**Percentuale**	S_{heat}	S_{disp}	W_R	S/V	DD	U_{op}	U_{tr}	η_H	η_H
1	6%	63.21	184.22	0.07	0.72	2639	1.32	4.91	0.73	0.65
2	5%	63.70	82.23	0.16	0.34	2647	0.87	2.36	0.51	0.37
3	4%	65.38	84.39	0.14	0.33	2646	1.31	5.04	0.80	0.33
4	3%	64.36	83.81	0.16	0.34	2644	1.11	2.61	0.76	0.36
5	7%	125.04	169.82	0.13	0.36	2654	1.08	3.12	0.71	0.64
6	4%	66.42	79.65	0.17	0.32	2639	1.10	4.64	0.55	0.35
7	6%	61.21	173.02	0.07	0.70	2644	1.27	4.93	0.69	0.33
8	5%	56.71	155.87	0.08	0.69	3131	1.30	3.56	0.70	0.49
9	6%	67.36	163.64	0.08	0.62	2709	0.58	2.18	0.78	0.71
10	7%	69.81	197.84	0.07	0.73	2687	0.69	2.45	0.61	0.47
11	7%	72.87	75.36	0.23	0.27	2641	0.93	3.23	0.56	0.65
12	7%	68.93	202.88	0.06	0.74	2648	1.23	2.51	0.76	0.72
13	4%	121.40	350.34	0.07	0.74	2719	0.79	2.32	0.76	0.69
14	5%	64.20	88.85	0.14	0.35	2644	1.25	2.49	0.79	0.71
15	5%	62.69	95.30	0.12	0.40	2635	1.05	2.65	0.57	0.69
16	6%	57.72	160.31	0.07	0.69	2626	1.31	2.61	0.69	0.40
17	7%	70.97	89.31	0.15	0.32	2640	1.24	4.90	0.74	0.68
18	6%	113.23	354.30	0.06	0.78	2695	1.34	3.85	0.71	0.56

Tabella 4.4: Centroidi *k-means*

Dopo aver confermato il *k-means* come metodologia più efficiente, sono stati valutati i centroidi identificativi di ogni cluster. In uno spazio *n-dimensionale*, il centroide corrisponde a un vettore contenente le media dei valori di ogni *n-esimo* attributo del cluster. Esso rappresenta un edificio fittizio di riferimento per il suo raggruppamento e viene visualizzato attraverso un grafico a coordinate parallele (Figura 4.14), dove viene rappresentato come una linea spezzata. I valori

sono normalizzati in z-score per garantire una visualizzazione qualitativa del peso di ogni attributo sulla diversificazione del cluster. A causa del numero ingente di dimensioni, molti cluster risultano molto simili fra loro ad eccezione di pochi attributi divergenti. I cluster 13 e 18 ad esempio, hanno un andamento pressoché analogo ad eccezione della U_{op} e della U_{tr} per la quale il cluster 13 descrive valori più elevati rispetto al cluster 18. In Tabella 4.4 sono riportati i valori reali, non normalizzati, dei centroidi e la distribuzione percentuale degli elementi del dataset, dalla quale si evince come i vari attestati siano distribuiti omogeneamente in ogni cluster, non essendocene uno che presenti una cardinalità preponderante rispetto agli altri.

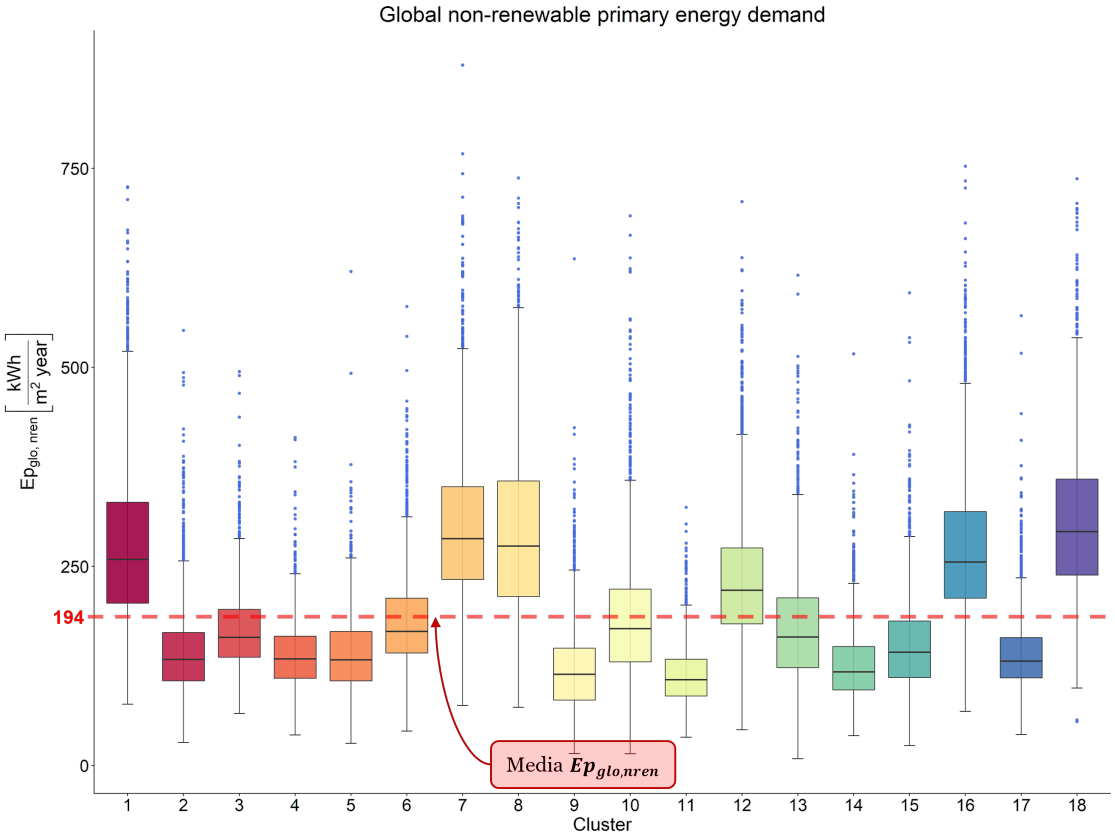


Figura 4.15: Distribuzione $Ep_{glo,nren}$ per cluster

Gli input scelti per l'analisi sono fortemente determinanti il valore dell'$Ep_{glo,nren}$, per cui i cluster caratterizzano gli edifici anche dal punto di vista delle prestazioni. Dalla Figura 4.15, che rappresenta le distribuzioni dell'indice di prestazione in funzione del cluster, è possibile osservare come i raggruppamenti 7, 8 e 18 identifichino edifici dal consumo elevato, superiore alla media dell'indice di prestazione valutata su tutto il dataset, diversamente dal 8, 11 e 14 per i quali il valore dell'$Ep_{glo,nren}$

è tendenzialmente al di sotto. Spiccano anche i cluster 1 e 16 per la loro somiglianza; nonostante essi presentino dei valori dell'indice di consumo analoghi, la configurazione degli attributi che li ha condotti verso questo risultato è differente. In Figura 4.16 si desume come gli attributi determinanti la divergenza dei due cluster siano principalmente la conducibilità termica dell'involucro trasparente e il rendimento globale per il riscaldamento. Il cluster 1 è caratterizzato da edifici con valori tendenzialmente più elevati di tali attributi rispetto al 16

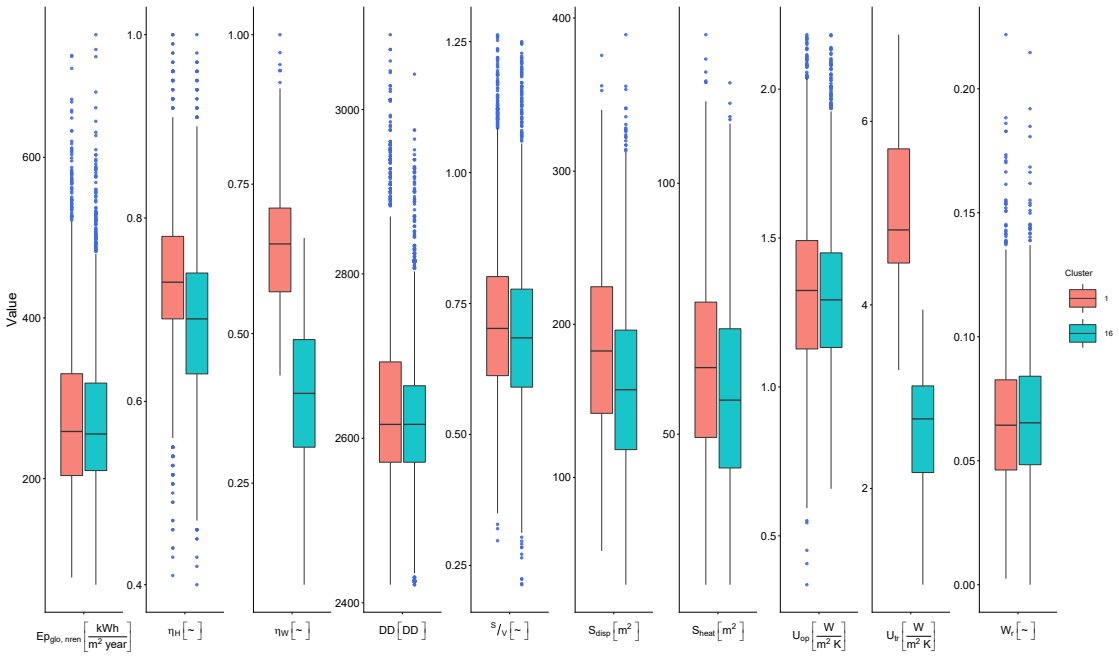


Figura 4.16: Confronto cluster 1 e 16

Considerazioni simili possono essere effettuate anche per tutti gli altri cluster che presentano andamenti equiparabili, ma che differiscono per la configurazione dei propri attributi.

Processo di assegnazione di un nuovo oggetto ad un cluster

Lo scopo dell'analisi è di riuscire a sviluppare uno strumento in grado di collocare dei nuovi edifici nel cluster corretto che li caratterizzi maggiormente. Noto il funzionamento dell'algoritmo *k-means* alla base del metodo di clustering, è stato sviluppato un algoritmo in grado di calcolare la minima distanza multidimensionale fra il vettore costituito dagli attributi della nuova istanza che si vuole valutare e i

centroidi dei cluster del dataset. Il nuovo edificio sarà dunque collocato all'interno del cluster identificato dal centroide più vicino ad esso. La Figura 4.17 riporta un esempio di come due nuove istanze siano state attribuite a un determinato cluster in base alla similitudine con il rispettivo centroide.

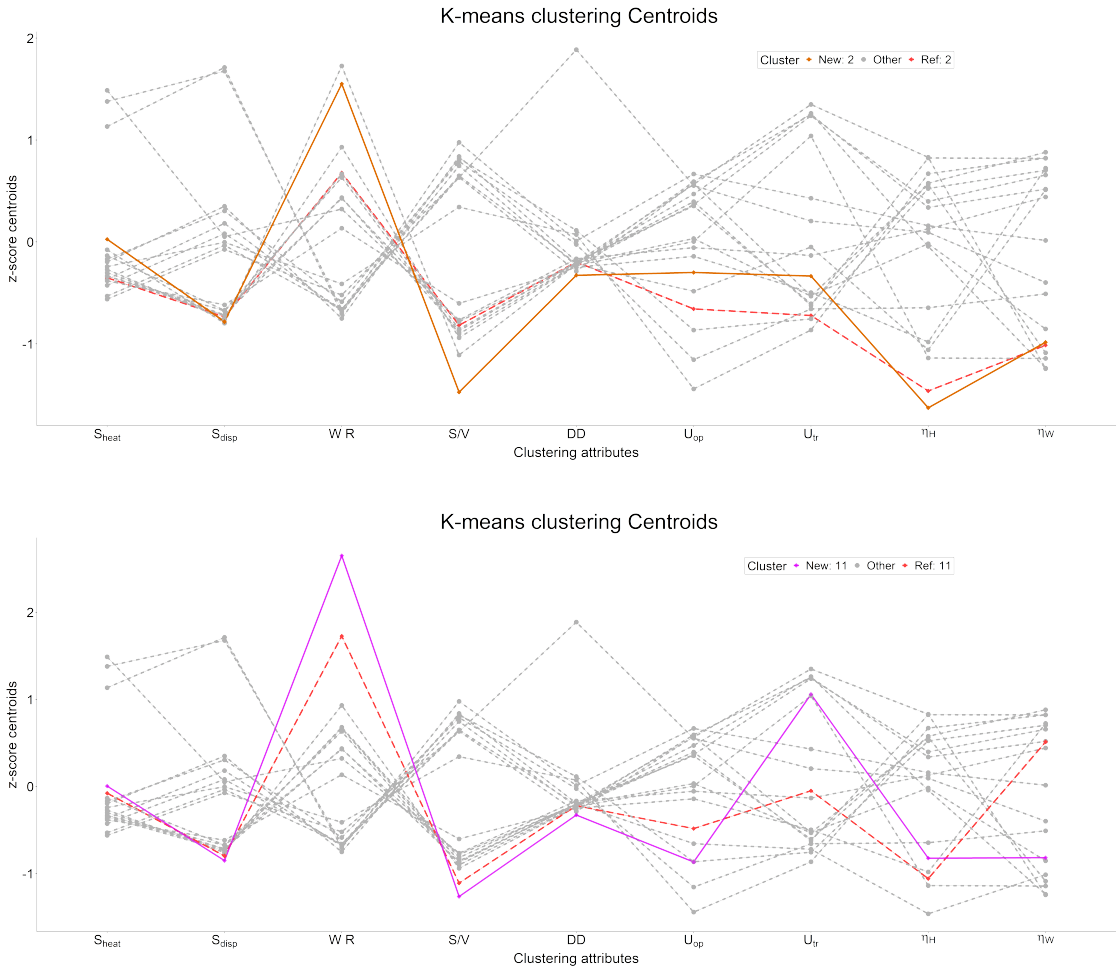


Figura 4.17: Assegnazione di un cluster a un nuovo edificio

4.4 Obiettivo 4

L'obiettivo dei modelli di regressione è stimare il valore continuo dell'$Ep_{glo,nren}$ di un edificio in funzione degli attributi di input descritti in Tabella 3.2. La definizione di un modello data driven di stima, risulta cruciale sia nella fase di progettazione di

edifici di nuova costruzione, per determinare come le variabili progettuali influenzino la prestazione energetica, sia nel caso di edifici esistenti, per la valutazione della fattibilità e dell'impatto di un piano di riqualificazione. Indipendentemente dallo scopo da perseguire, stimare le prestazioni energetiche degli edifici in modo rapido e affidabile, assume un importanza sostanziale. Nell'intento di ottenere il risultato più soddisfacente l'analisi è stata sviluppata seguendo due approcci differenti. Il primo consiste in una metodologia di stima regressiva diretta dell'output, la seconda prende ispirazione dalla metodologia HEDEBAR sviluppata da Attanasio et al. [9], nella quale l'analisi regressiva è stata differenziata in relazione al range di consumo dell'edificio, al fine di incrementare la precisione del modello predittivo.

4.4.1 Approccio a un livello di analisi per lo sviluppo di un modello previsionale regressivo

Per l'approccio regressivo singolo sono stati valutati sei algoritmi di machine learning sulla base di quattro metriche statistiche:

- Coefficiente di determinazione R^2: è un indicatore della validità del modello di regressione, esso esprime la proporzione della varianza dei risultati alla luce della varianza delle variabili indipendenti di input. Assume un valore compreso fra 0 e 1 e, tendenzialmente, maggiore è il valore di R^2, migliori sono le performance del modello considerato;

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y}_i)^2}$$

- *RMSE (Root Mean Squared Error)*: misura l'errore medio commesso dal modello nella stima del risultato. Matematicamente corrisponde alla radice della differenza quadratica media fra il valore reale e il valore stimato dal modello, rapportata al numero totale di osservazioni;

$$RMSE = \sqrt{MSE} = \sqrt{\frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{n}}$$

- *MAE (Mean Absolute Error)*: in maniera analoga all'RMSE, misura di quanto si discosta in media e in maniera assoluta il valore predetto rispetto a quello reale. Il MAE è però meno sensibile alla varianza dei risultati rispetto all'RMSE;

$$MAE = \frac{\sum_{i=1}^{n} |y_i - \hat{y}_i|}{n}$$

- *MAPE (Mean Absolute Percentage Error)*: esprime lo stesso concetto del MAE, ma in forma percentuale.

$$MAPE = \frac{\sum_{i=1}^{n} |y_i - \hat{y}_i|}{n} 100$$

L'analisi è stata sviluppata tramite il pacchetto *Caret* [39] del software statistico *R*, contenente numerosi metodi per l'implementazione di modelli predittivi basati sia sulla regressione che sulla classificazione. Per ogni algoritmo di machine learning implementabile con *Caret*, viene fornito il set di iperparametri che è possibile regolare per ottimizzare il modello. Sfruttando l'interoperabilità del pacchetto, questo procedimento può essere svolto seguendo uno schema logico costante e indipendente dal modello selezionato e dai suoi parametri, rendendo più semplice e scalabile la fase di addestramento.

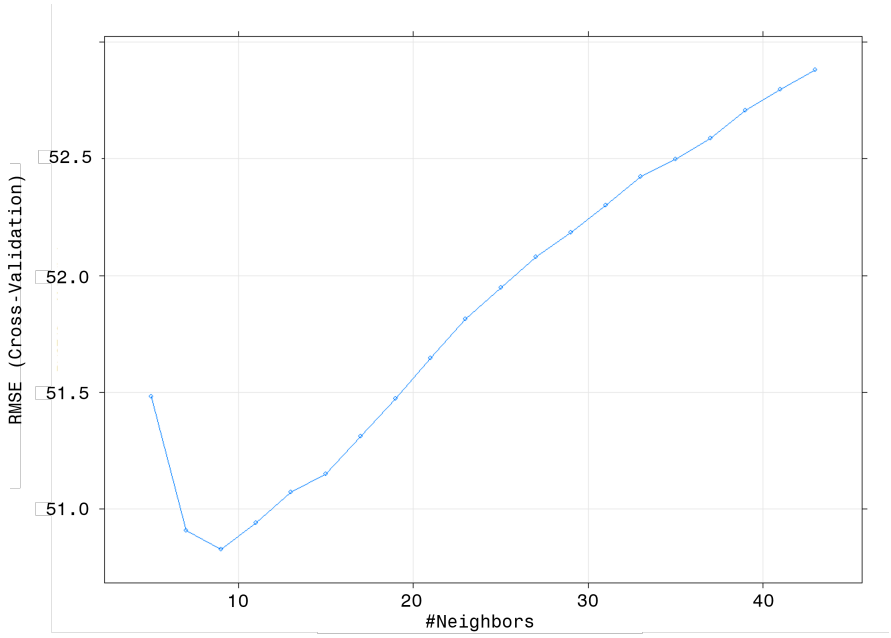


Figura 4.18: *k-fold cross validation* applicata all'iperparametro K dell'algoritmo *KNN*

Ciascun iperparametro, per ogni modello, è stato opportunamente testato tramite la tecnica di *k-fold cross validation*, che valuta le performance del modello, ottenute su sottoinsiemi diversi del dataset di train, calcolando il tasso di errore medio della predizione ottenuta, rispetto al valore reale. I passaggi svolti dall'algoritmo sono i seguenti:

1. Partizionamento casuale del dataset di train in k sottoinsiemi, con k impostato fra 5 e 10;
2. Esclusione di un sottoinsieme e allenamento del modello sulle restanti $k-1$ ripartizioni;
3. Validazione del modello ottenuto sul *k-esimo* sottoinsieme escluso dall'allenamento e calcolo dell'errore di predizione;
4. Ripetizione del processo fino a che ogni *k-esimo* sottoinsieme è stato utilizzato come test per il modello;
5. Computazione dell'errore di cross validation ricavato dalla media dei k errori registrati, utilizzato come metrica di validazione del modello allenato.

Questa metodologia permette di testare diversi valori dei parametri del modello, in modo semplice e veloce, al fine di ottenere la configurazione ottimale. La tecnica della *k-fold cross validation* può essere ripetuta più volte (*repeated k-fold cross validation*), al fine di rendere più robusto il risultato ottenuto. Per la modellazione del BART e della rete neurale artificiale, sono state utilizzate altre librerie di *R* dedicate, rispettivamente *bart.machine* e *keras*, poiché *Caret* non dispone di tutte le funzionalità per poter ottimizzare al meglio i loro iperparametri.

Modello	$R^2[-]$	$RMSE[\frac{kWh}{m^2year}]$	$MAE[\frac{kWh}{m^2year}]$	$MAPE[\%]$
KNN	0.76	49.89	33.52	18.33
RT	0.68	56.01	39.74	22.82
RF: Bagging	0.80	44.60	30.27	17.08
Boosting	0.80	43.82	30.19	16.74
BART	0.83	43.06	27.01	14.95
ANN	0.71	68.26	46.59	22.08
Ensembling	0.82	42.34	28.12	15.48

Tabella 4.5: Risultati approccio ad un livello di analisi

La Tabella 4.5 riassume le metriche di validazione delle performance di ciascun modello sviluppato. L'algoritmo $BART$ produce l'errore più basso dell'$RMSE = 43.06\frac{kWh}{m^2year}$, del $MAE = 27.01\frac{kWh}{m^2year}$ e del $MAPE = 14.95\%$, riscontrando anche il valore di $R^2 = 0.83$ migliore rispetto a tutti gli altri. Oltre ai sei algoritmi sviluppati, è stata proposta una metodologia di ensembling dei modelli più efficienti al fine di incrementare ulteriormente i livelli di accuratezza nella stima regressiva. Per fare ciò sono state valutate diverse combinazioni fra i modelli con il tasso di errore minore, per calcolare un nuovo valore predetto attraverso la media ponderata sul $MAPE$ delle predizioni effettuate dai singoli modelli [40].

Fra le diverse combinazioni testate, l'unione del *BART* e del *RF: Bagging* ha prodotto il risultato migliore, esprimibile attraverso la seguente equazione:

$$Ep_{glo,nren}^{Ensemble} = \frac{Ep_{glo,nren}^{BART} MAPE^{BART} + Ep_{glo,nren}^{RF} MAPE^{RF}}{MAPE^{BART} + MAPE^{RF}}$$

La valutazione alla base del metodo ensembling proposto, si basa sul principio di compensazione degli errori, secondo il quale unendo due o più algoritmi diversi, ponderandone i risultati si potrebbe ottenere un'accuratezza globale migliore rispetto a quella ottenuta dai singoli modelli costituenti. Seguendo questa procedura si ottengono dei valori di $RMSE = 42.34 \frac{kWh}{m^2 year}$, del $MAE = 28.12 \frac{kWh}{m^2 year}$ e del $MAPE = 15.48\%$. Nonostante l'$RMSE$ sia il minore valutato, la metodologia ensemble non presenta dei grossi vantaggi rispetto al *BART*, che risulta essere il miglior algoritmo per la stima del valore dell'$Ep_{glo,nren}$.

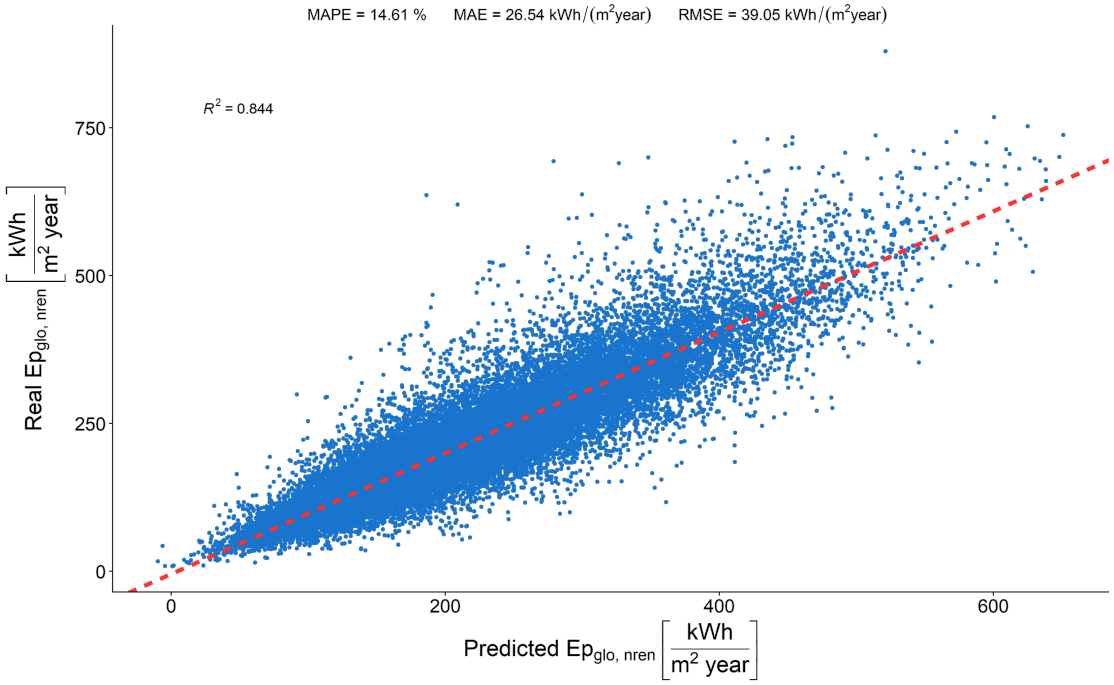


Figura 4.19: Scatter plot dei valori dell'$Ep_{glo,nren}$ reale e predetto con il modello regressivo BART

La Figura 4.19 rappresenta la dispersione dei valori predetti dal *BART* intorno alla retta di linearità. Ogni punto identifica un valore predetto e la distanza rispetto alla bisettrice costituisce l'errore commesso dal modello. Dunque, più i punti sono sovrapposti a tale retta, più precisa sarà la stima effettuata. Dal grafico

si osserva una maggiore dispersione per i valori elevati dell'$Ep_{glo,nren}$, determinata da una scarsa accuratezza del modello nel predire tali entità, e una maggiore concentrazione per i valori più bassi. La tendenza del modello di stimare con maggiore accuratezza i consumi bassi, può essere giustificata dalla scarsa cardinalità nel dataset di edifici con consumi elevati. Circa l'80% dei certificati riporta valori dell'$Ep_{glo,nren}$ inferiori a $250\frac{kWh}{m^2 year}$, dunque il modello è abituato a riconoscere le relazioni tipiche di tali edifici, rispetto a quelli caratterizzati da consumi elevati.

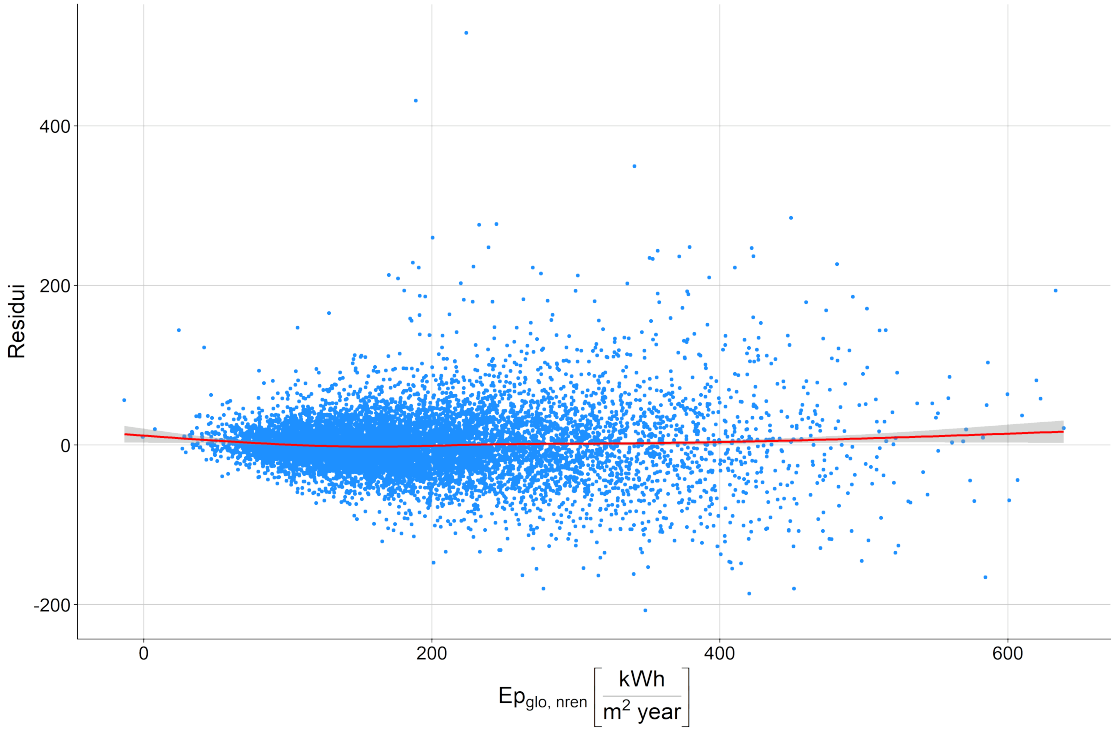


Figura 4.20: Grafico dei residui comparati alla stima dell'$Ep_{glo,nren}$

Un ulteriore metodo di validazione del modello regressivo adottato consiste nella verifica dell'omogeneità della varianza dei residui. In Figura 4.20 tramite un grafico a dispersione, sono rappresentati i residui rispetto ai valori stimati dell'indice di consumo; la linea rossa definisce la linea di tendenza di tali punti. Per accettare l'ipotesi di omoschedasticità, i residui devono disporsi in modo casuale attorno al valore nullo, dunque, più la linea di tendenza approssima una linea retta passante per lo zero, più i residui sono distribuiti omogeneamente. Nel caso in cui tale assunzione non venisse verificata, sarebbe possibile rilevare un legame fra l'andamento dei residui e il valore assunto dalla variabile dipendente. Se ad esempio i residui avessero un andamento monotono crescente, il modello tenderebbe a sottostimare con costanza i valori elevati dell'$Ep_{glo,nren}$, non fornendo una

rappresentazione accurata.

4.4.2 Approccio a due livelli di analisi per lo sviluppo di un modello previsionale regressivo

Come introdotto nel capitolo 3 il primo passaggio svolto nell'approccio a due livelli di analisi consiste nella suddivisione della distribuzione dell'$Ep_{glo,nren}$ in specifici range di consumo con lo scopo di raggruppare gli edifici aventi simili prestazioni energetiche, attraverso il metodo di partizionamento adattivo utilizzato in [33].

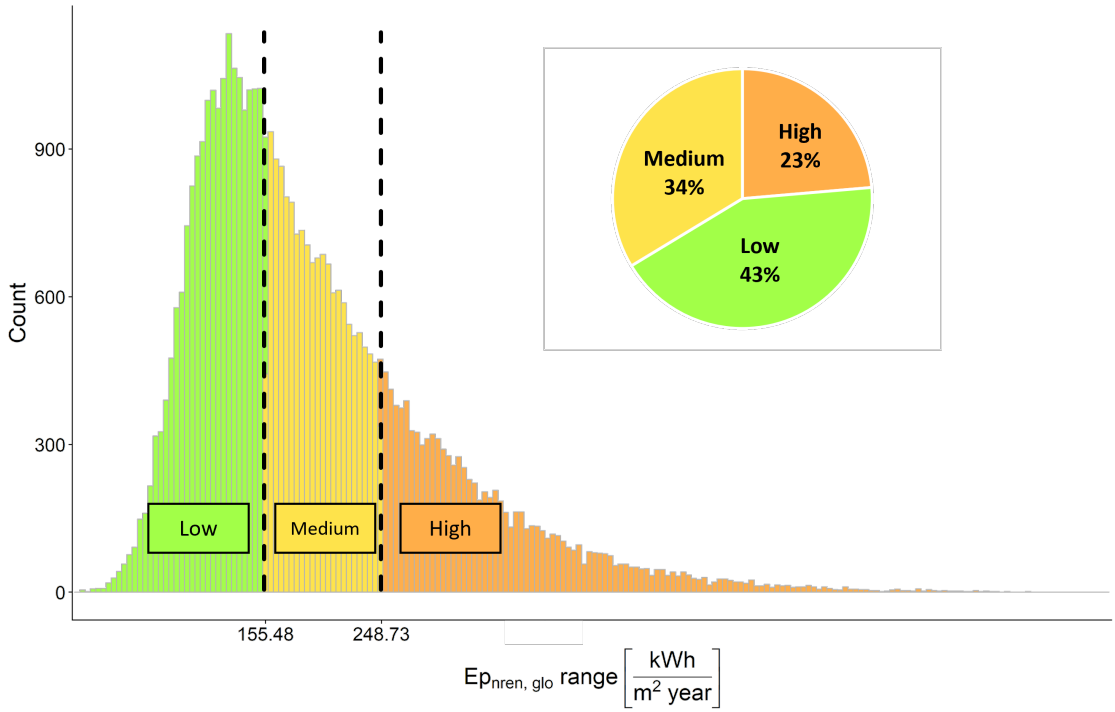


Figura 4.21: Suddivisione $Ep_{glo,nren}$ in tre range di consumo *Low*, *Medium* e *High*

In particolare sono stati evidenziati tre segmenti della distribuzione identificanti:

- edifici con bassi consumi energetici etichettati come *Low*, caratterizzati da un $Ep_{glo,nren} \leq 155 \frac{kWh}{m^2 year}$;
- edifici con consumi medi *Medium*, compresi fra $155 \frac{kWh}{m^2 year} \leq Ep_{glo,nren} \leq 249 \frac{kWh}{m^2 year}$;

- ed edifici con consumi alti *High*, contraddistinti da $Ep_{glo,nren} \geq 249 \frac{kWh}{m^2 year}$.

La scelta di ripartizionare il dataset in tre segmenti garantisce la presenza di un numero significativo di attestati in ciascun gruppo. All'intervallo *Low* appartiene infatti il 43% degli attestati presenti nel dataset, seguito dal *Medium* con il 34% e infine da *High* con il 23%. La scelta di un numero maggiore di segmenti avrebbe generato dei gruppi con un numero esiguo e sproporzionato di elementi, con una valenza statistica non sufficiente per sviluppare un modello di predizione. D'altro canto, un numero minore avrebbe reso insensato l'approccio a due livelli di analisi, ricadendo nel caso descritto nel paragrafo precedente.

Successivamente è stato sviluppato il modello di classificazione, con lo scopo di assegnare il corretto range di consumo ad ogni edificio di nuova osservazione. A tal proposito sono stati vagliati cinque dei sei algoritmi analizzati nell'approccio regressivo (*KNN*, *Classification Tree*, *Random Forest: Bagging*, *Boosting* e *ANN*), che si prestavano anche a problemi di stima categorica. Per validare i risultati del classificatore, anche in questo caso, sono state utilizzate delle metriche specifiche al fine di valutare il modello in termini di accuratezza. Nei problemi di classificazione non vi è un valore numerico continuo con il quale valutare i residui, la bontà del modello si basa sul conteggio delle istanze classificate correttamente, rispetto al numero totale di stime effettuate. Inoltre, sono state utilizzate altre tre metriche per una validazione più accurata del classificatore, al fine di arginare la generazione di bias nell'accuratezza dovuti al non perfetto bilanciamento dei range:

- *Precision*: definisce la proporzione di classi correttamente predette dal modello, rispetto al totale di predizioni effettuate per quella classe;

$$Precision = \frac{VP}{VP + FP}$$

- *Recall*: definità anche come sensitività, stabilisce la proporzione di classi correttamente individuate dal modello rispetto al totale di elementi appartenenti alla suddetta classe;

$$Recall = \frac{VP}{VP + FN}$$

- *F1-score*: matematicamente corrisponde alla media armonica della precision e del recall, tendendo dunque a fondere insieme i contributi delle due metriche, che considerate singolarmente possono dare informazioni fuorvianti;

$$F1\text{-}score = \frac{2}{Precision^{-1} + Recall^{-1}} = \frac{VP}{VP + 1/2(FP + FN)}$$

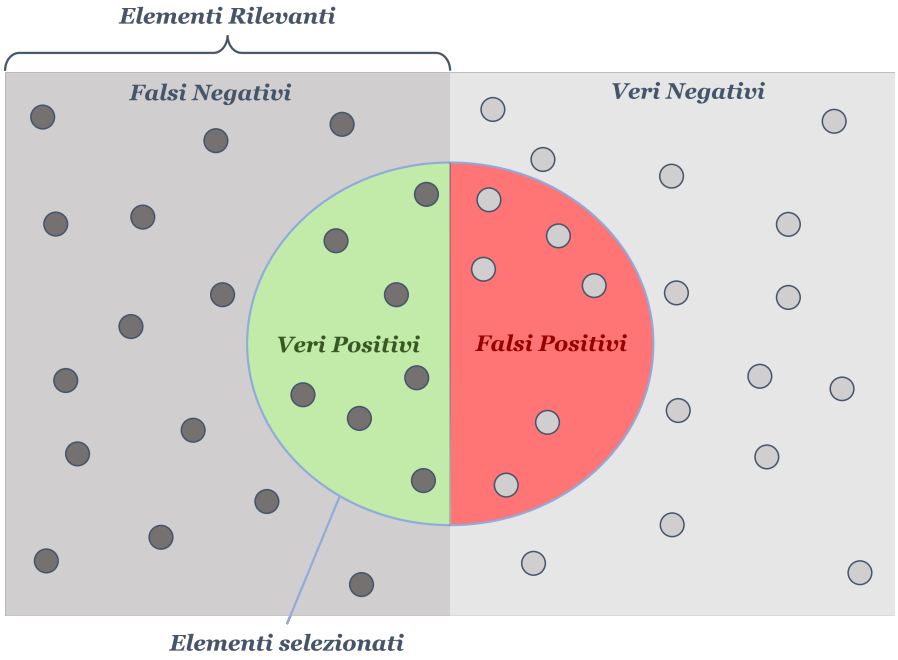


Figura 4.22: Concettualizzazione metriche accuratezza

Per comprendere meglio la formulazione matematica delle metriche sovraesposte si propone in Figura 4.22 un esempio di classificatore binario per semplicità, ma il concetto è estendibile anche ai modelli multiclasse. All'interno dell'ellisse vi sono gli elementi selezionati dal modello come appartenenti a una determinata categoria. Fra di essi vi sono elementi realmente appartenenti alla categoria stimata, definiti *Veri Positivi (VP)* ed elementi che in realtà non vi appartenevano, per questo identificati come *Falsi Positivi (FP)*. Restano dunque fuori dall'insieme di selezione del modello quegli elementi appartenenti alla categoria ricercata, ma che non sono stati identificati come tali, *Falsi Negativi (FN)*, e gli elementi giustamente esclusi perché non afferenti ad essa, *Veri Negativi (VN)*. Volendo formalizzare anche il concetto di accuratezza la si definisce matematicamente come:

$$Accuratezza = \frac{VP + VN}{VP + VN + FP + FN}$$

La Tabella 4.6 riassume i risultati ottenuti per i cinque algoritmi testati. L'algoritmo di *Random Forest: Bagging* ha raggiunto l'accuratezza maggiore dell'80.5% seguito quasi a pari prestazioni dal*Boosting* con il 79.5 % e dalla *ANN* con il 78.2%. I primi tre algoritmi si sono distinti anche per le ottime metriche di Precision, Recall e F1-score. Da una prima osservazione si evidenzia come il range dei valori *Low* sia stato quello meglio caratterizzato da tutti i modelli, a differenza del

Metrica	KNN	CT	RF: Bagging	Boosting	ANN
			Globale		
Accuracy (%)	74.7	73.8	80.5	79.5	78.2
			Segmento: Low		
Preision (%)	81.2	80.2	85.7	86.1	84.0
Recall (%)	84.1	83.6	86.6	86.6	86.4
F1 (%)	82.6	81.9	86.2	86.4	85.2
			Segmento: Medium		
Preision (%)	62.4	62.5	69.3	69.5	68.4
Recall (%)	67.8	62.7	72.5	71.8	68.1
F1 (%)	65.0	62.6	70.8	70.7	68.2
			Segmento: High		
Preision (%)	83.1	78.4	83.9	82.1	81.7
Recall (%)	67.6	72.0	76.7	77.5	78.0
F1 (%)	74.6	75.0	80.2	79.7	79.8

Tabella 4.6: Risultati classificazione

Medium che si assesta come l'intervallo meno accurato da stimare. Ciò è dovuto al fatto che esso rappresenti il range di intermezzo fra *Low* e *High* e dunque la possibilità di classificare erroneamente un elemento ad esso afferente sono duplici rispetto agli altri due intervalli. La maggior accuratezza dei classificatori rispetto alla classe *Low* può essere ulteriormente argomentata, non solo perché essa rappresenta la maggioranza di istanze del dataset, ma poiché i valori da essa identificati sono condensati in un intervallo di massima frequenza. Di contro, la classe *High* presenta un'elevata dispersione dei suoi valori, che si traducono in un'accuratezza minore (Figura 4.21).

Per la tipologia di analisi che si vuole condurre, alti valori di accuratezza globale e specifica di ogni segmento sono fondamentali per la corretta predizione dell'$Ep_{glo,nren}$ effettuata nel livello successivo. Il modello regressivo è allenato sulla base delle somiglianze presenti fra gli elementi appartenenti a una determinata classe, e la scorretta classificazione inficerebbe sul risultato finale. All'errore commesso dal regressore si sommerà quello effettuato a priori dal classificatore. Viene scelto dunque l'algoritmo di *Random Forest: Boosting*, in base ai livelli di accuratezza raggiunti, come metodo di classificazione del dataset e dunque, come primo modello della metodologia a due livelli di analisi.

Per la modellazione regressiva è stato valutato l'algoritmo *BART* risultato essere

MAPE BART			
Low	*Medium*	*High*	**Globale**
12.1%	9.1%	11.1%	10.88%

Tabella 4.7: Risultati Regressione algoritmo BART sui range reali

il più accurato nell'approccio a un singolo livello di analisi. Nella fase di allenamento, sono stati sviluppati tre algoritmi indipendenti, ognuno addestrato sulla base di dati relativa a un determinato intervallo reale. I risultati del MAPE ottenuto dal modello sono riassunti in Tabella 4.7. Essi fanno riferimento al modello allenato e testato su intervalli reali, non stimati dal classificatore. Applicando in sequenza il modello di classificazione e regressione le prestazioni ottenute sono riassunte in Tabella 4.8.

$R^2[-]$	$RMSE[\frac{kWh}{m^2 year}]$	$MAE[\frac{kWh}{m^2 year}]$	$MAPE[\%]$
Globale			
0.79	45.12	29.90	15.80
Segmento: Low			
0.51	26.81	17.19	14.59
Segmento: Medium			
0.28	47.58	33.27	16.75
Segmento: High			
0.54	65.53	48.57	16.67

Tabella 4.8: Risultati Regressione algoritmo BART sui range predetti dal classificatore

La dispersione dei valori predetti con i modelli di regressione rispetto ai valori reali è rappresentata nella Figura 4.23. Anche per la regressione il segmento con i valori predetti più accurati è il *Low*, per le medesime motivazioni argomentate nell'analisi del classificatore. Inoltre il modello regressivo è influenzato dagli errori commessi nella fase di classificazione, che riducono ulteriormente la sua accuratezza finale. Si evince una forte perdita di accuratezza dovuta alla somma degli errori generatasi dall'unione dei due modelli. Il risultato globale dell'approccio a due livelli di analisi (MAPE = 15.8%) è inferiore a quello raggiunto dal solo modello regressivo (MAPE = 14.95%). Dalle analisi sperimentali svolte sulla tipologia di dataset su cui è stato condotto il lavoro di tesi, l'approccio a un solo livello di analisi puramente regressiva, è risultata essere la metodologia più efficace per la

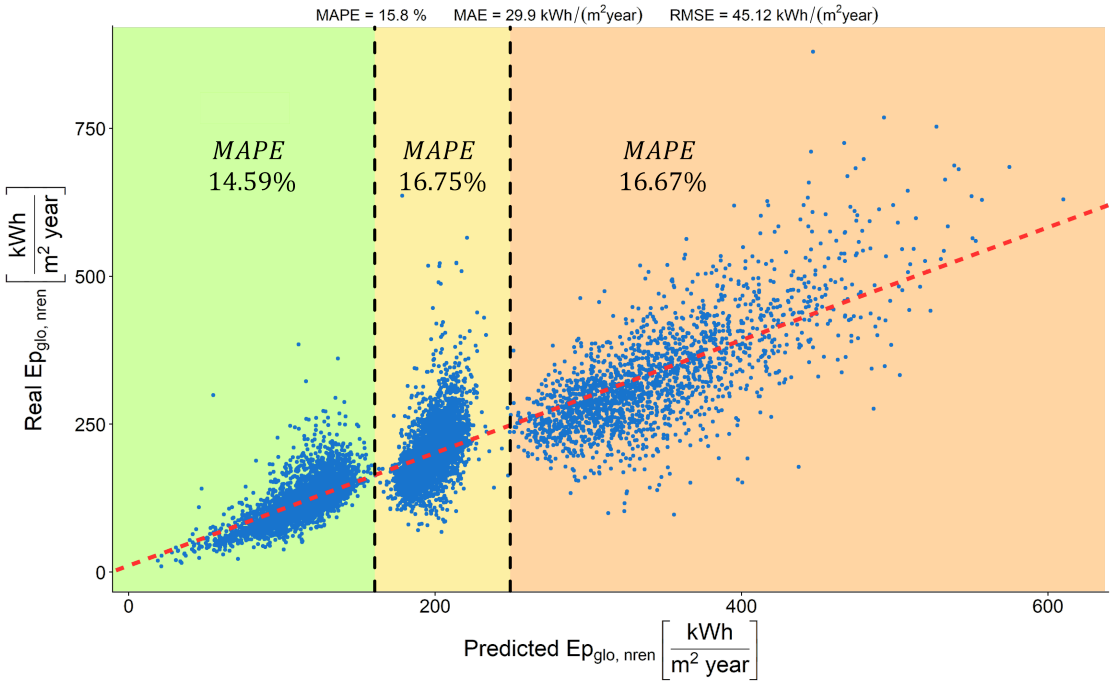


Figura 4.23: Scatter plot dei valori dell'$Ep_{glo,nren}$ reale e predetto con il modello regressivo BART, applicato a tre intervalli di consumo

stima del valore dell'indice di consumo globale non rinnovale di un edificio.

Oltre agli aspetti di accuratezza gli ulteriori vantaggi di questo approccio sono identificabili al livello di tempistiche e complessità computazionali, poiché allenare un solo modello è più semplice e meno dispendioso. In Tabella 4.9 sono espressi i tempi computazionali dei vari modelli sviluppati in relazione al tempo dell'algoritmo KNN, scelto come riferimento per la sua scalabilità e semplicità di parametrizzazione. In questo modo i risultati sono espressi in maniera indipendente dall'hardware utilizzato per le simulazioni.

Modello	**KNN**	**CART**	**RF:Bagging**	**Boosting**	**BART**	**ANN**
Regressione	1	0.64	147.54	0.65	3.93	0.98
Classificazione	1	0.20	10.50	0.54	-	0.64

Tabella 4.9: Tempi computazionali relativi al KNN

4.5 Obiettivo 5

Questo obiettivo mira a fornire un livello di analisi in grado di spiegare il perché delle stime effettuate dall'algoritmo *BART*. Il funzionamento del modello regressivo sarà analizzato attraverso diversi metodi di XAI, forniti dalla libreria DALEX [41] di R. A livello globale, l'*ALE plot* evidenzierà come ciascun attributo influenza la predizione, e a livello locale tramite l'algoritmo *breakDown* [42], verrà mostrato il contributo additivo di ogni attributo alla predizione della specifica osservazione, che può assumere verso e peso differente. La somma di tutti i contributi delle singole variabili coincide con la stima effettuata dal modello black-box per quella determinata istanza.

4.5.1 Global Methods

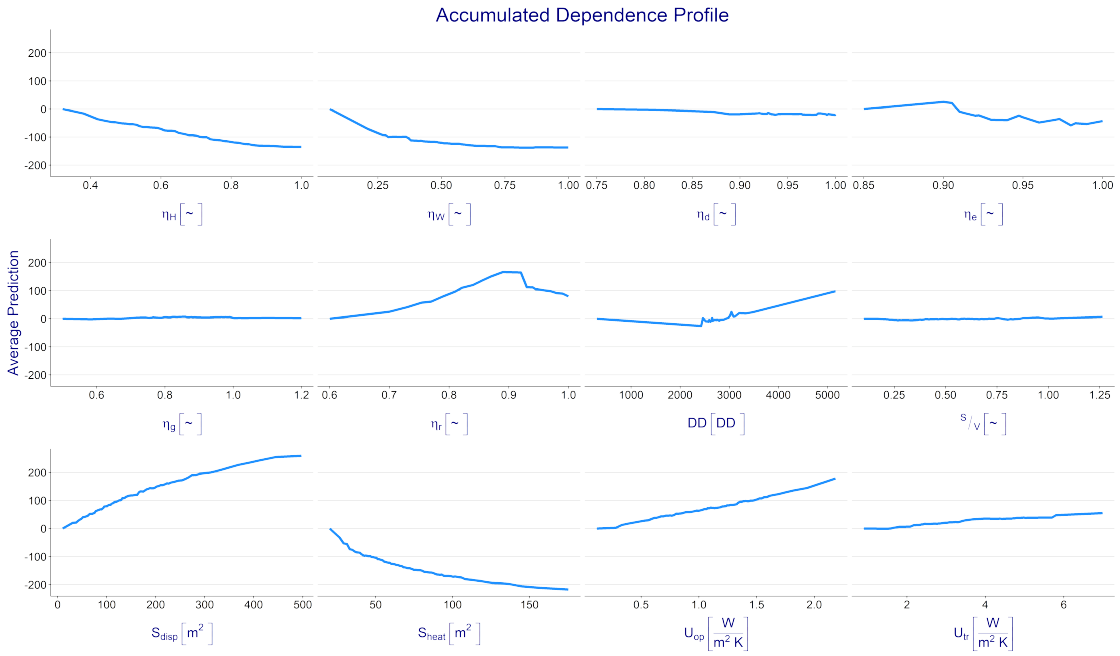


Figura 4.24: ALE plot BART

Analizzando la Figura 4.24 si osserva come la variazione di η_d, η_g, η_e e del fattore di forma non determina un cambiamento della predizione. Attributi come la conducibilità termica di involucro, sia opaca sia trasparente, e la superficie disperdente seguono un andamento monotono crescente, una loro variazione positiva comporta

un aumento del valore stimato. Un comportamento analogo sarebbe stato presumibile anche per la superficie riscaldata, ma il modello ha osservato che essa assuma un andamento inversamente proporzionale alla stima del consumo, per cui un'analisi approfondita dovrebbe essere condotta per verificare questa controtendenza. Il crescere del valore dei rendimenti η_H ed η_W influenza negativamente la crescita dei consumi. I gradi giorno e il rendimento del sottosistema di regolazione η_r, a differenza degli altri attributi, non presentano un andamento monotono. L'aumento del primo comporta una leggera diminuzione della predizione intorno ai 2500 DD, superati i quali la tendenza ritorna verosimilmente positiva. L'η_r, invece, presenta un comportamento inverso; i consumi tendono ad aumentare fino al valore di 0.9, per poi decrescere. Gli attributi che presentano degli andamenti inattesi possono essere stati influenzati dagli effetti di correlazione con altre variabili di ingresso, per cui la relazione rappresentata fra essi e la stima del consumo non è costituita solamente dagli effetti principali.

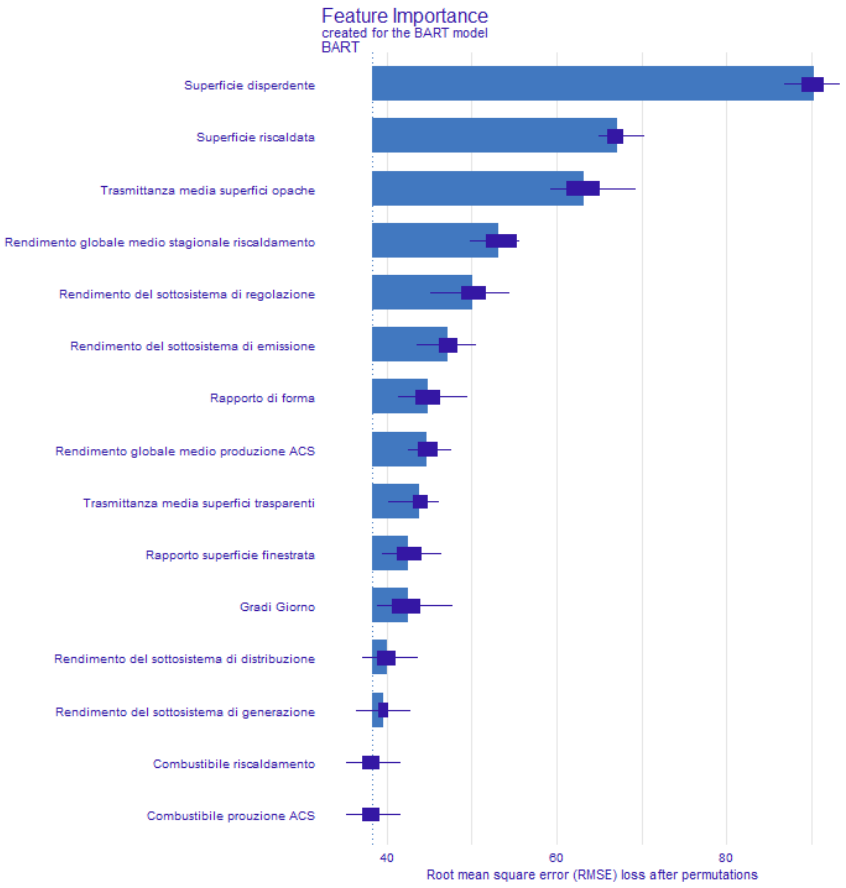


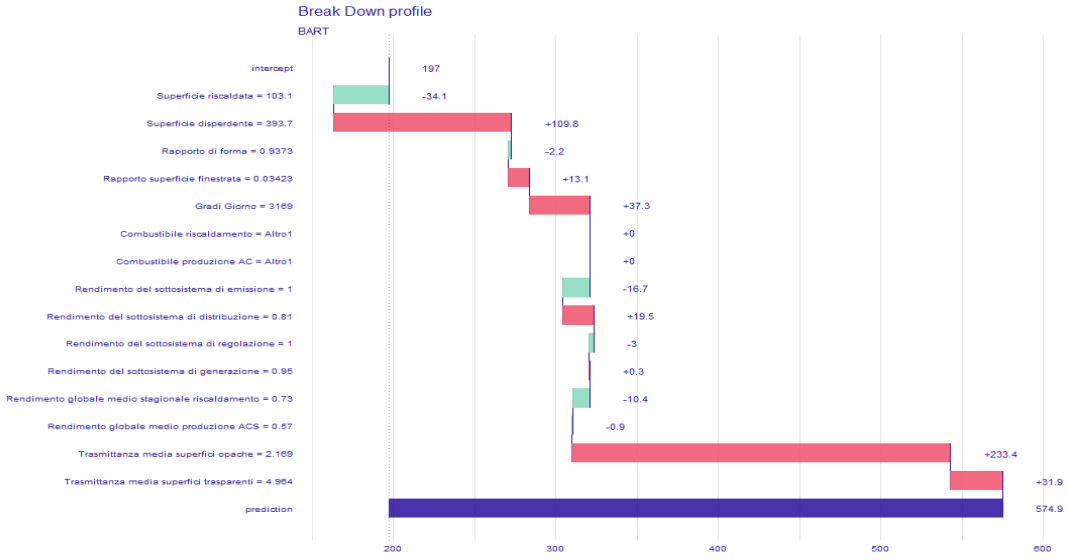
Figura 4.25: Feature importance plot BART

In Figura 4.25 è possibile visualizzare quali variabili sono ritenute più importanti ai fini della predizione, da parte del modello *BART*. Le variabili geometriche indicative delle superfici risultano essere le più influenzanti, seguite dalla trasmittanza delle superfici opache e dai rendimenti. Nell'ottica del retrofit è conveniente che variabili come le conducibilità termiche e i rendimenti di impianto siano rilevanti, perché in questo modo un loro miglioramento comporterebbe una variazione significativa dei consumi stimati, con elevate possibilità di guadagno.

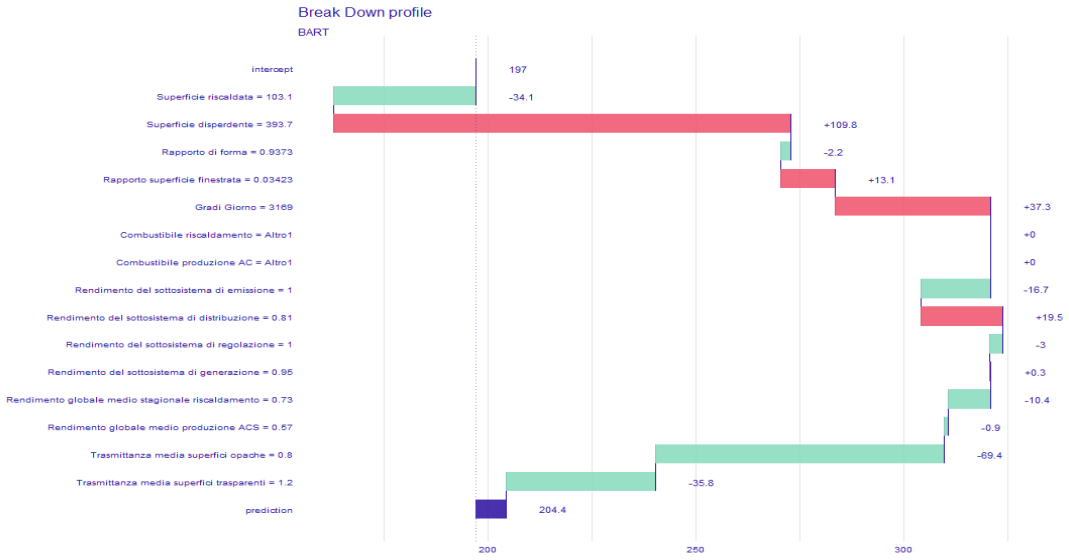
4.5.2 Local Methods

Concentrando l'attenzione sulla singola predizione, è stata selezionata un'istanza con un valore dell'$Ep_{glo,nren}$ stimato molto accurato rispetto al valore effettivo, al fine di rappresentare la realtà il più verosimilmente possibile. Inoltre, è stato selezionato un edifico caratterizzato da consumi mediamente elevati, per poter poi evidenziare in maniera più incisiva i vantaggi ottenibili da un eventuale azione di riqualificazione energetica, e su di esso è stato sviluppato il *breakDown plot* rappresentato in Figura 4.26a. Tali grafici sono utilizzati per rappresentare la decomposizione della previsione del modello in contributi attribuibili alle diverse variabili esplicative. Sull'asse delle ordinate sono rappresentati gli attributi e i valori ad essi associati. Il primo elemento costituisce l'intercetta, ovvero la media delle predizioni dell'indice di consumo. Successivamente, a cascata, vengono rappresentate le variazione della media avvenute in seguito all'imposizione nel dataset del valore dell'attributo dell'istanza. La barra rossa indica che il valore specifico assunto dall'attributo ha determinato un aumento del consumo, mentre la barra verde identifica una diminuzione dello stesso. A fianco ad ogni barra è indicata la variazione rispetto all'iterazione predente. Infine tutti i contributi delle variabili convergono nella predizione finale dell'$Ep_{glo,nren}$, i cui valori sono riportati in ascissa. Il criterio di ordinamento degli attributi è stato scelto tenendo conto di quanto sia possibile modificarne il valore in funzione di un ipotetica azione di retrofit. Per questo motivo le variabile geometriche sono poste per prime, e i rendimenti e le conducibilità termiche per ultimo, cosicché una loro modifica renda apprezzabile la variazione del loro contributo nella predizione.

La Figura 4.27a rappresenta la stessa istanza, ma per la quale non è stato imposto un ordine aprioristico. La sequenza di variabili rappresentate segue l'algoritmo euristico di ordinamento, precedente discusso nel capitolo 3, secondo il quale le variabili sono riportate in funzione dell'effetto che esse hanno sulla variazione della predizione, nel momento in cui il proprio valore viene imposto a tutte le istanze del dataset. Come è possibile notare ciò non influisce sul valore della predizione finale, che assume il valore di $574.9\frac{kWh}{m^2 year}$ in etrambi i grafici, ma sul contributo



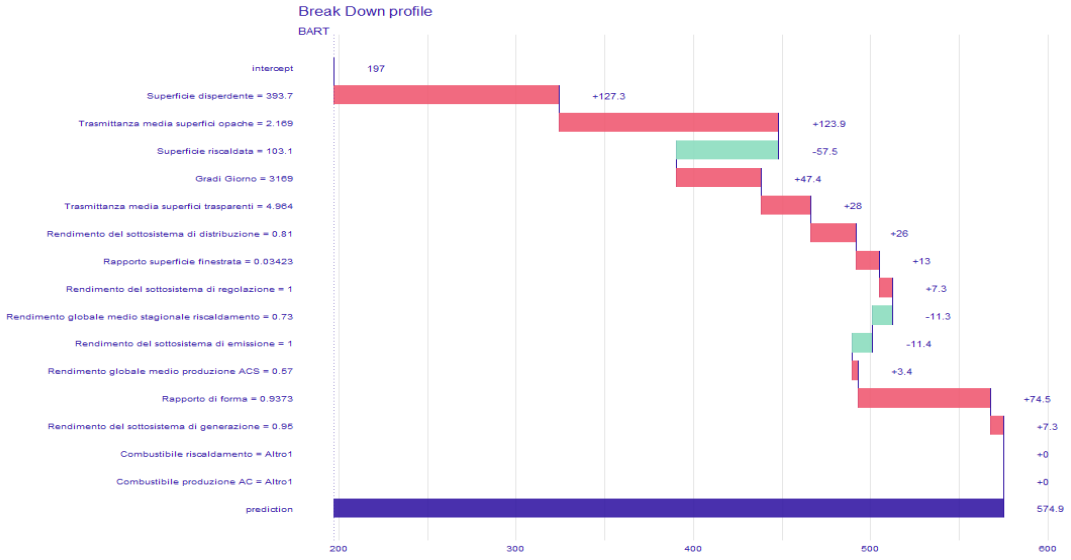
(a) Pre-Retrofit



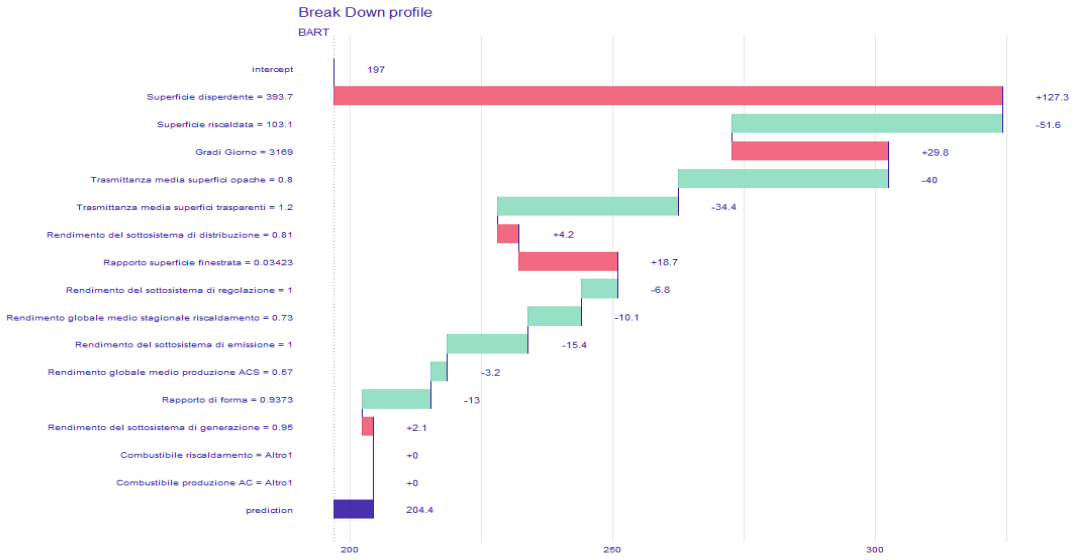
(b) Post-Retrofit

Figura 4.26: BreakDown plot per un edificio con alto consumo stimato, con ordinamento delle variabili predeterminato

specifico di ogni attributo. Il caso più evidente per questo esempio è costituito dal rapporto di forma: nella Figura 4.26a esso compare come terza variabile e il valore che assume comporta un decremento, seppur minimo di soli $2.2\frac{kWh}{m^2 year}$, della predizione. Nella Figura 4.27a invece, viene collocato come quartultima variabile, e il



(a) Pre-Retrofit



(b) Post-Retrofit

Figura 4.27: BreakDown plot per un edificio con alto consumo stimato, con ordinamento delle variabili euristico

suo valore decreta un aumento di $74.5 \frac{kWh}{m^2 year}$. Dal punto di vista assoluto, queste incongruenze rendono il breakDown plot difficile da interpretare, se non si conosce a fondo il suo algoritmo di funzionamento. Per tale motivo l'utilizzo di questa metodologia è stata demandata principalmente alla visualizzazione del mutamento

relativo del contributo degli attributi di uno stesso edificio, a monte e a valle delle azioni di retrofit. Per tale scopo è evidente come l'ordinamento aprioristico sia la metodologia di visualizzazione migliore per apprezzare la variazione del valore stimato in seguito alla modifica delle variabili di interesse.

Dai grafici prodotti dal modello si osserva come il contributo preponderante verso l'aumento del valore stimato è determinato dalla trasmittanza dell'involucro opaco. Si decide dunque di intervenire immaginando un intervento di riqualificazione dell'involucro edilizio, che comporti un aumento dell'isolamento termico. Si impostano dunque dei nuovi valori per la $U_{op} = 0.8\frac{W}{m^2K}$ e per la $U_{tr} = 1.2\frac{W}{m^2K}$, relativi a stratigrafie e ad infissi con alte performance, e si rivaluta il contributo degli attributi. Nel momento in cui vengono modificati i valori della trasmittanza dell'involucro opaco e trasparente, si genera di fatto una nuova istanza, diversa da quella precedente. Il modello la rielabora come se fosse un edificio distinto e valuta i pesi dei vari attributi in base alla nuova configurazione. Con il metodo di ordinamento euristico dunque, le variabili vengono riordinate e non è possibile apprezzare in maniera diretta il contributo della specifica variazione (Figura 4.27a e Figura 4.27b). L'ordinamento predeterminato cerca di ovviare a tale problema imponendo un ordine fisso per le variabili che non hanno subito modifica, in modo che il loro contributo pre e post retrofit resti immutato, riconducendo la variazione del valore finale stimato direttamente alla modifica degli attributi selezionati.

Capitolo 5

Conclusioni e Discussione dei Risultati

Il framework metodologico proposto in questo lavoro di tesi restituisce come risultati degli strumenti con cui poter soddisfare tre richiesta specifiche: analisi statistica del database di attestati di prestazione energetica, benchmarking energetico e stima dei consumi futuri di un edificio, evidenziandone le opportunità di miglioramento e efficientamento. Sebbene questi obiettivi finali siano stati esposti distintamente all'interno della trattazione, essi possono coesistere in un unico ambiente che possa fungere da strumento di supporto, per il certificatore energetico, o di verifica per gli enti di competenza. A tal proposito i vari obiettivi del framework sono stati riorganizzati e condensati, per semplificarne la fruizione ad un ipotetico utente finale.

Per poter redigere un certificato energetico il professionista deve disporre delle informazioni geometriche e termofisiche dell'edificio per poter valutare anche solo in maniera approssimativa le prestazioni energetiche dell'immobile. La maggior parte di esse sono reperibili in seguito a un sopralluogo o alla revisione dei progetti di costruzione, ma ciò richiede un investimento di tempo e risorse, in contrasto con l'esigenza di aver un risultato rapido. Sfruttando l'Obiettivo 2, l'utente è guidato nell'imputazione del valore di uno specifico attributo di cui non possiede informazioni, attraverso la visualizzazione delle distribuzioni monovariate o inerenti a determinate tipologie edilizie e periodo di costruzione. Nel caso in cui ad usare lo strumento fosse un ente predisposto alla verifica e al controllo dei certificati energetici, la sua funzionalità consisterebbe nel confronto immediato dei dati presenti

nel certificato rispetto a un campione statistico rilevante. Nel caso in cui un determinato valore si discosti di molto dalla media della sua distribuzione, un controllo più dettagliato dovrebbe essere effettuato al fine di giustificare tale discrepanza.

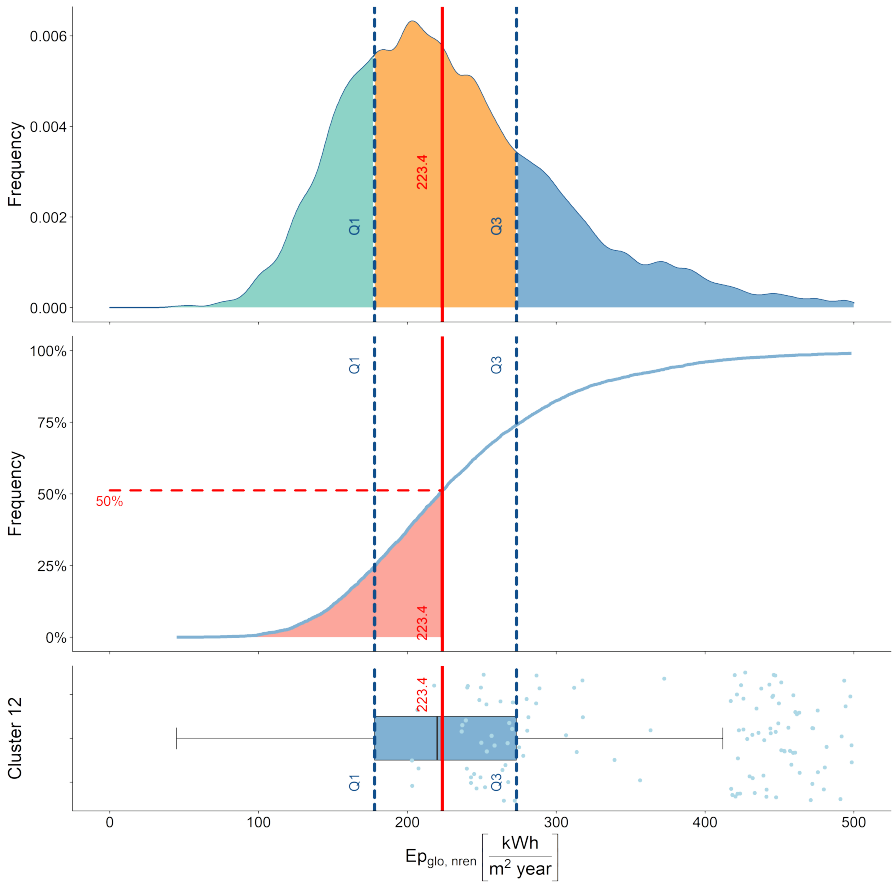


Figura 5.1: Rappresentazione valore stimato $Ep_{glo,nren}$ nel suo cluster di riferimento

Successivamente alla fase di inserimento dei dati di input inerenti a un nuovo edificio, l'utente ha la possibilità di conoscere la stima del valore dell'$Ep_{glo,nren}$ e di come esso si collochi all'interno della distribuzione dei consumi relativi ad edifici simili. In questa operazione vengono condensate le metodologie di clustering, visualizzazione statistica e stima regressiva. In Figura 5.1 è possibile osservare come il valore stimato dal *BART* dell'indice di consumo si posiziona rispetto ai valori del cluster di edifici simili a quello in esame, individuati dall'algoritmo *k-means*. Sono evidenziati tre range di valori delimitati dal primo e terzo interquartile, ciò significa che il segmento centrale caratterizza il 50% dei valori dell'$Ep_{glo,nren}$ catalogati nei

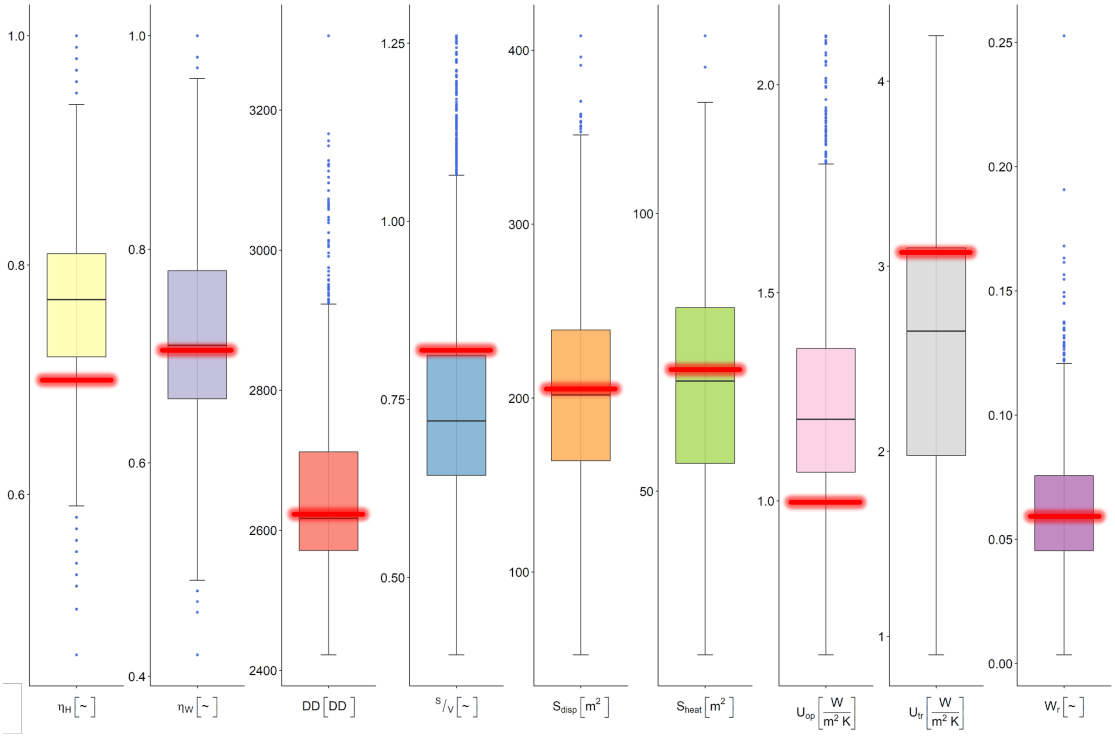


Figura 5.2: Attributi dell'edificio i cui consumi sono rappresentati in Figura 5.1

certificati. Un valore stimato appartenete a tale intervallo risulta essere frequente nella distribuzione del cluster di riferimento, mentre gli intervalli di sinistra e di destra rappresentano relativamente edifici dai consumi inferiori e maggiori rispetto alla media del campione. L'intersezione del valore stimato con la curva *CDF* determina che il 50% di edifici del campione di riferimento ha un valore dell'$Ep_{glo,nren}$ inferiore rispetto a quello stimato per l'edificio in esame. L'analisi dei suoi attributi di input in Figura 5.2 permette di giustificare l'andamento dell'output, osservando come anche essi assumano i valori più frequentemente presenti nel cluster. L'edificio in esame presenta delle caratteristiche molto vicine al centroide del suo cluster di appartenenza, salvo qualche variabile, come ad esempio il rendimento globale medio stagionale per il riscaldamento e le conducibilità termiche di involucro.

5.1 Prospettive future di implementazione

L'estrazione, la caratterizzazione e la valorizzazione delle informazioni contenute nel database di certificati energetici, si pongono come obiettivi principali della

metodologia proposta in questo lavoro di tesi. Nonostante essa proponga un ampio livello di applicazione e di robustezza dei risultati ottenuti, diverse possibili migliorie sono applicabili per incrementarne l'efficacia e le possibilità di utilizzo.

Il primo obiettivo da perseguire in futuro riguarda l'integrazione dei dataset di certificati energetici provenienti da più regioni italiane. Ad oggi, ogni regione provvede all'invio degli *APE* alla sede centrale nazionale di *ENEA* tramite il formato xml e mette a disposizione gli open data in formato csv. Questi ultimi contengono solamente delle informazioni parziali, spesso non coincidenti fra regioni diverse, il che rende complicata l'aggregazione di file provenienti da regioni differenti in un unico database nazionale. L'estrazione degli attributi di interesse direttamente dai file xml potrebbe arginare questo problema.

Inoltre, un ulteriore passaggio verso una maggiore scalabilità, prevederebbe l'estensione dell'analisi anche agli edifici non residenziali, facenti parte delle categorie edilizie differenti dalla E1(1), e a servizi energetici afferenti alla climatizzazione estiva e alla produzione da fonti rinnovabili, qualora si riesca a disporre di una base statistica rilevante.

Per quando riguarda l'analisi di benchmarking, altri algoritmi di clustering più efficienti potrebbero essere testati con differenti configurazioni degli attributi in ingresso, al fine di ridurre il numero di raggruppamenti individuati e semplificare il confronto con i centroidi. La stima delle performance energetiche potrebbe essere integrata con algoritmi e metodologie meno generalizzate e adattate alla tipologia di dataset di cui si dispone, ottenendo un modello in grado di individuare in maniera più accurata le relazioni presenti fra input e output.

L'unione degli obiettivi presentata come un unico strumento nel paragrafo precedente potrebbe essere presentata come un *Software as a Service (SaaS)*, scalabile, intuitivo e di ampia fruizione da parte degli utenti. I dati necessari alla computazione dei modelli vengono inseriti attraverso un interfaccia grafica, dalla quale è possibile osservare i trend di distribuzione degli attributi. Successivamente l'utente seleziona la variabile dell'output di interesse (oltre all'$Ep_{glo,nren}$ altri indici rilevanti potrebbero riguardare la quota dell'indice di consumo dovuta al riscaldamento o il fabbisogno di energia primaria dell'involucro edilizio), e ne inserisce il valore, qualora ne sia già a disposizione (caso di verifica dell'*APE*), o sceglie di ricavarlo attraverso la stima effettuata dal modello regressivo. Indipendentemente dal fatto che l'output sia stato inserito o stimato, il valore verrà collocato all'interno del cluster di riferimento e visualizzato rispetto ad esso, proponendo una visualizzazione analoga anche per i valori delle variabili di input.

Successivamente, sfruttando il layer di XAI, è possibile conoscere il contributo

di ogni variabile ai fini della predizione finale, e valutare una possibile configurazione di retrofit, modificando le variabili maggiormente determinanti l'aumento dei consumi, relative all'involucro edilizio e agli impianti installati. Il software rielaborerà una nuova stima e valuterà una nuova configurazione degli attributi, in funzione delle modifiche apportate, in questo modo l'utente può confrontare come sono variati i consumi stimati attraverso i due *breakDown plot* relativi alle stime effettuate precedentemente e posteriormente gli interventi di riqualificazione.

A valle delle considerazioni svolte, il lavoro di tesi intende fornire una metodologia innovativa per la valutazione delle prestazioni energetiche degli edifici a supporto di interventi di riqualificazione energetica, capace di sfruttare le conoscenze estratte da dataset di *APE* e di elaborarle attraverso l'uso di metodi di intelligenza artificiale. La metodologia proposta, implementata all'interno di uno strumento diffuso nell'ambito della progettazione edile, potrebbe essere potenzialmente capace di velocizzare e snellire il processo di valutazione delle prestazioni energetiche degli edifici dei progetti di retrofit energetico.

Bibliografia

[1] Decreto del presidente della repubblica 26 agosto 1993, n. 412. URL https://www.gazzettaufficiale.it/eli/id/1993/10/14/093G0451/sg.

[2] Comunicato stampa. Ondata di ristrutturazioni: raddoppiare il tasso di ristrutturazione per abbattere le emissioni, stimolare la ripresa e ridurre la povertà energetica. URL https://ec.europa.eu/commission/presscorner/detail/it/ip_20_1835.

[3] Miguel Molina-Solana, María Ros, M. Dolores Ruiz, Juan Gómez-Romero, and M.J. Martin-Bautista. Data science for building energy management: A review. *Renewable and Sustainable Energy Reviews*, 70:598–609, 2017. ISSN 1364-0321. doi: https://doi.org/10.1016/j.rser.2016.11.132. URL https://www.sciencedirect.com/science/article/pii/S1364032116308814.

[4] Björn Hårsman, Zara Daghbashyan, and Parth Chaudhary. On the quality and impact of residential energy performance certificates. *Energy and Buildings*, 133:711–723, 2016. ISSN 0378-7788. doi: https://doi.org/10.1016/j.enbuild.2016.10.033. URL https://www.sciencedirect.com/science/article/pii/S0378778816312695.

[5] Stefano Cozza, Jonathan Chambers, Chirag Deb, Jean-Louis Scartezzini, Arno Schlüter, and Martin K. Patel. Do energy performance certificates allow reliable predictions of actual energy consumption and savings? learning from the swiss national database. *Energy and Buildings*, 224:110235, 2020. ISSN 0378-7788. doi: https://doi.org/10.1016/j.enbuild.2020.110235. URL https://www.sciencedirect.com/science/article/pii/S0378778820303315.

[6] Chris van Dronkelaar, Mark Dowson, E. Burman, Catalina Spataru, and Dejan Mumovic. A review of the energy performance gap and its underlying causes in non-domestic buildings. *Frontiers in Mechanical Engineering*, 1, 2016. ISSN 2297-3079. doi: 10.3389/fmech.2015.00017. URL https://www.frontiersin.org/article/10.3389/fmech.2015.00017.

[7] Patrick X.W. Zou, Xiaoxiao Xu, Jay Sanjayan, and Jiayuan Wang. Review of 10 years research on building energy performance gap: Life-cycle and stakeholder perspectives. *Energy and Buildings*, 178:165–181, 2018. ISSN 0378-7788. doi: https://doi.org/10.1016/j.enbuild.2018.08.040. URL https://www.sciencedirect.com/science/article/pii/S0378778818309460.

[8] Oleksii Pasichnyi, Jörgen Wallin, Fabian Levihn, Hossein Shahrokni, and Olga Kordas. Energy performance certificates — new opportunities for data-enabled urban energy policy instruments? *Energy Policy*, 127:486–499, 2019. ISSN 0301-4215. doi: https://doi.org/10.1016/j.enpol.2018.11.051. URL https://www.sciencedirect.com/science/article/pii/S0301421518307894.

[9] Antonio Attanasio, Marco Savino Piscitelli, Silvia Chiusano, Alfonso Capozzoli, and Tania Cerquitelli. Towards an automated, fast and interpretable estimation model of heating energy demand: A data-driven approach exploiting building energy certificates. *Energies*, 12(7), 2019. ISSN 1996-1073. doi: 10.3390/en12071273. URL https://www.mdpi.com/1996-1073/12/7/1273.

[10] Antonio Galli, Marco Savino Piscitelli, Vincenzo Moscato, and Alfonso Capozzoli. Bridging the gap between complexity and interpretability of a data analytics-based process for benchmarking energy performance of buildings.

[11] Xuefeng Gao and Ali Malkawi. A new methodology for building energy performance benchmarking: An approach based on intelligent clustering algorithm. *Energy and Buildings*, 84:607–616, 2014. ISSN 0378-7788. doi: https://doi.org/10.1016/j.enbuild.2014.08.030. URL https://www.sciencedirect.com/science/article/pii/S0378778814006720.

[12] Geoffrey K.F. Tso and Kelvin K.W. Yau. Predicting electricity energy consumption: A comparison of regression analysis, decision tree and neural networks. *Energy*, 32(9):1761–1768, 2007. ISSN 0360-5442. doi: https://doi.org/10.1016/j.energy.2006.11.010. URL https://www.sciencedirect.com/science/article/pii/S0360544206003288.

[13] Zhun Yu, Fariborz Haghighat, Benjamin C.M. Fung, and Hiroshi Yoshino. A decision tree method for building energy demand modeling. *Energy and Buildings*, 42(10):1637–1646, 2010. ISSN 0378-7788. doi: https://doi.org/10.1016/j.enbuild.2010.04.006. URL https://www.sciencedirect.com/science/article/pii/S0378778810001350.

[14] A.P. Melo, D. Cóstola, R. Lamberts, and J.L.M. Hensen. Development of surrogate models using artificial neural network for building shell energy labelling. *Energy Policy*, 69:457–466, 2014. ISSN 0301-4215. doi: https:

//doi.org/10.1016/j.enpol.2014.02.001. URL https://www.sciencedirect.com/science/article/pii/S0301421514000883.

[15] Isacco Simion. L'evoluzione della legislazione italiana ed europea: verso il risparmio energetico, 2013. URL https://www.expoclima.net/.

[16] Nicola Furcolo. Certificazione energetica: quadro normativo e glossario. URL https://biblus.acca.it/focus/attestato-prestazione-energetica-ape/#La_leggen_3731976.

[17] Siape - sistema informativo nazionale degli ape. URL http://www.portale4e.it/centrale_dettaglio_pa.aspx?ID=7.

[18] Decreto legislativo 10 giugno 2020, n. 48. URL https://www.gazzettaufficiale.it/eli/id/2020/06/10/20G00066/sg.

[19] Pnrr, energia da fonti rinnovabili: D.lgs. 199/2021 di attuazione della direttiva red ii. URL https://www.ecolstudio.com/it/news-normative/1035-pnrr-energia-rinnovabile-dlgs-8-novembre-2021-199.html.

[20] Cosa s'intende per indice di prestazione energetica (epgl, ipe, epgl,nren). URL https://www.studiocardilloetripodi.it/home/energia/.

[21] Funzione di densità per variabili casuali continue. URL http://progettomatematica.dm.unibo.it/Prob2/6funzionedidens.html.

[22] A. Kassambara. *Practical Guide to Cluster Analysis in R: Unsupervised Machine Learning*. Multivariate Analysis. STHDA, 2017. ISBN 9781542462709. URL https://books.google.it/books?id=-q3snAAACAAJ.

[23] Shraddha K Popat and M Emmanuel. Review and comparative study of clustering techniques. *International journal of computer science and information technologies*, 5(1):805–812, 2014.

[24] D Opitz and R Maclin. Popular ensemble methods: An empirical study. *The Journal of artificial intelligence research*, 11:169–198, 1999. ISSN 1076-9757.

[25] Hugh A Chipman, Edward I George, and Robert E McCulloch. Bart: Bayesian additive regression trees. 2008.

[26] François Chollet. *Deep learning with R / François Chollet with J.J. Allaire.* Manning Publications, Shelter Island, N.Y, 2018. ISBN 1-61729-554-X.

[27] Tim Miller. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267:1–38, 2019. ISSN 0004-3702. doi: https://doi.org/10.1016/j.artint.2018.07.007. URL https://www.sciencedirect.com/science/article/pii/S0004370218305988.

[28] Been Kim, Rajiv Khanna, and Oluwasanmi Koyejo. Examples are not enough, learn to criticize! criticism for interpretability. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, page 2288–2296, Red Hook, NY, USA, 2016. Curran Associates Inc. ISBN 9781510838819.

[29] Christoph Molnar. *Interpretable Machine Learning*. 2019.

[30] Jerome H. Friedman. Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29(5):1189–1232, 2001. ISSN 00905364. URL http://www.jstor.org/stable/2699986.

[31] Janis Klaise, Arnaud Van Looveren, Giovanni Vacanti, and Alexandru Coca. Alibi explain: Algorithms for explaining machine learning models. *Journal of Machine Learning Research*, 22(181):1–7, 2021. URL http://jmlr.org/papers/v22/21-0017.html.

[32] Przemyslaw Biecek and Tomasz Burzykowski. *Explanatory Model Analysis*. Chapman and Hall/CRC, New York, 2021. ISBN 9780367135591. URL https://pbiecek.github.io/ema/.

[33] Ninh D Pham, Quang Loc Le, and Tran Khanh Dang. Hot asax: A novel adaptive symbolic representation for time series discords discovery. In *Intelligent Information and Database Systems*, Lecture Notes in Computer Science, pages 113–121. Springer Berlin Heidelberg, Berlin, Heidelberg. ISBN 9783642121449.

[34] Alfonso Capozzoli, Gianluca Serale, Marco Savino Piscitelli, and Daniele Grassi. Data mining for energy analysis of a large data set of flats. 2017.

[35] Tania Cerquitelli, Evelina Di Corso, Stefano Proto, Paolo Bethaz, Daniele Mazzarelli, Alfonso Capozzoli, Elena Baralis, Marco Mellia, Silvia Casagrande, and Martina Tamburini. A data-driven energy platform: From energy performance certificates to human-readable knowledge through dynamic high-resolution geospatial maps. *Electronics (Basel)*, 9(2132):2132, 2020. ISSN 2079-9292.

[36] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2021. URL https://www.R-project.org/.

[37] Decreto-legge 19 maggio 2020, n. 34. decreto rilancio. URL https://www.gazzettaufficiale.it/eli/id/1993/10/14/093G0451/sg.

[38] Garbage in, garbage out. URL https://www.worldwidewords.org/qa/qa-gar1.htm.

[39] Max Kuhn. The caret package, 2009.

[40] Cheng Fan, Fu Xiao, and Shengwei Wang. Development of prediction models for next-day building energy consumption and peak power demand using data mining techniques. *Applied Energy*, 127:1–10, 2014. ISSN 0306-2619. doi: https://doi.org/10.1016/j.apenergy.2014.04.016. URL https://www.sciencedirect.com/science/article/pii/S0306261914003596.

[41] Przemyslaw Biecek. Dalex: Explainers for complex predictive models in r. *Journal of Machine Learning Research*, 19(84):1–5, 2018. URL http://jmlr.org/papers/v19/18-416.html.

[42] Chaehan So. Understanding the prediction mechanism of sentiments by xai visualization. In *Proceedings of the 4th International Conference on Natural Language Processing and Information Retrieval*, NLPIR 2020, page 75–80, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450377607. doi: 10.1145/3443279.3443284. URL https://doi.org/10.1145/3443279.3443284.