POLITECNICO DI TORINO

Corso di Laurea Magistrale in Ingegneria Matematica

Tesi di Laurea Magistrale

Sviluppo di una metodologia data-driven per l'analisi delle certificazioni energetiche

Caso di studio: la città di Torino

Relatore:
Prof.ssa Tania Cerquitelli

Correlatore:
Dott.ssa Evelina Di Corso

Candidato:
Débhora Poggi Coutinho de Lira

Tutore Aziendale
Edison Spa
Ing. Silvia Casagrande

Anno Accademico 2018-2019

Ringraziamenti

Vorrei ringraziare innanzitutto la Prof.ssa Cerquitelli per avermi dato la possibilità di far parte di questo progetto di tesi, per la cortesia e la gentilezza dimostratami durante l'intero periodo trascorso insieme.

Non avrei potuto concludere questo percorso senza il sostegno, la disponibilità e la pazienza della Dott.ssa Evelina Di Corso, alla quale rivolgo un ringraziamento speciale. Desidero inoltre esprimere la mia gratitudine nei confronti del Dott. Daniele Mazzarelli per la disponibilità e la gentilezza dimostratami.

Questa tesi è stata svolta presso le Officine Edison, struttura all'interno della quale ho trovato una bellissima accoglienza e grande sostegno da parte di tutti i colleghi di lavoro. In particolare, vorrei ringraziare l'Ing. Silvia Casagrande per la professionalità e la disponibilità ricevuta durante tutto il progetto di tesi. Ringrazio inoltre il collega Diego Albergo per il carisma e la simpatia condivise.

Ringrazio le amiche di sempre che mi hanno continuamente sostenuta sia di persona che telefonicamente fino alla fine di questo percorso: Erika, Rosa e Oana. Vorrei inoltre ringraziare il mio compagno di avventure Alessio, per questa esperienza vissuta insieme e le giornate condivise in armonia all'interno delle Officine Edison.

Vorrei ringraziare di cuore tutta la mia famiglia. Vi ringrazio per aver permesso e favorito il raggiungimento di questo obiettivo. Grazie per avermi supportato, appoggiato e aiutato durante questo importante periodo della mia vita. Senza i vostri sacrifici non avrei potuto raggiungere questo traguardo.

La famiglia non è solo quella del sangue, ma anche quella del cuore. Per questo vorrei ringraziare infinitamente Mary e Francesco, le persone che mi hanno accolto nella loro vita come una figlia. Voglio ringraziarvi per essere sempre dolci con me e per tutto l'affetto che mi mostrate ormai da anni.

Ringrazio dal profondo del mio cuore la mia anima gemella, Andrea. Desidero ringraziarti per tutte le volte che sei stato al mio fianco, non solo durante il periodo universitario ma anche, e soprattutto, nei momenti più difficili della mia vita. Mi hai calmato quando l'ansia prendeva il sopravvento e, orgoglioso, mi hai incoraggiata a dare il meglio di me stessa sempre. Grazie per aver sempre creduto nelle mie capacità e perché fai di me, ogni giorno, una persona migliore. Se sono riuscita a concludere questo percorso è anche grazie a te.

Ogni persona che incontri sta combattendo una
una battaglia di cui non sai nulla.
Sii gentile, sempre.

(Platone)

Indice

Introduzione **1**

1 La Certificazione Energetica **3**
1.1 Storia delle Normative di Riferimento 3
1.1.1 Passaggio da ACE ad APE 4
1.2 Certificati nella Regione Piemonte 6
1.3 Attributi Caratteristici delle Certificazioni 6
1.3.1 Dati Catastali e Geografici 7
1.3.2 Dati Tecnici Generali e Specifici 8
1.3.3 Dati sui Rendimenti e sugli Impianti 10
1.3.4 Dati sui Fabbisogni e sugli Indici di Prestazione 12
1.4 Classificazione Energetica 13
1.4.1 Classe Energetica: ACE 13
1.4.2 Classe Energetica: APE 14

2 Framework TUCANA **16**
2.1 Pre-elaborazione dei Dati 18
2.1.1 Pulizia dei Dati 18
2.1.2 Selezione e Normalizzazione dei Dati 22
2.2 Analisi Esplorativa dei Dati 24
2.2.1 Algoritmo K-Means 26
2.2.2 Scelta dei Parametri 29
2.3 Interpretazione della Conoscenza Estratta 30
2.3.1 Caratterizzazione della Conoscenza 30
2.3.2 Visualizzazione della Conoscenza 34
2.3.3 Applicazione Web Sviluppata 36
2.4 Processo di Generalizzazione 38
2.4.1 Modelli di Regressione 40
2.4.2 Modelli di Classificazione 45

3 Risultati Sperimentali **48**
3.1 Preparazione dei Dati 50
3.1.1 Pulizia dei Dati Geografici 51
3.1.2 Correzioni di Scala e Filtri sulle Variabili 54
3.1.3 Eliminazione degli Outlier 57
3.1.4 Selezione e Normalizzazione delle Variabili 58
3.2 Applicazione dell'Algoritmo K-Means 60
3.2.1 Parametri di Input 60
3.2.2 Risultati del K-Means 61
3.2.3 Caratterizzazione degli Edifici 62
3.3 Visualizzazione dei Risultati Ottenuti 66
3.3.1 Realizzazione delle Mappe 67
3.3.2 Interfaccia dell'Applicazione Web 69
3.4 Modelli di Generalizzazione 72
3.4.1 Implementazione del Modello Semi-Completo 73
3.4.2 Implementazione del Modello Completo 77
3.4.3 Certificati Recuperati 81

4 Conclusioni e Sviluppi Futuri **84**
4.1 Applicazione: Nuovi APE di Torino 85
4.2 Estensione: Nuovo Caso di Studio 85

Bibliografia **87**

Elenco delle figure

1.1 Pagina 1/5 del Fac simile di APE [11]. 5
1.2 Zona climatica e Gradi giorno secondo il Decreto del Presidente della Repubblica 26 agosto 1993, n. 412 [17]. 13

2.1 Architettura sviluppata per il *framework* TUCANA. 16
2.2 Le fasi del processo di *Knowledge Discovery in Databases* (KDD) [20]. 17
2.3 Esempio DBSCAN: l'algoritmo colora di verde i *core point*, di blu i *border point* e di rosso gli *outlier* ©Tan, Steinbach, Kumar, Introduction to Data Mining, McGraw Hill 2006. 21
2.4 Esempio *K-Distance Plot*: setting dei parametri per DBSCAN ©Tan, Steinbach, Kumar, Introduction to Data Mining, McGraw Hill 2006. 22
2.5 Esempio di suddivisione in *cluster* ©Tan, Steinbach, Kumar, Introduction to Data Mining, McGraw Hill 2006. 24
2.6 Esempio di *clustering* partizionale ©Tan, Steinbach, Kumar, Introduction to Data Mining, McGraw Hill 2006. 25
2.7 Esempio di *clustering* gerarchico con il relativo dendrogramma ©Tan, Steinbach, Kumar, Introduction to Data Mining, McGraw Hill 2006. 26
2.8 Esempio 1 - Effetto di una buona scelta random dei centroidi ©Tan, Steinbach, Kumar, Introduction to Data Mining, McGraw Hill 2006. 27
2.9 Esempio 2 - Effetto di una scelta random inadeguata dei centroidi ©Tan, Steinbach, Kumar, Introduction to Data Mining, McGraw Hill 2006. 27
2.10 Esempio di curva ottenuta con l'*elbow method*. 29
2.11 Metodologia per la scelta del parametro K nell'utilizzo del *K-Means*. 30
2.12 Esempio di boxplot con indicazioni statistiche. 31
2.13 Esempio di CART con tre nodi foglia e tre classi (setosa, virginica e versicolor) sul *dataset* Iris. 32
2.14 Esempio di radar chart per due gruppi diversi di variabili. 33

2.15 Esempio di mappa coropletica: tasso di disoccupazione in Italia nel 2017. Fonte Dati Istat, immagine esibita da BBC. 34
2.16 Esempio di mappa *scatter* con i simboli grafici detti *marker*. 35
2.17 Confronto tra una mappa *scatter* e la mappa *marker-cluster*. 36
2.18 Semplice esempio di modello *client-server* ©TechTerms.com. 37
2.19 Esempio di predizione di una variabile quantitativa e qualitativa. . . 39
2.20 Esempio di suddivisione del *dataset* per la validazione dei modelli: *leave-one-out* in alto e una fase dell'algoritmo *K-fold* con $K = 6$ in basso. 39
2.21 Metodo dei minimi quadrati su una serie di dati ©James, Witten, Hastiea, Tibshirani, An Introduction to Statistical Learning. 41
2.22 Esempio di funzione non lineare per la regressione polinomiale ©James, Witten, Hastiea, Tibshirani, An Introduction to Statistical Learning. 42
2.23 Esempio di output del K-NN, a sinistra la scelta dei $K = 3$ vicini, a destra il limite decisionale definito per $K = 3$ ©James, Witten, Hastiea, Tibshirani, An Introduction to Statistical Learning. 45
2.24 Esempio di matrice di confusione. 46

3.1 Distribuzione delle province della Regione Piemonte negli *open data*. 49
3.2 Distribuzione delle destinazioni d'uso degli immobili negli *open data*. 50
3.3 Distribuzione delle *trasmittanze opache*, *trasparenti* e del *fattore forma*. 56
3.4 Distribuzione dei quattro *rendimenti* dei sottosistemi. 56
3.5 Risultato del gESD sull'attributo *anno di costruzione*. La distribuzione dei dati (in blu) si discosta dagli *outlier* (in rosso). 57
3.6 *K-Distance Plot* per la scelta dei parametri per il DBSCAN. 58
3.7 Matrice di correlazione degli attributi scelti per le analisi. 59
3.8 Scelta automatica del K per il *K-Means* con l'*elbow method*. 61
3.9 Riepilogo delle circoscrizioni che suddividono la città di Torino. . . 62
3.10 *Boxplot* delle variabili di analisi di un *cluster* performante. 63
3.11 *Boxplot* delle variabili di analisi di un *cluster* non performante. . . . 64
3.12 *Radar chart* di un centroide relativo ad un *cluster* performante (a sinistra) e di un centroide corrispondente ad un *cluster* non performante (a destra). 64
3.13 Esempio di una parte del CART generato dal *framework*. 65
3.14 Mappa coropletica per la *superficie disperdente totale*. 67
3.15 Esempio di mappa *scatter* generata dal *framework*. 68

3.16 Mappa *marker-cluster* ad un livello di dettaglio inferiore (a sinistra) e ad un livello di dettaglio maggiore (a destra). 68
3.17 Pagina html dell'applicazione web per l'autenticazione dell'utente. . 70
3.18 Esempio di statistiche specifiche sui *cluster*, presenti sul sito web e visualizzabili solo dagli utenti con un certo grado di *expertise*. . . . 70
3.19 Pagina html con la mappa coropletica per l'attributo *rendimento climatizzazione invernale* nella sezione per l'utente pubblica amministrazione. 71
3.20 *Grid-search* per l'algoritmo K-NN nel modello semi-completo con tutte le variabili di analisi. 74
3.21 *Grid-search* per l'algoritmo K-NN nel modello semi-completo con le sole variabili geometriche. 76

4.1 Distribuzione delle province e delle destinazioni d'uso degli *open data* nella Regione Piemonte. 86

Elenco delle tabelle

1.1 Elenco delle destinazioni d'uso stabilite nel Decreto del Presidente della Repubblica 26 agosto 1993, n. 412 [17]. 9
1.2 Limiti normativi per l'EP_L secondo il D.p.r. n. 412/1993 [17]. 14
1.3 Scala di classificazione della classe energetica secondo il Decreto n. 162/2015. 15

3.1 Riepilogo di alcune librerie in R utilizzate dal *framework*. 48
3.2 Riepilogo di alcune librerie in Python utilizzate dal *framework*. . . . 49
3.3 Esempi di risoluzione degli indirizzi tramite il confronto con il viario. 53
3.4 Esempi di risoluzione degli indirizzi con il servizio Geocoding API. . 54
3.5 Risultati ottenuti con la procedura di pulizia dei dati geografici. . . 54
3.6 Filtri sulla base dei range di accettabilità. 55
3.7 Dettaglio della percentuale di certificati energetici eliminati dal *database* per ogni attributo fuori dal range di ammissibilità. 55
3.8 Riepilogo delle variabili significative scelte per le analisi. 59
3.9 Percentuale di certificati suddivisi per *cluster* e circoscrizioni. 61
3.10 Classificazione e caratterizzazione dei gruppi di edifici trovati. . . . 66
3.11 Valori di precisione e richiamo per il modello semi-completo con tutte le variabili di analisi. 75
3.12 Valori di precisione e richiamo medio per il modello semi-completo con tutte le variabili di analisi. 75
3.13 Valori di precisione e richiamo per il modello semi-completo con tutte le variabili di analisi. 77
3.14 Valori di precisione e richiamo per il modello semi-completo con tutte le variabili di analisi. 77
3.15 *Grid-search* per la regressione con il K-NN nel modello completo con l'attributo di analisi *ETAH* mancante. 79
3.16 *Grid-search* per la regressione Lasso nel modello completo con l'attributo di analisi *ETAH* mancante. 79

3.17 Risultato della *grid-search* nel modello completo con l'attributo di analisi *ETAH* mancante. 79
3.18 Confronto tra gli algoritmi di regressione nel modello completo con l'attributo di analisi *ETAH* mancante. 80
3.19 *Grid-search* per la regressione Lasso e K-NN nel modello completo con l'attributo di analisi *Rapporto S/V* mancante. 81
3.20 Confronto tra gli algoritmi di regressione nel modello completo con l'attributo di analisi *Rapporto S/V* mancante. 81
3.21 Numero di certificati energetici eliminati dalla base di dati per ogni attributo fuori dal range di ammissibilità. 82
3.22 Numero di certificati energetici recuperati con i modelli di regressione. 82

Introduzione

Il tema dell'energia è un argomento di interesse sia nello studio e sviluppo di tecnologie future, che nel trascorrere quotidiano di ognuno di noi. Nello specifico, osserviamo che si cercano sempre di più delle soluzioni e dei modelli che migliorino l'efficienza delle nostre abitazioni, minimizzando l'impatto ambientale. A tal proposito, il Consiglio Europeo ha convenuto di istituire nel 2015 un'*Unione dell'energia*, la quale ha come strategia principale trasformare l'Europa in un'economia con un'efficienza energetica elevata e a basse emissioni di carbonio [1].

In questo nuovo scenario diventa di fondamentale importanza la certificazione energetica degli edifici e, in particolare, l'*Attestato di Prestazione Energetica* (APE), documento che permette di stimare le performance energetiche di un edificio basandosi sulle sue caratteristiche termofisiche e geometriche. Con l'introduzione delle nuove normative in Italia per gli APE, viene garantita un'archiviazione standard e centralizzata dei dati energetici nel *Sistema Informativo sugli Attestati di Prestazione Energetica* (SIAPE). Di conseguenza, la libera disponibilità dei dati consente l'utilizzo di tecniche di *data mining* al fine di estrarre conoscenza utile e creare modelli decisionali [3].

Questo lavoro di tesi propone una metodologia per la caratterizzazione energetica degli edifici della città di Torino, mediante l'utilizzo di tecniche di *data mining* applicate sui dati degli attestati di prestazione energetica. Il primo obiettivo principale della tesi si basa sulla creazione di un *framework* che consenta di estrarre informazioni utili dai dati degli APE con tecniche di analisi opportune, caratterizzare i risultati ottenuti e visualizzarli tramite mappe interattive e geolocalizzate. Inoltre, il *framework* proposto consente di generalizzare la conoscenza estratta dai risultati ottenuti nelle fasi di analisi. Si pone come secondo obiettivo fondamentale invece, la progettazione e lo sviluppo di un'applicazione web che permetta di esplorare i risultati sperimentali ottenuti e di rendere la conoscenza estratta disponibile e facilmente consultabile dai diversi utenti finali. Al fine di soddisfare i propositi appena citati, si articola la presente tesi in 5 diversi capitoli di cui vengono illustrati una breve descrizione.

Nel **Capitolo 1** si introduce il concetto di certificazione energetica e si presenta la storia delle normative di riferimento negli anni. In seguito, viene fatta una descrizione approfondita delle variabili principali che definiscono le certificazioni energetiche;

Nel **Capitolo 2** viene presentata l'architettura del *framework* sviluppato e descritte tutte le fasi che lo compongono. In particolare, vengono discussi nel dettaglio i concetti teorici che sono stati utilizzati per progettare ogni fase di estrazione della conoscenza dalle certificazioni energetiche.

Nel **Capitolo 3** si descrivono i risultati sperimentali ottenuti dall'applicazione di ogni componente del *framework* sul *database* delle certificazioni energetiche. Si utilizzano diversi grafici, tabelle e mappe al fine di esporre al meglio le informazioni estratte con le tecniche di *data mining*;

Nel **Capitolo 4** vengono riassunti gli aspetti più importanti del *framework*, tratte le conclusioni sul lavoro svolto e suggeriti gli eventuali sviluppi futuri.

Capitolo 1

La Certificazione Energetica

La certificazione energetica è un documento che attesta i consumi energetici medi annuali di un immobile e che consente di classificarlo tramite la cosiddetta *classe energetica*. Il certificato energetico di un edificio riporta le sue varie caratteristiche geometriche e termofisiche, oltre che altri aspetti tipici della sua zona climatica di appartenenza. Poiché questo documento fornisce informazioni chiare e trasparenti sulla qualità energetica dell'immobile, esso diventa uno strumento fondamentale qualora si voglia ottimizzare le sue prestazioni energetiche.

Il consumo energetico degli edifici è costantemente aumentato dal 2008, rappresentando dal 20% al 40% del consumo totale di energia primaria nei paesi sviluppati [4]. La tendenza al rialzo del consumo di energia è dovuta alla crescita della popolazione, all'aumento della domanda dei servizi di costruzione e dei livelli di comfort, e all'aumento delle ore trascorse all'interno degli edifici [5].

L'attenzione al settore energetico è sempre più giustificata dalle riunioni organizzate dal Consiglio europeo, che chiede di riesaminare e sviluppare la normativa relativa alla riduzione delle emissioni di gas serra, all'efficienza energetica nell'edilizia abitativa e alle energie rinnovabili per sostenere il conseguimento degli obiettivi concordati per il 2030 [6]. La certificazione energetica diventa, quindi, un sistema di valutazione attraverso cui studiare delle misure efficaci per migliorare la sicurezza energetica e ridurre i costi energetici dell'Unione Europea.

1.1 Storia delle Normative di Riferimento

Prima degli anni '70, in Italia, non esistevano obblighi legislativi nell'ambito del settore energetico ma, a seguito della crisi energetica petrolifera di quegli anni, venne emanata la prima direttiva di riferimento, la Legge n. 373/1976. Essa

fu la prima norma a limitare la dispersione termica verso l'esterno degli edifici residenziali e industriali. Nonostante una maggiore attenzione verso l'isolamento termico degli edifici, fu con la Legge n. 10/1991 (attuata successivamente con il D.P.R. n. 412/1993) che vennero inclusi dei vincoli nelle modalità progettuali e di gestione del sistema edificio-impianto [7]. Oltre all'individuazione dei criteri di progettazione energetica, con il D.P.R. n. 412/1993, vennero introdotte la classificazione del territorio nazionale in funzione del numero di gradi giorno e degli edifici in base alla loro destinazione d'uso.

Nel 2002 venne emanata dal Parlamento Europeo la Direttiva 2002/91/CE il cui intento era orientare i paesi membri verso un concetto di efficienza energetica che consenta di perseguire gli obiettivi rivolti alla riduzione dell'impatto ambientale [8]. Tale norma venne recepita in Italia con i D.Lgs. n. 192/2005 e D.Lgs. n. 311/2006 che introducono i requisiti minimi delle prestazioni energetiche e i criteri per la certificazione energetica, sia per gli edifici di nuova costruzione che per quelli in via di ristrutturazione [7].

Con il D.P.R. n. 59/2009 vennero definite le metodologie, i criteri e i requisiti minimi degli impianti relativamente alla climatizzazione invernale ed estiva, alla predisposizione di acqua calda sanitaria e all'illuminazione artificiale [8]. Vennero introdotte inoltre le norme tecniche UNI/TS 11300-1 e UNI/TS 11300-2, riconosciute a livello nazionale, per il calcolo delle prestazioni energetiche. Le regioni e le province italiane avevano spesso direttive personalizzate quindi, con il D.M. 26/06/2009, vengono emanate le direttive per la certificazione energetica degli edifici valide su tutto il territorio nazionale.

Inoltre, l'attuazione della Direttiva 2009/28/CE tramite il D.Lgs. n. 28/2011 introdusse l'obbligo di riportare, in sede di compravendita o locazione di un immobile, l'indice di prestazione e le relative informazioni riguardanti la propria certificazione energetica [9]. Altra novità importante venne introdotta dal D.M. 22/12/2012 che esclude la possibilità di autocertificazione in classe energetica G da parte dei proprietari degli immobili [7].

1.1.1 Passaggio da ACE ad APE

Il Decreto Eco-bonus/Energia identificato con D.L. n. 63/2013 sostituisce l'Attestato di Certificazione Energetica (ACE) con l'Attestato di Prestazione Energetica (APE) ed effettua altri cambiamenti, facendo pensare alla realizzazione di un vero e proprio "censimento" del profilo energetico del patrimonio immobiliare [7]. Questo decreto è stato convertito con modificazioni nella Legge n. 90/2013 attuata successivamente con i decreti ministeriali D.M. 26/06/2015. Dal 1° ottobre

2015 si trovano in vigore le nuove regole sui requisiti minimi per la redazione del documento APE [10], di cui si riporta un esempio in Figura 1.1.

La principale differenza tra i due certificati risiede nella norma europea a cui fanno riferimento, in particolare, l'ACE recepisce la Direttiva 2002/91/CE mentre l'APE recepisce la Direttiva 2010/31/UE. Nello specifico, per questo nuovo attestato, viene definito un nuovo metodo per l'identificazione della classe energetica degli edifici (maggiori dettagli sono presentati nella sezione 1.4), inserito un nuovo indicatore della prestazione energetica invernale ed estiva dell'involucro edilizio e specificate le nuove linee guida per le procedure di calcolo [12].

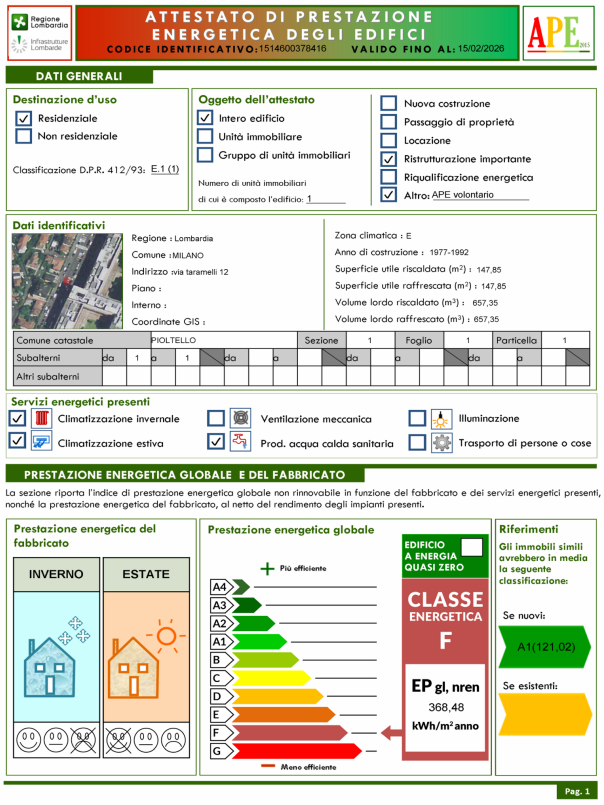

Figura 1.1: Pagina 1/5 del Fac simile di APE [11].

Nel panorama più recente entra in vigore la nuova Direttiva 2018/844/UE del 30 maggio 2018 che modifica la Direttiva 2010/31/UE sulla prestazione energetica nell'edilizia. La direttiva, approvata dal Consiglio Europeo, ha diversi obiettivi tra cui rendere più efficaci le strategie di ristrutturazione degli immobili a lungo termine per un settore edile idealmente decarbonizzato entro il 2050, sostenere lo sviluppo delle infrastrutture per l'elettromobilità, migliorare la trasparenza

delle metodologie di calcolo della prestazione energetica e accrescere il ruolo dei consumatori, informandoli e proteggendoli dalla povertà energetica [13]. Viene ripreso inoltre il concetto di edifici ad energia quasi zero, già introdotti con la Direttiva 2010/31/UE, e indicati come quegli edifici con fabbisogno energetico quasi nullo, coperto in misura significativa da fonti rinnovabili, prodotte all'interno del confine energetico dell'edificio [12].

1.2 Certificati nella Regione Piemonte

Con la delibera D.G.R n. 14-2119/2015 la Regione si è uniformata alle disposizioni nazionali emanate in attuazione della Direttiva 2010/31/UE sulla prestazione e certificazione energetica degli edifici. Il nuovo provvedimento introduce l'obbligo di allegare l'APE ai documenti che riguardano sia gli edifici di nuova costruzione sia quegli che sono oggetti di compravendita o locazione. Nel 2015 la Regione Piemonte ha inoltre realizzato, con il supporto di CSI-Piemonte[1], il cosiddetto *Sistema informativo per la Prestazione Energetica degli Edifici* (SIPEE) [14], con l'intento di gestire l'elenco regionale dei soggetti abilitati al rilascio dei certificati energetici, i dati inseriti negli APE e la raccolta degli attestati trasmessi dai professionisti. Un aggiornamento ulteriore è stato fatto nelle procedure di controllo della qualità delle attestazioni. Infatti, è stato previsto che l'*Agenzia Regionale per la Protezione Ambientale* (ARPA) effettui controlli a campione allo scopo di verificare la validità dei certificati e il rispetto dei vincoli imposti [15].

Gli attestati energetici oggetto di analisi in questa tesi sono stati rilasciati dal CSI-Piemonte seguendo i temi degli *open data*[2]. Ciò significa, che si hanno a disposizione le stesse informazioni richieste durante la compilazione della procedura telematica attraverso il sistema informativo SIPEE, da parte dei certificatori autorizzati. Nello specifico, vengono considerate le certificazioni energetiche dagli anni 2009 al 2014 (identificate ancora con ACE) e dal 2016 al primo semestre del 2018 (sostituite con l'acronimo APE).

1.3 Attributi Caratteristici delle Certificazioni

In questo paragrafo vengono descritte le variabili maggiormente interessate durante la compilazione dei certificati APE all'interno del SIPEE. In particolare,

[1]Consorzio per il Sistema Informativo, http://www.csipiemonte.it/web/it/.

[2]Dati pubblicati sul web che sono liberamente consultabili e utilizzabili da altri enti informativi.

gli attributi dei certificati energetici considerati durante le analisi possono essere suddivisi tra: dati catastali e geografici, dati tecnici generali e specifici sul fabbricato, dati sui rendimenti e gli indici di fabbisogno e i dati sugli impianti dello stabile.

Osservazione 1.1. *All'interno della base di dati sono presenti anche alcuni dati personali di natura sensibile come il nome del soggetto certificatore, regolarmente iscritto sul portale SIPEE, insieme al suo numero identificativo.*

1.3.1 Dati Catastali e Geografici

I dati identificativi catastali sono attributi che permettono di individuare univocamente un'unità immobiliare [16]. Fra i dati necessari per effettuare la ricerca di un immobile all'interno degli archivi catastali nazionali, si citano quelli presenti nei certificati energetici analizzati.

- **Foglio**: numero che indica la porzione di territorio comunale che il catasto rappresenta nelle proprie mappe cartografiche;
- **Particella**: numero che rappresenta una porzione di terreno, o un fabbricato;
- **Subalterno**: numero che identifica la singola unità immobiliare esistente su una particella (quando il fabbricato è costituito da un'unica unità immobiliare il subalterno potrebbe essere assente).

L'insieme costituito da questi tre attributi [Foglio, Particella, Subalterno] permette di riferirsi in maniera univoca ad un'unità immobiliare presente all'interno della base di dati su cui verranno fatte le analisi.

Osservazione 1.2. *In presenza di più certificazioni APE per la stessa un'unità immobiliare, si è deciso di considerare l'attestato più recente rispetto alla data di compilazione nel portale SIPEE.*

Oltre ai dati appena esposti, sono di fondamentale importanza anche gli attributi che permettono di caratterizzare gli edifici dal punto di vista geografico. Si elencano di seguito tali variabili.

- **Indirizzo**: campo testuale libero che identifica la via, strada, piazza, contrada o zona dell'unità immobiliare a cui fa riferimento il certificato APE;
- **Numero civico**: campo testuale libero che riconosce il codice numerico o alfanumerico identificativo dell'immobile nel contesto di un certo indirizzo;

- **Codice di Avviamento Postale**: campo numerico formato da cinque cifre, noto abitualmente con la sigla CAP, che si aggiunge all'indirizzo;
- **Comune**: campo testuale libero corrispondente al comune in cui si trova l'unità immobiliare a cui si riferisce l'APE;
- **Provincia**: campo testuale libero corrispondente alla provincia (ente locale avente una competenza su un gruppo di comuni) a cui appartiene l'immobile;
- **Latitudine**: numero che indica la distanza angolare, misurata in gradi, lungo l'arco di meridiano compreso tra l'Equatore e il parallelo passante per il punto considerato;
- **Longitudine**: numero che indica la distanza angolare, misurata in gradi, lungo l'arco di parallelo compreso tra il Meridiano di Greenwich e il meridiano passante per il punto considerato.

1.3.2 Dati Tecnici Generali e Specifici

Esistono altri attributi che caratterizzano gli attestati APE, come i dati tecnici generali che esprimono delle informazioni sulla struttura dell'edificio e i dati tecnici specifici sul fabbricato, ovvero indicazioni dal punto di vista geometrico e fisico.

- **Anno di costruzione**: indica l'anno di realizzazione dell'unità abitativa;
- **Anno ultima ristrutturazione**: coincide con l'anno delle ultima opera edile importante effettuata sull'edificio ai fini di riportarlo al rispetto delle normative del periodo;
- **Destinazione d'uso**: descrive la finalità di utilizzo dell'immobile secondo la classificazione data dal *D.p.r. n. 412/1993* e illustrata nella Tabella 1.1;
- **Tipologia di edificio**: si riferisce alle caratteristiche costruttive dell'edificio, cioè se ad esempio è un appartamento, una villetta a schiera o altro;
- **Superficie utile riscaldata** [m^2]: si intende la superficie interessata al funzionamento del servizio energetico di climatizzazione invernale;
- **Superficie utile raffrescata** [m^2]: si intende la superficie interessata al funzionamento del servizio energetico di climatizzazione estiva;
- **Superficie utile** [m^2]: rappresenta l'unione delle superfici riscaldate e/o raffrescate dell'unità immobiliare;

Destinazione d'uso	Descrizione
E.1	Edifici di tutte le tipologie 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 1.1: Elenco delle destinazioni d'uso stabilite nel Decreto del Presidente della Repubblica 26 agosto 1993, n. 412 [17].

- **Volume lordo riscaldato** [m^3]: si tratta della somma dei volumi lordi di tutti i vani riscaldati dell'edificio;
- **Volume lordo raffrescato** [m^3]: equivale alla somma dei volumi lordi di tutti i locali raffrescati dell'edificio;

- **Superficie disperdente totale** [m^2]: si tratta della somma di tutte le superfici dell'involucro riscaldato;
- **Fattore forma o Rapporto S/V** [m^{-1}]: indica il rapporto tra la superficie disperdente totale e il volume lordo riscaldato dell'immobile [18].

Osservazione 1.3. *Si hanno valori elevati del fattore forma per le superfici disperdenti più estese, le quali determinano maggiori dispersioni termiche. Questo parametro quindi, può esprimere la compattezza di un edificio sottolineando la propensione alla scelta di edifici con forme semplici, senza molte sporgenze.*

Esistono ancora altri due importanti attributi strutturali che caratterizzano gli attestati APE, le trasmittanze opache e trasparenti. Queste variabili fanno riferimento al concetto di trasmittanza termica, cioè il flusso di calore medio che passa, per metro quadrato di superficie, attraverso una struttura che delimita due ambienti a diversa temperatura[3]. Questi attributi sono molto utili per calcolare le dispersioni termiche che avvengono attraverso l'involucro di un edificio e, per questo motivo, possono influenzare notevolmente le sue prestazioni energetiche.

- **Trasmittanza media delle superfici opache** [W/m^2K]: misura la trasmittanza termica media ponderata attraverso le strutture opache che delimitano l'edificio con l'ambiente esterno;
- **Trasmittanza media delle superfici trasparenti** [W/m^2K]: misura la trasmittanza termica media ponderata degli elementi trasparenti dell'edificio confinanti con l'ambiente esterno.

1.3.3 Dati sui Rendimenti e sugli Impianti

Il rendimento è un parametro fisico con cui si esprime il grado di efficienza di una macchina. Quando si utilizzano impianti reali esistono sempre delle perdite di efficienza rispetto ai rendimenti teorici, perché i sistemi di riscaldamento reali non sono in grado di eliminare completamente alcune perdite di calore [19]. Per tale ragione, il miglioramento delle prestazioni energetiche di un edificio dipende dall'analisi dei quattro rendimenti che caratterizzano gli impianti di riscaldamento. I rendimenti descritti di seguito sono quelli medi stagionali, gli unici interessanti ai fini del risparmio energetico.

[3]Informazioni reperite dall'Agenzia Nazionale Efficienza Energetica (ENEA), sezione dedicata ai tecnici nel sito http://www.acs.enea.it/.

- **Rendimento di generazione**: rapporto fra il calore utile prodotto dal generatore nella stagione di riscaldamento e l'energia fornita nello stesso periodo sotto forma di combustibile ed energia elettrica [19];
- **Rendimento di emissione**: rapporto fra il calore richiesto per il riscaldamento dei locali con un sistema di emissione teorico di riferimento (in grado di fornire una temperatura ambiente perfettamente uguale e uniforme nei diversi ambienti) ed il sistema di emissione reale [19];
- **Rendimento di regolazione**: rapporto fra il calore richiesto per il riscaldamento dei locali una regolazione teorica perfetta ed il calore richiesto per il riscaldamento degli stessi ambienti con un sistema di regolazione reale [19];
- **Rendimento di distribuzione**: rapporto fra la somma del calore utile emesso dai corpi scaldanti e del calore disperso dalla rete di distribuzione, e il calore in uscita dall'impianto di generazione ed immesso nella rete di distribuzione [19].

Osservazione 1.4. *Per poter disporre, ai fini delle analisi successive, di un attributo che riassuma in modo sintetico i rendimenti dei quattro sottosistemi relativi alla climatizzazione invernale, viene generata la variabile ETAH calcolata come il prodotto dei quattro rendimenti appena citati, indicati rispettivamente con ETAG, ETAE, ETAR ed ETAD.*

Oltre ai rendimenti, l'attestato di prestazione energetica si riferisce ad altre variabili per il calcolo della prestazione energetica dell'edificio, come alcuni attributi relativi agli impianti dell'abitazione.

- **Tipologia di impianto**: nei certificati energetici ACE indicava il tipo di impianto di riscaldamento presente (centralizzato o autonomo) ma, con il nuovo attestato APE, è diventato un campo libero in cui si possono riportare anche più di due impianti (e.g. caldaia standard, riscaldamento elettrico);
- **Combustibile**: indica la tipologia di combustibile utilizzata per il servizio di riscaldamento e generazione di acqua calda sanitaria (gas naturale, energia elettrica o altro);
- **Potenza termica nominale** [kW]: se presente, indica la potenza termica dell'impianto effettivamente resa all'ambiente, può essere riportata sul relativo libretto di istruzioni.

1.3.4 Dati sui Fabbisogni e sugli Indici di Prestazione

Per poter determinare la classe energetica di un edificio, spiegata in dettaglio nel paragrafo 1.4, sono necessari ulteriori attributi energetici. Al fine di comprendere meglio le dinamiche con cui vengono assegnate le classi energetiche degli immobili, si riportano di seguito le variabili intese come fabbisogni di energia e indici di prestazione energetica[4].

- **Fabbisogno di energia termica utile** [kWh/m^2]: indica la quantità di energia primaria richiesta per mantenere negli ambienti riscaldati la temperatura di progetto. Per semplicità lo si indica con $\mathbf{Q_{H,nd}}$;
- **Fabbisogno di energia primaria per acqua calda sanitaria** [kWh/m^2]: indica la quantità di energia necessaria ad assicurare l'erogazione dell'acqua calda sanitaria, tenendo conto dell'efficienza del sistema che la produce. Per semplicità lo si indica con $\mathbf{Q_{ACS,w}}$;
- **Prestazione energetica invernale** [kWh/m^{2}anno]: rappresenta l'indice di prestazione energetica invernale dell'involucro e si calcola come il rapporto tra $\mathsf{Q}_{\mathsf{H,nd}}$ e la superficie utile. Per semplicità lo si indica con $\mathbf{EP_{H,nd}}$;
- **Prestazione energetica acqua calda sanitaria** [kWh/m^{2}anno]: rappresenta l'indice di prestazione energetica per la preparazione dell'acqua calda sanitaria e si calcola come il rapporto tra $\mathsf{Q}_{\mathsf{ACS,w}}$ e la superficie utile. Per semplicità lo si indica con $\mathbf{EP_{ACS,w}}$;
- **Indice di prestazione energetica invernale** [kWh/m^{2}anno]: calcolato come il rapporto tra $\mathsf{EP}_{\mathsf{H,nd}}$ e il rendimento globale medio stagionale dell'impianto di riscaldamento. Di solito si indica con il simbolo $\mathbf{EP_H}$;
- **Indice di prestazione energetica acqua calda sanitaria** [kWh/m^{2}anno]: calcolato come il rapporto tra $\mathsf{EP}_{\mathsf{ACS,w}}$ e il rendimento globale medio stagionale dell'impianto di produzione per l'acqua calda sanitaria. Di solito si indica con il simbolo $\mathbf{EP_{ACS}}$.

Osservazione 1.5. *L'elenco degli attributi appena presentati non risulta essere esaustivo per determinare la classe energetica dei nuovi attestati di prestazione energetica (APE), si veda nel dettaglio la sezione 1.4.2.*

[4]Informazioni reperite dal documento *Metodo di Calcolo Semplificato della Prestazione Energetica Globale degli Edifici Residenziali Esistenti.*

1.4 Classificazione Energetica

La certificazione energetica degli edifici consiste in una procedura che termina con la classificazione energetica regionale dell'immobile. La classe energetica è un indicatore alfanumerico che esprime in modo sintetico il grado di efficienza energetica di un edificio. Il metodo per la definizione dell'etichetta di classe negli attestati APE è cambiato rispetto a quello utilizzato nei certificati ACE, perciò, può essere utile esaminare brevemente le due diverse metodologie.

1.4.1 Classe Energetica: ACE

Per ottenere la classe energetica degli edifici nei certificati ACE si dovevano considerare alcuni attributi caratteristici, utilizzati fino al 1° ottobre 2015 e descritti di seguito. A tali attributi si associa inoltre l'immagine presentata in Figura 1.2.

- **Gradi giorno**: si tratta della somma, estesa a tutti i giorni di un periodo annuale convenzionale di riscaldamento, delle sole differenze positive giornaliere tra la temperatura dell'ambiente (fissata a 20°C per convenzione) e la temperatura media esterna giornaliera [17]. Maggiori sono i gradi giorno (GG) e più rigido è il clima;
- **Zona climatica**: indica la suddivisione del territorio nazionale in sei diverse zone climatiche in funzione dei gradi giorno, che non sono individuate dalla loro temperatura indipendentemente dalla localizzazione geografica.

Gradi Giorno	Ore/giorno	Periodo	Zona
GG ≤ 600	6	1/12 – 15/03	A
600 < GG ≤ 900	8	1/12 – 31/03	B
900 < GG ≤ 1400	10	15/11 – 31/03	C
1400 < GG ≤ 2100	12	15/11 – 31/03	D
2100 < GG ≤ 3000	14	15/10 – 15/04	E
GG > 3000	–	–	F

Figura 1.2: Zona climatica e Gradi giorno secondo il Decreto del Presidente della Repubblica 26 agosto 1993, n. 412 [17].

L'**Indice di prestazione energetica globale lordo (EP_L)** definito come la somma tra l'indice di prestazione energetica per la climatizzazione invernale e

quello per la produzione di acqua calda per usi igienici e sanitari:

$$\mathsf{EP_L} = \mathsf{EP_H} + \mathsf{EP_{ACS}}\,,$$

consente di determinare la classe energetica di un edificio confrontando i valori dell'$\mathsf{EP_L}$ con i limiti imposti dalla normativa D.P.R. n. 412/1993, come evidenziato dalla Tabella 1.2. Si osservi che, nei certificati ACE, questi due indici di prestazione tengono conto dei gradi giorno (GG) relativi alla localizzazione dell'edificio, in questo modo si permette il confronto fra abitazioni distribuite in zone diverse.

Limiti $\mathsf{EP_L}$ $[\mathsf{kWh/m^2anno}]$	**Classe energetica**
$\mathsf{EP_L} < 27$	A+
$27 \leq \mathsf{EP_L} < 44$	A
$44 \leq \mathsf{EP_L} < 82$	B
$82 \leq \mathsf{EP_L} < 143$	C
$143 \leq \mathsf{EP_L} < 201$	D
$201 \leq \mathsf{EP_L} < 249$	E
$249 \leq \mathsf{EP_L} < 300$	F
$300 \leq \mathsf{EP_L} < 436$	G
$436 \leq \mathsf{EP_L}$	NC

Tabella 1.2: Limiti normativi per l'$\mathsf{EP_L}$ secondo il D.p.r. n. 412/1993 [17].

1.4.2 Classe Energetica: APE

Con l'introduzione degli attestati APE, il metodo di calcolo per l'assegnazione della classe energetica si basa sulla definizione dell'**Indice di prestazione energetica globale non rinnovabile ($\mathbf{EP_{gl,nren}}$)** ottenuto dalla somma degli indici di prestazione non rinnovabile dei singoli servizi energetici:

$\mathbf{EP_{H,nren}}$: prestazione energetica per la climatizzazione invernale;

$\mathbf{EP_{W,nren}}$: prestazione energetica per la produzione di acqua calda sanitaria;

$\mathbf{EP_{C,nren}}$: prestazione energetica per la climatizzazione estiva;

$\mathbf{EP_{V,nren}}$: prestazione energetica relativo alla ventilazione;

$\mathbf{EP_{L,nren}}$: prestazione energetica relativo all'illuminazione;

$\mathbf{EP_{T,nren}}$: prestazione energetica relativo al trasporto di persone o cose.

La determinazione della classe energetica dipende dal confronto tra l'indice calcolato sulla base dei dati reali dell'edificio in esame ($EP_{gl,nren}$) e l'indice relativo ad un corrispondente edificio di riferimento ($EP_{gl,nren,rif}$), come viene mostrato nella Tabella 1.3.

Limiti $EP_{gl,nren}$ [kWh/m^2anno]	**Classe energetica**
$EP_{gl,nren} \leq 0{,}4\ EP_{gl,nren,rif}$	A4
$0{,}4\ EP_{gl,nren,rif} < EP_L \leq 0{,}6\ EP_{gl,nren,rif}$	A3
$0{,}6\ EP_{gl,nren,rif} < EP_L \leq 0{,}8\ EP_{gl,nren,rif}$	A2
$0{,}8\ EP_{gl,nren,rif} < EP_L \leq 1{,}0\ EP_{gl,nren,rif}$	A1
$1{,}0\ EP_{gl,nren,rif} < EP_L \leq 1{,}2\ EP_{gl,nren,rif}$	B
$1{,}2\ EP_{gl,nren,rif} < EP_L \leq 1{,}5\ EP_{gl,nren,rif}$	C
$1{,}5\ EP_{gl,nren,rif} < EP_L \leq 2{,}0\ EP_{gl,nren,rif}$	D
$2{,}0\ EP_{gl,nren,rif} < EP_L \leq 2{,}6\ EP_{gl,nren,rif}$	E
$2{,}6\ EP_{gl,nren,rif} < EP_L \leq 3{,}5\ EP_{gl,nren,rif}$	F
$EP_{gl,nren} > 3{,}5\ EP_{gl,nren,rif}$	G

Tabella 1.3: Scala di classificazione della classe energetica secondo il Decreto n. 162/2015.

L'*edificio di riferimento* serve per determinare il valore di energia primaria limite di legge che l'edificio di progetto deve rispettare [12]. Il vantaggio dell'APE nell'utilizzo di questo edificio di riferimento è che i limiti normativi vengono modellati sull'edificio oggetto della certificazione energetica.

Capitolo 2

Framework TUCANA

In questo capitolo vengono presentati i concetti teorici e le metodologie che hanno contribuito alla progettazione e allo sviluppo del *framework* TUCANA (TUrin Certificates ANAlysis). L'architettura progettata ha lo scopo di caratterizzare e visualizzare la conoscenza utile presente all'interno degli attestati di prestazione energetica relativi alla città di Torino. Nei paragrafi successivi verranno descritti nel dettaglio le diverse fasi di analisi mostrate nella Figura 2.1.

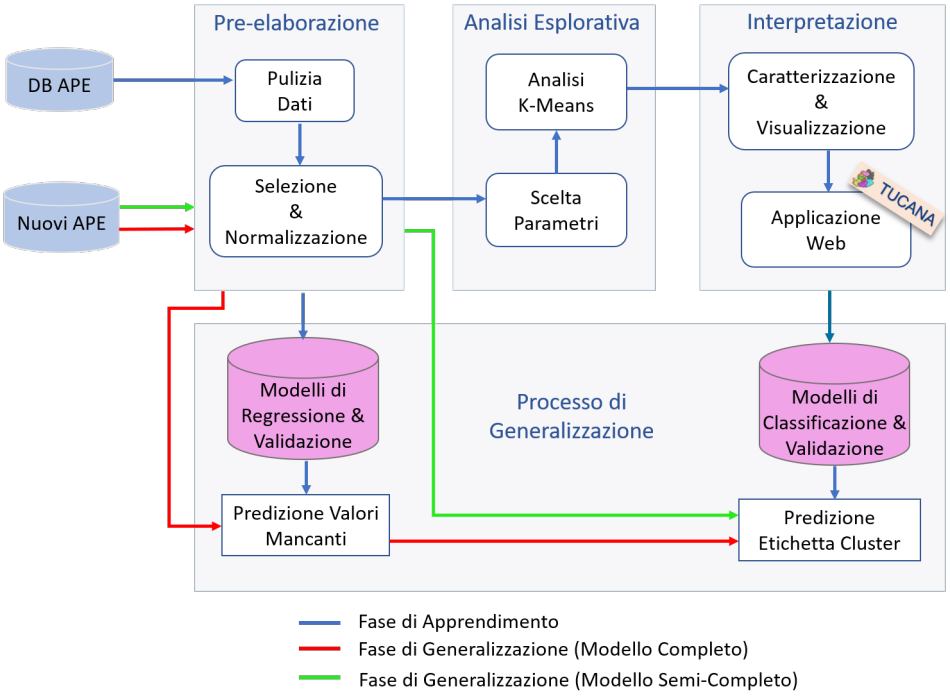

Figura 2.1: Architettura sviluppata per il *framework* TUCANA.

Il *framework*, mostrato in Figura 2.1, è composto da tre blocchi principali che richiamano le seguenti fasi del processo di *Knowledge Discovery in Databases* (KDD): *Pre-elaborazione e Trasformazione dei dati, Estrazione della Conoscenza, Interpretazione e Visualizzazione della conoscenza estratta* (Figura 2.2). Il quarto blocco del *framework* rappresenta il Processo di Generalizzazione, che caratterizza gli attestati di prestazione energetica che non fanno parte della base di dati.

Con il termine KDD si intende un processo iterativo ed interattivo che esprime la conoscenza come prodotto finale di un processo di identificazione di *pattern*[1] non banali, potenzialmente utili e comprensibili nei dati [20]. Dal 1989 ad oggi le fasi del KDD rappresentano l'estrazione della conoscenza ad alto livello a partire dai dati grezzi, al fine di ottenere informazioni spendibili per il *decision-making*[2].

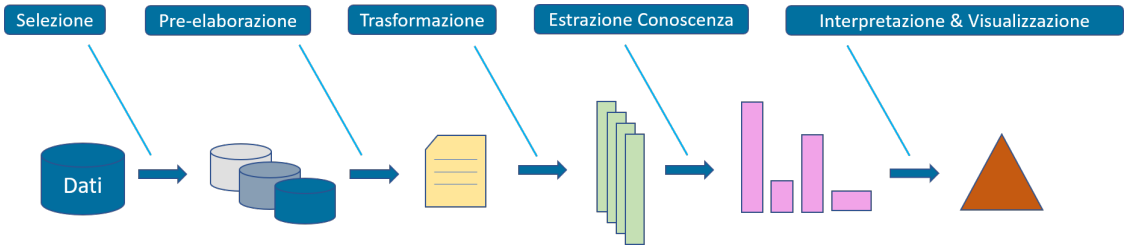

Figura 2.2: Le fasi del processo di *Knowledge Discovery in Databases* (KDD) [20].

Alcune volte il termine *data mining* viene utilizzato impropriamente come sinonimo del processo di estrazione della conoscenza presentato in Figura 2.2, tuttavia questa parola indica una particolare fase dell'intero processo di KDD. Nello specifico, il *data mining* individua la scelta di uno o più algoritmi opportuni per estrarre i *pattern* e soddisfare due obiettivi [20] principali esposti di seguito.

- **Descrizione**: ricercare e analizzare i pattern significativi al fine di presentare la conoscenza finale in modo facilmente interpretabile;
- **Predizione**: studiare e scoprire nuovi pattern con lo scopo di predire i comportamenti futuri degli attributi di interesse.

All'interno del *framework* vengono perseguiti entrambi gli obiettivi, il primo attraverso l'analisi esplorativa dei dati presentata nella sezione 2.2, mentre il secondo tramite il processo di generalizzazione esposto nella sezione 2.4.

[1]Un *pattern* è uno schema ricorrente che viene osservato e utilizzato per descrivere un comportamento specifico di un insieme di dati.

[2]Il processo di *decision-making* è il risultato di diversi processi mentali, cognitivi ed emozionali, che determinano la scelta di una determinata strategia dopo aver valutato le alternative proposte.

2.1 Pre-elaborazione dei Dati

La base di dati a disposizione contiene le due tipologie di certificazioni energetiche introdotte nel capitolo precedente. In particolare, sono presenti i dati relativi alle certificazioni ACE dal 2009 al 2014 e quelli relativi agli attestati APE dal 2015 al primo semestre del 2018, forniti dal CSI-Piemonte. I dati reali, come gli *open data* a disposizione, possono essere incompleti, inconsistenti, duplicati o rumorosi. Per questo motivo, il *framework* propone questa fase iniziale con l'obiettivo di garantire la qualità dei dati utilizzati nel processo di *data mining*, al fine di fornire risultati attendibili nelle fasi successive.

2.1.1 Pulizia dei Dati

A questo livello si adottano diverse strategie per la pulizia delle certificazioni energetiche raccolte. Grazie ai suggerimenti dell'esperto di dominio, viene effettuato uno studio univariato sugli attributi ritenuti fondamentali ai fini delle analisi. Nello specifico, si applicano dei filtri su alcune variabili termofisiche (e.g. trasmittanza media delle superfici opache) e sugli attributi legati all'efficienza del sistema di riscaldamento (e.g. rendimento di emissione). Una simile metodologia viene riconosciuta come *domain-driven*, tuttavia il *framework* dispone anche dei due passi descritti di seguito, i quali vengono indicati con il termine *data-driven*[3].

Pulizia degli Indirizzi

Uno degli obiettivi finali del *framework* sviluppato si basa sulla rappresentazione della conoscenza estratta tramite mappe geolocalizzate. Per questo motivo è stata implementata una procedura di pulizia dei dati geografici come l'indirizzo, il numero civico, il CAP e le coordinate geografiche (latitudine e longitudine).

Al fine di poter correggere in modo efficace e automatico errori di battitura e/o formato del campo di testo libero *indirizzo*, viene proposto l'utilizzo delle applicazioni offerte da Google Maps come le Geocoding API[4], le quali riescono a restituire sia l'indirizzo corretto che altre informazioni come il CAP e le coordinate geografiche. In principio, questo servizio permetteva di effettuare un numero limitato di richieste gratuite al server, tuttavia, essendo un servizio molto preciso ed affidabile, negli anni le sue prestazioni sono diventate a pagamento. In particolare,

[3]Aggettivo usato per indicare che il progresso in un'attività viene spinto dai dati, piuttosto che dall'intuizione o dall'esperienza personale.

[4]Consultabile su https://developers.google.com/maps/documentation/geocoding/intro.

fino ad un massimo di 100 000 richieste si pagano \$0.005/chiamata. Immaginando di rendere il *framework* utilizzabile anche per una considerevole mole di dati, si affianca alla metodologia precedente una soluzione alternativa, basata sul confronto degli indirizzi presenti nella base di dati con l'elenco delle vie che compongono lo stradario della città di Torino[5]. Per confrontare gli indirizzi presenti all'interno del *database* delle certificazioni con quelli presenti nello stradario si utilizza la *distanza di Levenshtein* [21], tramite la quale è possibile calcolare l'indice di similarità fra gli indirizzi. La distanza di Levenshtein esprime il numero minimo di modifiche che si ottengono fra inserimento, cancellazione o sostituzione di un carattere, necessarie a trasformare la prima stringa nella seconda, e si calcola in questo modo:

$$Lev_{dist} = 2\,(NS) + NI + NC\,,$$

dove NS indica il numero di sostituzioni necessarie a trasformare la prima stringa nella seconda, NI indica il numero di inserimenti di caratteri ed NC rappresenta il numero di cancellazioni di caratteri. A questo punto si passa al calcolo dell'indice di similarità che permette di capire quanto le stringhe degli indirizzi siano simili. Identificando con Len_{sum} la somma delle lunghezze delle due stringhe, l'indice di similarità si calcola così:

$$Sim_{index} = \frac{Len_{sum} - Lev_{dist}}{Len_{sum}}\,.$$

L'indice di similarità assume valori compresi nell'intervallo $[0, 1]$, per cui 1 indica una totale similarità tra due stringhe (sono cioè uguali) mentre 0 indica la completa dissimilarità fra le stringhe. Per poter valutare la somiglianza fra gli indirizzi, si sceglie una soglia tra 0 e 1 oltre la quale l'indirizzo presente nel *database* viene considerato sufficientemente simile all'indirizzo trovato nel viario, e viene sostituito con le informazioni presenti in esso come la via, il numero civico, il CAP e le coordinate geografiche. Soltanto quando l'indice di similarità fra gli indirizzi risulta essere inferiore alla soglia definita, si utilizza il servizio di Geocoding API messo a disposizione da Google Maps. Questa metodologia di sanitizzazione degli attributi geografici permette di restringere l'uso dell'applicazione offerta da Google e, quindi, di richiedere la risoluzione a pagamento di un indirizzo solo quando si ritiene effettivamente necessario. Qualora l'indirizzo presente nel *database* non riesca a risolversi con uno dei due metodi presentati, il corrispondente attestato di prestazione energetica viene eliminato dalla base di dati [22].

[5]Stradario delle vie consultabile nel sito http://geoportale.comune.torino.it/web/.

Eliminazione degli Outlier

Gli *open data* a disposizione provengono dal mondo reale perciò possono essere caratterizzati dalla presenza di *outlier*, cioè valori estremi e lontani dalle altre osservazioni [23]. Diversi sono i motivi che possono generare gli *outlier*, ad esempio un errato inserimento del dato o una codifica scorretta in fase di caricamento. Il *framework* sviluppato utilizza sia le tecniche univariate (gESD, Rimozione tramite Percentili) che quelle multivariate (DBSCAN) per la rilevazione e la rimozione automatica degli *outlier*, si riportano di seguito le descrizioni relative a ciascuna metodologia impiegata.

- **gESD (generalized Extreme Studentized Deviate)**: metodo parametrico usato per identificare valori anomali dei dati che seguono approssimativamente una distribuzione normale [24]. Per ottenere dei buoni risultati, occorre settare un numero massimo di *outlier* a priori e confermare che la distribuzione dei dati sia effettivamente gaussiana. Sia r il limite superiore scelto, allora il gESD esegue r test statistici separati definiti come:
  - H_0: non sono presenti *outlier* nei dati
  - H_1: sono presenti fino a r *outlier* nei dati

$$R_i = \frac{\max_j |x_j - \bar{x}|}{s}, \quad i = 1, \dots, r \quad \text{dove:} \quad \bar{x} = \frac{\sum_{j=1}^{n_i} x_j}{n} \quad \text{e} \quad s^2 = \frac{\sum_{j=1}^{n_i} (x_j - \bar{x})^2}{n_i - 1}$$

  Si calcolano le r statistiche del test rimuovendo ad ogni passo l'osservazione che massimizza la quantità $|x_j - \bar{x}|$, perciò R_1 viene calcolata con $n_1 = n$ osservazioni, R_2 a sua volta con $n_2 = n - 1$ osservazioni e si procede così fino a R_r che viene calcolata con $n_r = n - r + 1$ osservazioni [25]. Per determinare i valori critici del test si specificano sia il livello di significatività α che i parametri β, $\lambda(\beta)$ tali per cui:

$$\mathbb{P}\left[R_i > \lambda_i(\beta) \,|\, H_0\right] = \beta, \quad i = 1, \dots, r \qquad \text{e} \qquad \mathbb{P}\left\{ \bigcup_{i=1}^{r} \left[R_i > \lambda_i(\beta) \,|\, H_0\right] \right\} = \alpha$$

  Se tutte le statistiche soddisfano $R_i \leq \lambda_i(\beta)$ allora non sono presenti degli *outlier* invece, se esiste una qualche statistica tale per cui $R_i > \lambda_i(\beta)$, il numero di *outlier* viene determinato dal più grande valore di i in grado di rispettare questa maggiorazione.

- **Rimozione tramite Percentili**: metodologia basata sull'analisi della distribuzione dei dati con l'intento di identificare e rimuovere gli *outlier* che si presentano nel primo percentile [26]. Si consideri un insieme ordinato di osservazioni, si definisce il percentile p quella misura tra 0 e 100 sotto la quale si verifica una certa percentuale dei dati. Ad esempio, il decimo percentile indica il valore sotto il quale si trovano il 10% delle osservazioni [27].

- **DBSCAN (Density-Based Spatial Clustering of Applications with Noise)**: algoritmo in grado di individuare all'interno dei dati zone ad alta e bassa densità, queste ultime vengono riconosciute come le aree contenenti rumore cioè quelle in cui sono presenti gli *outlier* [28]. Per poter di definire il concetto di densità il DBSCAN richiede due parametri: *minPts* ed *Eps*. Un'area si dice densa se all'interno del suo raggio *Eps* esiste un numero di punti maggiore o uguale a *minPts*. Dato un punto p di definisce:

$$N_{eps}(p) = \{q \in D : dist(p,q) \leq Eps\}\,, \quad \text{come il vicinato di } p\,.$$

  L'algoritmo riesce a suddividere i punti presenti nella base di dati in gruppi (*cluster*) distinti. All'interno di ogni gruppo esistono i *core point* e i *border point* che restano sul bordo, si veda un esempio in Figura 2.3.

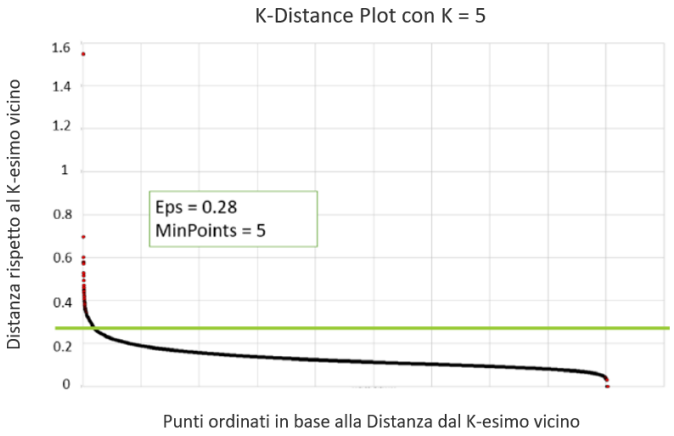

Figura 2.3: Esempio DBSCAN: l'algoritmo colora di verde i *core point*, di blu i *border point* e di rosso gli *outlier* ©Tan, Steinbach, Kumar, Introduction to Data Mining, McGraw Hill 2006.

In letteratura esistono dei metodi che aiutano a settare i parametri di input richiesti dal DBSCAN, in questo caso il *framework* utilizza una rappresentazione grafica chiamata *K-Distance Plot* per la scelta dei parametri *minPts* ed *Eps*. Si generano diverse curve del tipo *K-Dist*, si veda un esempio in Figu-

ra 2.4, dove si plottano le distanze tra le varie coppie di punti considerando differenti valori di K (parametro corrispondente al *minPts*). Dopodiché si prende la curva risultante stabile tra due diversi valori di K e si sceglie il valore di *Eps* in corrispondenza del "gomito". L'idea di base è che i punti all'interno di uno stesso gruppo siano vicini tra loro mentre gli *outlier* siano punti lontani e, perciò, che si trovano al di fuori del raggio*Eps* scelto nella curva *K-Distance Plot*. I parametri vengono determinati dal *framework* generando il *K-Distance Plot* ed estraendo in modo automatico i valori considerati migliori per l'identificazione degli *outlier*.

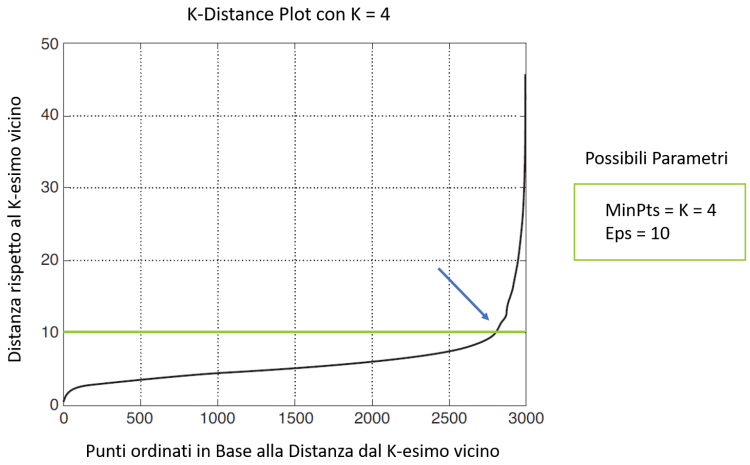

Figura 2.4: Esempio *K-Distance Plot*: setting dei parametri per DBSCAN ©Tan, Steinbach, Kumar, Introduction to Data Mining, McGraw Hill 2006.

2.1.2 Selezione e Normalizzazione dei Dati

In molti casi di studio reali, non tutte le variabili contenute all'interno della base di dati risultano essere rilevanti ai fini delle analisi. Per questo motivo, il *framework* sviluppato ricorre a diverse tecniche che possano aiutare a ridurre il numero degli attributi coinvolti. In questo modo, si possono velocizzare le procedure di *data mining* e ottenere risultati più significativi anche nelle altre fasi del *framework*.

Analisi della Correlazione

Il *database* a disposizione è caratterizzato da un numero elevato di attributi quindi, per ridurre la dimensionalità dei dati, il *framework* effettua la correlazione tra le variabili presenti in modo da conoscere quanto una variabile implichi l'altra e, in seguito, procede con la scelta di quelle da utilizzare nella fase di analisi

esplorativa dei dati. Dal punto di vista teorico, prese due variabili X, Y di tipo numerico, la correlazione viene calcolata mediante il coefficiente di correlazione lineare [29], conosciuto anche come *coefficiente di Pearson* e definito così:

$$\rho_{X,Y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} = \frac{Cov\,(X,Y)}{\sqrt{Var(X)\,Var(Y)}}\,,$$

dove $\bar{x}$ e $\bar{y}$ indicano rispettivamente le medie delle variabili X e Y. Il coefficiente di correlazione lineare è definito di modo che $\rho_{X,Y} \in [-1, 1]$. In particolare, se X e Y sono fortemente correlate negativamente oppure positivamente allora $\rho_{X,Y} = -1$ oppure $\rho_{X,Y} = 1$ invece, quando X e Y sono scorrelate allora $\rho_{X,Y} = 0$. Capire quali attributi sono altamente correlati o meno permette di includere nelle analisi successive solo quelli rilevanti. La correlazione fra gli attributi può essere visualizzata tramite una *matrice di correlazione* generata dal *framework*.

Normalizzazione delle Variabili di Analisi

In questa fase di *data-driven* vengono normalizzati gli attributi rilevanti selezionati precedentemente con la matrice di correlazione. Come nella sezione 2.1.1 dedicata alla pulizia del dato, il *framework* tiene conto di alcune scelte intraprese dall'esperto di dominio proponendo così l'approccio *domain-driven*. La normalizzazione delle variabili è utile affinché l'unità di misura dei diversi attributi non incida sull'analisi [30], inoltre si tratta di un passaggio necessario quando si utilizzano algoritmi di *clustering*, come in questo caso. Durante lo sviluppo di *framework* sono state prese in considerazione le seguenti tipologie di normalizzazioni.

- **Max-Min**: caratterizzata dall'effettuare una trasformazione lineare sui dati. Infatti, sia A una certa variabile di analisi e min_A, max_A i rispettivi minimi e massimi valori che vengono assunti da A. Allora, questa normalizzazione trasforma un valore v_i di A in v_i' calcolando:

  $$v_i' = \frac{v_i - min_A}{max_A - min_A}(max_{A,\,new} - min_{A,\,new}) + min_{A,\,new}\,.$$

  Tale trasformazione fa in modo che il valore v_i' possa variare all'interno del nuovo range definito come $[min_{A,\,new}, max_{A,\,new}]$. Questa tipologia di normalizzazione viene adottata dal *framework* perché tende a mantenere la relazione presente fra i dati originali e di facilitare la lettura dei risultati.

- **Z-Score**: in questo caso la normalizzazione di un attributo A considera la media e la deviazione standard. Un qualsiasi valore v_i dell'attributo A viene normalizzato in v_i' mediante il seguente calcolo:

$$v_i' = \frac{v_i - \bar{A}}{\sigma_A},$$

dove $\bar{A}$ indica la media dell'attributo e σ_A la sua deviazione standard. Questo tipo di normalizzazione può essere utilizzato quando non si conoscono i valori di massimo e minimo di un attributo di analisi. Rispetto alla metodologia precedente, la *Z-Score* risulta essere meno sensibile alla presenza di possibili *outlier* nella base di dati.

2.2 Analisi Esplorativa dei Dati

Dopo aver concluso la fase di pre-elaborazione dei dati, il *framework* procede con la fase dedicata all'applicazione di tecniche di *data mining* che consentono di estrarre conoscenze utili dai dati delle certificazioni energetiche. Il *framework* proposto applica la tecnica di *clustering* [31] che ha lo scopo di individuare sottogruppi omogenei di dati (*cluster*) con membri aventi caratteristiche simili, e membri appartenenti a gruppi diversi con caratteristiche dissimili. L'algoritmo di *clustering* utilizza il concetto di similarità che, in molti casi, equivale ad una misura di distanza dei dati nello spazio multidimensionale.

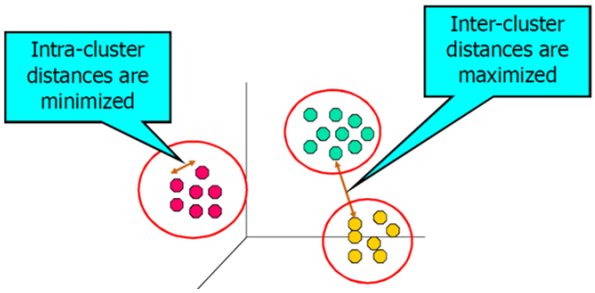

Figura 2.5: Esempio di suddivisione in *cluster* ©Tan, Steinbach, Kumar, Introduction to Data Mining, McGraw Hill 2006.

Inoltre, come mostrato in Figura 2.5, si vuole minimizzare la distanza *intra-cluster*, cioè quella fra i dati appartenenti al gruppo, e massimizzare la distanza

inter-cluster, cioè quella tra i gruppi. La procedura di *clustering* si basa sul concetto di *unsupervised learning* [29], questo termine indica che non si conoscono a priori le informazioni che verranno apprese dai dati e, perciò, si tende a parlare di output e apprendimento non supervisionati.

In generale gli algoritmi di *clustering* suddividono tutti i punti del *dataset* in gruppi ad intersezione vuota. Tuttavia esistono anche delle tecniche, note come *soft clustering* o *fuzzy clustering*, in cui uno stesso punto può appartenere a gruppi diversi in percentuali variabili. Inoltre, è possibile trovare in letteratura alcune categorie distinte di *clustering*, descritte brevemente di seguito [32]:

- **Partizionale**: dato un insieme di dati con cardinalità n, i punti vengono suddivisi in K gruppi disgiunti generando una partizione del *dataset* originale. Ogni gruppo rappresenta un *cluster* e contiene un insieme di elementi minore o uguale a n. Fanno parte di questa categoria l'algoritmo *K-Means* e sue varianti, ad esempio il *K-Medoids* [33];
- **Gerarchico**: genera diversi *cluster* annidati rappresentabili tramite una struttura ad albero chiamata dendrogramma, in grado di evidenziare la sequenza di aggregazioni effettuata;
- **Basato su densità**: categoria in cui i *cluster* si formano aumentando la loro densità fino a quando il numero di punti all'interno non supera una certa soglia. Rientra in questa tipologia di *clustering* l'algoritmo DBSCAN, accennato nella sezione 2.1.2;

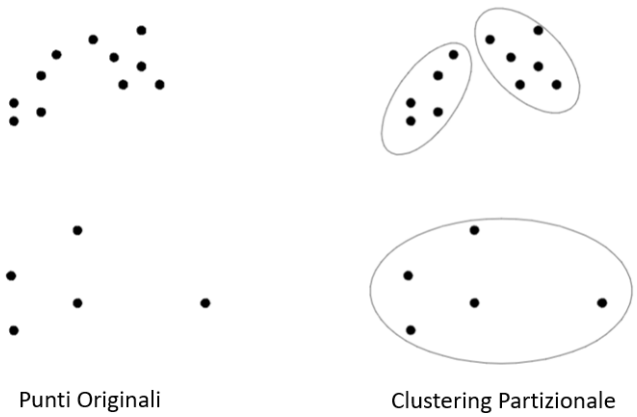

Figura 2.6: Esempio di *clustering* partizionale ©Tan, Steinbach, Kumar, Introduction to Data Mining, McGraw Hill 2006.

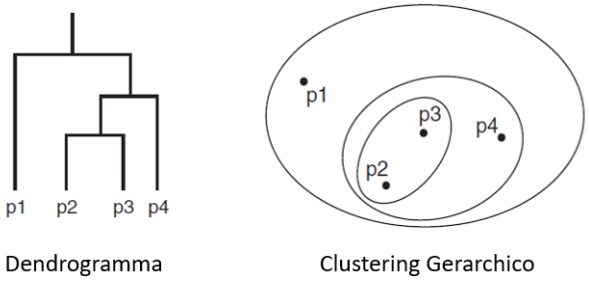

Figura 2.7: Esempio di *clustering* gerarchico con il relativo dendrogramma ©Tan, Steinbach, Kumar, Introduction to Data Mining, McGraw Hill 2006.

Il *framework* propone per l'analisi esplorativa dei dati l'algoritmo *K-Means* discusso nel dettaglio nel paragrafo successivo. L'architettura sviluppata mostrata in Figura 2.1 evidenzia la scelta dei parametri di input da fornire al *K-Means* e poi il suo utilizzo. Tuttavia, ai fini teorici di presentazione di questo lavoro, inizialmente verranno esposti i concetti fondamentali alla base dell'algoritmo e, successivamente, le tecniche adottate per la scelta dei parametri di input.

2.2.1 Algoritmo K-Means

L'idea di raggruppare edifici con proprietà termofisiche e geometriche simili consente sia l'esplorazione di informazioni non banali presenti all'interno dei certificati energetici, sia una loro possibile interpretazione dal punto di vista energetico. In questa sezione viene quindi descritta la metodologia utilizzata per suddividere il *dataset* degli attestati di prestazione energetica a disposizione.

Nell'algoritmo *K-Means* il parametro K indica il numero di *cluster* che compongono la partizione del *dataset*. Inoltre, si tratta di un algoritmo *centroid-based*, nel senso che utilizza il punto centrale del *cluster* per realizzare il partizionamento dei gruppi. In letteratura il punto centrale viene chiamato *centroide*[6] ed è calcolato come la media dei punti al suo interno. L'algoritmo si sviluppa scegliendo K punti casuali, i centroidi iniziali dei *cluster*, dopodiché i punti restanti vengono assegnati al gruppo il cui centro risulta essere più vicino. A questo punto si ricalcolano i centroidi dei gruppi ottenuti e si rifanno le assegnazioni considerando tutti i punti del *database*. L'algoritmo si conclude quando si hanno punti assegnati allo stesso *cluster* per due o più iterazioni successive.

[6]Per definizione il *centroide* non è mai un punto del *dataset*, ma esistono altri algoritmi in cui il punto centrale è sempre un punto del *database* e si chiama *medoide*, e.g. nel *K-Medoids*.

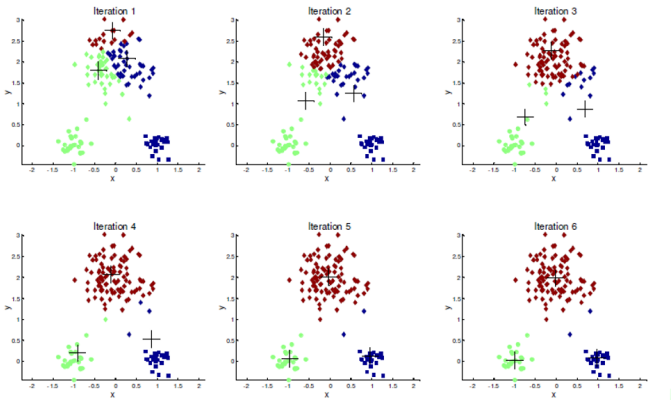

Figura 2.8: Esempio 1 - Effetto di una buona scelta random dei centroidi ©Tan, Steinbach, Kumar, Introduction to Data Mining, McGraw Hill 2006.

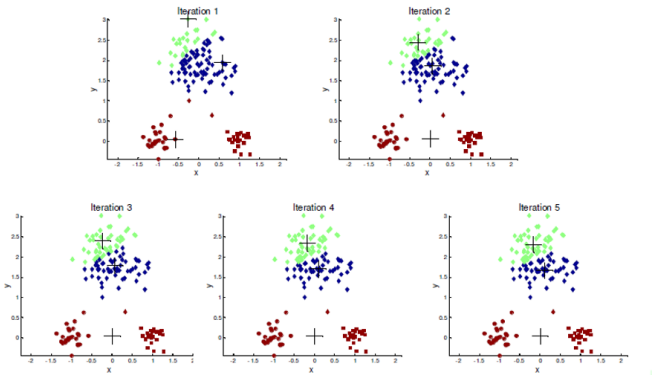

Figura 2.9: Esempio 2 - Effetto di una scelta random inadeguata dei centroidi ©Tan, Steinbach, Kumar, Introduction to Data Mining, McGraw Hill 2006.

Nell'algoritmo *K-Means* il concetto di vicinanza fra i punti e i centroidi si esprime mediante il calcolo della *distanza euclidea* [34]. Questo tipo di distanza è fra le più diffuse e viene spesso utilizzata per confrontare dati di tipo numerico. Siano $p = (x_{p1}, \ldots, x_{pm})$ e $q = (x_{q1}, \ldots, x_{qm})$ due punti multidimensionali descritti da m attributi numerici, allora la distanza euclidea fra questi oggetti si calcola in questo modo:

$$d(p, q) = \sqrt{(x_{p1} - x_{q1})^2 + (x_{p2} - x_{q2})^2 + \cdots + (x_{pm} - x_{qm})^2}\,.$$

Osservazione 2.1. *Poiché gli algoritmi di clustering, come il K-Means, utilizzano il concetto di distanza, la fase di normalizzazione dei dati presentata nella sezione 2.1.2 risulta fondamentale per ottenere dei buoni risultati.*

I primi centroidi dell'algoritmo sono scelti in modo random. Per cui i *cluster* variano ad ogni iterazione e si hanno risultati diversi fino alla stabilizzazione dei centroidi. La scelta iniziale dei centroidi è molto importante e può determinare la bontà dei risultati, si faccia ad esempio il confronto tra la Figura 2.8 e la Figura 2.9. L'algoritmo appena descritto può essere scritto in modo sintetico:

Algoritmo 1 K-Means ©Tan, Steinbach, Kumar

1: Seleziona K punti random come centroidi iniziali
2: **ripeti**
3: Crea K gruppi assegnando ogni punto al centroide più vicino
4: Ricalcola il centroide di ogni gruppo
5: **finché** Centroidi non cambiano

Gli algoritmi di *clustering* sono tipicamente espressi tramite una funzione obiettivo che deve minimizzare la distanza *intra-cluster* e massimizzare la distanza *inter-cluster*, come evidenziato nella Figura 2.5. In questo caso, il *K-Means* è caratterizzato da una funzione obiettivo che deve ridurre al minimo il quadrato della distanza di ciascun punto dal suo centroide più vicino [31]. Quindi, la bontà dei gruppi $C_1, \ldots, C_K$ ottenuti con questo algoritmo, può essere misurata tramite SSE (*Sum of Squared Error*), ovvero la somma degli errori quadratici:

$$\text{SSE} = \sum_{k=1}^{K} \sum_{p \in C_k} dist(p, c_k)^2 \,,$$

dove il termine $dist(p, c_k)$ indica la distanza euclidea tra un punto p appartenente a C_K e il suo centroide c_k. Date due diverse partizioni generate da due successive iterazioni del *K-Means*, viene presa in considerazione la partizione che produce il più piccolo SSE, in questo modo vengono scelti i *cluster* che meglio rappresentano la distribuzione dei punti al loro interno. Questa procedura garantisce l'esistenza di un minimo locale ma non di un minimo globale perché vengono fatte delle scelte specifiche per i centroidi e, di conseguenza, per i *cluster*. Dal punto di vista computazionale, minimizzare l'SSE guardando tutte le possibili scelte per i centroidi determina un problema di tipo *NP-hard* con complessità $\mathcal{O}(n^{dK+1} \log n)$, dove d è la dimensione dello spazio, n il numero di punti presenti e K il numero di *cluster*. L'algoritmo *K-Means* risulta molto efficiente quando processato in una grande collezione di dati, infatti la sua complessità è dell'ordine di $\mathcal{O}(nKt)$, dove t indica il numero di iterazioni. Solitamente si ha t tale che $t << n$.

Osservazione 2.2. *L'algoritmo K-Means ha la predisposizione a trovare gruppi con cardinalità simili e di forma globulare, quindi presenta una scarsa capacità di suddividere efficacemente gruppi di forma non globulare e dimensioni variabili. Inoltre, poiché il centroide indica la media dei punti appartenenti ad un gruppo, il K-Means è influenzato dalla presenza di rumori o outlier, motivo per cui risulta fondamentale la fase di pulizia spiegata nella sezione 2.1.1.*

2.2.2 Scelta dei Parametri

Oltre alla definizione e la normalizzazione degli attributi di interesse, presentate nella sezione 2.1.2, per poter applicare l'algoritmo *K-Means* è necessario specificare il parametro K, ovvero il numero di gruppi omogenei desiderati. Esistono diverse tecniche che consentono di stimare il migliore K, tra cui il cosiddetto *elbow method* o metodo del "gomito" [35], il quale considera la percentuale di variabilità spiegata, ovvero i diversi valori di SSE, in funzione del numero di *cluster* prodotti, si consideri ad esempio la Figura 2.10 seguente.

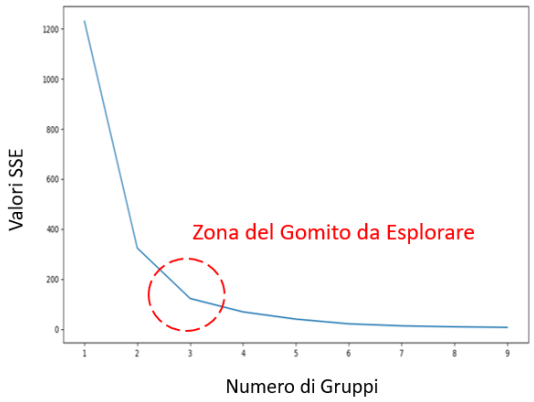

Figura 2.10: Esempio di curva ottenuta con l'*elbow method*.

Da questa immagine si può osservare che i primi *cluster* forniscono molta informazione perché c'è un'alta variabilità ma, aumentando il valore del parametro K, il guadagno di informazioni comincia a decrescere significativamente generando un gomito nella curva tra $K = 3$ e $K = 4$. L'idea alla base di questo metodo vede l'analisi dei risultati ottenuti con il *K-Means* per tutti quei valori di K che si trovano nella zona del gomito. Per cercare di rendere automatica la scelta del K migliore fra quelli presenti nel gomito, si utilizza la metodologia riportata in Figura 2.11. In

pratica, si calcolano le proiezioni ortogonali dei punti che appartengono alla curva sulla retta che congiunge il punto iniziale e quello finale, e si sceglie il parametro K nell'asse delle ascisse corrispondente al punto p che ha distanza massima dalla retta che congiunge il punto iniziale e finale della curva [36].

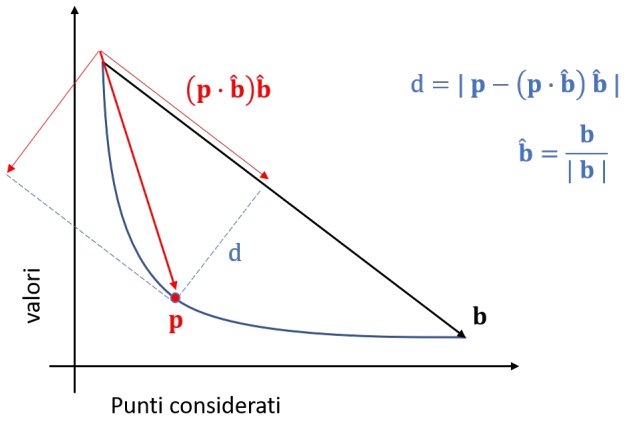

Figura 2.11: Metodologia per la scelta del parametro K nell'utilizzo del *K-Means*.

2.3 Interpretazione della Conoscenza Estratta

I modelli ottenuti dall'applicazione delle tecniche di *data mining* richiedono un processo di approfondimento, studio e verifica delle informazioni acquisite. Per questo motivo, in questa fase del *framework* vengono descritte le metodologie utilizzate per caratterizzare e visualizzare la conoscenza estratta dal modello di *clustering* definito nella fase precedente. Inoltre, vengono esposte dal punto di vista teorico le fasi di progettazione e sviluppo dell'applicazione web TUCANA, creata con lo scopo di rendere facilmente fruibili le informazioni apprese grazie alla caratterizzazione della conoscenza.

2.3.1 Caratterizzazione della Conoscenza

In questa parte del lavoro, vengono interpretate mediante alcuni strumenti di conoscenza, le informazioni restituite dall'algoritmo di *clustering*. Si analizzano quindi i risultati e i *pattern* estratti seguendo siano le specifiche e il giudizio dell'esperto di dominio, siano le tecniche statistiche e alcuni indici di valutazione. Questo passo è anche fondamentale per poter indirizzare l'ultima fase del *framework*, quella relativa alla generalizzazione della conoscenza, come mostrato in Figura 2.1.

Metodo Grafico: Boxplot

Con il termine *boxplot* [36], o *diagramma a scatola e baffi*, si intende uno strumento grafico che permette di visualizzare la distribuzione di un campione di dati attraverso i suoi quartili. Si tratta di un rettangolo rappresentabile sia verticalmente che orizzontalmente in cui vengono indicati il primo quartile (Q_1), la mediana (Q_2) e il terzo quartile (Q_3) della distribuzione di dati. Come mostrato nella Figura 2.12, il *boxplot* riporta anche altri importanti valori come i massimi e minimi definiti rispettivamente come $Q_3 + 1.5 * IQR$ e $Q_1 - 1.5 * IQR$, dove IQR indica la dimensione del rettangolo, ovvero il range interquartile definito come $Q_3 - Q_1$. Vengono invece classificati come *outlier* quei valori che si trovano al di sopra o al di sotto dei valori di massimo e minimo appena citati. In questo modo, si riesce a capire come caratterizzare i diversi gruppi ed identificare le eventuali differenze fra un *cluster* e l'altro.

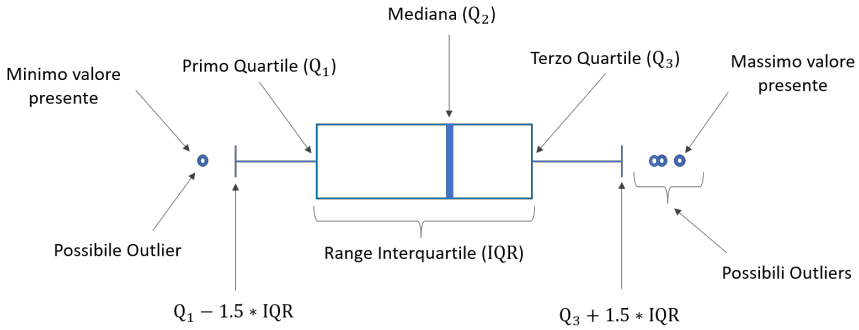

Figura 2.12: Esempio di boxplot con indicazioni statistiche.

Rispetto ad altri elementi grafici come l'*istogramma*, il *diagramma a scotola e baffi* può essere utile per valutare rapidamente determinati macro comportamenti di un insieme di dati. In genere fornisce meno informazioni rispetto alla distribuzione cumulata del campione, ma diventa molto utile quando si vogliono confrontare sottogruppi di dati. Nel caso specifico, il *framework* utilizza questo metodo grafico per valutare l'andamento delle variabili di analisi all'interno dei diversi *cluster*, generati durante la fase esplorativa dei dati.

Albero di Decisione: CART

Un albero di decisione [38] evidenzia un processo decisionale multifase assimilabile ad un *flow chart*, dove ogni nodo interno denota una regola applicata su un certo attributo del *dataset*, e ogni ramo il risultato del test effettuato su quel

nodo. Esistono sia gli alberi di regressione che quelli di classificazione, i primi vengono usati per predire variabili di tipo categorico mentre i secondi vengono impiegati nella predizione di variabili quantitative. Come struttura ad albero il *framework* utilizza il CART (*Classification and Regression Tree*) [39] per generare alberi di classificazione, si veda un esempio in Figura 2.13. L'obiettivo è quello di partizionare i dati in un certo numero di classi e confrontare il risultato con l'output fornito dal *clustering*.

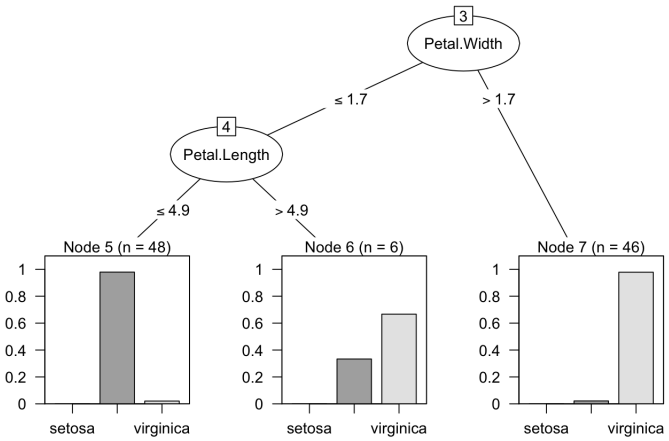

Figura 2.13: Esempio di CART con tre nodi foglia e tre classi (setosa, virginica e versicolor) sul *dataset* Iris.[7]

L'albero è caratterizzato dai suoi nodi foglia che riportano l'etichetta di classe e dal nodo radice, ovvero il primo nodo dell'albero. Il CART si basa sulla costruzione ricorsiva dell'albero utilizzando un approccio *top-down*, perché inizia nella parte superiore dell'albero e *greedy*, ciò significa che ad ogni *split* viene scelta la migliore suddivisione dell'attributo in quel particolare passo. Il CART utilizza come criterio di *splitting* l'*indice di Gini* [29]. L'idea di base è quella di ridurre la varianza generata durante la creazione dell'albero.

Al fine di soddisfare tale richiesta si calcola una misura della varianza totale presente fra le K classi tramite la seguente formula:

$$G = \sum_{k=1}^{K} \hat{p}_{mk}(1 - \hat{p}_{mk}) \,,$$

[7]Il *dataset* Iris sono 150 istanze di Iris misurate da Edgar Anderson e classificate secondo tre specie: Iris setosa, Iris virginica e Iris versicolor.

dove $\hat{p}_{mk}$ indica la probabilità che le osservazioni di *training* nella m-esima regione provengano dalla k-esima classe. Le regioni sono le porzioni spaziali determinate dalle suddivisioni che si effettuano sugli attributi di interesse durante le fasi di *split*, mentre le classi definiscono la variabile categorica di risposta da visualizzare nei nodi foglia. Dalla formula presentata, si osserva che l'indice di Gini assume un basso valore quando tutti i $\hat{p}_{mk}$ sono vicini a zero oppure a uno e, per questo motivo, viene indicato come una misura della "purezza" dei nodi foglia. Infatti, quando la misura G è piccola allora il nodo contiene per la maggior parte osservazioni provenienti da una singola classe.

Metodo Grafico: Radar Chart

Una rappresentazione grafica molto utile in grado di mostrare i dati rispetto a diverse variabili è il *radar chart*, o il *diagramma di Kiviat*, sempre più utilizzato in diversi ambiti dell'analisi dei dati [40] in quanto consente di valutare rapidamente attributi multipli su un grafico bidimensionale.

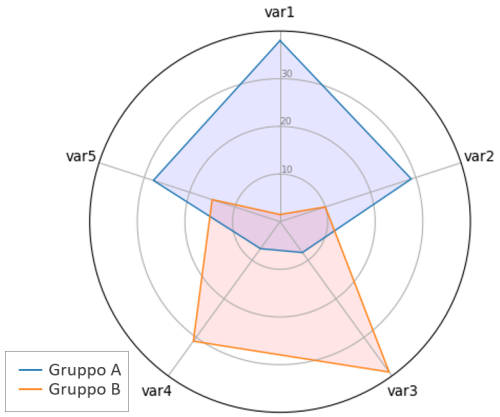

Figura 2.14: Esempio di radar chart per due gruppi diversi di variabili.[8]

Il *radar chart* è composto da una sequenza di raggi con origine da un centro comune, dove ogni raggio indica una variabile di analisi. Ogni cerchio concentrico determina dei valori particolari assunti dalle variabili rappresentate. Per questo motivo, tale metodo grafico risulta essere molto flessibile e, di conseguenza, molto utile per la caratterizzazione di alcuni gruppi di dati. Infatti, sul *radar chart* si possono raffigurare contemporaneamente sia le variabili di tipo categorico che quelle di tipo numerico.

[8]Grafico tratto da https://python-graph-gallery.com/390-basic-radar-chart/.

2.3.2 Visualizzazione della Conoscenza

Lo scopo di questa fase del *framework* è quello di visualizzare le informazioni estratte durante l'analisi esplorativa dei dati. Le rappresentazioni tipiche della statistica, come quelli presentati nella sezione precedente, sono comprensibili soprattutto dagli esperti del settore e, perciò, possono risultare di difficile interpretazione. Per questo motivo, sono state scelte per la visualizzazione della conoscenza delle mappe dinamiche e geolocalizzate descritte brevemente di seguito che, per definizione, sono di facile esposizione e conseguente traduzione sia da parte dei non esperti di dominio sia da parte di un pubblico generico.

Mappe Coropletiche

Le mappe coropletiche sono mappe tematiche suddivise per diverse aree geografiche e colorate secondo i risultati di calcoli statistici effettuati su di esse [41]. Queste mappe possono essere utilizzate ad esempio per evidenziare la densità di una popolazione o per visualizzare la distribuzione di una qualsiasi variabile di analisi. Si consideri ad esempio la mappa in Figura 2.15, si osserva come sia possibile rappresentare una grande collezione di dati in modo semplice ed intuitivo.

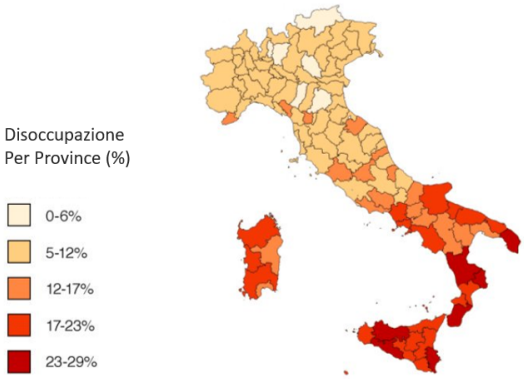

Figura 2.15: Esempio di mappa coropletica: tasso di disoccupazione in Italia nel 2017. Fonte Dati Istat, immagine esibita da BBC.

Questa tipologia di mappa viene proposta con lo scopo di approfondire la distribuzione di una delle variabili di interesse sulla mappa della città di Torino, suddivisa nelle sue rispettive circoscrizioni. Le mappe coropletiche generate dal *framework* si colorano in base al valore maggioritario dell'attributo di analisi considerato. Nel capitolo successivo verranno presentati alcuni risultati ottenuti con le variabili di analisi utilizzate durante l'analisi esplorativa dei dati.

Mappe Scatter

Le mappe geografiche di tipo *scatter* consentono di visualizzare i dati a disposizione come punti sulla mappa, ovvero come simboli geografici chiamati *marker*. Ad esempio, in Figura 2.16, i *marker* rossi sulla mappa rappresentano una specifica informazione in corrispondenza di quelle coordinate geografiche.

Figura 2.16: Esempio di mappa *scatter* con i simboli grafici detti *marker*.[9]

Questo tipo di mappe è stato utilizzato per visualizzare nel dettaglio le informazioni riportate nella base di dati dei certificati energetici. In particolare, poiché ogni attestato di prestazione energetica è identificato anche dagli attributi geografici di latitudine e longitudine dell'abitazione a cui si riferisce, essi possono essere visualizzati e, successivamente, caratterizzati tramite i *marker* sulla mappa di Torino. In questo caso, la visualizzazione del dato risulta essere molto dettagliata, tuttavia, la numerosità degli attestati presenti nel *database* può causare un problema di *override* informativo, si veda l'immagine a sinistra nella Figura 2.17.

Mappe Marker-Cluster

Un modo per gestire la presenza di un numero eccessivo di *marker* in una determinata area è quello di creare degli aggregatori di posizione. Un aggregatore viene visualizzato come un cerchio con al suo interno il numero di *marker* presenti in una certa area [42], si consideri l'immagine a destra in Figura 2.17. La visualizzazione fatta utilizzando gli aggregatori di *marker* consente di rappresentare il

[9]Immagine tratta da Google OR-Tools, riperibile nel sito https://developers.google.com/optimization/routing/tsp.

raggruppamento effettuato nella fase di *clustering*. Per questo motivo, la mappa viene indicata con il termine *marker-cluster* infatti, a seconda del livello di zoom impostato sulla mappa, l'informazione può risultare aggregata al livello dei *cluster* oppure al livello di dettaglio dei *marker* finali.

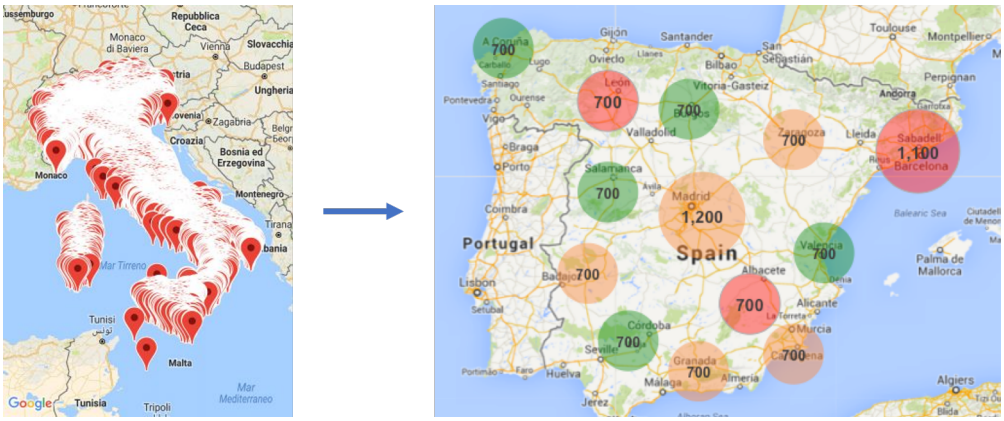

Figura 2.17: Confronto tra una mappa *scatter* e la mappa *marker-cluster*.

Come verrà presentato nel capitolo successivo, questo tipo di mappa trasmette rapidamente due tipi di informazioni relative al processo di analisi: il colore (legato alla performance energetica) dei *cluster* generati durante la fase di analisi esplorativa dei dati e la loro cardinalità espressa anche dalla dimensione dei cerchi generati sulla mappa. Un altro aspetto interessante della mappa *marker-cluster* riguarda la possibilità di rappresentare un'informazione multivariata in modo semplice ed intuitivo. Infatti, mentre una mappa coropletica esprime il valore assunto da un singolo attributo, in una mappa *marker-cluster* si visualizzano i risultati ottenuti da un'analisi fatta considerando più attributi del *database*.

2.3.3 Applicazione Web Sviluppata

Il *framework* propone una fase di sviluppo di un'applicazione web al fine di rendere fruibile le parti di caratterizzazione e visualizzazione dei risultati ottenuti dalle analisi. Diversi potrebbero essere gli *stakeholder* interessati a queste informazioni [22], ad esempio i privati cittadini che desiderano approfondire le caratteristiche energetiche di un dato edificio, la pubblica amministrazione che pretende di apportare alcuni miglioramenti in una determinata zona della città di Torino, oppure i fornitori di servizi energetici, potenzialmente intenzionati a individuare delle zone in cui gestire la propria erogazione energetica.

Una qualsiasi applicazione web utilizza la cosiddetta un'architettura *client-server*, si veda ad esempio la Figura 2.18, dove la porzione *client* interagisce con quella *server* effettuando alcune richieste di servizi. Il processo *client* si dedica maggiormente all'interazione con l'utente finale, mentre quello lato *server* è reattivo cioè svolge un'azione soltanto quando richiesta da un *client*, perciò è sempre pronto a ricevere le eventuali richieste [49].

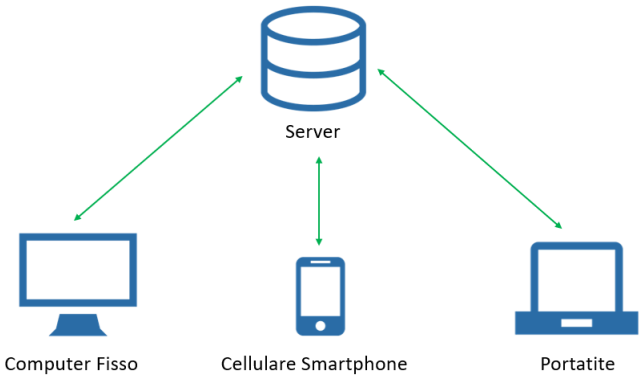

Figura 2.18: Semplice esempio di modello *client-server* ©TechTerms.com.

Il *client* può accedere alle funzionalità applicative giacenti sull'*application server* facendo eseguire i codici dal *web browser*. Questa parte con cui interagisce l'utente costituisce l'interfaccia dell'applicazione web, chiamata *front-end*. In un'applicazione web esiste anche la parte *back-end*, ovvero l'insieme di codici che vengono eseguiti lato *server* per rispondere alle richieste dell'utente. Il *back-end* include inoltre un database utilizzato per storicizzare in modo persistente i dati dell'applicazione web. In questo caso, non si hanno interventi che possano impattare sull'interfaccia grafica e, quindi, il *back-end* non viene visto dall'utente finale.

L'applicazione web proposta diventa quindi uno strumento tecnologico di distribuzione di informazioni e raccolta di possibili *feedback* da parte delle persone interessate. Nel caso specifico del *framework*, la parte *back-end* include sia la base di dati contenente gli attestati di prestazione energetica che tutti gli algoritmi utilizzati durante le fasi di analisi. Infatti, a seconda delle richieste effettuate dall'utente, l'applicazione web TUCANA fornisce i risultati desiderati. Per quanto riguarda la parte *front-end*, oltre alla visualizzazione dei risultati ottenuti dall'analisi esplorativa dei dati, vengono proposte alcune statistiche sulla base di dati, in modo da poter guidare l'utente finale all'acquisizione dei concetti legati agli attestati di prestazione energetica a disposizione.

2.4 Processo di Generalizzazione

All'inizio di questo capitolo, si è parlato di un altro obiettivo perseguibile grazie alle tecniche di *data mining*, quello di studiare nuovi *pattern* per predire i possibili comportamenti futuri delle variabili di interesse. Nel caso del *framework* specifico, l'intento è quello di considerare nuovi attestati energetici e caratterizzarli, utilizzando la conoscenza estratta dalle certificazioni energetiche che attualmente compongono il *database* [26]. Perciò, al fine di generalizzare il processo di conoscenza agli edifici non presenti nella base di dati, il *framework* propone due possibili modelli con percorsi evidenziati in Figura 2.1.

- **Modello completo**: gestisce la situazione in cui si predicono alcune fra le variabili di analisi mancanti e si assegna un'etichetta di *cluster*;
- **Modello semi-completo**: considera il caso in cui si assegna direttamente una delle possibili etichette di *cluster*.

Questi due modelli fanno riferimento alle procedure di *supervised learning* [29] perché, rispetto alla fase di analisi esplorativa dei dati, lo scopo ora è quello di prevedere una o più variabili di risposta note a priori, per questo motivo si parla di output supervisionato. In generale, la creazione di un modello di tipo *supervised* consiste nell'esprimere una variabile di risposta Y in funzione di p diversi attributi di interesse, chiamati anche *predittori* o *feature*, $X_1, \ldots, X_p$:

$$Y = f(X_1, \ldots, X_p) + \epsilon\,, \tag{2.1}$$

dove f è una funzione incognita tramite la quale i predittori spiegano la variabile di risposta Y, e il simbolo ϵ indica un termine di errore indipendente dalle variabili di input. Questo termine è necessario perché un modello è sempre una semplificazione della realtà, perciò il vero il legame fra i predittori e la variabile *target* può venire al più approssimato. All'interno delle metodologie di tipo *supervised* si distinguono due possibili casistiche, i problemi di *regressione* e di *classificazione* [43]. Nel primo caso l'output cercato è una variabile di tipo quantitativo, mentre nel secondo caso la variabile *target* è di tipo qualitativo o categorico. Una semplice immagine mostrata in Figura 2.19 evidenzia la differenza fra questi tipi di variabili. All'interno delle due metodologie di generalizzazione proposte dal *framework*, vengono applicati sia alcuni algoritmi di regressione che di classificazione infatti, come verrà mostrato nel capitolo successivo, gli attributi utilizzati per le analisi sono delle variabili numeriche, mentre l'etichetta di *cluster* è una variabile categorica.

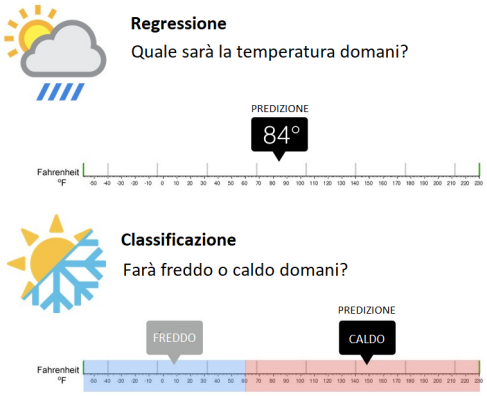

Figura 2.19: Esempio di predizione di una variabile quantitativa e qualitativa.[10]

Validazione dei Modelli Predittivi

Per valutare le performance di un modello predittivo si suddivide casualmente il *dataset* in due sottoinsiemi disgiunti. Si chiama *training set* [55] l'insieme delle osservazioni che costituiscono la fase di "addestramento" del modello, e *test set* l'insieme di punti della base di dati che non sono stati coinvolti nella creazione del modello e che servono per valutare l'errore di predizione commesso. Nei casi reali, tipicamente il 75% dei dati viene usato come *training set* e il restante 25% come *test set*, in questo caso si parla di *leave-one-out cross validation*.

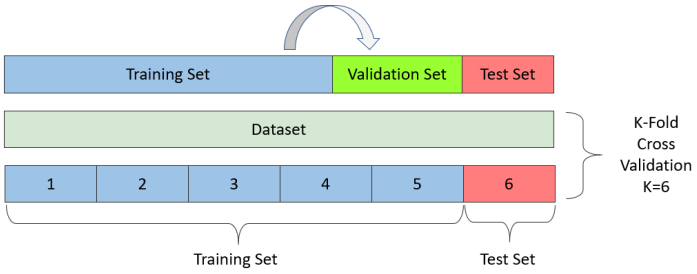

Figura 2.20: Esempio di suddivisione del *dataset* per la validazione dei modelli: *leave-one-out* in alto e una fase dell'algoritmo *K-fold* con $K = 6$ in basso.

Quando si hanno molti dati a disposizione, la migliore suddivisione è quella che considera anche un insieme di dati di validazione, detto *validation set*, utile per la valutazione ulteriore del modello di predizione scelto rispetto a tutti gli altri. Quando la numerosità del *dataset* non è sufficientemente grande, la suddivisione

[10]Immagine tratta da *Regression vs Classification*, Ali Reza Kohani 2017.

semplice in un insieme di *training* e un insieme di *test* può dare dei risultati fuorvianti per il calcolo dell'errore, perché magari lo si calcola su una partizione non significativa dei dati [32]. Per evitare questo problema si può utilizzare la procedura *K-fold cross validation* [56], la quale suddivide il *training set* in K sottoinsiemi disgiunti e ripete la procedura di validazione usando, ad ogni fase, un sottoinsieme diverso come *test set* e tutti gli altri come *training set*. Si ottengono in questo modo, diverse stime del tasso di errore il cui valore medio si chiama tasso di errore cross-validato. In Figura 2.20 vengono rappresentate le due metodologie di validazione per un *dataset*.

Nei paragrafi successivi, verranno spiegati nel dettaglio i concetti teorici fondamentali che sono stati utilizzati per l'implementazione dei modelli di predittivi di generalizzazione presenti all'interno del *framework* TUCANA. Inoltre, verranno presentate le misure con cui sono state effettuate le validazioni per i modelli predittivi di regressione e classificazione.

2.4.1 Modelli di Regressione

Con il termine regressione [44] si intende una tecnica di modellazione predittiva che studia la relazione tra una variabile numerica dipendente Y e una o più variabili indipendenti $X_1, \ldots, X_p$. Si consideri un *dataset* con n differenti punti, allora il termine x_{ij} rappresenta il valore del predittore X_j per l'i-esima osservazione di *training* dove $j = 1, \ldots, p$ e $i = 1, \ldots, n$. Inoltre, il termine y_i indica il valore della variabile *target* per l'osservazione i-esima. Esistono diversi tipi di regressione che si possono utilizzare su una serie di dati, tuttavia in questa sezione vengono descritti le principali metodologie utilizzate all'interno del *framework*.

Regressione Lineare

Il modello di regressione lineare [45] rimane uno dei più importanti strumenti di analisi predittiva. L'idea è quella di predire una risposta quantitativa Y sulla base di una o più caratteristiche. La regressione lineare semplice prevede la definizione di un'unica variabile indipendente X, mentre la regressione lineare multipla è l'estensione di quella semplice e considera un certo numero p di predittori quest'ultimo modello assume la seguente forma:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon\,. \tag{2.2}$$

In questa formula X_j indica il j-esimo predittore mentre β_j è un parametro che quantifica l'associazione tra questa variabile e la risposta Y. I parametri $\beta_0, \ldots, \beta_p$

sono incogniti e, perciò, devono essere stimati in modo da poter fare le predizioni mediante la formula seguente:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_p x_p \,, \tag{2.3}$$

dove $\hat{y}$ indica una previsione di Y sulla base dei valori $X_1 = x_1, \ldots, X_p = x_p$. Esistono diverse metodologie per stimare i parametri della funzione lineare espressa nella (2.3), la più popolare utilizza il metodo dei minimi quadrati [46], il quale seleziona i coefficienti $\beta_0, \ldots, \beta_p$ in modo da minimizzare la somma dei residui quadratici RSS (*Residual Sum of Squares*) definita come:

$$\text{RSS} \;=\; \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \;=\; \sum_{i=1}^{n}(y_i - \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \cdots + \hat{\beta}_p x_{ip}) \,. \tag{2.4}$$

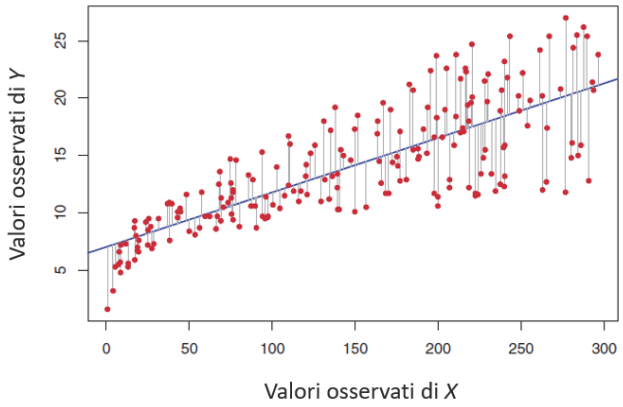

Figura 2.21: Metodo dei minimi quadrati su una serie di dati ©James, Witten, Hastiea, Tibshirani, An Introduction to Statistical Learning.

La differenza tra y_i e $\hat{y}_i$ indica il *residuo* i-esimo ed esprime l'errore di predizione quando si sostituisce il valore osservato di risposta y_i con il valore predetto $\hat{y}_i$ dal modello lineare. In Figura 2.21 viene mostrato un esempio di come vengono calcolati i residui nel caso di un modello lineare semplice, in pratica per ogni punto del *dataset* viene calcolata la sua distanza quadratica dalla retta definita da $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$. Perciò, la funzione lineare che meglio approssima l'insieme di punti, è quella tale per cui tutti i punti risultano avere distanza quadratica minima.

Regressione Polinomiale

Il modello di regressione lineare espresso nella (2.2) ipotizza una relazione di tipo lineare la variabile *target* e le variabili di input. Tuttavia, in alcuni casi, il vero legame tra la variabile Y e gli attributi $X_1, \ldots, X_p$ può essere non lineare. Allora, si può estendere la regressione lineare usando la cosiddetta regressione polinomiale, che utilizza l'interpolazione polinomiale [47] dei dati invece di quella lineare definita da una semplice retta. La regressione polinomiale si basa sull'aggiunta di ulteriori predittori, che si ottengono elevando a potenza le *feature* originali, in questo caso il modello espresso nella (2.2) assume la seguente forma:

$$Y = \beta_0 + \sum_{j=1}^{p} f_j(X_j) + \epsilon = \beta_0 + f_j(X_1) + f_j(X_2) + \cdots + f_j(X_p) + \epsilon\,. \qquad (2.5)$$

Perciò, il modo per consentire le relazioni non lineari tra ogni *feature* e la variabile di risposta, si basa sulla sostituzione di ogni componente lineare $\beta_j X_j$ con una funzione polinomiale $f_j(X_j)$ di grado d. Di solito non si considerano funzioni con $d > 4$ perché, un grado polinomiale elevato può generare una curva molto flessibile con forme strane.

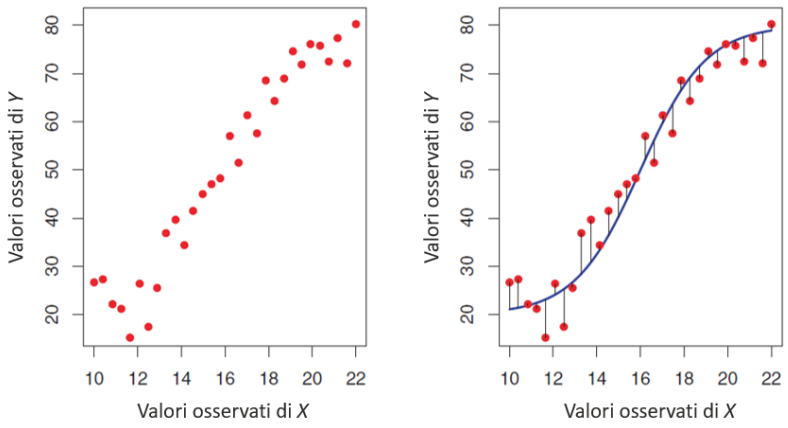

Figura 2.22: Esempio di funzione non lineare per la regressione polinomiale ©James, Witten, Hastiea, Tibshirani, An Introduction to Statistical Learning.

Questo è un esempio di modello GAM (*Generalized Additive Models*) [48], chiamato additivo perché si calcola una f_j diversa per ogni caratteristica X_j con $j = 1, \ldots, p$ e poi si aggiungono i loro contributi insieme. Un possibile limite associato alla regressione polinomiale è la sua definizione additiva infatti, se si hanno a disposizione molte variabili di input, si possono perdere alcune importanti

interazioni fra la variabile di risposta e i diversi predittori. Esistono anche altre estensioni della regressione polinomiale ottenute sostituendo le funzioni polinomiali con le funzioni spline [46].

Regressione Lasso

Il modello lineare definito dall'equazione (2.2) può essere migliorato sostituendo al metodo dei minimi quadrati alcune procedure alternative di "fitting" della funzione incognita f. La regressione *lasso* (*least absolute shrinkage and selection operator*) [50] è un metodo alternativo che considera tutti i predittori p e utilizza una tecnica che limita (o regolarizza) le stime dei coefficienti. In questo caso, i parametri del modello vengono scelti in modo tale da minimizzare una quantità simile all'RSS, la differenza è che ora viene aggiunto un nuovo termine:

$$\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \lambda\sum_{j=1}^{p}|\beta_j| \;=\; \mathrm{RSS} + \lambda\sum_{j=1}^{p}|\beta_j|\,. \tag{2.6}$$

La quantità che viene aggiunta all'espressione (2.4) è un termine di penalità dove λ è un parametro di *tuning* che va scelto durante le prove sperimentali. La regressione lasso stima i parametri $\beta_0, \ldots, \beta_p$ minimizzando l'RSS, tuttavia il termine di penalità diventa piccolo quando questi coefficienti sono vicini a zero. Quando il parametro λ è sufficientemente grande, alcune stime possono essere esattamente uguali a zero, ciò significa che il metodo lasso può essere usato anche per eseguire una selezione di variabili significative per la predizione. L'effetto dell'aggiunta di questa penalità può ridurre significativamente la varianza presente nella stima dei parametri del modello lineare. In letteratura, esiste anche un altro metodo molto simile a questo chiamato regressione *ridge* [51] che utilizza nel termine di penalità la norma l_1 al posto della norma l_2:

$$\lambda\sum_{j=1}^{p}|\beta_j| = \lambda\,\|\beta\|_1 \quad \text{(lasso)} \qquad \text{e} \qquad \lambda\sum_{j=1}^{p}\beta_j^2 = \lambda\,\|\beta\|_2^2 \quad \text{(ridge)}$$

Regressione K-NN

Le regressioni lineari e polinomiali accennate in precedenza hanno una semplice costruzione matematica e i coefficienti stimati sono facili da interpretare. Tuttavia, questi metodi hanno lo svantaggio di imporre delle forti ipotesi sulla forma della funzione $f(X_1, \ldots, X_p)$, se le ipotesi fatte non corrispondono alla realtà la predizione della variabile *target* sarà poco accurata. In letteratura, esistono altri approcci che

risolvono un problema di regressione senza dover dare delle indicazioni a priori sulla forma della funzione incognita f, come ad esempio il metodo K-NN (*K-Nearest Neighbors*) [52]. La regressione K-NN identifica per ogni nuova osservazione x_0 le K osservazioni del *dataset* più vicine ad essa (vicinato $\mathcal{N}_0$) e, successivamente, procede con la stima di $f(x_0)$ data la media dei valori della variabile di risposta in $\mathcal{N}_0$. In altre parole il valore predetto si ottiene così:

$$\hat{f}(x_0) = \frac{1}{K} \sum_{x_i \in \mathcal{N}_0} y_i \tag{2.7}$$

All'aumentare del numero di predittori p questo metodo restituisce delle performance peggiori rispetto al modello di regressione lineare. Ciò avviene perché, in elevate dimensioni, l'insieme di punti del *dataset* sono più isolati e, quindi, più distanti tra di loro. Questa situazione, conosciuta in letteratura come *curse of dimensionality* [53] rende difficile la definizione del vicinato $\mathcal{N}_0$.

Osservazione 2.3. *In letteratura esistono altri algoritmi che risolvono problemi di regressione. Per esempio, l'SVR (Support Vector Regression) [54], utilizzato in questo framework come termine di confronto fra quelli altri appena accennati.*

Misura di Validazione

La qualità di un modello predittivo di regressione può essere misurata utilizzando la statistica *R-quadro* [57]. Nel caso della regressione lineare per esempio, questa statistica indica quanto i dati sono vicini alla retta di regressione generata. Il calcolo della statistica R^2 si ottiene mediante la seguente formula:

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

dove con il termine TSS (*Total Sum of Squares*) si intende la varianza totale della risposta Y, ovvero la variabilità legata all'attributo *target* prima della creazione del modello di regressione. D'altra parte, il termine RSS misura la variabilità non spiegata che rimane dopo aver eseguito la regressione. Perciò, la variabilità di Y spiegata dal modello usando i diversi predittori determina il valore di R^2.

In generale, l'R^2 rappresenta una percentuale di varianza della variabile *target* e, quindi, assume valori tra 0 e 1 indipendentemente dai valori assunti dalla variabile di risposta Y. Un valore di R^2 prossimo a zero indica che il modello di regressione

non riesce a spiegare quasi nulla della variabilità della risposta Y attorno alla sua media. Invece un valore di R^2 prossimo a 1 significa che il modello di regressione spiega quasi tutta la variabilità della variabile *target* attorno alla propria media.

2.4.2 Modelli di Classificazione

Come introdotto all'inizio di questo paragrafo, con il termine classificazione [58] si intende una tecnica di modellazione predittiva che vuole predire il comportamento di una certa variabile categorica Y assegnandole un'etichetta, sulla base della conoscenza a priori di una o più attributi $X_1, \ldots, X_p$. La procedura di classificazione si basa sulla costruzione di un classificatore utilizzando un insieme di osservazione di *training* $(x_1, y_1), \ldots, (x_n, y_n)$. L'idea è che il classificatore sia in grado di assegnare un'etichetta di classe non solo a questi dati ma anche a nuovi dati che non sono stati usati durante la fase di apprendimento del modello. Esistono diversi metodi che permettono di risolvere un problema di classificazione su una serie di dati, tuttavia in questa sezione viene descritto l'algoritmo utilizzato all'interno del *framework* per l'implementazione dei modelli *completo* e *semi-completo*.

Classificatore K-NN

Il classificatore *K-Nearest Neighbors* [59] è un metodo semplice e intuitivo che funziona molto bene quando applicato a problemi di vario genere. Il K-NN utilizza la distanza euclidea per calcolare il vicinato $\mathcal{N}_0$, di una data osservazione di *test* x_0, e assegnarle un'etichetta di classe fra quelle presenti in $\mathcal{N}_0$.

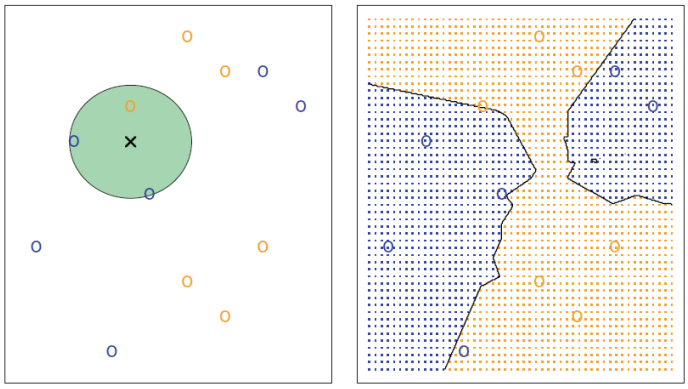

Figura 2.23: Esempio di output del K-NN, a sinistra la scelta dei $K = 3$ vicini, a destra il limite decisionale definito per $K = 3$ ©James, Witten, Hastiea, Tibshirani, An Introduction to Statistical Learning.

La classe che viene assegnata all'osservazione x_0 è quella più frequente all'interno del suo vicinato $\mathcal{N}_0$, costituito dalle K osservazioni di *training*. Nello specifico, il classificatore K-NN stima la probabilità condizionata che una certa classe j provenga dall'osservazione x_0 guardando la frazione di punti in $\mathcal{N}_0$ il cui valore di risposta è proprio dato dalla classe j:

$$\mathbb{P}\left(Y = j \,|\, X = x_0\right) = \frac{1}{K} \sum_{i \in \mathcal{N}_0} I(y_i = j)\,. \tag{2.8}$$

In particolare, il K-NN applica il teorema di Bayes [60], un concetto tale per cui l'etichetta di classe che viene assegnata a x_0 è quella che restituisce il valore più alto della probabilità condizionata della formula precedente. Uno degli aspetti positivi del K-NN è la sua facile implementazione e il fatto che non si devono fare delle assunzioni a priori sul *dataset* a disposizione.

Osservazione 2.4. *Poiché l'algoritmo K-NN utilizza la distanza euclidea per il calcolo della vicinanza fra le osservazioni, quindi è molto importante scegliere una normalizzazione opportuna per i dati a disposizione.*

Misure di Validazione

Prima di utilizzare il classificatore su un insieme di dati nuovi, bisogna validarlo per capire quanto sia effettivamente performante il modello costruito. Questo passaggio risulta essere molto importante anche per la scelta del parametro K di vicini da definire per poter implementare l'algoritmo. Dopo aver suddiviso il *dataset* nell'insieme di *training* e di *test*, si può precedere con il calcolo di alcune misure di validazione del modello.

	Predicted class		
Actual Class		Class = Yes	Class = No
	Class = Yes	True Positive	False Negative
	Class = No	False Positive	True Negative

Figura 2.24: Esempio di matrice di confusione.[11]

Per semplicità, si supponga di avere un classificatore binario ovvero un problema di classificazione in cui si hanno solo due possibili etichette di classi. Si definisce

[11]Immagini e informazioni reperibili al sito internet https://blog.exsilio.com/all/accuracy-precision-recall-f1-score-interpretation-of-performance-measures/.

matrice di confusione una tabella che viene utilizzata per descrivere le prestazioni del modello di classificazione su un insieme di dati di *test* per i quali sono note le vere etichette di classe, si veda un esempio in Figura 2.24 fra classi reali e classi predette. Data una certa classe di interesse, ad esempio $Class = Yes$, le osservazioni che appartengono a questa classe sono dette positivi (P) mentre tutte le altre osservazioni con $Class = No$ sono dette negativi (N). In questo modo si può confrontare la classe predetta dal classificatore con quella la classe vera dell'osservazione. Le quantità presentate in Figura 2.24 si definiscono così:

- Vero Positivo (TP): numero di osservazioni etichettate come P nella base di dati e classificate correttamente dal modello;
- Vero Negativo (TN): numero di osservazioni etichettate come N nella base di dati e classificate correttamente dal modello;
- Falso Positivo (FP): numero di osservazioni etichettate come N nella base di dati ed erroneamente classificate come P dal modello;
- Falso Negativo (FN): numero di osservazioni etichettate come P nella base di dati ed erroneamente classificate come N dal modello.

A partire da questi dati presenti nella matrice di confusione, calcolata in modo automatico, si possono effettuare le misure di prestazione [32] del modello di classificazione che sono le seguenti:

$$Accuratezza = \frac{TP + TN}{TP}, \quad Precisione = \frac{TP}{TP + FP}, \quad Richiamo = \frac{TP}{TP + FN}$$

L'accuratezza del modello indica la percentuale di osservazioni che sono state classificate in modo corretto dal classificatore. Perciò, tale misura dà un'indicazione dell'affidabilità del modello di classificazione generato. Può anche essere utilizzata per scegliere i parametri del modello, come ad esempio il numero di vicini K del K-NN. In alcuni casi, la sola accuratezza non è sufficiente a valutare le prestazioni di un modello di classificazione, in questi casi si ricorre alle altre due misure viste nelle formule precedenti. La precisione indica la percentuale di osservazioni che sono state classificate come P e che, nella realtà, erano veramente tali. Il richiamo indica la percentuale di osservazioni di tipo P classificate effettivamente come P dal classificatore. Quindi, la precisione dà un'indicazione di quanto ha funzionato bene il modello mentre il richiamo rappresenta una percentuale di quanto le classi, singolarmente, vengano correttamente classificate.

Capitolo 3

Risultati Sperimentali

In questo capitolo vengono introdotti i risultati sperimentali ottenuti all'interno delle diverse componenti del *framework*. Per ognuno dei blocchi principali, viene discussa la metodologia implementata e l'esito sperimentale del suo utilizzo. Inoltre, vengono esposti gli ambienti di sviluppo utilizzati per l'elaborazione del *framework* e gli strumenti impiegati durante lo svolgimento delle analisi.

Strumenti Tecnici Utilizzati

L'architettura di base del *framework* è stata sviluppata in *Python* [63] e integrata con *R* [64], un altro linguaggio di programmazione specifico per l'analisi statistica dei dati. Gli ambienti di sviluppo *open-source* utilizzati come IDE (*Integrated Development Environment*) sono `PyCharm` [65] ed `R Studio` [66]. Questi due strumenti di sviluppo sono stati affiancati anche da altri *tool* si supporto come `Excel`[1] e `Rapidminer`[2], un software *open-source* che consente di applicare alcune tecniche di *data mining* in modo automatico.

Librerie	Descrizione
ggplot	generazione di alcune statistiche, come i boxplot
rpart	costruzione di alberi di regressione e classificazione, e.g CART
distances	calcolo della distanza spaziale e della similarità
MLmetrics	metriche di performance dei modelli costruiti

Tabella 3.1: Riepilogo di alcune librerie in R utilizzate dal *framework*.

[1]Informazioni su Excel https://support.office.com/it-it/excel

[2]Maggiori dettagli su Rapidminer https://rapidminer.com/get-started/

Librerie	Descrizione
pandas, csv	manipolazione dei file CSV, tramite DataFrame
matplotlib	creazione di istogrammi, diagrammi a torta, radar chart ecc
scikit-learn	*data mining* (*clustering*, predizione, classificazione)
folium	realizzazione delle mappe geolocalizzate navigabili
shapely	elaborazione e analisi di oggetti geometrici nel piano cartesiano
pygeoj, pyproj	lettura, scrittura e modifica dei file geojson
Levenshtein	calcolo delle somiglianze fra le stringhe
geocoder	gestione di servizi per la risoluzione degli indirizzi

Tabella 3.2: Riepilogo di alcune librerie in Python utilizzate dal *framework*.

Per lo sviluppo dell'applicazione web sono stati utilizzati inoltre: `HTML 5` [67] come linguaggio di *markup*[3], `CSS 3` [68] per la formattazione delle pagine web, `Javascript` [69] e `Flask`, [70], due strumenti fondamentali per la realizzazione della componente *back-end* del sito web.

Collezioni di Dati Open

Il *framework* TUCANA è stato validato sperimentalmente su una collezione reale di certificati energetici raccolti e rilasciati in modalità *open* dal CSI-Piemonte. Le certificazioni considerate sono state emesse negli anni tra il 2009 e il 2018 e si distinguono tra ACE ed APE[4], come mostrato in Figura 3.1.

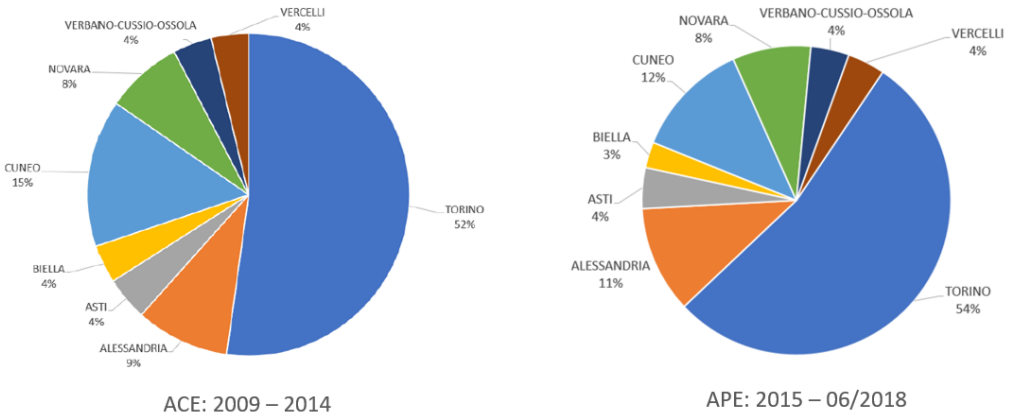

Figura 3.1: Distribuzione delle province della Regione Piemonte negli *open data*.

[3]Insieme di regole usate per definire la struttura di rappresentazione di un testo.

[4]La differenza fra i due certificati è stata trattata e riportata nella sezione 1.1.1.

Questa collezione di dati *open* comprende **circa 270 000 certificati** energetici degli edifici e appartamenti situati in Piemonte, una delle principali regioni italiane. Ciascuna certificazione energetica è caratterizzato da 132 attributi, tra cui attributi relativi all'aspetto energetico e attributi termofisici, suddivisi in 89 variabili categoriche e 43 variabili quantitative.

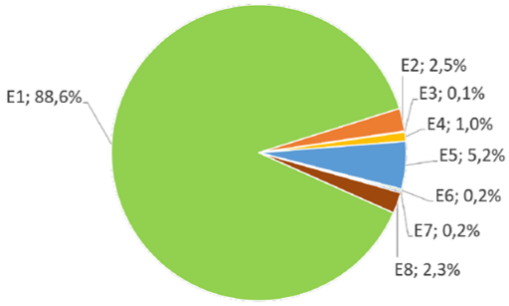

Figura 3.2: Distribuzione delle destinazioni d'uso degli immobili negli *open data*.

Analizzando la distribuzione delle province all'interno dell'insieme di dati, si osserva una maggioranza di certificati emessi per la provincia di Torino. Inoltre, dalla Figura 3.2, si nota che circa il 90% delle finalità di utilizzo degli immobili[5] è di tipo E.1(1), ovvero edifici residenziali a carattere continuativo. Per questo motivo, i risultati presentati in questo lavoro di tesi si focalizzano sull'analisi delle certificazioni energetiche della città di Torino selezionando i certificati relativi alle unità abitative di tipo E.1(1). Sono stati ottenuti, così, **circa 50 000 certificati** utilizzati come *dataset* su cui condurre le analisi successive. Le abitazioni di categoria catastale A presenti a Torino sono approssimativamente 500 000 [61], poiché il *dataset* di partenza contiene più o meno 50 000 dati, il *database* ottenuto rappresenta quasi il 10% degli edifici reali di Torino.

3.1 Preparazione dei Dati

In questa sezione, vengono riportati nel dettaglio le implementazioni e i risultati sperimentali, ottenuti con l'applicazione del primo blocco principale del *framework* TUCANA considerando *dataset* del caso di studio. In particolare, viene trattata la prima fase fondamentale realizzata dal *framework*, ovvero la fase di *Pre-elaborazione dei Dati*, spiegata dal punto di vista teorico nella sezione 2.1.

[5]Per maggiori dettagli sulle *destinazioni d'uso* consultare la Tabella 1.1.

Data la natura dei dati *open* a disposizione, diversi possono essere i problemi riscontrati all'interno della base di dati. In generale, se i dati su cui vengono effettuate le analisi sono non attendibili, la qualità dei risultati finali potrebbe essere scarsa. All'interno del *framework* vengono gestite le seguenti situazioni:

- *Dati incompleti*: serie di valori mancanti o non valorizzati;
- *Dati rumorosi*: valori anomali con errori oppure *outlier*;
- *Dati inconsistenti*: serie di valori in cui è presente un forte divario;
- *Dati duplicati*: varie ripetizioni delle stesse quantità.

Al fine di limitare gli effetti dovuti alla presenza di una di queste casistiche, il *framework* utilizza sia la metodologia *data-driven*, basata sull'applicazione di tecniche di *data mining*, che quella *domain-driven*, guidata dall'esperto di dominio.

3.1.1 Pulizia dei Dati Geografici

All'interno del *database*, ottenuto dal CSI-Piemonte, sono stati rintracciati diversi problemi legati al campo libero *indirizzo* delle certificazioni energetiche. Come descritto nella sezione 2.1.1, *framework* sviluppato è in grado di effettuare una fase preliminare di *address resolution*, caratterizzata da due operazioni separate:

- **Confronto con il viario**: viene implementato un algoritmo che effettua il confronto della via presente nel *database* con quelle del viario di Torino. Nello specifico, si correggono gli attributi *indirizzo, numero civico, CAP, latitudine, longitudine* quando il confronto va a buon fine e, cioè, quando il valore dell'indice di *similarità di Levenshtein* supera la soglia di 0.95. Un esempio di risoluzione di indirizzi ottenuto con questa prima procedura è riportato nella Tabella 3.3.
- **Utilizzo Geocoding API**: quando l'indice di *similarità di Levenshtein* fra le due stringhe 'via del database' e 'via del viario' restituisce un valore al di sotto della soglia posta a 0.95, allora viene utilizzata l'applicazione Geocoding API offerta da Google Maps. Si riporta un esempio di output ottenuto con questa seconda procedura in Tabella 3.4.

Per quanto riguarda la prima operazione descritta, l'algoritmo memorizza in un dizionario le informazioni all'interno del viario di Torino, in modo da confrontarle con gli attributi geografici presenti nel *database*. Questo algoritmo prevede anche

altre fasi preliminari come la conversione dei caratteri in formato ASCII. Tuttavia, per ragioni espositive di questo lavoro, si propone il `pseudocodice` della funzione che effettua il confronto con il viario di Torino, riportata nell'Algoritmo 2.

Algoritmo 2 Ricerca della via nella base di dati nel viario di Torino

Input: $viaUtente, dizionarioViario$
Output: $viaTrovata, maxLevValue$

```
1:  maxLevValue ← 0
2:  lenViaUtente ← lunghezza viaUtente
3:  for viaViario, listaPermutazioni in dizionarioViario do
4:      levValue ← 0
5:      for Permutazione in listaPermutazioni do
6:          lenPermutazione ← lunghezza permutazione
7:          if levValue !=1  and lenViaUtente ≥ lenPermutazione then
8:              viaConcatenata ← Concatena parole Permutazione
9:              simValue ← Calcola simLev tra viaUtente e viaConcatenata
10:             if simValue > levValue then
11:                 levValue ← simValue
12:             end if
13:         end if
14:     end for
15:     if levValue ≥ maxLevValue then
16:         maxLevValue ← levValue
17:         Aggiungi viaViario alla listaVie
18:         Aggiungi maxLevValue alla listaLevenshtein
19:     end if
20: end for
21: viaTrovata ← prendi ultima via in listaVie
22: diff ← calcola differenza tra lunghezza viaUtente e viaTrovata
23: while listaLevenshtein[indiceviaTrovata] == maxlevValue do
24:     viaTrovata ← listaVie[viaTrovata-1]
25:     if differenza tra lunghezza viaUtente e viaTrovata < diff then
26:         Aggiorna diff
27:         Aggiorna viaTrovata
28:     end if
29: end while
30: return viaTrovata, maxLevValue
```

La funzione considera la via all'interno del *database* e calcola la *similarità di Levenshtein* con tutte le possibili permutazioni delle parole che compongono le vie del viario. Viene poi memorizzato il *valore di Levenshtein* più elevato e la via

del viario tale per cui si ottiene questo valore. Nel caso in cui siano presenti più valori uguali, vengono salvati tutti quanti e poi selezionato il valore che corrisponde alla via del viario con lunghezza più vicina a quella della via del *database*. Se, ad esempio, si considera 'VIA PO' di lunghezza 2 dal *database*, l'algoritmo confronterà tale stringa con le permutazioni sia di 'VIA MONTEU DA PO' di lunghezza 4, che con quelle di 'VIA PO', perché entrambe le vie sono presenti all'interno del viario di Torino. In questo caso, la funzione è in grado di associare correttamente la via del viario perché seleziona quella con lunghezza più vicina alla via del *database*.

Quando il valore della *similarità di Levenshtein* scelto supera la soglia 0.95, l'algoritmo di sanitizzazione dei dati geografici controlla anche che il *numero civico* presente nella base di dati trovi corrispondenza nel viario per la via trovata. Se la risposta è positiva, gli attributi geografici memorizzati nel *database* vengono automaticamente sostituiti con quelli presenti nel viario di Torino, altrimenti si prendono le informazioni associate al civico più vicino trovato nel viario. Nella situazione in cui il *numero civico* è un campo vuoto, contiene errori di battitura oppure lettere, si assegnano le informazioni geografiche relative al civico più frequente presente nel viario per quella via trovata.

Indirizzo Originale	**Civico**	**Indirizzo Risolto**	**Civico**	**Lev**
VIPACCO	45	VIA VIPACCO	45	1
CORSO ROSSELLI	155	CORSO CARLO E NELLO ROSSELLI	155	1
CORSO E.GAMBA	36-S	CORSO ENRICO GAMBA	36	0.95
LUNGO PO DIAZ	6	LUNGO PO ARMANDO DIAZ	6	1
VIA EANDI	28	VIA ABATE ANTONIO VASSALLI EANDI	28	1
VIA ARTOM	9	VIA EMANUELE ARTOM	9	1
CORSO MARONCELLI PIETRO	9	CORSO PIERO MARONCELLI	9	0.98
GABETTI	12	CORSO GIUSEPPE GABETTI	12	1

Tabella 3.3: Esempi di risoluzione degli indirizzi tramite il confronto con il viario.

Le richieste effettuate con l'applicazione Geocoding si eseguono tramite la libreria *geocoder* di Python, la quale richiede l'indirizzo da correggere e l'API Key, ovvero un codice identificativo univoco associato all'account Google su cui vengono addebitate le richieste. Grazie all'applicazione dell'algoritmo di sanitizzazione degli attributi geografici, il numero di richieste effettuate sono state soltanto circa 1 800 perché l'algoritmo ha risolto in totale una buona parte del *dataset*, come mostrato nella Tabella 3.5. Con questa seconda procedura di *address resolution* è stato eliminato circa lo 0.14% di certificati dal *database*, questo perché alcuni indirizzi originali, corretti con l'applicazione Geocoding, sono risultati fuori Torino e, perciò, sono stati automaticamente eliminati, visto il caso di studio in esame.

Sono stati eliminati, inoltre, un'altra percentuale di 0.067% certificati dal *database* durante la prima procedura di *address resolution*, in questi casi veniva riportato il campo *indirizzo* vuoto. Nel complesso, grazie all'applicazione dell'intera procedura, oltre il **99%** degli indirizzi presenti nella base di dati sono stati geograficamente **corretti**.

Indirizzo Originale	**Civico**	**Indirizzo Risolto**	**Civico**	**Lev**
VIA DAUBR	10	VIA ADOLPHE DAUBREE	10	0.84
VIO BOVIO	3	VIA BOBBIO	3	0.88
CORSO SIRACURA	166	CORSO SIRACUSA	166	0.92
VIA UGOLINI EVIA RIGOLA	1-3	VIA GIUSEPPE RIGOLA	1	0.70
MARTIGNANA	25	VIA MARTINIANA	25	0.90
VIA GIOSU BORSI	73	VIA GIOSUE BORSI	73	0.93
C.SO SICILA	21	CORSO SICILIA	21	0.86

Tabella 3.4: Esempi di risoluzione degli indirizzi con il servizio Geocoding API.

Metodologia	**Risolti**	**Eliminati**
Viario di Torino	96%	0.067%
Geocoding API	3%	0.14%

Tabella 3.5: Risultati ottenuti con la procedura di pulizia dei dati geografici.

3.1.2 Correzioni di Scala e Filtri sulle Variabili

Dopo aver ripulito gli attributi geospaziali presenti nel *dataset*, il *framework* propone l'individuazione e l'eliminazione degli *outlier* mediante la tecnica *domain-driven*. In particolare, vengono attuate in cascata le seguenti operazioni:

- **Scalatura**: viene effettuata una riduzione di scala per alcuni attributi, perché molti valori si presentavano con scale diverse da quelle permesse. Ad esempio, se un rendimento (variabile definita tra 0 e 1) assume valore 89 allora tale quantità viene corretta con 0.89;
- **Filtraggio**: sulla base dei suggerimenti dell'esperto di dominio e dei documenti energetici, ad esempio il documento dell'IREN [72], vengono applicati dei filtri personalizzati ai valori assunti da parte di alcuni attributi: *trasmittanze opache* e *trasparenti*, *fattore forma* e i quattro *rendimenti* che caratterizzano gli impianti di riscaldamento.

Nella Tabella 3.6 si riportano nel dettaglio i filtri applicati agli attributi critici appena citati. Le certificazioni energetiche con valori fuori dai range di ammissibilità definiti dall'esperto di dominio, sono state eliminate dalla base di dati. Per meglio comprendere la distribuzione dei dati prima della fase di filtraggio, si riportano sia la Figura 3.3, con le distribuzioni di *fattore forma*, *tramittanze opache* e *trasparenti*, che la Figura 3.4 con le distribuzioni dei *rendimenti*.

Sigla	**Attributo**	**Range di ammissibilità**
S/V	Fattore Forma	$[0.1 - 2]$
ETAD	Rendimento Distribuzione	$[0.75 - 1.25]$
ETAE	Rendimento Emissione	$[0.85 - 1]$
ETAG	Rendimento Generazione	$[0.65 - 1.1]$
ETAR	Rendimento Regolazione	$[0.6 - 1]$
UO	Trasmittanza Opaca	$[0.1 - 3]$
UW	Trasmittanza Trasparente	$[0.9 - 7]$

Tabella 3.6: Filtri sulla base dei range di accettabilità.

Come evidenziato dalla Tabella 3.7, l'applicazione dei filtri alle variabili *fattore forma*, *tramittanze opache* e *trasparenti* non ha impattato molto sul *dataset* infatti, in questa fase, sono stati scartati poco più che l'1% dei certificati presenti. D'altra parte, i filtri sugli attributi *rendimento di distribuzione*, *emissione*, *generazione* e *regolazione*, hanno causato un impatto maggiore perché, dopo questa operazione, sono rimasti a disposizione circa il 67% dei dati originari.

Sigla	**Attributo**	**Eliminati (%)**
S/V	Fattore Forma	0.18
UW	Trasmittanza Trasparente	0.22
UO	Trasmittanza Opaca	0.15
ETAD	Rendimento Distribuzione	0.88
ETAE	Rendimento Emissione	0.72
ETAG	Rendimento Generazione	6.31
ETAR	Rendimento Regolazione	0.20

Tabella 3.7: Dettaglio della percentuale di certificati energetici eliminati dal *database* per ogni attributo fuori dal range di ammissibilità.

Osservazione 3.1. *Visto l'impatto che l'applicazione dei range di ammissibilità ha avuto sui dati, alcuni certificati energetici eliminati in questa fase vengono recuperati tramite le metodologie presentate nel paragrafo 3.4 e nella sezione 3.4.3.*

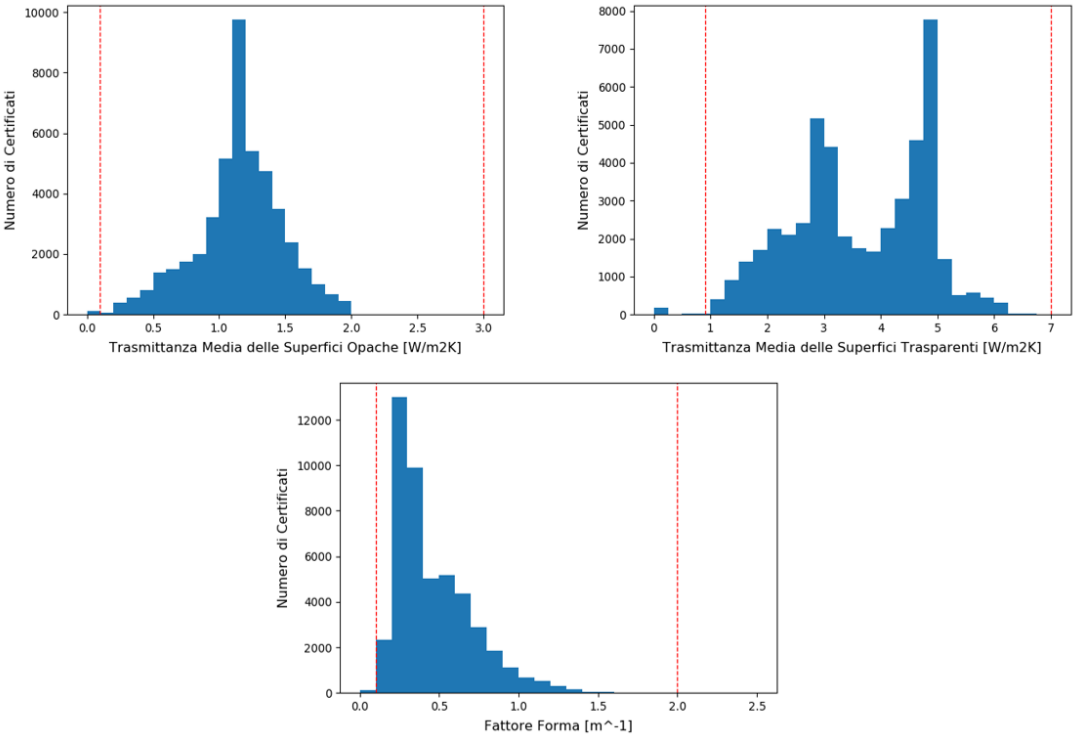

Figura 3.3: Distribuzione delle *trasmittanze opache*, *trasparenti* e del *fattore forma*.

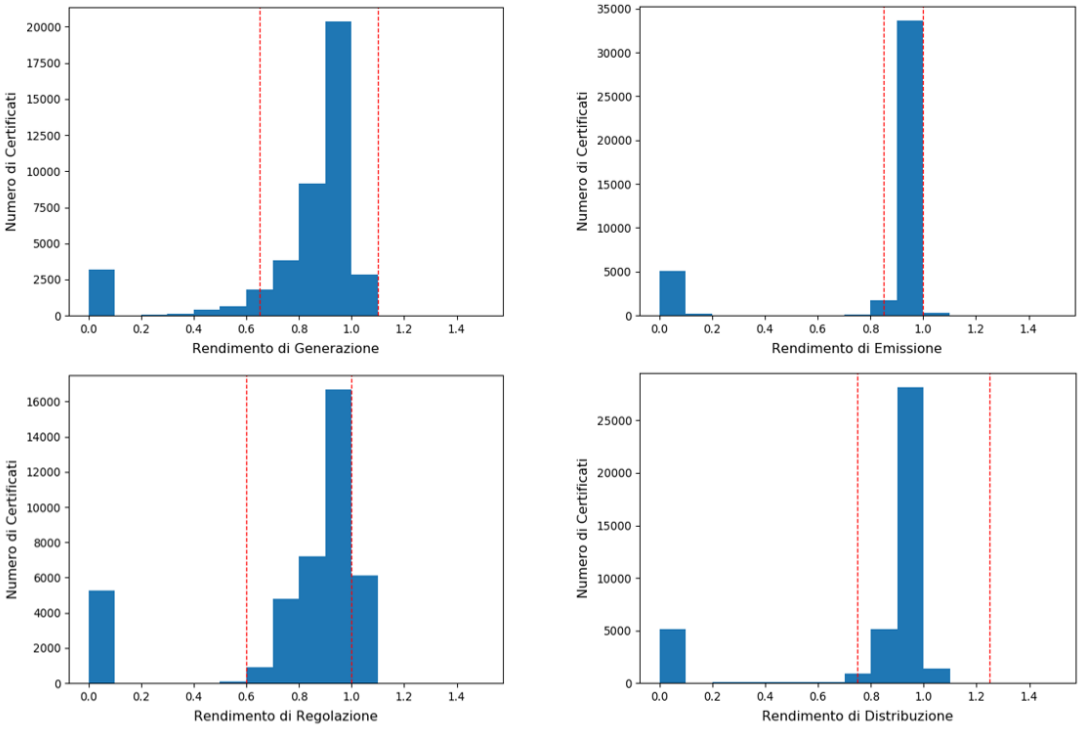

Figura 3.4: Distribuzione dei quattro *rendimenti* dei sottosistemi.

3.1.3 Eliminazione degli Outlier

Il *framework* propone strategie *data-driven* per l'eliminazione di eventuali *outlier* dalle distribuzioni di alcuni attributi del *database*, ritenuti importanti per le analisi successive. Come tecniche univariate, si riportano i risultati sperimentali ottenuti applicando il metodo gESD ed eliminando quei valori al di sotto del percentile 1%. Gli attributi su cui viene applicato il gESD sono *superficie riscaldata*, *superficie disperdente totale*, *volume lordo riscaldato*, *emissioni gas serra*, *anno di costruzione* e *fabbisogno di energia termica utile*. In Figura 3.5, viene mostrato un esempio del funzionamento del gESD sull'attributo *anno di costruzione*. In questo caso, il gESD ha funzionato bene perché riesce a identificare come *outlier* diversi punti (in rosso) che si discostano in modo significativo dalla distribuzione dei dati.

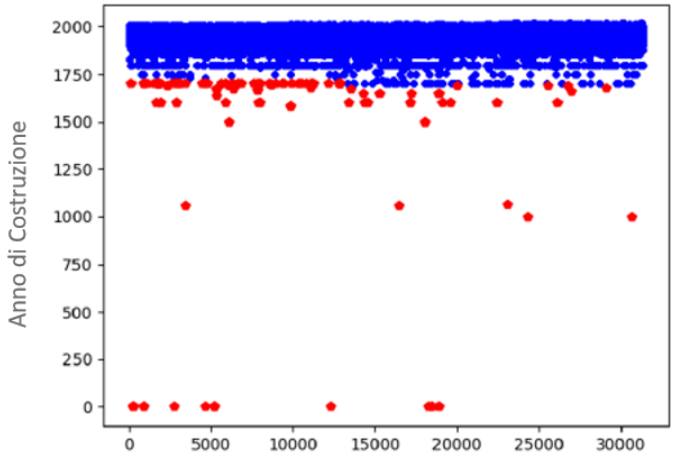

Figura 3.5: Risultato del gESD sull'attributo *anno di costruzione*. La distribuzione dei dati (in blu) si discosta dagli *outlier* (in rosso).

La metodologia gESD è stata utilizzata considerando come parametro "numero massimo di *outlier*" lo 0.5% dei certificati presenti all'interno del *database*, invece per il parametro di significatività si è posto $alpha = 1\%$. Per alcuni attributi come la *superficie disperdente totale* e il *il volume lordo riscaldato*, il gESD non ha fornito degli output soddisfacenti. Per questo motivo, viene applicata la seconda metodologia univariata, basata sulla rimozione tramite percentile, al fine di eliminare i valori poco realistici rimasti e che si trovano nelle code delle distribuzioni. In seguito all'utilizzo di queste tecniche univariate per la procedura di *outlier detection*, il *framework* individua ed elimina circa il 4% dei certificati identificati come *outlier* dal *dataset*.

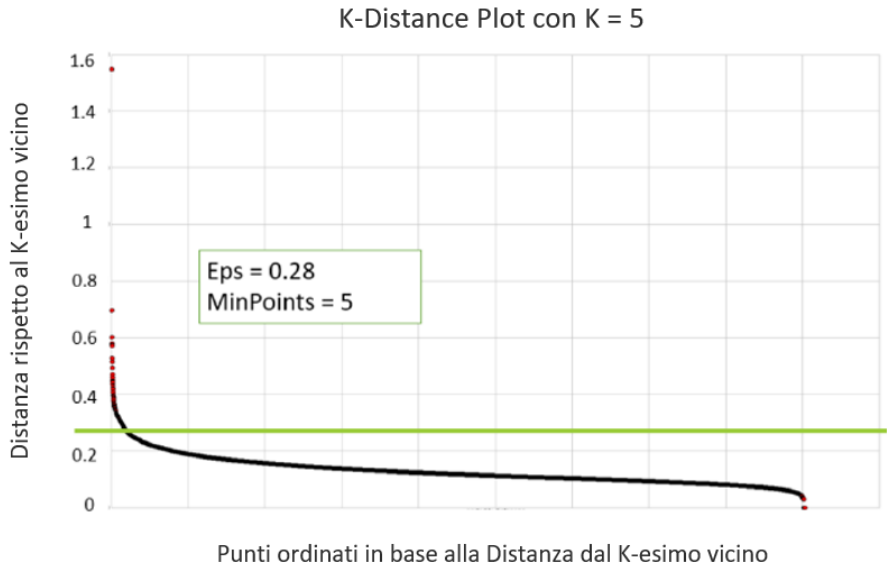

Figura 3.6: *K-Distance Plot* per la scelta dei parametri per il DBSCAN.

Al fine di garantire che i dati analizzati siano il più possibile privi di *outlier*, la metodologia univariata viene affiancata da quella multivariata, in modo da ridurre ulteriori anomalie. In particolare, viene applicato l'algoritmo DBSCAN con parametri $Eps = 0.28$ e $MinPoints = 5$, come evidenziato dalla Figura 3.6, ottenuta grazie allo strumento *Data to Similarity*[6] messo a disposizione da `Rapidminer`. In seguito all'utilizzo di questo algoritmo, alcuni pochi certificati sono stati eliminati dalla base di dati, che contiene a questo punto **circa 30 000 record**.

3.1.4 Selezione e Normalizzazione delle Variabili

In questa fase, il *framework* effettua un'analisi delle correlazioni fra diversi attributi energetici ritenuti importanti per la determinazione delle prestazioni energetiche di un'abitazione. L'analisi della matrice di correlazione generata dal *framework* viene guidata dai suggerimenti dell'esperto di dominio, attraverso i quali si decidono le variabili da utilizzare negli algoritmi di *data mining*. In Figura 3.7, viene mostrata la matrice di correlazione degli attributi scelti come più significativi ai fini delle analisi. La scelta di questi attributi viene fatta considerando le variabili che risultano essere poco correlate fra di loro.

[6] https://docs.rapidminer.com/latest/studio/operators/modeling/similarities/data_to_similarity.html

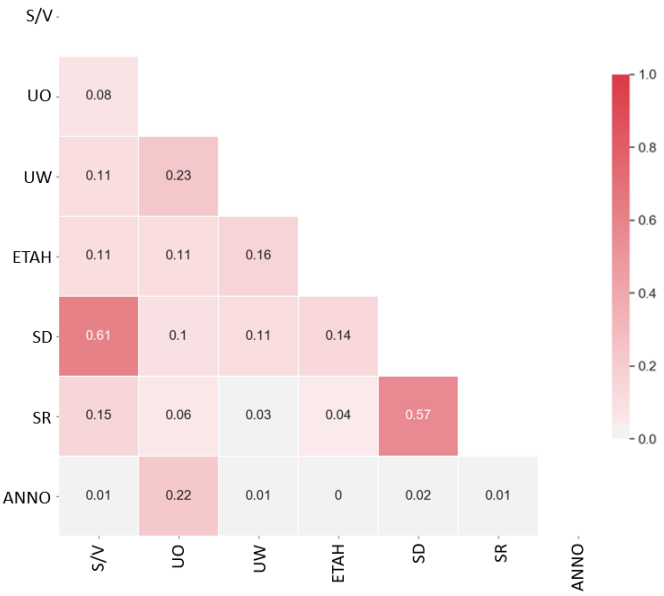

Figura 3.7: Matrice di correlazione degli attributi scelti per le analisi.

Dall'immagine in Figura 3.7, si osserva proprio che la maggior parte degli attributi scelti tendono ad assumere bassi valori di correlazione. Queste considerazioni sono fondamentali per indirizzare una valida estrazione della conoscenza. Infatti, quando alcune variabili sono altamente correlate fra di loro, esse tendono ad esprimere la stessa informazione [71]. Tenere queste variabili all'interno delle analisi, spinge i modelli generati a spiegare con un peso maggiore lo stesso concetto più volte, quando invece l'obiettivo è quello di estrarre conoscenza utile dai dati.

Sigla	**Nome Variabile**	**Unità di misura**
S/V	Fattore Forma o Rapporto S/V	$[m^{-}1]$
UW	Trasmittanza Media Superfici Trasparenti	$[W/m^2K]$
UO	Trasmittanza Media Superfici Opache	$[W/m^2K]$
SR	Superficie Riscaldata	$[m^2]$
SD	Superficie Disperdente Totale	$[m^2]$
ETAH	Rendimento Climatizzazione Invernale	[–]
ANNO	Anno Costruzione	[–]

Tabella 3.8: Riepilogo delle variabili significative scelte per le analisi.

Anche la fase di selezione delle variabili di analisi risulta essere un'importante procedura, ai fini dell'estrazione della conoscenza dai dati. Il *framework* considera la normalizzazione delle variabili in Tabella 3.8 attraverso la metodologia *Max-Min*,

descritta nella sezione 2.1.2. La normalizzazione degli attributi di interesse è un'operazione necessaria sia per poter applicare l'algortimo di *clustering* descritto nel paragrafo successivo, sia per creare i modelli di generalizzazione trattati nel paragrafo 3.4 dove, in particolare, si adotta una diversa tecnica di normalizzazione.

3.2 Applicazione dell'Algoritmo K-Means

I passi preliminari di *Pre-elaborazione dei Dati* hanno permesso di ripulire il *database* in esame dai casi anomali, in modo da non impattare negativamente sulle analisi che vengono effettuate in questa fase del *framework*. In particolare, in questo paragrafo, viene applicato l'algoritmo di *data mining* scelto per l'estrazione della conoscenza dalla base di dati, ed applicate alcune metodologie statistiche per la caratterizzazione dell'informazione ottenuta. Il *framework* propone come algoritmo di *clustering* il *K-Means*, la cui procedura teorica è stata trattata nella sezione 2.2.1. La scelta dei metodi da utilizzare per esplorare una collezione di dati rappresenta uno degli step più delicati del KDD (*Knowledge Discovery in Database*), perché le prestazioni degli algoritmi dipendono dalla tipologia di dati considerati. In questo lavoro di tesi, è stato scelto come algoritmo il *K-Means* perché, confrontandosi con i metodi proposti allo stato dell'arte, esso viene utilizzato molto spesso per la sua semplicità e agevolezza computazionale. L'unica questione da gestire quando si utilizza questo algoritmo è la scelta a priori del parametro K desiderato per il numero di *cluster*, condizione risolta con la metodologia automatica dell'*elbow method* presentata nella sezione 2.2.2.

3.2.1 Parametri di Input

Gli input da fornire all'algoritmo *K-Means* sono gli attributi di interesse da utilizzare per la creazione dei gruppi con caratteristiche simili, e il parametro K, ovvero il numero di gruppi omogenei che si desiderano ottenere con l'applicazione di questo algoritmo. Le variabili di analisi utilizzate per questa fase di *clustering* sono: *fattore forma, trasmittanza trasparente, trasmittanza opaca, superficie riscaldata, superficie disperdente totale, rendimento di climatizzazione invernale e anno di costruzione*, come riportato nella Tabella 3.8. Per poter trovare il parametro K ottimale invece, si utilizzata la metodologia basata sul metodo del "gomito". In pratica, l'algoritmo viene fatto girare sulla collezione di dati per determinati valori di K, in modo tale da poter ricavare i valori di SSE corrispondenti ad ogni passo e generare la curva presentata in Figura 3.8.

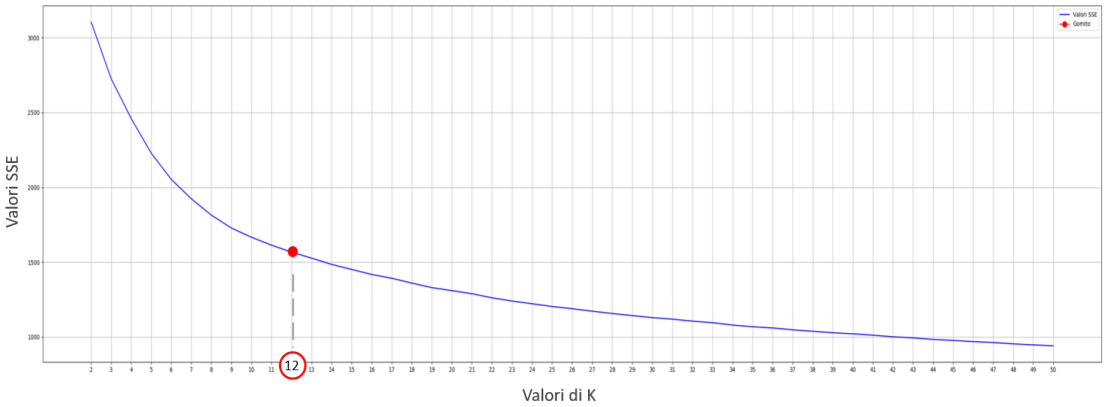

Figura 3.8: Scelta automatica del K per il *K-Means* con l'*elbow method*.

Come si evince dal grafico in Figura 3.8 vengono ottenuti dei valori di SSE dal *K-Means* per K= $1, \ldots, 50$, un insieme di valori scelto dall'esperto di dominio ritenuto sufficientemente adeguato per valutare gruppi di edifici con caratteristiche simili. Il risultato ottimale per K si trova in corrispondenza del punto della curva in cui si presenta una maggiore decrescita del valore dell'SSE. Secondo il metodo automatizzato del "gomito", tale valore viene raggiunto esattamente per $\mathbf{K = 12}$.

3.2.2 Risultati del K-Means

L'algoritmo *K-Means* con il parametro K= 12 scelto nella fase precedente, restituisce dodici *cluster* diversi, ognuno dei quali contenente un certo numero di certificati con caratteristiche termofisiche e geometriche simili.

	Circostrizione								
Label	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**Totale**
0	0.34	0.82	1.10	0.72	0.94	0.74	0.57	0.75	6%
1	0.77	0.96	1.03	0.83	0.43	0.45	0.48	1.05	6%
2	0.30	0.79	0.88	0.94	0.87	0.37	0.65	0.80	6%
3	0.83	0.28	0.30	0.26	0.07	0.14	0.36	0.69	3%
4	0.73	1.3	1.7	1.0	1.0	0.90	0.97	1.4	9%
5	1.43	0.61	0.78	0.55	0.11	0.12	0.35	0.87	5%
6	1.3	2.5	2.3	1.6	1.3	1.0	1.2	2.5	14%
7	1.39	1.44	2.12	1.60	1.38	2.08	1.17	1.71	12%
8	1.5	2.5	2.9	2.2	2.0	1.5	1.7	2.3	16%
9	2.60	1.97	2.14	1.57	1.17	0.91	1.19	1.85	12%
10	2.1	0.0068	0.027	0.047	0.0033	0.030	0.18	0.26	3%
11	0.85	1.07	1.46	0.81	1.0	0.89	0.97	1.34	8%

Tabella 3.9: Percentuale di certificati suddivisi per *cluster* e circoscrizioni.

Nella Tabella 3.9 si può notare come si distribuiscono i certificati del *database* fra i diversi gruppi creati, cioè vengono riportate le percentuali approssimative di certificati all'interno di ogni gruppo e, in contemporanea, all'interno di ogni circoscrizione. Si veda la Figura 3.9 per approfondire i quartieri di Torino che compongono le diverse circoscrizioni.

Circoscrizione	Quartieri
1	Centro - Crocetta
2	Santa Rita - Mirafiori
3	San Paolo - Cenisia - Pozzo Strada - Cit Turin - Borgata Lesna
4	San Donato - Campidoglio - Parella
5	Borgo Vittoria - Madonna di Campagna - Lucento - Vallette
6	Barriera di Milano - Regio Parco - Barca - Bertolla - Falchera - Rebaudengo - Villaretto
7	Aurora - Vanchiglia - Sassi - Madonna del Pilone
8	San Salvario - Cavoretto - Borgo Po - Nizza Millefonti - Lingotto - Filadelfia

Figura 3.9: Riepilogo delle circoscrizioni che suddividono la città di Torino.

In generale, dalla Tabella 3.9, si osserva che i *cluster* più numerosi sono identificati con le etichette 6,7,8 e 9. Tuttavia, le numerosità dei diversi gruppi sembrano essere abbastanza vicine fra di loro. Come accennato all'inizio del paragrafo, nonostante il *K-Means* sia un algoritmo semplice da utilizzare, restituisce come conoscenza estratta un insieme di gruppi omogenei che necessitano di una fase di interpretazione. In particolare, per poter identificare le performance energetiche dei diversi *cluster*, il *framework* prevede una fase di caratterizzazione dei gruppi trovati, attraverso le metodologie approfondite nella sezione 2.3.

3.2.3 Caratterizzazione degli Edifici

In questa fase, il *framework* mette a disposizione degli analisti e degli esperti di dominio alcuni strumenti di analisi, che possono essere utilizzati per riuscire ad interpretare la conoscenza estratta durante la fase di *clustering*. Le diverse tecniche proposte possono essere utilizzate separatamente oppure congiuntamente, in modo tale da avere una panoramica completa delle informazioni relative ai diversi *cluster*. Nello specifico, sono stati utilizzati i seguenti metodi grafici:

- **Boxplot**: una volta considerati i dati appartenenti ad ogni *cluster*, vengono generati i *boxplot* per ciascuna variabile di analisi, in modo tale da caratterizzare le distribuzioni degli attributi di interesse nei gruppi trovati;

- **Radar Chart**: dopo aver calcolato i centroidi dei *cluster* generati, si utilizzano i *radar chart* per visualizzare congiuntamente i valori delle variabili di analisi, assunti dal centroide di ogni gruppo;
- **CART**: l'albero decisionale viene utilizzato per trovare alcune regole e implicazioni logiche fra gli attributi di interesse. Questo semplice metodo grafico diventa essenziale per gli esperti del settore energetico, perché facilita la classificazione delle performance energetiche degli edifici.

Di seguito verranno proposti alcuni esempi di come è stato indirizzato sia il processo di interpretazione, che di classificazione dei gruppi di edifici con caratteristiche termofisiche e geometriche simili. Quindi si confrontano i *boxplot* e i *radar chart* di alcuni gruppi omogenei e, infine viene presentato un esempio di regole ottenute dalla generazione del CART sul *dataset* a disposizione.

Osservazione 3.2. *Gli esempi e alcuni processi decisionali esposti in questa sezione, utilizzano delle nozioni energetiche di base utili alla definizione delle performance energetiche di un determinato edificio. In generale, secondo gli esperti di dominio, un'abitazione risulta efficiente dal punto di vista delle prestazioni energetiche quando le variabili* rapporto S/V, trasmittanze opache, trasmittanze trasparenti, superficie disperdente *e* superficie riscaldata *tendono ad assumere valori bassi e, quando gli attributi* rendimento di climatizzazione invernale *e* anno di costruzione *assumono valori elevati.*

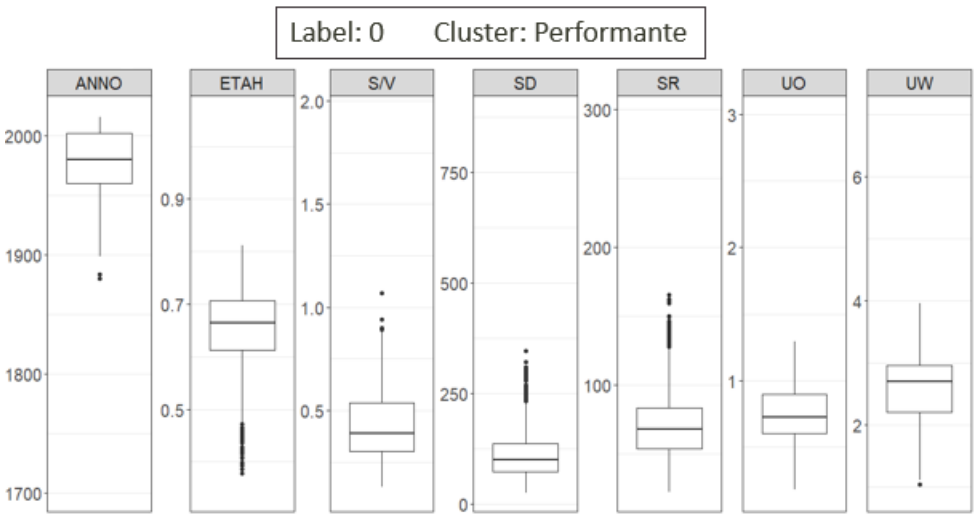

Figura 3.10: *Boxplot* delle variabili di analisi di un *cluster* performante.

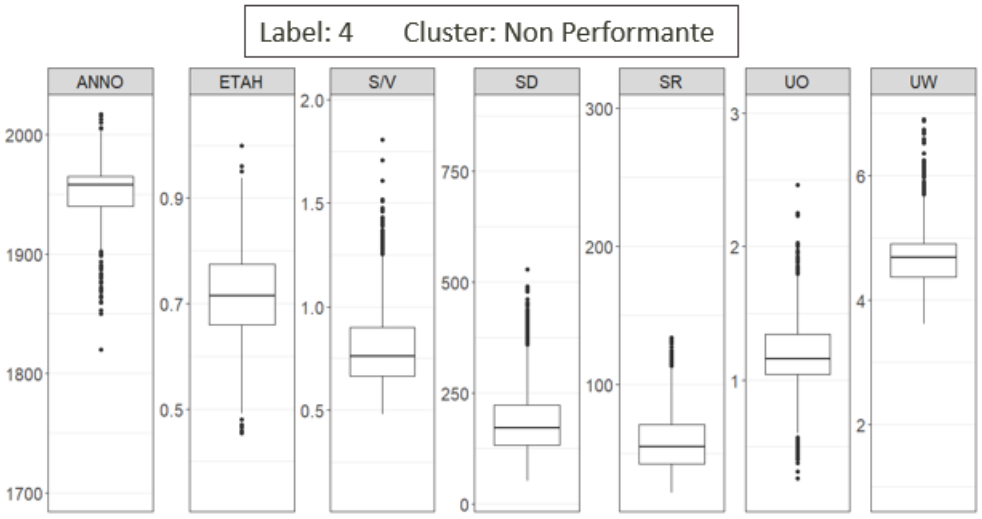

Figura 3.11: *Boxplot* delle variabili di analisi di un *cluster* non performante.

L'immagine presentata in Figura 3.10 evidenzia la distribuzione dei valori di ciascuna variabile di analisi all'interno del *cluster 0*, il quale presenta edifici abbastanza recenti con *rapporto S/V*, *superficie disperdente* e *trasmittanze trasparenti* tendenti a valori più bassi, rispetto al *cluster 4*. Ciò significa che, gli edifici che appartengono al gruppo etichettato con la *label 0*, sembrano essere più performanti rispetto a quelli appartenenti al gruppo etichettato con *label 4*.

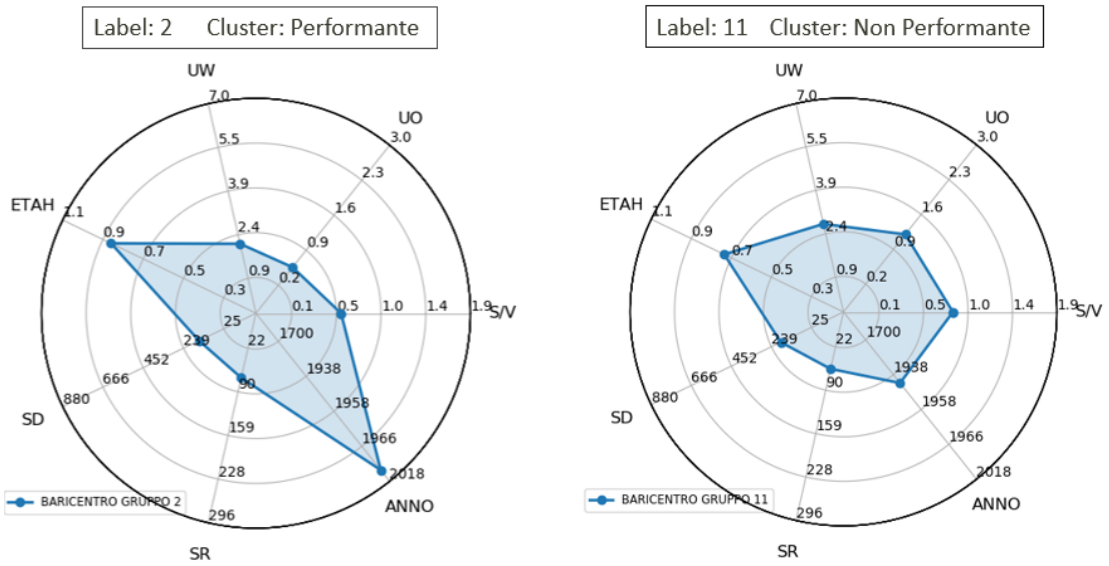

Figura 3.12: *Radar chart* di un centroide relativo ad un *cluster* performante (a sinistra) e di un centroide corrispondente ad un *cluster* non performante (a destra).

Per quanto riguardano i *radar chart*, la Figura 3.12 rappresenta un possibile confronto fra centroidi. Si può osservare che il centroide del *cluster 2* presenta un valore elevato sia per il *rendimento di climatizzazione invernale* che per l'*anno di costruzione*, ma valori bassi per le *trasmittanze opache, trasparenti* e per la variabile *rapporto S/V*. Invece, il centroide del *cluster 11* si comporta in modo opposto. In altre parole, esso mostra un valore più basso sia per il *rendimento di climatizzazione invernale* che per l'*anno di costruzione*, e valori più alti per le variabili: *trasmittanze opache, trasparenti* e *rapporto S/V*. Perciò, si può concludere che il *cluster 2* comprende, in media, edifici recenti con buone caratteristiche geometriche e termofisiche. D'altra parte, il *cluster 11* tiene conto, in generale, di edifici più antichi caratterizzati da un involucro poco performante.

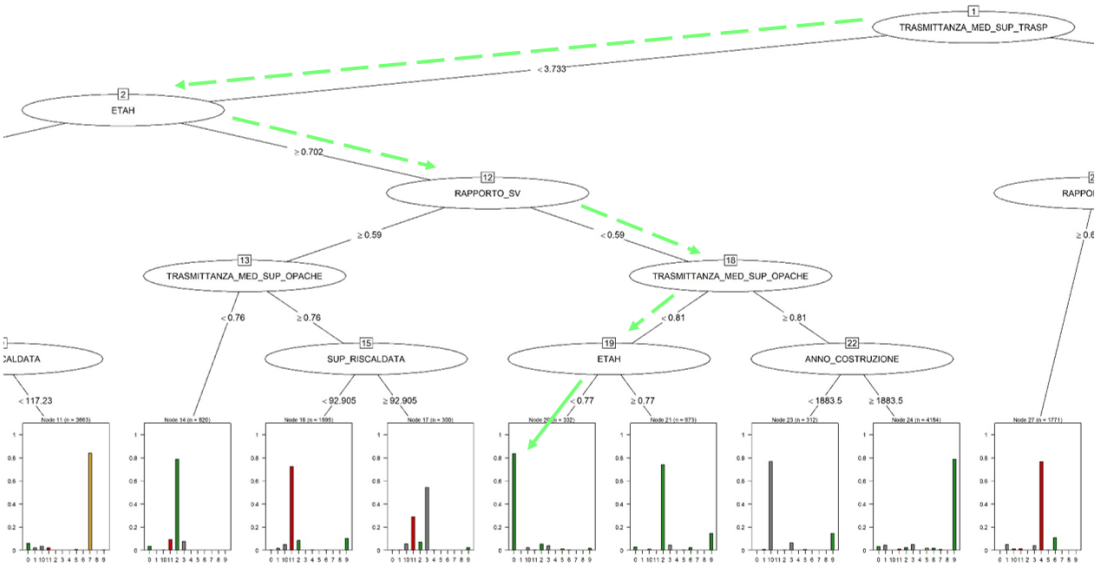

Figura 3.13: Esempio di una parte del CART generato dal *framework*.

L'albero decisionale presentato in Figura 3.13 rappresenta una porzione dell'intero CART restituito dal *framework*, e viene utilizzato soprattutto per caratterizzare gli edifici appartenenti ad un dato *cluster*. Questo tipo di metodo grafico è molto intuitivo e affidabile infatti, per quanto riguarda la validazione del CART, il *framework* implementa un albero di classificazione che raggiunge un'accuratezza dell'80%. La lettura di un albero di classificazione avviene partendo dal nodo radice e seguendo le frecce che collegano ogni nodo a quello successivo, si veda ad esempio il percorso evidenziato in Figura 3.13. Per esempio, il nodo radice dell'albero è rappresentato dalla *trasmittanza media delle superfici trasparenti*,

essa indica quindi la migliore variabile di analisi in grado di separare i *cluster* sulla base delle loro prestazioni energetiche. Questa metodologia consente all'analista di determinare alcune regole di tipo **if-then** all'interno dei gruppi omogenei che sono stati generati nella fase di *clustering*. Una possibile regola che si può estrarre dal CART è quella individuata dal percorso in mostrato nella Figura 3.13:

```
if  UW < 3.733, ETAH in [0.702 0.77], S/V < 0.59, UO < 0.81:
                then  Label Cluster = 0
```

Lo studio e gli approfondimenti effettuati con queste tecniche grafiche presentate, consentono di identificare le prestazioni energetiche degli edifici appartenenti a ciascun *cluster*. La Tabella 3.10 rappresenta il risultato finale della combinazione dei concetti ottenuti grazie al processo di caratterizzazione della conoscenza, implementato nel *framework* TUCANA.

Label	Performance	Descrizione
0	Alta	Involucro Performante, Impianto Mediamente Performante
1	Non Classificabile	Involucro Scarso, Fattore Forma Basso
2	Alta	Involucro e Impianto Performanti
3	Non Classificabile	Edifici con Ampia Metratura
4	Bassa	Involucro Scarso, Fattore Forma Alto
5	Media	Involucro Scarso, Impianto Mediamente Performante, Fattore Forma Basso
6	Alta	Involucro Scarso, Impianto Performante, Fattore Forma Basso
7	Media	Involucro Performante, Impianto Scarso, Fattore Forma Basso
8	Media	Involucro Mediamente Performante, Impianto Scarso, Fattore Forma Basso
9	Alta	Involucro e Impianto Mediamente Performanti, Fattore Forma Basso
10	Non Classificabile	Edifici Storici
11	Bassa	Involucro e Impianto Mediamente Performanti, Fattore Forma Alto

Tabella 3.10: Classificazione e caratterizzazione dei gruppi di edifici trovati.

3.3 Visualizzazione dei Risultati Ottenuti

In questo paragrafo vengono illustrate le metodologie adottate dal *framework* per presentare i risultati ottenuti nelle fasi precedenti. In particolare, vengono proposti due strumenti di visualizzazione differenti. Il primo riguarda una rappresentazione geolocalizzata sia degli edifici presenti nel *database* che delle variabili di analisi, tramite l'utilizzo di mappe dinamiche e facilmente navigabili. Il secondo invece, si basa sulla progettazione e lo sviluppo di un'applicazione web per la visualizzazione della conoscenza acquisita in tutte le fasi di analisi precedenti. Gli obiettivi principali della creazione di queste due tipologie di visualizzazione sono:

1. valorizzare la conoscenza estratta fornendo una corretta interpretazione dei dati a disposizione, utilizzando strumenti per un'analisi interattiva;
2. interpretare la conoscenza a diversi livelli di dettaglio e trasformare i dati in conoscenza utile, al fine di guidare i processi del *decision-making*;
3. distinguere le informazioni ottenute durante le analisi, allo scopo di comunicare i risultati sperimentali a diverse tipologie di utenti finali;
4. rendere i dati ottenuti dalle certificazioni energetiche facilmente fruibili, in modo che possano essere consultabili in qualsiasi momento.

3.3.1 Realizzazione delle Mappe

Poiché le certificazioni energetiche sono caratterizzate dagli attributi geografici, esse possono essere rappresentate su mappe dinamiche e geolocalizzate. Questo tipo di visualizzazione consente di ottenere un'informazione immediata sull'efficienza energetica di un edificio, e anche sulle prestazioni energetiche della sua zona di appartenenza. All'interno del *framework* si distinguono tre tipologie di mappe diverse realizzate in *Python* grazie alla libreria *folium*. Come riportato nella sezione 2.3.2, ogni tipologia di mappa interattiva comunica un'informazione specifica all'utente finale. Di seguito verranno presentati alcuni esempi delle mappe implementate dal *framework* TUCANA.

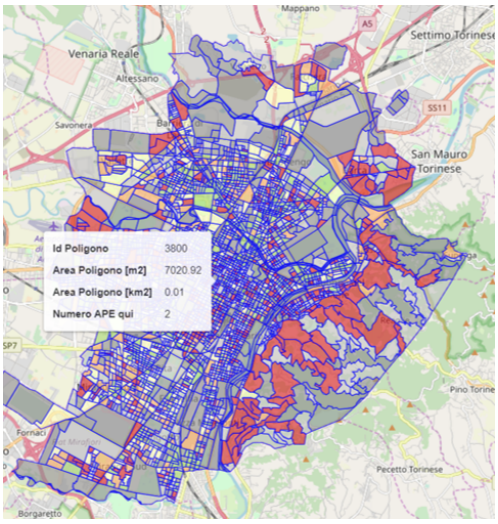

Figura 3.14: Mappa coropletica per la *superficie disperdente totale*.

L'immagine presentata a sinistra visualizza la mappa di Torino, suddivisa per isolati, chiamati all'interno dell'architettura sviluppata anche con il termine "poligoni". Il colore di ogni poligono viene assegnato in base al valore più frequente dell'attributo *superficie disperdente totale* all'interno del poligono stesso. Il *framework* prevede inoltre, la possibilità di conoscere l'area di ogni poligono e il numero di certificati presenti al suo interno, tramite la navigazione dinamica da parte dell'utente finale. I poligoni colorati di grigio, in questo caso, indicano una zona in cui la presenza di certificati è irrisoria.

Questo è un esempio di mappa di tipo *scatter*, ovvero una mappa in cui le certificazioni energetiche vengono rappresentate con un simbolo grafico detto *marker*. Quindi, in corrispondenza di ogni coordinata geografica visualizzata sulla mappa, è presente un edificio la cui certificazione fa parte del *database* in esame. La mappa a destra rappresenta, in particolare, un certo numero di edifici con performance energetica alta, secondo la Tabella 3.10 mostrata in precedenza. In particolare, si visualizzano alcuni certificati all'interno della circoscrizione 1 (Centro - Crocetta).

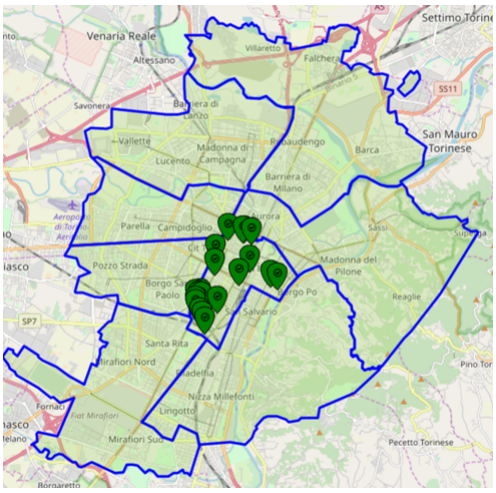

Figura 3.15: Esempio di mappa *scatter* generata dal *framework*.

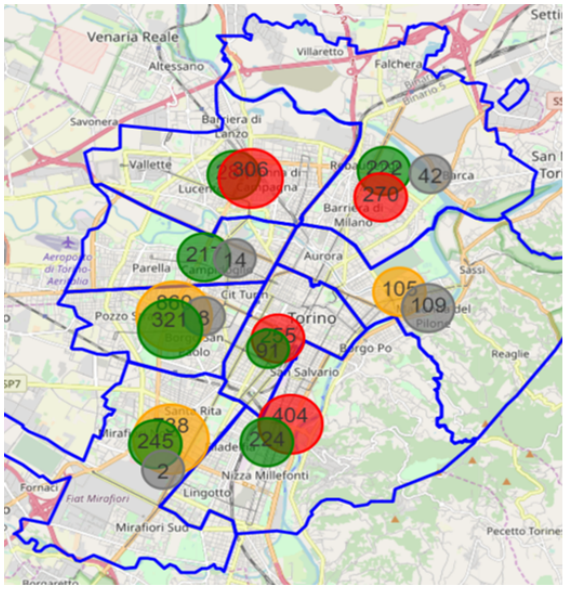

Figura 3.16: Mappa *marker-cluster* ad un livello di dettaglio inferiore (a sinistra) e ad un livello di dettaglio maggiore (a destra).

Come mostrato in Figura 3.16, le mappe *marker-cluster* sono strumenti di visualizzazioni che permettono di rappresentare informazioni aggregate a diversi livelli di dettaglio. Questo fa sì che una grande quantità di informazioni, come quelle utilizzabili dal *framework*, possano essere rese disponibili e navigabili in modo semplice ed intuitivo. La mappa *marker-cluster* infatti risolve il problema

dell'*override* informativo, perché un numero elevato di *marker* nella stessa mappa potrebbe rallentare la navigazione, limitando l'acquisizione delle conoscenze utili da parte dell'utente finale. Con questa metodologia invece, tale difficoltà viene superata tramite l'utilizzo dei simboli *marker-cluster*[7] personalizzabili tramite la libreria *folium*. Nel caso in esame, questi simboli rappresentano i *cluster* generati durante le fasi di analisi precedenti, e vengono colorati secondo le loro performance energetiche, determinate nella Tabella 3.10. Un'altra informazione immediata che viene restituita da questo tipo di mappa è la cardinalità di ogni gruppo, infatti il numero all'interno dei *marker-cluster* indica la numerosità di certificati energetici contenuti in quel determinato *cluster*.

3.3.2 Interfaccia dell'Applicazione Web

Lo scopo dell'applicazione web TUCANA è quello di rendere disponibile, a diverse tipologie di utenti, le informazioni acquisite durante la fase di caratterizzazione della conoscenza estratta. Vengono integrate inoltre, siano le mappe geolocalizzate che le statistiche di base sui dati *open* delle certificazioni energetiche. La realizzazione di questo lavoro è stata preceduta da una fase di studio approfondito su come distinguere le informazioni per i diversi utenti finali o *stakeholder*. Il sito web potrebbe essere di supporto a tre principali tipologie di utenti:

1. **Privato cittadino**: interessato a conoscere le prestazioni energetiche di alcune zone della città di Torino, usa il sito come strumento orientativo qualora fosse interessato ad acquistare oppure ad affittare un immobile;

2. **Pubblica amministrazione**: orientata verso l'efficienza energetica delle strutture edilizie, utilizza le informazioni presentate per individuare le aree critiche della città in cui poter apportare dei miglioramenti;

3. **Fornitore di servizi energetici**: interessato a conoscere le prestazioni energetiche attualmente presenti, sfrutta il portale web per ricercare diverse aree della città ed effettuare analisi di tipo comparativo.

Le rappresentazioni grafiche tipiche della statistica, come quelle presentate durante la fase di *Caratterizzazione degli Edifici*, sono del tutto comprensibili solo ad un numero ridotto di utenti. Per questo motivo, l'applicazione web è stata sviluppata creando diverse sezioni personalizzate, a seconda della tipologia di utente che lo visita, si veda la Figura 3.17.

[7] https://python-visualization.github.io/folium/plugins

Figura 3.17: Pagina html dell'applicazione web per l'autenticazione dell'utente.

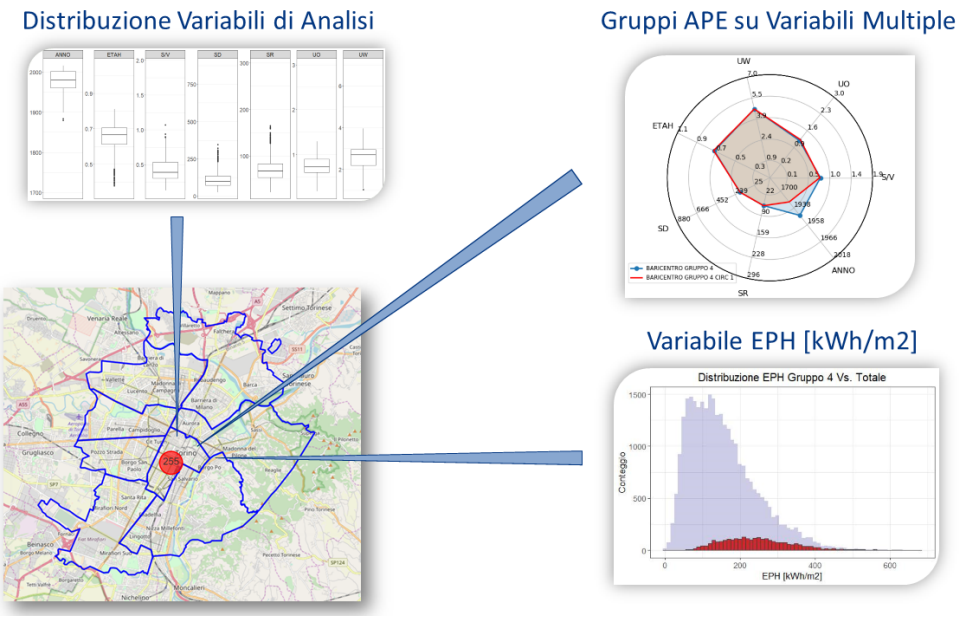

Figura 3.18: Esempio di statistiche specifiche sui *cluster*, presenti sul sito web e visualizzabili solo dagli utenti con un certo grado di *expertise*.

Mentre le mappe coropletiche e quelle *scatter* presentate nella sezione precedente, sono di facile interpretazione e possono essere potenzialmente visualizzate sia da utenti esperti che da un pubblico generico, la mappa *marker-cluster* presenta un concetto più di alto livello perché introduce il concetto di *clustering*. Per questo motivo, alcune sezioni sono visualizzabili solo agli analisti energetici o agli esperti della pubblica amministrazione, ma non al cittadino. Un esempio di informazione legata ai *cluster* e non visualizzabile dalla *sezione cittadino* è presentato in Figura 3.18. Per ogni sezione creata all'interno dell'applicazione web, sono presenti sia la descrizione di ciò che si sta visualizzando che l'immagine corrispondente, si veda ad esempio la mappa coropletica per l'attributo *rendimento di climatizzazione invernale* mostrata in Figura 3.19.

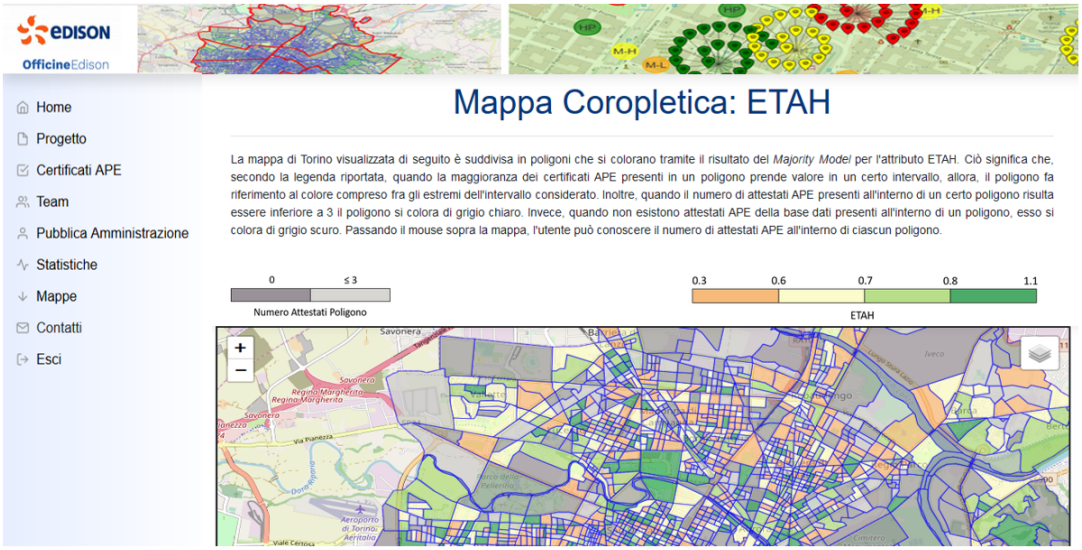

Figura 3.19: Pagina html con la mappa coropletica per l'attributo *rendimento climatizzazione invernale* nella sezione per l'utente pubblica amministrazione.

Come accennato precedentemente, l'applicazione web dispone anche di una *sezione statistiche*, come si nota dalla Figura 3.19 a sinistra, all'interno della quale vengono proposti alcuni grafici statistici e informazioni di base, che possono aiutare gli utenti finali a conoscere il *dataset* dei dati *open* utilizzato durante le analisi. In questo modo, gli utenti possono prepararsi ad una navigazione più mirata e consapevole delle componenti presenti all'interno dell'applicazione web.

3.4 Modelli di Generalizzazione

Il *framework* propone una metodologia per generalizzare la conoscenza estratta dalle certificazioni energetiche con lo scopo di prevedere il comportamento energetico degli edifici che non sono presenti nel *database* a disposizione. Al fine di raggiungere tale obiettivo, vengono gestite due possibili predizioni:

- **Variabili di analisi mancanti**: situazione in cui il nuovo attestato energetico non presenta una delle sette variabili usate per la caratterizzazione della conoscenza estratta, si veda la Tabella 3.8;
- **Etichetta di cluster**: condizione in cui viene gestita l'assenza dell'etichetta di *cluster* per il nuovo attestato energetico. In particolare, si suppone che il nuovo certificato abbia o tutti gli attributi di interesse, oppure, solo le 3 variabili geometriche di analisi.

Il processo di generalizzazione può seguire due percorsi separati, il primo parte dalla gestione di alcune variabili mancanti fino ad arrivare alla predizione dell'etichetta di *cluster* l'altro, invece, considera solo la fase di classificazione dei nuovi edifici tramite l'etichetta di *cluster* predetta. Le tecniche utilizzate per la predizione dell'etichetta di *cluster* costituiscono il cosiddetto *modello semi-completo*, spiegato in dettaglio nella sezione 3.4.1. D'altra parte, siccome lo scopo finale è quello di caratterizzare le performance energetiche di nuovi edifici in base alla Tabella 3.10, la gestione delle variabili di interesse mancanti è fondamentale, in modo tale che poi possa essere eseguita la fase di etichettatura degli attestati. Questa procedura viene formalizzata attraverso l'implementazione del *modello completo*, esposto in dettaglio nella sezione 3.4.2.

Fase Preliminare: Training Vs Test

Per valutare le prestazioni dei modelli di classificazione e regressione implementati dal *framework*, si è confrontata la metodologia *K-fold cross validation* con la *leave-one-out cross validation*. La scelta del K è stata fatta considerando la numerosità del *dataset* (circa 30 000 tuple) e alcune prove sperimentali, da cui si è arrivati a $K = 3$. Per il modello semi-completo, la *leave-one-out* prende il 75% dei certificati come dati di *training* e il restante 25% come dati di *test*, invece, nel modello completo i risultati si ottengono considerando l'80% dei dati per il *training* e il 20% per il *test*. I risultati restituiti da entrambi i metodi si sono dimostrati confrontabili indicando, così, che non esistono partizioni più significative

del *dataset* rispetto ad altre. Di conseguenza, si è deciso di utilizzare la metodologia *leave-one-out cross validation*, essendo meno onerosa a livello computazionale.

Una fase preliminare importante riguarda la normalizzazione dei due sottoinsiemi *training set* e *test set*. In particolare, dopo aver scelto le variabili di input da fornire agli algoritmi, viene applicata la normalizzazione *Z-Score* sui dati di *training*, la quale considera la media e la deviazione standard della percentuale di osservazioni che costituiscono il *training set*. Anche sul *test set* viene attuata la normalizzazione *Z-Score*, ma si usa sempre la media e la deviazione standard dei dati di *training* [62]. Quest'operazione consente di valutare quanto bene il modello, che si genera sul *training set*, riesce ad essere generalizzato sul sottoinsieme di *test*. Infatti, se una particolare osservazione di *test* presenta un valore anomalo rispetto alla distribuzione dei dati di *training* questo valore verrà trattato come tale grazie ad una migliore affidabilità della statistica *R-quadro*.

3.4.1 Implementazione del Modello Semi-Completo

La predizione dell'etichetta di *cluster* di un nuovo edificio, non considerato nelle fasi di analisi, viene attuata in due casi diversi: quando il nuovo attestato energetico presenta le 7 variabili di analisi riportate in Tabella 3.8, oppure quando presenta soltanto quelle geometriche. Poiché la *label* che definisce l'etichetta di un gruppo di edifici è una variabile categorica, il modello semi-completo viene implementato utilizzando un metodo di classificazione, nello specifico l'algoritmo *k-Nearest Neighbors*. Come spiegato nel capitolo precedente, tale algoritmo richiede la normalizzazione delle variabili, di conseguenza si applica la normalizzazione *Z-Score* mediante la metodologia accennata nella fase preliminare. Di seguito si riportano, separatamente, i risultati sperimentali ottenuti.

Caso 1: Tutte le Variabili di Analisi

Nel caso in cui un'osservazione di *test* presenti tutti gli attributi di interesse cioè: *fattore forma, trasmittanza media delle superfici trasparenti, trasmittanza media delle superfici opache, superficie riscaldata, superficie disperdente, anno di costruzione* ed *ETAH*, viene applicata una metodologia a due livelli:

- **Vicinato geografico**: come primo passo si sceglie un certo numero di osservazioni di *training* vicine geograficamente all'osservazione di *test*. Nello specifico, si considerano gli attributi [latitudine, longitudine] e si calcola la distanza euclidea fra l'osservazione di *test* e le osservazioni di *training*;

- **K-NN similarità**: una volta selezionato il vicinato dell'osservazione di *test*, si applica il classificatore *K-Nearest Neighbors* al fine di considerare solo le osservazioni di *training* che risultano effettivamente più vicine all'osservazione di *test*. Quindi si calcola la distanza euclidea tra tutte le variabili di interesse dell'osservazione di *test* e quelle del vicinato geografico.

La predizione dell'etichetta di *cluster* per l'osservazione di *test* si realizza prendendo la *label* maggioritaria all'interno del vicinato selezionato dal K-NN. Al fine di decidere il numero di osservazioni da considerare come vicinato geografico e il parametro K dell'algoritmo K-NN, viene attuata la cosiddetta *grid-search*, la quale prende diversi possibili valori per questi due parametri e misura l'accuratezza del modello K-NN risultante.

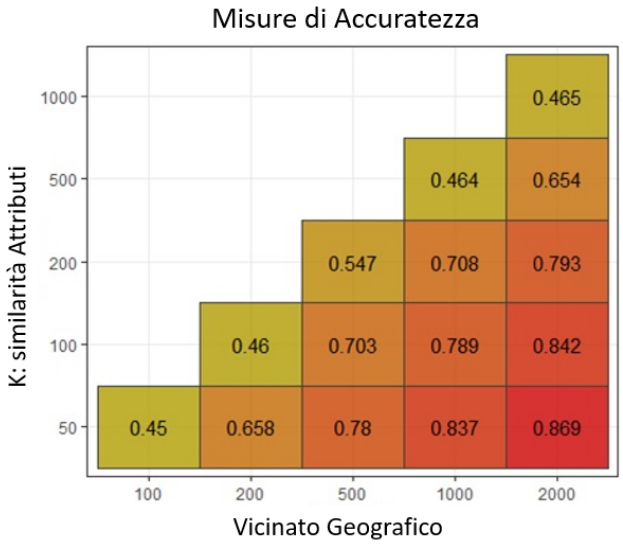

Figura 3.20: *Grid-search* per l'algoritmo K-NN nel modello semi-completo con tutte le variabili di analisi.

Sono riportati in Figura 3.20 tutti i valori di accuratezza ottenuti effettuando le diverse combinazioni. Si può osservare che, scegliendo **1000** osservazioni di *training* come vicinato geografico e $\mathbf{K = 50}$ si ha un valore di accuratezza pari a **0.837**. Questo risultato indica che il modello con questi parametri ha una buona performance e, nonostante il valore più elevato di accuratezza si abbia per un vicinato di 2000 punti, il guadagno che si ottiene è minimo rispetto al valore scelto. Si presentano, inoltre, nella Tabella 3.12 i valori medi di precisione e richiamo e

nella Tabella 3.11 i valori di precisione e richiamo per ogni etichetta di *cluster*. Questi risultati sperimentali confermano che questa metodologia di predizione della *label* si dimostra robusta.

Label	**Precisione**	**Richiamo**
0	0.917	0.576
1	0.951	0.480
2	0.942	0.662
3	0.962	0.481
4	0.897	0.861
5	0.820	0.580
6	0.839	0.950
7	0.787	0.950
8	0.842	0.989
9	0.765	0.954
10	0.963	0.792
11	0.829	0.863

Tabella 3.11: Valori di precisione e richiamo per il modello semi-completo con tutte le variabili di analisi.

Accuratezza	**Precisione Media**	**Richiamo Medio**
0.837	0.876	0.761

Tabella 3.12: Valori di precisione e richiamo medio per il modello semi-completo con tutte le variabili di analisi.

Caso 2: Solo le Variabili Geometriche

D'altra parte, quando un'osservazione di *test* presenta soltanto gli attributi geometrici di analisi cioè: *fattore forma, superficie riscaldata* e *superficie disperdente*, si utilizza comunque la metodologia a due livelli spiegata in precedenza ma, rispetto al **Caso 1**, si hanno a disposizione meno variabili per la predizione dell'etichetta di *cluster*. In particolare, si eseguono le due fasi:

- **Vicinato geografico**: come prima, si sceglie il vicinato geografico costituito dalle osservazioni di *training*. Quindi, si considerano gli attributi [latitudine, longitudine] e si calcola la distanza euclidea fra l'osservazione di *test* e il suo vicinato con le osservazioni di *training*;

- **K-NN similarità**: all'interno del vicinato geografico, viene applicato l'algoritmo *K-Nearest Neighbors* con il concetto di distanza soltanto sulle variabili geometriche a disposizione. Anche in questo caso si selezionano le osservazioni del vicinato che risultano essere più simili all'osservazione di *test*.

Una volta considerate le osservazioni di *training* più simili dal punto di vista delle variabili geometriche, l'algoritmo K-NN assegna all'osservazione di *test* la *label* più frequente presente all'interno di esse. Per poter scegliere i parametri migliori da dare al modello semi-completo si attua *grid-search*, i cui risultati sono riportati in Figura 3.21.

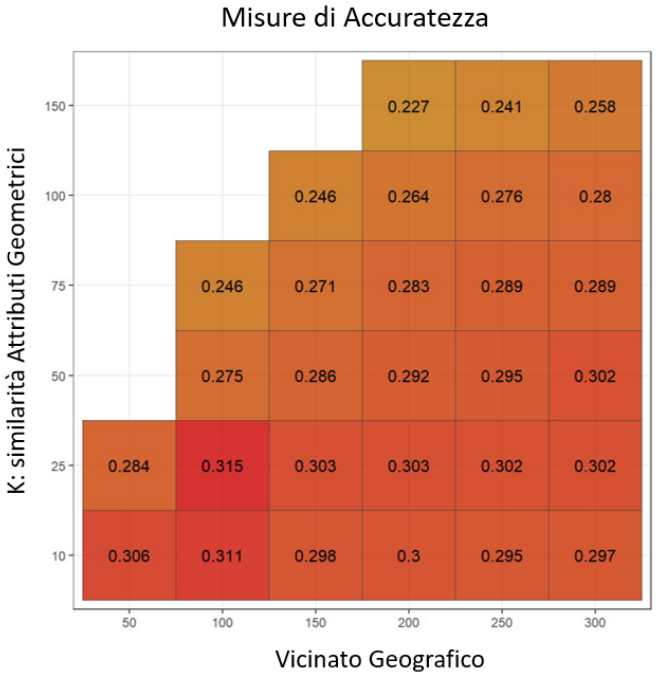

Figura 3.21: *Grid-search* per l'algoritmo K-NN nel modello semi-completo con le sole variabili geometriche.

Nel complesso la Figura 3.21 riporta dei valori di accuratezza molto più bassi rispetto ai valori mostrati nel **Caso 1**, ciò significa che questa procedura risulta essere meno accurata. Questo effetto si spiega con la mancanza delle variabili termofisiche di analisi (*trasmittanza media delle superfici trasparenti, trasmittanza media delle superfici opache, anno di costruzione, ETAH*), senza le quali diventa più difficile predire l'etichetta di *cluster*.

Nonostante questo metodo sia meno robusto, il *framework* propone un metodo per classificare la performance energetica di edifici che presentano nel proprio certificato energetico solo alcune informazioni necessarie alle analisi. Si riportano inoltre la Tabella 3.14 per i valori di precisione e richiamo medi, e la Tabella 3.13 se si considerano questi valori per ogni etichetta di classe. Entrambe le tabelle sono state ottenute scegliendo un numero di vicini geografici di *training* pari a **100** e un numero di vicini simili pari a $\mathbf{K = 25}$ perché, come evidenziato dalla *grid search*, per questi valori si ha il valore più elevato di accuratezza.

Label	**Precisione**	**Richiamo**
0	0.292	0.204
1	0.117	0.051
2	0.319	0.261
3	0.790	0.299
4	0.358	0.415
5	0.497	0.467
6	0.245	0.306
7	0.247	0.185
8	0.274	0.413
9	0.203	0.142
10	0.381	0.564
11	0.397	0.429

Tabella 3.13: Valori di precisione e richiamo per il modello semi-completo con tutte le variabili di analisi.

Accuratezza	**Precisione Media**	**Richiamo Medio**
0.315	0.343	0.311

Tabella 3.14: Valori di precisione e richiamo per il modello semi-completo con tutte le variabili di analisi.

3.4.2 Implementazione del Modello Completo

Il modello completo si occupa della predizione di alcuni attributi numerici di analisi e della successiva predizione dell'etichetta di *cluster*, secondo la procedura spiegata nel paragrafo precedente. In questo caso, le variabili *target* sono quantitative, perciò si applicano i metodi di regressione introdotti nella sezione 2.4.1. Nello specifico, si considera il caso in cui ad un'osservazione di *test* manchi o la variabile

fattore forma, anche detta *rapporto S/V*, oppure l'attributo *rendimento per la climatizzazione invernale*, indicato con *ETAH*. Anche per il modello completo il *framework* prevede la fase preliminare di normalizzazione del *dataset* con la metodologia *Z-Score*, presentata nella sezione 3.4. Per maggiore semplicità espositiva di questo lavoro, vengono presentati separatamente le analisi dei risultati ottenuti.

Caso 1: ETAH mancante

Il *rendimento per la climatizzazione invernale* è ottenuto dal prodotto dei rendimenti dei sottosistemi che caratterizzano gli impianti di riscaldamento (per maggiori dettagli si riveda la sezione 1.3.3):

$$ETAH = ETAG * ETAE * ETAR * ETAD\,. \tag{3.1}$$

Mentre l'*ETAH* indica di una variabile calcolata ai fini delle analisi, i rendimenti scritti nella formula (3.1) sono attributi in genere presenti negli attestati di prestazione energetica. Di conseguenza, il *framework* propone una metodologia per la predizione di *ETAH* nei casi in cui, per un'osservazione di *test*, una fra queste quattro variabili non sia presente.

Le variabili di input utilizzate per "addestrare" i modelli di regressione sono tutte le variabili di analisi riportate in Tabella 3.8, tranne la variabile di risposta *ETAH*. Poiché si suppone che un'osservazione di *test* abbia a disposizione tre rendimenti tranne uno, per ottenere dei buoni risultati, si devono aggiungere anche questi attributi come regressori. La variabile *target* si esprime come il prodotto dei rendimenti perciò, generare un modello di regressione sommando questi attributi insieme alle variabili di analisi, porterebbe ad un modello non performante. Per questo motivo, si aggiunge come ulteriore regressore il prodotto dei rendimenti presenti nell'osservazione di *test*. Quindi se, per esempio, ad un dato di *test* manca l'attributo *ETAE*, viene generato il seguente modello di regressione:

$$ETAH \sim S/V + UW + UO + SR + SD + ANNO + ETAG * ETAR * ETAD\,.$$

I modelli di regressione implementati dal *framework*, spiegati dal punto di vista teorico nella sezione 2.4.1, riguardano gli algoritmi per la regressione *lineare, polinomiale, lasso* e *K-Nearest Neighbors*. Per poter confrontare le performance di questi metodi nella predizione di *ETAH*, sono state attuate due *grid search*: una per la scelta del parametro K del K-NN - risultati riportati in Tabella 3.15 - e l'altra per la selezione del parametro di *tuning* λ che costituisce il termine di penalità nella regressione lasso, si veda la Tabella 3.16.

K	R^2 (ETAD)	R^2 (ETAE)	R^2 (ETAG)	R^2 (ETAR)
1	0.82	0.92	0.55	0.42
3	0.89	0.94	0.68	0.60
5	0.89	0.95	0.72	0.64
8	0.90	0.95	0.74	0.64
10	0.90	0.95	0.74	0.67
≥ 11	0.90	0.74	0.65	0.65

Tabella 3.15: *Grid-search* per la regressione con il K-NN nel modello completo con l'attributo di analisi *ETAH* mancante.

λ	R^2 (ETAD)	R^2 (ETAE)	R^2 (ETAG)	R^2 (ETAR)
0.00001	0.90	0.97	0.74	0.67
0.0001	0.90	0.97	0.74	0.67
0.001	0.91	0.98	0.74	0.67
0.01	0.90	0.97	0.74	0.67
0.1	0.88	0.96	0.72	0.65

Tabella 3.16: *Grid-search* per la regressione Lasso nel modello completo con l'attributo di analisi *ETAH* mancante.

Parametri	ETAD	ETAE	ETAG	ETAR
λ	0.001	0.001	0.001	0.001
K	8	5	8	10

Tabella 3.17: Risultato della *grid-search* nel modello completo con l'attributo di analisi *ETAH* mancante.

Come si evince dalle *grid-search*, si hanno valori di R^2 più elevati per i modelli di regressione in cui le variabili mancanti sono *ETAD* oppure *ETAE*. Tuttavia, i modelli di regressione generati con gli altri rendimenti mancanti possono essere comunque ritenuti accettabili, anche se meno performanti. In particolare, si riporta nella Tabella 3.17 il risultato della *grid-search* per ogni rendimento in esame. Una volta decisi i valori di λ e K, si possono confrontare alcuni modelli di regressione, presentati nella Tabella 3.18 e scegliere quello più performante per ognuno dei rendimenti. Inoltre, poiché *ETAH* dipende in modo cubico dagli altri tre rendimenti con cui vengono generati i modelli di regressione, si applica un algoritmo per la regressione polinomiale utilizzando un polinomio di grado 3. In questo caso, il modello generato è più complesso perché vengono fatte tutte le combinazioni possibili fra le variabili di input al fine di ottenere dei polinomi di grado 3, motivo

per cui in seguito a questo algoritmo viene applicata la lasso, in grado di mettere alcune variabili a zero in modo da semplificare il modello.

Modello	R^2 (ETAD)	R^2 (ETAE)	R^2 (ETAG)	R^2 (ETAR)
Lineare	0.91	0.97	0.73	0.67
Lasso	0.91	0.97	0.72	0.65
Polinomiale + Lasso	0.91	0.97	0.74	0.68
K-NN	0.90	0.95	0.74	0.67

Tabella 3.18: Confronto tra gli algoritmi di regressione nel modello completo con l'attributo di analisi *ETAH* mancante.

Considerando i casi in cui l'osservazione di *test* non possiede *ETAD* oppure *ETAE*, l'algoritmo di regressione scelto dal *framework* si basa sulla metodologia **lasso**. Infatti, a parità dei valori R^2, la regressione lasso è una tecnica che rende i modelli più interpretabili. Per quanto riguarda la mancanza degli altri due rendimenti, quando *ETAR* è assente il *framework* utilizza la funzione *pipeline*[8], implementata in *Python*, la quale permette di definire una **regressione polinomiale seguita da lasso**, mentre quando manca *ETAG* non c'è una distinzione fra l'utilizzo del modello polinomiale di grado 3 insieme a lasso oppure il K-NN. Dal punto di vista sperimentale sono state fatte anche delle prove con il modello SVR tuttavia non ha mai prodotto un risultato perché computazionalmente oneroso.

Caso 2: Rapporto S/V Mancante

L'altra situazione che viene gestita dal *framework* è quella in cui un'osservazione di *test* non presenti l'attributo *fattore forma*, il quale diventa la variabile *target* che dev'essere predetta dal modello di regressione. In questo caso, le variabili di input da considerare per generare i modelli di regressione sono tutte le variabili di analisi riassunte in Tabella 3.8, ad eccezione della variabile *rapporto S/V* che è la variabile di risposta. In generale, viene costruito un modello di questo tipo:

$$S/V \sim UW + UO + SR + SD + ETAH + ANNO\,.$$

Prima di poter scegliere l'algoritmo che ha la migliore performance in termini di R^2, viene attuata la *grid-search* per la scelta dei parametri degli algoritmi di regressione. Il *framework* considera i metodi di regressione *lineare, lasso* e *K-Nearest Neighbors*,

[8] https://scikit-learn.org/stable/modules/generated/sklearn.pipeline.Pipeline

per cui si riporta nella Tabella 3.19 il *tuning* dei parametri λ per la regressione lasso e K per il K-NN. Dai risultati si evince che, in mancanza della variabile *fattore forma*, i parametri migliori sono $\boldsymbol{\lambda = 0.001}$ e $\mathbf{K = 10}$. La Tabella 3.20 riporta invece i valori di R^2 per i diversi modelli testati, la regressione **K-Nearest Neighbors** è quella che dà le prestazioni migliori in questo caso, ed è quindi quella che viene implementata dal *framework*.

λ	R^2 Lasso	K	R^2 K-NN
0.00001	0.75	1	0.78
0.0001	0.75	3	0.84
0.001	0.76	5	0.85
0.01	0.75	8	0.85
0.1	0.70	≥ 10	0.86

Tabella 3.19: *Grid-search* per la regressione Lasso e K-NN nel modello completo con l'attributo di analisi *Rapporto S/V* mancante.

Modello	Valore R^2
Lineare	0.74
Lasso	0.76
K-NN	0.86

Tabella 3.20: Confronto tra gli algoritmi di regressione nel modello completo con l'attributo di analisi *Rapporto S/V* mancante.

Caso 3: Altri Attributi Mancanti

Nei due casi precedenti il *framework* implementa una metodologia di predizione delle variabili *rendimento per la climatizzazione invernale* e *fattore forma*. In questo lavoro di tesi sono stati esaminati anche dei modelli di regressione per predire gli altri attributi di interesse, come ad esempio le *trasmittanze medie delle superfici opache* e *trasparenti*. Tuttavia, nei risultati sperimentali i modelli si sono dimostrati poco performanti. Per questo motivo, non sono stati presi in considerazione dato che una loro predizione rischiava di inserire dati rumori all'interno della base di dati delle certificazioni energetiche.

3.4.3 Certificati Recuperati

I modelli di classificazione e regressione sono stati costruiti per generalizzazione la conoscenza estratta dalle certificazioni energetiche. Ciò significa che, possono

essere adoperati per caratterizzare edifici non presenti all'interno della base di dati in esame. Durante la fase di *Preparazione dei Dati*, sono state escluse alcune tuple dal *database*, tramite la metodologia *domain-driven* spiegata nella sezione 3.1.2. In questo scenario, si è ritenuto opportuno sfruttare i modelli predittivi implementati dal *framework* per provare a recuperare alcuni attestati di prestazione energetica. Infatti, se un record possiede un attributo fuori dai range di ammissibilità, si può pensare che esso abbia il valore dell'attributo mancante e, quindi, che si possa usare uno dei modelli di generalizzazione per la predizione di tale variabile.

Sigla	**Attributo**	**Eliminati (%)**
S/V	Fattore Forma	0.18
UW	Trasmittanza Trasparente	0.22
UO	Trasmittanza Opaca	0.15
ETAD	Rendimento Distribuzione	0.88
ETAE	Rendimento Emissione	0.72
ETAG	Rendimento Generazione	6.31
ETAR	Rendimento Regolazione	0.20

Tabella 3.21: Numero di certificati energetici eliminati dalla base di dati per ogni attributo fuori dal range di ammissibilità.

Facendo riferimento alla Tabella 3.21, esposta nella sezione 3.1.2 e riporta qui per maggiore facilità di lettura, si osserva che, si possono provare a recuperare gli attestati che sono stati esclusi per avere valori di *rapporto S/V*, oppure di uno dei quattro rendimenti *ETAD, ETAE, ETAG, ETAR*, fuori dai range di ammissibilità. In generale, le prestazioni fornite dai modelli di regressione nel **Caso 1**, sono migliori quando mancano i rendimenti *ETAD* oppure *ETAE*. Pertanto, questi modelli sono stati utilizzati per il recupero delle certificazioni che sono state eliminate durante la fase dei *Filtri sulle Variabili*.

Sigla	**Attributo Mancante**	**Variabile Target**	**Recuperati (%)**
S/V	Rapporto S/V	S/V	0.18
ETAD	Rendimento Distribuzione	ETAH	0.84
ETAE	Rendimento Emissione	ETAH	0.67

Tabella 3.22: Numero di certificati energetici recuperati con i modelli di regressione.

La tabella dei certificati eliminati mostra che ci sono circa l'1.78% di record potenzialmente recuperabili. Tuttavia, dai risultati riportati nella Tabella 3.22, si

nota che, in totale, sono stati **recuperati** circa l'1.69% degli attestati energetici. Questa differenza di percentuale, viene spiegata dall'eliminazione di alcuni record che presentavano valori anomali per una o più variabili di analisi. Infatti, tramite l'analisi dei *boxplot* degli attributi di interesse per i certificati inizialmente recuperati, sono stati identificati alcuni possibili *outlier* all'interno della distribuzione dei dati. Come passo finale dell'implementazione del *modello completo*, gli edifici recuperati possono essere caratterizzati secondo la loro performance energetica. Infatti, per la predizione dell'etichetta di *cluster* si utilizzano i modelli di classificazione trattati nella sezione 3.4.1, mentre per la caratterizzazione dell'edificio si fa riferimento alla Tabella 3.10.

Osservazione 3.3. *Fra i modelli di regressione costruiti nel* ***Caso 1****, ci sono anche quelli per la predizione di ETAH con i rendimenti ETAG oppure ETAR mancanti. Però, questi modelli risultano essere meno performanti rispetto agli altri, ragion per cui non sono stati recuperati i certificati eliminati per ETAG ed ETAR fuori dai range di dominio. Tuttavia, se si è disposti ad ammettere un grado di imprecisione considerevole, come nel modello di classificazione costruito per il* ***Caso 2****, il framework consente comunque il loro utilizzo.*

Capitolo 4

Conclusioni e Sviluppi Futuri

L'obiettivo della presente tesi è stato quello di progettare e sviluppare una metodologia per la caratterizzazione energetica degli edifici siti nella città di Torino, in grado di supportare l'analista durante l'analisi di grandi volumi di dati relativi ad edifici geolocalizzati. A tal proposito è stato sviluppato un *framework* di nome TUCANA (TUrin Certificates ANAlysis) che include tutti i blocchi principali per analizzare in maniera automatica, l'efficienza energetica degli edifici residenziali presenti nella città di Torino, attraverso tecniche di *data mining*.

Grazie alla la fase di pulizia, sono stati estratti, uniformati e resi affidabili, dati provenienti da diverse fonti *open*, per ottenere un *dataset* il più possibile integro ed utilizzabile per analisi future.La fase di analisi ha permesso di estrarre la conoscenza e presentarla poi, in maniera semplificata, ai vari *stakeholder* attraverso la visualizzazione su diverse mappe interattive. Il processo di analisi include una fase automatica di *clustering* per la suddivisione degli edifici in gruppi omogenei con caratteristiche termofisiche simili. Inoltre, il *framework* sviluppato garantisce un valido supporto allanalista nel *decision-making*, attraverso la rappresentazione dei dati con mappe geolocalizzate, consentendogli di analizzare velocemente insiemi di dati più interessanti, rilevandone gli aspetti più significativi.

Successivamente, la visualizzazione è stata resa fruibile ai diversi *stakeholder* attraverso lo sviluppo di una applicazione web, fornendo un insieme di strumenti grafici che consentono di fruire facilmente della conoscenza nascosta all'interno dei dati a diversi livelli di dettaglio. Questa conoscenza può essere infatti utilizzata dai differenti attori in maniera diversa, adeguandola alle proprie esigenze.

1. *Cittadini Privati*: potrebbero essere interessati all'analisi delle prestazioni energetiche di edifici siti in una particolare area della città, così come ai diversi attributi che caratterizzano la geometria degli edifici, in modo da

poter attuare una scelta mirata sulle caratteristiche più rilevanti su dove comprare o affittare casa;

2. *Pubblica Amministrazione*: potrebbe essere interessata nell'identificare zone scarsamente performanti dal punto di vista energetico ed attuare degli interventi migliorativi, aumentando il benessere della cittadinanza;

3. *Esperti di Dominio*: gli esperti di dominio e gli analisti energetici potrebbero utilizzare TUCANA per caratterizzare, sia con tecniche *supervised* sia con tecniche *unsupervised*, gruppi di edifici con caratteristiche comuni, in modo da effettuare analisi comparative.

4.1 Applicazione: Nuovi APE di Torino

Per rendere generalizzabile la conoscenza estratta, TUCANA include un blocco di analisi *Machine Learning*. L'obiettivo principe della generalizzazione è relativo alla creazione di un modello che includa un apprendimento automatico, ovvero che sia in grado di generalizzare dalla propria esperienza estraendone ragionamenti induttivi. In questo contesto, per generalizzazione, si intende l'abilità del *framework* di supportare l'analista nell'etichettare nuovi edifici in maniera accurata, dopo aver fatto esperienza su un insieme di dati di apprendimento relativi alle certificazioni energetiche raccolte dalla città di Torino.

Un possibile sviluppo futuro riguarda la raccolta e l'etichettatura, tramite le procedure implementate nel *framework*, di nuovi edifici residenziali della città di Torino che non sono presenti nella base di dati. Si possono inoltre, pensare a delle possibili estensioni per la parte di generalizzazione. Per esempio, l'integrazione in TUCANA di nuovi modelli di generalizzazione possono aumentare l'accuratezza dei modelli proposti e, di conseguenza, migliorare la determinazione delle prestazioni energetiche dei nuovi edifici.

4.2 Estensione: Nuovo Caso di Studio

Sebbene i risultati ottenuti in questo lavoro di tesi siano soddisfacenti, si possono definire diversi sviluppi futuri. In particolare, possibili estensioni riguardano l'integrazione in TUCANA di nuovi casi di studio. Facendo riferimento alla Figura 4.1, si osserva che i dati raccolti dal CSI-Piemonte comprendono altre province della regione Piemonte e altri edifici con destinazioni d'uso diverse. In questo contesto, si può pensare di:

- Estendere la metodologia proposta anche ad edifici con altre tipologie di destinazioni d'uso, e non solo a quelli residenziali, e caratterizzare l'intera regione Piemonte;
- Integrare i dati del catasto con i dati degli impianti, così da valutare altre componenti determinanti per valutare l'efficienza energetica;
- Valutare altri algoritmi di *clustering*, come quelli gerarchici, potrebbe permettere un'ulteriore caratterizzazione degli edifici, soprattutto se le finalità di utilizzo sono differenti dalla sola residenza a carattere continuativo;
- Estendere la metodologia anche ad altre città per aumentare la base di conoscenza estratta e analizzare le performance delle diverse zone d'Italia;
- Progettare di integrare la metodologia proposta anche in altri domini di dati *open* (ad esempio inquinamento, condizioni meteorologiche, mobilità intelligente) verso città più trasparenti e comprensibili dove i cittadini detengono il ruolo chiave ed integrarli nell'applicazione web sviluppata.

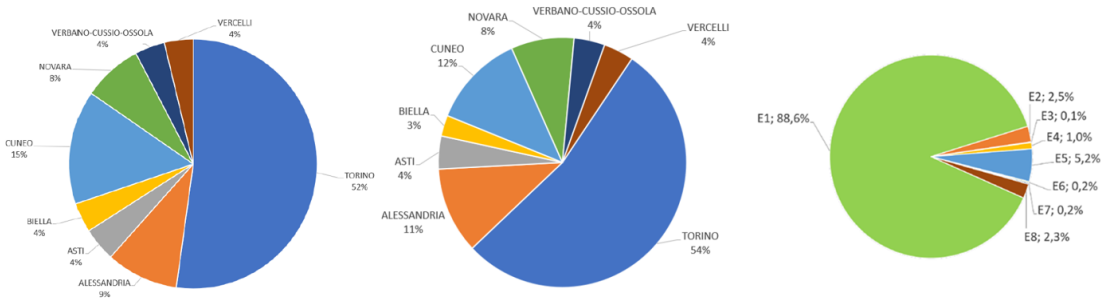

Figura 4.1: Distribuzione delle province e delle destinazioni d'uso degli *open data* nella Regione Piemonte.

Bibliografia

[1] Consiglio dell'Unione Europea, *Affrontare i cambiamenti climatici nell'UE*, https://www.consilium.europa.eu/it/policies/climate-change/.

[2] Gazzetta Ufficiale dell'Unione Europea, *Regolamento 2018/1999 del Parlamento Europeo e del Consiglio*, 11 dicembre 2018.

[3] Capozzoli, et al. *Data mining for energy analysis of a large data set of flats*, 2015, pp. 1–16.

[4] Perez-Lombard L, Ortiz J and Pout C, Universidad de Sevilla, *A review on building energy consumption information*, Energy and Buildings 2008, 40(3): 394-396.

[5] Ren, X., Yan, D. and Hong, T, *Data Mining of Space Heating System Performance in Affordable Housing*, Building and Environment, 89: 1-13, 2015.

[6] Segretario generale del Consiglio europeo, *Conclusioni sull'Unione dell'energia*, Bruxelles, 20 marzo 2015.

[7] Dott. per. ind. Pietro Rotiroti, *La normativa vigente in materia di Certificazione Energetica in Italia.*

[8] Biblus-net, *Certificazione Energetica: Quadro Normativo e Glossario*, http://biblus.acca.it/focus/attestato-prestazione-energetica-ape/.

[9] Blog Apros, *Attestato di Prestazione Energetica: Quadro Normativo*, http://blog.apros.it/2017/04/03/attestato-prestazione-energetica-parte-1-quadro-normativo/.

[10] Comitato Termotecnico Italiano, *Efficienza e Attestazione della Prestazione Energetica degli edifici in Italia*, https://www.cti2000.eu/.

[11] Infrastrutture Lombarde S.p.A., *Attestato di Prestazione Energetica*, http://www.cened.it/cose.

[12] Andrea Denza, *Come cambia l'APE dal 1 Ottobre 2015 con il nuovo D.m. 26/06/2015*, https://www.certificato-energetico.it/news/nuovo-ape-2015.html

[13] Redazione QualEnergia.it, *Edilizia efficiente, in vigore da lunedì 9 luglio 2018 la direttiva Ue*, https://www.qualenergia.it/articoli/edilizia-efficiente-in-vigore-da-lunedi-la-direttiva/.

[14] Sistema Piemonte, *Ambiente e Energia*, http://www.sistemapiemonte.it.

[15] *Delibera della Giunta Regionale del 21 settembre 2015 n. 14-2119*, https://www.cti2000.eu.

[16] Avv. Alessandro Gallucci, *Foglio, Particella e Subalterno*, https://www.condominioweb.com/dati-identificativi-catastali.13620.

[17] *Decreto del Presidente della Repubblica 26 agosto 1993 n. 412*, http://www.gazzettaufficiale.it/eli/id/1993/10/14/093G0451/sg.

[18] Settore Sviluppo energetico sostenibile, *Attestato di Prestazione Energetica - Vademecum per i Cittadini*, Regione Piemonte.

[19] Assotermica, *Impianti Termici - Concetti Innovativi della Normativa Vigente*, Federazione delle Associazioni Nazionali dell'Industria Meccanica Varia e Affine.

[20] Usama Fayyad, Gregory Piatetsky-Shapiro and Padharaic Smyth, *Knowledge Discovery and Data Mining: Towards a Unifying Framework*. In: (1996), pp. 82– 88, https://www.aaai.org/Papers/KDD/1996/KDD96-014.pdf.

[21] Vladimir I Levenshtein, *Binary codes capable of correcting deletions, insertions and reversal*. In: Soviet Physics-Doklady (1966), pp. 707–710.

[22] Tania Cerquitelli, Evelina Di Corso, Stefano Proto, Alfonso Capozzoli, Fabio Bellotti, Maria G. Cassese, Elena Baralis, Marco Mellia, Silvia Casagrande and Martina Tamburini, *Exploring Energy Performance Certificates through Visualization*. In: Workshop Proceedings of the EDBT/ICDT 2019.

[23] G. S. Maddale, *Outliers*, Second Edition, Macmillan Publishing Company, New York: MacMillan. pp. 88–96, 1992. isbn: 0-02-374545-2.

[24] J. E. Seem, *Using Intelligent Data Analysis to Detect Abnormal Energy Consumption in Buildings*, Energy and buildings, 39(1):52–58, 2007.

[25] Bernard Rosner, *Percentage points for a generalized ESD many-outlier procedure*. In: Technometrics 25 (1983), pp. 165–172.

[26] Tania Cerquitelli, Evelina Di Corso, Stefano Proto, Alfonso Capozzoli, Daniele Mazzarelli, Andrea Nasso, Elena Baralis, Marco Mellia, Silvia Casagrande, Martina Tamburini, *Visualising High-Resolution Energy Maps through the Exploratory Analysis of Energy Performance Certificates*.

[27] Sheldon M. Ross, *Probabilità e Statistica per l'Ingegneria e le Scienze*, Maggioli Editore, 2015.

[28] Martin Ester et al, *A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise*. In: KDD (1996).

[29] Gareth James, Daniela Witten, Trevor Hastiea and Robert Tibshirani, *An Introduction to Statistical Learning with Applications in R*, Springer 2014.

[30] Jiawi Han, Micheline Kamber and Jian Pei, *Data Mining: Concepts and Techniques*, Morgan Kaufmann Publishers, 2012.

[31] Pang-Ning Tan, Michael Steinbach, Anuj Karpatne and Vipin Kumar, *Introduction to Data Mining*, McGraw Hill, Second Edition, 2006.

[32] Susi Dulli, Sara Furini e Edmondo Peron, *Data mining, Metodi e Strategie*, Springer, 2009.

[33] Jin X. and Han J., *K-Medoids Clustering*, Encyclopedia of Machine Learning and Data Mining, Springer 2017. In: Sammut C., Webb G.I. (eds).

[34] Jiawei Han, Micheline Kamber and Jian Pei, *Data Mining: Concepts and Techniques*, Elsevier, 2012.

[35] Purnima Bholowalia and Arvind Kumar, *EBK-Means: A Clustering Technique based on Elbow Method and K-Means in WSN*. In: International Journal of Computer Applications 105 (2014), pp. 17–24.

[36] Ville Satopaa, Jeannie Albrecht, David Irwin and Barath Raghavan, *Finding a "Kneedle" in a Haystack: Detecting Knee Points in System Behavior*, 31st International Conference on Distributed Computing Systems, 2011.

[37] John W. Tukey. *Box-and-whisker plots*. In: Exploratory data analysis (1977), pp. 39–43.

[38] Leo Breiman, *Classification and Regression Trees*, Routledge, 2017.

[39] Roger J. Lewis, *An Introduction to Classification and Regression Tree (CART) Analysis*. In: Annual Meeting of the Society for Academic Emergency Medicine (2000), San Francisco, California.

[40] M. Joan Saary, *Radar plots: A Useful Way for Presenting Multivariate Health Care Data*, Journal of Clinical Epidemiology, Elsevier, 2008.

[41] Michael Friendly, *Milestones in the History of Thematic Cartography, Statistical Graphics, and Data Visualization*, In: yorku.ca record (2009), pp. 1–79.

[42] EasyTask (Programmare con facilità), *ANDROID: Google Maps con Aggregatori (Cluster) di Posizione*, Aprile 2018, http://easytask.altervista.org/android-google-maps-con-aggregatori-cluster-di-posizione/.

[43] Trevor Hastie, Robert Tibshirani and Jerome Friedman, *The Elements of Statistical Learning. Data Mining, Inference, and Prediction*, Springer Second Edition, 2008.

[44] Raymond H. Myers, *Classical and Modern Regression with Applications*, Second Edition, Duxbury 1990.

[45] Douglas C. Montgomery, Elizabeth A. Peck and G. Geoffrey Vining, *Introduction to Linear Regression Analysis*, Wiley 2012.

[46] Giovanni Monegato, *Metodi e Algoritmi per il Calcolo Numerico*, CLUT 2008.

[47] Alfio Quarteroni, *Matematica Numerica, Esercizi, Laboratori e Progetti*, Seconda edizione, Springer 2013, pp. 199-252.

[48] T. Hastie and R. Tibshirani, *Generalized Additive Models: Some Applications*, Journal of the American Statistical Association, Taylor & Francis 1987.

[49] Mary Ercolini, *Introduzione alle Applicazioni Web*, Università di parma.

[50] Robert Tibshirani, *Regression Shrinkage and Selection Via the Lasso*, Royal Statistical Society 1996.

[51] Donald W. Marquardt and Ronald D. Snee, *Ridge Regression in Practice*, Taylor & Francis 1975.

[52] Rui Li and Guan Gong, *K-Nearest-Neighbour Non-Parametric Estimation of Regression Functions in the Presence of Irrelevant Variables*, The Econometrics Journal (2008), pp. 396-408.

[53] M. Brown, K.M. Bossley, D.J. Mills and C.J. Harris, *High Dimensional Neurofuzzy Systems: Overcoming the Curse of Dimensionality*, IEEE 1995.

[54] Debasish Basak, Srimanta Pal and Dipak Chandra Patranabis, *Support Vector Regression*, Department of Instrumentation and Electronics Engineering, Heritage Institute of Technology, Kolkata 2007.

[55] Jason Brownlee, *What is the Difference Between Test and Validation Datasets?*, Machine Learning Mastery 2017, https://machinelearningmastery.com/difference-test-validation-datasets/.

[56] Ronald A. Fisher, *Annals of Eugenics*, The Use of Multiple Measurements in Taxonomic Problems, 7(2):179–188, 1936.

[57] Jeremy Miles, *R Squared, Adjusted R Squared*, John Wiley & Sons, Encyclopedia of Statistics in Behavioral Science 2005.

[58] Charu C. Aggarwal, *Data Classification: Algorithms and Applications*, Taylor & Francis Group 2015.

[59] Leif E. Peterson, *K-Nearest Neighbor*, Scholarpedia 2009.

[60] K. Ming Leung, *Naive Bayesian Classifier*, Polytechnic University 2007, http://cis.poly.edu/ mleung/FRE7851/f07/naiveBayesianClassifier.pdf.

[61] Area Edilizia Residenziale Pubblica, *Osservatorio Condizione Abitativa - XIV Rapporto Anno 2017*, Città di Torino.

[62] Chris Albon, *Python Machine Learning Cookbook*, O'Reilly Media 2017.

[63] Guido Rossum, *Python Reference Manual*, Centre for Mathematics and Computer Science, Amsterdam 1995.

[64] W. N. Venables, D. M. Smith and the R Core Team, *R Language Definition*, Spring-Verlag 2000.

[65] Claudio Davide Ferrara, *PyCharm: Python IDE per i professionisti*, 19 marzo 2018, https://www.html.it/19/03/2018/pycharm-python-ide-per-i-professionisti/.

[66] R-Tools Technology Inc 2019, *R-Studio, Users Manual*.

[67] *Guida HTML5*, https://www.html.it/guide/guida-html5/.

[68] *Guida CSS3*, https://www.html.it/guide/guida-css3/.

[69] David Flanagan, *JavaScript: The Definitive Guide*, O'Reilly 2006.

[70] Miguel Grinberg, *Flask Web Development: developing web applications with Python*, O'Reilly 2018.

[71] Rajan Sambandam, *Cluster Analysis Gets Complicated*, American Marketing Association 2003, https://www.trchome.com/docs/5-cluster-analysis-gets-complicated/file.

[72] *Fattori di conversione in energia primaria dell'energia termica fornita ai punti di consegna della rete di teleriscaldamento della rete di Torino*, 3 dicembre 2018, https://www.gruppoiren.it/documents/21402/69847/PEF+2018+-+Torino.pdf/4bc7fbb3-2748-4319-bfa3-44530f3559ba.