ALMA MATER STUDIORUM – UNIVERSITÀ DI BOLOGNA
CAMPUS CESENA
SCUOLA DI INGEGNERIA E ARCHITETTURA

CORSO DI LAUREA MAGISTRALE IN INGEGNERIA E SCIENZE INFORMATICHE

TITOLO DELLA TESI

SPERIMENTAZIONE DI METODI PER LA PREVISIONE DELLA DOMANDA TERMICA IN APPLICAZIONI LEGATE AL TELERISCALDAMENTO

Tesi in

RICERCA OPERATIVA L-A

Relatore

Prof. Daniele Vigo

Presentato da

Claudia Capelli

Correlatori

Angelo Gordini

Matteo Pozzi

Sessione III

Anno Accademico 2013-2014

Indice

Introduzione

Abraham Lincoln diceva: "*Se potessimo sapere a priori dove siamo e dove stiamo tendendo, potremmo giudicare meglio cosa fare e come farlo*". Ecco il perchè del forecasting. Da sempre desiderio recondito dell'essere umano, prevedere il futuro si rivela di grande importanza nell'ottica, per esempio, di un'efficiente pianificazione aziendale.

Il vantaggio competitivo di un azienda si misura anche nella capacità di simulare situazioni future, interpretarle ed essere in grado di organizzarsi di conseguenza. Metodi statistici e avanzati strumenti di analytics fanno sì che tutto possa essere oggetto di forecasting. Alla base di un'analisi previsionale c'è un principio semplice: utilizzare le informazioni del passato come strumento per la simulazione del futuro. Più nel dettaglio, si assume che il mondo in cui le situazioni si sono evolute fino ad oggi continui anche nel domani, e si cercano particolari pattern nei dati storici, con l'obiettivo di proiettarli avanti nel tempo.

In letteratura è presente un elevato numero di articoli e saggi che analizzano in modo approfondito i diversi aspetti di questa disciplina. Attraverso la spiegazione dettagliata di alcune tecniche statistiche che guidano la formulazione delle previsioni, si mette in luce quanto sia indispensabile una comprensione preliminare di queste. Per garantire un buon processo previsionale, infatti, è necessario che non

sia implementato un sistema di tipo "*black box*", sistema di cui gli addetti ai lavori si trovano a lavorare con numeri di cui non conoscono l'origine, generati da un programma di cui non conoscono il funzionamento.

Molto vicino all'analisi statistica, il data mining si distingue per il massiccio utilizzo di più tecniche di apprendimento computerizzate per la generazione ed il confronto di modelli, al fine di identificare eventuali relazioni, trend e pattern presenti nei dati stessi. Il data mining, infatti, generalmente utilizza grandi quantità di dati che, nell'area del business e del mercato, è conosciuto anche come "Big Data". Tale peculiare approccio di analisi non è tuttavia sinonimo di magia: l'induzione di un modello o di una relazione partendo dai dati, comporta un risultato che è anche legato alla qualità dei dati stessi. L'analista non solo deve padroneggiare le tecniche e gli algoritmi che la disciplina mette a disposizione, ma deve anche possedere una profonda conoscenza del fenomeno che ha generato i dati oggetto dell'analisi. Gli specialisti del settore affermano che due sono le chiavi del successo del data mining: la precisa formulazione del problema da analizzare e l'utilizzo di dati "buoni".

Proprio in quest'ottica si inquadra il seguente lavoro: partendo dalla previsione basata sullo storico di dati termici per stimare il fabbisogno della rete di teleriscaldamento urbano, si è operata una sperimentazione di metodi, capaci di garantire un miglioramento dei risultati.

Nel primo capitolo verrà trattata la formulazione del data mining, dalle tecniche su cui si basa, fino al processo di scoperta della conoscenza, per assicurarne l'efficace applicazione. Nei capitoli successivi l'attenzione si sposterà sugli aspetti di natura predittiva; in particolare, nel secondo capitolo, verrà illustrato

l'approccio tradizionale al tema della previsione della domanda, presentando le principali tecniche statistiche. Saranno spiegati alcuni metodi quantitativi (basati sulle serie storiche e causali) e qualitativi.

Nel terzo capitolo sarà approfondito lo stato del modello presente nel contesto aziendale, mostrando una descrizione degli applicativi di cui l'azienda fa uso. Nel capitolo successivo, si forniranno le analisi compiute per poter apportare soluzioni di ottimizzazione.

Infine, nel capitolo conclusivo, saranno illustrati gli obiettivi perseguiti nell'analisi del processo previsionale, e sarà mostrato che quanto detto nei capitoli precedenti può garantire, non solo un estensione del modello esistente ma anche nuovi spunti d'analisi, per raggiungere livelli prestazionali sempre più elevati.

Capitolo 1

Modelli predittivi nel data mining

1.1 Introduzione all'analisi dei dati aziendali

Negli ultimi anni, le organizzazioni hanno effettuato investimenti significativi per migliorare la loro capacità di raccogliere dati. La quantità dei dati disponibile esplode, mentre il numero di scienziati, informatici, tecnici in grado di analizzarli rimane costante. Questo è il *Data Gap* definito da Grossmann. (*figura*)

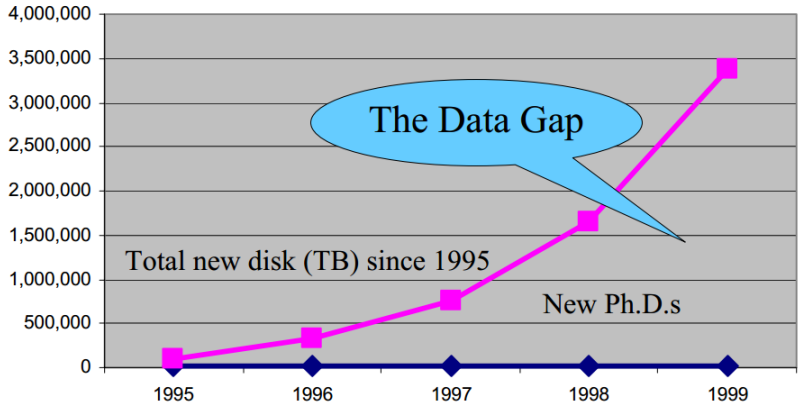


Figura 1.1: Fonte: R. Grossman, C. Kamath, V. Kumar - Data Mining for Scientific and Engineering Applications

Nell'era di Internet, delle smart cities e dei sensori, un'immensa mole di dati circola dunque nel *cyberspazio* e va ad aumentare la densità di quella che è stata definita *infosfera*. Il continuo incremento dell'utilizzo dei sistemi informativi e procedure automatizzate, insieme al carico di *Big Data* che portano con sè, sono il futuro: vanno capiti e gestiti. Le nuove opportunità di condivisione di informazioni, legate alla digitalizzazione e all'evoluzione dei sistemi informativi, sono in aumento e in costante progresso, così come lo scenario competitivo che osserviamo oggi nei mercati. La velocità di cambiamento e la complessità che lo caratterizzano, obbligano sempre più i dirigenti delle aziende a prendere decisioni in maniera molto rapida.

1.2 Big Data e Business Analytics

La spiegazione del termine Big Data può partire dalla definizione della Gartner Inc [1]:

Big data is high-volume, high-velocity and high-variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making.

Si parla dunque di Big Data quando si ha un dataset di grande volume da richiedere strumenti non convenzionali per estrapolare, gestire e processare informazioni intuitive e di ottimizzazione, entro un tempo ragionevole. Basti pensare che fino a non molto tempo fa, le organizzazioni che desideravano condurre un'analisi dei dati erano limitate alla quantità di informazioni contenute in un floppy disk. Oggi, al contrario, grazie a questo nuovo modo di analizzare grandi quantità di informazioni di qualità, strutturate (transazioni, log file), e non (e-mail, immagini), è

possibile comprendere in modo più approfondito e rapido le dinamiche del proprio mercato di riferimento, anticipando, per esempio, alcune decisioni di marketing con un vantaggio competitivo rispetto ai concorrenti.

In questo contesto, i *Big Data* rientrano a pieno titolo nelle dieci tecnologie strategiche di grande rilevanza in questi anni. Sono infatti utilizzati da aziende di medio-grandi dimensioni che, grazie anche all'integrazione di servizi di cloud computing, iniziano ad apprezzarne i benefici e l'utilità.

Questa nuova ed enorme quantità di dati che il mondo aziendale si trova a dover gestire, costituisce la parte principale della *Business Analytics*. Tale metodo di supporto alle decisioni aziendali, consente di studiare analiticamente la straordinaria mole di dati che ogni azienda produce, segnando una svolta rispetto al passato dove l'istinto e l'esperienza erano le guide migliori per ogni dirigente e imprenditore.

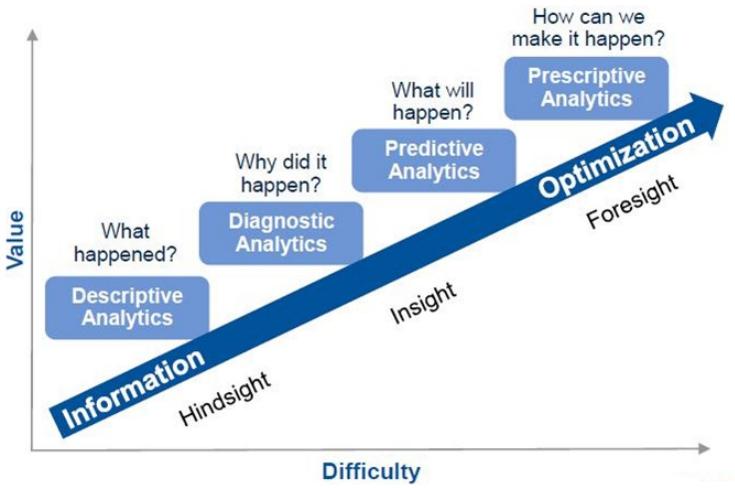


Figura 1.2: Fonte Gartner: Analytics Moves To The Core

Tra le svariate tipologie di approccio che aiutano ad esplorare i dati si utilizzano i 4 approcci illustrati: descrittivo, esplorativo, predittivo e prescrittivo. Del tipo

predittivo, in particolare, se ne parlerà in modo approfondito nel secondo capitolo.

1.3 Data Mining

Per assistere in modo intelligente e automatico gli utenti decisionali, accresce dunque l'esigenza di tecniche e strumenti che sappiano estrarre gli elementi di conoscenza dei dati.

A questo proposito, il Data Mining si propone come disciplina principe per l'analisi dei dati: essa raccoglie tecniche e algoritmi sviluppati nell'ambito della statistica classica e dell'intelligenza artificiale, al fine di offrire all'analista gli strumenti più all'avanguardia.

In un contesto così ampio, può risultare difficile definire in maniera precisa un'area in continua evoluzione, quale è sicuramente quella del data mining. Un fattore, che certo non aiuta l'individuazione di un approccio univoco, è il vasto campo di applicazione: si spazia dal marketing alle diverse tipologie di business, dalla medicina alle applicazioni biomediche e biologiche, dalla ricerca mirata all'individuazione di truffe al text mining e web mining. Così, in assenza di una definizione universalmente accettata, si considerano diverse definizioni. La prima tratta da Marcel Holshemier & Arno Siebes (1994) [2]:

Data mining is the search for relationships and global patterns that exist in large databases but are 'hidden' among the vast amount of data, such as a relationship between patient data and their medical diagnosis. These relationships represent valuable knowledge about the database and the objects in the database and, if the database is a faithful mirror, of the real world registered by the database.

Nella seconda, invece, l'analogia con il processo di estrazione è descritta dalla Guida per l'utente di SPSS Clementine come:

Data mining refers to "using a variety of techniques to identify nuggets of information or decision-making knowledge in bodies of data, and extracting these in such a way that they can be put to use in the areas such as decision support, prediction, forecasting and estimation. The data is often voluminous, but as it stands of low value as no direct use can be made of it; it is the hidden information in the data that is useful."

A parte la pittoresca azione, dedotta dal nome, di setacciare i dati al fine di trovare l'oro, il data mining identifica l'insieme di tecniche e metodologie che hanno per oggetto l'estrazione di un sapere o di una conoscenza, a partire da grandi quantità di dati, e l'utilizzo scientifico, industiale o operativo di questo sapere per prendere cruciali decisioni di business [3].

Contrariamente a quanto si potrebbe essere portati a credere, il data mining è una disciplina molto più evoluzionaria, che non rivoluzionaria. I vari filoni che portano su questa strada iniziano negli anni '60 con gli studi di Frank Rosenblatt sul machine learning: l'obiettivo che lo scienziato si poneva, era di fare in modo che un computer, partendo da un certo numero di situazioni conosciute, potesse sviluppare un insieme di regole sottostanti, universalmente vere. Egli sviluppò "*Perceptron*" [4], predecessore delle attuali reti neurali, su cui si riversarono grandi aspettative, ma che si rivelò un fallimento. Le tecniche ed i miglioramenti che ne seguirono confluirono nel 1969 nelle ricerche sulla *Knoledge Discovery* fino ad arrivare negli anni '80 con la creazione di nuove e più complesse reti neurali, applicate in nuovi campi come quello della valutazione dei risultati nelle campa-

gne di marketing. Nel 1993 G. Piatetsky-Shapiro, C.J.Matheus e P.K.Chan [5] identificarono cinque passi nel processo per il *Knowledge Discovery in Databases* (*figura*), che rese popolare il termine ”data mining” come analisi matematica eseguita su database di grandi dimensioni.

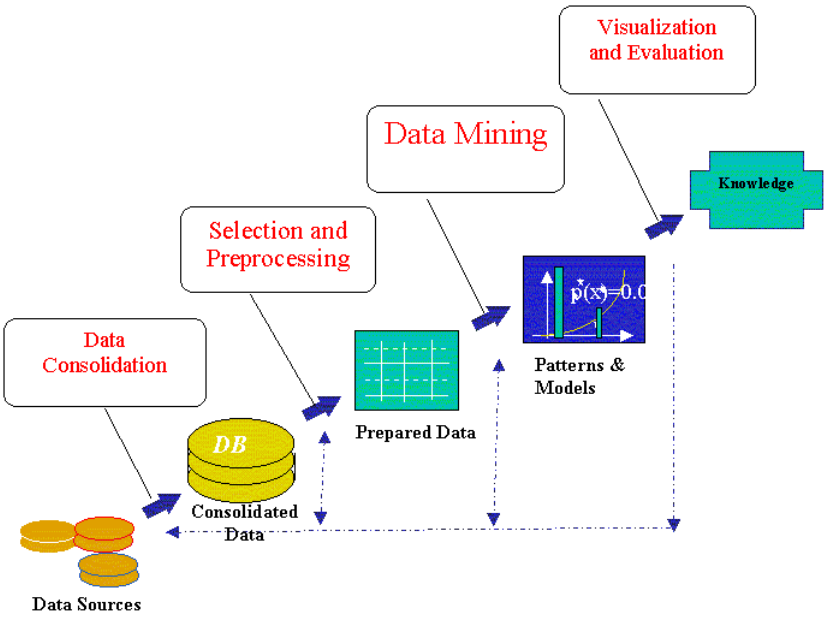


Figura 1.3: Data Mining, cuore del processo di Knowledge Discovery in Databases (KDD)

Il data mining è un peculiare passo in questo processo. Esso, insieme agli altri step del processo di KDD, come ad esempio la preparazione, la selezione, la pulizia dei dati, l’incorporamento di conoscenza già acquisita, l’interpretazione dei risultati, assicurano e garantiscono che la conoscenza estratta sia effettivamente valida.

Una volta focalizzato il punto di integrazione tra il data mining e altre procedure, l’attenzione vuole essere riportata sulla peculiarità dell’approccio d’analisi proprio del data mining.

Oggi, infatti, il data mining ha una duplice valenza [3]:

- estrazione complessa di informazioni implicite, precedentemente sconosciute e potenzialmente utili dai dati;
- esplorazione e analisi, per mezzo di sistemi automatici e semi-automatici, di grandi quantità di dati allo scopo di scoprire pattern significativi.

In entrambe le definizioni i concetti di informazione e di significato sono legati strettamente al dominio applicativo in cui si esegue il data mining.

Tuttavia non risulta semplice individuare la differenza tra il data mining e le altre tecniche di analisi dei dati. In generale, possiamo dire che quando si conoscono i confini ed i contenuti approssimativi di ciò che stiamo cercando, non abbiamo a che fare con un problema di data mining. Esso esce dal range di tecniche tradizionali al fine di scoprire regolarità non note o ignorate precedentemente. Molto vicino all'analisi statistica, il data mining si distingue per il massiccio utilizzo di più tecniche per la generazione e il confronto di modelli, al fine di identificare eventuali relazioni, trend e pattern presenti nei dati stessi.

1.4 CRISP-DM: ciclo di vita di un progetto

Un progetto di data mining richiede un approccio strutturato in cui la scelta del miglior algoritmo è solo uno dei fattori di successo.

Non è possibile sperare di utilizzare un algoritmo di data mining direttamente sui dati e pretendere di ottenere risultati significativi. La scoperta di conoscenza è un processo che richiede un certo numero di passi necessari per assicurare l'efficace applicazione del data mining.

Secondo il modello CRISP-DM (*Cross Industry Standard Process for Data Mining*) [6], il ciclo di vita di un progetto di data mining è articolato in sei fasi non strettamente rigide e talvolta ricorsive, come si illustra in figura.

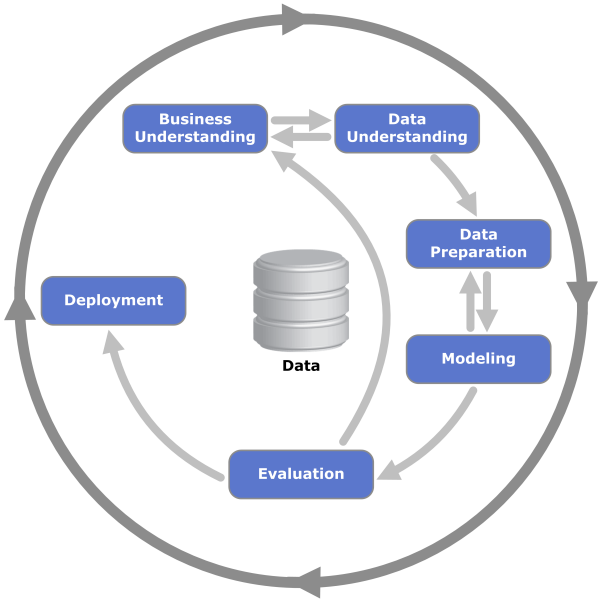


Figura 1.4: Le fasi definite dal CRISP-DM

1. Comprensione del dominio applicativo
2. Analisi esplorativa dei dati
3. Preparazione dei dati come input del modello
4. Sviluppo del modello
5. Valutazione del modello
6. Uso del modello e delle informazioni trovate

Per sottolineare l'importanza di un approccio sistematico a prescindere dal tipo di prodotto di analisi o dalla bontà degli algoritmi a disposizione, vengono brevemente descritte di seguito le varie fasi del modello.

1. *Comprensione del dominio applicativo*: il prerequisito fondamentale per la ricerca di una nuova conoscenza è la comprensione delle informazioni in possesso. Senza un'adeguata chiarezza del problema, nessun algoritmo per quanto sofisticato possa essere può restituire un buon risultato: il rischio di un errata formulazione dell'obiettivo impedisce la scelta del modello corretto, genera leggerezza nella preparazione dei dati e difficoltà dell'interpretare i risultati finali. Si deve pertanto tradurre il problema dell'utente in un problema di data mining, definendo un primo piano di progetto.

2. *Analisi esplorativa dei dati*: gli specialisti del settore ritengono che il tempo e le risorse per la preparazione delle informazioni di input di data mining dovrebbero incidere dal 50% al 90% rispetto all'intera analisi. Tale raccolta preliminare dei dati è pertanto finalizzata a identificare le caratteristiche salienti del modello.

3. *Preparazione dei dati come input del modello*: in tale fase si procede alla selezione delle variabili da utilizzare nel modello (attributi e record). Comprende anche le attività di trasformazione e pulizia dei dati, necessarie per creare il dataset finale.

4. *Sviluppo del modello*: tale fase è iterativa dal momento che il modello finale corrisponde a quello che risolve in maniera migliore il problema di analisi rispetto a un numero di alternative. Diverse tecniche di data mining possono dunque essere applicate al dataset, anche con parametri diversi, al fine di individuare quella che permette di costruire il modello più accurato.

5. *Valutazione del modello*: una volta costruito il modello, è necessario valutare l'effettiva bontà dei risultati ottenuti, al fine di verificare che rispondano adeguatamente agli obiettivi dell'utente.

6. *Uso del modello e delle informazioni trovate*: in quest'ultima fase di deployment il modello generato e la conoscenza acquisita sul fenomeno devono essere messi a disposizione dell'utente attraverso, per esempio, la creazione di un report oppure implementando un sistema di data mining controllabile direttamente dall'utente.

1.5 Pattern

Le tecniche di data mining sono fondate su specifici algoritmi. I *pattern* identificati possono essere, a loro volta, il punto di partenza per ipotizzare, e quindi verificare, nuove relazioni di tipo causale tra fenomeni; in generale possono servire in senso statistico per formulare previsioni su nuovi insiemi di dati.

Un pattern deve essere [7]:

- Valido sui dati con un certo grado di confidenza;
- Comprensibile dal punto di vista sintattico e semantico affinchè l'utente lo possa interpretare;
- Precedentemente sconosciuto e potenzialmente utile affinchè l'utente possa intraprendere azioni di conseguenza.

Il data mining trova pattern e relazioni tramite la costruzione di *modelli*. I modelli, come può essere una cartina stradale, sono delle rappresentazioni astratte della realtà. Una cartina può modellare le vie di una città, ma non può mostrare un incidente che rallenta il traffico. Un modello non si dovrebbe mai confonder con la realtà, però un buon modello è un'utile guida alla comprensione dei problemi con i quali si ha a che fare, e suggerisce le azioni che possono essere intraprese

per raggiungere i propri scopi.

Esistono due tipi principali di modelli. Il primo, di tipo *predittivo*, usa i dati che rappresentano fatti o eventi per predirne esplicitamente la loro evoluzione; il secondo, di tipo *descrittivo*, modella i pattern esistenti che possono essere utilizzati per prendere decisioni.

I pattern estratti dal processo possono essere considerati l'output di cinque metodologie di base: *Classificazione*, *Regressione*, *Serie temporali*, *Regole associative*, *Clustering*. Modelli di classificazione, regressione e serie temporali sono principalmente usati per la predizione, mentre i modelli di clustering e di associazione sono soprattutto utilizzabili per la descrizione.

1.5.1 Classificazione

Si ha una classificazione quando vengono identificati schemi o insiemi di caratteristiche che definiscono il gruppo cui appartiene un dato elemento. Il task di classificazione prende in input un insieme di item ognuno già classificato e incluso in un insieme determinato. Mentre fornisce in output un modello (classificatore) che stabilisce l'appartenenza, basandosi sul valore che possiedono gli altri attributi. Formalmente un task di classificazione è effettuato attraverso tre passi:

1. *Training step* - si costruisce il nuovo modello a partire da un insieme di record già classificati (*dati storici*);
2. *Test step* - si verifica la bontà del modello su un insieme di dati precedentemente non noti;
3. *Prediction step* - si usa il modello per predire la classe dei dati che si vogliono classificare.

I modelli di classificazione sono realizzati nella maggior parte dei casi tramite alberi di decisione.

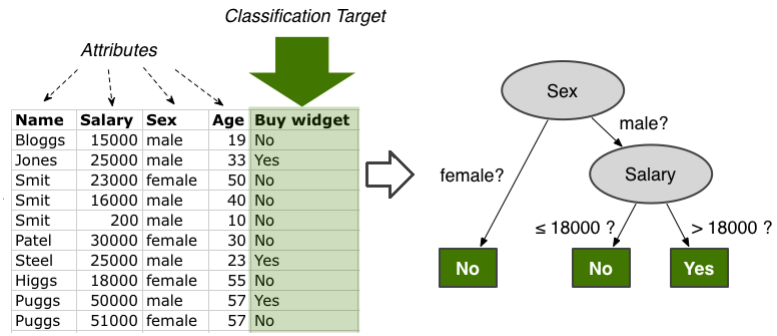


Figura 1.5: Esempio di classificazione

In alto, un esempio molto semplice di un insieme di dati, di cui ogni riga costituisce una istanza del dateset che può essere utilizzato per creare il modello di classificazione.

1.5.1.1 Alberi Decisionali

L'albero di decisione [21] è un modello utilizzato per la classificazione che viene costruito analizzando i vari attributi che descrivono i record che stiamo processando. Nel caso in cui l'albero interessi variabili discrete, è detto di *classificazione*; nel caso invece di variabili continue è detto di *regressione*.

Grazie ad un test sul valore di un attributo che identifica la variabile più rilevante, possiamo suddividere i dati, navigando l'albero dalla radice fino a giungere ad una foglia che è l'etichetta di classe. Gli archi rappresentano il risultato del test sull'attributo. Il predicato che si associa ad ogni nodo interno è chiamato condizione di split. A tal proposito, tra gli elementi caratterizzanti per definire un albero decisionale si distinguono [7]:

- *la condizione di split*: che può essere influenzata sia dal tipo di attributo, che dal tipo di split. Questo può essere a due o più vie senza violare l'ordi-

namento dei valori. Per gestire la complessità della ricerca dei punti di split ottimali è utilizzata una tecnica di discretizzazione statica (una sola volta) o dinamica (ad ogni passo di ricorsione).

- *il criterio che definisce lo split migliore*: che utilizza parametri noti in letteratura come l'errore di classificazione, Gini index e la variazione di entropia.

- *il criterio per interrompere lo splitting*: avviene quando tutti i record appartengono alla medesima classe.

Un esempio, mostrato di seguito, illustra la percentuale di sopravvissuti dei passeggeri sul Titanic. Ogni nodo interno corrisponde ad una delle variabili di ingresso; ogni foglia rappresenta un valore della variabile target.

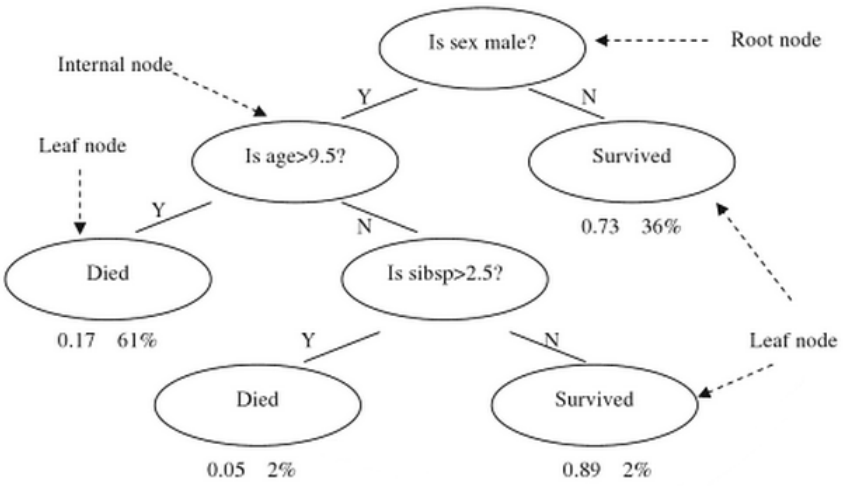


Figura 1.6: Esempio di albero: con "sibsp" si vuole indicare il numero di fratelli o coniugi a bordo. [8]

Nel data mining, gli alberi di decisione possono essere descritti anche come la combinazione di tecniche matematiche e computazionali per facilitare la descri-

zione, classificazione di un set di dati. I dati, in particolare, sono modellati nella forma:

$$(x,Y) = (x_1,x_2,...,x_k,Y) \tag{1.1}$$

La variabile dipendente, Y , è la variabile target che si deve scoprire, classificare o generalizzare. Il vettore x è composto dalle variabili di input $x_1...x_n$ che sono usate per la funzione. Tra le algoritmi di costruzione degli alberi decisionali vi sono:

- *Bagging (Breiman,1996)* [9]: algoritmo che costruisce più alberi decisionali campionando ripetutamente i dati di training e combinando le classificazioni predette. Utilizza la tecnica del reibussolamento (bootstrap) per costruire il classificatore.

- *Boosting (Friedman,1999)*[10]: utilizzato per problemi di regressione o di classificazione. Così come il bagging costruisce i classificatori modificando il training set concentrandosi, in particolare, sui record classificati in modo non corretto. Una delle tecniche di boosting più utilizzate è *AdaBoost* [11] da Adaptive Boosting.

- *C 4.5 (Quinlan,1993)*[12]: estende l'algoritmo ID3 [13]. Esso compie una ricerca hill-climbing attraverso l'insieme di tutti i possibili alberi di decisione partendo dall'ipotesi più semplice e cercandone una più complessa.

1.5.2 Regressione

Simile alla classificazione, la regressione è specializzata nella previsione di variabili quantitative. Molti risultati provengono dalla statistica, ambito in cui il problema della regressione è stato ampiamente studiato: il concetto base su

cui poggia l'intera teoria è la possibilità di approssimare la vera funzione che ha generato le osservazioni. La previsione è verificata sugli scarti d'errore generati dal modello rispetto ai veri valori.

La prima forma di regressione nella storia fu il *metodo dei minimi quadrati* pubblicato da Legendre nel 1805 [14] e da Gauss nel 1809 [15]. Entrambi applicarono il metodo al problema di determinare l'orbita dei pianeti attorno al sole. Il termine '*regressione*' venne coniato nel diciannovesimo secolo per descrivere la legge dell'eredità filiale stabilita da F. Galton [16], secondo cui i figli tenderebbero a variare nello stesso senso dei genitori, ma con intensità minore rispetto ai loro avi. Seppur il termine sia nato con significato biologico, poi venne ampiamente esteso in termini statistici.

L'analisi della regressione può essere usata per effettuare previsioni (ad esempio per prevedere dati futuri di una serie temporale), per testare ipotesi o modellare delle relazioni di dipendenza. Ci si occuperà della sua descrizione formale nel capitolo successivo.

1.5.3 Serie Temporali

Le serie temporali [20] esprimono la dinamica di un certo fenomeno nel tempo e vengono studiate sia per interpretare il fenomeno, individuando componenti di trend, ciclicità, stagionalità, sia per prevedere il suo andamento futuro in sequenze di dati complesse . Nel contesto del data mining lo studio delle serie temporali permette sia di effettuare misure di similarità che di filtrare e analizzare i processi stocastici descritti dai dati.

Un vantaggio della scoperta di sequenze temporali simili è che molti pattern possono essere scoperti senza nessuna condizione particolare da imporre, tranne

che i dati devono essere quantitativi e dipendenti dal tempo. Le serie temporali, o anche dette serie storiche, verranno ampiamente trattate nel capitolo a seguire.

1.5.4 Regole associative

Si ha un'associazione [17] quando più eventi o fatti vengono collegati da una relazione di causalità. La ricerca di regole di associazione all'interno di un database, pur essendo una delle tecniche concettualmente più semplici e intuitive, permette di ottenere eccellenti risultati in diversi campi, dall'ambito del marketing alla ricerca scientifica. Una regola di associazione è un legame di causalità valido tra gli attributi dei record di un database, cioè un espressione del tipo:

$$\mathbf{X} \Rightarrow \mathbf{Y} \tag{1.2}$$

dove X e Y sono insiemi di attributi e rispettivamente detti antecedente e conseguente. L'approccio consiste dunque in un insieme di *regole*, ossia relazioni tra attributi del tipo " *Se è presente l'attributo X, allora l'osservazione conterrà anche l'attributo Y*".

Gli algoritmi di associazione sono veloci ed efficienti nel ricavare le regole. Le difficoltà, piuttosto, nascono quando si deve giudicare della validità ed importanza delle regole. A questo riguardo metriche per la valutazione delle regole associative sono: il *supporto* e la *confidenza*.

Formalmente il supporto indica la probabiltà della presenza di X e Y nella transazione:

$$P(X,Y) = \frac{\sigma(X,Y)}{N} \tag{1.3}$$

E la *confidenza* misura quante volte gli elementi di Y appaiono in transizioni che contengono X, ovvero la probabilità della presenza di Y condizionata a quella di X.

$$P(X|Y) = \frac{P(X,Y)}{P(X)} = \frac{\sigma(X,Y)}{N}\frac{N}{\sigma(X)} = \frac{\sigma(X,Y)}{\sigma(X)} \tag{1.4}$$

Regole con supporto e confidenza elevati sono molto più significative di quelle con tali valori bassi.

L'applicazione più importante delle regole associative è senza dubbio il *market basket analisys* [18] (letteralmente analisi del paniere) che riguarda il processo di affinità tra i prodotti acquistati dai clienti in un supermercato, molto utile per l'adozione di strategie di marketing ad hoc. Individuata la regola associativa, si dice che l'antecedente se è tolto dal mercato condiziona il prodotto a lui collegato, per il conseguente bisogna invece capire cosa fare per incrementare la vendita.

Nell'ambito del data mining, tra i numerosi algoritmi per la scoperta delle regole di associazioni, l'algoritmo *Apriori* [19] ne risulta essere uno tra i più usati e funzionali.

1.5.5 Clustering

Il clustering, detto anche analisi dei gruppi, è simile alla classificazione, ma consente di produrre nuovi raggruppamenti di elementi omogenei in precedenza non definiti [22]. Il clustering deriva dal partizionamento del database in modo che i membri di ogni gruppo siano simili secondo alcuni criteri o metriche. L'output di un algoritmo di clustering può essere meglio interpretato se si utilizza una visualizzazione grafica che utilizza, per esempio, la distanza euclidea come misura di similarità se gli attributi dei punti assumono valori continui.

Esistono numerose tecniche per realizzare il clustering, alcune di queste sono:

- *algoritmi aggregativi*: nei quali i gruppi si formano aggregando le tuple a centri predeterminati in funzione di qualche criterio di scelta. Si tratta, per esempio, del più noto algoritmo di clustering, quello delle *k-means* [23].
- *scissori*: si basano sulla partizione dell'insieme iniziale in due sottoinsiemi e sulle successive suddivisioni che soddisfano qualche criterio di ottimalità.
- *reti neurali* [4]: definiscono un modello metematico per la simulazione di una rete di neuroni biologici.

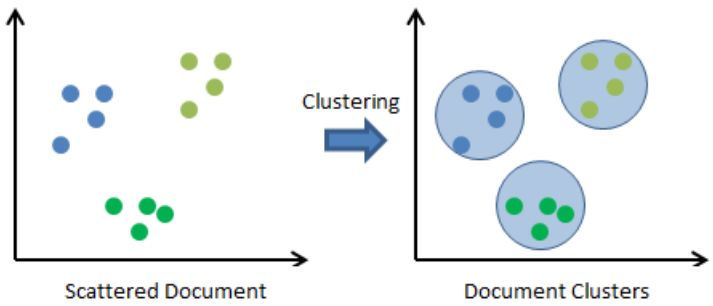


Figura 1.7: Esempio di rappresentazione del clustering

L'analisi dei cluster è oggigiorno utilizzata in numerose applicazioni, comprese il riconoscimento di pattern, il trattamento delle immagini e la ricerca di mercato. In economia, il clustering può aiutare gli operatori a scoprire gruppi distinti di clienti caratterizzandoli in base ai loro acquisti. In biologia, esso può essere utilizzato per derivare le tassonomie di piante e degli animali, per categorizzare i geni con funzionalità simili ed esaminare le varie caratteristiche delle popolazioni. Il clustering può essere utile anche nella ricerca degli outlier (valori molto lontani da ciascun cluster); in alcuni studi sono più importanti dei valori comuni. Si pensi,

ad esempio, alla ricerca delle frodi nelle carte di credito oppure al minitoring delle attività criminali nel commercio elettronico. Seppur si tratti di una disciplina scientifica giovane, è un campo della ricerca affascinante e in enorme sviluppo.

Capitolo 2

Forecasting

2.1 Introduzione ai modelli predittivi

Uno degli obiettivi principali del data mining è la *predizione*, ossia il tentativo di validare i risultati, applicando i modelli ottenuti a nuovi sottoinsiemi di dati.
Il *predictive modeling* [23] è simile all'esperienza dell'apprendimento umano, dove usiamo le osservazioni per creare un modello delle caratteristiche essenziali, sottostanti ad un certo fenomeno. Il modello deve essere in grado di fornire la risposta corretta di fronte ad alcuni casi precedentemente risolti, prima che possa esser utilizzato su nuove osservazioni (approccio definito "*supervised learning*").

La struttura complessiva di un generico sistema di previsione, a titolo di esempio, è mostrata in figura nella pagina a seguire.
L'obiettivo del forecasting è quello di determinare in anticipo il risultato più probabile di una variabile incerta. Nei sistemi di pianificazione e logistica, per esempio, il forecasting assume un ruolo dominante nella previsione delle attività aziendali dalla formulazione di strategie, alla pianificazione di quantitativi per la distribuzione, fino alla gestione delle scorte.

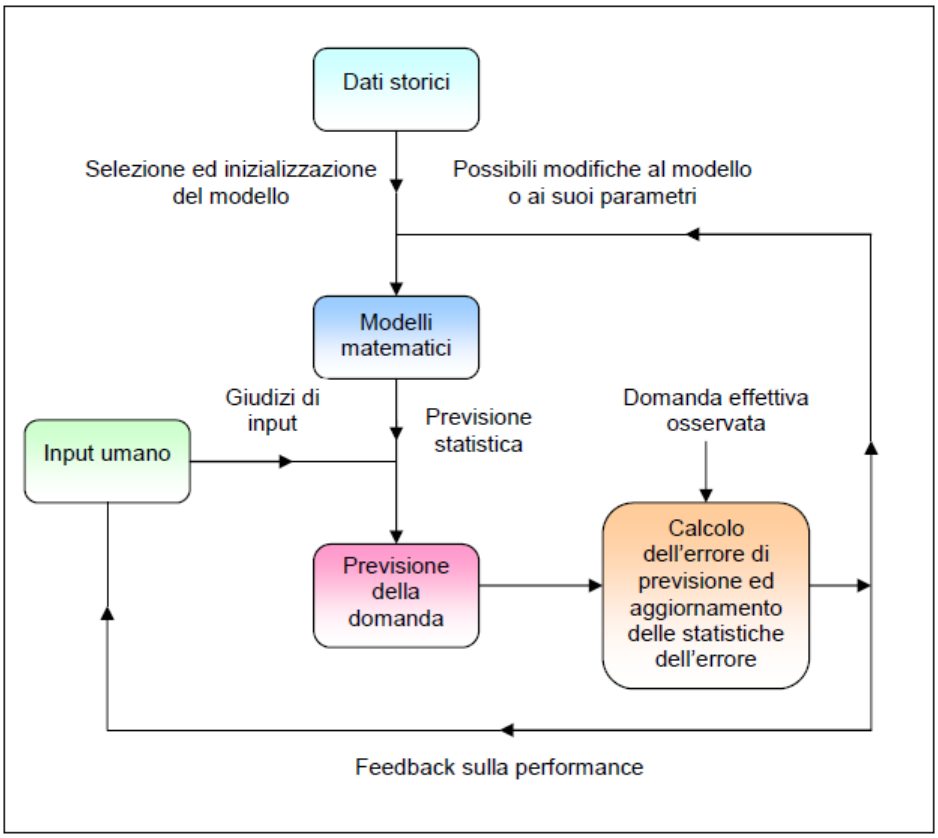


Figura 2.1: Esempio di architettura di un sistema di previsione

Dalla letteratura si definiscono tre orizzonti temporali che classificano le diverse tecniche previsionali in base al periodo di riferimento [24].

Le previsioni a *lungo termine* coprono un periodo da uno a cinque anni. Previsioni di questo tipo risultano meno affidabili e sono spesso generate per un intero gruppo di servizi piuttosto che sul singolo. Le previsioni a *medio termine* coprono un periodo da qualche mese fino ad un intero anno. Infine le previsioni di *breve termine* si sviluppano su al massimo qualche settimana. La figura schematizza quali metodi vengono solitamente utilizzati nei diversi range temporali.

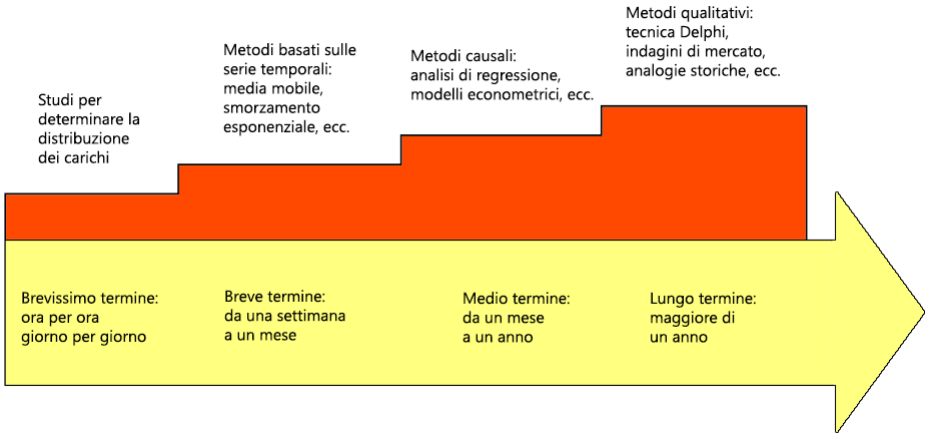


Figura 2.2: Applicazione di tecniche previsionali a diversi orizzonti temporali

Nel brevissimo termine, possono essere sufficienti degli studi per determinare le distribuzioni di carico orarie o giornaliere, assicurandosi che il sistema rimanga stabile. I metodi basati sulle serie temporali sono adatti ad orizzonti previsionali di breve termine mentre per le previsioni di lungo raggio sono da preferirsi i metodi qualitativi, come il metodo Delphi, che sottopone ad una serie di questionari un gruppo di esperti.

2.2 Tecniche previsionali

In letteratura i modelli previsionali sono classificati in *metodi qualitativi* o soggettivi, basati sul parere di esperti o indagini di mercato, e in *metodi quantitativi* o oggettivi che impiegano modelli matematici e dati storici [25].

Questi ultimi possono a sua volta essere distinti in due gruppi: *metodi causali* e *metodi basati sulle serie storiche*. I primi sono basati sull'ipotesi che la domanda futura dipenda dal passato o dal valore corrente di alcune variabili, prendendo in esame il rapporto tra l'evento causale e quello da prevedere. Tali metodi in-

cludono: *la regressione, modelli econometrici, modelli di input-output, modelli di simulazione al computer e le reti neurali*. Molti di questi approcci sono difficili da implementare, anche per le grandi aziende. In pratica, solo la regressione è tra tutte la più utilizzata. Metodi basati sulle serie storiche, invece, presuppongono che le caratteristiche della domanda non cambino, facendo esclusivamente riferimento all'andamento temporale. Ciò si esprime in diversi approcci come: *la media mobile, tecniche di smorzamento esponenziale, metodo Box - Jenkins*.

La scelta della tecnica quantitativa di forecasting più adatta dipende dal tipo di dati disponibili e dal tipo di prodotto o (servizio). Tuttavia, generalmente, si preferisce scegliere quella più semplice. Questo principio si basa sulle due osservazioni seguenti:

- le previsioni ottenute utilizzando tecniche semplici sono anche più facili da capire e spiegare. Questo è un aspetto fondamentale quando, sui processi di decision-making, sono investite grandi quantità di denaro.

- in un'ottica di business, raramente procedure di forecasting complesse rendono più di quelle semplici.

Queste due pratiche regole vengono spesso tenute in considerazione, come conferma il sondaggio effettuato nel Nord America e in Europa (*vedi Tabella*) [24].
La frequenza di utilizzo riportata nella seconda e terza colonna deve essere calcolata tenendo in considerazione il livello di familiarità degli analisti rispetto ai differenti metodi di forecasting (colonna 4). Per esempio, se si paragona la tecnica di decomposizione con quella più complessa di Box - Jenkins nel medio termine, si analizzano due differenti livelli di familiarità su tali approcci (rispettivamente del 57% e 37%). Ciò significa che la percentuale di utilizzo, se gli analisti conoscessero entrambe le tecniche, sarebbe del 21% ($\frac{12}{0{,}57}$) e 13,5% ($\frac{5}{0{,}37}$).

Forecasting method	Use (%) in short term	Use (%) in medium term	Level (%) of familiarity
Decomposition	7	12	57
Elementary technique	19	14	84
Moving average	33	28	96
Exponential smoothing	20	17	83
Regression	25	26	83
Box–Jenkins	2	5	37

Figura 2.3: Utilizzo e familiarità dei metodi di forecasting. Fonte: Ghiani - Logistic Systems Planning and Control

Nei paragrafi successivi, a partire dall'approccio causale, verranno descritte nel dettaglio prima le tecniche qualitative (metodi causali e basati su serie storiche) e poi quelle quantitative.

2.2.1 Metodi causali

2.2.1.1 Metodo di Regressione

La regressione, già introdotta nel primo capitolo, è un metodo statistico che mette in relazione una variabile dipendentente y (che rappresenta, per esempio, la domanda futura d_{t+1}) a una o più variabili indipendenti $x_1, x_2 \ldots x_n$ il cui valore è noto o può esser previsto.

Lo studio della regressione consiste nella determinazione di una funzione matematica che esprima la relazione tra le variabili. La funzione più utilizzata, soprattutto se i dati rilevati sono numerosi, è la funzione lineare; si parla allora di *regressione lineare*.

Pertanto, nella situazione più generale vi è un set di n campioni (X_i, Y_i) con i=1,2,...,n governati da una relazione lineare, stabilita dalla formula [26]:

$$Y = a + bX + e \tag{2.1}$$

in cui a e b coefficienti dell'equazione di regressione, e l'errore ovvero la deviazione dell'osservazione dalla relazione lineare.

Attraverso una rappresentazione grafica, che rappresenta le coppie dei valori rilevanti (x_i, y_i), si ottiene un diagramma a dispersione. Se esiste una relazione lineare, i punti si distribuiscono vicino alla retta, di regressione appunto, come nel grafo riportato.

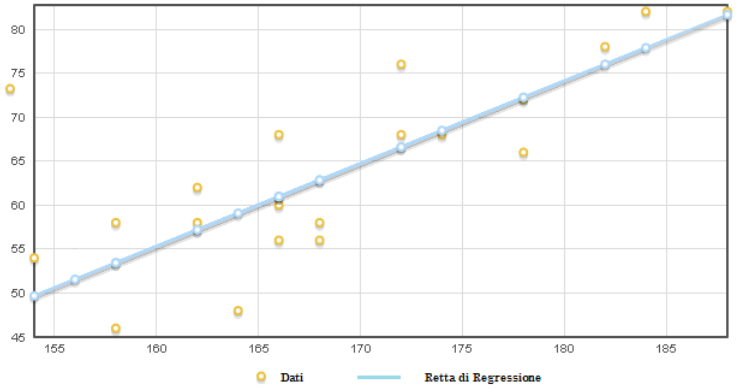


Figura 2.4: Retta di regressione

L'esistenza di un legame lineare è confermata da un valore elevato del *coefficiente di correlazione lineare* r, così chiamato come abbreviazione del termine regressione. Esso può variare tra 0, nel caso in cui non vi è correlazione, e $r = \pm 1$ nel caso di perfetta correlazione lineare. Nello specifico, quando $r > 0$ si dice che le due variabili sono correlate positivamente (al crescere di una, cresce anche l'altra); quando $r < 0$ sono correlate negativamente.

Il coefficiente di correlazione r_{XY} viene calcolato mediante la formula matematica:

$$r_{XY} = \frac{Cov_{XY}}{S_X S_Y} \tag{2.2}$$

in cui:

Cov_{XY} = la covarianza tra X e Y, ovvero $\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$

$S_X S_Y$ = la deviazione standard rispettivamente di X e di Y.

È possibile verificare la bontà di adattamento (in inglese *fitting*) della retta di regressione alla serie delle osservazioni in esame mediante il *coefficiente di determinazione* R^2. In statistica indica la proporzione tra la variabilità dei dati e la correttezza del modello statistico usato e si esprime come:

$$R^2 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2} \tag{2.3}$$

Coefficiente che varia tra 0 e 1 e deve assumere valore 1 per avere un buon adattamento.

2.2.2 Metodi basati sulle serie storiche

L'ipotesi di base assume che il comportamento della domanda segua nel futuro una legge determinata dall'andamento passato. In particolare, una **serie storica** si definisce formalmente come *una sequenza di valori $D_1 \ldots D_n$ assunti da una grandezza misurabile e osservati in corrispondenza di specifici intervalli temporali di norma equidistanti [27].*

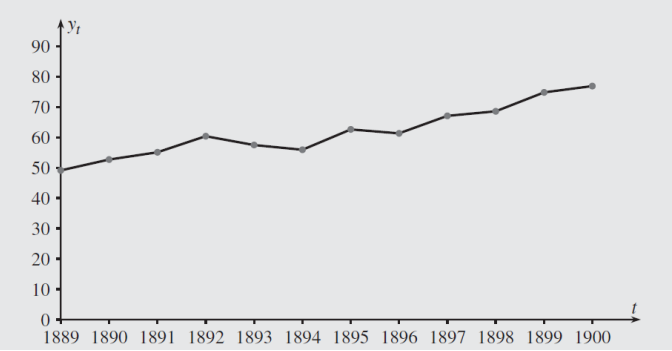


Figura 2.5: Rappresentazione grafica della serie storica del prodotto nazionale lordo (PNL in miliardi di dollari) negli USA dal 1889 al 1900. [28]

Il modo più semplice ed intuitivo per studiare una serie storica è quello di usare una rappresentazione grafica (diagramma Cartesiano (t, y_t)), che consente di interpretare i dati più rapidamente che in una tabella. Analizzando l'andamento della serie storica si individua l'esistenza di cinque componenti della domanda che riguardano: (1) variazioni sistematiche dei dati dovute alla stagionalità, (2) eventuali andamenti ciclici, (3) tendenze e (4) tasso di crescita di queste tendenze (5) outliers.

Prima di applicare il modello basato sulle serie storiche, la variabile da prevedere deve essere analizzata rispetto alle prime tre componenti. Una volta individuate le singole componenti delle previsioni, i metodi basati sulle serie storiche partono dal presupposto che il futuro sarà simile al passato, cioè che l'andamento esistente continuerà invariato nel futuro.
Questo presupposto è solitamente abbastanza giusto nel breve termine ed è qui che l'applicazione di queste tecniche risulta più adatta. Tuttavia, queste tecniche non consentono previsioni precise, a meno che l'andamento della domanda non sia piuttosto stabile.

Verranno considerate in ordine di complessità crescente quattro tecniche fondamentali e cioè: la media mobile, la media mobile ponderata, lo smorzamento esponenziale e infine il modello ARMA/ARIMA [29] [24].

2.2.2.1 Media mobile

D'ora in avanti si suppone il tempo diviso in periodi t = 1,2,3... e viene indicato con d_t la domanda nel periodo t come dato storico noto, mentre d_τ la previsione in un periodo τ futuro.

Precursore (o caso particolare) del metodo della media mobile è il *metodo elementare* [48] dove la previsione della domanda per domani è semplicemente data dalla domanda osservata oggi:

$$p_{t+1} = d_t \tag{2.4}$$

Tale metodo non fa altro che replicare l'effettiva domanda con un ritardo di un periodo. Questo, in generale, genera una previsione molto scarsa e riduttiva.

Diversamente la previsione del fabbisogno per un periodo futuro a media mobile si determina attraverso la media aritmetica delle richieste osservate durante n periodi (ad esempio mesi o settimane) anteriori. La media si dice «mobile» perchè viene costantemente aggiornata sostituendo via via l'ultimo dato disponibile al più lontano del tempo. Analiticamente il calcolo della media mobile applicata alla previsione per un periodo futuro indicato con t+1 risulta dall'espressione:

$$p_{t+1} = \sum_{k=0}^{r-1} \frac{d_{t-k}}{r} \tag{2.5}$$

dove:

p_{t+1} = la previsione del periodo successivo

d_t = la domanda effettiva nel periodo precedente

r = il numero di periodi

Notevole importanza ha la scelta del numero ottimale di periodi presi in considerazione nel calcolo della media mobile. Per elevati valori di r la variabilità dei tempi anteriori risulterà più attenuata; per valori piccoli risulterà, al contrario, enfatizzata.

La media mobile, anche se è facile da calcolare, ha una scarsa reattività nel riflettere i cambiamenti e richiede l'archivio e l'aggiornamento di una elevata quantità di dati storici.

2.2.2.2 Media mobile ponderata

Con la tecnica precedente viene dato lo stesso peso alla serie storica, con quella della media mobile ponderata [49], invece, è possibile attribuire un peso a ciascun elemento, garantendo, ovviamente, che la somma di tutti i pesi risulti uguale a uno. Pertanto in questo caso la formula è la seguente:

$$p_{t+1} = p_1 * d_{t-1} + p_2 * d_{t-2} + ... + p_n * d_{t-n} \tag{2.6}$$

Questa soluzione ha numerosi vantaggi rispetto alla media mobile semplice, tuttavia è più sconveniente e costosa. Un ulteriore passo in avanti è costituito dallo smorzamento esponenziale (*exponential smoothing*).

2.2.2.3 Smorzamento esponenziale

Un'altra versione più affidabile e facilmente automatizzabile è quella che prevede lo smorzamento esponenziale [42] del peso dei dati più lontani (*Modello di Brown*). Si tratta dunque di un particolare tipo di media mobile, ove i fattori sono ponderati in base ad un parametro sottoposto a continua revisione.

Analiticamente la domanda media stimata di un periodo futuro t+1, viene espressa dalla funzione:

$$p_{t+1} = \alpha d_t + (1 - \alpha) p_t \tag{2.7}$$

dove:

p_{t+1} = la previsione del periodo t

d_t = la domanda effettiva nel periodo precedente

α = il coefficiente di attenuazione esponenziale (*smoothing parameter*)

Uno dei problemi legati a questa tecnica sta nella determinazione del valore da attribuire al fattore alfa. Si può infatti notare che:

- $\alpha \to 1$ aumenta la sensibilità al cambiamento della previsione che diventa molto reattiva;
- $\alpha \to 0$ rallenta la rapidità con cui la previsione riflette il cambiamento e implica, pertanto, una reazione lenta o ritardata.

Di conseguenza, l'uso dello smorzamento esponenziale non elimina la neccessità di prendere decisioni di merito. In effetti, chi si occupa della previsione, nel stabilire il fattore alfa, deve tener conto dei vantaggi e svantaggi che comporta.

Si può mostrare infine che il numero di periodi N significativi, cioè necessari per individuare il nuovo valore della domanda prevista, dipende da α secondo la relazione:

$$N = \frac{2-\alpha}{\alpha} \tag{2.8}$$

Quindi già per $\alpha = 0,2$ solo gli ultimi 9 termini di una serie sono significativi per la previsione; ovviamente il numero di periodi da prendere in esame aumenta al diminuire di α.

Questa tecnica può essere efficacemente impiegata quando la domanda è stazionaria. Nel caso, invece, in cui è presente un trend si introduce un secondo coefficiente di attenuazione esponenziale denominato β. Questo nuovo coefficiente riduce l'impatto dell'errore che si verifica tra la previsione e il valore attuale. Questo tipo di approccio lineare denominato anche *modello di Holt*, per $\mathsf{T} = 1, ..., T$ assume la forma:

$$p_t(\mathsf{T}) = a_t + b_t \mathsf{T} \tag{2.9}$$

Per tale calcolo della previsione della domanda con correzione di trend, si procede studiando inizialmente il livello della domanda prevista:

$$a_t = \alpha d_t + (1-\alpha)(a_{t-1} + b_{t-1}) \tag{2.10}$$

Successivamente viene calcolato il trend corrente, valutato sulla base della differenza rilevata tra i livelli previsti della domanda nei due periodi adiacenti e il trend calcolato nel periodo precedente.

$$b_t = \beta(a_t - a_{t-1}) + (1-\beta)b_{t-1} \tag{2.11}$$

Quest'ultimo metodo consente di calcolare le nuove previsioni in modo rapido, con forte dipendenza dalla scelta dei valori delle costanti di smorzamento. Tuttavia, questo vantaggio teorico rappresenta anche il suo più grande limite. In effetti, i modelli di Holt sono spesso caratterizzati da un'eccessiva sensibilità dovuta alla difficoltà di ponderare le singole componenti della previsione, con conseguenti problemi di attendibilità delle previsioni stesse [30].

2.2.2.4 Modello ARMA/ARIMA

Box e Jenkins [31] hanno concepito un modello più complesso di quelli visti precedentemente, ma che rappresenta un'opportunità per una previsione più accurata.

Il modello autoregressivo a media mobile (acronimo di *Autoregressive Moving Average*, ARMA) valuta la domanda nel periodo attuale d_t attraverso una somma pesata di domande passate e di componenti casuali non prevedibili.

Data appunto una serie storica di valori, il modello ARMA consiste di due parti, ossia una parte autoregressiva (AR) e di una parte di media mobile (MA); solitamente indicato con *ARMA(p,q)* dove p è l'ordine della parte autoregressiva e q di quella media mobile.

Formalmente si dice che una serie storica d_t segue un modello ARMA(p,q) se soddisfa la relazione [32]:

$$d_t = a_0 + a_1 d_{t-1} + ... + a_p d_{t-p} + \varepsilon_t \tag{2.12}$$

dove

$$\varepsilon_t = u_t + b_1 u_{t-1} + ... + b_q d_{t-q} \tag{2.13}$$

è detto processo a media mobile di ordine q o MA(q) e dove gli errori u_t sono un *rumore bianco*, cioè una successione di variabili aleatorie a media zero e varianza finita.

Un modello ARMA ha diverse caratteristiche che lo rendono semplice da analizzare:

- *linearità* moltiplicando i valori in ingresso per un fattore k anche l'uscita risulterà moltiplicata per tale valore;
- *tempo invarianza*: il sistema tende a dimenticare il passato, in modo da esserne influenzato in maniera esponenzialmente decrescente nel tempo.

Nel caso in cui i dati evidenzino la presenza di non stazionarietà, è possibile rimuovere tale proprietà attraverso la trasformazione in differenze prime,$d_t - d_{t-1}$.

Il modello ARMA(p,q) applicato ai dati così trasformati prende il nome di modello ARIMA (*Autoregressive Integrated Moving Average*) con parametri (p,1,q).

La trasformazione dei dati in differenze prime può essere applicata $d \geq 0$ ottenendo così il modello ARIMA(p,d,q). In particolare, il modello ARIMA (p,0,q) coincide con il modello ARMA(p,q).

Un semplice esempio di modello ARIMA(0,1,0) è dato dalla serie storica:

$$d_t = d_t - 1 + u_t \tag{2.14}$$

2.2.3 Metodi qualitativi

L'utilizzo di tali metodi qualitativi in orizzonti temporali di lungo periodo, a differenza delle previsioni di medio/breve termine, non sfrutta dati storici su cui basare il forecasting.

Nell'ambito manageriale vengono adottati per prendere le più importanti decisioni basandosi su cosa le persone pensano, su come reagiscono ai test di mercato e sull'analogia con situazioni simili. Per questo motivo le previsioni di lungo termine sono di natura più qualitativa. Tra le tecniche di maggior rilievo è nota la metodologia Delphi [33] [34] [35] [36] che, affidandosi ad un pool di esperti, si snoda nei seguenti passi:

- ogni esperto del gruppo formula una previsione indipendente sotto forma di breve asserzione;
- un coordinatore pubblica e spiega queste asserzioni;
- il coordinatore fornisce una serie di domande, che vengono alla fine combinate tra loro.

Solitamente i membri mantengono l'anonimato e la loro identità resta segreta. Questo impedisce ai partecipanti di usare la loro autorità e personalità durante il processo per dominare gli altri. Ognuno è libero quindi di esprimere il proprio punto di vista e di fare aperte critiche.

Le prime applicazioni del metodo Delphi furono nel campo della scienza e della previsione tecnologica. Successivamente si è diffuso in altre aree relative

alla *policy pubblica* [37], come l'andamento economico e la salute, riscontrando poi successo nelle previsioni di business.

2.3 Scelta del metodo previsionale

Come illustrato, sono presenti molte tecniche previsionali e, quindi, un buon sistema di previsione richiede la scelta di una tecnica matematica o statistica appropriata. Esistono vari modi per valutare le tecniche previsionali alternative [38].

Makridakis e Wheelwright [39] suggeriscono i seguenti criteri per la valutazione dell'applicabilità di una determinata tecnica:

1. il grado di precisione
2. l'orizzonte temporale di previsione
3. l'importanza delle previsioni
4. la disponibilità dei dati
5. il tipo di andamento dei dati
6. l'esperienza degli utenti nel campo delle previsioni

Ogni tecnica previsionale alternativa deve essere valutata in termini qualitativi e quantitativi a fronte di questi sei criteri.

In questi ultimi decenni, le tecniche previsionali si sono evolute con l'incorporazione di funzioni statistiche e analitiche avanzate. Lo sviluppo di queste tecniche si è basato sull'ipotesi che la maggior complessità avrebbe garantito una maggiore precisione nelle previsioni. Studi recenti indicano invece che in molti casi l'alternativa migliore è quella più semplice. Le tecniche più avanzate non sempre danno

risultati significativamente migliori rispetto alle tecniche più semplici, soprattutto quando si tiene conto delle maggiori risorse richieste sia a livello informatico sia a livello di competenza [40].

In altre parole, è compito dei responsabili del sistema previsionale scegliere la tecnica (o più di una) che dimostra, attraverso prove e simulazioni di dare migliori risultati [41].

2.4 Monitoraggio delle previsioni

A seguito della scelta del metodo, attuare un sistema di monitoraggio risulta una componente fondamentale nel processo previsionale. Un buon sistema di monitoraggio, infatti, si basa su presupposti di semplicità, sinteticità e flessibilità in grado di garantire accuratezza previsionale (*forecast accuracy*).

Tuttavia, per poter parlare di accuratezza è prima necessario definire il concetto di *errore di previsione* [29] che, per il periodo t, è definito come differenza tra il valore effettivo della domanda e il valore previsto per quel periodo:

$$E_t = D_t - P_t \tag{2.15}$$

La stima puntuale dell'errore di previsione della domanda per un singolo periodo di tempo è di per sè di poco aiuto; occorre quindi tener sotto controllo una serie di indicatori sintetici che possono informare circa il tipo e l'entità degli errori per migliorare il processo di previsione.

Tra i principali indicatori, i più frequentemente usati sono i seguenti:

- **Errore Medio** (*o anche Mean Error - ME*): è la media aritmetica degli errori commessi

$$ME = \frac{1}{n}\sum_{t=1}^{n} D_t - P_t \tag{2.16}$$

- **Errore Medio Assoluto** (*o anche Mean Absolute Deviation - MAD*) rende prima di tutto ogni errore positivo, prendendone il valore assoluto, poi ne fa la media.

$$MAD = \frac{1}{n}\sum_{t=1}^{n} |D_t - P_t| \tag{2.17}$$

- **Errore Quadratico Medio** (*o anche Mean Square Error - MSE*) rende analogamente gli errori positivi mediante l'elevazione a quadrato

$$MSE = \frac{1}{n}\sum_{t=1}^{n} |D_t - P_t|^2 \tag{2.18}$$

- **Errore Medio Assoluto Percentuale** (*o anche Mean Absolute Percentage Deviation - MAPD*)

$$MAPD = \frac{100}{n}\sum_{t=1}^{n} \frac{|D_t - P_t|}{D_t} \tag{2.19}$$

Mediante il calcolo del MAPD è possibile valutare la qualità del metodo previsionale. Vengono riportati in tabella i valori che tale indicatore può assumere.

MAPD	QUALITY OF FORECAST
$\leq 10\%$	Very Good
$> 10\%, \leq 20\%$	Good
$> 20\%, \leq 30\%$	Moderate
$> 30\%$	Poor

Tabella 2.1: *Range di valore del MAPD - Dati tratti da Ghiani - Forecasting Logistic Requirements*

Capitolo 3

Stato ”As is” delle previsioni della domanda termica in Optit

3.1 Il teleriscaldamento

Prima di illustrare il contesto del modello aziendale esistente, è bene definire il significato ed il funzionamento del teleriscaldamento urbano.

Nel glossario dell'Autorità per l'Energia Elettrica ed il Gas si trova la seguente definizione [43]: *sistema di riscaldamento a distanza di un quartiere o di una città che utilizza il calore prodotto da una centrale termica, da un impianto di cogenerazione, da una sorgente geotermica o da altre fonti.*

Il teleriscaldamento è una soluzione alternativa, rispettosa dell'ambiente, sicura ed economica per la produzione di acqua igenico sanitaria e il riscaldamento degli edifici residenziali, terziari e commerciali. Il termine « teleriscaldamento» sottolinea la peculiarità del servizio, ossia la distanza esistente tra il punto di produzione del calore e i punti di utilizzo. Il cuore del sistema risiede, infatti, in una

o più centrali in cui viene prodotta l'acqua calda che viene inviata negli edifici anche ad alcuni chilometri di distanza: una soluzione innovativa rispetto ai sistemi di riscaldamento tradizionale che prevedono l'installazione di caldaie in ciascun edificio.

L'acqua calda, trasportata attraverso una rete di tubazioni, giunge fino agli edifici allacciati. Qui, tramite uno scambiatore di calore, l'acqua cede il calore all'impianto condominiale e/o individuale e consente di riscaldare gli ambienti e avere l'acqua calda per impieghi domestici. In alcune zone della città, nei mesi estivi, il teleriscaldamento si trasforma in refrigerazione.

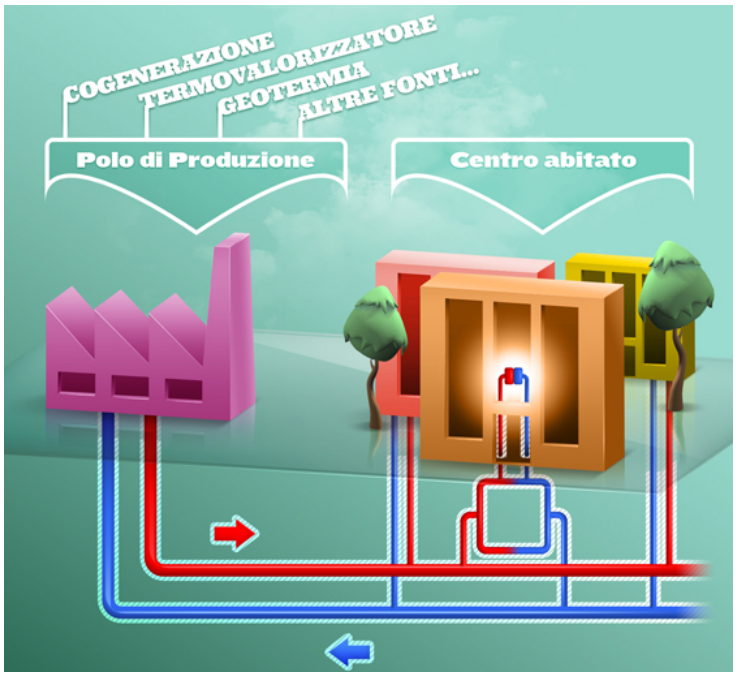


Figura 3.1: Rappresentazione di un sistema di teleriscaldamento - Fonte Gruppo Hera [44]

Come evidenziato in figura, l'impianto di teleriscaldamento è composto da una *centrale termica* (nel polo di produzione), una *rete di distribuzione* e da un *sistema di sottocentrali* (nel centro abitato). La centrale di produzione del calore,

alimentata da combustibili convenzionali ad alto rendimento, fornisce in uscita acqua calda o surriscaldata, alla temperatura circa di 90°. La rete di distribuzione, costituita da tubazioni interrate preisolate, trasporta e distribuisce l'acqua alle utenze e la riconvoglia alla centrale dopo che è avvenuta la cessione del calore. Le sottocentrali sono una serie di scambiatori collocati nei singoli edifici, per il trasferimento all'impianto interno del calore necessario.

Il calore è solitamente prodotto in una centrale di cogenerazione termoelettrica a gas naturali o biomasse, oppure utilizzando il calore proveniente dalla termovalorizzazione dei rifiuti solidi urbani. Oltre alle biomasse, le altre fonti di energia rinnovabile utilizzate per il teleriscaldamento sono la geotermia e il solare termico. L'immagine illustra alcuni di questi esempi.

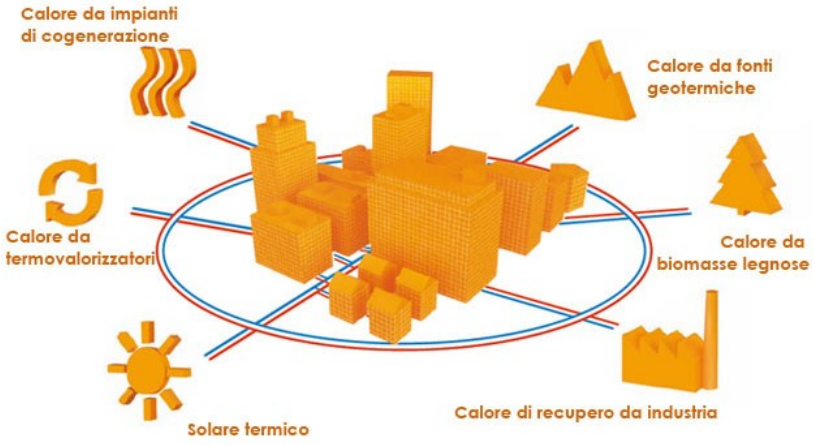


Figura 3.2: Fonti rinnovabili del Teleriscaldamento - Fonte AIRU [45]

Il teleriscaldamento è un servizio molto diffuso in Europa (Centro e Nord), in Giappone e nei paesi dell'Est. Il primo impianto di teleriscaldamento urbano nel mondo è stato quello di New York, risalente al 1876 (oggi quasi la totalità di Manhattan è teleriscaldata), mentre il primo impianto europeo è stato installato nel 1893 ad Amburgo. In Italia una tra le prime città italiane a dotarsi di un siste-

ma di teleriscaldamento è stata Brescia, seguita da Torino, che alla fine del 2011 possedeva la rete di teleriscaldamento più estesa d'Italia e fra le maggiori del continente. La regione Emilia Romagna, oggi in particolare, obbliga (delibera n.1366 del 26/09/2011) la predisposizione delle opere necessarie a favorire il collegamento a reti di teleriscaldamento, in presenza di tratte di rete a una distanza inferiore a 1000 metri, nel caso di nuove costruzioni, di ristrutturazioni, nuove installazioni di impianti di climatizzazione in edifici esistenti.

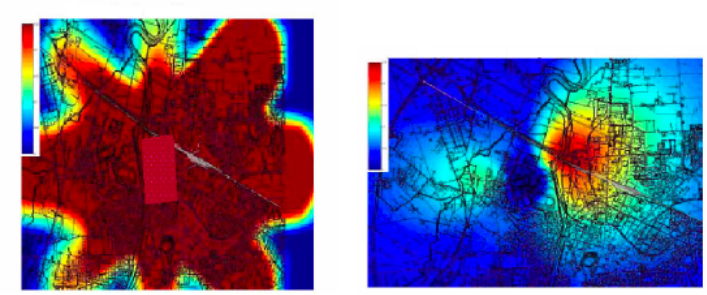
I vantaggi che il teleriscaldamento può offrire, rispetto alle forme tradizionali di produzione di energia termica, essenzialmente possono essere ricondotti a:

- risparmio energetico e benefici ambientali (benefici collettivi)
- vantaggi economici e semplicità d'uso per gli utenti (benefici individuali)

Nel dettaglio, per quanto concerne gli aspetti energetici ed ambientali, il teleriscaldamento urbano consente di utilizzare tutte le fonti energetiche disponibili; infatti nella centrale è possibile bruciare combustibili diversi a seconda della maggiore convenienza economica e disponibilità.

Dato l'utilizzo di un'unica centrale, inoltre, sono garantiti maggiori controlli sui gas di scarico, rispetto a quelli effettuati sulle singole caldaie, in minima misura.

Per dare una idea più concreta sul vantaggio ambientale del teleriscaldamento rispetto al riscaldamento alimentato con caldaie, si riportano in figura i risultati di uno studio condotto dal consorzio OPET SEED [46] per un impianto realizzato nel comune di Cesena.



Impatto ambientale con caldaiette *Impatto ambientale con teleriscaldamento*

Figura 3.3: Caso studio "Impianto di teleriscaldamento a cogenerazione realizzato nel Comune di Cesena - Fonte OPET SEED")

Per quanto concerne, invece, i benefici all'utente finale, il servizio è semplice da usare, sicuro ed economico.

Tuttavia l'ostacolo principale alla diffusione dei sistemi di teleriscaldamento non è certo legata agli aspetti tecnologici, semplici e collaudati, e ambientali, piuttosto agli aspetti finanziari, normativi e culturali. Alquanto sconosciuto e spesso confuso con i sistemi di riscaldamento centralizzati che non godono di buona fama, dovrebbe beneficiare, al contrario, di una maggiore ed opportuna politica di promozione.

3.2 Optit-EPM

Con il progetto di tesi si vuole elaborare una soluzione per ottimizzare la previsione della domanda termica, coniugando perfettamente sia i contenuti di analytics che attraverso l'uso di software di forecasting.

Le serie storiche dei dati termici sul quale contestualizzare il lavoro, dato un problema di real-life, vengono fornite dall'applicativo **OptitEPM** (E*nergy Production Management*). Questa soluzione è sviluppata da Optit srl per utilities che

gestiscono impianti per l'alimentazione di reti di teleriscaldamento urbano.

Altro non è che un applicativo che consente di modellare il processo di ottimizzazione dei sistemi complessi co-trigenerativi. Questi impianti possono essere, per esempio, composti al suo interno da:

- *motori a cogenerazione* che bruciano gas per produrre EE (Energia Elettrica) e C (Calore)
- *sistemi di assorbimento* che prendono il C e lo traducono in F (Freddo)
- *caldaie di integrazione* che bruciano gas e producono C
- *gruppi frigo* che prendono in ingresso EE e la traducono in F

Si tratta dunque di sistemi complessi che data una certa domanda termica (C), frigogena (F) e elettrica (EE) sono fortemente influenzati da: (1) prezzo dell'energia elettrica, (2) prezzo del gas, (3) vincoli e funzioni di trasformazione all'interno del sistema, (4) prezzo di uscita che, per esempio, è differente se indirizzata a un ospedale o a Enel.

Mediante l'applicativo OptitEPM è dunque possibile gestire il processo di ottimizzazione di tali impianti di cogenerazione. Il sistema OptitEPM, in particolare, è una soluzione di supporto di programmazione di breve, medio e lungo termine che mette a disposizione tre differenti funzionalità. In primo luogo, consente di effettuare il forecasting della richiesta di carico termico della rete di teleriscaldamento urbano in esame. Secondariamente, elabora una previsione ottimizzata di generazione di energia elettrica dell'impianto sul breve termine (il giorno seguente) e medio-lungo termine (fino all'intero anno) sulla base del fabbisogno termico da erogare. Infine, verifica condizioni alternative di funzionamento che soddisfino

a pieno, o con scostamenti minimi, le esigenze di erogazione delle curve termiche ed elettriche date.

Dato il problema di ottimizzazione basato sullo storico dei dati termici, il problema decisionale si esprime a partire dalla previsione che stima il fabbisogno della rete di teleriscaldamento, dalla previsione di disponibilità delle macchine degli impianti e dalla stima del prezzo dell'energia elettrica. Sulla base di queste informazioni è possibile creare una pianificazione ottimizzata dell'impianto in esame a N giorni.

L'elemento centrale di questo studio e fornito dunque da tale applicativo, risulta pertanto il *calore* (C). I dati a consuntivo prelevati dai diversi impianti rappresentano in particolare il calore prodotto su base oraria.

3.3 Software Weka

Weka, acronimo di *Waikato Enviroment for Knowledge Analysis* [47], è un ambiente software interamente scritto in Java che, attraverso metodi di apprendimento automatico, consente di avere una previsione del comportamento dei dati.

Senza dubbio la scrittura del codice sorgente in linguaggio Java ha portato notevoli vantaggi che riguardano i punti di forza del linguaggio stesso, come la portabilità, la gestione della memoria e la programmazione orientata ad oggetti.

Weka è stato concepito per esser facilmente usabile: lo sforzo di sviluppare interfacce semplici e intuitive è davvero apprezzabile, sebbene talvolta sia limi-

tante rispetto alle potenzialità del prodotto.

Figura 3.4: Interfaccia GUI Chooser Weka

Come mostra la figura, è possibile scegliere diversi tipi di GUI attraverso cui sfruttare diverse caratteristiche di Weka:

- *Simple CLI*: permette l'esecuzione degli algoritmi da linea di comando, da cui appunto il nome «command line interface». Tale modalità simula l'esecuzione dell'algoritmo prescelto come fosse un programma a parte. Sebbene tale modalità sia nata per esigenze di test in fase di sviluppo, è certamente utile avere la possibilità di eseguire un algoritmo da linea di comando che costringano l'interazione con l'utente.

- *Explorer*: rappresenta l'interfaccia grafica di riferimento attraverso cui è possibile avviare la fase di analisi mediante i numerosi algoritmi di machine learning. Questa interfaccia, in particolare, si presenta all'utente come un insieme di pagine sovrapposte, ciascuna delle quali offre le varie fasi di analisi: la preparazione dei dati, la classificazione, il clustering, l'associazione, la selezione automatica degli attributi e la visualizzazione grafica dei dati.

Nella pagina *Preprocess*, l'utente è guidato alla selezione dei dati da utilizzare per l'analisi. Una volta importati i dati, il dataset è pronto per essere modificato in modo manuale o automatico mediante l'applicazione dei filtri.

Le restanti pagine aiutano l'utente nell'applicazione delle varie tecniche di data mining al dataset. La struttura di ciascuna è essenzialmente la stessa: i possibili algoritmi tra cui scegliere, le opzioni di configurazione dell'algoritmo prescelto e infine la finestra con l'output finale del modello costruito, accompagnato dalle valutazioni sulla bontà dei risultati ottenuti.

Si mostra di seguito, in *figura*, un esempio di interfaccia Explorer utilizzata nella pagina di Preprocess per l'analisi dei dati di un impianto.

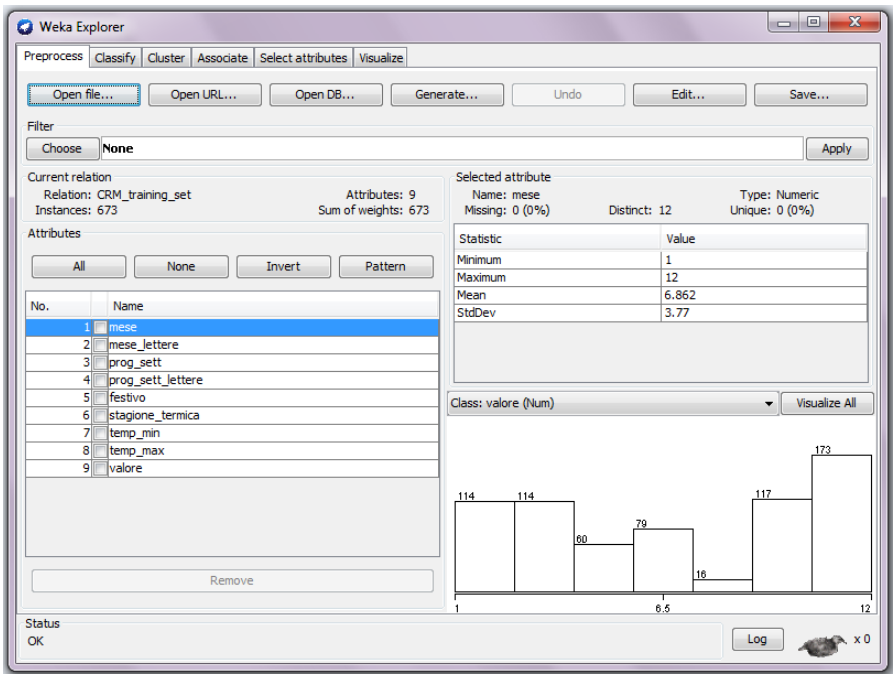


Figura 3.5: Interfaccia Explorer Weka

- *Experimenter*: questa interfaccia grafica si propone come soluzione di rapida configurazione al problema dell'esecuzione sincrona di più esperimenti: l'obiettivo è quello di applicare contemporaneamente diversi algoritmi di data mining su uno stesso dataset di ingresso e di poter analizzare successivamente i risultati ottenuti.

- *KnowledgeFlow*: è uno strumento grafico per la modellazione del data-flow, ossia la sequenza di operazioni successive di cui si compone l'analisi. L'utente compone uno schema mediante drag-and-drop di icone che individuano una certa operazione e la definizione della successione delle operazioni da effettuare tramite collegamenti grafici tra icone. Grazie alla forza comunicativa che possiede il linguaggio visivo, è possibile descrivere in modo semplice la procedura d'analisi.

Il formato utilizzato in Weka per la lettura dei dataset è l'*ARFF* (*Attribute Relationship File Format*) che corrisponde all'equivalente di una tabella di un database relazionale. In questo file, i campi vengono strutturati nel seguente modo:

```
@relation CRM_training_set
@attribute mese integer
@attribute mese_lettere {apr, ago, dic, feb, gen, giu, lug, mag,
    mar, nov, null, ott, set}
@attribute prog_sett integer
@attribute prog_sett_lettere {domenica, giovedì, lunedì, martedì
   , mercoledì, null, sabato, venerdì}
@attribute festivo {f, null, t}
@attribute stagione_termica {f, null, t}
@attribute temp_min integer
@attribute temp_max integer
@attribute valore real
@data
```

```
9,set,1,lunedì,f,f,16.4,21.8,24065
9,set,2,martedì,f,f,18,23.3,25097
9,set,3,mercoledì,f,f,18,23.7,24152
9,set,4,giovedì,f,f,18.2,26,23875
9,set,5,venerdì,f,f,16.5,25.6,22848
9,set,6,sabato,f,f,16.1,26.2,22680
9,set,7,domenica,t,f,14,27.5,24073
...
```

Il file inizia con una riga contenente il tag *@relation*, che indica il nome o la descrizione del dataset. Seguono tante righe precedute dal tag *@attribute* quanti sono gli attributi di ciascuna osservazione. Per ogni attributo è specificato il nome e il tipo come elenco dei possibili valori: real se numerico, string per il testo libero e date (seguito dall'eventuale formato se diverso da quello ISO). La sezione dedicata alle osservazioni è segnalata dalla riga con il tag *@data*.

I classificatori addestrati possono essere salvati su file che sono definiti modelli. In questo modo è sempre possibile ricaricare un modello, magari rieseguendolo su un nuovo dataset mediante l'opzione di test "supplied test set". Weka, ed in particolare il package `weka.classifier`, contiene l'implementazione degli algoritmi più comunemente utilizzati per la classificazione e predizione numerica. La tabella mostra i metodi definiti.

Categoria di classificatore	Strumenti usati
Bayes	Reti bayesiane
Functions	Refressione lineare, Reti neurali, SVM
Lazy	Somiglianza
Meta	Bagging, Boosting, Cost sensitive classificatiom, etc.
Mi	Algoritmi per dati multi-istanze
Misc	InputMapped, Serialized
Rules	Regole associative
Tree	Alberi decisionali

Tabella 3.1: Tipi di classificazione

3.4 Il modello esistente

Ad oggi l'azienda Optit srl, spin-off accreditato dell'università di Bologna nata nel 2007 con sede operativa a Cesena, si colloca tra le aziende leader nello sviluppo di metodologie e soluzioni (*Decision Support Systems*) di forecasting, data analysis, simulazione e ottimizzazione dei sistemi complessi.

Nello specifico di questa tesi, grazie alla serie di dati storici forniti dall'applicativo aziendale OptitEPM, è stato possibile costruire il modello avendo a disposizione le informazioni riguardanti:

1. mese (in lettere e numerico)
2. giorno della settimana (in lettere e numerico)
3. giorno festivo o feriale
4. stagione termica (che corre dal 15 ottobre al 15 aprile)
5. temperatura minima e massima
6. calore rilevato

L'ultima informazione riguarda il calore misurato in uscita dall'impianto, ovvero il dato a consuntivo di tipo certo.

Mediante la classificazione, a partire dall'insieme determinato di classi (training set), si costruisce il modello predittivo generando regole associative da alberi di decisione, attraverso il metodo denominato **M5Rules**, estensione dell'algoritmo base M5.

M5, inventato da Quinlan (1992) [50] e migliorato da Yong Wang (1996) [51], è un popolare approccio che deriva il modello matematico specifico da un quadro generale, per questo più adatto per prevedere il comportamento di sistemi complessi e di data mining. Il modello ad albero M5 combina funzionalità di classificazione e di regressione in un modo semplice ed efficace; una regressione strutturata ad albero si basa sul presupposto che la dipendenza funzionale tra i valori di ingresso e uscita non è costante nel dominio delle variabili esplicative, ma può essere considerato come tale in sottodomini più piccoli. Pertanto, M5 rappresenta la divisione del dominio in input in sottodomini.

Metodo utilizzato in passato dall'azienda per il *progetto SPRINT* [52] riguardante l'ottimizzazione delle risorse agli sportelli di Hera, viene illustrato in figura mediante sia il partizionamento del dominio di ingresso sia la costruzione del rispettivo albero per ogni specifico sottodominio. Qui vengono costruiti diversi modelli a seconda del giorno della settimana e sul livello dell'attività di fatturazione, mentre sull'asse y è rappresentato il numero totale di arrivi.

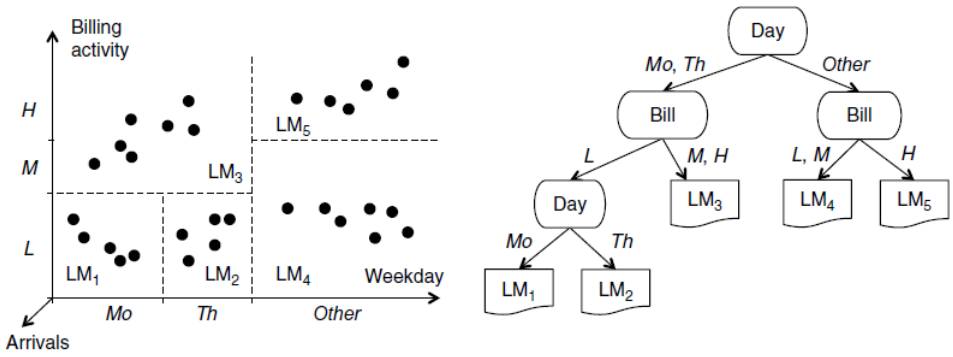


Figura 3.6: Algoritmo M5 - Fonte articolo: S*PRINT: Optimization of Staff Management for Desk Customer Relations Services at Hera* - Vigo, Caremi, Gordini

Weka fornisce l'estensione del metodo M5 attraverso l'implementazione della

classe M5Rules. Approfondito da Holmes, Hall e Prank (1999) questo metodo genera regole da un model tree e opera nel seguente modo: un model tree viene applicato interamente al dataset in modo da sviluppare un albero con potature (pruned). Successivamente, si va alla ricerca della miglior foglia tramite le regole ai nodi. Tutte le istanze corrispondenti alla regola sono rimosse dal dataset. Il processo è applicato ricorsivamente alle istanze rimanenti e termina quando tutte sono coperte da una o più regole. Questo approccio genera una lista di decisione per la regressione sfruttando la politica del «divide et impera»; anzichè costruire una singola regola, come si fa abitualmente, si genera un modello di sottoalbero da ogni nodo che possiede una regola per trovar la miglior foglia.

La costruzione di alberi parziali consente di migliorare l'efficienza computazionale, allo stesso modo non si intacca l'accuratezza e il peso delle regole.

A fronte di una mole di dati reali, modellati attraverso metodi forniti dal software Weka e non solo, si vuole fornire un'attenta analisi esplorando diverse strategie. Il capitolo successivo illustrerà dunque la valutazione sperimentale che ha permesso di individuare la più valida configurazione, che conferisce i migliori risultati.

Capitolo 4

Analisi e Valutazioni

4.1 Prima Analisi: M5 con meno attributi

L'obiettivo di questa prima analisi riguarda la validazione del metodo M5 già modellato in passato, con particolare cura all'analisi degli attributi. Si vuole, infatti, comprendere la perdita (o il guadagno) ottenuta/o dall'eliminazione di alcune dimensioni.

Sono disponibili i dati consuntivi su uno storico di qualche anno di sei impianti di teleriscaldamento, che per motivi di privacy si esprimono come *P1,P2...P6*. Nel dettaglio le dipendenze del fenomeno riguardano:

- **mese valore numerico** (1,2,3, ...)
- **mese in lettere** (gennaio, febbraio, ...)
- **settimana valore numerico** (1,2,3, ...)
- **giorno della settimana** (lunedì, martedì, ...)
- giorno festivo o feriale (true, false)

- **stagione termica** (true, false)
- temperatura minima
- temperatura massima
- valore calore (numerico in KWh)

Questi dati vengono estratti dal database di OptitEPM e convertiti in un file *arff* per la generazione del modello in Weka.

Nella prima fase di pre-processing del programma è possibile effettuare l'analisi sugli attributi. Alcuni di questi, evidenziati in grassetto, vengono eliminati poichè si pensa che possano essere poco significativi e ridondanti. In Weka tale funzione è svolta dal filtro *Remove* che cancella appunto uno specifico set di attributi dal dataset di partenza.

Dopo aver preparato i dati in input e creato il modello di previsione con Weka, vengono collezionati i risultati del forecasting in un file *csv*. Su questo file, grazie ai dati giornalieri consuntivi e previsti , è possibile calcolare:

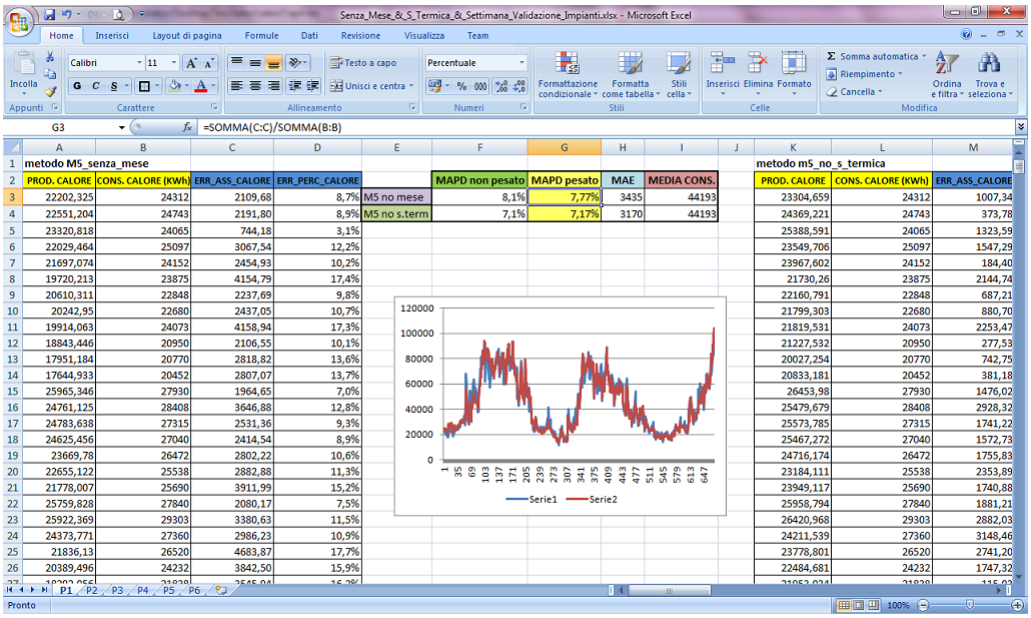
- l'*errore previsionale assoluto* (AD) corrispondente a $|D_t - P_t|$ (differenza tra il valore effettivo della domanda e il valore previsto in un determinato giorno)
- l'*errore previsionale percentuale* (PD) corrispondente a $\frac{100*|D_t-P_t|}{D_t}$ (rapporto percentuale tra l'errore previsionale e il valore effettivo della domanda in un determinato giorno)
- *MAPD non pesato* corrispondente a $\frac{100}{n}\sum_{t=1}^{n}\frac{|D_t-P_t|}{D_t}$ (media di PD)

- *MAPD pesato* definito come $100 * \frac{\sum_{t=1}^{n} AD}{\sum_{t=1}^{n} D_t}$(rapporto percentuale tra la somma di AD e la somma del valore effettivo nell'intero periodo)

- l'*errore previsionale medio assoluto* (MAD) corrispondente a $\frac{1}{n}\sum_{t=1}^{n} |D_t - P_t|$ (media dell'errore assoluto)

- la *media del consuntivo* corrispondente a $\frac{1}{n}\sum_{t=1}^{n} D_t$

Per una migliore comprensione di questi calcoli, si riporta un esempio esplicativo.



	A	B	C	D	E	F	G	H	I	J	K	L	M
1	metodo M5_senza_mese										metodo m5_no_s_termica		
2	PROD. CALORE	CONS. CALORE (KWh)	ERR_ASS_CALORE	ERR_PERC_CALORE		MAPD non pesato	MAPD pesato	MAE	MEDIA CONS.		PROD. CALORE	CONS. CALORE (KWh)	ERR_ASS_CALORE
3	22202,325	24312	2109,68	8,7%	M5 no mese	8,1%	7,77%	3435	44193		23304,659	24312	1007,34
4	22551,204	24743	2191,80	8,9%	M5 no s.term	7,1%	7,17%	3170	44193		24369,221	24743	373,78
5	23320,818	24065	744,18	3,1%							25388,591	24065	1323,59
6	22029,464	25097	3067,54	12,2%							23549,706	25097	1547,29
7	21697,074	24152	2454,93	10,2%							23967,602	24152	184,40
8	19720,213	23875	4154,79	17,4%							21730,26	23875	2144,74
9	20610,311	22848	2237,69	9,8%							22160,791	22848	687,21
10	20242,95	22680	2437,05	10,7%							21799,303	22680	880,70
11	19914,063	24073	4158,94	17,3%							21819,531	24073	2253,47
12	18843,446	20950	2106,55	10,1%							21227,532	20950	277,53
13	17951,184	20770	2818,82	13,6%							20027,254	20770	742,75
14	17644,933	20452	2807,07	13,7%							20833,181	20452	381,18
15	25965,346	27930	1964,65	7,0%							26453,98	27930	1476,02
16	24761,125	28408	3646,88	12,8%							25479,679	28408	2928,32
17	24783,638	27315	2531,36	9,3%							25573,785	27315	1741,22
18	24625,456	27040	2414,54	8,9%							25467,272	27040	1572,73
19	23669,78	26472	2802,22	10,6%							24716,174	26472	1755,83
20	22655,122	25538	2882,88	11,3%							23184,111	25538	2353,89
21	21778,007	25690	3911,99	15,2%							23949,117	25690	1740,88
22	25759,828	27840	2080,17	7,5%							25958,794	27840	1881,21
23	25922,369	29303	3380,63	11,5%							26420,968	29303	2882,03
24	24373,771	27360	2986,23	10,9%							24211,539	27360	3148,46
25	21836,13	26520	4683,87	17,7%							23778,801	26520	2741,20
26	20389,496	24232	3842,50	15,9%							22484,681	24232	1747,32

Figura 4.1: Esempio foglio di calcolo

Successivamente, mediante lo studio del **MAPD pesato** per ogni impianto, vengono confrontati i valori per ogni impianto.

La tabella che segue riporta il risultato percentuale, per ogni impianto di teleriscaldamento (P1...P6), per i quali si calcola la media. Nella seconda colonna, invece, si riportano i valori ottenuti dal modello M5Rules con mese, da cui si ef-

fettua il confronto percentuale (DELTA). La colonna evidenziata indica il risultato migliore.

	M5 CON MESE	**M5 NO MESE**	*DELTA*	**M5 NO S.TERM**	*DELTA*	**M5 NO SETT.**	*DELTA*
P1	7,44%	7,77%	0,33%	7,17%	-0,27%	7,05%	-0,39%
P2	17,97%	21,34%	3,36%	19,39%	1,42%	19,34%	1,37%
P3	22,56%	24,01%	1,45%	23,34%	0,78%	22,50%	-0,06%
P4	8,35%	9,71%	1,37%	8,36%	0,01%	8,42%	0,07%
P5	14,41%	17,82%	3,41%	15,96%	1,55%	15,57%	1,16%
P6	8,72%	10,43%	1,71%	8,98%	0,27%	8,71%	-0,01%
MEDIA	13,24%	15,18%	**1,94%**	13,87%	**0,63%**	13,60%	**0,36%**

Tabella 4.1: *risultati prima analisi - M5 con meno attributi*

Come si evince dal *DELTA*, definito appunto come la differenza percentuale tra il modello preso in esame e il modello M5Rules, la perdita dell'informazione dal dataset mese/mese in lettere, settimana/settimana in lettere oppure stagione termica non esibisce alcun miglioramento consistente. Probabilmente, però, delle tre dimensioni l'attributo riguardante la settimana risulta quello più ridondante; l'eliminazione della settimana, infatti, nel file *arff* non provocherebbe una perdita significativa.

4.2 Seconda Analisi: M5 vs Linear Regression

L'algoritmo ibrido M5Rules adottato nella prima analisi e presente in Weka, combina in modo semplice ed efficace caratteristiche di classificazione e regressione. Per validare ulteriormente tale tecnica si è scelto di applicare la seconda

analisi allo studio della regressione. Per la sua versatilità questa tecnica trova impiego nel campo delle scienze applicate: dalla chimica, biologia, ingegneria, nonchè nelle scienze sociali: economia, psicologia.

La creazione del modello in Weka è stata effettuata mediante la classe `Linear Regression` del package `weka.classifier.functions`. Tale metodo ha come scopo l'individuazione di una combinazione lineare di variabili indipendenti (o predittori) per predire in modo ottimale il valore assunto dalla variabile dipendente (o variabile di risposta).

L'obiettivo della seconda analisi verte, pertanto, in una prova del metodo Linear Regression (LR), al fine di validare l'utilizzo del metodo M5 con gli attributi rilevanti. Come ulteriore test, infatti, si eliminano le stesse dimensioni prese in esame precedentemente.

	LR CON MESE	**LR NO MESE**	*DELTA*	**LR NO S.TERM**	*DELTA*	**LR NO SETT.**	*DELTA*
P1	9,58%	12,45%	5,01%	9,58%	2,15%	9,58%	2,14%
P2	19,27%	28,70%	10,73%	19,94%	1,97%	19,91%	1,94%
P3	25,41%	29,07%	6,52%	25,35%	2,80%	24,71%	2,16%
P4	12,49%	18,93%	10,58%	12,49%	4,14%	12,56%	4,22%
P5	16,76%	21,15%	6,74%	17,06%	2,65%	16,75%	2,34%
P6	13,15%	18,35%	9,63%	13,20%	4,48%	13,19%	4,47%
MEDIA	16,11%	21,44%	**8,20%**	16,27%	**3,03%**	16,12%	**2,88%**

Tabella 4.2: *risultati seconda analisi - M5 vs Linear Regression*

Dalla tabella sopra riportata, osservando il DELTA, non emergono notevoli miglioramenti. Tale risultato lo si può vedere anche attraverso i grafici che mo-

strano l'andamento qualitativo della previsione di calore rispetto al dato reale dell'impianto.

La prima immagine, infatti, validata con il metodo M5Rules, mostra una previsione piuttosto fedele al dato a consuntivo.

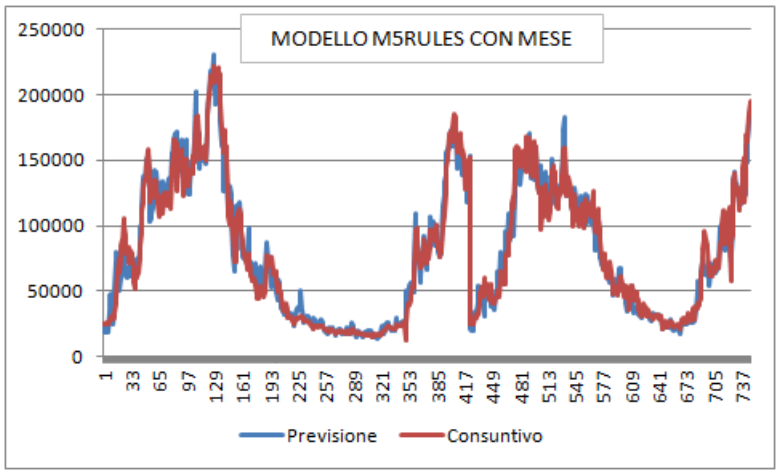


Figura 4.2: Grafico modello M5 - Impianto P4

La seconda previsione, invece, che è calcolata con il metodo LR, è chiaramente distante dall'osservazione reale vista la presenza di diversi outliers (picchi).

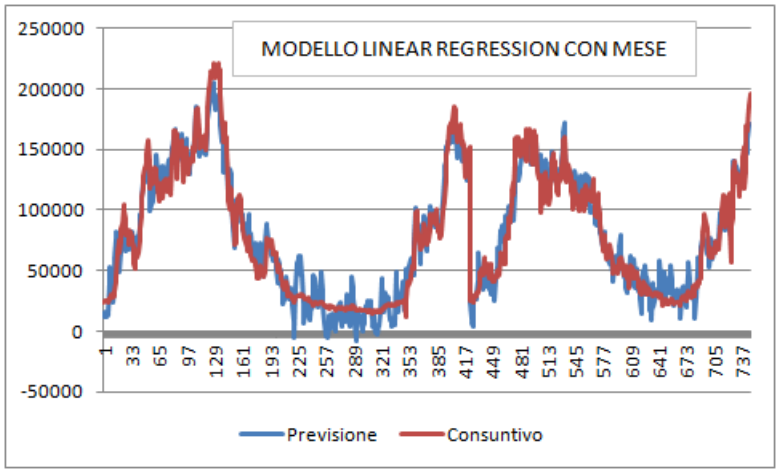


Figura 4.3: Grafico modello LR - Impianto P4

Un semplice confronto tra i risultati di previsione ottenuti fino a questo punto, mette in luce che il modello M5Rules con Weka sia migliore in termini di attendibilità delle previsioni. Tuttavia, data la presenza in letteratura di altri validi metodi, vengono ora presi in esame la *media mobile* e lo *smorzamento esponenziale*.

4.3 Terza Analisi: M5 vs altri metodi

Dalla letteratura sono descritte numerose metodologie previsionali, ma non tutte sono ugualmente costose da adottare e mantenere. In genere, le tecniche d'analisi delle serie storiche, come la media mobile e lo smorzamento esponenziale, sono relativamente economiche da implementare, grazie anche alla drastica riduzione dei costi della capacità automatica di calcolo.

Per questo valido motivo si studiano, in ordine di complessità, le previsioni a media mobile e, a seguire, lo smorzamento esponenziale. Espressa la formula matematica nel secondo capitolo, si riporta subito un esempio elaborato per il primo impianto (P1) nel giorno 3 settembre 2012, sulla base della domanda termica, pari a 24312 e 24743 kilowattora nei due giorni precedenti:

$$P_{(3Settembre2012)} = \frac{24312 + 24743}{2} = 24527,5[kWh] \tag{4.1}$$

Notevole importanza ha la scelta del numero ottimale r che corrisponde ai giorni presi in considerazione. Un periodo troppo lungo potrebbe provocare una scarsa reattività nel riflettere i cambiamenti. La tabella mostra quanto enunciato.

GIORNO	D_T	P_T r = 2	P_T r = 7
01/09/2012	24312		
02/09/2012	24743		
03/09/2012	24065	24527	
04/09/2012	25097	24404	
05/09/2012	24152	24581	
06/09/2012	23875	24624	
07/09/2012	22848	24013	24156
08/09/2012	22680	23361	23922
09/09/2012	24073	22764	23827
10/09/2012	20950	23376	23200
11/09/2012	20770	22511	23382
12/09/2012	20452	20860	22764
...			
MAPD pesato		**8,08%**	**11,35%**

Tabella 4.3: *Previsione della domanda su media mobile con r = 2 e r = 7*

Già graficamente appare evidente quale sia il valore di r più corretto, ma per determinarlo in maniera univoca si ricorre al calcolo dell'errore minimo di previsione, come ad esempio, il MAPD pesato (evidenziato in tabella).

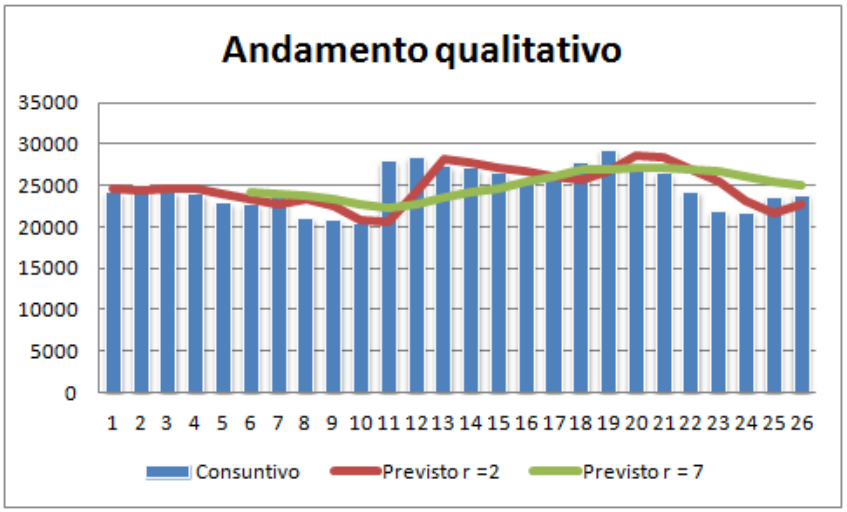


Figura 4.4: Media mobile con periodo uguale a 3 o 7 giorni

Come mostrato in figura, all'aumentare dell'ordine della media, si ottiene un maggiore smussamento della variabilità dei dati e un effetto ritardato, nel senso che i massimi e minimi della media mobile si presentano sfalsati in ritardo rispetto a quelli presenti nella serie storica. Per questi motivi si è scelta di adottare una serie di ordine 2.

In un secondo tempo, come affinamento di questa tecnica, è stato introdotto lo smorzamento esponenziale (Exponential Smoothing o anche detto Modello di Brown) che rappresenta un tipo di media mobile ponderata.

A titolo esemplificativo, sempre in riferimento all'impianto di teleriscaldamento P1, si suppone che la previsione della domanda termica per il periodo trascorso fosse di 24527 kWh, mentre la domanda effettiva fosse stata di 24065 kWh. Supposto, inoltre, che il fattore alfa sia di 0,2%.

Il calcolo si presenta come segue:

$$P_{(4Settembre2012)} = 0,2 * 24065 + (1 - 0,2) * 24527 = 24435[kWh] \qquad (4.2)$$

Uno dei problemi di questa tecnica risulta ancora una volta la determinazione del valore da attribuire al fattore α. Con un fattore 1, per esempio, il risultato che si ottiene è quello di assumere come previsione della domanda il valore del dato reale dell'ultimo periodo trascorso. L'uso di un valore molto basso, al contrario, avrebbe l'effetto di ridurre la previsione, in pratica, a una semplice media mobile. Qui di seguito viene riportato un esempio che evidenzia gli impatti di una differente scelta del parametro alfa. La colonna evidenziata mostra, come sempre, il miglior risultato.

GIORNO	D_T	P_T alfa = 0,2	P_T alfa = 0,7
01/09/2012	24312		
02/09/2012	24743		
03/09/2012	24065	24527	
04/09/2012	25097	24435	
05/09/2012	24152	24567	
06/09/2012	23875	24484	
07/09/2012	22848	24362	24156
08/09/2012	22680	24059	23122
09/09/2012	24073	23783	23827
10/09/2012	20950	23841	23787
11/09/2012	20770	23263	21801
12/09/2012	20452	22764	21079
...			
MAPD pesato		**10,65%**	**7,45%**

Tabella 4.4: *Previsione della domanda su modello di Brown con alfa = 0.2 e alfa = 0.7*

Per misurare l'accuratezza di entrambi i metodi di previsione rispetto al metodo M5Rules, si mostra una sintesi del MAPD pesato con valore $r = 2$ e $\alpha = 0,7$ per tutto lo storico di ogni impianto a disposizione.

	M5 CON MESE	**M.MOBILE** R = 2	*DELTA*	**S.ESPON.** $\alpha = 0.7$	*DELTA*
P1	7,44%	8,08%	0,64%	7,45%	0,01%
P2	17,97%	19,59%	1,61%	17,93%	-0,05%
P3	22,56%	11,26%	-11,29%	10,32%	-12,23%
P4	8,35%	8,48%	0,13%	7,88%	-0,46%
P5	14,41%	11,47%	-2,94%	10,76%	-3,65%
P6	8,72%	7,76%	-0,96%	7,17%	-1,55%
MEDIA	13,24%	11,11%	**- 2,13%**	10,25%	**- 2,99%**

Tabella 4.5: *risultati terza analisi - M5 vs altri metodi*

Dalla tavola appena illustrata emerge che questa terza analisi apporta dei miglioramenti più significativi relativamente alla previsione effettuata con la tecnica

di smorzamento esponenziale. Tuttavia i calcoli eseguiti non prevedono l'utilizzo di Weka, ma di un semplice foglio di calcolo Excel e, per questo motivo, non sono fortemente correlabili con le precedenti prove. Ciò non toglie che, in un futuro, grazie all'affidabilità di questo metodo, possa essere reso facilmente automatizzabile.

4.4 Quarta analisi: effetto profondità del dataset

Dopo aver analizzato alcune metodologie previsionali e validato il metodo M5Rules come approccio che ottiene una maggiore qualità della previsione in Weka, si introduce la quarta analisi.

Avendo a disposizione tutto l'orizzonte storico per ogni impianto di teleriscaldamento, si vuole intendere come profondità l'informazione che verrà aumentata o diminuita rispetto al numero di record. In particolare, tale analisi sul forcasting EPM mediante Weka, si scinderà in:

- algoritmo che amplia il tracciato del file *arff* introducendo nel dataset valore e temperature di n giorni precedenti ($A4_1$)

- algoritmo che esclude le condizioni di anomalie perchè fisicamente impossibili per gli impianti ($A4_2$)

L'obiettivo di tali prove riguarda ancora il miglioramento dell'accuratezza previsionale, sperimentando nuovi approcci implementativi.

4.4.0.1 Algoritmo A4(1)

Per quanto concerne la prima analisi, denominata $A4_1$, si può misurare la forza con cui le variabili si associano linearmente attraverso il coefficiente di correlazione, introdotto nel secondo capitolo.

Formalmente, avendo $(x_1, y_1)...(x_n, y_n)$ dati sperimentali, il coefficiente di correlazione risulta:

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2 * \sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{4.3}$$

In particolare, si ha che:

- $r_{xy} > 0$ le variabili sono correlate positivamente
- $r_{xy} < 0$ le variabili sono correlate negativamente
- $r_{xy} = 0$ le variabili non sono correlate

Per calcolarlo in Excel si fa uso della funzione di CORRELAZIONE[SERIE1, SERIE2] dove *serie1* rappresenta il valore di calore in un determinato giorno, e *serie2* riporta il valore di calore per gli n giorni precedenti. Mediante l'esaminazione di questa funzione su un numero statisticamente significativo di dati sperimentali, si ottiene una rappresentazione grafica (diagramma a dispersione) che possa suggerire quanti giorni siano più adatti a descrivere il fenomeno. Sull'asse x si trovano i giorni, mentre sull'asse y i coefficienti di correlazione.

Il valore di r è > 0, e quindi la correlazione è positiva; inoltre, il coefficiente assume un valore abbastanza alto, e ciò dimostra che la correlazione è buona. In altri termini, le due variabili vanno di pari passo, nel senso che quando aumenta

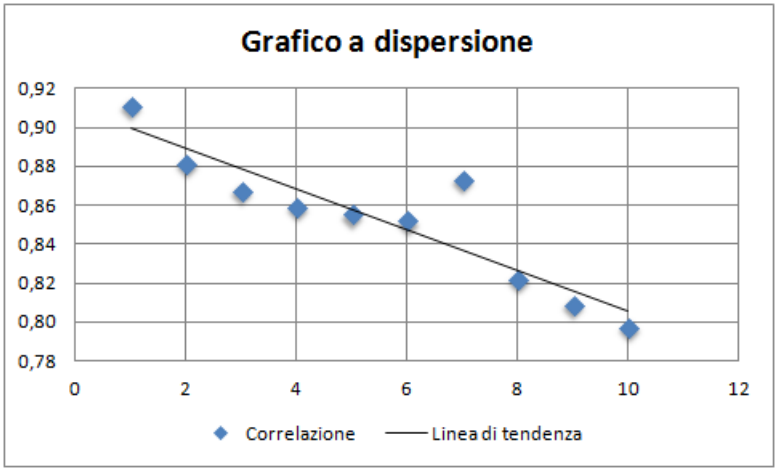


Figura 4.5: Correlazione per le serie di n giorni prima

il valore dell'una aumenta proporzionalmente anche il valore dell'altra (giorno precedente).

Da questo studio si decide di ampliare il tracciato del file *arff* inserendo il valore e la temperatura fino ai 7 giorni precedenti, settimo giorno che tuttavia, rispetto alla linea di tendenza, si trova più spostato.

La creazione del modello in Weka è stata dunque effettuata mediante la classe `M5Rules` del package `Weka.classifiers.rules`. Al fine di ottenere la miglior prestazione previsionale, sono state testate tre differenti configurazioni:

1. dataset esteso con le informazioni (valore e temperatura minima e massima) di esattamente 7 giorni precedenti
2. dataset esteso con informazioni del giorno precedente
3. dataset esteso con informazioni del settimo giorno precedente

Ancora una volta, per verificare la bontà dell'algoritmo viene calcolato il MAPD pesato per ogni impianto, sul quale poi si esegue la media.

	M5	**info -7**	*DELTA*	**info -1**	*DELTA*	**info solo 7**	*DELTA*
P1	7,44%	4,67%	-2,77%	4,79%	-2,65%	6,58%	-0,86%
P2	17,97%	12,76%	-5,22%	14,02%	-3,95%	17,02%	-0,95%
P3	22,56%	8,30%	-14,26%	8,53%	-14,03%	14,64%	-7,92%
P4	8,35%	4,40%	-3,95%	4,67%	-3,68%	7,13%	-1,22%
P5	14,41%	9,96%	-4,45%	10,49%	-3,91%	11,95%	-2,46%
P6	8,72%	4,25%	-4,47%	4,31%	-4,41%	7,55%	-1,17%
MEDIA	13,24%	7,39%	**-5,85%**	7,80%	**-5,44%**	10,81%	**-2,43%**

Tabella 4.6: *risultati quarta analisi - Effetto profondità del dataset A4(1)*

È evidente che l'effetto sulla profondità del dataset, relativo alle informazioni della settimana appena precedente, determina in media un buon risultato. Per di più, l'aggiunta di un solo giorno garantisce comunque una buona stima. Al contrario, l'effetto del settimo giorno sul dataset, migliora in minor parte, circa come il risultato ottenuto dallo smorzamento esponenziale.

4.4.0.2 Algoritmo A4(2)

A causa della presenza di valori molto elevati dell'errore previsionale percentuale (PD) nella costruzione dei diversi modelli in Weka, si è pensato di classificare tali giornate come «anomale». Seppur l'applicativo OptitEPM selezioni e scarti a monte le anomalie, risulta comunque opportuno eliminare i record con $PD > 50\%$.

Sebbene il calcolo del MAPD evidenzi un margine di miglioramento, non si considera quest'ultima analisi particolarmente significativa, soprattutto perchè non è verificata da tutti gli impianti.

	M5	no anomalie	*DELTA*
P1	7,44%		
P2	17,97%	14,35%	-3,62%
P3	22,56%	16,83%	-5,72%
P4	8,35%	8,15%	-0,20%
P5	14,41%	13,57%	-0,84%
P6	8,72%	7,48%	-0,85%
MEDIA	13,24%	12,16%	**-2,24%**

Tabella 4.7: *risultati quarta analisi - Effetto profondità dello storico A4(2)*

Emerge, tuttavia, che sia necessario in futuro svolgere un'approfondita analisi su tali *noisy data*. Si tratta, infatti, di dati con valori decisamente più alti da quelli che sarebbe lecito attendersi. Le osservazioni in cui occorrono questi noisy values vengono chiamate «outliers» e la loro presenza dipende da diversi fattori, che richiederanno in futuro uno studio più approfondito .

Analogamente a queste prove, infine, si è svolta un'analisi che ha limitato lo storico a soli tre mesi (ottobre, novembre, dicembre). Per questo motivo, non potendola confrontare con le prove effettuate sull'intero arco temporale, si è scelto di illustrarla nella sezione *Appendice A* di questa tesi, con il nome " effetto profondità dello storico ".

4.5 Utilizzo di altro sofware con Metodo ARIMA

L'analista non solo deve padroneggiare le tecniche e gli algoritmi che il data mining mette a disposizione, ma deve anche possedere una conoscenza dei mezzi che il mercato offre per sviluppare modelli predittivi sempre più precisi. In

particolare, per fornire nuove funzionalità al problema di business, si è ritenuto opportuno eseguire un'ultima analisi utilizzando il software avanzato e versatile SPSS.

SPSS, acronimo di *Statistical Package for Social Science*, è un sistema di analisi dei dati molto potente che offre un'esauriente gamma di funzioni per la gestione e la trasformazione dei dati, la classificazione e l'analisi statistica, la cui prima versione è stata realizzata nel 1968. Oggi è uno tra i programmi più utilizzati, grazie ad un ambiente grafico intuitivo che consente di utilizzare menu descrittivi e semplici finestre di dialogo per eseguire automaticamente svariate operazioni.

L'Editor dei dati, infatti, offre uno strumento simile ad un foglio di calcolo semplice ed efficiente per l'immissione di dati e la visualizzazione dei file in esame.

Il modulo «Previsioni» fornisce due procedure che possono essere utilizzate per creare modelli e previsioni:

1. La procedura *Modelli serie storiche* crea modelli per le serie storiche e genera previsioni. La procedura comprende anche *Expert Modeler* che stabilisce automaticamente il modello migliore per ciascuna serie storica. Gli analisti esperti, per avere un maggior grado di controllo, possono usare questo strumento anche per creare modelli personalizzati.

2. La procedura *Applica modelli di serie storiche* applica i modelli delle serie storiche esistenti, creati nella procedura precedente. Ciò consente di effettuare previsioni per le serie che contengono dati nuovi o rivisti senza dover ricreare il modello.

La procedura *Modelli di serie storiche*, in particolare, consente di stimare i modelli di livellamento esponenziale, i modelli ARIMA e di generare previsioni. Tale procedura include un *Expert Modeler* che identifica e stima automaticamente il modello più adatto per una o più serie di variabili dipendenti ed eliminando quindi la necessità di identificare un modello appropriato tramite prove ed errori.

L'obiettivo di quest'ultima analisi si è dunque focalizzato sul confronto tra il metodo ARIMA, descritto nel secondo capitolo e presente in SPSS, e il metodo M5Rules modellato in Weka.

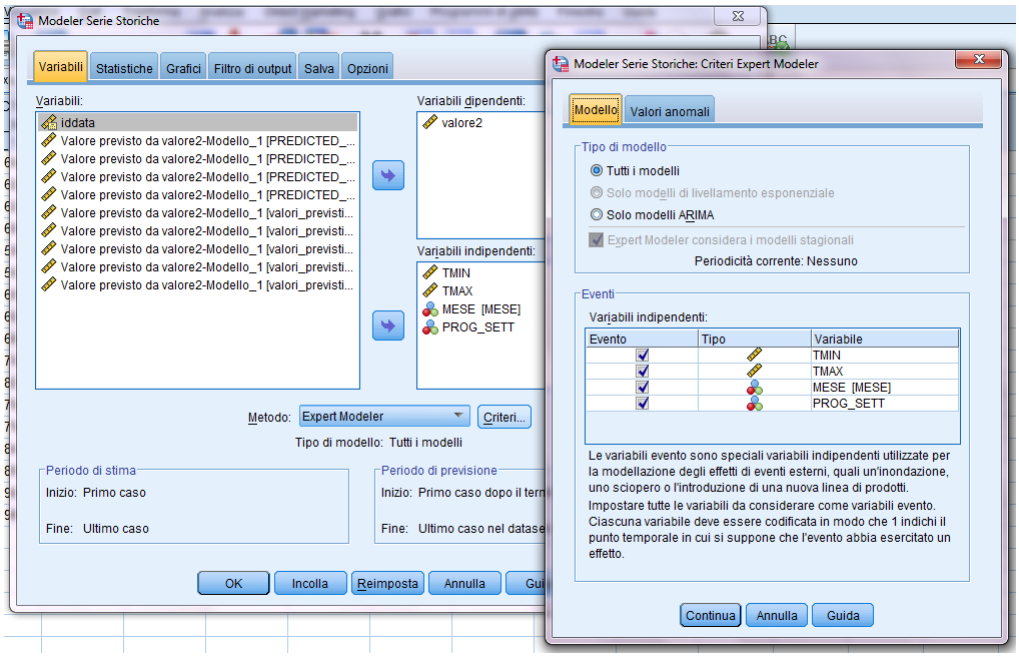


Figura 4.6: Modeler Serie Storiche Impostazioni

Per utilizzare il modello, dalla finestra di Editor dei dati è stato possibile scegliere il percorso (Analizza -> Previsioni -> Crea modelli..). Scelta la variabile *Valore* come dipendente, nei modelli ARIMA si considerano anche le variabili indipendenti *Tmin*, *Tmax*, *Mese* e *Progressivo della settimana*. Una considerazione sui

dati riguarda il fatto che le variabili devono essere numeriche, limite che non consente di modellare tutte le dipendenze del fenomeno.

Nei modelli ARIMA, inoltre, è possibile attivare il rilevamento automatico dei valori anomali (outliers). Si tratta, infatti di cambiamenti di livello di una serie storica che non è possibile spiegare. Tali osservazioni sono incoerenti rispetto al resto della serie, incidendo considerevolmente sull'analisi e, di conseguenza, influiscono sulla capacità di previsione del modello.

La figura che segue mostra tipi di valori anomali rilevati di frequente nelle serie storiche e analizzati nella prova. Le righe blu rappresentano una serie priva di valori anomali, mentre le righe rosse suggeriscono lo schema che potrebbe essere presente se la serie contenesse dei valori anomali.

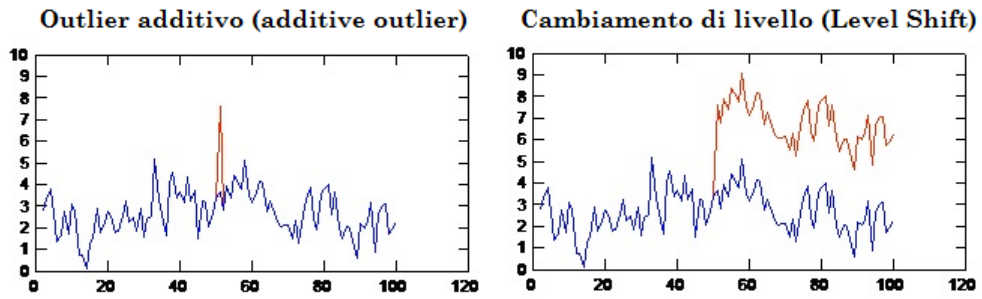


Figura 4.7: Modeler Serie Storiche Impostazioni

Per impostazione predefinita, tali outliers non vengono generalmente rilevati, ma ai fini dell'analisi svolgeremo anche la previsione che li considera, denominandola *ARIMA no anomalie*.

I risultati dell'esecuzione di tale procedura statistica vengono visualizzati nel *Viewer*. L'output può essere generato sottoforma di tabella, grafico o testo, a

seconda delle opzioni scelte per l'esecuzione della procedura. In particolare, la tabella descrittiva del modello contiene una voce per ciascun modello stimato. I tipi di modello ARIMA vengono elencati utilizzando la notazione standard di ARIMA (p,d,q)(P,D,Q), dove p è l'ordine dell'autoregressione, d è l'ordine delle differenze (o delle integrazioni) e q è l'ordine delle medie mobili. P, D e Q sono i corrispondenti stagionali.

Registrando i valori stimati del modello si può calcolare il MAPD pesato e confrontarlo con i risultati ottenuti con M5.

	M5	**ARIMA**	*DELTA*	**ARIMA no anomalie**	*DELTA*
P1	7,44%	5,98%	-1,46%	5,81%	-1,63%
P2	17,97%	16,30%	-1,67%	14,39%	-3,59%
P3	22,56%	23,24%	0,69%	22,54%	-0,01%
P4	8,35%	12,79%	4,44%	6,06%	-2,29%
P5	14,41%	13,57%	-0,84%	13,27%	-1,14%
P6	8,72%	6,73%	-1,99%	4,64%	-4,08%
MEDIA	13,24%	13,10%	**-0,14%**	11,12%	**-2,12%**

Tabella 4.8: *risultati ultima analisi - Utilizzo di altro software con metodo ARIMA*

Le ultime prove svolte hanno portato al raggiungimento di un medio risultato, che non supera quello ottenuto con i classificatori in Weka. Tuttavia, il potenziale dell'applicativo non esclude la possibilità che se ne possano effettuare altre migliorative, a fronte del fatto che l'efficace workbench, offra un'ampia gamma di tecniche di creazione di modelli di ottimizzazione tra i più competitivi.

Conclusioni

Il data mining rappresenta senza dubbio una nuova frontiera di analisi, in continua espansione nelle diverse aree di business. I suoi pilastri, e cioè le tecniche, i dati e la modellazione, rappresentano tre aree strategiche di conoscenze indispensabili perchè l'analisi dei dati sia efficace.

Nel presente lavoro di tesi, svolto in collaborazione con Optit srl, sono stati perseguiti parallelamente diversi obiettivi. In primo luogo è stata effettuata una validazione del modello M5Rules adottato. Tale algoritmo, presente in Weka, utilizza in modo semplice ed efficace caratteristiche di classificazione e regressione. Analizzando, in un primo tempo, le dipendenze del fenomeno sul dataset e poi, successivamente, la perdita derivata dall'utilizzo della Regressione Lineare, si è dimostrato che l'algoritmo M5Rules è l'unico mudulo in Weka in grado di elaborare le previsioni, adattandosi alle particolarità della serie storica in input.

In secondo luogo, l'obiettivo successivo, si è esplicato nell'evoluzione del metodo M5Rules. L'inserimento di nuove informazioni all'interno del dataset (relative alla temperatura e al valore del calore prodotto dalla centrale di teleriscaldamento) ha permesso di abbassare complessivamente gli errori previsionali, assicurando il best fitting del modello. Inoltre, nuove soluzioni di forecasting, senza far uso del software Weka, sono state implementate anche grazie allo studio

di algoritmi, come la media mobile e lo smorzamento esponenziale che, mediante calcoli puramente analitici, hanno permesso di utilizzare le informazioni del passato come valido strumento per la simulazione del futuro.

In ultimo, si è conseguita una breve analisi sullo strumento di calcolo previsionale. Weka e SPSS rappresentano oggi delle soluzioni di alto livello per la maturità del software e per i risultati brillanti, che hanno captato l'attenzione di una vasta comunità scientifica. Seppur non sia possibile definire una casistica sulla base della quale effettuare la scelta di Weka piuttosto che SPSS: ci si basa solo sul contesto di analisi per prediligere lo strumento migliore. A tal proposito, le analisi modellate in ambiente SPSS con metodo ARIMA non hanno messo in luce risultati significativi come quelli ottenuti in Weka. Tuttavia, vista la potenzialità del prodotto, è auspicabile che in un futuro vengano compiute delle analisi più approfondite.

A margine di tali prove, si sono aperti, infine, nuovi orizzonti di analisi che catturano, sia la sensitivity alla lunghezza di una serie storica su periodi di tempo definiti (*appendice A*), sia alla bontà del profilo orario a disposizione (*appendice B*).

Appendice A

Effetto profondità dello storico (ridotto)

Questa analisi, marginale ma pur sempre rilevante, si è svolta limitando il training set a un arco temporale di soli 12 mesi, per ogni impianto. Tale storico, infatti, corre dal mese di dicembre 2013 a dicembre 2014. Per i mesi di dicembre, novembre ed ottobre 2014 si addestra il modello in Weka con le seguenti configurazioni:

1. lo storico intero fino al mese precedente di riferimento
2. lo storico dei 12 mesi precedenti
3. solamente lo storico del mese precedente

Relativamente a questi tre differenti casi, si collezionano i risultati e si calcola il MAPD complessivo degli ultimi 3 mesi come si può vedere in tabella.

	caso 1	caso 2	caso 3
P1	8,1%	8,9%	18,1%
P2	19,7%	18,9%	35,4%
P3	81,9%	82,9%	82,2%
P4	9,0%	11,4%	23,1%
P5	18,1%	13,0%	22,2%
P6	8,9%	9,6%	27,2%
MEDIA	24,3%	24,1%	34,7%

Tabella A.1: *risultati analisi - caso 1 - 2 - 3*

Prove di questo tipo, focalizzate solo su alcuni mesi, potrebbero aprire nuovi orizzonti d'analisi. La verifica, infatti, di quanto impatta nella capacità previsiva la sensibilità alla lunghezza della serie storica è un aspetto da non sottovalutare in futuro.

Appendice B

Previsioni in corso di giornata

Ad oggi, il forecasting di OptitEPM si distingue su due fondamentali aspetti: (1) la produzione giornaliera, (2) il profilo orario. Ciò significa che viene prima effettuato il forecasting in termini di produzione giornaliera poi viene riclassificato in base al profilo orario. Per esempio, se si prevede una domanda termica in un giovedì, si verifica nelle serie storiche tutti i giovedì che hanno prodotto calore, per poi calcolarne la media. Questo rappresenterà il profilo orario di tale giorno. Inoltre tale riprofilazione viene declinata su fasce orarie che si riferiscono al singolo quarto d'ora.

In subordine alle precedenti analisi, sui sei impianti di teleriscaldamento di riferimento, si è deciso di effettuare un confronto di curve sempre in termini di MAPD (pesato) complessivo.

Avendo a disposizione un foglio di calcolo Excel, si è determinato il valore totale a consuntivo per ciascuna fascia di ogni singolo giorno che indichiamo con il simbolo T_g.

Il confronto si è poi sviluppato distinguendo tre situazioni significative.

La prima determina il valore ottenuto moltiplicando il totale del giorno per le percentuali del giorno precedente alla stessa ora. Il calcolo si esprime, per chiarezza, nel seguente modo:

$$A = T_g . \frac{C_{g-1}}{T_{g-1}} \tag{B.1}$$

dove C_{g-1} e T_{g-1} sono rispettivamente il valore e il totale a consuntivo del giorno precedente.

La seconda, invece, ottiene il valore moltiplicandolo per le percentuali di 7 giorni precedenti alla stessa ora. Quindi il calcolo si esprime come:

$$B = T_g . \frac{C_{g-7}}{T_{g-7}} \tag{B.2}$$

Analogamente ai casi appena descritti, l'ultima prova calcola il valore moltiplicandolo per le percentuali dei 7,14,21,28 giorni precedenti:

$$C = T_g . \frac{\frac{C_{g-7}}{T_{g-7}} + \frac{C_{g-14}}{T_{g-14}} + ... + \frac{C_{g-28}}{T_{g-28}}}{4} \tag{B.3}$$

Bibliografia

[1] *Gartner Inc,* HTTP://WWW.GARTNER.COM/IT-GLOSSARY/BIG-DATA/

[2] *Holsheimer e M Siebes (1994),* DATA MINING: THE SEARCH FOR KNOWLEDGE IN DATABASES *Report CS-R9406. CWI. Amsterdam, The Netherlands*

[3] *Definizione Web,* HTTP://IT.WIKIPEDIA.ORG/WIKI/DATA-MINING

[4] *Rosenblatt Frank (1957),* THE PERCEPTRON A PERCEIVING AND RECOGNIZING AUTOMATON, *Report 85-460-1, Cornell Aeronautical Laboratory*

[5] *Usama Fayyad, Gregory Piatetsky-Shapiro e Padhraic Smyth (1994),* FROM DATA MINIG TO KNOWLEDGE DISCOVERY IN DATABASES, *American Association for Artificial Intelligence, Magazine*

[6] *Gary Miner, Robert Nisbet, John Elder,* HANDBOOK OF STATISTICAL ANALYSIS AND DATA MINING APPLICATIONS, *Elsevier*

[7] *Matteo Golfarelli,* INTRODUZIONE AL DATA MINING, *slide del corso 2013-2014*

[8] *Guandong Xu,Yu Zong,Zhenglu Yang (2013),* APPLIED DATA MINING, *CRC Press*

[9] *Leo Breiman (1996),* BAGGING PREDICTORS, *articolo*

[10] *J.H. Friedman (1999),* GREEDY FUNCTION APPROXIMATION: A GRADIENT BOOSTING MACHINE, *articolo*

[11] *Friendman, Hastie, Tibshirani (1998)* ADDITIVE LOGISTIC REGRESSION: A STATISTICAL VIEW OF BOOSTING, *articolo*

[12] *Quinlan (1993),* C4.5: PROGRAMS FOR MACHINE LEARNING, *Morgan Kaufmann Publishers*

[13] *Quinlan (1986),* INTRODUCTION OF DECISION TREE, *Machine learning 81-106*

[14] *Legendre (1805),* NOUVELLES METHODES POUR LA DETERMINATION DES ORBITES DES COMETES, *Firmin Didot, Paris*

[15] *Gauss (1809,* THEORIA MOTUS CORPORUM COELESTIUM IN SECTIONIBUS CONICIS SOLEM AMBIENTUM

[16] *Francis Galton (1877),* TYPICAL LAWS OF HEREDITY, *Nature 15*

[17] *Agrawal, R.; Imielinski, T.; Swami, A. (1993),* MINING ASSOCIATION RULES BETWEEN SETS OF ITEMS IN LARGE DATABASES, *SIGMOD international conference on Management of data*

[18] *Demystifying Market Basket Analysi (2009)*

[19] *Rakesh Agrawal and Ramakrishnan Srikant (1994),* HTTP://RAKESH.AGRAWAL-

FAMILY.COM/PAPERS/VLDB94APRIORI.PDF, *Proceedings of the 20th International Conference on Very Large Data Bases, VLDB*

[20] *Imdadullah,*HTTP://ITFEATURE.COM/TIME-SERIES-ANALYSIS-AND-FORECASTING/TIME-SERIES-ANALYSIS-FORECASTING*(2014), Basic Statistics and Data Analysis*

[21] *Rokach, Lior; Maimon, O. (2008),* DATA MINING WITH DECISION TREES: THEORY AND APPLICATIONS, *World Scientific Pub Co Inc*

[22] *Bailey, Ken (1994),* NUMERICAL TAXONOMY AND CLUSTER ANALYSIS, *Typologies and Taxonomies. p. 34*

[23] *Geisser, Seymour (1993),* PREDICTIVE INFERENCE: AN INTRODUCTION. NEW YORK: CHAPMAN & HALL

[24] *Ghiani, Laporte e Musmanno,* INTRODUCTION TO LOGISTIC SYSTEMS PLANNING AND CONTROL, *Wiley*

[25] *Chambers, Mulick e Smith (1971)* HOW TO CHOOSE THE RIGHT FORECASTING TECHNIQUE, *Harvard Business Review, pag 47-76*

[26] *D.C. Boes, F.A. Graybill, A.M. Mood (1988),* INTRODUZIONE ALLA STATISTICA, *McGraw-Hill Libri Italia*

[27] *Damiano Milanato(2008),* PROCESSI, METODOLOGIE E MODELLI MATEMATICI PER LA GESTIONE DELLA DOMANDA COMMERCIALE, *Google eBook*

[28] *Gianpaolo Ghiani, Gilbert Laporte, Roberto Musmanno (2013),* INTRODUCTION TO LOGISTICS SYSTEMS MANAGEMENT, *John Wiley & Sons*

[29] *Steven Wheelwright, Spyros Makridakis (1980),* FORECASTING METHODS FOR MANAGEMENT, *John Wiley & Sons*

[30] *Winters (1960),* FORECASTING SALES BY EXPONENTIALLY WEIGHTED MOVING AVERAGES, *Management science*

[31] *George E. P. Box; Gwilym Jenkins (1971),* HYPOTHESIS TESTING IN TIME SERIES ANALYSIS

[32] *Asteriou, Dimitros; Hall, Stephen G. (2011),* ARIMA MODELS AND THE BOX JENKINS METHODOLOGY, *Palgrave MacMillan*

[33] *Norman Dalkey, Olaf Helmer (1963),* AN EXPERIMENTAL APPLICATION OF THE DELPHI METHOD TO THE USE OF EXPERTS. MANAGEMENT SCIENCE

[34] *Bernice B. Brown (1968),* DELPHI PROCESS: A METHODOLOGY USED FOR THE ELICITATION OF OPINIONS OF EXPERTS

[35] *Sackman, H. (1974),* DELPHI ASSESSMENT: EXPERT OPINION, FORECASTING AND GROUP PROCESS

[36] *Harold A. Linstone, Murray Turoff (1975),* THE DELPHI METHOD: TECHNIQUES AND APPLICATIONS, *Addison-Wesley*

[37] *Thomas Birkland (2001),* AN INTRODUCTION TO THE POLICY PROCESS

[38] *Hogarth e Makridakis (1981),* FORECASTING AND PLANNING: AN EVALUATION, *Management Science*

[39] *Makridakis e Wheelwright(1997),* FORECASTING: ISSUES AND CHALLENGES FOR MARKETING MANAGEMENT, *Journal of marketing*

[40] *Scott Armstrong(1984),* FORECASTING BY EXTRAPOLATION: CONCLUSIONS FROM 25 YEARS OF RESEARCH, *Interfaces*

[41] *Smith e Wight (1978),* FOCUS FORECASTING: COMPUTER TECHNIQUES FOR INVENTORY CONTROL

[42] *Gardner (1985),* EXPONENTIAL SMOOTHING: THE STATE OF THE ART*Journal of forecasting*

[43] *Gse,* WWW.GSE.IT

[44] *Gruppo Hera Dossier,* HTTP://WWW.GRUPPOHERA.IT/GRUPPO/COM-MEDIA/DOSSIER-TLR/

[45] *Associazione Italiana Riscaldamento Urbano,* HTTP://WWW.AIRU.IT/TELERISCALDAMENTO/

[46] *Opet Seed,* HTTP://AMBIENTE.COMUNE.FORLI.FC.IT/

[47] *Weka sito ufficiale,* HTTP://WWW.CS.WAIKATO.AC.NZ/ML/WEKA/

[48] *Booth et al. (2006),* HYDROLOGIC VARIABILITY OF THE COSUM-NES RIVER FLOODPLAIN, *San Francisco Estuary and Watershed Science*

[49] WEIGHTED MOVING AVERAGES: THE BASICS, *Investopedia*

[50] *Quinlan (1992),* LEARNING WITH CONTINUOUS CLASSES

[51] *Yong Wang e Ian Witten (1996),* INDUCTION OF MODEL TREES FOR PREDICTING CONTINUOUS CLASSES

[52] *Daniele Vigo, Claudio Caremi, Angelo Gordini, Sandro Bosso, Giuseppe D'Aleo, Beatrice Beleggia (2014),* SPRINT: OPTIMIZATION OF STAFF MANAGEMENT FOR DESK CUSTOMER RELATIONS SERVICES AT HERA*, Interfaces*

[53] *Ian H.Witten e Eibe Frank,* MINING PRACTICAL MACHINE LEARNING TOOLS AND TECHNIQUES*, Elsevier Inc*

[54] *Michael J.A. Berry e Gordon S.Linoff,* DATA MINING TECHNIQUES FOR MARKETING, SALES, AND COSTOMER RELATIONSHIP MANAGEMENT*, Second Edition Wiley*

Elenco delle figure

Elenco delle tabelle

Glossario

Elenco dei termini e delle abbreviazioni più ricorrenti nel testo:

KDD

Knowledge Discovery in Databases: abbreviazione per il processo di apprendimento dei dati.

CRISP-DM

Cross Industry Standard Process for Data Mining: modello che descrive le fasi standard di un progetto di data mining.

ARMA/ARIMA

Autoregressive (Integrated) Moving Average: modello autoregressivo (integrato) a media mobile ideato da Box and Jenkins.

ME

Mean Average: abbreviazione per descrivere l'"errore medio".

MAD

Mean Absolute Deviation: abbreviazione per descrivere l'"errore medio assoluto".

MAPD

Mean Absolute Percentage Deviation: abbreviazione per descrivere l'"errore medio assoluto percentuale".

MSE

Mean Square Error: abbreviazione per descrivere l'"errore quadratico medio".

Optit-EPM

Energy Production Management di Optit: una soluzione di supporto alla programmazione di breve, medio e lungo termine degli impianti di cogenerazione realizzato dall'azienda Optit.

Weka

Waikato Enviroment for Knowledge Analysis: ambiente software open source, interamente scritto in Java, per l'apprendimento automatico.

ARFF

Attribute Relationship File Format: formato utilizzato dal software Weka per la lettura dei dati.

LR

Linear Regression: abbreviazione per descrivere la "Regressione Lineare".

SPSS

Statistical Package for Social Science: ambiente software realizzato da IBM per l'analisi dei dati.