Alma Mater Studiorum · Università di Bologna

SCUOLA DI SCIENZE
Corso di Informatica per il Management

PROGETTAZIONE E IMPLEMENTAZIONE DI UN DATA WAREHOUSE DI SUPPORTO ALLA PROFILAZIONE DEI CONSUMI ENERGETICI DOMESTICI

Relatore: Chiar.mo Prof.
MARCO DI FELICE

Presentata da:
TOMMASO PECORELLA

Sessione III
Anno Accademico 2013-2014

I predict the Internet will soon
go spectacularly supernova and
in 1996 catastrophically collapse.

Robert Metcalfe, 1995

Introduzione

Il presente elaborato è il risultato del lavoro di ricerca, sviluppo e analisi di un dataset campione, composto da circa 3 mln di entry ed estratto da un data warehouse di informazioni riguardanti il consumo energetico di diverse smart home.
L'attività svolta ha portato, in modo forse provocatorio, ma non fuori luogo, a scegliere come citazione iniziale di questo scritto un'affermazione di Robert Metcalfe, inventore di *Ethernet*, fondatore di 3Com (oggi un'azienda del gruppo Hewlett-Packard) e formulatore della *legge di Metcalfe*.
E' proprio grazie alle tecnologie sviluppate per le connessioni di rete che oggi possiamo trasmettere informazioni da un nodo all'altro, ed è proprio grazie al fallimento dell'affermazione di Metcalfe che oggi viviamo seguendo un trend di interconnessione continuamente in via di sviluppo, nel bene e nel male.
E' così che oggi la totalità degli utenti connessi trasmette e riceve informazioni in qualunque momento e tramite una varietà incredibile di mezzi a disposizione: notebook, convertibili, smartphone, ma anche sistemi di navigazione, smartwatch e non solo; un elenco molto generoso che possiamo terminare con il concetto di smart home.
Il presente lavoro tratta il tema energetico e le tecniche di *data cleaning* e *prediction* dei consumi futuri, che portano inevitabilmente ad esaminare il *Big Data*, il dato e la sua tipologia.
Grazie alle tecniche di Data Mining, infine, si è cercato di estrarre valore da un significativo volume di dati. Un percorso, il suddetto, volto a contestualizzare l'analisi e lo sviluppo software affrontato, al fine di disporre di

uno strumento che permetta una prima lettura e catalogazione del dataset trattato.

La risultante delle considerazioni derivate dall'utilizzo di software per l'apprendimento automatico (*Weka*), per Business Intelligence (*Microsoft SSRS*) e il software sviluppato (*C#.NET* + DBMS: *MSSQL Server*), costituiranno le conclusioni del lavoro svolto.

Indice

Capitolo 1

Introduzione allo stato dell'arte

Il primo capitolo tratta gli argomenti relativi agli strumenti volti all'analisi, alla trasmissione dei dati e alla loro applicazione sul tema del consumo energetico.

1.1 Big Data

Nel 2010, il volume di dati disponibili nel web era di circa 500 miliardi. di gigabyte [1], le stime prevedono un loro aumento pari a 5 volte entro il 2015.

1.1.1 L'origine dei big data

L'elevato livello di connettività attuale è l'origine dell'immensa mole di informazioni che attraversa la *rete*, un insieme di dati caratterizzati però da una significativa presenza di rumore, ridondanza e informazioni pleonastiche che ne compromettono, in parte, il potenziale informativo.
Non solo persone ed eventi, quindi, ma anche tecnologia, grazie alla quale è possibile raccogliere le informazioni e analizzarle al fine di ottenere stime predittive.
Il *McKinsey Global Institute* ha definito i Big Data come *datasets whose size*

[1] Bollier, 2010

is beyond the ability of typical database software tools to capture, store, manage, and analyze [2].

Il termine Biga Data, dunque, trova la sua genesi nei volumi e nella complessità dei dati, in quanto caratterizzato da un immenso numero di entry e da una crescita esponenziale degli stessi.
La rappresenzazione dei Big Data coinvolge le seguenti tipologie di dato [4]:

- **Traditional enterprise data**: include informazioni relative ai clienti (o customer) che provengono da sistemi di tipo *customer relationship management* (CRM), nonchè dati transazionali provenienti da sistemi *ERP* (MS Dynamics AX, AS400, SAP, etc...) e transazioni eseguite su store online.

- **Machine-generated/sensor data**: include *Call Detail Records* (CDR), weblogs, **smart meters**, sensori, logs e dati provenienti da sistemi di trading.

- **Social data**: include dati provenienti da siti di micro-blogging, come *Twitter*, e da piattaforme di social media, come *Facebook*. [5]

Da un dettaglio maggiore è possibile ottenere il seguente elenco:

- Social networks e social media (Twitter, Facebook, blogs,forum, etc...)

- Email

- Transazioni

- Documenti cartacei digitalizzati

- Registrazioni video

- Registrazioni audio

[2] Manyika, et al., 2011

- Immagini
- Dati di geo-posizionamento (GPS)
- Dati generati da trasmettitori e sensori (cellulari, wifi,bluetooth, Rfid, NFC, etc.), o misuratori digitali (digital meters)
- M2M (Machine to Machine) data - Internet of Things
- Automazione processi produttivi
- Digitalizzazione dei processi di Ricerca e Sviluppo (nella bioinformatica e biogenetica, chimica, climatologia, etc...)
- Clickstream - Web Log

Ciò detto, è immediato pensare che i Big Data possano creare nuove opportunità di business per le aziende.

1.1.2 Le 3V: velocity, volume, variety

Le caratteristiche principali dei Big Data, si possono riassumere secondo quello che si definisce lo *schema delle 3v* [2]:

- **V**olume: rappresenta la dimensione effettiva del dataset
- **V**elocità: si riferisce alla velocità di generazione dei dati; tende ad effettuare analisi dei dati in tempo reale, o quasi
- **V**arietà: riferita alle varie tipologie di dati provenienti da fonti diverse (strutturate e non)
- **V**alore: i Big Data nascondono significative informazioni che necessitano di essere identificate, trasformate ed estratte per svolgere operazioni di *analysis* e *decision making*.

1.1.3 Il valore dei big data

I Big Data sono un tesoro per molti settori.
IDC[3] stima che entro il 2020 la mole di dati scambiati che saranno scambiati in un anno, sarà pari a 40 *zettabyte*[4].
Oggi, a credere maggiormente nel valore dei Big Data sono il settore delle Telecomunicazioni e Media (che ha scelto questi sistemi soprattutto per interagire meglio con l'utente finale e capire dove si trovano i clienti), dei servizi finanziari (algoritmi di trading e profilazione dei clienti) e della manifattura
.
Il settore dell'healthcare, invece, potrà analizzare le evoluzioni di malattie, di sindromi e anticipare eventuali azioni correttive.

[3]International Data Corporation
[4]1 000 000 000 000 000 000 000 byte = 10007 = 1021 byte = 1 triliardo di byte.

Concretamente, i Big Data vengono utilizzati per:

- analizzare i mercati
- analizzare dati provenienti da fonti eterogenee (dati non strutturati)
- analizzare le conversazioni legati al mondo Social
- sviluppare business di valore *marketing contestuale*, aumentando la conoscenza dei propri utenti

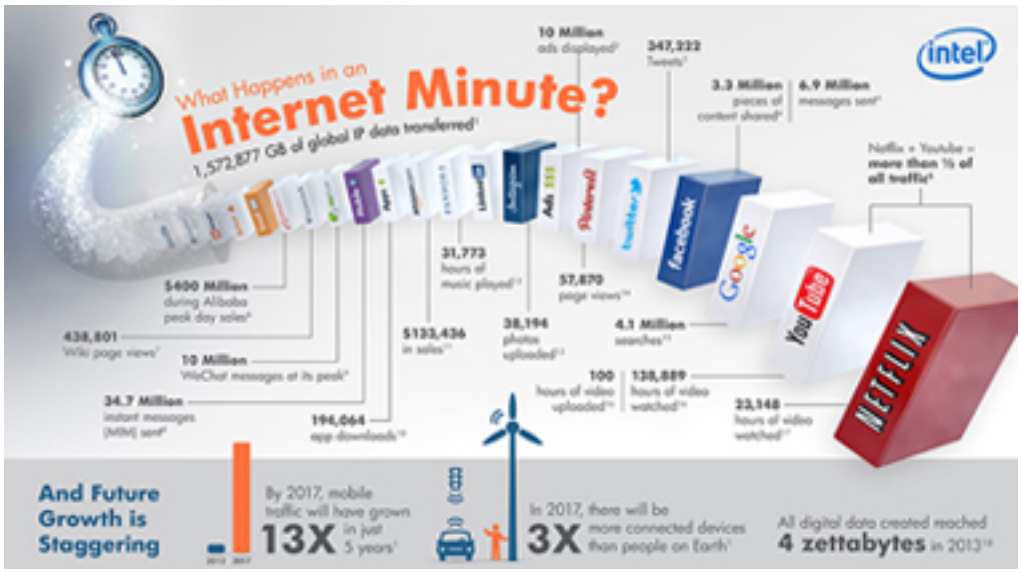


Figura 1.1: What Happens in an Internet Minute? - http://www.intel.com/

E' probabilmente in virtù dell'enorme potenziale analitico dei dati, che il padre del web, Tim Berners-Lee, ha coniato lo slogan del **Raw data now** , al fine di invitare tutti a rendere disponibili online le informazioni per aumentare sempre di più il patrimonio di Big Data.
Almeno in Italia, il percorso inerente i Big Data si sta gradualmente delineando, e non intraprenderlo comprometterebbe il miglioramento dei servizi e della competitività imprenditoriale delle aziende.

1.2 Data Mining

Il Data Mining è un processo di estrazione di conoscenza da banche dati di grandi dimensioni, per mezzo di applicazione di algoritmi che individuano le associazioni implicite (e apparentemente non individuabili tra le informazioni) e le rendono visibili.

Figura 1.2: Data Mining - Volume...to...Value

1.2.1 Applicazione e tecniche

Le attività di Data Mining conducono alla definizione delle strutture all'interno dei dati, dove per struttura si intendono **patterns**, modelli e relazioni. Questo processo, noto anche col nome **KDD** (*Knowledge Discovery in Databases*), consente di estrarre conoscenza in termini di informazioni significative, ed immediatamente utilizzabili, tramite l'applicazione di particolari tecniche ed algoritmi.

Le tecniche più utilizzate nel Data Mining sono:

- Clustering
- Reti neurali
- Alberi di decisione

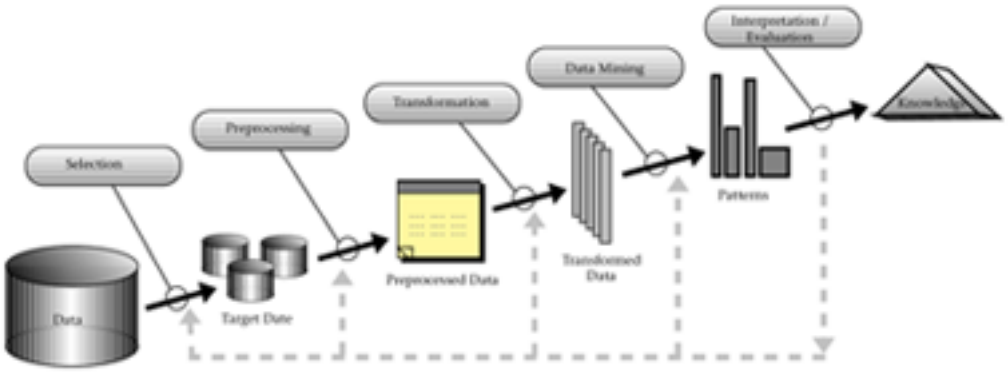


Figura 1.3: KDD - Knowledge Discovery in Databases

- Analisi delle associazioni
- Approccio con supervisione (*supervised*)

Ognuna delle suddette tecniche è caratterizzata da un insieme di metodi e di algoritmi che ha l'obiettivo di fare emergere patterns (sequenze ripetute, omogeneità, regole) dai dati che saranno utilizzati a scopo descrittivo e/o previsivo.

1.2.2 Ambiti applicativi

In campo economico-finanziario, le principali applicazioni sono assimilabili a:

- **segmentazione della clientela** (database marketing) applicazione di tecniche di clustering per individuare i raggruppamenti impliciti nei dati omogenei in termini di comportamento d'acquisto e di caratteristiche socio-demografiche.
- **customer retention** applicazione di tecniche previsive per individuare i clienti a rischio di abbandono.

- **fraud detection** individuazione di comportamenti fraudolenti.
- **analisi delle associazioni** (market basket analysis) individuazione dei prodotti acquistati congiuntamente (sequential patterns) individuazione di comportamenti ricorrenti in sequenze temporali di eventi.
- **competitive intelligence** applicazione di tecniche di clustering a documenti estratti da banche dati internazionali di tipo tecnico-scientifico, volte ad individuare le tecnologie emergenti, le loro relazioni, l'evoluzione temporale e le aziende coinvolte.
- **analisi testuale** (text mining) individuazione degli argomenti trattati da un set di documenti e dalle relazioni tra argomenti.

1.2.3 Text mining

Il text mining è una particolare applicazione di individuazione di sequenze di parole (*pattern*), in comune e con stesse caratteristiche, in un insieme di documenti (*raggruppamento tematico*).
Questo tipo di applicazione è particolarmente utile quando si deve analizzare il contenuto di una collezione di documenti (anche provenienti da fonti eterogenee).
L'individuazione di *gruppi tematici* consente di dare un'organizzazione all'informazione disponibile e di individuare argomenti minori, che anche ad una lettura attenta potrebbero sfuggire.
Le relazioni, inoltre, mettono in evidenza legami tra argomenti, apparentemente separati ma, che hanno una terminologia comune.
Nasukawa e Nagano hanno proposto un sistema noto come *Text Analysis and Knowledge Mining* (TAKMI) da utilizzare su database testuali.
Si tratta di un sistema interattivo che consente agli utenti di confermare facilmente il risultato delle analisi raffrontandolo con un documento originale. Inoltre, un'analisi statistica consente di ignorare i pattern minori così da consentire agli utenti, utilizzando il TAKMI, di trovare solo i maggiori pattern, escludendo i casi di *rumore* nei dati.[6]

1.3 L'energia elettrica

L'energia elettrica rappresenta sulla Terra una fonte di energia secondaria e, pertanto, deve essere prodotta a partire da fonti di energia primaria, attraverso un processo di trasformazione a rendimento (valore compreso tra 0 e 1) sempre inferiore al 100%. La produzione di energia elettrica avviene all'interno di centrali elettriche che dispongono di macchine fondamentali per il processo di trasformazione e produzione (turbina, alternatore, trasformatore).

Nel processo di trasformazione si assiste ad un significativo utilizzo dell'acqua, in forma liquida o aerea, ma comunque ad alta pressione, così da muovere le turbine e garantire un alto numero costante di giri necessari all'alternatore per produrre corrente alternata.

I problemi di reperibilità dell'acqua e di inquinamento termico rappresentano uno svantaggio conseguente al fatto di non essere, l'energia elettrica, una fonte di energia primaria e, pertanto, non è esclusa dalla lista anche la perdita relativa al trasporto lungo le linee elettriche.

Per concludere questo excursus sull'energia elettrica, si riportano le fonti di energia primaria necessarie al processo produttivo di energia elettrica:

- Combustibili fossili (Idrocarburi e carbon fossile)
- Combustibili rinnovabili, come il biogas, la biomassa, gli RSU o gli scarti di legname
- Nucleare
- Solare
- Eolica
- Idrica (idroelettrica, maree, moto ondoso, a osmosi)
- Geotermica

1.3.1 Il consumo energetico

Il consumo energetico nelle abitazioni domestiche è determinato da due principali fattori:

- la tipologia e il numero di dispositivi presenti
- il numero di persone presenti nell'abitazione

Infatti, nelle singole abitazioni riconducibili ad una stessa configurazione, è possibile riscontrare un ampio numero di dispositivi collegati, ed ognuno di questi è caratterizzato da un intervallo di utilizzo e, quindi, di consumo differente. E' necessario, inoltre, considerare l'influenza sull'andamento dei consumi determinata dai singoli individui presenti in un'abitazione, i quali introducono ed utilizzano i dispositivi nella stessa. E' proprio quest'ultimo aspetto che rende difficile fare una predizione dei consumi, soprattutto su un breve arco temporale (entro il range di 60 minuti). Tuttavia, è necessario capire il consumo energetico per poter fornire le indicazioni di efficienza dei consumi e la produzione di energia on-site.[8]

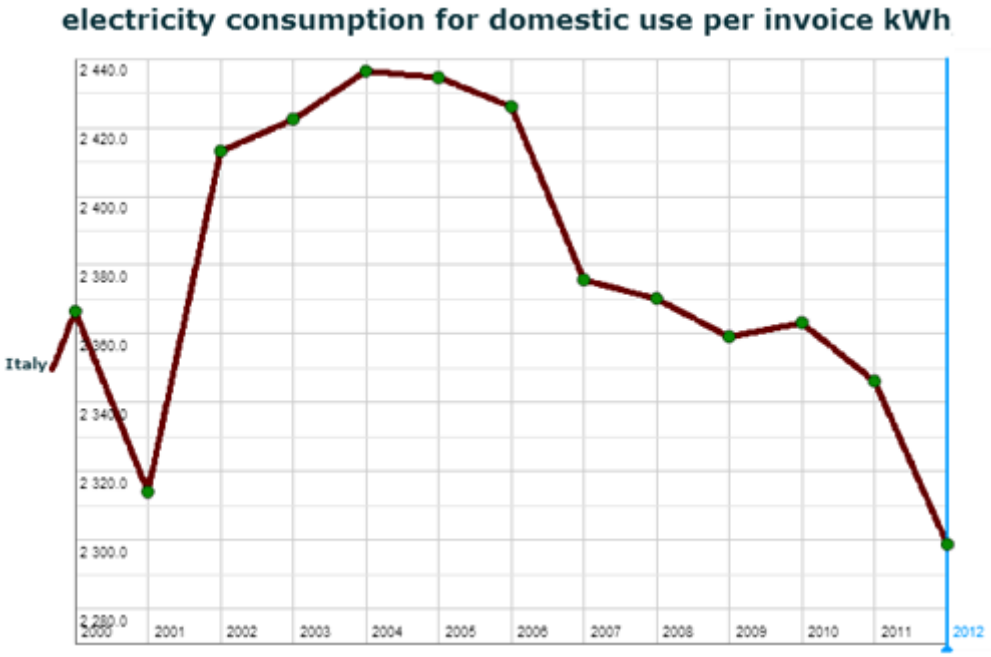


Figura 1.4: Analisi sui consumi domestici 2000-2012 (http://dati.istat.it/)

I dati Istat relativi ai consumi aggregati su base provinciale in Italia ci mostrano l'andamento del consumo energetico su un range di oltre 10 anni (2000-2012), come mostrato in *figura 1.1* [1]

1.3.2 Le politiche energetiche

La disponibilità di fonti di energia primaria limitata, le conseguenze dei processi produttivi e di distribuzione in termini di inquinamento, e il destino delle generazioni future hanno scatenato non pochi dibattiti ed hanno evidenziato la necessità di adottare misure incisive e programmate.
Nascono così le c.d. *politiche energetiche*.
Il problema dell'energia, dunque, è una delle sfide principali per l'Europa.
Mitigare i cambiamenti climatici e ridurre le emissioni (con azioni regolamentate dal **protocollo di Kyoto** fino al 2020) è necessario ed oneroso, in quanto occorre preparare l'infrastruttura energetica di oggi, per le esigenze

di domani. L'UE ha adottato a tal proposito **una strategia 20-20-20** da realizzare entro il 2020, strutturata come segue:

- 20% di riduzione delle emissioni di gas a effetto serra rispetto ai livelli del 1990
- 20% dell'energia consumata proveniente da fonti rinnovabili
- 20% di miglioramento delle prestazioni energetiche

L'obiettivo a lungo termine è ridurre le emissioni di gas serra dell'80-95%, rispetto ai livelli del 1990 ed entro il 2050, assicurando al tempo stesso l'approvvigionamento e la competitività.
Per migliorare l'efficienza, l'UE si concentra sui settori che offrono le maggiori possibilità di risparmio, ossia i trasporti pubblici e l'edilizia.
Inoltre, i contatori intelligenti e le etichette energetiche dell'UE per gli elettrodomestici aiutano a consumare meno energia.
Come si legge dal sito stesso dell'UE, una riconversione tecnologica e una forte collaborazione internazionale sono indispensabili per poter gestire un bacino di consumatori (europei) pari ad oltre 500 milioni. Gli obiettivi di politica energetica dell'UE sono quelli della maggior parte dei paesi:

- cambiamenti climatici
- accesso al petrolio e al gas
- sviluppo tecnologico
- efficienza energetica.

1.4 Tecnologia e raccolta dati

In questo capitolo saranno trattate le tecnologie e i dispositivi idonei alla raccolta dei dati di consumo energetico.

1.4.1 I dispositivi (Appliance)

Il consumo elettrico di ogni abitazione è determinato dalla sommatoria dei consumi dei singoli dispositivi (d'ora in poi definiti *appliance*) presenti ed operativi.
Gli appliance rilevati in questo progetto (in modo univoco con l'utilizzo del software WEKA) nell'analisi dei dati a disposizione sono riconducibili ad una classificazione in relazione alla loro modalità di utilizzo.
Seguendo uno schema presente in letteratura [8], è possibile aggregare gli appliance come da tabella.

Classificazione degli appliance		
Categoria	**Descrizione**	**Esempio**
Continua	In attivita continua e con un consumo costante.	Orologio Modem Router Antifurto
Standby	Accensione volontaria da parte dell'uomo. Quando non è utilizzato, il consumo energetico potrebbe non essere uguale a zero.	Tv-CRT Tv-Plasma Tv-Led Decoder HiFi Caricabatterie PC
Fredda	In attività continua e consumo ciclico tra zero e il livello massimo.	Frigorifero Freezer Congelatore
Attiva	Accensione volontaria da parte dell'uomo. Quando non è utilizzato, il consumo energetico è pari a zero.	Bollitore Fornelli Lavatrice Illuminazione

1.4.2 Il dispositivo *smart meter*

Per disporre di misurazioni puntuali e continue è necessario consentire agli appliance di *comunicare* con un sistema centrale di raccolta dati.
Questa necessità viene soddisfatta dagli smart meter.
Si tratta di una tecnologia di grande utilità nel campo dell'efficienza energetica che consente di misurare il consumo domestico identificandolo con un set di informazioni ben precise.
E' un sistema di controllo basato su reti di sensori *wireless* per il monitoraggio in tempo reale di luce, acqua e gas. I componenti tecnologici sono già esistenti da tempo e diffusi sul mercato a prezzi accessibili; queste caratteristiche permettono l'utilizzo di tali tecnologie a tutti i livelli di consumo, aspetto necessario per la realizzazione dei cosiddetti *smart grid*.
In Italia, tra il 2000 e 2005, Enel Spa ha dotato l'intera utenza (circa 30.000.000 di clienti) di smart meter, raggiungendo il più alto numero di smart meter nel mondo per singola compagnia di distribuzione energetica.
Un esempio di componenti caratterizzante i smart network include[7]:

- Un'infrastruttura di comunicazioni integrate che abilita uno scambio bidirezionale di informazioni e potenza energetica.
- dispositivi *smarter* per le misurazioni (inclusi *advanced metering infrastructure*) che registrano e comunicano informazioni con un dettaglio maggiore sul consumo energetico
- sensori e sistemi di monitoraggio in rete che tengano sotto controllo il flusso di energia nel sistema e le prestazioni della rete stessa
- controlli automatici che intercettano e riparano i problemi afferenti il network e lo riparano in autonomia (*self-healing solutions*)
- sistemi IT provvisti di applicazioni integrate e di analisi dei dati

1.4.3 Una rete tecnologica: *smart grid*

Il concetto di smart grid rappresenta un'evoluzione della rete elettrica così come la si conosce oggi.
In una rete di questo tipo, utenti e produttori vedono integrate intelligentemente le proprie *azioni* al fine di favorire la distribuzione di energia in modo efficiente e sostenibile.
Non si tratta più solo di un controllo centralizzato con linee, interruttori e trasformatori, ma anche di flussi bidirezionali, tecnologia elettronica, informatica e comunicazione.
In conclusione, si può realizzare un sistema bilanciato della domanda e dell'offerta di energia attraverso i diversi modelli software disponibili, tra cui[3]:

- **Smart Grid Interoperability Maturity Model** (**SGIMM**)
  SGIMM è stato sviluppato dal GridWise® Architecture Council (GWAC) per monitorare e misurare il livello di automazione in aree come quella di tramissione, distribuzione e richiesta di risorse.
  Le funzionalità disponibili sono:

  – Status/progress measuring statistics
  – Gap analysis
  – Prioritization of efforts

  Il principale obiettivo di questo modello è quello di crearne un altro per la misurazione che possa promuovere l'interoperabilità nelle aree chiave, come:

  – Configuration & evolution
  – Operation & performance
  – Security & safety nei sistemi elettrici

  SGIMM favorisce la qualità dei modelli SGIM.

- **Smart Grid Investment Model** (**SGIM**)
  SGIM è un modello di tipo finanziario che si occupa di calcolare l'impatto negli investimenti di diversi smart grid.
  Questo permette di valutare i costi e la tracciabilità dei benefici provenienti da investimenti fatti su smart grid in relazione allo spettro di distribuzione.
  SGIM identifica le tecnologie presenti in un smart grid e ne analizza i costi/benefici per l'utente per i prossimi 20 anni.

- **Smart Grid Maturity Model** (**SGMM**)
  SGMM è uno strumento di gestione atto alla valutazione del progresso del processo di implementazione di un smart grid nelle sue 8 fasi:

  - Strategy, Management, and Regulatory (SMR)
  - Organization and Structure (OS)
  - Grid Operations (GO)
  - Work and Asset Management (WAM)
  - Technology (TECH)
  - Customer (CUST)
  - Value Chain Integration (VCI)
  - Societal and Environmental (SE)

  SGMM si basa su questionari per ricavare una stima utile alla valutazione del livello di *maturità*:

  - Initiating
  - Enabling
  - Integrating
  - Optimizing
  - Pioneering

- **Smart Grid Conceptual Model** (**SGCM**)
  Il modello Smart Grid Conceptual fornisce la visualizzazione di un diagramma che mostra come i diversi componenti di un Smart Grid possono essere integrati. SGCM si compone di sei livelli:

  – Customers
  – Markets
  – Service Providers
  – Operations
  – Bulk Generation
  – Transmission
  – Distribution

  Le *feature* disponibili con questo modello sono:

  – Overview dello sviluppo di un smart grid
  – Un contesto di analisi dei diversi standard e interoperabilità.
  – Analisi dell'interazione tra i diversi livelli sopra elencati, utile alle attività di *Distribution Management System*(DMS)

  Questo modello, inoltre, si focalizza sui temi riguardanti *cyber security*, *network management*, *data management* e *application integration*.

1.4.4 L'architettura di rilevazione

La maggior parte delle tecniche di monitoring presenti nelle abitazioni (in questo progetto definite anche *HAG*) tracciano solo il profilo sotto l'aspetto totale del consumo energetico, quindi un dato poco adatto alla predizione dei consumi futuri. Diversi algoritmi di classificazione e disaggregazione sono

stati proposti in letteratura[5], ma questi possono contare su dati monitorabili su media o alta frequenza (almeno 1 Hz).

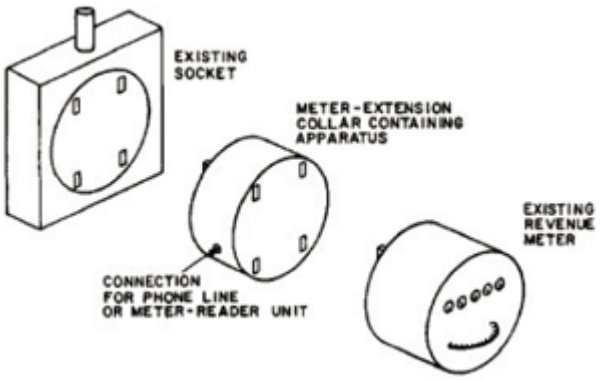


Figura 1.5: Schema di applicazione del sistema di misura NALM

Poichè il consumo energetico è il risultato dei consumi dei singoli appliance, sono stati identificati diversi metodi da adottare nelle abitazioni private. Berger distingue i tre principali approcci[6]:

- *Whole house continuous monitoring.* Questa tecnica comporta l'utilizzo di smart meter collegati al quadro generale dello stabile, al fine di poter misurare il consumo istantaneo dell'intera abitazione.

- *Non-Intrusive Load Monitoring (NILM).* Si tratta di un'insieme di tecniche che mira a riconoscere il consumo energetico di uno specifico appliance estratto dal consumo totale dell'abitazione.

- *Hardware-based sub-metering.* Questo metodo consiste nel collegare un modulo hardware per la misurazione, su ogni appliance dell'abitazione. In questo modo il consumo energetico del singolo appliance può facilmente essere memorizzato ed identificato successivamente.

[5] A.Zoha,A.Gluhak,M.A. Imran,S.Rajasegarar 'Non-Intrusive Load Monitoring Approaches for Disaggregated Energy Sensing: A Survey', Sensors vol 2012, 12 pp. 16838-16866

[6] non solo M.Berges, ma anche E.Goldman,H.Scott Matthews, L.Soibelman 'Training Load Monitoring Algorithms on Highly Sub-Meter Home Electricity Consumption Data'

Capitolo 2

Progettazione

La necessità di memorizzare le informazioni[1] in modo maggiormente strutturato ha portato allo sviluppo di un software *Raw Data Synchronizer* in grado di prendere in input file di tipo Excel (xsl o xslx) e memorizzarne il contenuto in una base dati dalla quale estrarre le prime informazioni da rendere consultabili e utilizzabili per eseguire una prima statistica descrittiva. Per poter sfruttare al meglio la scelta della tecnologia Microsoft è stato utilizzato, come OS, per ospitare l'intero ambiente di sviluppo **Windows Server 2012 R2 St. Ed.**

2.1 L'architettura del progetto

Il progetto oggetto di analisi in questo elaborato ha mostrato chiaramente e da subito le difficoltà insite nella organizzazione di una grande mole di dati semi-strutturata memorizzata in formato testuale, anche se organizzata in righe e colonne; è così che sono strutturati i dati provenienti dalla piattaforma Telecom Energy@home.

[1] I dati trattati in questo elaborato sono stati collezionati adottando un sistema distribuito di smart plug, dispositivi con connessione WiFi ed azionabili anche da remoto con propri magnetotermici in grado di spegnere l'appliance nel caso si verifichino particolari condizioni di surriscaldamento.

Si possono segmentare le necessità incontrate come segue:

- Gestione dei file excel memorizzati nel file system in strutture annidate di cartelle e sottocartelle
- Conversione dei file excel in file csv (Comma Separated Values)
- Lettura automatizzata dei file (Software)
- Implementazione di una base dati per lo *storaging* dei dati estratti (Database)
- Prima analisi dei dati (SSRS e Business Intelligence Tool)
- Analisi dati con il software *machine learning* Weka preprocessando i dati e classificandoli utilizzando i pattern forniti dal software stesso

Figura 2.1: Architettura del progetto

2.2 La base dati

La struttura della base dati è stata progettata per ospitare i dati presenti nei file CVS, generati da una conversione in *run-time* per mezzo del software *Raw Data Synchronizer*.

Ulteriori informazioni sono state ricavate anche dal nome stesso del file CSV, caratterizzato sempre dal medesimo paradigma. Per esempio, applicando il paradigma precedente ad un file, un esempio di n-pla di informazioni possibile è:

- **nome hag**: hag-0003
- **id hag**: 933
- **nome IT appliance**: Lavastoviglie
- **nome EN appliance**: DISH WASHER
- **data file**: 2012-01-02

I file CSV, invece, sono caratterizzati dal seguente tracciato:

- *APPL ID*
- *START TIME*
- *MIN PWR*
- *MAX PWR*
- *DELTA ENERGY*
- *DURATION*
- *MIN PWR TIME*
- *MAX PWR TIME*

Identificati, dunque, struttura e tracciato dei file, è possibile progettare la base dati che dovrà gestire l'univocità delle informazioni (le righe dei file CSV) e unire quelle presenti nel nome con quelle presenti nei singoli record, nonchè ospitarle in campi progettati, al fine di avere il *TYPE* corretto.

Il DMBS utilizzato è *Microsoft SQL Server 2012 Std. Ed.*[2]
La tabella *FilesDataRepository* presenta quindi questa struttura:

- `[hag]` [nchar](10) NOT NULL
- `[applianceid]` [bigint] NOT NULL
- `[start_time]` [datetime] NOT NULL
- `[min_pwr]` [bigint] NULL
- `[max_pwr]` [bigint] NULL
- `[delta_energy]` [bigint] NULL
- `[duration]` [bigint] NULL
- `[min_pwr_time]` [bigint] NULL
- `[max_pwr_time]` [bigint] NULL
- `[appliance_name_en]` [nvarchar](max) NULL
- `[appliance_name_it]` [nvarchar](max) NULL
- `[file_data]` [date] NULL

ed ha come **chiave primaria** (primary key, PK)

CONSTRAINT `[PK_FilesDataRepository]`
PRIMARY KEY CLUSTERED
(
`[hag]` *ASC*,
`[applianceid]` *ASC*,
`[start_time]` *ASC*)

[2]Microsoft SQL Server 2012 si è attestato quale database più sicuro per cinque anni consecutivi (*National Institute of Standards and Technology Comprehensive Vulnerability Database 4/17/2013, quota di mercato da IDC 2013*).

Dalla base dati così definita sono state realizzate 4 viste (*view*) per avere una prima segmentazione dei dati.

- **Delta Energy per Year and Month** (aggregazione del Delta Energy per Anno e Mese)
- **Entries per Year and Month and ApplianceName** (numero di record afferenti un appliance per Anno e Mese)
- **HagAppliance** (identificazione univoca degli HAG e APPLIANCE)
- **NewMeasurements** (nuove misurazioni per una gestione differenziale futura dei dati)

Figura 2.2: Il Software: Raw Data Synchronizer

2.3 Il software

La gestione automatizzata per la navigazione dell'intero *folder tree* finalizzata all'identificazione dei file utili (paradigma per identifcare il nome e 'xls' per identificare il tipo di file), loro lettura e controllo, memorizzazione attraverso query *Insert* nella base dati e prima analisi descrittiva è stata

affidata ad un software denominato *Raw Data Synchronizer* sviluppato con tecnologia **Microsoft .NET** e IDE proprietario *Visual Studio 2013 Ultimate Ed.* Il software è caratterizzato da uno *Splash Screen* (fig.2.2) che comunica la fase di *loading* dell'applicazione.
Dopo il caricamento viene eseguito un *Windows Form* stilizzato secondo il trend corrente in termini di design e UI applicativo. L'applicazione possiede anche alcune funzionalità esplorabili dall'icona presente nel *system tray* del sistema operativo.
Le funzionalità di *Raw Data Synchronizer* sono:

- **Loading** (fig. 2.3): strumento di caricamento dei dati. L'utente, specificando il folder *root* di partenza, può avviare la lettura e caricamento nella base dati dei file excel utili. Durante l'elaborazione è possibile seguire le notifiche in *real time* del software relative al file in corso di lettura e sue informazioni.

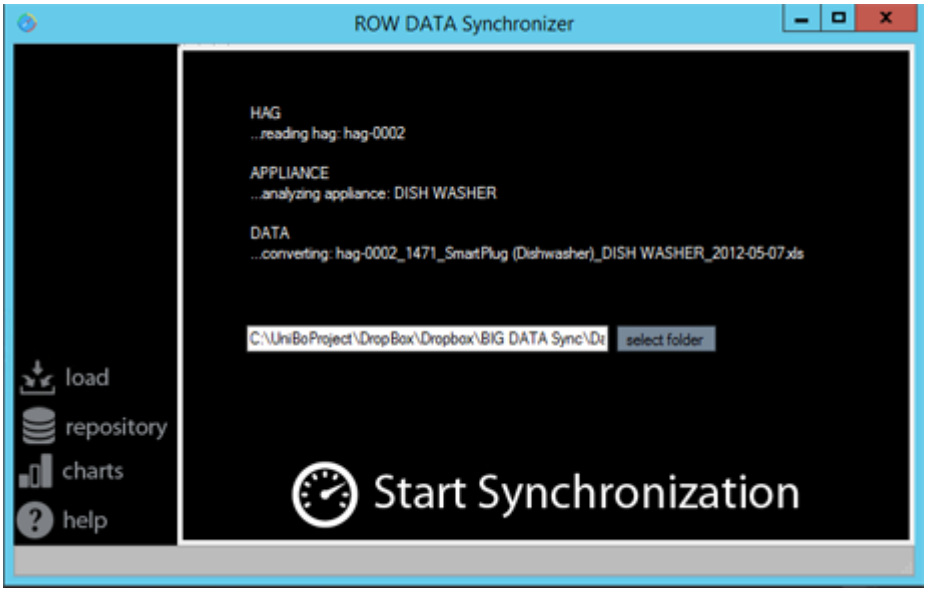


Figura 2.3: Il Software: *Loading*

- **Grid View** (fig. 2.4): utilizzando un componente .NET è possibile per l'utente esplorare l'intera tabella del database in tempo reale ed

eventualmente eseguire un *export* parziale dei dati.

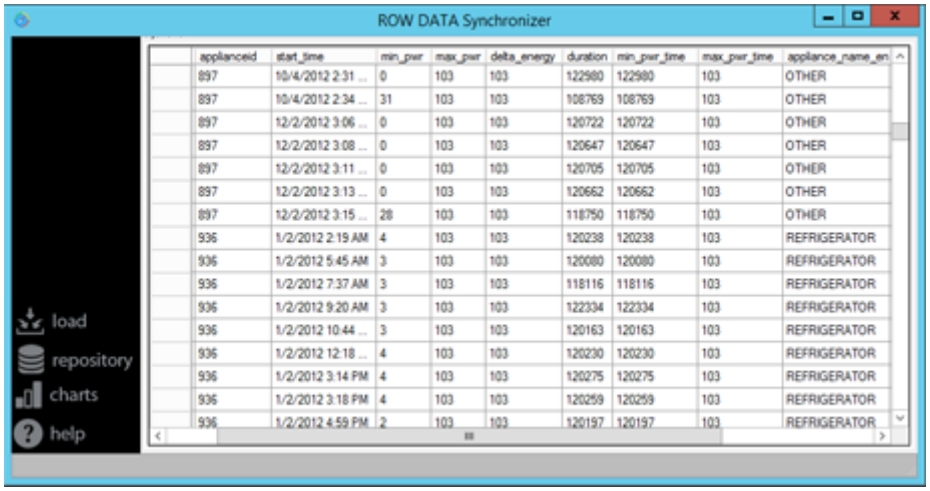


Figura 2.4: Il Software: *Grid View*

- **Business Intelligence** (fig. 2.5): l'utilizzo di *SQL SERVER Reporting Services* permette all'utente di eseguire *report* sviluppati con l'ausilio dello strumento *Business Intelligence Tool*.

2.4 Business Intelligence Tool

Utilizzando il software *Raw Data Synchronizer* sul totale dei file excel a disposizione si è scelto di eseguire un primo lavoro di BI a fronte di 2.500.000 entry[3].
La scelta del prodotto da utilizzare è ricaduta su SQL Server Reporting Services.
Reporting Services è una piattaforma di report basata su server che fornisce funzionalità di report complete per numerose origini dati. Include un set completo di strumenti per creare, gestire e recapitare report, nonché API che consentono di integrare o estendere l'elaborazione di dati e report in applicazioni personalizzate.

[3]dati provenienti dalla piattaforma Telecom Energy@home

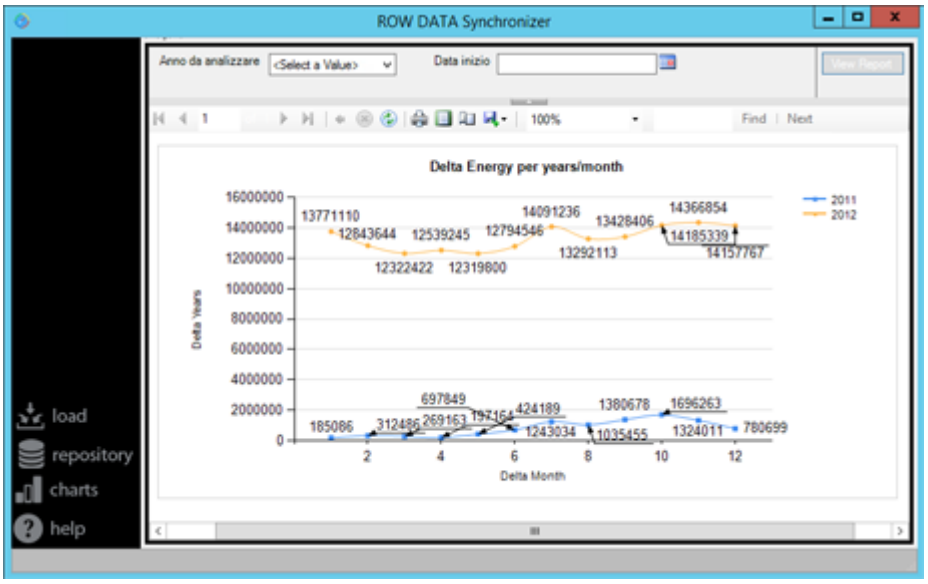


Figura 2.5: Il Software: *Business Intelligence*

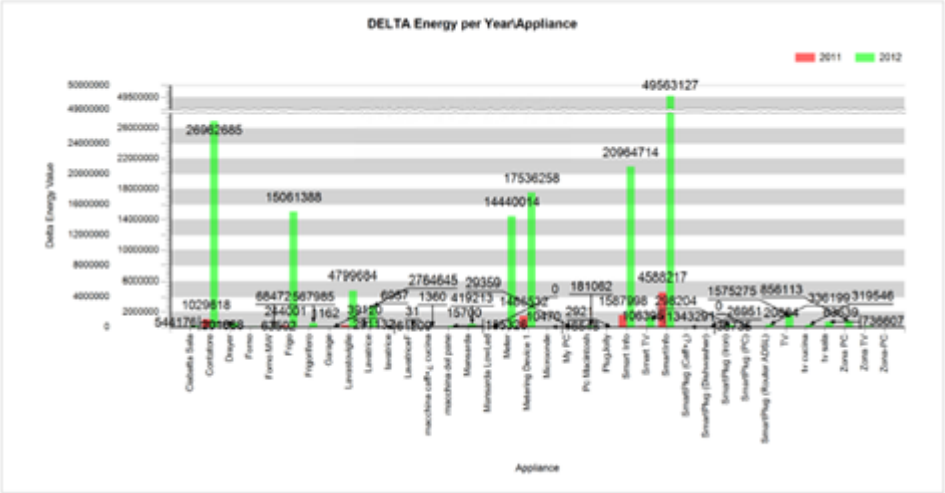


Figura 2.6: Il Software: *Un grafico sviluppato per SSRS*

Il risultato finale del lavoro di sviluppo del report è memorizzato in un file con estensione *rdl* che contiene il codice, le query e tutte le informazioni relative agli altri aspetti caratteristici del documento di reportistica.

Capitolo 3

Implementazione

L'implementazione dell'intero progetto software e i sistemi operativi utilizzati si è basata quasi completamente[1] su soluzioni Microsoft. Per lo sviluppo del software e la progettazione della base dati è stata utilizzata una macchina server[2] con sistema operativo *Microsoft Windows Server 2008 R2 St. Ed.*, 3Gb di Ram e un processore dual-core 1.5 GHz su architettura x64. L'elaborazione dei dati con l'utilizzo del software Weka, vista la mole di dati, ha richiesto l'utilizzo di un secondo server con sistema operativo Windows Server 2012 St. Ed, 12Gb di Ram e 2 processori quad-core su architettura x64.

3.1 Il codice

Sono stati selezionati alcuni snippet di codice significativi per trattare l'esigenza gestita e la soluzione ideata a tal fine.

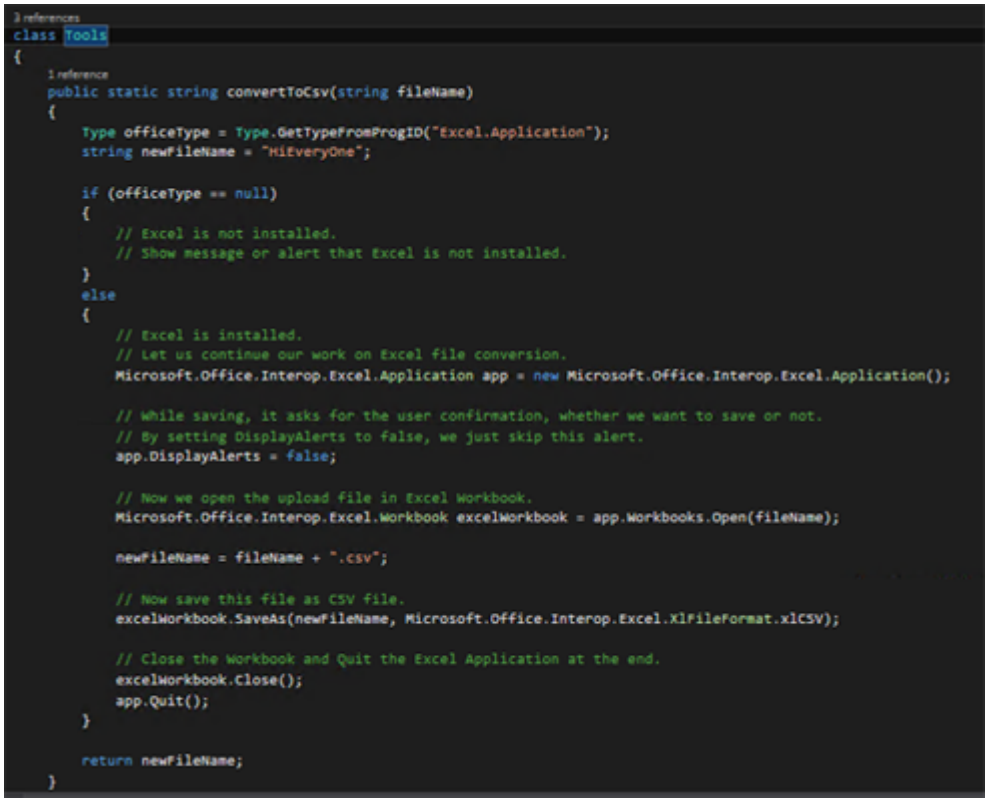
Software

L'applicazione sviluppata per il progetto oggetto di questo elaborato è costituita da uno *splash screen* per il caricamento dei componenti necessari

[1] Ad esclusione di Weka, software scritto in Java.

[2] Si tratta di un server virtualizzato con VMWare ESXi

tra cui il *Windows Form* principale che fornisce le funzionalità sviluppate.
Relativamente alla funzionalità di lettura dei file in formato *xlsx*, si è reso necessario distaccarsi il più possibile dal formato proprietario di Microsoft Excel scegliendo come obiettivo il formato testuale CSV.
La soluzione individuata è stata quella di eseguire una conversione *run-time*[3] dei file Excel in file CSV, mantenendo inalterata la struttura del nome del file convertito[4]



```
3 references
class Tools
{
    1 reference
    public static string convertToCsv(string fileName)
    {
        Type officeType = Type.GetTypeFromProgID("Excel.Application");
        string newFileName = "HiEveryOne";

        if (officeType == null)
        {
            // Excel is not installed.
            // Show message or alert that Excel is not installed.
        }
        else
        {
            // Excel is installed.
            // Let us continue our work on Excel file conversion.
            Microsoft.Office.Interop.Excel.Application app = new Microsoft.Office.Interop.Excel.Application();

            // While saving, it asks for the user confirmation, whether we want to save or not.
            // By setting DisplayAlerts to false, we just skip this alert.
            app.DisplayAlerts = false;

            // Now we open the upload file in Excel Workbook.
            Microsoft.Office.Interop.Excel.Workbook excelWorkbook = app.Workbooks.Open(fileName);

            newFileName = fileName + ".csv";

            // Now save this file as CSV file.
            excelWorkbook.SaveAs(newFileName, Microsoft.Office.Interop.Excel.XlFileFormat.xlCSV);

            // Close the Workbook and Quit the Excel Application at the end.
            excelWorkbook.Close();
            app.Quit();
        }

        return newFileName;
    }
```

Figura 3.1: Il Software: *Utilizzo di Office Primary Interop Assemblies*

La figura 3.1 mostra l'utilizzo dei componenti **Office Primary Interop Assemblies** di Microsoft per utilizzare le API (*Application programming*

[3]L'origine di questa necessità va ricercata nel garantire il livello massimo di automazione possibile da parte del software *Raw Data Synchronizer.*

[4]Nel II capitolo, sez. *Il Software* si specificano le informazioni reperibili nel nome del file.

interface), in questo caso, di MS Excel.
L'intero processo di elaborazione dei file è schematizzato con un diagramma di attività in figura 3.2

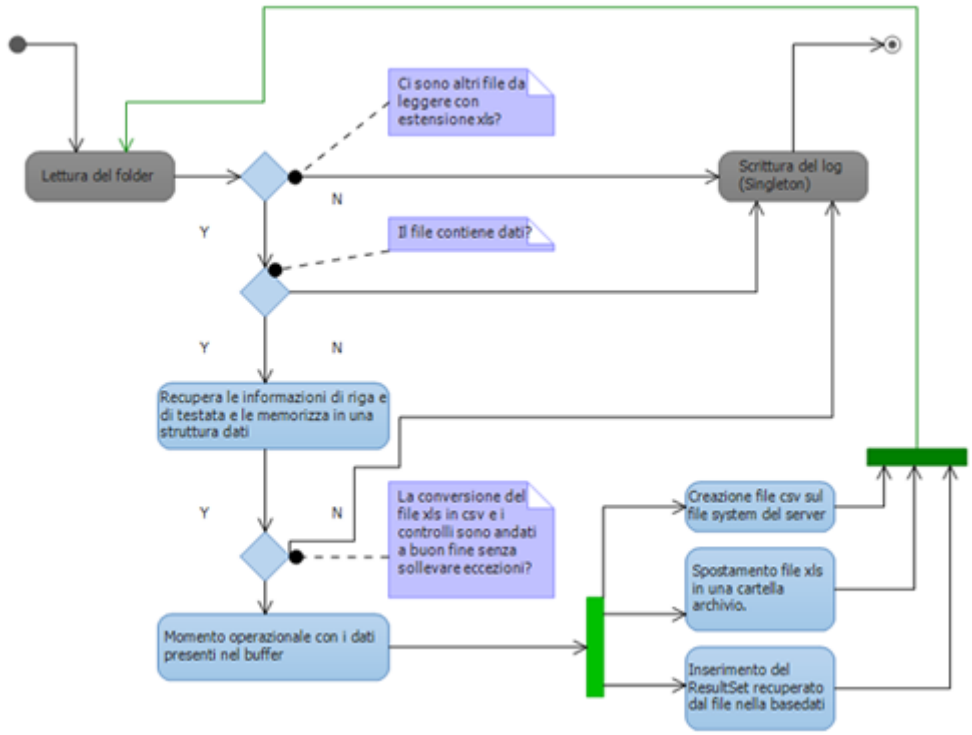


Figura 3.2: *Activity Diagramm: Elaborazione dei file.*

Base Dati

La base dati è stata progettata su DBMS Microsoft SQL Server 2012.
In figura 3.3 è visibile uno snippet di codice sql relativo alla realizzazione di una vista (*View*) utilizzata dal software Weka.

Report

Una prima analisi dei dati è stata eseguita con l'aiuto di SSRS.
Lo sviluppo di un report produce un file con estensione rdl (*Report Definition*

```
USE [UniBo]
GO

SET ANSI_NULLS ON
GO

SET QUOTED_IDENTIFIER ON
GO

CREATE VIEW [dbo].[FilesDataRepository4WEKA]
AS
    SELECT
            ROW_NUMBER() OVER (ORDER BY [applianceid]) AS rowid,
            hag,
            applianceid,
            start_time,
            min_pwr,
            max_pwr, delta_energy, duration, min_pwr_time, max_pwr_time,
            appliance_name_en, appliance_name_it, file_data,
            YEAR(start_time) AS Year,
            MONTH(start_time) AS Month
    FROM    dbo.FilesDataRepository
    WHERE   (hag IS NOT NULL) AND (applianceid IS NOT NULL)
GO
```

Figura 3.3: La base dati: *Create View*

Language),una rappresentazione XML di una definizione di un report di SQL Server Reporting Services.
Una definizione del report contiene informazioni sul layout e sul recupero dei dati per un report.
RDL è costituito da elementi XML che corrispondono a una grammatica XML creata per Reporting Services. La struttura RDL permette l'interoperabilità ed è quindi stato integrato nel software *Raw Data Synchronizer*.

Capitolo 4

Validazione

La predizione è il concetto generale che si divide in **Classificazione** quando la classe è un valore nominale e in **Regressione** quando la classe è un valore numerico.
Il software selezionato per la validazione dei dati è Weka[1], acronimo di **Waikato Environment for Knowledge Analysis**, un software per l'*apprendimento automatico* sviluppato nell'università di Waikato in Nuova Zelanda.
Si tratta di un SW open source, rilasciato con licenza GNU General Public License.

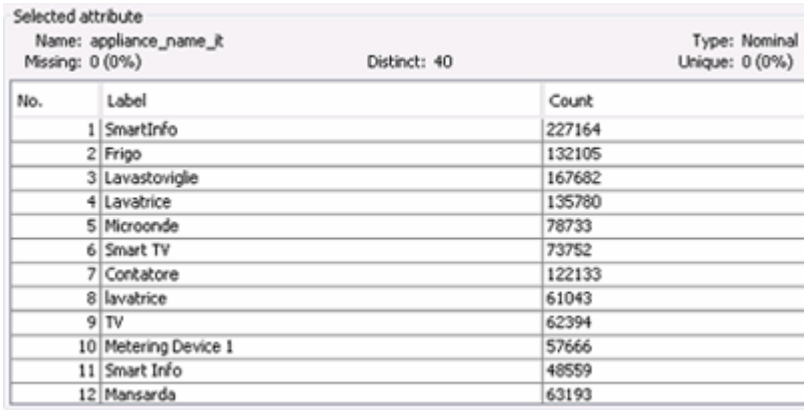
4.1 Weka: Experimenter

Experimenter è l'ambiente che consente di esplorare i dati attraverso i comandi Weka.
Le attività di validazione sono state svolte utilizzando **Preprocess**, **Classify** e **Visualize**.

[1]Curiosamente la sigla corrisponde al nome di un simpatico animale simile al Kiwi, presente solo nelle isole della Nuova Zelanda. (Wikipedia)

4.2 Preprocess

I dati memorizzati da *Raw Data Synchronizer* nella base dati sono stati preprocessati (*Data Cleaning*) e salvati in formato **ARFF**[2], formato utilizzato in Weka per la lettura dei dataset, simile al CSV[3] ed è equivalente alla tabella di un database relazionale.


Selected attribute
Name: appliance_name_it Type: Nominal
Missing: 0 (0%) Distinct: 40 Unique: 0 (0%)

No.	Label	Count
1	SmartInfo	227164
2	Frigo	132105
3	Lavastoviglie	167682
4	Lavatrice	135780
5	Microonde	78733
6	Smart TV	73752
7	Contatore	122133
8	lavatrice	61043
9	TV	62394
10	Metering Device 1	57666
11	Smart Info	48559
12	Mansarda	63193

Figura 4.1: Weka: *Appliance riconosciuti*

Dopo il caricamento dei dati, si nota subito che Weka ha identificato gli HAG e le APPLIANCE nel dataset costituito da quasi 2mln e mezzo di entry (fig 4.1).
Dopo aver rimosso gli attributi non necessari all'analisi si è applicato il filtro ***InterquartileRange*** *-R first-last -O 3.0 -E 6.0*

In seguito all'elaborazione dei dati, Weka definisce gli **Outlier** (date n istanze e il numero k, si determinano le k istanze più dissimili dalle altre) e **Extreme value**, come mostrato in figura 4.3

[2] **A**ttribute **R**elationship **F**ile **F**ormat
[3] **C**omma-**s**eparated **v**alues

Selected attribute

Name: hag Type: Nominal
Missing: 0 (0%) Distinct: 8 Unique: 0 (0%)

No.	Label	Count
1	hag-0007	483702
2	hag-0006	376722
3	hag-0003	325183
4	hag-0002	205771
5	hag-0010	94140
6	hag-0112	382624
7	hag-0011	161027
8	hag-0111	424124

Figura 4.2: Weka: *Hag riconosciuti*

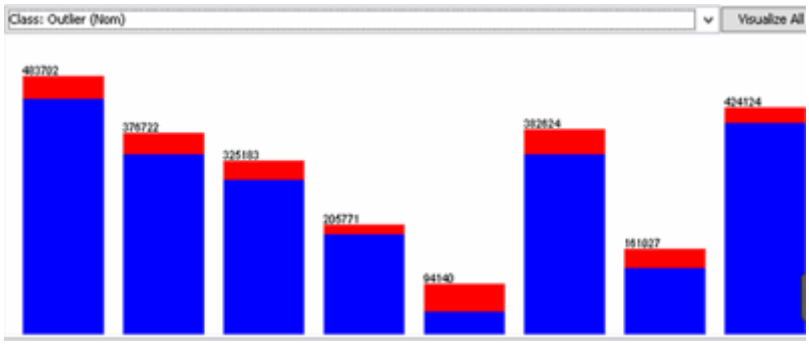


Figura 4.3: Weka: *Outlier*

4.3 Classify

Dopo aver preparato i dati sono stati applicati i seguenti algoritmi di classificazione e regressione:

- weka.classifiers.trees.**RandomTree**
  Un albero di classificazione si può trasformare in un insieme di regole di classificazione.
  Una regola è creata per ogni percorso dalla radice alle foglie .
  Ogni nodo interno del percorso è un test dell'antecedente.

- weka.classifiers.bayes.**NaiveBayes**
  Il classificatore Naive è un classificatore bayesiano semplificato con un modello di probabilità sottostante che fa l'ipotesi di indipendenza delle feature, ovvero assume che la presenza o l'assenza di una particolare feature non sia correlata alla presenza o assenza di altre feature.
  In formule, sia X una istanza da classificare, e C1,...,Cn le possibili classi. I *classificatori Bayesiani* calcolano la probabilità di Ci subordinata alla conoscenza di X come:

$$P(Ci|X) = \frac{P(X|Ci)P(Ci)}{P(X)} \tag{4.1}$$

  Si sceglie la classe Ci che massimizzi il risultato dell'equazione 4.1
  P(X) è uguale per tutte le classi per cui non occorre calcolarla.
  P(Ci) si può calcolare sull'insieme dei dati di addestramento e si conta la percentuale di istanze di classe Ci sul totale.

- weka.classifiers.bayes.**BayesNet**
  Matematicamente, una rete bayesiana è un grafo aciclico orientato[4] in cui:

  – I nodi rappresentano le variabili

[4]Un grafo orientato aciclico è spesso chiamato **DAG** (*Directed Acyclic Graph*)

– Gli archi rappresentano le relazioni di dipendenza statistica tra le variabili e le distribuzioni locali di probabilità dei nodi foglia rispetto ai valori dei nodi padre

Una rete bayesiana rappresenta la distribuzione della probabilità congiunta di un insieme di variabili.

classi vere\predette	*Classe 1*	*Classe 2*	*Classe 3*	*...*	*Classe k*
Classe 1	23	3	4		2
Classe 2	..	17	..		..
Classe 3	..	..	15		..
....					
Classe k	..	..	..		93

Figura 4.4: *Matrice di confusione*

Per la valutazione dei classificatori vengono usate le **Matrici di confusione** (fig 4.4), dov'è possibile visualizzare gli errori commessi dal classificatore.
L'*errore campionario* si ottiene dalla somma dei valori di tutte le celle, esclusa la diagonale.

forestTree, SVM (support vector a machine), decision table (regole if-then-alfa)

4.4 I risultati dell'analisi

La prima osservazione evidenzia come il software *Raw Data Synchronizer* e la struttura della base dati abbiano correttamente gestito l'univocità dei dati recuperati dai file excel a disposizione.
Nessuna eccezione è stata infatti sollevata dal software gestendo (*try...catch*) le operazioni da parte della base dati ed inoltre il totale dei record disponibili

è stato riconosciuto come unico da Weka.
Purtroppo, come anticipato dal Report sviluppato, dopo aver preprocessato i dati[5] emerge che solo gli anni 2011 e 2012 siano oggetto di analisi con una notevole disparità di informazioni.
Infatti, sia T il totale delle rilevazioni e r11, r12 le rilevazioni rispettivamente del 2011 e 2012:

$$T = \sum r11 + \sum r12 \qquad con \quad r12 \gg r11 \tag{4.2}$$

Tuttavia, grazie ai dati presenti per l'anno 2012, Weka ha correttamente individuato gli HAG e APPLIANCE con queste caratteristiche sui dati:
Missing: 0% e Unique: 0% .
Weka ha, quindi, individuato 40 tipologie distinte di APPLIANCE distribuiti su un totale di 8 HAG.
Purtroppo non sono disponibili ulteriori informazioni riguardanti le caratteristiche degli HAG (numero di occupanti, dimensione e numero di APPLIANCE formalmente attivi, ecc...)

Le attività svolte nella parte *Classify* di Weka, come descritto nel capitolo 4, sono riconducibili ai tre algoritmi utilizzati.

- **RandomTree**
  La dimensione dell'albero ottenuto è: 197
  L'appliance con il maggiore dettaglio e profondità del nodo generato da questo algoritmo è stato individuato nella lavatrice. Si riporta un frammento del risultato ottenuto.

  appliancename = *Lavatrice*

  — **minpwrtime** < 120092.5

  — — minpwrtime = no

[5] I file excel utilizzati in questo progetto sono solo una parte delle rilevazioni totali messe a disposizione dalla piattaforma Telecom Energy@home.

— — — applianceid < 843 : hag-0007 (6115/0)

— — — applianceid >= 843

— — — — applianceid < 1207 : hag-0003 (6707/0)

— — — — applianceid >= 1207 : hag-0010 (1993/0)

— — minpwrtime = yes

— — — duration < 118955.5

— — — — duration < 117299

— — — — — deltaenergy < 2 : hag-0006 (497/0)

— — — — — deltaenergy >= 2

— — — — — — deltaenergy < 7.5

— — — — — — — applianceid < 1374 : hag-0003 (12/0)

— — — — — — — applianceid >= 1374 : hag-0006 (19/0)

— — — — — — deltaenergy >= 7.5

— — — — — — — applianceid < 1207 : hag-0003 (63/0)

— — — — — — — applianceid >= 1207

— — — — — — — — applianceid < 1643 : hag-0010 (2/0)

— — — — — — — — applianceid >= 1643 : hag-0006 (2/0)

— — — — duration >= 117299

— — — — — year < 2011.5 : hag-0007 (125/0)

— — — — — year >= 2011.5

— — — — — — starttime < 1344420000000

— — — — — — — rowid < 706347 : hag-0007 (667/0)

— — — — — — — rowid >= 706347

— — — — — — — — rowid < 1786919.5 : hag-0003 (486/0)

— — — — — — — — rowid >= 1786919.5

— — — — — — — — — applianceid < 1643 : hag-0010 (2/0)

— — — — — — — — — applianceid >= 1643 : hag-0006 (3/0)

— — — — — — starttime >= 1344420000000

— — — — — — — minpwrtime < 118750.5

— — — — — — — — rowid < 1796752

— — — — — — — — — applianceid < 843 : hag-0007 (76/0)

— — — — — — — — — applianceid >= 843 : hag-0003 (306/0)

— — — — — — — — rowid >= 1796752

— — — — — — — — — starttime < 1345629600000 : hag-0006 (16/0)

— — — — — — — — — starttime >= 1345629600000

— — — — — — — — — — duration < 118407.5

— — — — — — — — — — — minpwrtime < 118015 : hag-0010 (9/0)

— — — — — — — — — — — minpwrtime >= 118015

— — — — — — — — — — — duration < 118376.5

— — — — — — — — — — — — starttime < 1354834800000

— — — — — — — — — — — — — rowid < 2430415 : hag-0010 (17/0)

— — — — — — — — — — — — — rowid >= 2430415 : hag-0006 (4/0)

— — — — — — — — — — — — — starttime >= 1354834800000 : hag-0006 (3/0)

— — — — — — — — — — — — duration >= 118376.5 : hag-0010 (5/0)

— — — — — — — — — — — duration >= 118407.5

— — — — — — — — — — — rowid < 2430497.5 : hag-0010 (27/0)

— — — — — — — — — — — rowid >= 2430497.5 : hag-0006 (44/0)

— — — — — — — minpwrtime >= 118750.5

— — — — — — — — applianceid < 843 : hag-0007 (193/0)

— — — — — — — — applianceid >= 843

— — — — — — — — — applianceid < 1207 : hag-0003 (134/0)

— — — — — — — — — applianceid >= 1207

— — — — — — — — — — rowid < 2430184 : hag-0010 (48/0)

— — — — — — — — — — rowid >= 2430184 : hag-0006 (40/0)

— — — duration >= 118955.5

— — — — applianceid < 1374

— — — — — starttime < 1349431200000

— — — — — — rowid < 711698 : hag-0007 (15/0)

— — — — — — rowid >= 711698 : hag-0003 (12/0)

— — — — — starttime >= 1349431200000 : hag-0003 (8/0)

— — — — applianceid >= 1374 : hag-0006 (2773/0)

— **minpwrtime** >= 120092.5

— — rowid < 717661 : hag-0007 (71377/0)

— — rowid >= 717661

— — — applianceid < 1207 : hag-0003 (43354/0)

— — — applianceid >= 1207

— — — — rowid < 2430047.5 : hag-0010 (257/0)

— — — — rowid >= 2430047.5 : hag-0006 (369/0)

appliancename = *Microonde* : hag-0007 (78733/0)

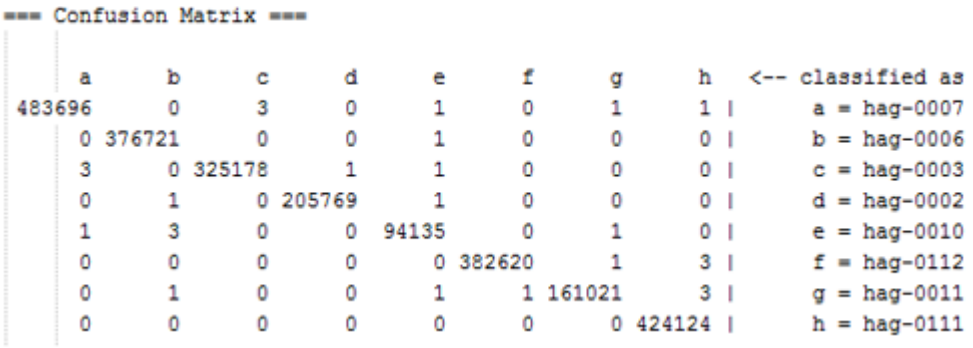
appliancename = *Smart TV* : hag-0007 (73752/0)

Il riepilogo dell'applicazione dell'algoritmo riporta le informazioni seguenti:

Correctly Classified Instances.....2453264 (99.9988 %)
Incorrectly Classified Instances....29 (0.0012 %)
Kappa statistic.....1
Mean absolute error.....0
Root mean squared error.....0.0017
Relative absolute error.....0.0014 %
Root relative squared error.....0.5128 %
Total Number of Instances.....2453293

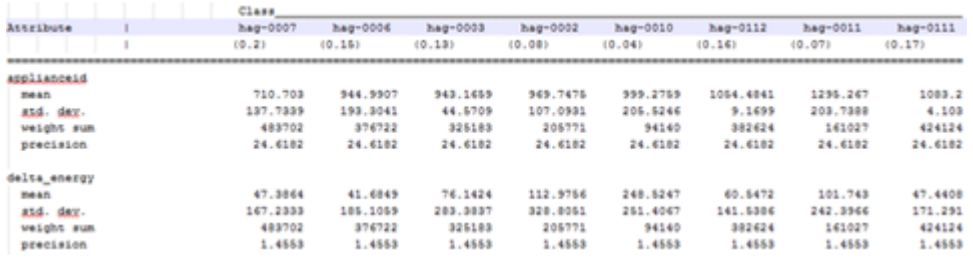
Per la matrice di confusione, si veda la *fig 5.1*

- **NaiveBayes**
  L'algoritmo sul modello semplificato di Bayes ha prodotto il seguente risultato, così come tabellato da Weka:



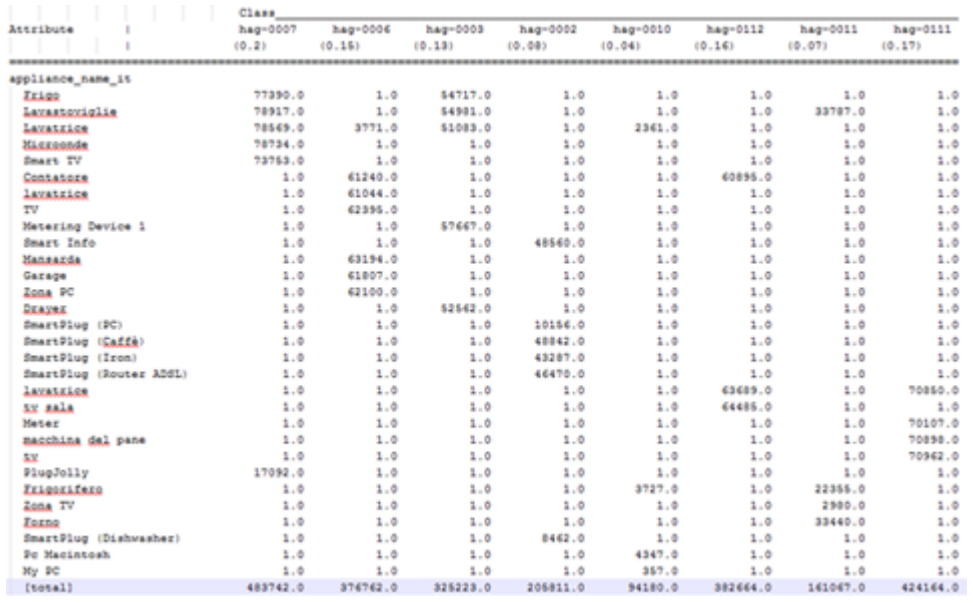
```
=== Confusion Matrix ===

     a      b      c      d      e      f      g      h   <-- classified as
483696      0      3      0      1      0      1      1 |      a = hag-0007
     0 376721      0      0      1      0      0      0 |      b = hag-0006
     3      0 325178      1      1      0      0      0 |      c = hag-0003
     0      1      0 205769      1      0      0      0 |      d = hag-0002
     1      3      0      0  94135      0      1      0 |      e = hag-0010
     0      0      0      0      0 382620      1      3 |      f = hag-0112
     0      1      0      0      1      1 161021      3 |      g = hag-0011
     0      0      0      0      0      0      0 424124 |      h = hag-0111
```

Figura 4.5: *Tree Folder: Matrice di confusione*



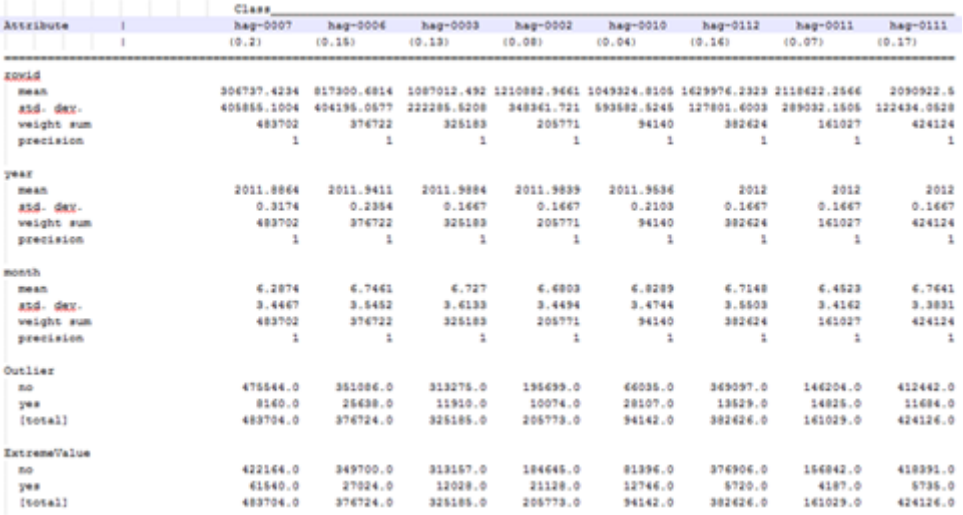
Attribute	Class: hag-0007 (0.2)	hag-0006 (0.15)	hag-0003 (0.13)	hag-0002 (0.08)	hag-0010 (0.04)	hag-0112 (0.16)	hag-0011 (0.07)	hag-0111 (0.17)
applianceid								
mean	710.703	944.9907	943.1659	969.7475	999.2759	1054.4841	1295.267	1083.2
std. dev.	137.7839	193.3041	44.5709	107.0931	205.5246	9.1699	203.7388	4.103
weight sum	483702	376722	325183	205771	94140	382624	161027	424124
precision	24.6182	24.6182	24.6182	24.6182	24.6182	24.6182	24.6182	24.6182
delta_energy								
mean	47.3864	41.6849	76.1424	112.9756	248.5247	60.5472	101.743	47.4408
std. dev.	167.2333	185.1069	283.3837	328.8051	251.4067	141.5386	242.3966	171.291
weight sum	483702	376722	325183	205771	94140	382624	161027	424124
precision	1.4553	1.4553	1.4553	1.4553	1.4553	1.4553	1.4553	1.4553

Figura 4.6: *Naive Bayes 1 di 3*



Attribute	Class: hag-0007 (0.2)	hag-0006 (0.15)	hag-0003 (0.13)	hag-0002 (0.08)	hag-0010 (0.04)	hag-0112 (0.16)	hag-0011 (0.07)	hag-0111 (0.17)
appliance_name_it								
Frigo	77390.0	1.0	54717.0	1.0	1.0	1.0	1.0	1.0
Lavastoviglie	78917.0	1.0	54981.0	1.0	1.0	1.0	33787.0	1.0
Lavatrice	78569.0	3771.0	51083.0	1.0	2361.0	1.0	1.0	1.0
Microonde	78734.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
Smart TV	73753.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
Contatore	1.0	61240.0	1.0	1.0	1.0	60895.0	1.0	1.0
lavatrice	1.0	61044.0	1.0	1.0	1.0	1.0	1.0	1.0
TV	1.0	62395.0	1.0	1.0	1.0	1.0	1.0	1.0
Metering Device 1	1.0	1.0	57667.0	1.0	1.0	1.0	1.0	1.0
Smart Info	1.0	1.0	1.0	48560.0	1.0	1.0	1.0	1.0
Mansarda	1.0	63194.0	1.0	1.0	1.0	1.0	1.0	1.0
Garage	1.0	61807.0	1.0	1.0	1.0	1.0	1.0	1.0
Zona PC	1.0	62100.0	1.0	1.0	1.0	1.0	1.0	1.0
Dryer	1.0	1.0	52562.0	1.0	1.0	1.0	1.0	1.0
SmartPlug (PC)	1.0	1.0	1.0	10156.0	1.0	1.0	1.0	1.0
SmartPlug (Caffè)	1.0	1.0	1.0	48842.0	1.0	1.0	1.0	1.0
SmartPlug (Iron)	1.0	1.0	1.0	43287.0	1.0	1.0	1.0	1.0
SmartPlug (Router ADSL)	1.0	1.0	1.0	46470.0	1.0	1.0	1.0	1.0
lavatrice	1.0	1.0	1.0	1.0	1.0	63689.0	1.0	70850.0
tv sala	1.0	1.0	1.0	1.0	1.0	64485.0	1.0	1.0
Meter	1.0	1.0	1.0	1.0	1.0	1.0	1.0	70107.0
macchina del pane	1.0	1.0	1.0	1.0	1.0	1.0	1.0	70898.0
tv	1.0	1.0	1.0	1.0	1.0	1.0	1.0	70962.0
PlugJolly	17092.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
Frigorifero	1.0	1.0	1.0	1.0	3727.0	1.0	22355.0	1.0
Zona TV	1.0	1.0	1.0	1.0	1.0	1.0	2980.0	1.0
Forno	1.0	1.0	1.0	1.0	1.0	1.0	33460.0	1.0
SmartPlug (Dishwasher)	1.0	1.0	1.0	8462.0	1.0	1.0	1.0	1.0
Pc Macintosh	1.0	1.0	1.0	1.0	4347.0	1.0	1.0	1.0
My PC	1.0	1.0	1.0	1.0	357.0	1.0	1.0	1.0
[total]	483742.0	376762.0	325223.0	205811.0	94180.0	382664.0	161067.0	424164.0

Figura 4.7: *Naive Bayes 2 di 3*



	Class							
Attribute	hag-0007	hag-0006	hag-0003	hag-0002	hag-0010	hag-0112	hag-0011	hag-0111
	(0.2)	(0.15)	(0.13)	(0.08)	(0.04)	(0.16)	(0.07)	(0.17)
rowid								
mean	306737.4234	817300.6814	1087012.492	1210882.9661	1049324.8105	1629976.2323	2118622.2566	2090922.5
std. dev.	405855.1004	404195.0577	222285.5208	348361.721	593582.5245	127801.6003	289032.1505	122434.0528
weight sum	483702	376722	325183	205771	94140	382624	161027	424124
precision	1	1	1	1	1	1	1	1
year								
mean	2011.8864	2011.9411	2011.9884	2011.9839	2011.9536	2012	2012	2012
std. dev.	0.3174	0.2354	0.1667	0.1667	0.2103	0.1667	0.1667	0.1667
weight sum	483702	376722	325183	205771	94140	382624	161027	424124
precision	1	1	1	1	1	1	1	1
month								
mean	6.2874	6.7461	6.727	6.6803	6.8289	6.7148	6.4523	6.7641
std. dev.	3.4467	3.5452	3.6133	3.4494	3.4744	3.5503	3.4162	3.3831
weight sum	483702	376722	325183	205771	94140	382624	161027	424124
precision	1	1	1	1	1	1	1	1
Outlier								
no	475644.0	351086.0	313275.0	195699.0	66035.0	369097.0	146204.0	412442.0
yes	8160.0	25638.0	11910.0	10074.0	28107.0	13529.0	14825.0	11684.0
[total]	483704.0	376724.0	325185.0	205773.0	94142.0	382626.0	161029.0	424126.0
ExtremeValue								
no	422164.0	349700.0	313157.0	184645.0	81396.0	376906.0	156842.0	418391.0
yes	61540.0	27024.0	12028.0	21128.0	12746.0	5720.0	4187.0	5735.0
[total]	483704.0	376724.0	325185.0	205773.0	94142.0	382626.0	161029.0	424126.0

Figura 4.8: *Naive Bayes 3 di 3*

- **BayesNet**

  Il grafo aciclico generato da questo algoritmo sul valore dell'attributo HAG è quello visibile in figura 5.5

  Applicando,invece, l'algoritmo considerando l'attributo APPLIANCE, si rilevano le seguenti informazioni preliminari prima del *fold building*:

  # attributes=9 # classindex=3

  Network structure (nodes followed by parents)

  hag(8): appliance name

  applianceid(56): appliance name

  delta energy(266): appliance name

  appliance name(40):

  rowid(56): appliance name

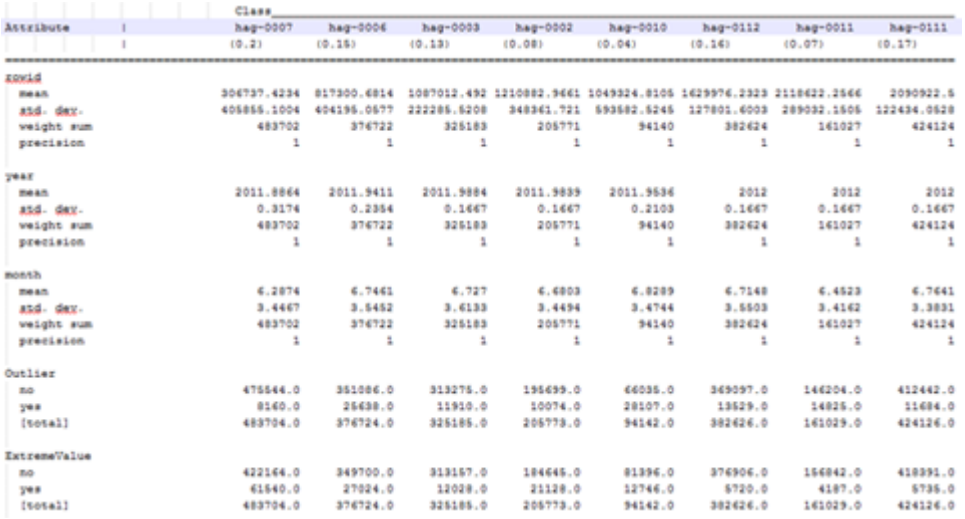
  year(2): appliance name

  month(11): appliance name

  Outlier(2): appliance name

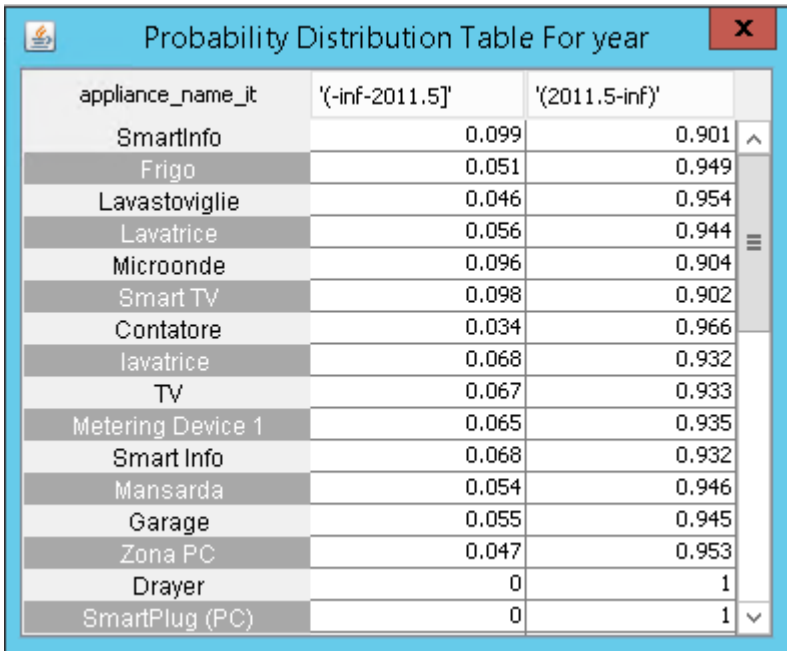
  Extreme Value(2): appliance name

  LogScore Bayes: -2.2214799124271825E7



Attribute		Class hag-0007 (0.2)	hag-0006 (0.15)	hag-0003 (0.13)	hag-0002 (0.08)	hag-0010 (0.04)	hag-0112 (0.16)	hag-0011 (0.07)	hag-0111 (0.17)
zoneid									
mean		306737.4234	817300.6814	1087012.492	1210882.9461	1049324.8105	1629976.2323	2118622.2566	2090922.5
std. dev.		405855.1004	404195.0677	222285.6208	348361.721	593582.5245	127801.6003	289032.1505	122434.0628
weight sum		483702	376722	325183	205771	94140	382624	161027	424124
precision		1	1	1	1	1	1	1	1
year									
mean		2011.8864	2011.9411	2011.9884	2011.9839	2011.9536	2012	2012	2012
std. dev.		0.3174	0.2354	0.1667	0.1667	0.2103	0.1667	0.1667	0.1667
weight sum		483702	376722	325183	205771	94140	382624	161027	424124
precision		1	1	1	1	1	1	1	1
month									
mean		6.2874	6.7461	6.727	6.6803	6.8289	6.7148	6.4523	6.7641
std. dev.		3.4467	3.5452	3.6133	3.4494	3.4744	3.5503	3.4162	3.3831
weight sum		483702	376722	325183	205771	94140	382624	161027	424124
precision		1	1	1	1	1	1	1	1
Outlier									
no		475544.0	351086.0	313275.0	195699.0	66035.0	369097.0	146204.0	412442.0
yes		8160.0	25638.0	11910.0	10074.0	28107.0	13529.0	14825.0	11684.0
[total]		483704.0	376724.0	325185.0	205773.0	94142.0	382626.0	161029.0	424126.0
ExtremeValue									
no		422164.0	349700.0	313157.0	184645.0	81396.0	376906.0	156842.0	418391.0
yes		61540.0	27024.0	12028.0	21128.0	12746.0	5720.0	4187.0	5735.0
[total]		483704.0	376724.0	325185.0	205773.0	94142.0	382626.0	161029.0	424126.0

Figura 4.9: *BayesNet: Hag*



Probability Distribution Table For year

appliance_name_it	'(-inf-2011.5]'	'(2011.5-inf)'
SmartInfo	0.099	0.901
Frigo	0.051	0.949
Lavastoviglie	0.046	0.954
Lavatrice	0.056	0.944
Microonde	0.096	0.904
Smart TV	0.098	0.902
Contatore	0.034	0.966
lavatrice	0.068	0.932
TV	0.067	0.933
Metering Device 1	0.065	0.935
Smart Info	0.068	0.932
Mansarda	0.054	0.946
Garage	0.055	0.945
Zona PC	0.047	0.953
Drayer	0	1
SmartPlug (PC)	0	1

Figura 4.10: *BayesNet(Appliance): Probability Distribution Table For Year*

LogScore BDeu: -2.2369379259557106E7

LogScore MDL: -2.2361189312302962E7

LogScore ENTROPY: -2.2244670170021296E7

LogScore AIC: -2.2260509170021296E7

Figura 4.11: *BayesNet: Appliance*

4.5 Gli algoritmi a confronto

		Weighted Avg.					
		TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area
HAG	NaiveBayes (Hag)	0.965	0.003	0.969	0.965	0.966	0.995
	RandomTree (Hag)	1.000	0.000	1.000	1.000	1.000	1.000
	BayesNet (Hag)	0.640	0.012	0.572	0.230	0.740	0.985
APP	NaiveBayes (Appl.)	0.972	0.005	0.909	0.865	0.936	0.965
	RandomTree (Appl.)	1.000	0.000	1.000	1.000	1.000	1.000
	BayesNet (Appl.)	0.680	0.014	0.666	0.680	0.640	0.971

Figura 4.12: *Gli algoritmi a confronto: Dettagli*

Dalle informazioni fornite da Weka in seguito all'elaborazione dei dati nella sezione *Classify*, è possibile visualizzare un confronto sulla precisione degli algoritmi rispetto alla classe HAG ed alla classe APPLIANCE.

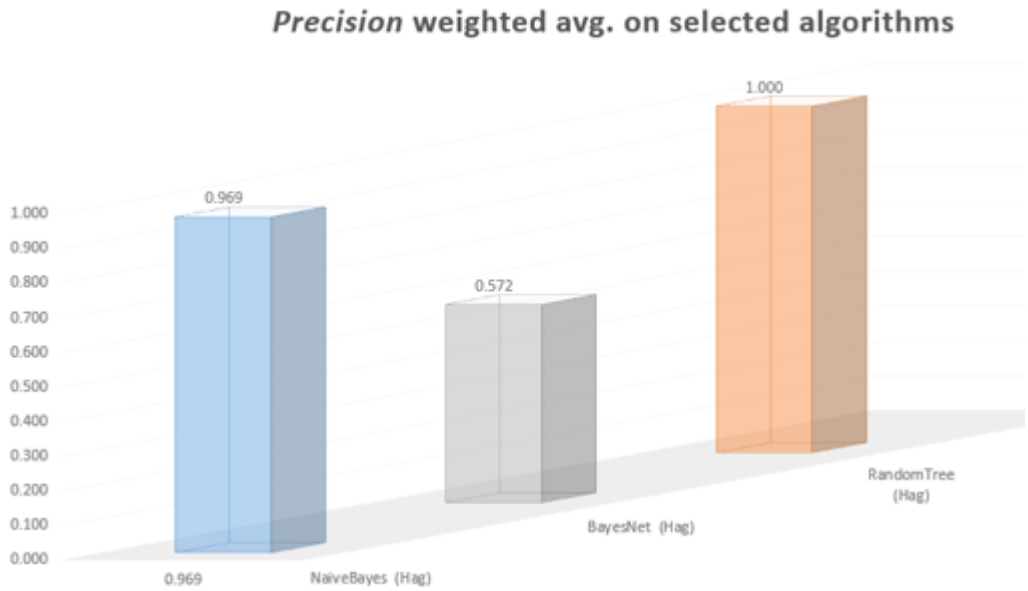


Figura 4.13: *Gli algoritmi a confronto: Precision on HAG*

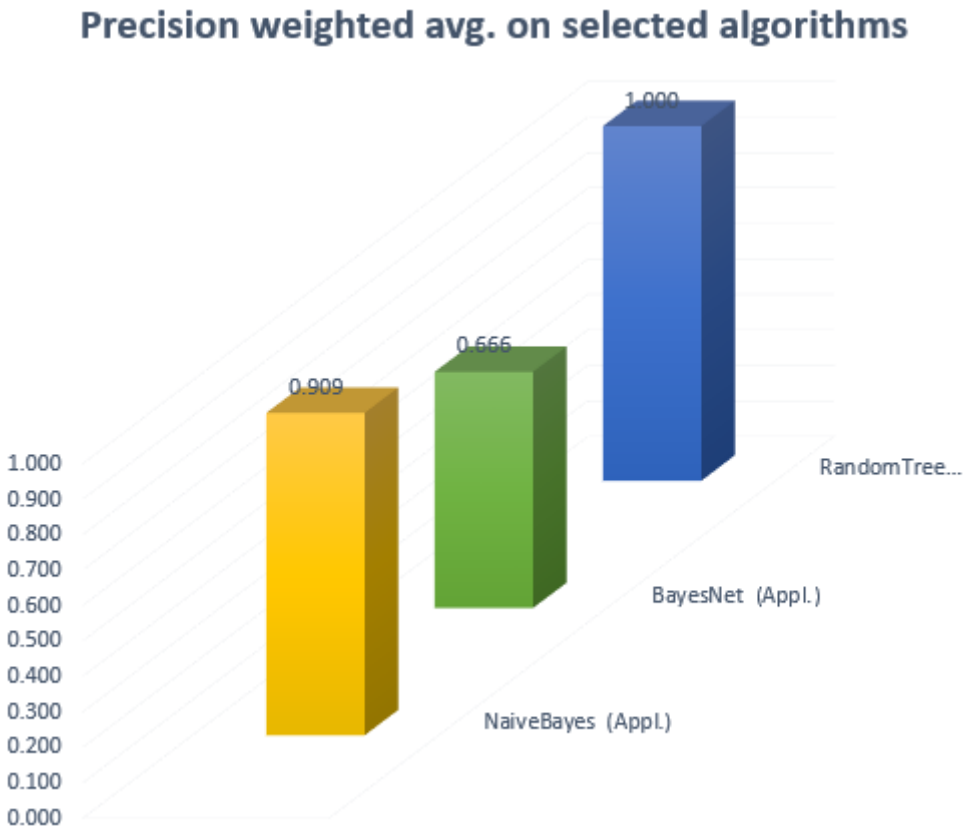


Figura 4.14: *Gli algoritmi a confronto: Precision on APPLIANCE*

Capitolo 5

Conclusioni

In questo capitolo saranno mostrati i risultati ottenuti nella fase di validazione e saranno altresì analizzate le criticità riscontrate e le ipotesi di sviluppi futuri possibili.

5.1 Difficoltà e criticità

Le difficoltà incontrate durante le diverse fasi del progetto possono essere elencate tenendo in considerazione il contesto ed eventuali soluzioni applicate.

- Sviluppo del software *Raw Data Synchronizer*
  Problema: la gestione dei file excel per via del formato proprietario.
  Soluzione: conversione in run-time in un file CSV utilizzando le API Interop di Microsoft.
  Problema: condivisione dei file del progetto software tra 2 server e un client.
  Soluzione: utilizzo di Drop Box per la sincronizzazione dei file.

- Progettazione della base dati
  Problema: Gestire correttamente il tipo di dato avendo come sorgente valori testuali (file CSV).

Soluzione: Demandare al software la conversione dei dati testuali nel corretto tipo (Convert.*metodo*)

- Validazione con Weka
  Problema: La RAM richiesta durante l'applicazione dei patter causava il crash di Weka.
  Soluzione: Creazione di una seconda macchina virtuale con 12Gb di RAM.

5.2 Sviluppi futuri e considerazioni

Portando avanti lo sviluppo del progetto software, si potrebbero implementare nuove feature quali:

- Realizzazione di un adapter per la connessione con altri DBMS
- Realizzazione di uno strumento di mapping dei file excel e/o CSV per la creazione di modelli di mapping da salvare e riutilizzare per i successivi import.
- Realizzazione di web service come ulteriore possibilità di input
- Migliorare l'integrazione di Weka con Raw Data Synchronizer utilizzando le API a disposizione

Ho ritenuto molto utile avere la possibilità di implementare un software per consentire lo storaging di dati provenienti da fonti eterogenee ed è stato significativo il tema dell'interoperabilità e integrazione dei software.
Durante la stesura di questo elaborato è stato immediato pensare all'immensa rete di comunicazione di cui oggi disponiamo così come, allo stesso modo, è stato sorprendente capire come ci sia ancora molto da *connettere* ed analizzare.
Aumentano le informazioni e la frequenza con la quale i sistemi le generano e si pensa a come migliorare le capacita di archiviazione e di analisi, ideando

nuovi modelli per strutturare gli obiettivi di ricerca.
Credo molto nel potenziale dei Big Data sul tema delle politiche energetiche e di come le buone idee, algoritmi e buon senso possano migliorare la qualità della vita per le generazioni future, ridando respiro ad un pianeta incautamente troppo sfruttato nelle sue risorse. Può funzionare.

Ringraziamenti

Ringrazio l'Alma Mater Studiorum, i docenti per la loro preparazione ed il Relatore di questa tesi, prof. Marco Di Felice, per la sua competenza e disponibilità.
Ringrazio Pyxis per la disponibilità dimostrata nel mio impegno universitario, Mauro per le lunghe chiacchierate su ogni esame durante le nostre trasferte e Sergio di Italtel per il suo interesse nel mio percorso.
In questa avvincente esperienza ho potuto incontrare i colleghi più disparati, ma con Nic e Bure è nato un team incredibile. Grazie a Lambdo e Beto per il sostegno morale durante le notti di studio.
Un grazie di cuore ai miei genitori per il patrimonio genetico e per avermi insegnato a combattere e grazie ai miei fratelli per il sostegno da remoto.
Ringrazio, inoltre, Ennio per il suo 'Ciao Tommà,...' post esame.
Un ringraziamento a parte va alla mia compagna di molte avventure,
Marlisa. Grazie per aver imparato tutti i miei esami (capisco l'urto psicologico causato da un integrale doppio ad un avvocato) e per avermi sostenuto nei momenti nevralgici di questo percorso. Da te ho appreso l'importanza della libertà per mezzo della cultura. A te dedico questo traguardo.

Bibliografia

[1] Analisi sui consumi domestici 2000-2012. `http://dati.istat.it/Index.aspx?DataSetCode=DCCV_CNSENRG#`.

[2] Mark Beyer. Gartner says solving 'big data' challenge involves more than just managing volumes of data. 2011.

[3] S.J. Bossart e J.E. Bean. Metrics and benefits analysis and challenges for smart grid field projects. In *Energytech, 2011 IEEE*, pp. 1–5, May 2011.

[4] Oracle Company http://www.oracle.com/us/products/database/big-data-for-enterprise 519135.pdf. Oracle: Big data for the enterprise. January 2012.

[5] Li Mu e Zhu Lei. Big data processing technology research and application prospects. In *Instrumentation and Measurement, Computer, Communication and Control (IMCCC), 2014 Fourth International Conference on*, pp. 269–273, Sept 2014.

[6] Naoya Otsuka e Mitsunori Matsushita. Constructing knowledge using exploratory text mining. In *Soft Computing and Intelligent Systems (SCIS), 2014 Joint 7th International Conference on and Advanced Intelligent Systems (ISIS), 15th International Symposium on*, pp. 1392–1397, Dec 2014.

[7] M.M. Rahman e A. Mto. Technologies required for efficient operation of a smart meter network. In *Industrial Electronics and Applications (ICIEA), 2011 6th IEEE Conference on*, pp. 809–814, June 2011.

[8] A. Wright-R. Wall S. Firth, K. Lomas. Identifying trends in the use of domestic appliances from houshold eletricity cunsumption measurements. 2007.