LUISS Guido Carli
LIBERA UNIVERSITA' INTERNAZIONALE DEGLI STUDI SOCIALI

Dipartimento di Scienze Politiche

Cattedra di Statistica

Tecniche e Metodologie di Data Mining: Analisi applicata alle Charities inglesi

RELATORE

Prof.ssa Livia De Giovanni

CANDIDATO

Filippo Fuggitti

Matricola 078262

ANNO ACCADEMICO 2015 / 2016

INDICE

INTRODUZIONE

Dalla metà degli anni Novanta in poi, il Web si è progressivamente affermato e diffuso come una piattaforma sempre più onnicomprensiva e capillare, verso la quale hanno finito per confluire, ad un ritmo impressionante e spesso caotico, una moltitudine di dati, a volte privi di significato intrinseco preciso.

In virtù di questa continua proliferazione di dati su Internet, connaturata essenzialmente alla definitiva affermazione dell'era digitale, la comunità accademica mondiale ha avvertito chiaramente l'esigenza di elaborare metodologie scientifiche, idonee a gestire ed interpretare questa crescente massa di dati a disposizione, per porli a servizio di scopi sociali, statistici ed economici, proponendo anche soluzioni di *business intelligence* innovative e, al tempo stesso, efficaci a beneficio delle comunità e delle aziende, comprese le organizzazioni non profit (Terzo settore).

Di ciò danno testimonianza le analisi comparate sui dati, tratti dalle ricerche promosse dalla Johns Hopkins University, relative al Terzo settore, che dimostrano come "il settore non profit ha assunto quasi ovunque…dimensioni economiche ed occupazionali assai rilevanti. Nei paesi oggetto dell'indagine, le spese complessive del settore ammontano a circa 1,4 miliardi di euro, una misura che ne farebbe l'ottava economia mondiale"[1].

In questo contesto, si sottolinea l'importanza fondamentale che ha assunto progressivamente la metodologia statistica del *Data Mining*, cuore pulsante del processo KDD (Knowledge Discovery in Database), nel rendere accessibili (o sarebbe più opportuno usare il termine "scavare nei dati") i dati "grezzi", presenti nei *database* delle organizzazioni non profit, estraendone informazioni, *patterns* e relazioni non immediatamente identificabili né conosciute a priori, ma utili per analizzare e comprendere gli aspetti sociali ed economici connessi alla realtà gestita al fine di razionalizzare ed ottimizzare la raccolta e la gestione delle risorse.

La presente ricerca ha delimitato il campo d'analisi alla raccolta ed alla gestione di dati delle organizzazioni non profit inglesi, in particolar modo delle *Charities*, cioè "enti privati d'interesse pubblico", diffuse capillarmente in tutta la Gran Bretagna (par.1.2), presentandone le caratteristiche distintive (par. 1.2.1), le differenti forme istituzionali (par.1.2.2) e il regime economico e finanziario (par. 1.3), evidenziando le principali fonti di finanziamento (par. 1.4).

[1] L.M. Salamon, H.K. Anheier, R. List, S. Toepler, W. Sokolowski et al., Global Civil Society. Dimensions of the Nonprofit Sector, Centre for Civil Society Studies, 1999, pp.252-253

In seguito, l'attenzione è stata indirizzata allo studio delle differenti e multiformi metodologie/tecniche di Data Mining (DM) (cap. 2), anello di congiunzione del processo di analisi dei dati, finalizzato alla scoperta di informazioni utili per comprendere e prevedere l'andamento di determinate variabili quantitative.

Tuttavia, il DM è un processo creativo che richiede una serie complessa di conoscenze e competenze ed attualmente non è ancora disponibile un approccio standardizzato, che aiuti a tradurre i problemi di *business* in compiti di DM, suggerisca appropriate interpretazioni dei dati e delle tecniche di data mining e, infine, fornisca gli strumenti necessari per valutare l'efficacia dei risultati, documentando l'esperienza in corso d'opera.

In tal senso, il progetto CRISP-DM (*Cross Industry Standard Process for Data Mining*) (cap. 3) propone un Modello procedurale, comprensivo ed intuitivo, idoneo per realizzare progetti di *Data Mining* maggiormente affidabili, reiterabili, gestibili, rapidi, a costi competitivi.

In conclusione, la presente ricerca propone un progetto di DM (cap. 4), che consiste nell'applicazione di modelli di analisi di serie storiche alle statistiche ufficiali sulle organizzazioni non profit *dell'HM Revenue & Customs,* l'Agenzia delle Entrate inglese, per effettuare delle previsioni sul numero di donatori e sull'importo annuo lordo delle donazioni ricevute da tali associazioni nel corso di 1990-2014.

CAPITOLO 1

1. Le organizzazioni non profit

Gli enti non profit, la cui origine risale storicamente ad alcuni secoli fa, hanno di recente riscoperto nuova importanza ed utilità sociale. Difatti, si avverte la necessità di ricercare percorsi e modelli che consentano di affrontare le nuove e urgenti problematiche legate alla crisi del *Welfare State,* ai mutamenti strutturali, sociali e culturali in atto. Tali esigenze spingono a superare il dualismo Stato-Mercato attraverso "la valorizzazione di sfere d'azione sottratte sia ai processi di mercificazione, quantunque strettamente intrecciate, che alla sfera autoritativa pubblica: sfere d'azione imperniate sul volontariato, sull'altruismo, sulla reciprocità, sulla solidarietà, sulla produzione non mercificata di relazionalità e socialità"[2].

Un'organizzazione *non profit* è un'associazione progettata per fini che esulano dal mero scopo di lucro ed in cui nessun provento dell'organizzazione è destinato a managers, soci, o funzionari appartenenti a quest'ultima. Gli enti non-profit sono spesso definiti "*non-stock corporation"*. Potranno, dunque, assumere la forma giuridica di società per azioni (privati daranno vita ad associazioni, sostenute principalmente dai contributi di beneficenza dei privati), di fondazioni e partnership (le quali si distinguono dalle S.p.A. sia per il fatto di poter ricevere sovvenzioni da un fondatore sia di poter assumere la forma giuridica di una amministrazione fiduciaria) o di condomini (in cui i proprietari di singole unità abitative stabiliranno consensualmente la comproprietà delle aree comuni, in base alle leggi statali vigenti). Dal momento in cui tali organizzazioni sono costituite, le stesse vengono organizzate e programmate in modo che non abbiano alcun fine di lucro e perseguano obiettivi ammessi legittimamente dagli statuti di tali associazioni.

Nel 1996, Lester M. Salamon e Helmut K. Anheier, membri accademici dell'Istituto di Studi Politici della Johns Hopkins University, stilarono la cosiddetta *International Classification of Nonprofit Organizations (ICNPO)*, basandosi sui dati raccolti dall'*International Standard Industrial Classification (ISIC)* redatta dalle Nazioni Unite e, modellando su di essa, una realistica configurazione del settore non-profit in 11 paesi che vennero coinvolti nella prima fase di questa ricerca accademica (U.S.A, U.K., Francia, Germania, Italia, Svezia, Giappone, Ungheria, Brasile, Ghana Egitto, India e Thailandia). La *ICNPO* individuò le principali e peculiari caratteristiche che contribuiscono a delineare il settore non profit.

In particolare, tali associazioni non profit possono essere:

- *Organizzate o parzialmente istituzionalizzate.* Tali associazioni presentano internamente una qualche forma o realtà istituzionale. Quest'ultima include un determinato grado di struttura organizzativa interna; comunanza di obiettivi e attività e limiti organizzativi significativi (ad esempio, la distinzione riconosciuta fra membri e non membri).

[2] U. Ascoli (cura di), *Il Welfare futuro. Manuale critico del Terzo settore*, Carocci, Roma, 1999, p. 13

- *Private o istituzionalmente separate dal governo centrale.* Le organizzazioni non profit non sono apparati governativi. Dunque, manifestano un'identità istituzionale distinta da quella dello Stato centrale, non sono enti pubblici dipendenti dal governo nazionale o locale e, dunque, non esercitano alcun potere governativo autoritativo.

- *Self-governing.* Sono assolutamente nelle condizioni di amministrare e controllare autonomamente le proprie attività o iniziative. Ciò implica necessariamente la costituzione di procedure di *governance* strettamente indipendenti da istituzioni o agenzie governative.

- *Senza fini di lucro.* Sono associazioni private non finalizzate a generare profitti, sia direttamente o indirettamente; inoltre non perseguono scopi di carattere commerciale o economico.

- *Volontarie.* Affinché tali associazioni possano essere incluse nel settore non-profit, quest'ultime dovranno incorporare il concetto di *voluntarism*, ossia sarà reso loro possibile di svolgere attività volontarie e gratuite a favore della collettività (in particolare a beneficio dei malati e dei bisognosi).

Tale concetto di *voluntarism* porta necessariamente a considerare due differenti aspetti. In primo luogo, tali enti dovranno coinvolgere i volontari nella gestione operativa della cooperativa (ad esempio, inserendo quest'ultimi nel *board* aziendale oppure impiegandoli nello *staff*). In secondo luogo, il servizio di "volontariato" è non coattivo, ossia non è richiesta alcuna iscrizione o registrazione di coloro che prestano servizio presso tali organizzazioni e parimenti, i privati cittadini potranno effettuare facoltativamente donazioni o, altrimenti, partecipare alle attività di volontariato senz'alcun obbligo di legge.

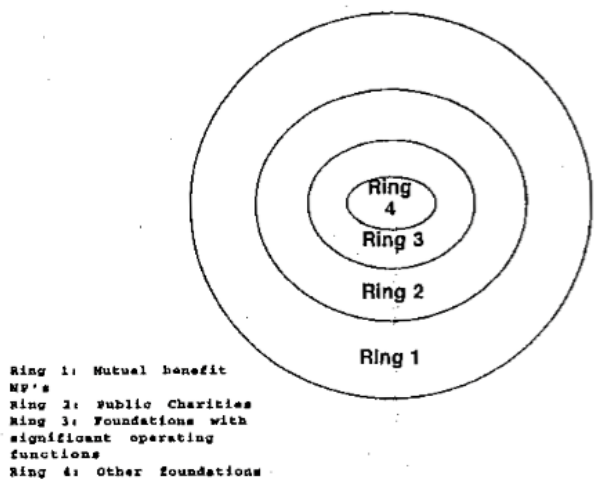


Figure 1.1. The Rings of Nonprofit Organizations. *Source:* John Simon, "The tax treatment of nonprofit organizations: A review of federal and state policies" in W. W. Powell, *The Nonprofit Sector*, New Haven: Yale University Press, 1989. p. 68.

Fig 1.1 Schema rappresentativo delle organizzazioni non profit

1.1 Cenni storici sulle organizzazioni non profit

Il settore della beneficenza e del volontariato annovera antichi precedenti, avendo fornito assistenza *not for profit* alle persone bisognose per diversi secoli, con una crescita spettacolare tra la fine del 20º secolo e gli inizi del 21º secolo.

In un'ottica europea, i paesi anglosassoni condividono la storia di uno stato sociale che, da sempre, ha riconosciuto l'importanza dell'iniziativa privata. "Del resto il concetto stesso di settore non profit o di Terzo settore nasce qui, nella tradizione della Common Law britannica, che premia innovazione e flessibilità, evitando una codificazione rigida degli assetti istituzionali e fiscali"[3].

La Charity Organization Society, fondata a Londra nel 1883, fu, a quel tempo, una fra le più importanti organizzazioni civili dell'Impero britannico e simili *network* di servizi umanitari e caritatevoli emersero in Italia, Germania, Francia, Australia e Giappone. Le organizzazioni di servizi sociali cattoliche e protestanti, "Caritas" e "Diakonie" sono attualmente fra i più importanti datori di lavoro in Germania con 1 milione di occupati; e UNIOPSS, federazione francese di organizzazioni nonprofit che si occupa di fornire servizi sociali e sanitari, conta più di 750.000 lavoratori.

Questi enti assistenziali non sono solamente protagonisti nell'ambito del sistema welfare e delle cure sanitarie, ma si distinguono per il loro operato anche in altre aree (di attività) socio-economiche. Ad esempio, in campo culturale ed artistico, la "Tate Modern" a Londra o il "Guggenheim" a Bilbao e, similmente, in ambito accademico, la "London School of Economics" sono considerate organizzazioni non-profit. Infine, è necessario citare le più significative associazioni umanitarie internazionali, fra cui l'organizzazione non governativa inglese (NGO) "Oxfam", l'associazione diritti umani "Amnesty International" e "Greenpeace" in Olanda.

1.2 La "Charity"

"For the purposes of the law of England and Wales, "charity" means an institution which—

a) is established for charitable purposes only, and
b) falls to be subject to the control of the High Court in the exercise of its jurisdiction with respect to charities."[4]

Il cardine del Terzo settore britannico è rappresentato dalle *Charities*, "enti privati d'interesse pubblico". La categoria, di origine giurisprudenziale, è stata consacrata, nell'ordinamento giuridico inglese, come dato positivo dalla legge sulle associazioni 1 luglio 1901. Quest'ultima, infatti, ha sancito, per gli enti

[3] S. Pasquinelli, Il Terzo settore nei diversi States europei, in U. Ascoli (a cura di), op. cit. p. 97

[4] Charities Act 2011, Part 1, Chapter 1: General

senza fini di lucro, (quali le associazioni e, in seguito, anche per le fondazioni) la possibilità di ottenere il riconoscimento di utilità pubblica, ovvero l'attestazione che i fini perseguiti sono considerati dallo Stato rilevanti per la collettività.

L'attribuzione di tale qualifica determina la possibilità di fruire di una serie di benefici, tra i quali ricevere finanziamenti pubblici per lo svolgimento delle proprie attività. Parimenti, il sistema anglosassone, in base a un provvedimento che risale all'anno 1061, conosce delle figure, chiamate *charities*, che identificano quella parte del terzo settore che ha finalità pubbliche: all'interno di tale categoria, in mancanza di una definizione normativa, si sogliono includere strutture di vario genere, quali organizzazioni di volontariato, fondazioni, istituzioni nazionali, associazioni, strutturate in forme altrettanto varie.

Il quadro è alquanto articolato, complesso e disomogeneo: a fianco di organizzazioni con elementi formali minimi, emergono associazioni fortemente strutturate ed orientate allo svolgimento di proprie attività e alla relativa raccolta di fondi (*fundraising*).

Ora, adottando un approccio sistematico e integrato, la dottrina stabilisce che è possibile individuare tre categorie: i *trust,* che determinano preliminarmente le finalità a cui destinare i fondi, spesso frutto di atti di liberalità; le associazioni create espressamente per la realizzazione di finalità *charitable* o *caritatevoli*; le *charitable companies.* Ad ogni modo, indipendentemente da tale sistematizzazione, il fattore rilevante è costituito dai tratti che accomunano le tipologie di *charities.*

Sinteticamente, esse esercitano senza scopo di lucro attività o iniziative per la comunità e, quindi, di pubblica utilità, in particolare nei settori socio-assistenziale, sanitario, educativo, culturale, ricreativo, ambientale, religioso.

"The of aged, impotent and poor people, the maintenance of sick and maimed soldiers and mariners, schools of learning, free schools and schools in universities, the repair of bridges, ports, havens, causewys, churches, sea-banks and highways, the education and preferment of orphans, the reliefs, stock or maintenance for houses of correction, the marriage of poor maids, the support, aid and help of young tradesmen, handicraftsmen and person decayed, the relief or redemption of prisoners or captives and the aid or ease of any poor inhabithants concerning payment of fifteens, setting out of soldiers and other taxes"[5].

I *charitable purposes,* menzionati nel Preambolo dello "Statute of Elisabeth I" e denominato anche "Charitable Uses Act 1601", non vanno considerati una categoria "chiusa", ma al contrario, rappresentano un elenco flessibile e dinamico in grado di accogliere ogni altro scopo caritatevole, purché, come sentenziato nel caso "Morice v. The Bishop of Durham," rientri nello *"spirit and intendment of the act".*

Di recente, il legislatore inglese è intervenuto col *Charity Act* dell'anno 2006, ove vengono elencate puntualmente le dodici finalità delle organizzazioni non profit anglosassoni, tra cui troviamo la prevenzione

[5] Charitable Uses Act 1601, Preambolo

e il sollievo dalla povertà, la tutela e l'avanzamento dell'istruzione, della religione, della salute, dell'arte, della cultura e della scienza, dello sport amatoriale, dei diritti umani, dell'ambiente e della protezione degli animali.

Secondariamente, esse presentano un'identità statutaria distinta da quella statale e si trovano in posizione d'indipendenza rispetto all'organizzazione statale, per quanto riguarda l'individuazione degli obiettivi e lo svolgimento delle attività.

Infine le *charities*, coerentemente con la loro natura non lucrativa, sono tenute a rispettare il vincolo di non distribuzione dei profitti; conseguentemente al riconoscimento del loro ruolo e al perseguimento di finalità pubbliche proprie di tali enti, l'ordinamento giuridico britannico riconosce alle *charities* uno *status* legale protetto.

Da un lato, ciò comporta la concessione di diversi benefici, in particolare sgravi fiscali e contributi statali, e dall'altro lato l'imposizione di un regime di responsabilità (*accountability*) che si esplica prevalentemente mediante obblighi di rendicontazione. In relazione a quest'ultimo aspetto, l'ordinamento britannico ha costituito sin dal 1853 un apposito organismo, la *Charity Commission*, istituzione pubblica ed indipendente sotto il controllo del Governo con funzioni di vigilanza, controllo e ispezione.

"The Commission has the following general functions—

1. *Determining whether institutions are or are not charities.*
2. *Encouraging and facilitating the better administration of charities.*
3. *Identifying and investigating apparent misconduct or mismanagement in the administration of charities and taking remedial or protective action in connection with misconduct or mismanagement in the administration of charities.*
4. *Determining whether public collections certificates should be issued, and remain in force, in respect of public charitable collections.*
5. *Obtaining, evaluating and disseminating information in connection with the performance of any of the Commission's functions or meeting any of its objectives.*
6. *Giving information or advice, or making proposals, to any Minister of the Crown on matters relating to any of the Commission's functions or meeting any of its objectives.*

The Commission has the following objectives—

1. *The public confidence objective*

   The public confidence objective is to increase public trust and confidence in charities.
2. *The public benefit objective*

   The public benefit objective is to promote awareness and understanding of the operation

of the public benefit requirement.

3. *The compliance objective*

   The compliance objective is to promote compliance by charity trustees with their legal obligations exercising control and management of the administration of their charities.

4. *The charitable resources objective*

   The charitable resources objective is to promote the effective use of charitable resources.

5. *The accountability objective*

   The accountability objective is to enhance the accountability of charities to donors, beneficiaries and the general public"[6].

In base a tali scopi, essa gestisce un Registro delle *Charities*, istituito nel 1960, per consentire alla *Charity Commission* di accertare che l'associazione richiedente l'iscrizione persegua effettivamente finalità *charitable* e adempia ai requisiti previsti dalla legge: *"the effective use of charitable resources by encouraging the development of better methods of administration, by living the charity trustees information or advice on any matter affecting the charity, and by investigating and checking abuses"*[7].

Dalla registrazione, inoltre, dipende la possibilità di avvalersi della qualifica di *charity* e di conseguenza alla sottoposizione alla relativa normativa. D'altro canto, le *charities* potranno richiedere supporto e affiancamento alla *Charity commission*, che offre un servizio di consulenza e orientamento in materia legale, gestionale, finanziaria e di governance, finalizzato ad incrementare l'effettività della loro azione nel rispetto della disciplina di legge vigente.

1.2.1 Le caratteristiche distintive

L'elemento primario che consente a una associazione di appartenere alla categoria delle charities è il perseguimento di uno scopo di pubblica utilità.

Tuttavia, vi sono altre caratteristiche rilevanti da considerare quali:

a) essere indipendenti;
b) essere "non profit" ovvero non avere scopo di lucro;
c) essere organizzazioni "non politiche".

[6] Charities Act 2011, Part 2, *The Charity Commission and the Official Custodian for Charities*, *The Commission*
[7] Charities Act 1993, Part 1, *The Charity Commissioners and the Official Custodian for Charities*, *The Commissioners*

Il concetto d'*indipendenza* fa riferimento al Governo e più in generale allo Stato. Il perseguimento e il mantenimento di uno stato d'indipendenza della Charity sono un dovere degli amministratori: ad essi è viene attribuita la responsabilità legale che tutte le attività delle charities siano realizzate per il perseguimento dello scopo istituzionale, e hanno il dovere di utilizzare i propri poteri e le risorse della charity per il raggiungimento dello stesso.

Essere non profit significa che non è possibile veicolare le risorse dell'organizzazione per scopi estranei all'organizzazione, ma che dovranno essere impiegate all'interno dell'organizzazione per il perseguimento della finalità pubblica. Gli amministratori dovranno essere sempre in grado di giustificare i costi sostenuti dalla charity rispetto alla finalità da essa prefissati.

In questo quadro, le attività commerciali e di business sono considerate lecite purché realizzano uno scopo *charitable.*

Le *organizzazioni "non politiche"* rispondono al requisito di apoliticità. Il *Charities Act 2006* richiede a tali associazioni e ai loro amministratori di evitare iniziative di natura politica. Pertanto, nessuna charity potrà ispirarsi a ideologie o programmi di partiti politici oppure celare obiettivi natura politica. Tuttavia, la legge consente a tali enti di partecipare alla campagne elettorali e contribuire al dibattito pubblico, purché direttamente connesso agli scopi di *charitable* perseguiti.

1.2.2 Le forme istituzionali: l'associazione, il trust, la charitable company

"Charitability is a status, not a legal form - an official badge wich may be attached to a range of different types of organisation".

Indipendentemente dalla sua struttura interna, una qualsiasi organizzazione di volontariato può essere denominata charity, dal momento in cui è ufficialmente riconosciuto il suo "Charitable Status", vale a dire il ruolo fondamentale ricoperto dall'elemento solidaristico nella gestione delle attività istituzionali, indirizzate a beneficio della società.

Le charities possono assumere diverse forme giuridiche, tuttavia, è possibile individuare tre forme giuridiche tipiche distinte a cui le charities si richiamano:

a) l'associazione;
b) il trust – charity trustees;
c) la charitable company.

L'*associazione* rappresenta il modello più semplice e intuitivo: si tratta di un insieme di membri cui non è riconosciuta personalità giuridica. Un regolamento definisce chiaramente lo scopo dell'ente ed assegna i rispettivi poteri agli organi preposti ad esercitarli; la responsabilità legale per gli obblighi assunti resta in

capo ai membri dell'associazione, sia individualmente sia collettivamente. Si ritiene che sia la forma più adeguata e conforme a gruppi informali che non possiedono strutture articolate e risorse ingenti.

La mancanza di personalità giuridica comporta un'enorme difficoltà per associazioni di minore entità che godono di beni patrimoniali. Proprio per aggirare tale ostacolo, le charities hanno assunto la forma giuridica di società a responsabilità limitata. Attualmente, in Gran Bretagna, sono numerose le charities che beneficiano di tale regime normativo, essendo registrate presso la Companies House, organismo pubblico di vigilanza delle società, ed essendo assoggettate alle sue regole. Da un lato, la normativa sulle società a responsabilità limitata costituisce un impedimento, poiché le charities osservano una doppia normativa, sia di carattere societario e ovviamente, quella tipica delle charities, dall'altro lato consente di fruire di indubbi vantaggi connessi alla fattispecie giuridica denominata "responsabilità limitata".

Il *trust* è lo status giuridico adottato dalle charities fin dal 1600 e, probabilmente, di natura o derivazione ecclesiastica. Si tratta essenzialmente di una promessa, giuridicamente tutelata, effettuata da un soggetto (B) nei confronti di un secondo soggetto (A) di beneficiare da un terzo (C) denaro o beni patrimoniali messi a disposizione da (A). Dunque, (B) non fruisce dei beni di (A) a scopo personale, bensì a beneficio di (C). In termini tecnici, s'individua un patrimonio, donato da parte di una persona fisica o giuridica considerata *charitable*, e, che viene amministrato da un "trust" per il perseguimento di scopi caritatevoli. Spesso i trust sono costituiti mediante un atto testamentario, il quale riporta anche lo scopo charitable prefisso, ma è bene precisare che, affinché possa essere istituito un trust, non è esplicitamente richiesto alcun atto scritto, in quanto gli elementi necessari e sufficienti per dare vita a tale entità giuridica sono il patrimonio e l'indicazione da parte del donatore dello scopo da perseguire: il trust crea, dunque, un legame fiduciario fra donatore, amministratori del trust e il beneficiario.

In particolar modo, il termine *charity trustees* identifica "*the persons having the general control and management of the administration of a charity*"; quindi, comprende non solo i *trustees*, nel senso comunemente inteso, ma anche ad esempio i *directors,* o, qualora sia opportuno o necessario, un *management committee*, come nel caso di una *charity incorporated.*

Una persona non ha le qualifiche per diventare un *charity trustee* o comunque un *trustee* se:

a) *he has been convicted of any offence involving dishonesty or deception;*
b) *has been adjudged bankrupt or sequestration of P's estate has been awarded and has not been discharged or is the subject of a bankruptcy restrictions order or deception;*
c) *has made a composition or arrangement with, or granted a trust deed for, creditors and has not been discharged in respect of it;*
d) *has been removed from the office of charity trustee or trustee for a charity by an order made by the Commission or the High Court on the griunds of any misconduct or mismanagement in the administration of the charity for which he was responsibleor to which he was privy, or which he by his conduct contributed to or facilitated;*

e) has been removed, under section 34(5)(e) of the Charities and Trustee Investment (Scotland) Act 2005 (asp 10) (powers of the Court of Session) or the relevant earlier legislation (as defined by section 179(6)), from being concerned in the management or control of any body;

f) he is a subject to a disqualification order or disqualification undertaking under the Company Directors Disqualification Act 1986 or the Company Directors Disqualification (Northern Ireland) Order 2002 (S.I. 2002/3150 (N.I.4)), or an order made under section 429(2) of the Insolvency Act 1986 (disabilities on revocation of county court administration order)[8].

Per concludere, la *charitable company* è una charity che costituisce a tutti gli effetti una azienda. Affinché essa possa acquisire il cosiddetto *charitable status* ed essere in questo modo considerata una *charitable company*, dovrà effettivamente perseguire *charitable purposes* illustrati nel paragrafo 1.2. Dunque, la *charitable company* manifesta, allo stesso tempo, analogie e differenze dalle imprese private che verranno descritte in seguito.

1.3 Regime economico e finanziario delle Charities

L'acquisizione del titolo di charity si concretizza materialmente in campo fiscale. Difatti, le Charities godono di un regime tributario fortemente agevolato. Tuttavia, un fattore da non sottovalutare è il sentimento di fiducia e di affiliazione che le charities suscitano nell'opinione pubblica, consentendo di attrarre cospicue donazioni da privati e la possibilità di ricevere finanziamenti pubblici.

In base ai relativi dati economici e finanziari, le charities possono essere ricondotte e classificate fra le organizzazioni non-profit in maniera formale e sostanziale. "Tali realtà hanno come finalità prevalente il soddisfacimento diretto di bisogni socialmente rilevanti, rispetto a cui la massimizzazione del reddito costituisce soltanto una finalità secondaria, del tutto strumentale al raggiungimento della prima. Mentre nell'impresa il perseguimento di situazioni di prevalenza dei ricavi sui costi, data la priorità del finalismo di carattere economico sugli altri, rappresenta l'obiettivo-guida di tutta l'attività, negli enti senza fini di lucro questa costituisce esclusivamente il presupposto che consente agli stessi di perseguire nel tempo la finalità sociale in condizioni di autonomia economica"[9].

[8] Charities Act 2011, Part 9, *Charities Trustees, Trustees and Auditors etc*, *Disqualification of charity trustees and trustees*

[9] A. Propersi, *Le aziende non profit. I caratteri, la gestione, il controllo*., Milano: RCS Libri, 1999, pp. 24-25

1.3.1 Elementi distintivi

Gli enti non profit, fra cui le charities, presentano caratteri propri distintivi: *unità*, *autonomia*, *durabilità* e *dinamismo*. L'*unità* consiste nella condivisione di comuni ideali e nell'impegno per il soddisfacimento di bisogni socialmente rilevanti. L'*autonomia*, garantita nelle imprese dagli interessi dei proprietari o azionisti, può essere presente, con diversi gradi di realizzazione, nelle organizzazioni non profit, tuttavia ciò dipende fortemente dalla dotazione patrimoniale, la capacità di fare raccolta fondi, l'adesione di soggetti "forti", ecc. La *durabilità* è legata strettamente al requisito dell'autonomia e discende principalmente dalle capacità manageriali.

Per concludere, il *dinamismo* richiede flessibilità, ossia la capacità di mantenere una certa fallibilità strategica, e al contempo, di pianificare, programmare e controllare le attività svolte.

1.3.2 Caratteristiche strutturali

Le organizzazioni non profit presentano le seguenti caratteristiche strutturali:

a) ricezione di capitali di significativo ammontare da parte di finanziatori che non richiedono alcuna contropartita;
b) l'esercizio di attività operative non finalizzate alla produzione di beni o fornitura di servizi da cui ricavare margini unitari di profitto;
c) l'assenza di interessi definiti che possano essere ceduti, trasferiti o riscattati da parte dei proprietari che offrano il diritto alla distribuzione delle risorse residuali provenienti dalla liquidazione dell'organizzazione.

Sono, altresì, da considerare altri aspetti marginali quali:

d) la difficoltà d'individuare e computare (calcolare), in termini quantitativi, sia il valore aggiunto sia il valore dell'attività svolta;
e) la collaborazione di volontari non remunerati che altruisticamente contribuiscono alla crescita dell'organizzazione, nel segno di ideali comunemente condivisi;
f) la pubblicizzazione dei valori propugnati da personalità influenti, tra cui finanziatori in grado di contribuire al bene pubblico;
g) la tendenza a creare e sviluppare strutture organizzative e di assicurarsi un patrimonio, garantendo la sopravvivenza e la crescita dell'organizzazione;
h) la gestione talvolta approssimata nella fase iniziale del business;
i) una maggiore flessibilità e discrezionalità nell'ambito di progetti realizzati dai componenti dell'associazione rispetto alle imprese private.

In un'economia di mercato, le organizzazioni non commerciali, ossia le charities, fissano prezzi al di sotto del costo di produzione e registrano utili per poter sovvenzionare le diverse iniziative promosse dall'ente, mentre, in alcuni casi quest'ultimo conviene di richiedere alcun corrispettivo in cambio. "E' sostanziale, per questo, sottolineare che l'attività delle organizzazioni non profit non è generalmente soggetta alla prova della diretta competizione nei mercati, come invece avviene per le imprese. Allorquando negli enti è svolta un'attività commerciale, questa è generalmente strumentale a fini istituzionali"[10].

1.3.3 Caratteristiche gestionali

Per quanto riguarda le peculiarità salienti della gestione delle organizzazioni non profit, tra cui le charities, si possono individuare le seguenti tipicità:

a) l'assenza del dato reddituale come indicatore di economicità, di efficienza e di efficacia, costituisce un grave problema ed ostacolo allo sviluppo di validi sistemi di controllo direzionale delle aziende non profit;
b) la tendenza ad assumere la natura economica tipica delle aziende che erogano servizi; dunque, in termini di controllo e gestione, un'azienda che produce e vende beni materiali presenta alcuni vantaggi, in quanto quest'ultimi possono essere immagazzinati e destinati ad altri scopi, mentre, nel caso dei servizi, l'assenza di domanda o nel caso di un evidente calo della domanda rispetto all'offerta, comporta la dissoluzione delle potenzialità organizzative/gestionali della struttura. Inoltre, le aziende di servizi sono organizzazioni ad alta intensità di lavoro e, come tali, richiedono una quota limitata di capitale per unità prodotta, costituendo così un vincolo ai fini del controllo e della valutazione dell'attività dell'ente. Occorre tenere presente quindi che anche la rendicontazione contabile di servizi è complessa e la qualità di essi può essere esaminata e valutata solo dopo che esso sia erogato;
c) gli enti non profit sono sottoposti a vincoli stringenti, quali la definizione di obiettivi e strategie. Essi sono costretti a seguire rigorosamente le direttive statutarie che specificano obblighi in materia di fornitura di servizi e finanziamenti a destinazione vincolata, con il risultato che la gestione è rigidamente monitorata;
d) i meccanismi e i processi di acquisizione di capitali e risorse finanziarie sono definiti primariamente da soggetti esterni all'organizzazione, i cosiddetti *donors*;
e) organizzazione di carattere gerarchico e verticalità nelle relazioni fra membri, venendo

[10] A. Propersi, G. Rossi, *Gli enti non profit,* Milano, 2006

meno una visione aziendalistica globale di sintesi;

f) le aziende senza finalità di lucro sono prive di centri di responsabilità chiaramente individuabili. Ciò è dovuto al fatto che non vi sono azionisti cui rendere conto del proprio operato ed il ruolo degli amministratori è confinato alla promozione di determinati valori etici e non tanto per la loro capacità di gestione e di governo, avendo una conoscenza sommaria e generica dei problemi aziendali ed una "agenda" organizzativa chiaramente definita e circoscritta, il che si riflette sulla qualità delle decisioni;

g) l'assenza di una direzione generale denota il fatto che in tali aziende non profit la responsabilità sociale dell'impresa è condivisa.

Le charities, presentando sia una costituzione formale sia una natura giuridica privata, operano per il diretto conseguimento del bene comune secondo logiche di autogoverno e d'interazione non sinallagmatica con l'ambiente di riferimento.

Esse, rispetto alle imprese, risultano in alcuni casi maggiormente vincolate (si pensi solamente ai vincoli statutari) nei processi di definizione e mutamento degli obiettivi e delle strategie. Si tratta, pertanto, di realtà ad elevato livello di rigidità strategica. Esse, spesso, sono dirette da "tecnici", più o meno impiegati full-time all'interno dell'organizzazione, dotati, non sempre, di competenze manageriali specifiche.

La gestione di questi enti è, talvolta, organizzativamente e amministrativamente approssimata, sorretta soltanto dallo slancio ideale o dalla generosità del fondatore o dal contributo volontario talvolta discontinuo dei simpatizzanti, in particolare nella fase pionieristica.

Le responsabilità spesso non risultano chiare, specie, nelle realtà di piccole dimensioni ed il rapporto che si instaura con la collettività di riferimento è di natura fiduciaria, sia per i servizi prestati che per i fondi ricevuti.

Figure 1.2. Management Schematic.

Fig. 1.2 Caratteristiche gestionali enti non profit

La gestione aziendale degli enti non profit può essere distinta in:

a) *gestione caratteristica istituzionale;*
b) *gestione delle attività connesse a quella istituzionale;*
c) *gestione patrimoniale;*
d) *gestione finanziaria e monetaria;*
e) *gestione della raccolta fondi.*

"La *gestione caratteristica* istituzionale è l'attività propria dell'ente volta all'attuazione degli scopi statutari. Essa impiega gli elementi patrimoniali necessari per il funzionamento dell'ente e i beni strumentali utilizzati nell'attività operativa (…)"[11]. Tale amministrazione coinvolge e dipende fortemente dalle gestioni cosiddetti "accessorie", ovvero ausiliarie a quella caratteristica, quali la gestione patrimoniale in senso stretto e quella delle attività strumentali finalizzate all'esercizio di attività commerciali. Nello specifico, la gestione accessoria può essere di competenza diretta dell'organizzazione oppure assegnata ad altri enti sui quali è esercitato un controllo indiretto.

La *gestione patrimoniale* amministra i cosiddetti "beni da reddito", ossia beni patrimoniali destinati a essere posseduti al solo scopo di generare rendite. La *gestione finanziaria e monetaria* consente di mantenere il bilancio contabile dell'associazione in pareggio e di rendere disponibili le risorse di cui l'ente necessita per gli obiettivi prefissati.

Il *fundraising* è un'iniziativa di marketing volta all'ideazione di campagne pubblicitarie o eventi al fine di ottenere finanziamenti destinati al perseguimento di determinati obiettivi "visibili" e riconosciuti da parte dell'organizzazione.

La *gestione della raccolta fondi*, diretta all'acquisizione di capitali finanziari e non, è destinata ad assumere maggiormente caratteri di sistematicità e professionalità.

Per concludere, tali gestioni non sono "ottimizzate"; di conseguenza, è necessario predisporre strumenti e dispositivi in grado di raggiungere i cosiddetti "massimi simultanei".

1.4 Le principali fonti di finanziamento delle organizzazioni non-profit

Ad eccezione di pochi grandi enti, il Terzo settore riunisce le istituzioni di natura privata operanti nel sistema economico ponendosi tra Stato e Mercato, pur non essendo riconducibili né all'uno né all'altro, e vede la presenza di centinaia di migliaia di organizzazioni di minori dimensioni spesso caratterizzate da strutture organizzative inadeguate, con sistemi contabili e di controllo evidentemente insufficienti.

[11] A. Propersi, *Le aziende non profit. I caratteri, la gestione, il controllo.*, Milano: RCS Libri, 1999, p. 67

La mancanza d'interessi proprietari, che ne indirizzino la gestione e promuovano l'efficienza, comporta il rischio di discontinuità e irregolarità dell'attività. In campo finanziario, pur tenendo in considerazione che il panorama delle associazioni è molto variegato e diversificato, si riscontra in molti casi una debolezza strutturale che si manifesta prevalentemente con bassa capitalizzazione, mancanza di sufficienti e adeguate garanzie e difficoltà di accesso al credito. In base ai dati statistici di settore, difatti, è stato accertato che sono poche le organizzazioni patrimonializzate o con cicli produttivi che consentano un'autosufficienza economica e finanziaria, col risultato che nascono problematiche in termini di ricerca e ottimizzazione delle fonti di finanziamento.

Non avendo di fronte un mercato, gli enti realizzano la propria *mission* secondo norme statutarie e non sono tenuti a esigere un prezzo per i servizi erogati (l'erogazione gratuita rientra proprio nelle loro finalità istituzionali).

Dunque, date le condizioni economiche e finanziarie individuate precedentemente, le aziende non profit presentano carenze di mezzi finanziari provenienti dal proprio ciclo produttivo ed è, pressoché, sempre necessario ricorrere ad altre fonti di sostegno. A questo punto, le donazioni rappresentano la quota maggioritaria delle entrate complessive delle organizzazioni del Terzo settore; queste ultime basano la propria capacità di sopravvivere e svolgere le attività per cui sono state costituite soprattutto sull'abilità nel raccogliere i fondi, sia nei riguardi di soggetti privati sia della pubblica amministrazione.

Dopo aver delineato il finanziamento nelle diverse configurazioni che può assumere, veniamo ad illustrare come lo stesso viene ad essere supportato o alimentato.

1.4.1 Fonti pubbliche di finanziamento

In Europa, l'erogazione privata dei servizi *Welfare* si completa con un massiccio finanziamento pubblico che garantisce l'accesso, pressoché universale, ai servizi stessi. L'entità di tali finanziamenti dipende marginalmente dalle strategie poste in essere dalle aziende non profit; difatti, gli stanziamenti pubblici sono condizionati dai bilanci, dalla necessità di contenere i disavanzi delle singole agenzie pubbliche, e solo secondariamente, da valutazioni di efficacia ed efficienza delle singole organizzazioni non profit.

Il consistente finanziamento pubblico sottolinea la significativa coesione e sinergia fra Stato e Terzo settore, che si manifesta in una vera e propria collaborazione dal punto di vista finanziario, favorendo la crescita e l'evoluzione dei tanti servizi di *Welfare*. "È proprio il compito di braccio operativo della pubblica amministrazione che consente, dunque, alle aziende non profit di svolgere il proprio ruolo redistributivo proprio per quanto riguarda la fornitura di servizi sociali, sanitari ed assistenziali alla generalità della popolazione o a soggetti che non sono in grado di pagare. Il finanziamento pubblico solitamente predilige aziende attive nei settori della sanità, dell'educazione e dei servizi sociali mentre le entrate di fonte privata

convergono in maniera rilevante nell'area della cultura e ricreazione, dell'ambiente e del sostegno allo sviluppo locale"[12].

Rientrano nelle fonti pubbliche di finanziamento i flussi finanziari derivanti dallo Stato o da istituzioni sovranazionali (Unione Europea, Banca mondiale, Unesco, Onu…).

Si distinguono le seguenti tipologie di processo:

a) *Finanziamento pubblico regolato*: l'organizzazione non profit riceve sussidi o contributi direttamente dallo Stato, senza alcun corrispettivo di scambio puntuale in termini di erogazione di servizi o produzione di beni, giacché è sufficiente essere iscritti in un albo a fronte di una verifica *ex ante* di specifiche prerogative a forte valenza burocratica. Questo tipo di finanziamento è prevalentemente a "pioggia", a forte parcellizzazione e basato prevalentemente sulla convenzione e consuetudine, piuttosto che sulla valutazione di risultati;

b) *Attività commerciale sul mercato dei servizi pubblici*: l'azienda non profit riceve un corrispettivo a fronte di servizi erogati o beni prodotti a favore della popolazione su cui insiste "istituzionalmente e geograficamente" un ente pubblico. È frequente che tale attività sia regolata da contratti che scaturiscono dall'espletamento di gare o a fronte di concessioni;

c) *Finanziamento pubblico saltuario*: l'associazione non profit riceve sovvenzioni dall'Unione Europea (o da altre istituzioni internazionali) al fine di potenziare la propria "*mission*" solidale, contribuire a iniziative economiche in aree "depresse", mettendo a disposizione il proprio "know-how" e progettare e gestire attività in paesi in via di sviluppo.

L'esame dei dati della *Charity Commission* sul finanziamento degli enti dimostra che lo Stato è, tuttora, il primo finanziatore del settore non profit.

L'iniziativa diretta degli enti in campo sociale, assistenziale, sanitario, di istruzione ha confermato/dimostrato di essere spesso inefficiente, ma presenta il vantaggio di poter essere effettuata seguendo una programmazione unitaria e razionale ed evitando la frammentazione e la dispersione che gestioni divise e non coordinate possono comportare.

Va, però, estendendosi l'attività d'indirizzo della pubblica amministrazione e di coordinamento delle attività svolte da terzi. Grazie a tali forme d'intervento (*outsourcing* della Pubblica amministrazione) si apre la strada a innovative forme d'indirizzo e coordinamento fra pubblico e settore non profit. Si stanno diffondendo non solo in ambito assistenziale, ma anche culturale, sportivo, artistico… fenomeni di vere e

[12] Barbetta G. P.: <*Il settore nonprofit italiano*>, Studi e Ricerche, il Mulino, 2000, pp. 59-62.

proprie *joint venture* del sociale. In particolar modo, nell'universo non profit, si sono affermati con successo i *social impact bonds*, strumenti finanziari assimilabili ai titoli obbligazionari ed impiegati per lo sviluppo di iniziative all'interno di aree di particolare fragilità sociale. Si afferma così una nuova modalità e impostazione di azione da parte della Pubblica Amministrazione che esercita una politica programmatica ottimale esternalizzata, collaborando attivamente con gli enti del privato sociale ed imponendosi come soggetti attivi, propositivi e partecipi della programmazione.

1.4.2 Aspetti legislativi e misure fiscali

Le più recenti misure d'incentivazione fiscale nei confronti del settore non profit, finalizzate, in particolar modo, all'incremento dei fondi privati come fonte di entrata accessoria strategica rispetto alle erogazioni pubbliche, prendono avvio dal modello giuridico-istituzionale anglosassone della *charity*.

Le organizzazioni senza scopo di lucro, forti di un apposito *charitable status* che garantisce molteplici vantaggi ed agevolazioni fiscali, beneficiano dell'insieme delle misure previste dalle varie *charity tax laws* approvate negli ultimi decenni, destinate ad assicurare le basi per un finanziamento congiunto e coordinato tra pubblico e privato, una vera e propria *joint venture* del sociale.

Le organizzazioni non profit, una volta certificato che l'attività è di pubblica utilità (*public benefit*), sono inserite nell'apposito *Charity Register* amministrato e costantemente aggiornato dalle rispettive commissioni governative (*Charity Commission for England and Wales, Office of the Scottish Charity Regulator, Charity Commission for Northern Ireland*).

Uno dei punti di forza del modello di charity anglosassone sta, pertanto, proprio nell'aver progettato un'unica categoria giuridico-istituzionale, contraddistinta da un corpus coerente, uniforme e unitario di leggi fiscali applicabili a una pluralità d'istituzioni eterogenee per quanto riguarda le aree di attività (sanità, diritti umani, educazione, scienza, tutela del patrimonio, arti e spettacolo ecc., purché rimanga nell'ambito di *public benefit*), lo stato giuridico (*trust, association, foundation, national company* ecc.) e le dimensioni istituzionali (organico, budget annuo, ecc.).

Una fotografia istantanea sulla composizione dei vari bilanci mostra la grande varietà di budget dell'insieme delle charities inglesi alla fine del 2010. Il settore è in continua evoluzione e questo rende necessario un periodico aggiornamento dei dati da parte delle commissioni nazionali.

In Inghilterra e nel Galles, il numero delle charities ufficialmente registrate nel dicembre 2010 era di 150.219 unità, cui si aggiungono 23mila organizzazioni scozzesi e poco più di 5mila istituzioni dell'Irlanda del Nord.

Il modello anglosassone delle charities, con una struttura piramidale e gerarchica che poggia su una cospicua base di realtà medio-piccole con un bilancio non superiore alle £ 10.000, dominata da un ristretto gruppo di associazioni protagoniste il cui bilancio annuo è superiore ai 5 milioni di sterline, grazie alla riuscita uniformazione degli incentivi fiscali al settore, costituisce una sorta di spina dorsale e di presupposto

giuridico-istituzionale per il consistente investimento privato anglosassone, prima crescita dei ricavi istituzionali, quantificabili in 51,7 miliardi di sterline per l'insieme delle organizzazioni appena analizzate.

1.4.3 "Charities and tax"

Gli amministratori delle charities possono richiedere e, quindi, ricevere dallo Stato inglese determinate agevolazioni fiscali (*tax relief*).

Tuttavia, per poter beneficiare di tali incentivi, occorre essere registrati presso *HM Revenue and Customs (HMRC)*, rappresentante l'autorità fiscale e doganale del Regno Unito e responsabile della raccolta e gestione del denaro, che viene reso disponibile sia per finanziare i servizi pubblici (*public services*) sia per assistere gli individui e le famiglie disagiate mediante un sostegno finanziario mirato.

Le charities non versano tasse su alcuna tipologia di entrata giacché utilizzano il denaro per assolvere *charitable purposes.* Tuttavia, nei casi in cui l'ente ottenga un guadagno che dà diritto ad agevolazioni fiscali oppure ha speso tale entrata in *non-charitable purposes*, è probabile che debba pagare le imposte.

Col termine *charitable expenditure* s'identifica quella parte di fatturato e di profitti di una charity che saranno esenti da imposte, ma soltanto se quest'ultimi sono destinati a *charitable purposes*, tra cui donazioni (*donations*), profitti derivanti da scambi commerciali (*profits from trading*), reddito da affitti o investimenti (*rental or investment income*), profitti derivanti dalla vendita o dismissione di un bene (*profits when you sell or 'dispose of' an asset*) e acquisto di proprietà (*buy property*).

Altrimenti, le charities versano aliquote fiscali sui dividendi provenienti da imprese britanniche (*dividends from UK companies*), sui profitti riguardanti lo sviluppo di un terreno o di una proprietà (*profits from developing land or property*), *purchases* ed, infine, è necessario pagare le imposte sulla cosidetta *'non-charitable expenditure'*, ovvero profitti indirizzati a *non-charitable purposes.*

Dopodiché, l'associazione può richiedere la *tax relief* qualora possieda i seguenti requisiti:

a) ha sede in UK, UE, Islanda, Liechtenstein o Norvegia;
b) è istituita solo per *charitable purposes*;
c) è registrata presso la *Charity Commissione* o un altro ente regolatore;
d) è amministrata da *"fit and proper persons"*;
e) è riconosciuta da *HM Revenue and Customs (HMRC).*

Per concludere, è obbligatorio compilare una dichiarazione dei redditi qualora la charity generi guadagni che non presentano i requisiti necessari per poter beneficiare di agevolazioni fiscali oppure una dichiarazione dei redditi annuale per le organizzazioni che realizzino un fatturato superiore alle £ 10.000.

1.4.4 Le fonti di natura privatistica

A tal proposito, bisogna segnalare l'apporto decisivo offerto dai cittadini privati o dalle imprese, che manifestano un palese interesse verso le attività del settore non profit. Tale sostegno si concretizza mediante la partecipazione di quest'ultimi ai costi della loro realizzazione o del loro mantenimento.

Difatti, in virtù delle risorse via via minori che lo Stato può destinare ad attività sociali, si registra una crescente ricerca da parte del mondo non-profit di fonti di finanziamento private integrative a quelle statali.

Un passo in tale direzione è l'introduzione di norme che stabiliscono interessanti agevolazioni fiscali, le cosiddette *tax reliefs*, per favorire e promuovere il Terzo settore, incentivando le donazioni da parte di privati.

Tuttavia, al fine di rendere durevole il rapporto tra organizzazione non profit e il mondo dei *donors* diventa essenziale accrescere la fiducia di quest'ultimi riguardo il corretto impiego dei fondi erogati. Gli enti non profit hanno finalmente preso coscienza del fatto che "da una buona raccolta fondi dipende non solo il destino di una campagna ma anche quello della stessa organizzazione; e che, affinché si possa realizzare una buona raccolta fondi, risulta necessario dissipare ogni dubbio circa il trasparente uso dei soldi raccolti. D'altro canto, i fund raiser avvertono che, in un ambito di attività così complesso e delicato, la buona reputazione è tutto"[13].

La trasparenza, dunque, è fondamentale per assicurare la pubblica fede, e, in particolare, rendere l'informazione veritiera e completa al fine di tutelare i donatori, attuali e potenziali, circa il corretto impiego dei mezzi raccolti.

Nonostante tali accorgimenti, non è sufficiente che vigili un'Authority sul Terzo settore (nel Regno Unito tale ruolo è svolto da *HM Revenue and Customs (HMRC)*), avente il compito di garantire l'uniforme e corretta applicazione della normativa fiscale e tutelare gli utenti da abusi compiuti dagli enti che si adoperano nella raccolta di fondi. È necessaria, anche, la presenza di revisori professionisti indipendenti che vigilino sul rispetto delle norme statutarie, delle leggi fiscali e sulla correttezza della gestione.

A tal fine, sono state adottate specifiche misure di regolamentazione quali, i documenti di rendicontazione (bilancio di esercizio e bilancio sociale), la presenza di un controllo esterno sull'ente (il cosiddetto revisore sociale, in Inghilterra ruolo svolto dalla *Charity Commission*), l'adozione di codici di autoregolamentazione, che assicurino trasparenza nella raccolta di fondi.

Il panorama delle iniziative volte a consolidare le attività e le strutture delle organizzazioni non profit è molto ampio e variegato: ciò è indubbiamente sintomo di una vitalità del settore e prelude a prospettive rosee di sviluppo dello stesso.

---

[13] Barbetta, G. P. & Maggio, F. (2008). *Nonprofit.* Bologna: Il Mulino, p.97

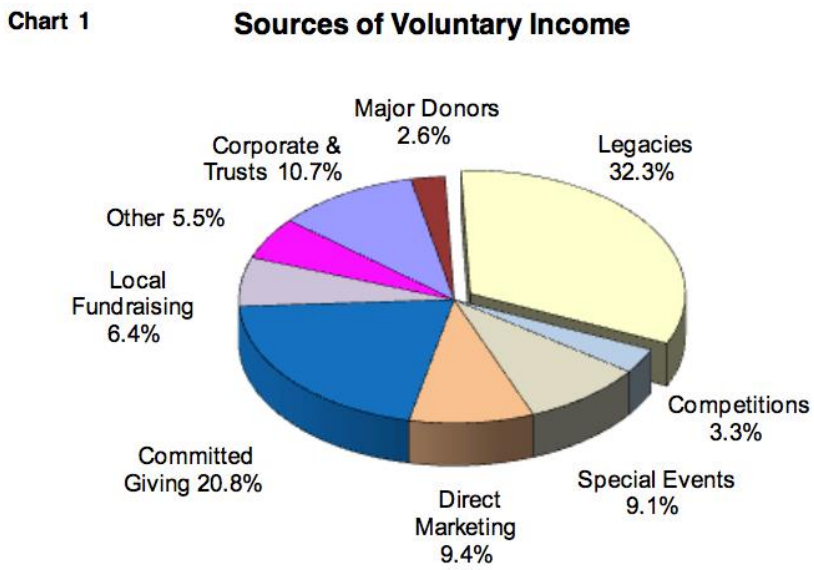


Fig. 1.3 Fonti di guadagno del Terzo settore in UK

1.4.5 Le principali fonti private di finanziamento

Per quanto riguarda le fonti di finanziamento di origine privatistica, provenienti da singoli individui, si distinguono in:

1. *Trading*: privati cittadini o esattamente "*clients"* offrono un corrispettivo per la fruizione di un servizio dell'organizzazione non profit o per l'acquisizione di un bene. La relazione fra "mercato" e scambio avviene sulla base prevalente dell'utilità reciproca, a meno che non esista una pregiudiziale positiva di altruismo e senso di solidarietà, che dà una valenza meno economicista o utilitarista dello scambio stesso in una logica di "equilibrio economico-finanziario" più che di "profitto" da raggiungere "*ad libitum*". È importante ribadire che il finanziamento derivante dagli introiti generati dalla vendita di beni e servizi, mette in evidenza le abilità o capacità intrinseche delle aziende non profit di produrre e offrire beni e servizi di qualità elevata, o almeno sufficiente a soddisfare la domanda pagante. La diffusione di questo sistema di finanziamento e il peso crescente che sta indiscutibilmente assumendo all'interno del complesso di risorse economiche delle aziende non profit, è testimoniato, in particolar modo, nei paesi anglofoni, dalla nascita dei *charity shops*, negozi dell'usato senza fini di lucro, che assolvono a compiti di *fundraising*, rafforzano la conoscenza del marchio, comunicano *mission* e progetti *charitable* a pubblici diversi, creano opportunità di volontariato e diffondono la cultura del dono e del riciclo.

2. *Financial intermediation*: membri dell'associazione non profit o *insiders*, che creano e autofinanziano la stessa, in base a una logica di copertura dei costi dell'attività posta in essere (*self financing*) e di autoconsumo dei servizi erogati. Questa tipologia di aziende non profit sono notoriamente autogestite e si configurano nel ruolo di supplenza privata rispetto alla carenza di interventi pubblici.

3. *Fund raising*: benefattori o "*donors*" devolvono somme di denaro (*donations*) ad una organizzazione, in base all'appartenenza a tipologie di sostentamento istituzionalizzato (*member, supporter, junior, senior* ecc…). Le "quote associative" o i "contributi" sono una forma piuttosto convenzionale di finanziamento, che richiede interventi mirati a rafforzare il clima motivazionale dell'organizzazione e il senso di appartenenza dei singoli membri.

Il fund raising si colloca trasversalmente alle funzioni tradizionali della gestione aziendale (marketing, controllo e finanza); è difatti una materia interdisciplinare che deve considerare e coniugare aspetti di carattere economico giuridico e sociale. In particolare, quest'ultimi sono diretti:

- *alla sollecitazione del bisogno di donare*;
- *alla capacità di utilizzare efficacemente ed efficientemente i mezzi disponibili per il conseguimento di risultati socialmente utili (persuasione circa la meritorietà dell'attività).*

È necessario ideare e predisporre un messaggio adeguato alle caratteristiche dei gruppi omogenei di donatori, tenendo presenti la diversa capacità contributiva e le motivazioni che spingeranno gli individui a donare.

La raccolta di risorse finanziarie ha assunto una funzione e un'importanza crescenti nel Terzo settore, ormai integrato e indispensabile alla struttura sociale di molti Stati.

Il fund raising, difatti, primeggia nei settori orientati alla tutela dei diritti, come nel caso delle organizzazioni ambientaliste, delle organizzazioni di tutela dei diritti umani, club ecc… Le *donations* giocano un ruolo fondamentale nell'area dell'aiuto economico internazionale e nelle organizzazioni d'intermediazione filantropica, come ad esempio le fondazioni di erogazione o *grant-making*.

Esse sono da considerarsi, di massima, come flussi di cassa della gestione corrente.

Questo tipo di supporto può essere:

- *consuntivo*: quando la contribuzione finanziaria è collegata alla fondazione della associazione non profit e struttura il capitale iniziale per la fase dello "start-up";

- *gestionale*: quando gli interventi donativi, i lasciti, i vitalizi hanno luogo proprio "durante" le fasi del ciclo di vita dell'organizzazione non profit;
- *individui volontari*: essi mettono il loro tempo al servizio dell'associazione non profit, finanziando indirettamente quest'ultima.

  In questo modo, essa è in grado di sostenere i costi di produzione di beni e l'erogazione servizi ove non esistano assetti finanziari sufficienti per coprire tali costi oppure ove s'integrano individui altamente qualificati o, al contrario, con bassa produttività e scarsa esperienza, in aziende non profit aventi strutture rigide di costo.

  Qualora non vi fosse l'ausilio e integrazione dei volontari, l'ente non sarebbe in grado di assicurare la copertura economica dei costi delle attività promosse, registrando di conseguenza saldi negativi. I volontari, dunque, partecipano e contribuiscono alla formazione e alla crescita dell'attività in una dimensione reddituale che dovrebbe tendere al pareggio o all'utile di gestione;
- *social impact bonds*: nell'universo non profit si è affacciata una nuova forma di *fund raising*, che viene a consolidare e completare l'ambito variegato complesso di strumenti di cui codeste aziende usufruiscono, per l'acquisizione di risorse finanziarie, ossia i *social impact bonds*. Tali strumenti finanziari sono assimilabili ai titoli obbligazionari e impiegati per lo sviluppo di iniziative all'interno di aree di particolare fragilità sociale. In tal contesto, un ente pubblico, segnatamente il governo, si impegna, a fronte del raggiungimento di un "risultato", a supportare economicamente la realizzazione di un progetto di interesse generale; affinché l'iniziativa possa essere finanziata, verranno emessi dei *bonds*, sottoscritti dai soggetti interessati a supportarne la realizzazione. Una volta raggiunto il "risultato" stabilito e conclusasi, quindi, con successo l'iniziativa, il governo erogherà le risorse necessarie a ripagare gli investitori e che deriveranno principalmente dal risparmio economico ottenuto dalla realizzazione del progetto.

La raccolta di fondi s'intraprende spesso, mediante strumenti di *marketing diretto*, inteso come un sistema di marketing interattivo che utilizza uno o più "media" (pubblicitari in senso classico, giornali, televisione, radio, contatto postale o telefonico) per ottenere risposte da un target o segmento specifico. Difatti, grazie ad una vera e propria analisi di mercato articolata e approfondita, è possibile operare la segmentazione dei diversi destinatari della comunicazione in categorie piuttosto omogenee.

L'attività di marketing interattivo è misurabile attraverso semplici indicatori o strumenti, quali:

a) *redemption*: intesa come rapporto fra il numero delle risposte ottenute e il numero dei contatti (informazione sui volumi dell'attività di contatto);
b) *costo per risposta*: ove s'indica il rapporto fra l'investimento pubblicitario e il numero delle risposte ottenute, calcolato moltiplicando il costo/contatto per il numero dei contatti rapportato al numero delle risposte;

c) *redditività degli investimenti pubblicitari sostenuti*: si esprime il rapporto fra finanziamenti/ donazioni offerte e investimenti pubblicitari effettuati a vario titolo. Questo indicatore consente di scegliere l'azione che produrrà la migliore redditività dell'investimento pubblicitario sostenuto.

Il marketing diretto attiva un comportamento, una risposta e una reazione a fronte della sollecitazione effettuata, tramite vari strumenti di contatto, tra cui i più diffusi sono il contatto postale e telefonico. L'obiettivo a cui mirano tali strumenti di *marketing diretto* è la strutturazione di una banca dati aggiornata e, dinamicamente aggiornabile, al fine di istituire e consolidare un circuito virtuoso fra le organizzazioni non profit e i potenziali *donors*.

Il fund raising rappresenta una attività cruciale in un ente non profit, in virtù della diminuzione delle entrate provenienti dal settore pubblico che ha determinato una ridefinizione della struttura organizzativa e l'emergere di nuove professionalità.

L'attività di raccolta fondi, pur avendo punti di contatto con la gestione caratteristica aziendale, si inquadra nell'ambito della gestione finanziaria, la quale, notoriamente, è rivolta/mira a coprire il fabbisogno finanziario dell'istituto.

Lo scopo principale di una strategia di fund raising è la realizzazione della propria *mission*, pertanto, la raccolta fondi costituisce un mezzo e non un fine proprio dell'organizzazione; gli ideali, dunque, sono preminenti sulla gestione finanziaria.

Nel definire una strategia di raccolta fondi, l'azienda *non profit* sarà, in primo luogo, impegnata ad individuare i segmenti a cui si vuole rivolgere, ovvero se si vuole che contribuiscano singoli privati, società a scopo di lucro o istituzioni governative e, in secondo luogo, selezionare le strategie adeguate al tipo di interlocutore scelto. La scelta non deve essere solo guidata da criteri economici (scegliere il segmento che garantisce maggior finanziamento), ma anche da un giudizio di opportunità e di senso di appartenenza e condivisione degli ideali dell'associazione da parte dei potenziali finanziatori, assicurando, in tal modo, un rapporto stabile e duraturo fra l'organizzazione e i *potential donors*.

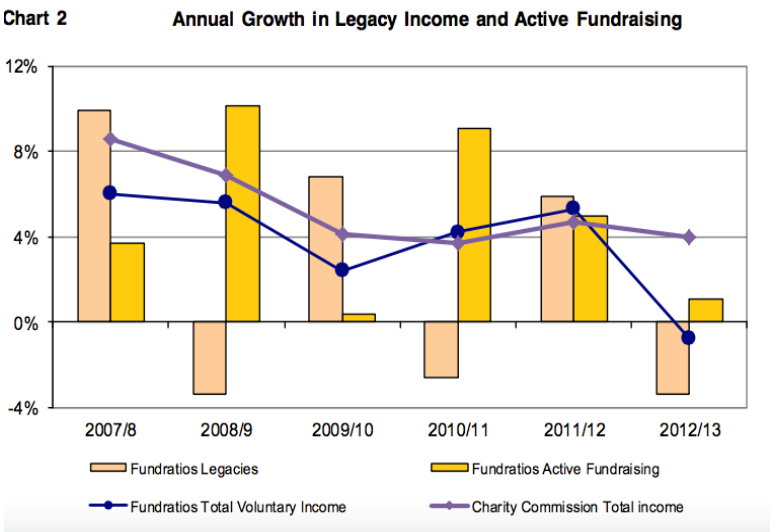


Fig. 1.4 Crescita annuale dei legati e del fund raising attivo

Ora, le strategie di *fund raising* possono essere raggruppate in quattro classi fondamentali:

a) *diretto generico*: quando il donatore decide di finanziare direttamente l'attività dell'ente;
b) *diretto su progetto*: quando il donatore finanzia direttamente in prima persona un determinato progetto;
c) *indiretto su progetto*: quando s'interpone un terzo soggetto che lega i *donors* e il "*deed of gift"* a un progetto specifico;
d) *indiretto generico*: quando il donatore si avvale di un soggetto terzo per devolvere indistintamente il proprio denaro.

La prima strategia (*diretto generico*) implica un rapporto fiduciario fra le parti, dipeso, spesso e volentieri, dall'immagine globale dell'azienda non-profit; tuttavia, in virtù del fatto che il donatore avverte maggiormente l'esigenza di avere un riscontro effettivo delle modalità d'impiego del proprio denaro, la seconda strategia/tecnica (*diretto su progetto*) si presta a soddisfare la necessità di una chiara correlazione fra fonte finanziaria ed il suo impiego.

In seguito, la terza strategia (*indiretto su progetto*) implica necessariamente un'elevata conoscenza e visibilità del soggetto terzo (istituti di credito e rete commerciale) e dell'organizzazione non profit; la raccolta fondi sarà fortemente condizionata dall'immagine dei partner e dall'attrattività del progetto da realizzare.

Infine, l'ultima tattica (*indiretto generico*) richiede un legame duraturo e un imprescindibile coinvolgimento di entrambi i partner.

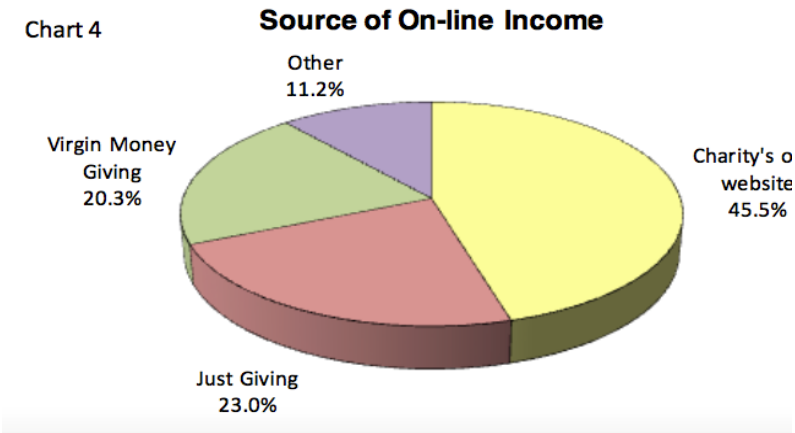


Fig. 1.5 Fonti di guadagno degli enti non profit provenienti dal Web

Un'ultima osservazione va riservata al ruolo del *Web*, che ha mutato radicalmente i metodi di *fund raising* convenzionali e si è imposto come canale supplementare di raccolta di fondi, in grado di raggiungere un target maggiore di potenziali donatori e di assicurare trasparenza riguardo i progetti o le iniziative da finanziare.

1.5 Le Joint Ventures

Il termine *joint venture* può descrivere una gamma di differenti accordi commerciali fra due o più entità distinte. Ogni parte, convogliando risorse alla *venture*, dà vita ad un nuovo *business*, in cui ogni membro collabora in maniera congiunta e sinergica e condivide i rischi e i benefici che la *venture* comporta. Ciascun socio può fornire un terreno, un capitale, proprietà intellettuale, staff altamente qualificato, equipaggiamento o qualsiasi altra forma di asset. Ciascuno possiede competenze o bisogni che sono centrali per lo sviluppo ed il successo del *business*, purché ogni membro condivida una *shared vison* sugli obiettivi, che la *joint venture* dovrà promuovere.

1.5.1 Joint Ventures tra le organizzazioni non profit e il settore pubblico

È importante distinguere la formazione di una vera e propria *JV* da meri accordi contrattuali, come, ad esempio, i contratti per la fornitura di beni, servizi o concessioni per cui lo Stato concede ad una terza parte (il *concessionaire*) il diritto di erogare servizi al pubblico a fronte di un ritorno economico. Le *Joint Ventures* sono istituite fra istituzioni che aspirano ad obiettivi complementari e condividono natura e propositi delle attività della *JV*. Al contrario, se il settore pubblico desidera concludere accordi che stabiliscono chiaramente alcuni specifici obiettivi e laddove non vi sia né alcuna prospettiva di crescita e diversificazione e né sia richiesto *risk sharing*, gli obiettivi perseguiti dallo Stato potranno essere realizzati grazie ad una procedura contrattuale più chiara e diretta.

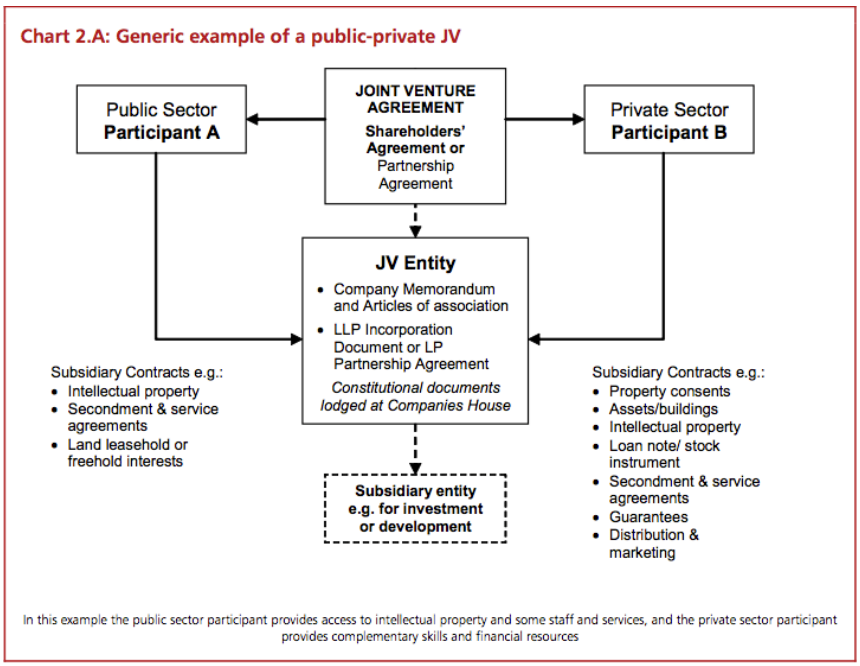
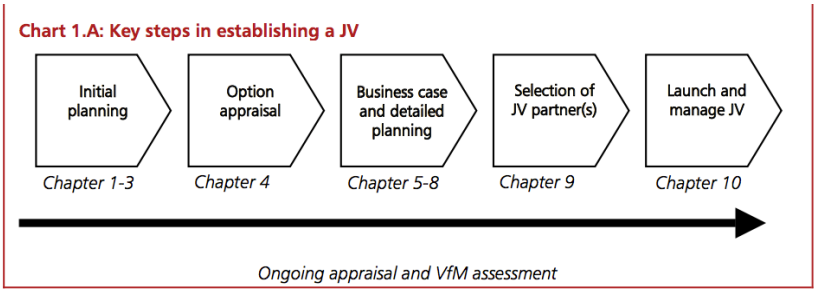


Fig. 1.6 Schema generale di una Joint Venture tra il settore pubblico e privato



I tipici *key steps* per la costituzione di una *Joint Venture* sono:

1. *initial planning*: è essenziale che il settore pubblico intraprenda un'analisi di valutazione della proposta di *JV* e che dimostri di essere un *business* credibile oppure eroghi servizi commerciali d'indubbia affidabilità;
2. *option appraisal*: lo Stato dovrà considerare le *Value for Money* issues, ossia "la combinazione ottima fra costi del capitale e di gestione e qualità dei beni e servizi offerti che rispetta i requisiti stabiliti per gli investimenti pubblici" (nota), e condurre una stima puntuale del *VfM* in conformità con le direttive del *HM Treasury* e dell'*OGC*. È opportuno che il settore pubblico effettui una valutazione appropriata dell'investimento ed uno studio che consideri la praticabilità e fattibilità di altri potenziali modelli di consegna (*delivery models*), come ad esempio le *concessions*, il *contractual service/supply contracts* ed il *PFI* (*Private Finance Initiative*) al fine di poter determinare se una *JV* è il modello di consegna che fornirà il miglior *VfM* e *long-term benefits* al settore pubblico. La valutazione del *VfM* e la stima dei principi che promuovono la *JV* dovranno attenersi ad un *business case approach*. Ciò comporta un processo simulato in cui, affinché vengano incrementati i livelli di garanzia, sono proposti casi urgenti e coerenti con l'azione proposta. Nelle fasi iniziali, sono definiti sia un *Strategic Outline Case* (*SOC*) sia un *Outline Business Case* (*OBC*), al fine di valutare se la *JV*, confrontata con altre opzioni di business, dimostra un potenziale appropriato, desiderabile e praticabile, riguardo i servizi di fornitura (*means of delivery*) compatibili con i risultati richiesti. Dopo ulteriori lavori ed un coinvolgimento formale dei partners possibili, il processo dovrebbe concludersi con un *Full Business Case* (*FBC*) presentato e, infine, approvato mediante idonei *governance arrangements* prima della formazione della *JV*. I capitoli più scottanti del *FBC* riguardano sia le linee guida del potenziale *VfM* e l'analisi e spiegazione dei benefits, degli inconvenienti e dei rischi di una *JV*, basandosi sul confronto di altre opzioni di consegna (*delivery options*);

3. *business case and detailed planning*: il governo dovrà discutere, con il dipartimento preposto a finanziare la *Joint Venture* (*HM Treasury*), qualsiasi proposta controversa o *JVs*, che vada oltre i limiti di approvazione delegati (*delegated approval limits*). La metodologia di valutazione quantitativa per valutare il *PFI*, rispetto al metodo di consegna del settore pubblico (*public sector*

delivery), utilizza un *public sector comparator* come riferimemto per stimare i rischi e scontare i costi e le entrate future mediante il *NPV approach*. È decisivo stabilire sia il *Joint Venture Agreement* ed i tipi di *equity* participation, affinché assicurino che lo schema organizzativo e la struttura decisionale della *JV* perseguano e realizzino gli obiettivi del settore pubblico ed individuino una soluzione commerciale desiderabile per l'apparato statale. Un partner della *JV* potrà dichiarare la propria intenzione di finanziare la *JV*, ma non vi sarà alcun impegno vincolante di donare fondi o assets da parte del partecipante alla *JV*, a meno che non lo richieda l'accordo. Il *JV agreement* è un contratto e, in quanto tale, è retto dalle leggi ordinarie del contratto. Durante il processo di selezione dei soci, il settore pubblico e i suoi *advisors* dovrebbero preparare sia una sintesi dei punti chiave del *JV agreement* o una bozza del *JV agreement* da diffondere ai potenziali partecipanti. Infine, i tipi di *equity participation*, per mezzo dei quali è possibile finanziare la *JV*, riguardano essenzialmente la sottoscrizione di un *ordinary equity shares* oppure di una *partnership capital*, che può essere suddivisa in classi diverse e con diritti differenti, in considerazione della distribuzione dei profitti e del capitale fra i soci;

4. *selezione dei JV partner(s)*: il settore pubblico dovrebbe normalmente unirsi ad un *JV* partner competitivo, sebbene potrebbero esservi alcune eccezioni, laddove sia preminente la giurisdizione dell'Unione Europea (*EU public procurement*);
5. *launch and manage* di una *JV*: infine, le *issues* concernenti l'avvio della *JV* e l'*on-going management* della partecipazione del settore pubblico alla *JV*. Il settore pubblico e privato partecipanti alla *JV* devono assolutamente decidere in che modo la *JV* vada gestita. La *JV* deve rendere conto ai partecipanti e, su alcune *issues*, occorrerà l'approvazione di quest'ultimi.

Tali questioni vengono spesso chiamate "*reserved matters*" o "*veto rights*" ed includono:

- approvazione dei business plans, dei budgets, dei contratti rilevanti e di qualsiasi difformità della *JV* da quei documenti;
- cambiamenti nella politica di distribuzione;
- introduzione di nuovi fondi e partecipanti;
- diritti di veto (*veto rights*), concernenti la nomina delle figure chiave della *JV*;
- mutamenti dei documenti costituzionali;
- cessazione o vendita degli *assets* della *JV*.

È opportuno considerare gli accordi di *governance*, che svolgono funzioni di controllo e protezione della *JV*, in modo particolare quando essa è almeno in parte fondata o altrimenti supportata dalle finanze pubbliche. Tali accordi dovrebbero minimizzare il rischio di un conflitto d'interesse e rassicurare gli *stakeholders* del settore pubblico, riguardo il senso di correttezza (*propriety*) degli accordi. La responsabilità per la supervisione e la gestione della *JV* e i suoi *business lies* con il Board della *JV* viene esclusa per quelle

materie che la legge sulle aziende anglosassoni pretende che venga decisa dai soci della *JV*. È necessario che sia assunta una decisione, riguardo la possibilità, per il Board, di essere coinvolto attivamente alle decisioni manageriali della *JV* oppure assumere un ruolo strategico e di sorveglianza. Il *chairman*, cardine su cui ruota la gestione della *JV*, dovrebbe essere selezionato sulla base della sua esperienza manageriale ed amministrativa, conoscenza del business, del mercato ad esso associato, leadership e il gradimento da parte dei soci, esercitando un giudizio indipendente e agendo in buona fede così da promuovere il successo della *JV*.

1.5.2 Joint Ventures tra le organizzazioni non profit e le imprese private

Le imprese private, secondo una classificazione proposta da un autorevole studioso, supportano le associazioni non profit tramite:

1. *Sponsorizzazioni*: intese come attività di sviluppo d'immagine e di notorietà dell'organizzazione non-profit, unitamente a donazioni non specifiche e non correlate a progetti, che sostengono l'attività della stessa. Si tratta di finanziamenti generalizzati che "ostentano" il ruolo di "fiancheggiamento" e di condivisione della "causa di solidarietà" svolta dall'impresa;

2. *Royalties*: sono legate all'ammontare di beni venduti sul mercato con il marchio dell'ente non profit. I prodotti possono essere di consumo o strumenti finanziari. In prospettiva, è probabile che abbia luogo un'estensione della concessione fra i marchi anche di beni durevoli. Il ritorno d'immagine, notorietà e di vendite è misurabile, mediante le rilevazioni di quote di mercato d'una o più linee di prodotti/ servizi mantenute o aumentate, sulle quali sono state computate le royalties;

3. *Donazioni finanziarie defiscalizzabili* menzionate nel paragrafo 1.4.3;

4. *Campagne di promozione*: in cui si sollecita l'opinione pubblica, in particolar modo i clienti dell'impresa, a donare quote di denaro a favore dell'organizzazione non-profit o percentuali sugli incassi (prevalentemente nel settore commerciale) in determinati periodi concordati fra l'azienda e l'associazione;

5. *Banche e intermediari finanziari*: erogano, per statuto, quote di finanziamento a favore degli enti non profit;

6. *Volontariato d'impresa*: in cui alcuni settori gestionali dell'impresa (amministrazione & controllo, marketing, organizzazione, logistica) contribuiscono allo sviluppo delle attività dell'organizzazione non profit strutturando, in questo modo, un finanziamento "indiretto";

7. *Gaming e Lotteries*: un innovativo metodo di *fund raising* è anche quello di collegare il fine solidaristico a giochi e lotterie nazionali o locali. In tal modo la raccolta dei fondi è ulteriormente incentivata, in quanto i donatori (benefattori) saranno attratti dalla prospettiva di vincere un montepremi considerevole e, al contempo, di aiutare il prossimo.

CAPITOLO 2

2. *Knowledge Discovery in Database*

Dalla metà degli anni Novanta in poi, sono stati raccolti e accumulati, a un ritmo impressionante, una moltitudine di dati, in seno ad un'ampia varietà di discipline. V'è, quindi, la necessità di elaborare nuove teorie computazionali e strumenti statistici per assistere gli individui nel processo di selezione delle informazioni utili da una moltitudine di dati digitali.

Tali teorie e strumenti sono il tema di un settore emergente quale è il *Knowledge Discovery in Database* (KDD).

A livello teorico, il campo di ricerca del KDD si concentra sullo sviluppo di metodologie e tecniche per ottenere dati che siano comprensibili dagli utenti. Il problema basilare, che il processo tenta di risolvere, riguarda la classificazione dei cosiddetti *low-level data* (i quali sono tipicamente troppo voluminosi da essere facilmente compresi ed assimilati) in altri formati che dovrebbero essere di dimensioni ridotte (ad esempio, un breve report), di carattere generale (per esempio, un'approssimazione descrittiva o un modello del processo che ha generato i dati) e di utilità pratica (ad esempio, un modello predittivo per stimare il valore di casi futuri). Il nucleo fondamentale del processo di KDD consiste nell'applicazione di specifiche metodologie di *data mining* per la scoperta, la selezione o l'estrazione di *pattern*.

Ora, nasce l'esigenza di incrementare le capacità analitiche degli individui, al fine di poter gestire un numero esteso di *bytes,* che potremo raccogliere sia in campo economico sia scientifico.

In ambito aziendale, l'impiego dei dati consente di guadagnare competitività, essere più efficienti, e fornire servizi maggiormente apprezzabili dai consumatori. I dati che riusciamo a cogliere sul mondo in cui viviamo dimostra che essi sono impiegati per costruire teorie e modelli di carattere scientifico. Grazie ai computer, l'uomo è stato in grado di accumulare più dati di quanti riusciamo ad assimilare ed è evidente che il passaggio alle tecniche computazionali ci aiuta a scoprire *pattern* e strutture significative dall'imponente volume di dati.

Dunque, il KDD è un tentativo di affrontare un problema quotidiano che l'era dell'informazione digitale è riuscita a far proprio: il sovraccarico di dati.

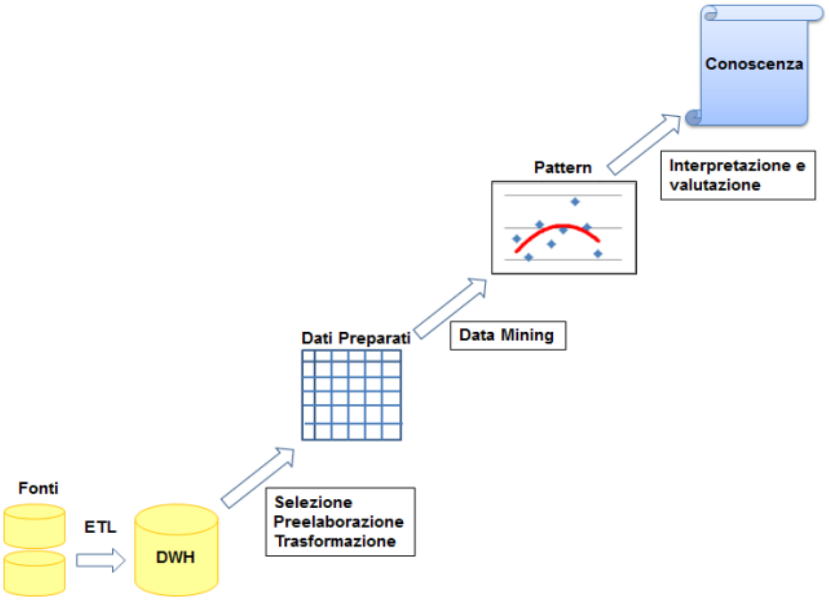


Fig. 2.1 Processo KDD

2.1 Il Data Mining e il Knowledge Discovery in Database

Storicamente, sono stati coniati termini diversi per indicare la ricerca, nei dati, di *patterns* preziosi per la ricerca scientifica, come ad esempio "*data mining*", "*knowledge extraction*", "*information discovery*", "*information harvesting*", "*data archaeology*" e "*data pattern processing*".

Il termine "*data mining*" è comunemente utilizzato da statistici, da analisti dei dati e dalle comunità di *management information system* (MIS).

L'espressione *Knowledge Discovery in Database* fu coniata nel 1989 durante il primo workshop sul KDD al fine di evidenziare che la conoscenza o le informazioni (*knowledge*) sono il prodotto finale di una scoperta guidata di dati.

Secondo Usama Fayyad, Piatetsky-Shapiro e Smyth, il KDD si riferisce all'intero processo di scoperta di conoscenza utile partendo dai dati, mentre il *data mining* è solamente un passaggio particolare di questo processo. Il *Data Mining* è l'applicazione di determinati algoritmi per l'estrazione di *pattern*, partendo dai dati.

Nel processo di KDD, vi sono fasi ulteriori quali la preparazione dei dati, la selezione dei dati, la pulizia dei dati, la valutazione e l'interpretazione dei risultati conseguiti, che costituiscono passaggi essenziali per assicurare che le informazioni ottenute derivino dai dati.

2.1.1 La natura interdisciplinare del Knowledge Discovery in Database

Il KDD è un processo in continua evoluzione, grazie all'interazione e compenetrazione di differenti campi di ricerca scientifica, come ad esempio il *machine learning*, l'identificazione di un pattern, i database, le statistiche, l'intelligenza artificiale (AI), l'acquisizione di informazioni per *expert systems*, la visualizzazione di dati e la performance di calcolo di dati. L'obiettivo comune di tali campi di ricerca consiste nell'estrarre informazioni di elevata qualità (*high-level knowledge*), partendo da dati di bassa qualità (*low-level data*), nel contesto di un vasto insieme di dati.

La fase di *data mining* del processo KDD dipende indissolubilmente da tecniche sperimentate e raffinate dal *machine learning*, dall'identificazione di *patterns* e dalla statistica che si occupa, in tal caso, della ricerca di *patterns*, partendo dai dati a disposizione degli esperti e analisti.

Una domanda che sorge spontanea è come sia possibile che il KDD differisca dall'identificazione di *patterns* o dal *machine learning*? La risposta risiede nel fatto che questi ambiti di ricerca scientifica forniscono alcune metodologie di *data mining* che sono successivamente impiegate nella fase di data mining del processo di KDD.

Il KDD si focalizza sull'intero processo di scoperta di conoscenza, partendo dai dati e includendo anche, come i dati siano immagazzinati e aperti, in che modo gli algoritmi possano essere estesi ad insiemi di dati su larga scala, pur mantenendo un livello ottimale di efficienza tecnico-produttiva, e come l'interazione globale uomo-macchina possa essere modellata e supportata proficuamente.

Il processo di KDD può essere immaginato come un'attività multidisciplinare che riunisce tecniche di analisi presenti in diverse discipline, ponendo una particolare enfasi sulla ricerca di *patterns* facilmente comprensibili, e, per tale motivo, poter essere considerati come informazioni utili o quantomeno interessanti.

La statistica riveste un ruolo fondamentale nel corso del complesso e delicato processo di scoperta d'informazioni partendo dai dati. Essa fornisce un linguaggio e un quadro teorico complessivo per quantificare l'incertezza, risultante dal tentativo di inferire *patterns* generali da un singolo campione di un'intera popolazione. Ora, è chiaro che il KDD fornisce gli strumenti necessari per automatizzare l'intero processo di analisi dei dati e il lavoro statistico di selezione d'ipotesi.

In secondo luogo, i *database* costituiscono la forza motrice del processo di KDD.

Difatti, la questione riguardante l'effettiva manipolazione dei dati, sebbene quest'ultimi non possano entrare nella memoria principale, è di fondamentale importanza per il KDD. Le tecniche di database per realizzare efficientemente l'accesso ai dati, raggruppare, ordinare le operazioni di accesso ai dati e ottimizzare i quesiti (*queries*) posti nel corso della ricerca, costituiscono le fondamenta per estendere gli algoritmi ad una più ampia sequenza di dati.

In generale, gli algoritmi di data mining, elaborati dalla statistica, dall'identificazione di pattern e dal machine learning assumono che i dati siano nella memoria principale e non prendano per nulla in considerazione il modo in cui l'algoritmo collassi, nel caso sia possibile soltanto una visione limitata dei dati.

Infine, il *data warehousing* è un filone di ricerca che ha avuto origine dai database e si riferisce ad una procedura di business, volta alla raccolta e pulizia di dati (di carattere economico o transattivo) preziosi per l'analisi on-line e il supporto alle decisioni aziendali. Nel corso del processo di KDD, il *data warehousing* contribuisce in maniera decisiva alla pulizia dei dati ed all'accesso ai dati.

2.1.2 L'evoluzione del Data Mining

Dal punto di vista logico e matematico, il *data mining* rappresenta un'area scientifica di recente formazione, che è sorta traendo essenzialmente spunto da quanto sviluppato in altri ambiti disciplinari, quali l'informatica, il marketing e la statistica.

In particolar modo, le basilari metodologie, applicate nel data mining, sono sorte principalmente da due comunità o filoni di ricerca accademica: la prima, squisitamente informatica, inerente al tema dell'apprendimento automatico (*machine learning*), la seconda dalla comunità degli statistici, i cui interessi s'incentravano soprattutto su temi di analisi dei dati e di statistica computazionale.

L'innovazione introdotta dal *data mining* è l'integrazione delle precedenti metodologie con gli aspetti applicativi o pratici.

Sebbene le tecniche di *data mining* ritaglino un ruolo rilevante in diversi ambiti applicativi (dalla genomica alla fisica nucleare, per citare i più noti), il più importante di essi, da un punto di vista storico, è l'ambito economico-aziendale. In tal contesto, l'attività di data mining deve elaborare risultati attendibili, interpretabili ed utilizzabili direttamente come supporto alle decisioni aziendali.

L'apprendimento automatico (*machine learning*) è un ambito di ricerca scientifica strettamente legato all'intelligenza artificiale, che si prefigge di estrapolare dai dati relazionali regolarità, alle quali poter fornire, in una prima fase, una valenza generale. Successivamente, l'obiettivo dell'apprendimento automatico è la riproduzione dei processi generatori di dati. Ciò consente la generalizzazione di quanto analizzato, al fine di prevedere l'andamento di determinate variabili in corrispondenza di casi non osservati. Il primo prototipo di *machine learning* venne proposto da Rosenblatt (1962) e chiamato "*percettrone multistrato*". In seguito, nella seconda metà degli anni '80, ebbero origine, da questo filone gli studi, le "*reti neurali*" e, poco più tardi, le *supporter vector machines* (Vapnik, 1998).

Nello stesso periodo, alcuni studiosi, sia di taglio informatico che statistico (Breiman *et al.*, 1984) perfezionarono la teoria degli *alberi decisionali*, applicata prevalentemente per problemi di classificazione e, in seguito, di regressione. La ricerca inerente agli alberi decisionali ha prodotto notevoli sviluppi, in particolar modo concernenti l'impiego di metodi di ricampionamento al fine di migliorare le capacità predittive dei modelli.

La statistica si è sempre dedicata alla costruzione di metodi e modelli per l'analisi dei dati, da un punto di vista puramente teorico; tuttavia, la comunità scientifica si è interrogata sugli aspetti computazionali inerenti all'applicazione della metodologia.

Dalla seconda metà degli anni '80, data la crescente importanza assunta da tali aspetti computazionali, ha avuto luogo un apprezzabile progressione della ricerca scientifica in questo ambito, conducendo inevitabilmente alla nascita di nuove metodologie statistiche fortemente "contaminate" da aspetti computazionali. Si pensi, ad esempio, agli studi accademici riguardanti il "*bootstrap*", i metodi di "*Monte Carlo*", basati sulle "*catene di Markov*", e i sistemi esperti probabilistici; parimenti, ha avuto luogo un crescente interesse della comunità statistica ai metodi di apprendimento automatico.

Un ultimo importante tassello, nell'evoluzione storica del data mining, è da attribuirsi alla seconda metà degli anni '90. Difatti, la rapida diffusione delle tecnologie informatiche ha costituito una chiave fondamentale per la rapida crescita della quantità di dati e informazioni disponibili, (spesso in formati non convenzionali). Si pensi, ad esempio, ai dati prodotti dalla navigazione in internet ("*web data*"); ai dati testuali ("*text data*") o ai dati multimediali ("*media data*"). I risultati di ricerca originali e innovativi di questi filoni di ricerca sono difficilmente classificabili nelle aree scientifiche tradizionali, ma facilmente ascrivibili al contesto del data mining.

Quanto finora premesso evidenzia che il data mining è una disciplina di recente sviluppo e, tuttora, in rapida espansione ed evoluzione.

2.1.3 Cos'è il Data Mining?

Il *data mining* è un nuovo termine il cui significato è da ricondursi ad un antico processo d'individuazione di *patterns* (schemi o modelli) nei dati.

Tipicamente, il compito di un "*data miner*" o "*data mining analyst*" è la risoluzione di un problema, solitamente di carattere aziendale, mediante i dati che si hanno a disposizione (*database*).

L'analista ha bisogno d'individuare i *patterns* nei dati, affinché possano essere utilizzati per risolvere il problema; successivamente, i *patterns* dovranno essere presentati, in modo tale da consentire un agevole interpretazione da parte dell'azienda ed evitare prassi o atteggiamenti giudicati sgraditi da parte dei loro clienti (*customers*).

Prima che diventasse popolare il *data mining*, numerosi filoni di ricerca accademica stavano lavorando con le medesime tecnologie, e, dunque, era estremamente difficile preservare la tipicità dei risultati ottenuti nei differenti ambiti disciplinari. Quindi, era inevitabile che i ricercatori, in campo statistico, di apprendimento automatico, di database e di econometria, si concentrassero sulla stessa tipologia di problemi, senza lavorare sinergicamente e promuovere compiutamente i loro rispettivi lavori. Le ricerche portate a termine mostravano lacune e frammentazioni difficilmente ricomponibili.

Il *data mining* unisce tutte queste discipline, presupponendo che si possa attribuire un maggior valore conoscitivo ai *database*.

La necessità del *data mining* è stata alimentata dalla *Business community*, che ha contribuito a rendere popolare il "*data warehousing*", ossia la collezione di immensi databases valevoli per l'analisi ed il

supporto decisionale, contrariamente ai "*transactional databases*", di cui si servono i managers per le operazioni di "*accounting*" e "*inventory*".

Negli ultimi anni, essendo cresciuto improvvisamente il numero di dati in circolazione per scopi commerciali, si è reso necessario predisporre strumenti per gestire tale fenomeno, creando una comunità di utenti che abbia la possibilità e le potenzialità per *scavare* (*to mine*) affondo questi databases e costruire modelli d'analisi di dati, i *patterns*; solo in tal modo è pensabile sia di migliorare il "*business process*" sia di analizzare dettagliatamente un esperimento scientifico (ad esempio, aiutare i dottori a comprendere meglio gli effetti di un trattamento medico).

2.1.4 Definizione di Data Mining

La comunità scientifica ha elaborato una definizione di *data mining*, sulla quale v'è accordo pressoché unanime. Per *data mining* s'intende il "*processo di selezione, esplorazione e modellazione di grandi masse di dati, al fine di scoprire regolarità o relazioni non note a priori, ed allo scopo di ottenere un risultato chiaro ed utile al proprietario del database*" (M. J. A. Berry, G. S. Linoff, Data Mining, Apogeo, 2001).

In ambito economico ed aziendale, si distingue il *data mining* dalla statistica applicata, non tanto per la quantità di dati che vengono analizzati o per le singolari tecniche che vengono impiegate, quanto per il bisogno di operare in condizioni in cui la conoscenza delle caratteristiche del database e le conoscenze di business sono fortemente integrate.

Difatti, il "*data mining process*" consiste in un processo metodologico integrato, in cui si evidenziano differenti operazioni logiche matematiche, in particolar modo la traduzione delle esigenze di business in una problematica da analizzare, l'estrazione del database necessario per l'analisi, fino alla pubblicazione di uno o più tecniche statistiche, implementate in un algoritmo informatico, al fine di produrre risultati significativi e funzionali al *decision making*.

Più dettagliatamente, il *data mining* non si limita alla mera applicazione di un algoritmo informatico o di una metodologia statistica, bensì di un processo di *business intelligence* che, in quanto tale, è volto all'utilizzo di quanto realizzato dalla tecnologia dell'informazione per supportare le decisioni aziendali.

Il *data miner* deve possedere solide conoscenze di base di tipo statistico, necessarie a mantenere una costante attenzione sugli aspetti statistico-metodologici e, infine, abilità in ambito informatico.

Il significato del "*data mining*" è stato per lungo tempo frainteso dagli statistici, assimilandolo a termini come "*data fishing*", "*data dredging*" o "*data snooping*". In tutti questi casi, l'interpretazione fornita ha voluto attribuire al *data mining* un'accezione negativa, che si basa fondamentalmente su due elementi. In primo luogo, nel data mining, non esiste un solo modello teorico di riferimento, ma numerosi modelli in competizione, selezionati sulla base dei dati in esame. Quindi, è plausibile che si individui un modello, sebbene complesso, che si adatti ai dati.

In secondo luogo, è verosimile che l'immensa quantità di dati complessi a disposizione possa indurre erroneamente a confermare relazioni inesistenti.

Tenendo in viva considerazione tali critiche, è necessario ribadire che le moderne metodologie di *data mining* prestano particolare attenzione al concetto di *generalizzabilità dei risultati*: ciò implica che, nella selezione di un modello, si tenga in debito conto la capacità predittiva, col risultato che vengono penalizzati modelli più complessi. In secondo luogo, è difficile disconoscere che risultati di maggior interesse, a livello applicativo, non sono noti a priori e, come tali, non quantificabili in un'ipotesi di ricerca. Ciò accade in presenza di database di considerevole entità (relativamente alle osservazioni, alle variabili o al numero di tabelle relazionali prese in esame).

Quest'ultimo aspetto è uno dei tratti che distingue il *data mining* dalla statistica applicata: l'analisi di dati primari, raccolti allo scopo di verificare determinate ipotesi di ricerca, è tradizionalmente una prerogativa dell'analisi statistica, mentre il *data mining* si concentra sui dati secondari rilevanti e, in seguito, raccolti per scopi differenti dal mero procedimento di analisi dei dati.

Infine, la natura di questi dati è radicalmente diversa. Il *data mining* si serve di dati di origine tipicamente osservazionale, mentre, in campo statistico, i dati possono essere definiti anche in forma sperimentale (frutto di un disegno di esperimenti che accoppia casualmente le unità statistiche a diverse tipologie di trattamenti).

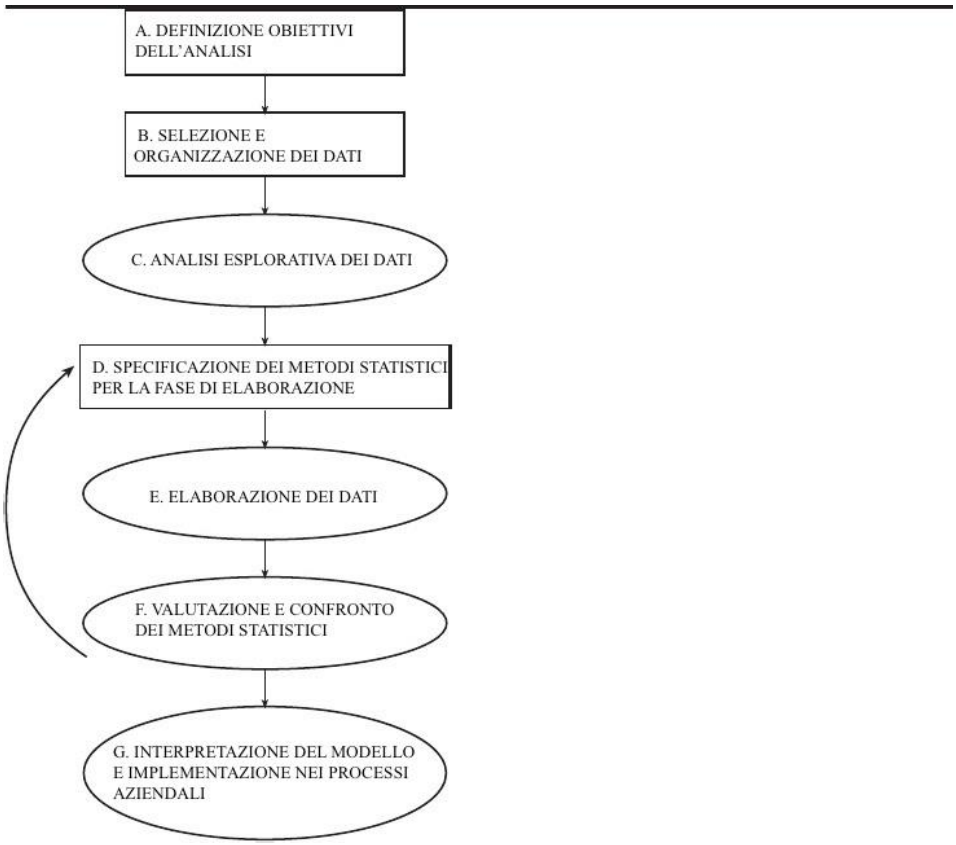


Fig. 2.2 Modello di processo del Data Mining

2.1.5 Le fasi dell'attività di KDD e Data Mining

Per identificare i *pattern* sono necessarie tre condizioni:

1. il database deve essere organizzato in maniera tale che ciascun dato o insieme di dati sia integrato con tutto il resto dell'informazione e non solo con una parte di essa;
2. i dati, così integrati, devono essere analizzati per il recupero dell'informazione;
3. l'informazione recuperata deve essere presentata in modo da rendere più immediato possibile la sua comprensione e il suo utilizzo.

Per soddisfare tali condizioni, un sistema di *data mining* (chiamato *data mining system*, DMS) ha bisogno di lavorare su una banca dati di tipo "*data warehouse*", che organizza le informazioni in modo da eliminare le ridondanze ed eventuali inconsistenze memorizzandone solo le componenti significative, in modo da facilitare l'analisi successiva.

Inoltre, un DMS deve essere in grado di scoprire automaticamente le informazioni "nascoste" nei dati: questi sistemi si chiamano "*discovery-driven*", ossia, orientati alla scoperta, e sono progettati con l'obiettivo di essere usati per migliorare la conoscenza.

Possiamo elencare formalmente il processo di KDD nei seguenti passi (figura 2.2):

1. *selezione*: estrazione di parte dei dati secondo alcuni criteri, dipendenti dall'obiettivo preposto all'analisi. Facendo riferimento alla metodologia statistica, si usa il termine *campionamento* dei dati;
2. *pre-elaborazione*: pulizia dei dati da informazioni ritenute inutili, in quanto possono rallentare le future interrogazioni. In questa fase, peraltro, i dati possono essere trasformati, per evitare eventuali inconsistenze dovute al fatto che dati simili possono provenire da sorgenti diverse e, pertanto, con metadati leggermente diversi (ad esempio in un database un bene o servizio economico può essere salvato come…);
3. *trasformazione*: I dati non sono semplicemente trasferiti da un archivio ad uno nuovo, ma sono trasformati in modo tale che sia possibile anche aggiungere informazioni, come, ad esempio, informazioni demografiche comunemente usate nella ricerca di mercato. Quindi, i dati vengono resi "usabili" e "navigabili";
4. *data mining*: questo stadio si occupa di estrarre modelli dai dati. Un modello può essere definito come segue: dato un insieme di fatti (i dati) F, un linguaggio L e alcune misure di certezza C, un modello è una dichiarazione S nel linguaggio L che descrive le relazioni che esistono tra i dati di un sottoinsieme G di F con una certezza c tale che S sia più semplice in qualche modo della enumerazione dei fatti contenuti in G;
5. *interpretazione e valutazione*: i modelli identificati dal sistema vengono interpretati, cosicché la

conoscenza che se ne acquisisce può essere di supporto alle decisioni, quali ad esempio la previsione, la classificazione degli elementi, la sintesi dei contenuti di un database o la spiegazione dei fenomeni osservati.

Sia la fase di pre-elaborazione che la fase di trasformazione si avvalgono di tecniche e strumenti software, ai quali si fa riferimento con il termine "*processo di ETL*".

Focalizzandoci sui risultati, il processo di *data mining* si può suddividere essenzialmente in due fasi.

1. *Esplorazione dei dati*. Questo passo, che precede la scoperta di un modello valido, comporta e necessita di una visualizzazione e di un'esplorazione controllata. Il suo scopo consiste di dare all'utente una prima visione dei dati, per evidenziare errori nella preparazione e nell'estrazione di quest'ultimi.
2. *Generazione di pattern*. Questo passaggio impiega la cosiddetta *regola di scoperta* (automatica o interattiva) e algoritmi di *scoperta di associazioni* per generare modelli. Questo passo coinvolge anche la convalida e l'interpretazione dei modelli scoperti.

2.1.6 Tipologie di "Data Mining Patterns"

La ricerca scientifica è riuscita a distinguere due tipologie di modelli, identificati dal *data mining*: *predittivo* ed *informativo*.

I *modelli predittivi* sono ideati per risolvere problemi, riguardanti l'identificazione di una o più entità presenti nel database, ad esempio il problema della conservazione della clientela (*customer retention*), descritto in precedenza. Tali modelli non sono sempre orientati a predire il futuro, bensì la loro caratteristica più importante è la capacità di proporre un'ipotesi plausibile sul valore di un'entità ignota, dati i valori delle altre variabili note.

Al contrario, i *modelli informativi* non sono finalizzati alla risoluzione di un problema specifico, ma, piuttosto, ad ideare un modello altamente innovativo, che possa offrire un contributo fondamentale alla ricerca scientifica.

È evidente che i modelli informativi sono molto più difficili da giudicare rispetto ai modelli predittivi, in virtù del fatto che i primi sono confinati alla ricerca puramente teorica e non sono in grado di suggerire alcuna soluzione pratica.

Tutti gli algoritmi di *data mining* incorporano una misura della bontà o attrattiva di un modello, consentendo, in tal modo, di individuare *patterns* che, successivamente, potranno essere conservati, eliminati o continuare ad esplorarli.

Giacché i modelli predittivi pronosticano il valore di un'entità o proprietà, ed in quanto quest'ultima esiste nel "*training database*", il metodo ordinario per valutare i modelli predittivi consiste nel confrontare

le loro previsioni con i valori correnti assunti dalle entità o proprietà, in seno al *training set*.

Misurando la frequenza e il numero degli errori commessi, l'algoritmo di *data mining* può giudicare i modelli.

Nel corso dell'applicazione di un algoritmo, le osservazioni dei modelli informativi non possono essere valutati in base alla qualità del proprio lavoro, in quanto quest'ultimo è condizionato dalle abilità dell'esperto. Invece, i principi matematici consentono sia di cogliere gran parte dei modelli potenzialmente interessanti sia di eliminare i *patterns* considerati poco stimolanti, dal punto di vista scientifico.

Inizieremo la nostra analisi delle diverse tipologie di "*data mining process*", esaminando i modelli informativi e i criteri di valutazione di quest'ultimi.

2.*2 Metodologie di Data Mining*

Il *Data Mining* non è un algoritmo, quanto piuttosto un processo fondato nella ricerca di *pattern* validi, nuovi, utili e comprensibili.

Tale processo si fonda su tre pilastri fondamentali:

1. *esplorazione;*
2. *modellazione;*
3. *valutazione.*

La fase di esplorazione è rappresentata dai dati su cui poggia il *data mining*: senza di essi il *data mining* non sarebbe possibile e potrebbe contare solamente su intuizioni più o meno corrette. La forza del *data mining* sta nel far leva sui dati ricavati durante un'attività di business, al fine di prendere decisioni più informate.

Il *data mining* ha il pregio di ridurre la complessità della realtà in una semplice tabella. Perché gli algoritmi funzionino in modo ottimale, è necessario che tutti i campi siano compilati e che i valori siano sensati. Il processo di raccolta di tutte le varie fonti dati e di estrazione delle informazioni che davvero interessano, rappresenta la sfida più impegnativa.

La fase di modellazione richiede una serie di competenze di modellazione per costruire modelli previsionali. Prima di tutto, occorre conoscere, in maniera approfondita, le diverse tecniche e i diversi ambiti su cui applicarle. La maturità raggiunta dalla disciplina permette una visione interdisciplinare del *data mining* che si può descrivere come l'intersezione tra sistemi basati sulla conoscenza, sistemi di autoapprendimento, statistica, visualizzazione, teoria delle basi di dati.

Infine, la fase di valutazione è necessaria per confrontare i risultati di alcuni modelli presi in esame. Solitamente, questi confronti si fanno su problematiche di tipo supervisionato; problematiche, cioè, dove si conoscono i risultati di un'applicazione. Nei modelli previsionali, i risultati effettivi, in genere, sono peggiori

delle previsioni, perciò poi, è necessario rimettere mano alla modellazione, per poter avere un determinato tipo di risposta.

In conclusione, il *data mining* non si confina ad una sola attività e il suo buon esito è legato necessariamente ai tre macro processi di esplorazione, modellazione e valutazione. Nonostante gli strumenti a disposizione si siano evoluti, i risultati non dipendono esclusivamente dalla tecnica quanto da un'elevata competenza e conoscenza del contesto applicativo. La vera sfida è quella di riuscire a gestire l'interdipendenza tecnica conservando l'importanza della conoscenza della problematica di business.

2.3 Trattamento preliminare dei dati

Il *data min*ing differisce da altri processi logici e analitici, nel senso che è apparentemente semplice applicare un "*data mining alghorithm*" a un *database* e, successivamente, ottenere alcuni risultati. Tuttavia, senza una chiara comprensione dei dati, quest'ultimi risulterebbero inutili. La preparazione dei dati assume, dunque, un ruolo cruciale nel processo di *data mining*, in quanto può influenzare, in maniera determinante, la bontà dei modelli.

Ora, la comunità accademica ha distinto quattro passaggi determinanti in questo processo di trattamento preliminare dei dati:

1. *data clearing* (pulizia dei dati): colmare i campi con i valori mancanti, attutire i dati rumorosi, eliminare i valori non realistici;
2. *data integration* (integrazione dei dati): l'integrazione dei dati provenienti da database multipli, chiarendo inconsistenze statistiche;
3. *data transformation* (trasformazione dei dati): la preparazione dei dati funzionali all'applicazione di particolari algoritmi di analisi;
4. *data reduction* (riduzione dei dati): la riduzione della mole di dati o del numero delle variabili in input, senza compromettere la validità delle analisi.

Il primo aspetto da considerare riguarda la qualità dei dati. Difatti, è molto probabile che, nel corso della raccolta dei dati, l'esperto possa misurarsi con informazioni mancanti. Pertanto, la fase di pre-elaborazione (dei dati) costituisce un passaggio fondamentale per migliorare la qualità della sorgente dei dati.

Assumendo che i dati da utilizzare per il *data mining* giungano direttamente dal *data warehouse*, si dà per scontato che il processo di pulizia, integrazione e uniformazione dei dati sia già stato compiuto.

In caso contrario, occorre valutare il grado di affidabilità e completezza dei dati e porre rimedio a specifici problemi di qualità (dei dati) (*data integration*, *data transformation*, *data reduction*), pena la costruzione di modelli destinati ad essere completamente inefficaci.

Successivamente, l'esperto affronta il problema legato ai valori mancanti degli attributi (*data missing*), che può essere sanato parzialmente nel data warehouse, mediante l'integrazione di fonti diverse, anche esterne all'azienda (si pensi all'acquisto di dati demografici o relativi al territorio da banche dati specializzate). Accade però spesso che, per alcuni attributi, non sia sempre possibile ottenere un valore. In questo caso, nel data warehouse, invece di presentare un valore NULL, si utilizzerà un valore di default che indica la mancanza del dato, tuttavia, questa soluzione non è indicata ottimale per il data mining ed è da gestire al momento della preparazione dei dati.

Alcuni algoritmi richiedono che i dati siano trasformati (*data transformation*, *data reduction*), di solito attraverso operazioni di vario genere: normalizzazioni, riduzione del numero di attributi, riclassificazione dei valori di un attributo.

Nei paragrafi seguenti, affronteremo alcune tecniche utilizzate per la preparazione dei dati.

2.3.1 Normalizzazioni

Gli algoritmi, basati sul calcolo della distanza tra punti nello spazio multidimensionale, beneficiano della normalizzazione dei valori ad un intervallo (per es. [-1,+1] oppure [0,1]). In caso contrario, gli algoritmi finirebbero per attribuire un peso eccessivo ai valori più grandi.

La normalizzazione di un attributo può avvenire in vari modi.

a) La normalizzazione minimo-massimo, dove il valore normalizzato è dato dalla seguente formula:

$$ValNorm_i = \frac{Val_i - \text{Min}(Val)}{Max(Val) - Min(Val)}$$

Questa metodologia non introduce distorsioni nei dati, tuttavia, non garantisce che, successivamente, non vi siano dati, il cui valore minimo o massimo sia fuori dal *range*, impiegato per la normalizzazione. Scegliendo tale modalità di normalizzazione, sarà determinante gestire i valori fuori range, mediante una tra le seguenti tecniche:

- nel caso in cui si usi un algoritmo in grado di gestire i valori fuori range, non è necessario compiere nessuna azione;
- prevedere anticipatamente abbastanza "spazio" per gli eventuali valori fuori range, utilizzando un valore superiore al massimo al posto di Max(Val) e inferiore al minimo al posto di Min(Val);
- effettuare un *clipping* dei valori che eccedono gli estremi, ponendoli pari o al minimo, se ne sono inferiori, o al massimo, se ne sono superiori.

b) La normalizzazione con la deviazione standard, dove il valore normalizzato è dato dalla distanza del valore originale e la media dei valori, espressa in termini di numero di deviazioni standard:

$$ValNorm_i = \frac{Val_i - \text{Media}(Val)}{DevStd(Val)}$$

Per le variabili discrete, la normalizzazione può avvenire con la stessa formula, nella quale, però la media è pari alla probabilità *p* dello stato considerato, mentre la deviazione standard è pari a *p*(1-p)*.

c) La normalizzazione con la funzione logistica. In questo caso, la funzione logistica è utilizzata per ottenere il valore normalizzato:

$$ValNorm_i = \frac{1}{1 + e^{-Val_i}}$$

La funzione logistica "schiaccia" i valori nell'intervallo (0,+1), come mostrato nella figura seguente.

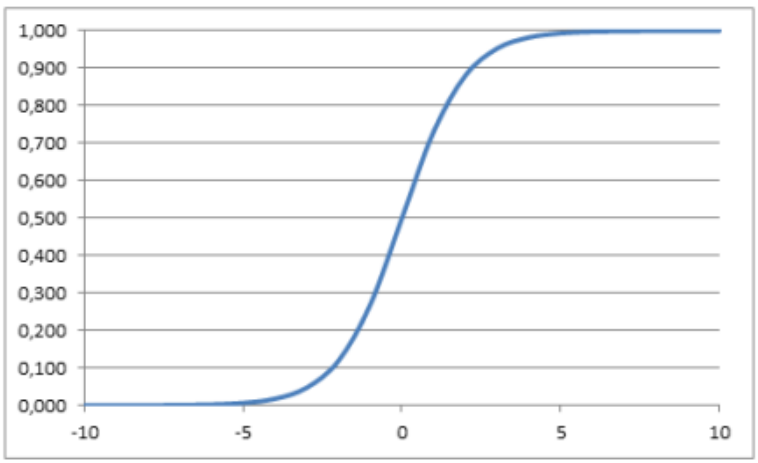


Fig. 2.3 Funzione logistica

Data la finita capacità massima dei computer di gestire i decimali, i valori vicini ai massimi e ai minimi, prodotti dalla distribuzione, sono arrotondati rispettivamente a 1 e a 0. Per evitare questa situazione è possibile fornire alla funzione logistica un valore trasformato (*ValTrasfi*), al posto del valore originale (*Vali*). Il valore trasformato è determinato con la formula seguente:

$$ValTrasf_i = \frac{Val_i - \text{Media}(Val)}{l \times (DevStd(Val)/2\pi)}$$

dove "*l*" è un coefficiente che esprime la dimensione della risposta lineare desiderata.

Per esempio utilizzando in input gli stessi valori usati per il grafico precedente, ma trasformando i dati, impostando il coefficiente l pari a 2, otteniamo questo grafico:

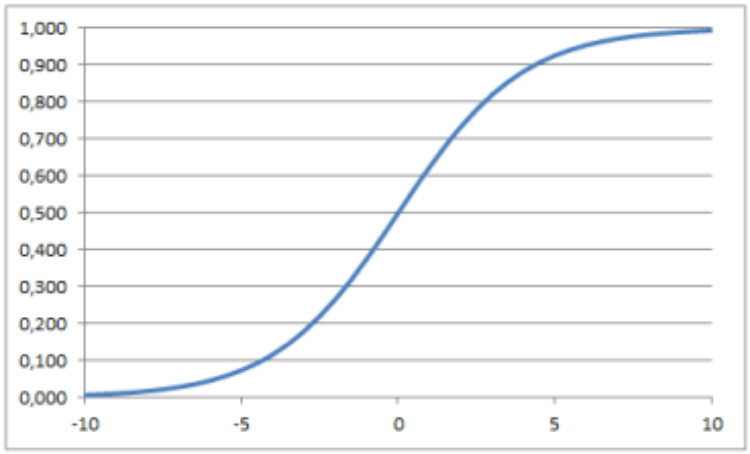


Fig. 2.4 Funzione logistica con valori trasformati

Confrontando la figura con quella originale, notiamo che la forma a S è meno marcata e le zone vicino al minimo e al massimo sono meno piatte, poiché la componente lineare è stata aumentata. Si conservano, quindi, più dettagli sugli estremi.

2.3.2 Smoothing

Un attributo numerico continuo potrebbe presentare un numero elevatissimo di valori distinti, pari anche al numero totale di elementi; inoltre, i valori, spesso, presentano o si differenziano solo per piccole variazioni. Un numero così alto di valori per un attributo può condurre a un degrado sia del risultato finale, sia delle performance di calcolo. Per questa ragione, è opportuno rendere più omogenei i valori, mediante tecniche di smoothing. La tecnica più semplice, laddove vi siano valori con diverse cifre decimali, è quella di eseguire un arrotondamento, avente l'effetto di rendere più uniformi i valori.

Per le serie temporali, una delle tecniche di smoothing disponibili è la *media mobile*, cioè la media degli ultimi *n* valori ottenuta variando via via l'istante temporale di riferimento, in modo da creare un'altra serie temporale di dati.

2.3.3 Trattamento dei dati mancanti

Come accennato precedentemente, i dati mancanti (*missing values*) sono una realtà connaturata anche al datawarehouse e, spesso, la costruzione di un dataset, avente tutti gli attributi completi, è possibile soltanto

selezionando una quantità molto limitata di dati.

Risulta evidente, pertanto, che non possiamo scartare ogni record che presenti un attributo mancante, ma occorre utilizzare una tecnica che mitighi tale problema.

La gestione dei valori mancanti può avvenire con le seguenti modalità:

- sostituzione del valore nullo con una costante;
- sostituzione del valore nullo con la media;
- sostituzione del valore nullo con un valore che non alteri la deviazione standard;
- sostituzione del valore nullo con la media per la stessa classe di elementi (questa tecnica è possibile sono quando i dati sono classificati a priori);
- creazione di nuovi record, basati su quello originale, che presentino gli stessi valori per gli attributi valorizzati e, per l'attributo mancante, uno dei valori del dominio di quell'attributo. La tecnica prevede la creazione di un nuovo record per ciascuno degli elementi del dominio;
- riempimento dei valori mancanti, in base a tecniche di correlazione tra gli attributi: in tal caso, si utilizzano metodologie di data mining, per scoprire le relazioni tra gli attributi di quei record che non hanno valori mancanti. Una volta creato il modello, viene impiegato per colmare i "buchi" nei record incompleti.

Tutte le tecniche, presentate in questo paragrafo, sono interessate dal medesimo problema: il valore, con cui si sostituisce il dato mancante, non corrisponde alla realtà e ciò può falsare le analisi. Perciò, è opportuno confrontare diversi modelli creati sia con le tecniche descritte sopra, sia solo coi record originariamente completi, per selezionare il metodo migliore per la gestione dei dati mancanti.

2.3.4 Trattamento degli outliers

In un dataset possono esservi elementi, gli *outliers*, decisamente diversi rispetto agli altri componenti. Gli outliers possono essere causati da errori nei dati, oppure possono essere un dato reale, nonostante siano espressione di un comportamento totalmente diverso rispetto al resto del dataset.

Gli algoritmi di data mining tendono a ridurre l'influenza degli outliers nel modello e, spesso, essi sono identificati a priori dall'analista ed eliminati, prima ancora di esaminare il modello. È bene puntualizzare che, in alcuni casi, l'identificazione degli outliers costituisce precisamente lo scopo del modello di data mining: si pensi, ad esempio, a un'applicazione per individuare l'utilizzo fraudolento di carte di credito.

L'identificazione degli outliers avviene principalmente coi seguenti sistemi:

- *soglie di deviazione standard*: i valori normali sono compresi tra $Media \pm 2 \times Dev\ standard$; questo

metodo prende in considerazione un attributo per volta.

- *distanza euclidea tra punti*: sono considerati outliers, quei punti che hanno un numero limitato di "vicini". La distanza tra punti multidimensionali a e b è calcolata, mediante la formula $\sqrt{\sum_i^n (a_i - b_i)^2}$.

Una volta calcolata la distanza, si stabilisce sia una soglia oltre la quale i punti possono essere considerati vicini, sia il numero minimo di punti vicini sotto al quale si classifica un elemento come outlier.

2.3.5 Riduzione del dataset

Abbiamo già menzionato la possibilità di ridurre il numero di attributi, impiegati per la costruzione di un modello di data mining.

I vantaggi che si ottengono dalla riduzione del dataset sono principalmente due:

- la diminuzione dei tempi di elaborazione;
- il miglioramento dell'accuratezza del modello, che dovrebbe beneficiare dell'eliminazione di variabili irrilevanti o ridondanti.

a) Riduzione della dimensionalità del dataset

Il primo intervento, che potremmo operare sul dataset, consiste nella selezione degli attributi da mantenere nel modello. La selezione può avvenire manualmente, basandosi sulla conoscenza del business e dei dati, da parte dell'analista.

La tecnica statistica più popolare, con cui si ottiene la riduzione della dimensionalità di un insieme di dati, è l'analisi delle componenti principali.

Tale metodologia mira a trasformare l'insieme iniziale dei dati in un altro insieme, per il quale le prime *n* dimensioni (con *n* piccolo) contengono la maggior parte delle informazioni rilevanti.

Quanto più è alta la varianza di un attributo, tanto più informazioni esso contiene.

Senza entrare nel minimo dettaglio del procedimento, la determinazione delle componenti principali prevede, come primo passo, il calcolo degli *autovalori della matrice di covarianza* (*o della matrice dei coefficienti di correlazione*) *delle variabili*, ottenendo un autovalore per ciascuna variabile. L'autovalore più grande indica l'attributo con la maggiore varianza: esso costituisce la varianza della componente principale 1. Proseguendo in ordine decrescente, il secondo autovalore indica la varianza della componente principale 2, e così via per gli tutti gli autovalori. A questo punto, si conosce l'ordine degli attributi, in base al valore decrescente dell'autovalore.

Infine, occorre soltanto decidere discrezionalmente quanta della varianza totale si vuole mantenere, attraverso l'impiego di alcuni attributi, partendo da quelli che corrispondono alle prime componenti. Poniamo di voler spiegare il 95% della varianza. Per ogni componente principale, si calcola il rapporto tra la

somma cumulata degli autovalori a partire dal primo fino ad arrivare a quello corrente e, poi dividiamo per la somma totale:

$$R = \frac{\sum_i^m V_i}{\sum_i^n V_i}$$

dove:

Vi è l'i-esimo autovalore;

n è il numero totale di autovalori (e quindi di attributi);

m è il numero di autovalori (e quindi di attributi) considerato per ogni calcolo di *R*. *m* parte da 1 (cioè il primo autovalore) e può arrivare a *n* (con *m*=*n*, R vale 1, cioè 100%)

S'impiegheranno gli attributi a partire dalla prima componente, fino a quello per cui il valore di *R* inizia a essere >= al coefficiente, che è stato stabilito dall'esperto essere 95%.

b) Riduzione del numero di valori nel dominio degli attributi

Gli attributi continui potrebbero esibire un numero altissimo di valori distinti, potenzialmente pari al numero di elementi. Gli algoritmi di data mining non operano agevolmente con un numero troppo elevato di valori per ciascun attributo, pertanto si rende necessaria una loro riduzione. La metodologia, impiegata a tal fine, consiste nella discretizzazione della variabile continua, ossia nella mappatura di intervalli a un numero limitato di valori discreti (ad esempio, una variabile continua, come l'attributo "reddito annuo", si può trasformare in un nuovo attributo, chiamato classe di reddito e che esprime degli intervalli di reddito, così riducendo drasticamente il dominio del nuovo attributo).

c) Riduzione del numero di record per la costruzione del modello.

Non è necessario utilizzare tutti i dati disponibili per la costruzione di un modello, ma si può effettuare un campionamento dei record. La tecnica ordinaria di campionamento è quella del campionamento casuale di un certo numero di elementi.

Tuttavia, com'è possibile sapere quanti elementi bisogna impiegare per costruire un campione efficace? Un metodo utile a determinare il numero di elementi è quello del *campionamento incrementale*: s'inizia con una percentuale bassa di elementi, successivamente, si costruisce il modello e lo si valuta; in seguito, si accresce la percentuale di campionamento, finché i risultati del modello s'assestano su valori ottimali. In tal modo, si stabilisce il campione ideale per la situazione oggetto di analisi.

2.4 Costruzione del modello

La costruzione del *modello di data mining* s'articola su più fasi:

- *La scelta dell'algoritmo di calcolo*. Si basa sull'analisi del problema di data mining da risolvere.
- *Il completamento della fase di preparazione dei dati*. Qualora l'algoritmo richieda elaborazioni particolari, è necessario ultimare la fase di preparazione dei dati, mediante le tecniche trattate nel paragrafo precedente.
- *La scelta dei parametri base di configurazione dell'algoritmo.*
- *La suddivisione dei dati disponibili in training set e test set.* Nel corso della costruzione di un modello di data mining, occorre operare una suddivisione dei dati disponibili in due insiemi: il primo, contenente un'ampia percentuale dei dati, il *training set*, ovvero l'insieme dei dati su cui l'algoritmo scelto è calibrato. Il secondo rappresenta il *test set*, cioè l'insieme di dati su cui si effettuerà il test del modello per verificarne la bontà. È ovvio che il test set contiene anche l'attributo o gli attributi, risultanti dall'attività predittiva del modello; in tal modo, sarà possibile confrontare i dati reali con quelli previsti e, in seguito, eseguire una valutazione.

  La suddivisione in training set e test set dovrebbe avvenire mantenendo la stessa distribuzione degli attributi in entrambi gli insiemi di elementi, in modo che essi siano ugualmente rappresentativi.
- *L'avvio della fase di training dell'algoritmo.* Nel corso della fase di training, l'algoritmo esamina le relazioni nascoste nei dati e imposta il modello di data mining.

2.4.1 Attività tipiche

Le attività, tipicamente oggetto di un processo di *data mining*, sono raggruppabili in categorie. Per ogni categoria, si possono individuare uno o più algoritmi di *data mining* che meglio si prestano a risolvere il problema. La tabella presenta una categorizzazione dei problemi di *data mining*, una breve descrizione e gli algoritmi più adatti a ciascuna categoria.

Problema	**Esempio**	**Algoritmo**
Stima di un attributo discreto: in questo caso si tratta di predire il valore di un particolare attributo sulla base dei valori degli altri attributi.	Stimare se il destinatario di una campagna di mailing diretto acquisterà un prodotto, sulla base di dati anagrafici e comportamentali di vario genere.	Decision Trees Bayesian classifier Clustering Neural Network

Stima di un attributo continuo.	Stimare le vendite dell'anno successivo	Time Series Neural Network
Ricerca di gruppi di elementi comuni nelle transazioni.	Utilizzare analisi di mercato sugli acquisti per suggerire a un cliente ulteriori prodotti da acquistare.	Association Rules Decision Trees
Ricerca di gruppi di elementi simili.	Segmentare i dati demografici in gruppi, con comportamenti d'acquisto simili	Clustering
Ricerca di anomalie nei dati	Per esempio la ricerca di utilizzi fraudolenti di strumenti di pagamento, come le carte di credito.	Clustering

Fig. 2.5 Problemi e algoritmi

2.4.2 Gli Algoritmi

In questa sede, si propone una panoramica dei metodi di calcolo utilizzati più di frequente dagli strumenti di *data mining*.

2.4.3 Il Clustering

Questo paragrafo illustra l'applicazione degli algoritmi di *clustering*. I record attraverso diversi algoritmi vengono raggruppati in base a delle analogie o delle omogeneità.

Nel *clustering* non esistono classi predefinite né tanto meno esempi di appartenenza ad una certa classe. Dunque, sta a chi applica l'algoritmo stabilire l'eventuale significato da attribuire ai gruppi che si sono formati.

2.4.3.1 Cluster Analysis

Con l'espressione "*Cluster Analysis*", s'intende definire quel processo che suddivide un insieme generico di *pattern* in gruppi di *patterns* o oggetti simili. Un *cluster* è un insieme di oggetti, che presentano tra loro delle similarità, tuttavia, presentano dissimilarità con oggetti in altri cluster. L'input di un algoritmo di clustering è costituito da un campione di elementi, mentre l'output è dato da un determinato numero di cluster, in cui gli elementi del campione sono suddivisi in base a una misura di similarità.

Gli algoritmi di *clustering* forniscono come output anche la descrizione delle caratteristiche di ciascun cluster.

I motivi principali del successo di questo tipo di algoritmi sono essenzialmente due:

- le tecniche di analisi dei gruppi sono largamente usate nei più svariati campi di ricerca (fisica, scienze sociali, economia, medicina, ecc.), in cui la classificazione dei dati disponibili è un momento essenziale della ricerca di modelli interpretativi dei fenomeni sociali, economici e scientifici;
- grazie all'evoluzione degli strumenti di calcolo automatico, la clustering analysis ha consentito di affrontare, senza difficoltà, la complessità computazionale, insita nella maggior parte dei metodi di classificazione e che, in precedenza, aveva spinto i ricercatori ad orientarsi verso quelle tecniche di analisi dei gruppi che erano più facilmente applicabili.

Si è resa così possibile la produzione di diversi algoritmi di classificazione, maggiormente complessi dal punto di vista computazionale, ed efficienti nel trarre informazioni dai dati, mediante una loro opportuna classificazione.

Tra le proprietà della Cluster Analysis, considerate desiderabili dagli analisti, sono l'esaustività, cioè suddivide in classi tutti i dati presi in considerazione e l'esclusività, in quanto genera delle partizioni sull'insieme originario.

2.4.3.2 Gli algoritmi di clustering

Nel momento in cui si procede al *clustering*, una delle esigenze più comuni è quella di raggruppare gli oggetti appartenenti ad un insieme dato, in modo tale da definire dei sottoinsiemi il più possibile omogenei. Gli algoritmi di *clustering* si dividono in due categorie principali:

1. algoritmi di *clustering* gerarchico;
2. algoritmi di *clustering* partizionale.

I primi organizzano i dati in sequenze nidificate di gruppi, rappresentabili in una struttura ad albero, mentre gli algoritmi di *clustering* partizionale determinano il partizionamento dei dati in cluster, riducendo quanto più possibile la dispersione all'interno del singolo *cluster* e, allo stesso tempo, di aumentare la dispersione tra i *cluster*.

Gli algoritmi di *clustering* partizionali sono più adatti a data set molto grandi, per i quali la costruzione di una struttura gerarchica dei cluster porterebbe a uno sforzo computazionale molto elevato.

Eseguire il *clustering* di un insieme assegnato, contenente oggetti descritti da un insieme di osservazioni, significa individuare gruppi di oggetti tali che:

- gli elementi appartenenti ad un cluster siano omogenei tra loro, ovvero simili sulla base delle osservazioni (alta similarità intraclasse);
- gli elementi appartenenti a cluster diversi siano disomogenei tra loro sulla base delle osservazioni (bassa similarità inter-classe).

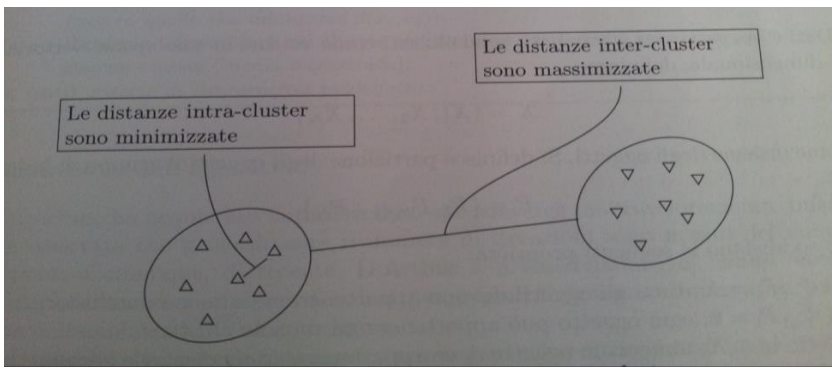


Fig. 2.6 Distanze intra-cluster e inter-cluster

La misura di similarità tra un elemento e un altro può essere calcolata con metodi diversi. Uno dei più semplici è la distanza euclidea tra due punti in uno spazio n-dimensionale. La sua formula è:

$$d(x_i, x_j) = \sqrt{\sum_{k=1}^{n} (x_{i\,k} - x_{j\,k})^2}$$

dove:

xi e *xj* sono due elementi del dataset rappresentati da vettori. Es.: *xi* = {*xi1, xi2, ..., xin*}

n è il numero di dimensioni

La formula è valida per le variabili continue, mentre per i valori booleani (vero/falso) o per le variabili categoriche (per es. il colore che potrebbe essere verde, bianco, rosso, ecc..) si utilizza un'altra formula:

$$d(x_i, x_j) = \frac{p - m}{p}$$

dove:

p è il numero di attributi booleani o categorici;

m è il numero di attributi con lo stesso valore tra *xi* e *xj*.

Il clustering gerarchico di tipo agglomerativo impiega l'algoritmo seguente per l'identificazione dei *cluster*:

1. ciascun elemento del dataset forma inizialmente un *cluster* separato;
2. ad ogni iterazione, sono accorpati i *cluster* più simili, estendendo via via la soglia di similarità;
3. l'iterazione termina nel momento in cui tutti gli oggetti sono stati accorpati in un unico *cluster*, ove tutti gli elementi sono considerati simili.

La formazione dei cluster è rappresentata attraverso un *dendogramma*, o "grafico a albero", come mostra la figura seguente:

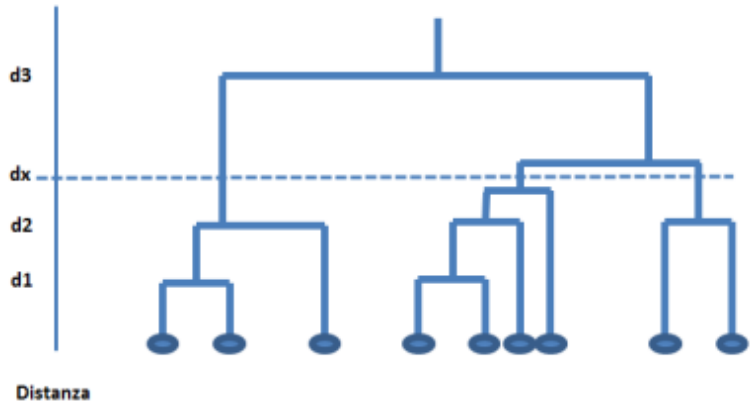


Fig. 2.7 Dendogramma

La linea tratteggiata, definita dall'esperto e rappresentata in figura 2.7 col valore di distanza *dx*, identifica il numero di *cluster*.

Dal punto di vista del calcolo, l'algoritmo risulta accurato, ma poco efficiente. Oltretutto, durante l'iterazione, non prevede una riassegnazione degli elementi ai cluster.

Nel corso dell'elaborazione del clustering partizionale, l'analista stabilisce inizialmente quanti cluster si intendono avere e s'indica tale numero con *K*. Il procedimento è il seguente:

1. partizioniamo l'insieme in *K* cluster, assegnando, a ciascuno di essi, degli elementi scelti casualmente;
2. calcoliamo i centroidi di ciascun cluster *k* con la formula seguente:

$$M_k = {}^{1}/_{n_k} \times \sum_{i=1}^{n_k} x_{ik}$$

dove:

Mk è il vettore delle medie, o centroide per il cluster *k*

nk è il numero di elementi del cluster *k*

xik è l'i-esimo elemento del cluster

1. calcoliamo la distanza degli elementi del cluster dal centroide, ottenendo un errore quadratico, mediante la formula:

$$e_k^2 = \sum_{i=1}^{n_k} (x_{ik} - M_k)^2$$

dove:

ek^2 è l'errore quadratico per il cluster *k*

xik è l'i-esimo elemento del cluster

nk è il numero di elementi del cluster *k*

Mk è il vettore delle medie, o centroide per il cluster *k*

La somma degli errori di tutti i cluster fornisce l'errore totale per la partizione, che rappresenta il valore da minimizzare.

2. A questo punto, si riassegnano gli elementi del campione in base al più vicino centroide;
3. si ripetono i passaggi 2,3 e 4, finché il valore minimo dell'errore totale non è raggiunto, o finché i membri dei cluster non si stabilizzano.

2.4.4 Le Regole di Associazione o "Association Rules"

Gli algoritmi di *association rule* sono concepiti per individuare relazioni tra elementi in basi dati estese. Una *association rule* è spesso rappresentata con un'implicazione, scritta nella forma "X -> Y" (X implica Y), nella quale X e Y sono due elementi di un insieme.

Tipicamente, questi algoritmi sono utilizzati per i problemi di "*market basket analysis"*, nei quali si tenta di identificare specifiche relazioni nei prodotti acquistati in combinazione da parte dei clienti. Tale metodologia permette, ad esempio, di riorganizzare la disposizione dei beni nei negozi della grande distribuzione (*charity shop*), di esaminare il "cross selling" e di progettare attività promozionali.

L'esempio canonico, per proseguire nella descrizione dell'algoritmo di association rules, consiste nell'analisi di un paniere di beni in un negozio, indicando, nella tabella seguente, il numero di transazioni effettuate:

Nr Transazione	Libri	Scarpe	Record	Calzini
1	1	0	1	0
2	0	1	0	0
3	0	0	0	1
4	1	1	1	0
5	1	0	0	0

La prima colonna contiene un numero identificativo della transazione, mentre le altre quattro colonne indicano la presenza nel paniere d'acquisto dei prodotti libri, scarpe, records e calzini.

Il procedimento per la determinazione di una regola di associazione è composto da due passaggi:

1. la definizione degli insiemi di elementi che compaiono con frequenza alta (*frequent itemset*), ossia superiore ad una certa soglia. La frequenza è data dal numero di occorrenze della combinazione di elementi divisa per il totale delle transazioni. Il numero così calcolato è detto "supporto". Per esempio il supporto di {libri, scarpe, records} è pari a 1/5, cioè il 20% delle transazioni contengono questo insieme di elementi;
2. l'estrazione delle regole che hanno un valore di confidenza superiore a una certa soglia. La confidenza di una regola X->Y è determinata dalla formula: $(X => Y) = \frac{Supporto(X \cup Y)}{Supporto(X)}$

   Ad esempio, la regola {libri, scarpe}->{records} ha una confidenza di 0,2/0,4 = 50%

La determinazione delle combinazioni di elementi aventi una frequenza alta non risulta difficoltosa, qualora sia eseguita su una base dati limitata.

Tuttavia, in presenza di database estesi, il calcolo la frequenza di tutte le combinazioni di elementi diviene un compito proibitivo.

Esistono algoritmi di selezione dei frequent itemset, che operano efficientemente. Uno di questi è l'algoritmo "*apriori*", che lavora con un approccio iterativo e consente di non considerare tutte le combinazioni di elementi nel database: tutti i sottoinsiemi non vuoti di un frequent itemset sono anch'essi frequenti e, di conseguenza, tutti gli insiemi che contengono sottoinsiemi poco frequenti, sono anch'essi poco frequenti. Tale metodo consente di ridurre progressivamente lo spazio di ricerca ad ogni successiva iterazione.

2.4.5 Alberi di decisione o "Decision Trees"

Gli alberi di decisione costituiscono il modo più semplice di classificare degli oggetti o pattern in un numero finito di classi. Il principio è la costruzione di un albero, ove i sottoinsiemi (di record) vengono chiamati nodi e quelli finali, foglie.

In particolare, i nodi sono etichettati col nome degli attributi, gli archi (i rami dell'albero) sono etichettati coi possibili valori dell'attributo. Un oggetto è classificato seguendo un percorso lungo l'albero che conduce dalla radice a una foglia. I percorsi sono rappresentati dai rami dell'albero, che forniscono una serie di regole.

Possiamo distinguere tre categorie di alberi decisionali:

- i *classification tree*, utilizzati per determinare l'appartenenza degli elementi di un insieme a classi

diverse;

- i *regression trees*, utilizzati per previsioni relative a un numero reale (per es. il valore di un indice azionario);
- i *classification & regression trees*, che uniscono le due tipologie appena descritte.

L'algoritmo si basa sul concetto di entropia, mutuato dalla teoria dell'informazione. S'ipotizzi di avere un insieme campione *T* di elementi e di voler individuare una regola di suddivisione degli elementi in un numero *k* di classi *Ck*, in base agli attributi degli elementi. L'unica informazione disponibile è la classe di appartenenza degli elementi del campione. Lo scopo è di trovare una regola di classificazione per i nuovi elementi.

L'informazione di un certo sotto insieme *I* è definita dalla seguente equazione da:

$$Info(I) = -\sum_{i=1}^{k}\left(\frac{freq(C_i, I)}{count(I)} \times log_2\left(\frac{freq(C_i, I)}{count(I)}\right)\right)$$

dove:

freq (Ci,I) è il numero di elementi della classe *Ci* presenti in *I*

count(I) è il numero totale di elementi di I

Una volta che l'insieme *T* è stato partizionato in *n* sotto insiemi I_i , secondo i valori di un attributo *x*, è possibile calcolare l'informazione totale:

$$Info_x(T) = -\sum_{i=1}^{n}\left(\frac{count(T_i)}{count(T)} \times Info(T)\right)$$

A questo punto, si può calcolare il guadagno di informazione, che s'ottiene dalla ripartizione in sottoinsiemi, mediante la formula: $Guagagno(X) = Info(T) - Info_x(T)$

Ora, per ogni attributo del nostro data set, occorre individuare quella ripartizione, basata sui valori dell'attributo, per la quale il guadagno è massimo. La ripartizione, che massimizza il guadagno, corrisponde al primo livello dell'albero, sotto l'insieme totale.

Infine, si ripete il processo per ogni sottoinsieme che, al suo interno, presenta elementi appartenenti a classi diverse; il processo s'arresta nel momento in cui i sottoinsiemi contengono elementi solo di una classe, o quando il proseguimento della suddivisione non comporta alcun miglioramento dell'accuratezza.

2.4.6 Classificazione Bayesiana o "Naïve Bayesian Classifier"

Questo approccio si basa sulla quantificazione del problema di decisione, impiegando la probabilità e i relativi costi, che accompagnano tale decisione.

In questa sezione sviluppiamo gli elementi fondamentali di questa teoria, e mostriamo come possa essere considerata semplice la formalizzazione del problema e, di conseguenza, quello della decisione. Il teorema definisce la probabilità condizionata (o a posteriori) di un evento rispetto ad un altro.

$$P(A|B) = \frac{P(B|A) \times P(A)}{P(B)}$$

dove:

P(A|B) è la probabilità condizionata di A rispetto a B

P(B|A) è la probabilità condizionata di A rispetto a B

P(A) è la probabilità "a priori" di A, che non tiene conto di nessuna informazione circa B

P(B) è la probabilità "a priori" di B che non tiene conto di nessuna informazione circa A

Le probabilità "a priori" possono essere stimate mediante la frequenza campionaria, rispettivamente per gli attributi discreti, mentre, per gli attributi continui, si assume che essi siano distribuiti, secondo la distribuzione *normale* e si usa, pertanto, la funzione di densità per il calcolo delle probabilità.

L'algoritmo *naïve bayesian classifier* presuppone che l'effetto di un attributo su una data classe sia indipendente dai valori degli altri attributi. Tale assunzione, chiamata indipendenza condizionale delle classi, è finalizzata a semplificare i calcoli e, per tale motivo, l'algoritmo prende il nome di "naïve".

L'algoritmo determina la classe di appartenenza, secondo le probabilità condizionali per tutte le classi in base agli attributi dei vari elementi. Si ha una classificazione corretta, quando la probabilità condizionale di una certa classe *C* rispetto agli attributi A_n ($P\{C|A1, A2, \ldots, An\}$) è massima.

La probabilità condizionale è data da:

$$P\{Cj\ |A\} = P\{Cj\ |A1, A2, \ldots, An\} = \frac{P\{A1, A2, \ldots, An|Cj\} \times P\{Cj\}}{P\{A1, A2, \ldots, An\}}$$

dato che assumiamo l'indipendenza degli attributi, per ogni classe C_j:

$$P\{A1, A2, \ldots, An|C_j\} = \prod_n P\{An|C_j\}$$

infine, massimizzare $P\{Cj\ |A\}$ equivale a massimizzare il numeratore $P\{A1, A2, \ldots, An|Cj\} \times P\{Cj\}$ espresso anche come:

$$\prod_n P\{An|C_j\} \times P\{Cj\}$$

La classificazione bayesiana incontra però un limite, rappresentato dall'assunzione dell'indipendenza degli attributi, tuttavia, quest'ultimo è superabile mediante l'uso di ulteriori algoritmi, basati sul teorema di Bayes, quali le "*Reti Bayesiane*", o altrimenti, di diversa derivazione, come quelli descritti nei paragrafi precedenti.

2.4.7 Analisi di Serie Temporali

Si definisce "serie temporale" o "serie storica", una sequenza di valori, frutto di misurazioni di una grandezza, effettuate secondo una sequenza temporale. Le misurazioni avvengono, solitamente, a intervalli regolari di tempo (minuti, ore, giorni, mesi, ecc..).

I database, contenenti serie di dati temporali, sono presenti in numerosi ambiti: mercati finanziari, registrazione di fenomeni naturali, dati medici, budgeting, ecc.

Una serie storica può essere analizzata sotto quattro aspetti:

1. *Trend* di lungo periodo: delinea la direzione in cui la serie si muove nel lungo periodo.
2. Movimenti ciclici: definiscono le oscillazioni dei valori attorno al trend principale. Le oscillazioni possono presentarsi a intervalli regolari o avere periodicità diverse.
3. Movimenti stagionali. I movimenti stagionali sono legati a date o periodi dell'anno specifici, durante i quali si registra un deciso incremento o diminuzione dei valori.
4. Movimenti casuali. Esistono poi movimenti legati a eventi casuali, che non sono determinati in alcun modo dalla stagionalità e che non si ripetono ciclicamente.

2.4.7.1 La regressione lineare e multipla

L'analisi delle serie storiche può avvenire mediante diverse tecniche statistiche, tra cui la regressione.

L'analisi della regressione è usata per spiegare la relazione esistente tra una variabile *Y*, detta variabile di risposta o dipendente, e una o più variabili esplicative o indipendenti, denominate anche "covariate", *"(x1, x2,... xk)"*. In termini di funzione, quindi, abbiamo:

$$Y=f(x1, x2,... xk)+\varepsilon$$

quest'ultima indica l'esistenza di un legame funzionale in media tra la variabile dipendente e i regressori, rappresentati dalla componente *"f(x1, x2,... xk)"* e alla quale suole dare il nome di "componente sistematica". A tale componente, va ad aggiungersi un'altra denominata "accidentale", casuale o erronea. Mentre la prima rappresenta la parte della variabile di risposta spiegata dai predittori, la seconda componente costituisce quella parte di variabilità della risposta che non può ricondursi a fattori sistematici oppure facilmente individuabili, ma dovuti al caso o, più in generale, a cause diverse non prese in considerazione nel modello di regressione.

In linea teorica, il legame funzionale può essere di qualsiasi tipo, tuttavia, nella prassi statistica, si preferisce utilizzare una funzione di tipo lineare e, pertanto, si definisce "regressione lineare o multipla" ed assume la seguente formulazione:

$$Y=\beta0+\beta1X1+...+\beta kXk +\varepsilon$$

dove $\beta0$ è detto termine noto, mentre $\beta1X1+...+\beta kXk$ sono detti coefficienti di regressione che, insieme alla varianza dell'errore ε, sono i parametri del modello da stimare, sulla base delle osservazioni campionarie.

Infine, nel caso in cui *Y* non dipende solo da un fattore, ma da un insieme di variabili esplicative ("*X1, X2...Xn*"), si ha la cosiddetta "regressione multipla", dove:

$$Y= \beta0+\beta1X1+\beta2X2 \ ...+\varepsilon$$

2.4.7.2 "Exponential Smoothing" o lisciamento esponenziale

L'"*Exponential Smoothing*" costituisce un essenziale strumento di previsione puntuale, in particolare quando si hanno a disposizione un dataset ridotto. Tale procedura presuppone che una ragionevole previsione del valore di una serie *X* al tempo *t* possa essere definita da una combinazione lineare della previsione operata sulla stessa serie nell'istante precedente. Tuttavia, tale combinazione lineare deve tener conto della variazione registrata nell'unità temporale precedente tra l'effettivo valore della serie e la previsione realizzata. Quindi, si ha:

$$Fn,1 = \delta \cdot Fn-1,1 + (1-\delta) \cdot yn$$

ove la nuova previsione (al tempo $n+1$) può essere considerata come la media pesata fra l'osservazione al tempo n (ultima osservazione disponibile) e la previsione precedente (relativa al tempo n, formulata al tempo $n-1$). Il valore assegnato al parametro δ è la chiave di lettura dell'analisi di lisciamento esponenziale. Se si desidera che la previsione sia stabile e che le variazioni casuali della serie siano "smussate", allora si sceglie un valore per δ vicino ad 1 e, viceversa, se si vuole attribuire maggior peso alle osservazioni più recenti, si opta per un valore vicino a 0.

Un'ulteriore tecnica di previsione, particolarmente proficua quando si devono modellare le serie storiche composte da una componente di trend, ed eventualmente, da un fattore stagionale, consiste nella metodologia "*Holt-Winters*". Questo metodo costituisce la naturale generalizzazione della tecnica del lisciamento esponenziale e, secondo tale approccio, la serie storica è il risultato di tre componenti: il livello al tempo n (Yn), il trend (Tn) e la stagionalità (Sn). Esistono due metodi stagionali proposti da Holt-Winters: additivo e moltiplicativo.

2.4.7.3 L'approccio di Box Jenkins

Nell'analisi di fenomeni socio-economici, non controllabili sperimentalmente, come il numero di donatori, l'ammontare lordo delle donazioni di una organizzazione non profit etc., il modello per il processo generatore dei dati, impiegato per inferire l'incognito, deve comprendere necessariamente una parte probabilistica εt, per generalizzare la parte deterministica.

I modelli, comunemente impiegati nell'analisi moderna delle serie storiche "univariate", sono quelli introdotti da Box e Jenkins (Box G. E., Jenkins G. M, 1976).

I due modelli fondamentali, proposti da questi studiosi, sono: il modello *autoregressivo* ("*AR*", acronimo di "*AutoRegressive*"), che armonizza una somma pesata di valori passati e uno shock casuale contemporaneo, ed il modello a *media mobile* ("*MA*" sta appunto per "*Moving Average*") che è il risultato di una serie di impulsi casuali.

Dalla combinazione di questi due modelli, discende il modello "*ARMA*" (*AutoRegressive Moving*

Average) ed assume che la variabile *Yt* dipenda linearmente sia da *p*, tempi precedenti della variabile stessa (componente "autoregressiva"), sia dai valori passati del termine di errore ε (componente a "media mobile") *q*. Dunque, il modello cosi descritto può essere formalizzato nell'espressione:

$$\varphi(B) \cdot Yt = \theta(B) \cdot \varepsilon t$$

dove:

"*B*" è l'operatore ritardo t.c. "$B \cdot yt = yt-1$";

$\varphi(B)=1-\varphi B-...-\varphi Bp\ 1$, è l'operatore autoregressivo non stagionale di ordine "*q*";

$\theta(B)=1+\theta B+...+\theta Bq\ 1$, è l'operatore a media mobile non stagionale di ordine "*q*";

εt è un processo "*white noise*" ($\varepsilon t \sim WN(0,\sigma 2)$).

L'interesse di armonizzare un modello *AR*(*p*) e un modello *MA*(*q*) in un modello *ARMA*(*p,q*) è nato dall'esigenza di descrivere, mediante un esiguo numero di parametri, una serie storica.

Sulla scia dei risultati ottenuti dal modello *ARMA*, Box e Jenkins estesero i processi *ARMA* ai modelli "*ARIMA*" (*p,d,q*) ("*Autoregressive Integrated Moving Average*"), che consentono di modellare i "processi non stazionari", ovvero quei processi che non mantengono la media o la variabilità costante. Una serie non stazionaria può essere resa stazionaria tramite *d* differenziazioni oppure tramite trasformazione, ad esempio quella logaritmica, qualora la serie non sia stazionaria in varianza.

Dato quindi un intero non negativo *d*, il processo *ARIMA(p,d,q)* può essere espresso come:

$$\varphi(B) \cdot (1-B)d \cdot Y = \theta(B) \cdot \varepsilon\, t$$

Se *d* è uguale a zero la serie non necessita di essere differenziata e si ritorna alla classe *ARMA*(*p,q*).

2.4.8 Reti neurali

Le "*reti neurali artificiali*" (*RNA*) sono modelli di calcolo che replicano, in maniera verosimile, il funzionamento del cervello umano. Così come il nostro cervello è formato da neuroni interconnessi da legami chiamati sinapsi, le RNA sono costituite da unità di calcolo (o neuroni artificiali) e da connessioni. Le RNA possono essere raffigurate con grafi, i cui nodi sono i neuroni e i cui archi sono le interconnessioni.

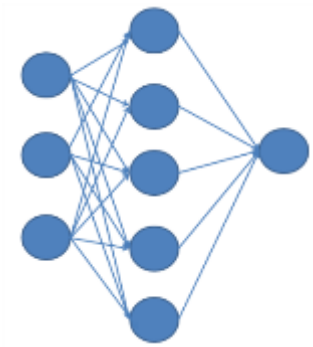


Fig 2.8 Grafo di una rete neurale

Ciascun nodo rappresenta un'unità di calcolo adattiva, in quanto il proprio output dipende da parametri modificabili. I calcoli eseguiti nei nodi possono essere lineari o non lineari, ed è proprio quest'ultima caratteristica che rappresenta uno dei tratti peculiari di questi algoritmi, assieme alla capacità di apprendere da campioni di dati. I neuroni artificiali, infatti, sono in grado di variare i propri parametri di calcolo, in base ai dati di training: a seguito del processo di apprendimento, la RNA formula un insieme di parametri ottimali che indicano la conoscenza del problema analizzato.

Le reti neurali, grazie alla loro flessibilità, si adattano a numerosi tipi di problemi, quali, per esempio: *analisi di marketing e di promozioni*, s*tima di fluttuazioni dei mercati finanziari*, *analisi di processi di produzione e industriali*, *diagnosi mediche*, *text mining, fraud detection.*

È possibile rappresentare la RNA come un insieme di strati (o *layers*) di neuroni.

Nella struttura della rete, distinguiamo tre tipi di *layer*:

- il layer di ingresso, formato dai neuroni di input, attraverso i quali sono forniti di dati;
- uno o più layer intermedi (o nascosti), che eseguono l'elaborazioni dei dati;
- un layer di output, che fornisce il risultato.

Ogni neurone ha uno stato di attivazione, rappresentato da un valore continuo compreso tra un massimo e un minimo, solitamente 0 e 1. Parimenti, i collegamenti tra neuroni possiedono un valore associato, detto peso sinaptico, avente lo scopo di amplificare o ridurre l'importanza che un dato neurone ha all'interno della rete.

Ciascun neurone riceve in ingresso la somma pesata dei pesi sinaptici e dei valori di attivazione degli altri neuroni ad esso collegati. Il singolo neurone calcola il proprio valore di attivazione, trasmesso, in seguito, al livello successivo della rete, per mezzo di una funzione di attivazione:

$$a_i = f\left(\sum_j w_{i\,j} \times a_j\right)$$

dove:

a_i= valore di attivazione del neurone *i*

f = funzione di attivazione

w_{ij}= peso sinaptico del j-esimo neurone collegato al neurone *i*

a_j = valore di output del neurone j-esimo

Esistono poi, diverse funzioni impiegate come funzione di attivazione, una tra le più popolari è la funzione logistica o sigmoide, già esposta in precedenza, la quale presenta un dominio dei valori di output compreso tra 0 e 1 (la tecnica di preparazione dei dati impiegata è detta normalizzazione, ed evita che attributi con valori molto elevati diventino predominanti rispetto ad altri con valori più piccoli). È importante che il layer di input riceva come primi dati in ingresso nella rete dei valori compresi nel dominio della funzione di attivazione.

La rete deve subire un processo di training, al fine di determinare i corretti pesi sinaptici, non noti a priori. L'apprendimento, utilizzando i dati del campione di training, consiste nella ricerca di un minimo in uno spazio a *n* dimensioni di una funzione, che esprime l'errore totale della rete, espresso come errore quadratico medio:

$$Err.tot = \frac{1}{2}\sum_i |d_i - y_i|^2$$

dove:

d_i è il valore desiderato;

y_i è il valore di output.

Per minimizzare l'errore totale, è opportuno modificare i valori dei rispettivi pesi sinaptici, impiegando la tecnica della discesa del gradiente, ossia operando piccoli spostamenti sulla funzione proporzionali alla sua derivata e di segno opposto, tentando di ottenere la convergenza a un minimo.

Successivamente, al fine di determinare le correzioni ai pesi, si utilizza la tecnica della back propagation, con la quale l'errore, emerso nel layer di output, è propagato all'indietro per determinare l'errore di ciascun neurone a ogni livello, utilizzando i pesi sinaptici per distribuire l'errore.

I pesi sono modificati con la seguente formula:

$$w_{i\,j} = w_{i\,j} + \varepsilon \times a_j \times \Delta_i$$

dove:

w_{ij} è il peso sinaptico utilizzato per l'input i del neurone j

ε è una costante, chiamata fattore di apprendimento; da essa dipende la dimensione degli spostamenti dei pesi

a_j è il valore di input j-esimo del neurone i

Δ_i è ottenuto dal prodotto dell'errore del neurone i-esimo con la derivata prima della funzione di attivazione, valorizzata con l'input del neurone stesso, ovvero:

Nella formula del delta, Err_i è dato dalla differenza tra il valore desiderato (ossia quello noto, proveniente dal campione di training) e il valore di output del neurone. Questa misurazione è valida soltanto per i neuroni di output, mentre l'errore dei layer intermedi è da calcolarsi differentemente. Infatti, per lo strato intermedio, l'errore Err_i del neurone i è dato dalla somma dei delta dello strato di output, pesata coi pesi sinaptici che lo collegano ai neuroni dello strato successivo:

$$Err_i = \sum_j w_{i\,j} \times \Delta_j$$

dove:

w_{ij} è il peso con cui il neurone i è collegato al neurone j dello strato successivo

Δ_j è il Δ del neurone j dello strato successivo

Il processo è ripetuto, finché l'errore totale sarà portato sotto a una soglia prestabilita.

Le fasi sono schematizzate nell'elenco che segue:

1. s'impiegano, come valori di ingresso, i dati del training set, e successivamente si calcolano i valori di ciascun neurone per ogni layer, fino ad ottenere il valore di output;
2. si procede al confronto tra il valore di output e il valore desiderato e si calcola l'errore totale;
3. si esegue la back propagation calcolando i valori di delta da applicare ai pesi;
4. si ripete il calcolo coi nuovi pesi e si ottiene il nuovo valore di output;
5. s'itera il processo fino a portare l'errore, al di sotto di una soglia prestabilita.

CAPITOLO 3

3.1 L'esigenza di un modello procedurale uniforme per l'industria DM

Nelle logiche di mercato, il *data mining* è considerato una tecnologia *push-button.* Tuttavia, in ambito scientifico, il data mining è un processo complesso, che richiede strumenti diversi e, allo stesso tempo, individui altamente qualificati. Il successo di un progetto di data mining dipende dalla giusta combinazione fra questi elementi; inoltre, si esigono una metodologia affidabile ed un *project management* efficace. Dunque, un modello procedurale uniforme può aiutare a comprendere e gestire le relazioni ed interazioni tra variabili differenti, lungo questo processo complesso.

Qualora vi fosse un modello procedurale condiviso dalla comunità scientifica, il mercato ne beneficerebbe indubbiamente, ergendo il *data mining* a una pratica ingegneristica consolidata e, conseguentemente, diventando un punto di riferimento comune per le aziende.

In pratica, i consumatori, avendo a che fare con tali strumenti e *service providers*, saranno in grado di scegliere il bene o servizio più confacente ai loro bisogni o desideri. Dall'altra parte, i produttori si avvarranno indirettamente dei vantaggi arrecati ai consumatori. Peraltro, i venditori potranno aggiungere valore alle loro merci (ad esempio, fornendo assistenza nel corso del processo) ed accrescere le qualifiche del personale.

Infine, perfino i *data miners' analists* potranno assicurarsi notevoli vantaggi da un modello procedurale uniforme di data mining. Esso costituirà una guida per coloro che muoveranno i primi passi in questa disciplina, aiutando a impostare il progetto e dando suggerimenti su ogni singola *task* (compito) del processo di data mining, mentre, per gli analisti più esperti, rappresenterà un modello di riferimento su cui potranno basare le loro ricerche, evitando che nessun compito sia trascurato o dimenticato.

Per concludere, è da evidenziare che un processo uniforme costituirà un anello di congiunzione fra differenti strumenti d'analisi e professionisti aventi competenze e backgrounds completamente diversi.

3.2 CRISP-DM: Towards a Standard Process Model for Data Mining

Il data mining è un processo creativo che richiede una serie di conoscenze e competenze. Attualmente, non vi è una cornice teorica uniforme che consenta di realizzare progetti di data mining. Ciò implica necessariamente che il successo o il fallimento di un progetto di data mining è da attribuirsi alla singola persona o al team che lo ha concepito e, inoltre, un programma di DM di successo è da considerarsi *one-shot*, ossia non potrà essere ripetuto più volte all'interno della stessa azienda. Il data mining ha bisogno di un approccio standardizzato che aiuti a tradurre i problemi di *business* in compiti di data mining, suggerendo appropriate interpretazioni dei dati e tecniche di data mining e, infine, fornendo gli strumenti necessari per valutare l'efficacia dei risultati e documenti l'esperienza in corso.

Il progetto CRISP-DM (Cross Industry Standard Process for Data Mining) propone un modello procedurale comprensivo ed intuitivo al fine di realizzare progetti di data mining.

Esso è indipendente sia dal settore industriale e dalla tecnologia impiegata. Lo scopo del CRISP-DM è di creare elaborati progetti di data mining maggiormente affidabili, reiterabili, gestibili, rapidi e a costi competitivi.

3.2.1 La Metodologia Cross Industry Standard Process for Data Mining (CRISP-DM)

La metodologia CRISP-DM si tratta di un modello procedurale di carattere gerarchico, composto da una serie di compiti (*tasks*) ordinati su quattro livelli di astrazione (dal generale al particolare): *phase*, *generic task*, *specialized task* e *process instance*.

Al vertice, il processo di data mining è organizzato in un determinato numero di fasi; ogni fase consiste di alcuni compiti generici di secondo piano.

Il secondo livello è definito "generico" perché si prefigge di disciplinare tutte le possibili situazioni di data mining. I compiti generici sono caratterizzati da completezza e stabilità. La completezza indica che intende regolare sia l'intero processo di data mining sia tutte le possibili applicazioni di data mining. Mentre, la stabilità suggerisce che il modello debba essere valido, nonostante gli sviluppi inattesi che la disciplina potrà manifestare (ad esempio nuove tecniche di realizzazione di modelli).

Successivamente, il terzo livello, lo *specialized task level*, si propone di descrivere il modo in cui le operazioni/azioni nei *generic tasks* debbono essere effettuate in situazioni specifiche. Ad esempio, nel corso del secondo livello esiste un *generic task*, propriamente detto *build model*; mentre, durante il terzo livello, è probabile che abbia luogo un'operazione chiamata *build response model,* la quale include attività specifiche ed attinenti al problema ed allo strumento di data mining selezionato.

L'identificazione e la descrizione delle fasi e dei compiti in passaggi distinti e compiuti in un determinato ordine rappresenta un'ideale sequenza di eventi.

Nella prassi, gran parte delle tasks possono essere eseguite in un ordine differente e, spesso, sarà necessario riesaminare compiti precedenti e ripetere determinate azioni. Il modello procedurale CRISP-DM non cerca di immortalare tutti questi passaggi, mediante il processo di data mining, in quanto ciò richiederebbe un modello statistico estremamente complesso, comportando in questo modo una progressiva riduzione dei benefici attesi.

Infine, il quarto livello, il *process instance level*, è un record/documentazione/testimonianza delle azioni, delle decisioni e dei risultati di una singola operazione (engagement) di data mining (corrente). Un *process instance* è organizzato secondo compiti circoscritti a livelli più elevati, ma ciò rappresenta cos'è effettivamente successo in un singolo procedimento, piuttosto che testimoniare cos'è accaduto in generale.

La metodologia CRISP-DM è distinta graficamente in due categorie, il *Reference Model* e l'*User Guide*. Il *Reference Model* presenta una rapida visione d'insieme elencando le fasi, i compiti e i loro risultati e descrive *cosa fare,* nel corso del progetto di data mining; invece, l'User Guide offre, da un lato, consigli puntuali per ogni singola fase e compito all'interno di essa e, dall'altro lato, raffigura *come realizzare*

effettivamente un progetto di data mining.

3.2.2 Le Fasi della metodologia Cross Industry Process for Data Mining

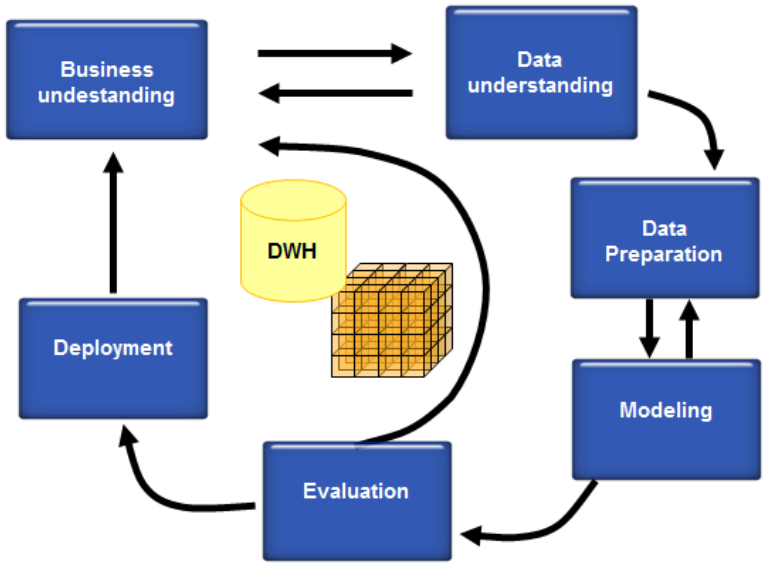


Fig. 3.1 La metodologia *Cross Industry Process for Data Mining (CRISP-DM)*

Il CRISP-DM, modello di riferimento per il data mining, fornisce una visione d'insieme del ciclo di vita di un progetto di DM. Esso include le fasi di un progetto, i loro rispettivi compiti e risultati.

Il ciclo di vita di un progetto di data mining è suddiviso in sei fasi che sono illustrate nella Figura 3.1.

L'ordine delle fasi è trascurabile. Le frecce indicano solo le relazioni fra fasi più importanti e frequenti, tuttavia, in un particolare progetto, tali corrispondenze dipendono sia dal risultato di ogni singola fase sia da quale fase dovrà essere effettuata successivamente.

Nella Figura 3.1, il cerchio esterno simboleggia la natura ciclica del data mining stesso. Il data mining non è terminato una volta che è stata disposta una soluzione. Le lezioni imparate durante il processo e dalla soluzione prevista possono innescare nuove questioni di business.

I successivi processi di data mining beneficeranno dalle esperienze precedenti.

Ora, delineiamo ciascuna fase del modello procedurale del CRISP-DM:

1. *Business Understanding*: lo stadio iniziale si focalizza sulla comprensione degli obiettivi e delle priorità del progetto DM da un punto di vista aziendale, in seguito, le informazioni acquisite si

traducono nella definizione di un problema di data mining, e finalmente, si procede ad una stesura di un *project plan* preliminare, affinché possano essere raggiunti gli obiettivi prefissi.

Task 1 - Determinare gli obiettivi di business: il primo obiettivo del *data analyst* consiste nel comprendere esaurientemente, da un punto di vista aziendale, cosa realmente desidera il cliente. Egli ha molti obiettivi contrastanti e dovranno essere bilanciati correttamente. L'obiettivo primario dell'analista consiste, quindi, nel rivelare i fattori decisivi che influenzeranno il risultato del progetto.

- *Obiettivi di business*: descrivere l'obiettivo primario del cliente, da un punto di vista imprenditoriale. Inoltre, vi possono essere ulteriori questioni di business, strettamente correlate con esso (all'obiettivo primario il cliente), che il cliente vorrebbe affrontare. Ad esempio, l'obiettivo primario di business potrebbe consistere nel preservare la clientela attuale, prevedendo quando essi sono propensi a passare alla concorrenza.

- *Realizzare il project plan*: descrivere il piano d'azione, per raggiungere gli obiettivi di data mining e, in tal modo, conseguire gli obiettivi aziendali. Il programma dovrebbe specificare tutte le fasi, che dovranno essere compiute nel corso del progetto, includendo la selezione iniziale degli strumenti e delle tecniche di DM impiegate.

- *Business success criteria*: presentare il criterio per valutare se il risultato del progetto sia utile dal punto di vista aziendale o meno.

Task 2 – Valutare la situazione: questo compito comporta una raccolta dati maggiormente dettagliata riguardo tutte le risorse, vincoli, assunzioni, ed altri elementi, che dovrebbero essere considerati per la definizione del project plan e dell'obiettivo dell'analisi dei dati.

- *Inventario delle risorse*: elencare le risorse disponibili per il progetto, quali il personale (esperti di business, esperti di dati, supporto tecnico, esperti di data mining), i dati (accesso a dati operativi o immagazzinati), i softwares (strumenti di data mining o ulteriori softwares rilevanti) e le risorse informatiche (piattaforme hardware).

- *Requisiti, assunzioni, e limiti*: elencare tutti i requisiti del progetto, includendo il programma di completamento (del progetto), comprensibilità e qualità dei dati e sicurezza. Successivamente, occorre enumerare le supposizioni sui dati, ideate dal progetto. Esse potranno essere verificate nel

corso del processo di DM, tuttavia, in presenza di assunzioni di carattere aziendale non verificabili e associabili al progetto, è importante individuarle e classificarle, qualora incidessero sulla validità dei risultati conseguenti. Infine, potrebbero emergere vincoli sulla disponibilità delle risorse e, allo stesso tempo, limiti tecnologici come ad esempio la dimensione del *dataset*.

- *Rischi e contingenze*: elencare i rischi o gli eventi che potrebbero ritardare o addirittura cagionare il fallimento del progetto. Dunque, è necessario predisporre piani d'azione contingenti, qualora dovessero manifestarsi potenziali rischi o eventi aleatori.

- *Costi e benefici*: elaborare un'analisi comparativa tra i costi del progetto e i potenziali benefici del business.

Task 3 – Definire gli obiettivi di Data Mining: un obiettivo di data mining illustra, in termini tecnici, gli obiettivi di un progetto. Ad esempio, un obiettivo di business potrebbe essere "incrementare le vendite dei clienti attuali". L'obiettivo corrispondente di data mining potrebbe essere "prevedere quanti articoli acquisterà un cliente, dati i loro acquisti durante gli ultimi tre anni, il prezzo dell'oggetto…".

- *Data Mining success criteria*: stabilire, in termini tecnici, un criterio valutativo del risultato che è stato conseguito, relativamente al progetto. Ad esempio, un determinato livello di accuratezza predittiva o un profilo di propensione all'acquisto per un dato grado di "lift".

Task 4 – Realizzare il project plan

- *Project Plan*: elencare le fasi che dovranno essere eseguite nel corso del progetto, considerando la loro durata, le risorse richieste, gli inputs, gli outputs e le *dependencies*. Laddove è possibile, è necessario rendere manifeste le reiterazioni su larga scala (*large-scale iterations*) nel processo di DM (per esempio, le ripetizioni delle fasi di modellazione e valutazione). Quale parte del project plan, è importante analizzare le dipendenze fra scadenze e rischi, che evidenzia esplicitamente i risultati di queste analisi nel project plan con azioni o raccomandazioni, qualora si manifestino eventuali rischi. Infine, il project plan è un documento in "continua evoluzione", pertanto ,è necessario che, al termine di ogni fase, vi sia un aggiornamento dei progressi e dei risultati raggiunti.

- *Valutazione iniziale degli strumenti e delle tecniche di Data Mining*: al termine della prima fase,

bisogna approntare una valutazione iniziale degli strumenti e delle tecniche impiegate dagli analisti, in quanto incideranno sull'intero progetto.

2. *Data Understanding*: individuati gli obiettivi del progetto di DM, ciò di cui disponiamo, per il raggiungimento di tali obiettivi, è rappresentato dai dati. La fase della comprensione dei dati prevede la raccolta iniziale dei dati e una serie di attività parallele sui dati stessi, al fine di acquisire familiarità con quest'ultimi, identificare problemi inerenti la qualità dei dati, esaminare le prime rilevazioni effettuate sui dati (calcolo delle statistiche di base, come ad esempio, medie, indici di variabilità…), o localizzare sottoinsiemi interessanti, tali da costituire ipotesi per informazioni nascoste.

   Task 1 – Raccolta iniziale dei dati: acquisire i dati (*access to the data*) elencati nelle risorse del progetto. Questa raccolta iniziale (di dati) include anche il caricamento dei dati, qualora fosse necessario per il data understanding. Ad esempio, se si impiega uno strumento specifico per la comprensione dei dati (data understanding), occorrerà caricare i dati all'interno di questo strumento (*tool*). Si noti che nel caso in cui si acquisissero fonti multiple di dati, la fase di integrazione (dei dati) è una ulteriore issue da tenere in considerazione.

   - *Report della raccolta iniziale di dati*: elencare le fonti di dati acquisiti, la loro provenienza, i metodi impiegati per acquisirli e qualsiasi problema cui si è imbattuti. Tale compito faciliterà sia l'interazione futura del progetto sia l'esecuzione di simili progetti futuri.

   Task 2 – Descrizione dei dati: esaminare le caratteristiche complessive dei dati acquisiti e il documento (*report*) che riporta i risultati.

   - *Report sulla rappresentazione (o descrizione) dei dati*: descrivere i dati che sono stati raccolti, includendo il formato dei dati, la quantità di dati (ad esempio, il numero di *records* e *fields* in ogni tabella), le generalità dei fields, e qualsiasi altra caratteristica di superficie (*surface features*) che è stata scoperta. Infine, è bene valutare se i dati acquisiti soddisfino i requisiti previsti.

   Task 3 – Esplorazione dei dati: questa task affronta questioni di data mining, che impiegano tecniche di *querying*, di visualizzazione e di *reporting*. Esse includono la distribuzione delle variabili chiave (ad esempio, il *target attribute* di una task sulla predizione), le relazioni fra coppie o un piccolo numero di variabili, i risultati delle aggregazioni semplici, le caratteristiche di sotto-popolazioni significative e semplici analisi statistiche. Tali analisi potrebbero indirizzarsi direttamente ad obiettivi di DM; oppure potrebbero contribuire a migliorare o raffinare la qualità dei reports e la descrizione dei dati, dando vita alla trasformazione (dei dati) e tracciando gli steps

di preparazione dei dati, preziosi per ulteriori analisi.

- *Report sull'esplorazione dei dati*: esaminare i risultati raggiunti in questa task, includendo le prime rilevazioni o ipotesi iniziali sui dati ed il loro impatto sull'avanzamento del progetto. È possibile inserire grafici e disegni per indicare le caratteristiche dei dati, che potranno essere soggette ad ulteriori esami.

Task 4 – Verifica della qualità dei dati: occorre esaminare la qualità dei dati, formulando tali questioni fondamentali: i dati sono completi, ovvero richiamano tutti i casi rilevanti? Sono corrette contengono errori? Se presentano errori, con quale frequenza si manifestano? Vi sono valori mancanti (*missing values*) nei dati? Se sono presenti missing values, come sono rappresentati, ove si verificano e con quale frequenza (si manifestano)?

- *Report sulla qualità dei dati*: elencare i risultati della verifica della qualità dei dati; nel caso in cui si presentano problemi sulla qualità (dei dati), bisogna provvedere a formulare le possibili soluzioni, condizionate fortemente dai dati e dalla conoscenza del business.

Vi è uno stretto legame fra il Business Understanding ed il Data Understanding. La formulazione del problema di data mining e del project plan richiedono almeno la comprensione dei dati a disposizione.

3. *Data Preparation*: lo stage di preparazione dei dati include tutte quelle attività, che rendono possibile la realizzazione del *dataset* finale (ossia quell'insieme di dati "estratti" dall'insieme iniziale di dati grezzi, che, naturalmente, svilupperà gli strumenti di modellazione/*modeling*). La fase di preparazione dei dati dovranno probabilmente essere compiuti più volte e senza un ordine prestabilito. Le tasks includono prospetti, records, selezione delle proprietà (*attributes*), pulizia dei dati, costruzioni di nuovi attributi o variabili e trasformazione dei dati, compatibili con gli strumenti di modellazione.

Task 1 – Selezione dei dati: decidere quali dati analizzare. Il criterio, che è alla base di tale scelta, tiene in considerazione la rilevanza degli obiettivi di data mining, la qualità, i vincoli tecnici, come, ad esempio, i limiti sul volume o tipologia di dati. Si noti che la *data selection* riguarda sia la selezione delle variabili (colonne) sia la selezione dei records (file) in una tabella.

- *Base logica per l'inclusione e l'esclusione dei dati*: lo scopo primario di tale compito consiste nel rubricare i dati, al fine di essere esclusi inclusi nell'analisi, giustificando queste scelte.

Task 2 – Pulizia dei dati: migliorare la qualità dei dati, elevandola al livello richiesto dalle tecniche di analisi selezionate. Ciò potrebbe comportare la pulizia di sottoinsiemi di dati, l'inserimento d'impostazioni predefinite adeguate, o l'impiego di tecniche maggiormente ambiziose come, ad esempio, la stima dei dati mancanti (*missing data*), mediante tecniche di modellazione.

- *Report sulla pulizia dei dati*: tale report include la descrizione di quali decisioni e azioni debbano essere prese, per affrontare i problemi riguardanti la qualità dei dati emersi durante la fase di Data Understanding. È necessario tenere in considerazione l'impatto dell'analisi (statistica) delle trasformazioni dei dati operate per scopi di ripulitura (dei dati stessi) sui risultati ottenuti.

Task 3 – Costruzione dei dati: tale compito include operazioni essenziali per la preparazione dei dati, come, ad esempio, la produzione di variabili derivate, di records completi o di valori trasformati per le variabili preesistenti.

- *Variabili derivate*: le variabili derivate sono nuove variabili, costruite da una o più variabili esistenti nel medesimo record. Ad esempio: Area= lunghezza x ampiezza.

- *Records generati*: tale risultato implica la descrizione dei processi di ideazione e creazione di nuovi records. Ad esempio, creare records per i clienti che non hanno compiuto alcun acquisto, nel corso dell'anno precedente; ai fini della fase di modellazione, è necessario prendere in esame anche i clienti, che non hanno fatto acquisti.

Task 4 – Integrazione dei dati: vi sono metodi che consentono di combinare le informazioni ottenute con databases multipli, tabelle o records, al fine di creare nuovi valori o records.

- *Unione o fusione di dati*: le *merging tables* consentono l'unione di due o più tabelle, che possiedono differenti informazioni riguardo i medesimi oggetti. Ad esempio: una catena di vendita al dettaglio possiede una tabella che illustra le caratteristiche generali di ogni singolo negozio (es.: tipologia di…), una seconda (tabella) che presenta i dati delle vendite (profitto…) e, infine, una terza (tabella) che mostra i dati demografici della zona circostante. Ciascuna di queste tabelle contiene un record per ogni singolo negozio. Queste tabelle possono essere fuse insieme in una nuova tabella con un record per ogni store, combinando diversi settori (*fields*), partendo dalle tabelle di riferimento.

  L'unione o fusione di dati si occupa anche delle *aggregations*. Le *aggregations* sono operazioni, i

cui valori sono calcolati in base ad informazioni riassuntive, provenienti da records multipli e/o tabelle.

4. *Modeling*: nel corso di questa fase, sono selezionate ed applicate una serie di tecniche di modellazione, i cui parametri sono calibrati su valori ottimali per ottenere modelli statistici. Tipicamente, vi sono differenti tecniche per un unico tipo o genere di problema di data mining. Infine, alcune tecniche richiedono specifici formati di dati (*data formats*), per cui è spesso opportuno ripetere la fase di preparazione dei dati, per modificare il data-set iniziale e adattarlo alla tecnica specifica che si vuole utilizzare.

   Task 1 – Selezione delle tecniche di modellazione: nel corso del primo stadio della modellazione, è necessario selezionare la tecnica specifica di modellazione che verrà impiegata, ad esempio, la costruzione di alberi decisionali mediante il software C5.0. Nel caso in cui si debbano applicare più tecniche di modellazione, bisogna eseguire questa task separatamente per ogni singola tecnica (di modellazione selezionata).

   - *Supposizioni formulate dalle tecniche di modellazione*: diverse tecniche di modellazione formulano specifiche supposizioni riguardo i dati, ad esempio, tutte le variabili presentano valori uniformi, non sono consentiti *missing values* ecc… È necessario, quindi, registrare qualsiasi tipo di supposizione che viene formulata.

   Task 2 – Generare un test design: prima di costruire il modello, è bene progettare una procedura o un meccanismo, che verifichi la qualità e la validità del modello. Ad esempio, nel corso di compiti di data mining (controllati), come la classificazione, si suole impiegare tassi di errore (*error rates*) che equivalgono a *quality measures* per il modelli di DM. Dunque, si separa tipicamente il *data set* in *file* e *test sets* e, successivamente, si costruisce il modello sul *train set* di riferimento, mentre si valuta la sua qualità attraverso un *test set indipendente*.

   - *Test design*: il test design presenta un piano d'azione, ideato per la verifica e la valutazione dei modelli. Una componente rilevante di questo piano si occupa di determinare in che modo dividere il dataset in *training*, *test* e *validation datasets*.

   Task 3 – Costruzione del modello: applicare lo strumento di modellazione su un predisposto *data set*, al fine di creare uno o più modelli.

   - *Impostazioni del parametro*: per qualsiasi strumento di modellazione, vi sono, spesso, un numero

elevato di parametri che possono essere coordinati. Dunque, occorre elencare i parametri e scegliere quali valori devono assumere, in conformità con le impostazioni del parametro selezionate.

- *Descrizioni del modello*: descrivere i risultati del modello. Registrare la spiegazione di ogni singolo modello e documentare dettagliatamente qualsiasi difficoltà incontrata.

Task 4 – Valutazione del modello: lo specialista di data mining (il *data scientist*) interpreta i modelli in base al proprio settore di competenza, ai criteri chiave di successo del data mining e al *test design* realizzato. L'analista valuta il successo dell'applicazione della modellazione; successivamente, contatta i business analysts ed gli esperti del settore per confrontarsi sui risultati del DM in un contesto aziendale (di business).

In primo luogo, lo specialista di data mining tenta di classificare i modelli, giudicandoli secondo specifici criteri di valutazione. In secondo luogo, egli tiene in considerazione gli obiettivi aziendali e i criteri chiave del successo di un business. Infine, nella maggior parte dei progetti di DM, l'analista applica una singola tecnica più di una volta oppure produce risultati di data mining con numerose tecniche di modellazione, confrontando quest'ultimi secondo specifici criteri di valutazione.

- *Risultati sulla valutazione del modello*: sintetizzare i risultati, enumerare le qualità dei modelli generati (ad esempio, in termini di accuratezza) e classificare i modelli, in termini qualitativi.

- *Configurazioni del parametro riveduto e corretto*: secondo quanto stabilito dalla task riguardante la valutazione del modello (*model assessment*), si procede alla revisione delle configurazioni del parametro e, in seguito, a regolarli adeguatamente per la fase successiva. La costruzione e valutazione di un modello sono processi iterati, di conseguenza, sarà necessario documentare ogni singola revisione e valutazione (del modello).

Intercorre uno stretto legame fra la Data Preparation e il Modeling. È frequente che la prima formalizzi i cosiddetti *data problems,* mentre entrambe saranno preziose per la costruzione di nuovi dati.

5. *Evaluation*: nel corso di questo stage del progetto, abbiamo costruito uno o più modelli che sembrano presentare una qualità elevata, da un punto di vista di analisi dei dati. Prima di procedere all'impiego finale del modello, occorre valutarlo accuratamente e controllare gli stadi precedenti, che hanno condotto alla realizzazione del modello, affinché venga assicurato che quest'ultimo consegua pienamente gli obiettivi di business previsti. È evidente che, a questo punto, sarà importante identificare potenziali *business issues* che non sono state approfondite sufficientemente. Al termine

di questa fase, sarà importante assumere una decisione sull'impiego dei risultati raggiunti dal data mining.

Task 1 – Valutazione dei risultati: nel corso di questa fase, si valuta il grado di conformità del modello con gli obiettivi di business aziendali, tentando di determinare se il modello presenta lacune o meno. Inoltre, la fase di *Evaluation* si occupa di giudicare gli ulteriori risultati di data mining generati. Questi ultimi si richiamano a modelli che sono necessariamente legati agli obiettivi aziendali originali e ad ulteriori scoperte o esiti, che potrebbero risultare determinanti in futuro.

- *Valutazione dei risultati di data mining con riguardo ai criteri di successo del business*: sintesi dell'analisi condotta sulla valutazione dei risultati del processo di data mining, con riguardo ai criteri di successo del business (*assessment of data mining results with respect to business success criteria*), includendo una dichiarazione formale sulla conformità del progetto con gli obiettivi iniziali del business.

- *Modelli approvati*: i modelli che avranno superato la fase di valutazione e che, di conseguenza, saranno compatibili con i criteri stabiliti dagli analisti, potranno essere considerati approvati (*models approved*).

Task 2 – Processo/attività di revisione: a questo punto, i modelli, approvati dagli esperti, possono essere ritenuti soddisfacenti e idonei alle necessità dell'azienda. Ora, è necessaria un'accurata revisione del processo complessivo di data mining, perché si accerti che nessun fattore o task sia stata trascurata.

Task 3 – Definizione steps successivi: dopo aver effettuato la valutazione dei risultati di DM e l'attività di revisione, il *project team* valuta se sia il caso di terminare questo progetto, procedendo all'applicazione dei risultati, oppure di impostare nuovi progetti di data mining. Questa task include un'analisi approfondita delle risorse rimanenti e del budget, i cui risultati potrebbero ripercuotersi sulle decisioni future

- *Elencare tutte le possibili azioni e procedere alla fase decisionale*.

6. *Deployment*: le informazioni acquisite dovranno essere organizzate e presentate, in maniera tale da consentire al consumatore di poterne usufruire. Dunque, secondo le esigenze degli analisti, lo stadio

del deployment potrà risultare semplice (ad esempio, realizzare un report) oppure complesso (per esempio, implementare un processo di data mining reiterabile). In molti casi, sarà compito dell'*user* e non del *data mining analyst* effettuare tale stadio. In conclusione, sarà decisivo conoscere chiaramente quali azioni dovranno essere compiute dall'user per rendere possibile l'impiego dei modelli creati.

Task 1 – Plan deployment: nel corso di questa task, la valutazione/giudizio dei/sui risultati del processo di data mining sono impiegati per definire una strategia aziendale.

Task 2 – Monitoraggio e mantenimento del plan: qualora i risultati del processo di data mining diventassero parte integrante delle dinamiche aziendali, il compito di monitoraggio e mantenimento del plan saranno *issues* cruciali per l'organizzazione. In particolar modo, un'attenta strategia di mantenimento (del plan) contribuisce a evitare un impiego non corretto dei risultati del data mining, mentre, l'ideazione di un *monitoring process plan* dettagliato manterrà inalterata l'efficacia applicativa/ pratica immediata del progetto (di data mining).

Task 3 – Produrre report finale: al termine del progetto di data mining, il *project team* redige un report finale (*final report*).

- *Final report*: il final report documenta l'iter "creativo" del progetto di data mining, includendo un'analisi dettagliata dei risultati definitivi (del progetto di data mining).

- *Presentazione finale*: nel corso della presentazione finale il project team avrà modo di illustrare al cliente (*customer*) i risultati del progetto di data mining.

Task 4 – Revisione del progetto: valutazione a posteriori dei pregi e difetti dei risultati del progetto di data mining evidenziati dagli analisti.

- *Documentazione dell'esperienza acquisita nel corso del data mining project*.

CAPITOLO QUARTO

In questo capitolo, si fornirà un'analisi esplicativa e previsionale degli andamenti del numero di donatori e dell'ammontare di donazioni effettuate da privati e imprese alle organizzazioni non profit in Inghilterra dal 1990 al 2014, mediante l'analisi statistica della regressione.

4.1 Ulteriori approfondimenti delle tecniche di regressione lineare e multipla

Secondo quanto è stato precedentemente esaminato (vedi paragrafo 2.4.7.1), l'analisi della regressione è impiegata per spiegare la relazione esistente tra una variabile "*Y*", detta variabile di risposta o dipendente, e una o più variabili esplicative o indipendenti, denominate anche "covariate", *"(X1, X2,... Xk)"*. In termini di funzione, quindi, si ha:

$$Y=f(X1, X2,... Xk)+\varepsilon$$

quest'ultima indica l'esistenza di un legame funzionale, in media, tra la variabile dipendente e i regressori, rappresentati dalla componente *"f(X1, X2,... Xk)"* e alla quale suole dare il nome di "componente sistematica". A tale componente, va ad aggiungersi un'altra denominata "accidentale", casuale o erronea. Mentre la prima rappresenta la parte della variabile di risposta spiegata dai predittori, la seconda componente costituisce quella parte di variabilità della risposta che non può ricondursi a fattori sistematici oppure facilmente individuabili, ma dovuti al caso o, più in generale, a cause diverse non prese in considerazione nel modello di regressione.

In linea teorica, il legame funzionale può essere di qualsiasi tipo, tuttavia, nella prassi statistica, si preferisce utilizzare una funzione di tipo lineare e, pertanto, si definisce "regressione lineare o multipla" ed assume la seguente formulazione:

$$Y= \beta 0 + \beta 1x +\varepsilon$$

oppure in maniera equivalente,

$$E(Y|x) = \beta 0 + \beta 1x$$

dove $\beta 0$ e $\beta 1$ sono costanti numeriche dette, rispettivamente, intercetta e coefficiente angolare: la prima rappresenta il valore di $E(Y|x)$ quando $x = 0$; la seconda indica la variazione che subisce $E(Y|x)$ per un incremento unitario di x.

Infine, nel caso in cui *Y* non dipende solo da un fattore, ma da un insieme di variabili esplicative ("*X1, X2...Xn*"), si ha la cosiddetta "regressione multipla", dove:

$$Y= \beta0+\beta1x1+ \beta2X2+...+\beta kxk$$

Il primo passo da compiere nell'analisi della regressione è quello di rappresentare il campione osservato con un diagramma di dispersione, per accertare l'idoneità del modello di regressione a esprimere la relazione statistica tra le due grandezze.

Successivamente, adottando il modello di regressione lineare, i valori osservati di *Y*, ossia le quantità *y1, y2, ..., yn*, vanno considerati come determinazioni delle seguenti variabili casuali:

$$Y1 = \beta0 + \beta1x1 +\varepsilon1$$
$$Y2= \beta0 + \beta2x2 +\varepsilon2$$
$$.$$
$$.$$
$$.$$
$$Yn= \beta0 + \beta1xn +\varepsilon n.$$

La stima puntuale dei parametri del modello di regressione richiede che le variabili casuali *Y1, Y2, ..., Yn*:

1. siano indipendenti;
2. abbiano la stessa varianza, vale a dire siano tali che

$$Var(Y1)=Var(Y2)=...=Var(Yn)= \sigma^2$$

La seconda ipotesi va sotto il nome di omoschedasticità. Queste assunzioni, unitamente a quella implicita nella retta di regressione, sono riferite alle εi, ossia alla componente aleatoria, richiedendo che $\varepsilon1,\varepsilon2,...,\varepsilon3$ siano variabili casuali indipendenti con media 0 e varianza costante σ^2.

Sia, dunque *(x1, y1), (x2, y2), ..., (xn, yn)* il campione di osservazioni generato dal modello di regressione. La questione fondamentale è la determinazione dei parametri $\beta0$ e $\beta1$ della retta di regressione, che consente di utilizzare il modello a scopi predittivi.

Ora, il metodo che storicamente si è affermato per la stima dei parametri del modello di regressione è il metodo dei minimi quadrati denominato anche *Ordinary Leasted Squares* (*OLS*). Tale stimatore, sulla base di alcune particolari condizioni (ipotesi di *Gauss-Markov*), gode di proprietà desiderabili, quali la correttezza e la consistenza.

Qualora fossero verificate le ipotesi di *Gauss-Markov*, ossia "*sotto le ipotesi classiche del modello di regressione lineare semplice, gli stimatori (B0,B1) dei minimi quadrati per i parametri (b0,b1) sono lineari, non distorti e i più efficienti nella classe degli stimatori lineari e non distorti*", lo stimatore *OLS* è *Best Linear Unbiased Estimator* (*BLUE*). Le stime *b0* e *b1 OLS* per i parametri *β0* e *β1* si ottengono, minimizzando la somma dei quadrati degli scarti e i tra valori osservati *yi* e valori teorici *yˆi*:

$$Sq = \sum_{i=1}^{n} (yi - b0 - b1)\text{^}2$$

Una volta ottenute le stime b0 e b1 dei parametri *β0* e *β1*, si dispone della retta di regressione stimata:

$$\hat{y} = b0 + b1x$$

La varianza del termine di errore, presente nel modello di regressione, è una quantità generalmente non nota che ha un rilievo particolare perché da essa dipendono le deviazioni standard degli stimatori *b0* e *b1*. Occorre quindi stimarla con i dati del campione osservato.

Si consideri:

$$\hat{Y}i = B0 + B1x$$

e si osservi che sussiste l'identità

$$E(\hat{Y}i) = E(Yi|xi) = \beta0 + \beta1xi$$

Tale identità mette in luce come $\hat{Y}i$ sia uno stimatore non distorto del valore medio della variabile dipendente *Yi* associato al valore *xi* della variabile indipendente.

Si prenda, ora, la variabile casuale:

$$\varepsilon i = Yi - \hat{Y}i$$

che è la differenza tra la variabile casuale *Yi* e lo stimatore del suo valore atteso $\hat{Y}i$. Tale differenza, chiamata residuo, ha un ruolo fondamentale nell'inferenza sul modello di regressione. Il residuo εi può essere considerato come la quantità con cui è possibile predire la componente di errore εi.

La stima puntuale dei parametri del modello è basata sulle ipotesi di indipendenza e di omoschedasticità per le variabili casuali *Y1, Y2, ..., Yn.* Per la costruzione degli intervalli di confidenza e per la verifica delle ipotesi sui parametri *b0* e *b1* si rendono necessarie le distribuzioni di probabilità degli stimatori *b0* e *b1*, distribuzioni che si ottengono agevolmente se si assume che le variabili casuali *Y1, Y2, ..., Yn* abbiano distribuzione normale.

Dunque, se si assume che le variabili casuali *Y1, Y2, ..., Yn* siano indipendenti, omoschedastiche e distribuite normalmente, i rapporti

$$Tb0 = \frac{b0 - \beta 0}{\hat{\sigma} b0} \text{ e } Tb1 = \frac{b1 - \beta 1}{\hat{\sigma} b1}$$

hanno entrambi distribuzione t di Student con *n - 2* gradi di libertà.

Sulla base della proposizione precedente, si determina che le variabili casuali campionarie

$$L1 = b0 - t1\text{-}\alpha/2\ \hat{\sigma} b0 \text{ e } L2 = b0 + t1\text{-}\alpha/2\ \hat{\sigma} b0$$

contengono il parametro *b0* con probabilità *1-α,* da cui si ricavano gli estremi dell'intervallo di confidenza per il parametro:

$$l1 = b0 - t1\text{-}\alpha/2\ \hat{\sigma} b0 \text{ e } l2 = b0 + t1\text{-}\alpha/2\ \hat{\sigma} b0$$

In modo analogo, si ricavano gli estremi dell'intervallo di confidenza per $\beta 1$:

$$l1 = b1 - t1\text{-}\alpha/2\ \hat{\sigma} b1 \text{ e } l2 = b1 + t1\text{-}\alpha/2\ \hat{\sigma} b1$$

Infine, sia H*0*:=*β1* = *β1* l'ipotesi nulla, ove *β1* è un particolare valore del coefficiente angolare della retta di regressione. Sia H*0*: *β1* = *β1* l'ipotesi alternativa. Perché possano essere verificate le ipotesi sui coefficienti, è necessario impiegare la statistica test seguente:

$$t = \frac{B_j - \overline{\beta}_j}{\hat{\sigma}_{B_j}}.$$

che, se è vera l'ipotesi nulla, per la proposizione precedentemente esposta, ha distribuzione t di Student con *n - 2* gradi di libertà. Con le usuali considerazioni, si perviene alla seguente zona di rifiuto

$$R = \{t\ b1: |tb1| > t1\text{-}\alpha/2\},$$

dove

$$tb1 = \frac{b1-\beta1}{\sigma\hat{}B1}$$

è il valore osservato della statistica test. Analogamente, si ottengono le zone di rifiuto nel caso di ipotesi alternative unidirezionali.

Prima d'impiegare un modello di regressione per fare inferenza, bisogna verificare che le ipotesi alla base del modello di regressione siano rispettate e non vi siano dati anomali che possano inficiare sui risultati. A tal fine, si effettuano determinati test di ipotesi statistica, attraverso cui si decide se accettare o meno l'ipotesi formulata sulla base delle risultanze campionarie.

Se le assunzioni presentate precedentemente sono vere, ossia se il modello è ben specificato, i residui *eˆi* rifletteranno le proprietà attribuite ai termini di errore *εi*.

Le tecniche di inferenza trattate fin qui sono basate su assunzioni, riguardanti il processo generatore dei dati: l'indipendenza delle variabili casuali *Y1, Y2 ,..., Yn*, l'omoschedasticità, ossia l'uguaglianza delle varianze di queste variabili casuali, e la normalità distributiva delle componenti di errore *ε1 , ε2, ..., εn.*

È importante dunque accertare se tali assunzioni trovano riscontro nei dati osservati.

Il grafico di dispersione, che viene generalmente costruito prima di procedere alla stima dei parametri del modello, fornisce di per sé indicazioni importanti, non solo sulla idoneità del modello a rappresentare la relazione statistica tra la variabile risposta e la variabile indipendente, ma anche sulla validità dell'assunzione di omoschedasticità. Tuttavia, lo strumento fondamentale per esaminare la validità del modello è

rappresentato dall'analisi dei residui. Si tratta dello studio delle proprietà dei residui, ossia delle quantità

$$ei = Yi - \hat{Y}i, \quad i = 1, 2, ..., n$$

L'analisi dei residui costituisce la fase finale dello studio della regressione lineare.

In primo luogo, per verificare l'omoschedasticità dei residui, viene applicato il test di Breusch-Pagan.

L'ipotesi nulla di questo test è la presenza di omoschedasticità, mentre l'ipotesi alternativa è la presenza di eteroschedasticità.

In secondo luogo, per diagnosticare l'autocorrelazione tra i residui, s'impiega la statistica di Durbin-Watson.

$$d = \frac{\sum_{t=2}^{T}(e_t - e_{t-1})^2}{\sum_{t=1}^{T} e_t^2},$$

La statistica di Durbin-Watson assume valori compresi tra 0 - autocorrelazione positiva massima - e 4 - autocorrelazione negativa massima -, con DW=2, che corrisponde al caso di incorrelazione, ossia non appare presente alcuna autocorrelazione. Come affermato precedentemente, la correlazione può essere positiva o negativa, nel primo caso un errore positivo di un'osservazione aumenta la probabilità di un errore positivo in un'altra osservazione, viceversa, se l'autocorrelazione è negativa, un errore positivo di un'osservazione aumenta la possibilità di un errore negativo in un'altra osservazione.

In terzo luogo, si accerta che i dati provengano da una distribuzione normale, mediante il test di Shapiro-Wilk:

$$W = \frac{\left(\sum_{i=1}^{n} a_i x_{(i)}\right)^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

il test è ottenuto dividendo il quadrato di un'appropriata combinazione lineare dei valori campionari ordinati, per l'usuale stima della varianza ed assume valori tra 0 e 1. Gli *xi* sono i dati ordinati di un campione *n*-dimensionale, gli *ai* sono costanti generate dalla media, varianza e covarianza di statistiche ordinate per un campione di dimensione *n* preso da una distribuzione normale. L'ipotesi di normalità viene rifiutata al livello α se il valore osservato di *W* è minore di $w\alpha$, α-imo quantile della statistica test. Dato un campione, il *p-value* di *W* è quindi la probabilità di ottenere un valore di *W* più grande di quello calcolato per i dati, in questo

modo si può rilevare che i valori minimi del *p-value* indicano la presenza di non normalità e, viceversa, per valori elevati. Congiuntamente al test di Shapiro-Wilk, è stato costruito il QQ-plot, rappresentazione grafica per la verifica della distruzione normale dei residui, basato sul confronto tra la distribuzione della variabile osservata con la distribuzione cumulata della variabile.

I punti della distribuzione si addensano sulla diagonale che va dal basso verso l'alto e da sinistra verso destra. Più i residui si distribuiscono normalmente e più i punti dovrebbero disporsi lungo la retta, la cui pendenza è 45° gradi.

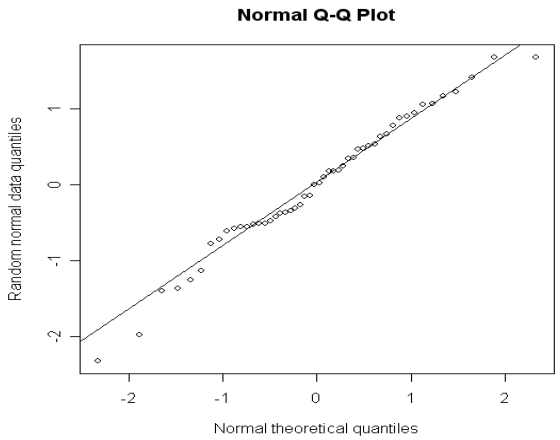


Fig. 4.1 Rappresentazione grafica del QQ-plot

Successivamente, è possibile che, in un modello di regressione multipla, le variabili esplicative *X* siano altamente correlate tra loro, per cui i coefficienti di regressione risultano spesso instabili e le statistiche *t* per le variabili risultano errate. In tal caso, si parla di multicollinearità e ciò non solo comporta che se un valore di una delle variabili esplicative viene minimante modificato, i coefficienti di regressione stimati potranno cambiare sensibilmente, ma anche che il test *F* per la verifica di ipotesi complessiva sul modello risulterà significativo, nonostante ciascuno dei test *t* per i singoli parametri apparirà al contrario non significativo. Per diagnosticare la multicollinearità, si ricorre al VIF (*variance inflation factor*).

$$\mathrm{VIF} = \frac{1}{1 - R_i^2}$$

I VIF, dunque, sono stime di quanto la multicollinearità aumenta la varianza di un coefficiente stimato. Un *VIF* = 1 significa che quella determinata variabile non è coinvolta in nessuna situazione di multicollinearità, mentre un VIF superiore a 1 indica la presenza di un minimo di multicollinearità. Non

esiste alcun criterio universalmente riconosciuto per stabilire la grandezza di VIF dimostrante una seria multicollinearità, tuttavia, la comunità accademica suggerisce di considerare VIF superiori a 10 come indicativi.

Infine, l'analisi della varianza, mediante la tabella ANOVA (*Analysis Of Variance*) è fondamentale per testare le differenze tra medie campionarie e, perciò ,si prendono in considerazione le rispettive varianze. In questo test, l'ipotesi nulla prevede che i dati di tutti i gruppi abbiano la stessa media, mentre l'ipotesi alternativa assume che almeno due delle medie siano tra loro differenti.

L'analisi del modello di regressione si conclude definitivamente con la valutazione della bontà di adattamento del modello ai dati, mediante l'indice di determinazione *R^2*. Esso è definito come il rapporto tra la somma dei quadrati spiegata (*ESS*) e la somma dei quadrati totali (*RSS*) e può assumere valori compresi fra 0 e 1. Se pari ad 1, esiste una perfetta relazione lineare fra il fenomeno analizzato e la sua retta di regressione, mentre se è pari a 0 non esiste alcuna relazione lineare fra le due variabili; infine, i valori compresi fra 0 e 1 forniscono una indicazione sull'efficacia della retta di regressione di sintetizzare l'oggetto dell'analisi.

$$R^2 = \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS}$$

Visto che il coefficiente di determinazione è sensibile al numero di regressori presenti nel modello, nel caso dell'analisi di regressione multipla, all'aumentare del numero di variabili dipendenti *x* incluse nel modello, aumenta il valore di *R^2*, malgrado l'inclusione di ulteriori indicatori non determini un miglioramento della capacità descrittiva del modello stesso. Pertanto, l'*R^2-adjusted* consente di confrontare la diversa capacità d'adattamento ai dati di modelli con un diverso numero di regressori.

4.2 Progetto

A questo punto, s'intende verificare in che misura la relazione tra l'ammontare totale delle donazioni e il numero dei donatori sussista, perciò, è necessario costruire due distinti modelli di regressione lineare per individuare la relazione tra le variabili dipendenti, vale a dire il totale delle donazioni, il numero di donatori e il rapporto tra l'ammontare totale delle donazioni e il numero di donatori, e la variabile indipendente, ossia l'anno. Successivamente, una volta stimati i parametri della regressione, si misurerà la bontà di ogni singolo modello ai dati, mediante l'indice di determinazione *R^2-adjusted* e, infine, si accerterà che siano valide le ipotesi di base, che abbiamo esposto in precedenza, tramite opportuni test statistici, in modo da poter impiegare i modelli a fini predittivi.

4.2.1 Modello di regressione lineare per l'ammontare totale delle donazioni

Variabile dipendente

- *Ammontare totale delle donazioni.*

Variabile indipendente

- *Arco temporale relativo al periodo 1990-2014.*

Stima ed analisi dei parametri del modello di regressione lineare

lm(formula = Total ~ Year, data = donDF)

Residuals:

Min	1Q	Median	3Q	Max
-744.17	-109.02	98.08	205.52	528.43

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-415546.40	20307.07	-20.46	2.93e-16 ***
Year	209.03	10.14	20.61	2.52e-16 ***

Residual standard error: 365.7 on 23 degrees of freedom

Multiple R-squared: 0.9486, Adjusted R-squared: 0.9464

F-statistic: 424.7 on 1 and 23 DF, p-value: 2.516e-16

Analisi dell'omoschedasticità, test di Breusch-Pagan

BP = 0.11608, Df = 1

p-value = 0.7333

Analisi indipendenza dei residui, statistica di Durbin-Watson

DW = 0.4972, p-value = 4.089e-07

alternative hypothesis: true autocorrelation is greater than 0

QQ-plot

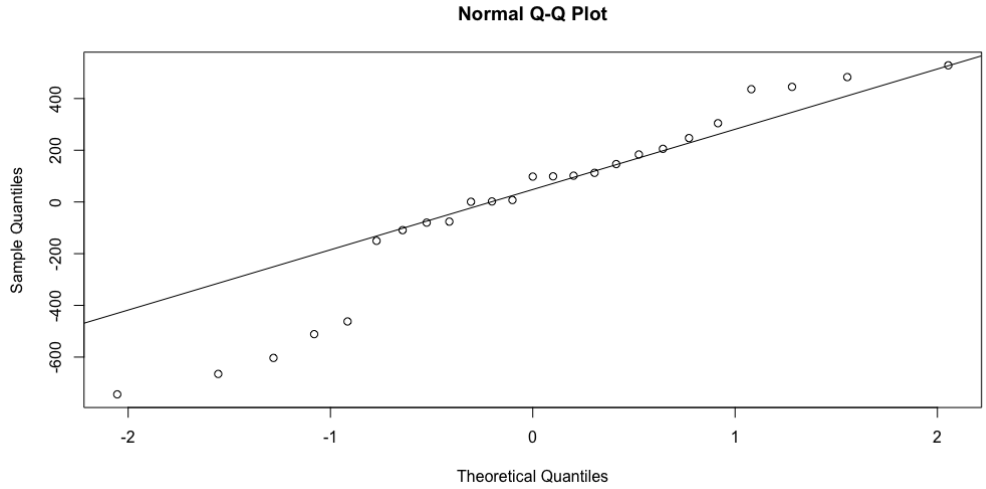


Fig. 4.2 Rappresentazione grafica del QQ-plot

I punti della distribuzione si addensano sulla diagonale, si può dunque accettare l'ipotesi di distribuzione normale dei residui.

Test di Shapiro-Wilk

W = 0.92844, p-value = 0.08

Tabella ANOVA

Response: Total

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Year	1	56802021	56802021	424.68	2.516e-16 ***
Residuals	23	3076323	133753		

4.2.2 Considerazioni finali sul modello di regressione per l'ammontare totale delle donazioni

In base ai risultati del modello di regressione per l'ammontare totale delle donazioni, il valore assunto dall'indice di determinazione (*R^2-adjusted* = 0.9464) denota un buon adattamento della retta di regressione ai punti osservati, difatti circa il 95% della variabilità di *Y,* ossia l'ammontare totale delle donazioni, è spiegata dalla retta di regressione.

Dalla rappresentazione grafica (fig. 4.3) viene confermato l'ottimo adattamento del modello ai dati.

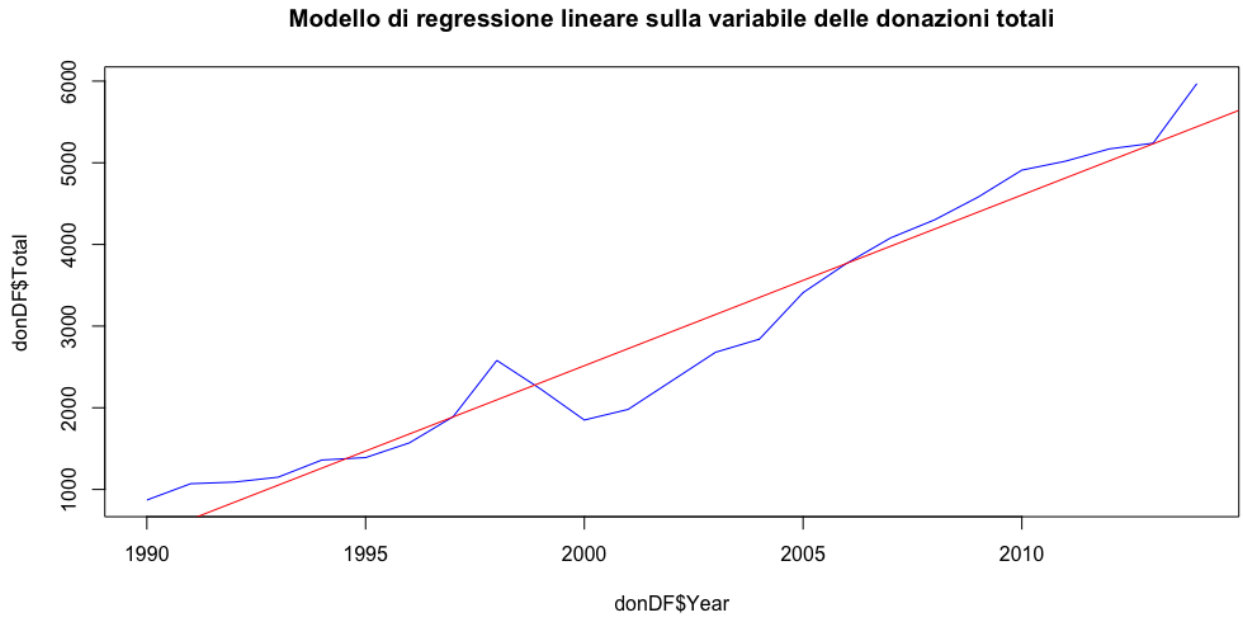


Fig. 4.3 Grafico del modello di regressione lineare sulla variabile donazioni totali

Secondo questo modello, sussiste una relazione lineare tra le donazioni totali ricevute dalle organizzazioni non profit inglesi e l'arco temporale preso in considerazione, dunque, nel corso degli anni, le donazioni, espresse in milioni di sterline, aumenteranno stabilmente. Tale risultato dimostra che il Terzo settore, in Inghilterra, è florido e in costante crescita. Inoltre, tutti i test di specificazione del modello hanno dato esito positivo, quindi si può affermare che le ipotesi alla base del modello di regressione sono valide.

Per concludere, conseguentemente al discreto valore ottenuto dall'indice di adattamento, è stato possibile impiegare il modello di regressione lineare a scopo previsionale, in modo da stimare l'andamento dell'ammontare del totale delle donazioni nel futuro quinquennio 2015-2020 e alla sua successiva rappresentazione (fig.4.4). I risultati confermano l'elevata significatività del modello o retta di regressione lineare rispetto ai dati a disposizione.

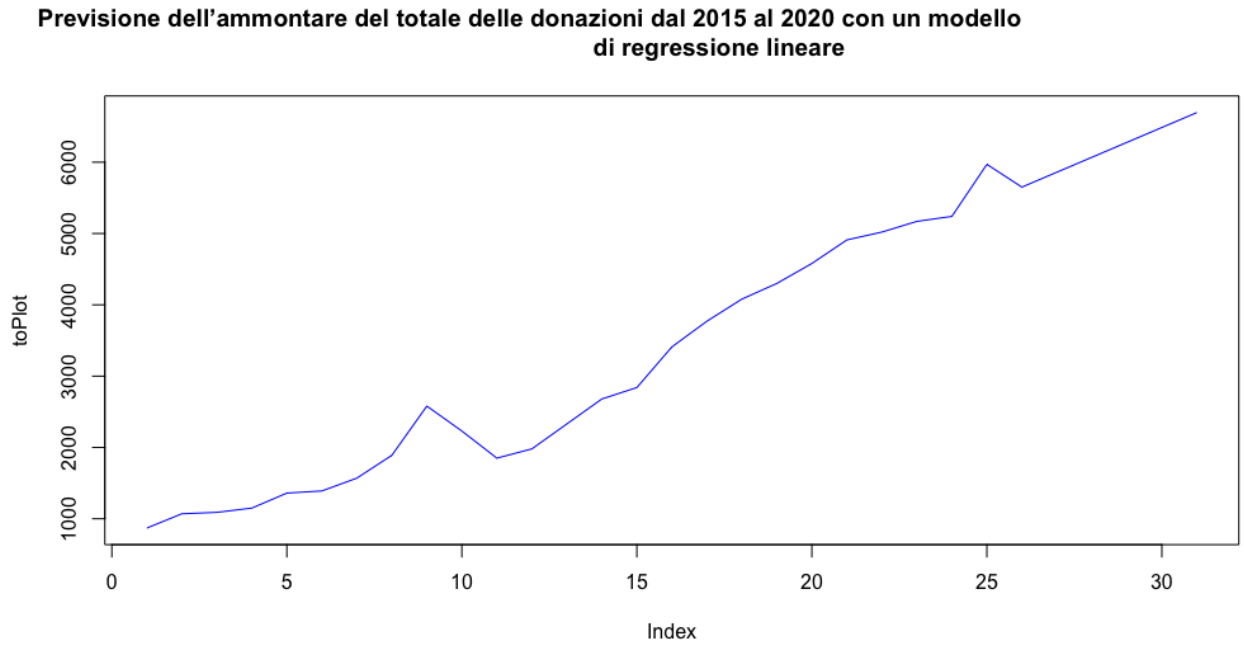


Fig. 4.4 Grafico sulla previsione dell'ammontare del totale delle donazioni dal 2015 al 2020 con un modello di regressione lineare

4.2.3 Modello di regressione lineare per il numero di donatori

Variabile dipendente

- *Numero di donatori.*

Variabile indipendente

- *Arco temporale relativo al periodo 1990-2014.*

Stima ed analisi dei parametri del modello di regressione lineare

lm(formula = Donors ~ Year, data = donDF)

Residuals:

Min	1Q	Median	3Q	Max
-132.57	-67.79	-42.94	49.53	281.99

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-64415.680	5780.341	-11.14	9.54e-11 ***
Year	32.463	2.887	11.24	8.01e-11 ***

Residual standard error: 104.1 on 23 degrees of freedom

Multiple R-squared: 0.8461, Adjusted R-squared: 0.8394

F-statistic: 126.4 on 1 and 23 DF, p-value: 8.013e-11

Analisi dell'omoschedasticità, test di Breusch-Pagan

BP = 0.77187, df = 1, p-value = 0.3796

Analisi indipendenza dei residui, statistica di Durbin-Watson

DW = 1.0876, p-value = 0.003528

alternative hypothesis: true autocorrelation is greater than 0

QQ-plot

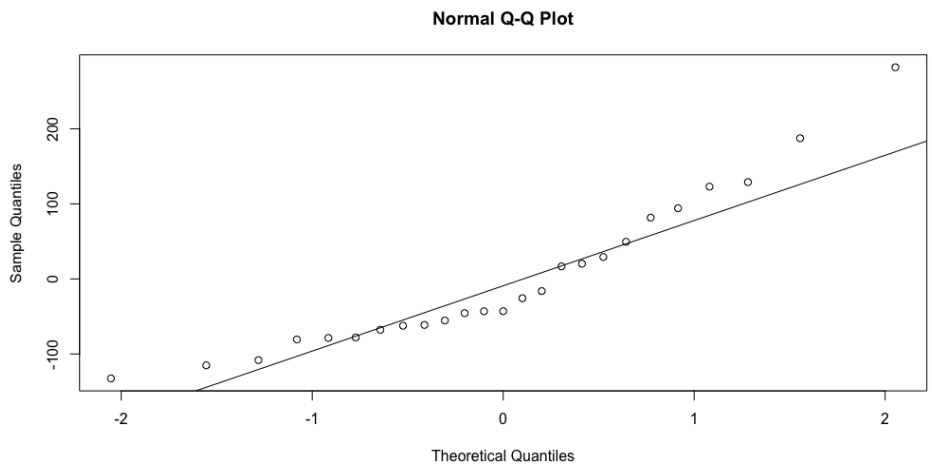


Fig. 4.5 Rappresentazione grafica del QQ-plot

Test di Shapiro-Wilk

W = 0.90464, p-value = 0.02319

Tabella ANOVA

Response: Donors

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Year	1	1370007	1370007	126.42	8.013e-11 ***
Residuals	23	249255	10837		

4.2.4 Considerazioni finali sul modello di regressione per il numero di donatori

In base ai risultati del modello di regressione per il numero di donatori, il valore assunto dall'indice di determinazione (*R^2-adjusted* = 0.8394) denota un discreto adattamento della retta di regressione ai punti osservati, difatti circa l'84% della variabilità di *Y*, vale a dire il numero di donatori, è spiegata dalla retta di regressione.

Dalla rappresentazione grafica (fig. 4.5) viene confermato il sufficiente adattamento del modello ai dati.

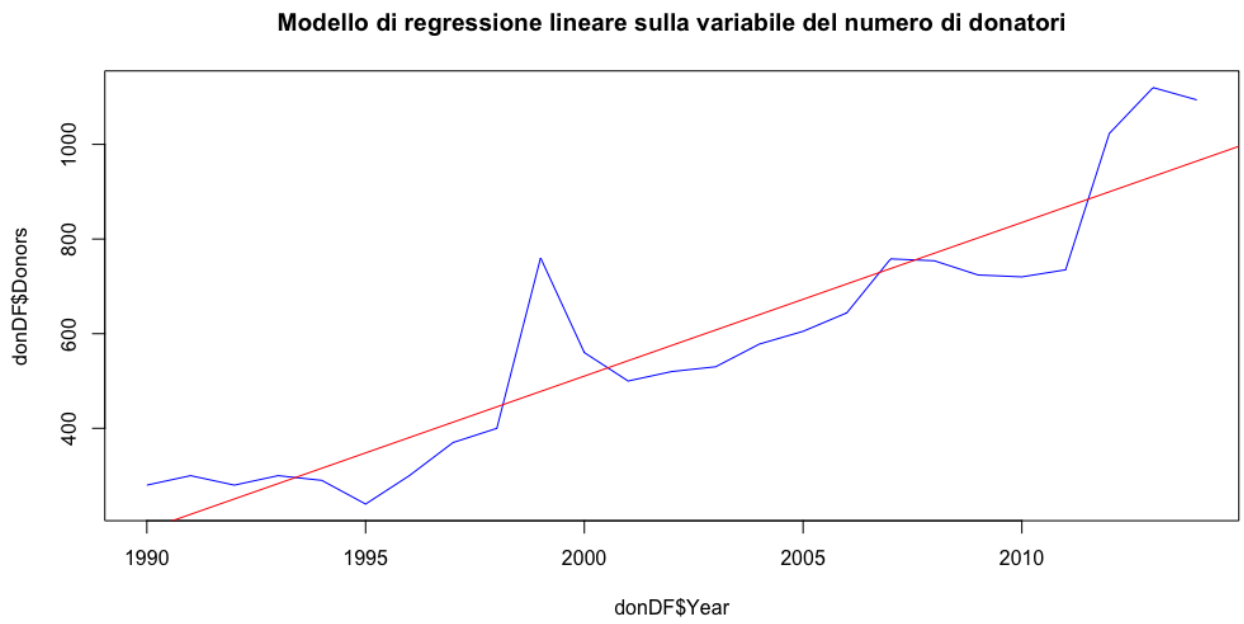


Fig. 4.6 Grafico del modello di regressione lineare sulla variabile numero di donatori

In effetti, non sussiste una perfetta relazione lineare tra il numero di donatori delle organizzazioni non profit inglesi e l'arco temporale preso in considerazione, in quanto il modello ha un andamento sinusoidale e tende a sovrastimare e sottostimare i dati. In particolare, dall'anno 2000, le autorità inglesi hanno riportato problemi di registrazione del numero di donatori partecipanti ai fondi congiunti attivati dalle organizzazioni non profit. Tale risultato dimostra che occorrerà effettuare ulteriori passi avanti nella collaborazione tra le organizzazioni del Terzo settore inglese, per monitorare il flusso di donatori e, conseguentemente, individuare strategie adeguate per coinvolgere e favorire il reperimento di risorse dai privati.

Peraltro, tutti i test di specificazione del modello hanno dato esito positivo, quindi si può affermare che le ipotesi alla base del modello di regressione sono valide.

Visto il discreto valore ottenuto dall'indice di adattamento, è stato possibile impiegare il modello di regressione lineare a scopo previsionale, operando una stima dell'andamento del numero di donatori nel futuro quinquennio 2015-2020 e alla sua successiva rappresentazione (fig.4.7). I risultati confermano la

modesta significatività del modello o retta di regressione lineare rispetto ai dati a disposizione.

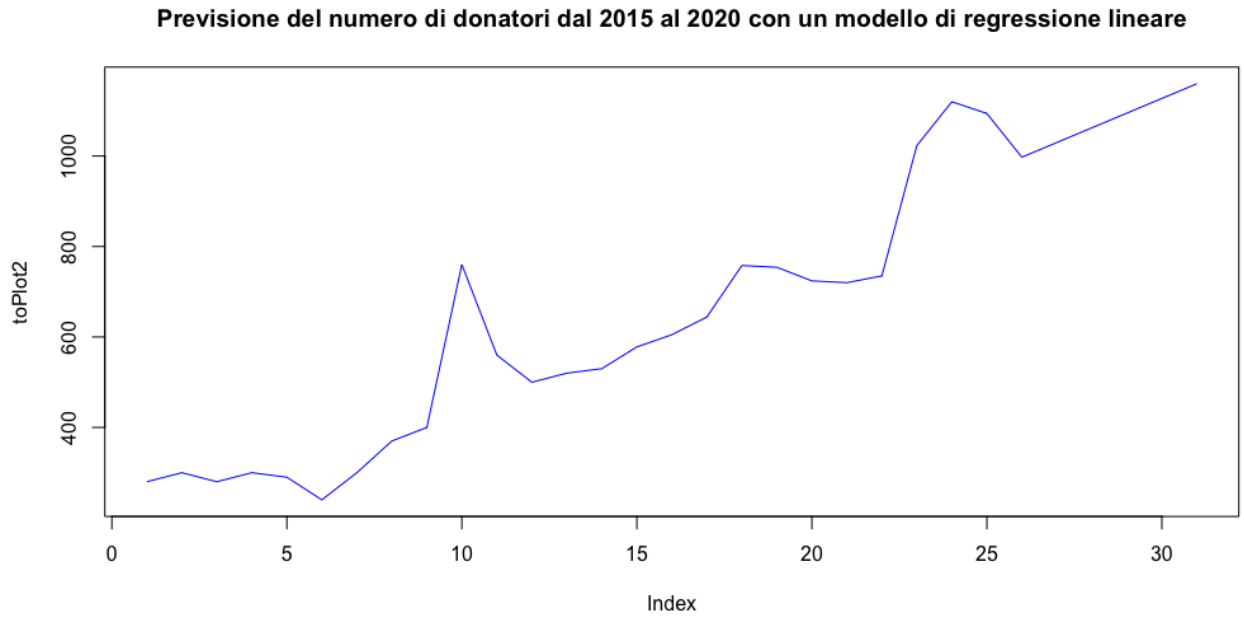


Fig. 4.7 Grafico sulla previsione del numero di donatori dal 2015 al 2020 con un modello di regressione lineare

4.2.5 Modello di regressione lineare per l'ammontare di donazione per singolo donatore

Variabile dipendente

- *Rapporto tra l'ammontare totale delle donazioni e il numero di donatori.*

Variabile indipendente

- *Arco temporale relativo al periodo 1990-2014.*

Stima ed analisi dei parametri del modello di regressione lineare

lm(formula = totMean ~ Year, data = donDF)

Residuals:

Min	1Q	Median	3Q	Max
-1.76516	-0.55789	-0.01842	0.58415	1.83834

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-170.62142	51.90150	-3.287	0.00323 **
Year	0.08770	0.02592	3.383	0.00256 **

Residual standard error: 0.9347 on 23 degrees of freedom

Multiple R-squared: 0.3323, Adjusted R-squared: 0.3032

F-statistic: 11.44 on 1 and 23 DF, p-value: 0.002562

Analisi dell'omoschedasticità, test di Breusch-Pagan

BP = 0.038245, df = 1, p-value = 0.845

Analisi indipendenza dei residui, statistica di Durbin-Watson

DW = 1.1519, p-value = 0.006222

alternative hypothesis: true autocorrelation is greater than 0

QQ-plot

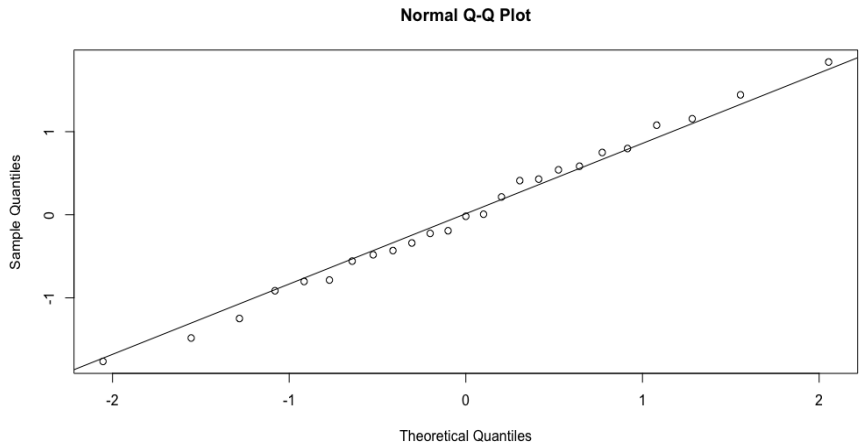


Fig. 4.7 Rappresentazione grafica del QQ-plot

Test di Shapiro-Wilk

W = 0.99135, p-value = 0.9982

Tabella ANOVA

Response: totMean

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Year	1	9.9996	9.9996	11.445	0.002562 **
Residuals	23	20.0954	0.8737		

4.2.6 Considerazioni finali sul modello di regressione per l'ammontare di donazione per singolo donatore

In base ai risultati del modello di regressione per l'ammontare di donazione per singolo donatore, il valore assunto dall'indice di determinazione (*R^2-adjusted* = 0.3032) non denota un buon adattamento della retta di regressione ai punti osservati, difatti circa il 30% della variabilità di *Y,* ossia l'ammontare di donazione per singolo donatore, è spiegata dalla retta di regressione.

Dalla rappresentazione grafica (fig. 4.7) si può osservare l'insufficiente adattamento del modello ai dati.

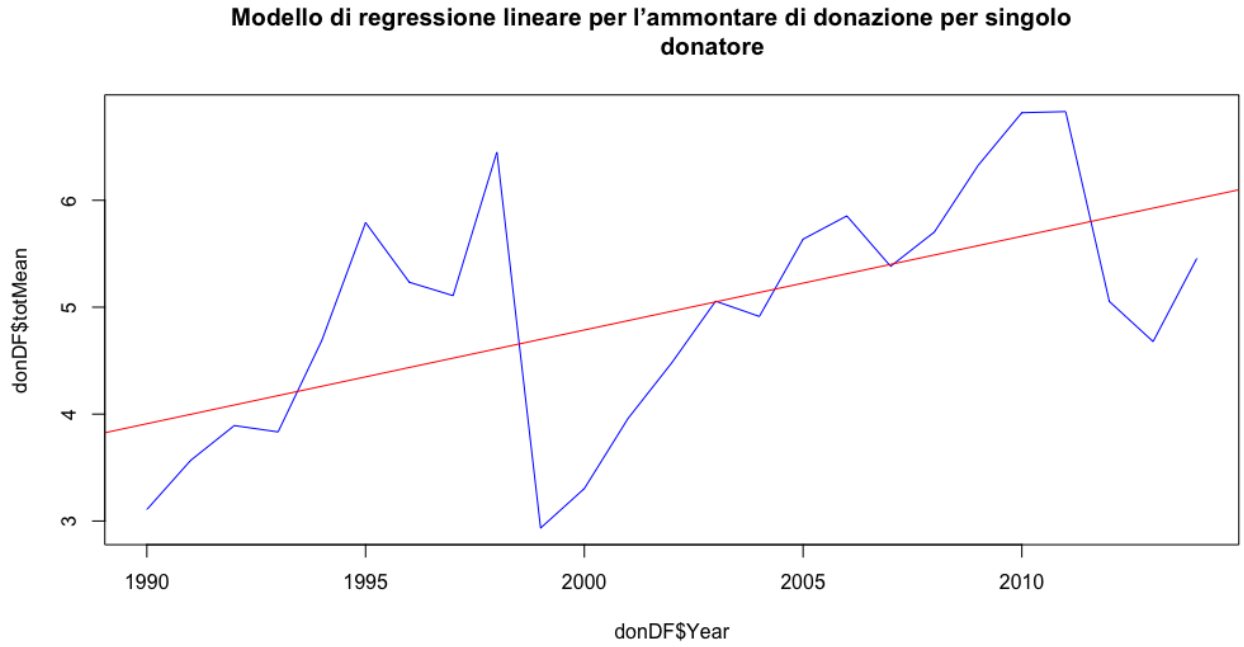


Fig. 4.8 Grafico del modello di regressione lineare per l'ammontare di donazione per singolo donatore

È importante notare che la variabile dipendente consiste di un rapporto tra l'ammontare totale delle donazioni ed il numero di donatori. Nel corso del biennio 1998-2000, il rapporto subisce una vistosa diminuzione, determinata principalmente da una sensibile riduzione dei donatori e, contemporaneamente, da un sostanziale incremento delle donazioni totali. Tuttavia, non sussiste una relazione lineare tra l'ammontare di donazione per singolo donatore ricevuto dalle organizzazioni non profit inglesi e l'arco temporale preso in considerazione. Tale esito dimostra che il Terzo settore, in Inghilterra, ha bisogno di adottare iniziative di solidarietà che inducano non solo i privati, ma anche le imprese a contribuire maggiormente all'operato delle associazioni non profit, a livello economico. Inoltre, dall'analisi risulta che tutti i test di specificazione del modello hanno dato esito positivo, quindi, si può affermare che le ipotesi alla base del modello di regressione

sono valide.

In conclusione, visto l'insufficiente adattamento del modello ai dati osservati, non è stato possibile realizzare una previsione affidabile riguardo l'ammontare di donazione per singolo donatore negli anni successivi al 2014.

NOTE CONCLUSIVE

Con il presente lavoro ci siamo prefissi l'obiettivo di interpretare e gestire la crescente massa di dati relativi alle organizzazioni non profit operanti nel Regno Unito, resi disponibili dall'Agenzia delle Entrate britannica, l'H&M Revenue & Customs, mediante le tecniche e le metodologie di Data Mining, al fine di ricavarne risultati utili sotto il profilo sociale, statistico ed economico e di proporre soluzioni di business intelligence innovative e, al tempo stesso, efficaci a beneficio delle comunità e delle aziende, nonché delle stesse organizzazioni non profit.

In effetti, il Terzo Settore, per le dimensioni economiche ed occupazionali assai rilevanti che ha assunto negli ultimi anni, sta diventando un elemento imprescindibile, e non occasionale, dell'organizzazione della comunità, assolvendo una funzione fondamentale nel promuovere politiche sociali attive. Ed è proprio grazie all'esistenza di tali organizzazioni, alternative allo Stato ed al mercato, che determinati bisogni possono trovare, e trovano, il loro soddisfacimento.

In tal senso, il Regno Unito vanta una tradizione culturale assai antica e radicata, che ha permesso l'instaurarsi di un Terzo Settore solido ed efficiente, presente ed attivo su tutto il territorio nazionale, che fornisce un forte sostegno economico e sociale all'intera società.

In virtù di tali considerazioni iniziali, utilizzando una tra le principali tecniche di data mining, cioè l'analisi della regressione lineare, abbiamo cercato di mettere in luce le criticità ed al tempo stesso le notevoli potenzialità degli enti non profit inglesi, le Charities, individuando e prevedendo gli andamenti del numero dei donatori e dell'ammontare di donazioni da parte di privati ed imprese alle organizzazioni non profit.

L'analisi statistica effettuata ha permesso di evidenziare che il Terzo settore, in Inghilterra, è florido ed in costante crescita. Le numerose iniziative solidaristiche, promosse dalle Charities, favoriscono la crescita ed il consolidamento di un'economia sociale dei servizi, riducendo le disuguaglianze e, conseguentemente, promuovono il dialogo e la tolleranza tra la cittadinanza, sviluppando un'organizzazione flessibile di rapporti sociali, in un quadro di tutele e garanzie fortificato. Oltretutto, dall'analisi dei bilanci sull'andamento dell'ammontare totale delle donazioni, emerge che le autorità inglesi contribuiscono in maniera decisiva allo sviluppo del Terzo Settore, attribuendogli valenza strategica per la promozione di uno Stato Sociale equo e solidale.

Tuttavia, l'analisi dei risultati evidenzia anche gravi anomalie di coordinamento tra le organizzazioni non profit, che, per raggiungere un'auspicabile autonomia finanziaria, dovrebbero proporsi l'obiettivo di rafforzare la collaborazione e la comunicazione tra di loro. A livello previsionale, l'analisi dei dati, sviluppata

con modelli statistici multivariati, mostra, infatti, un trend altalenante nel flusso dei donatori, che impone la necessità di individuare strategie efficaci di business che coinvolgano e favoriscano il reperimento delle risorse dai privati.

Perseguendo tali strategie, le organizzazioni non profit potranno non solo giocare un ruolo sempre più decisivo nella società e nelle istituzioni, ma anche costruire un solido legame con le comunità all'interno delle quali assolvono alla loro funzione sociale e solidaristica.

BIBLIOGRAFIA & SITOGRAFIA

CAPITOLO PRIMO

Anthony R. N., D. W. Young, Controllo di gestione per gli enti pubblici e le organizzazioni non profit, Milano, 1992

Ascoli, U. (cura di), *Il Welfare futuro. Manuale critico del Terzo settore*, Carocci, Roma, 1999, p. 13

Bancone, V. (2009). Trust ed enti non commerciali. Profili comparativi e potenziali applicazioni. Santarcangelo di Romagna: Maggioli editore.

Barbetta, G. P. & Maggio, F. (2008). Nonprofit. Bologna: Il Mulino.

Barbetta G. P., Il settore nonprofit italiano, Studi e Ricerche, il Mulino, 2000,

Charities Act 2011

Charities Act 1993, Part 1, The Charity Commissioners and the Official Custodian for Charities, The Commissioners

Charitable Uses Act 1601, Preambolo

Fries, R. (1999). Charity and the Charity Commission Atti del Convegno I controlli sulle organizzazioni non profit. Bologna: Il Mulino.

Giambrone F. (2014), Politiche per la cultura in Europa: modelli di governance a confronto, Milano: Franco Angeli

M. Grumo, Introduzione al management delle aziende non profit, Milano, 2001

Helmut K. Anheier (2014), Nonprofit Organizations: Theory, Management, Policy, New York: Routledge

Propersi A. (2011), Gestione e bilanci degli enti non profit, Milano: Franco Angeli

Propersi, A. (2004). Il sistema di rendicontazione negli enti non profit. Dal bilancio d'esercizio al bilancio di missione. Milano: Vita e Pensiero.

Propersi, A. (1999). Le aziende non profit. I caratteri, la gestione, il controllo. Milano: RCS Libri.

Propersi A., Rossi G. (2010). Gli enti non profit. Milano: Il sole 24 Ore.

Sessa, V. M. (2007), Gli enti privati di interesse generale. Giuffré Editore

Sharon M. Oster (1995), Strategic Management for Nonprofit Organizations: Theory and Cases, New York: Oxford University Press

Zamagni S., Il libro bianco sul terzo settore, Bologna: Il Mulino

https://www.gov.uk/charities-and-tax

https://www.law.cornell.edu/wex/non-profit_organizations

https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/225321/06_joint_venture_guidance.pdf

http://www.spes.lazio.it/europa/regno_unito.pdf

CAPITOLO SECONDO

Box G.E., Jenkins G. M. (1976): Time Series Analysis, Forecasting and Control, Holden-Day, San Francisco

Brachman, R.J.; Anand, T. (1996). The Process of Knowledge Discovery in Databases: A Human-Centered Approach. In Fayyad, U.M., Piatetsky-Shapiro, G., Smyth, P., & Uthurasamy, R. (eds.). Advances in Knowledge Discovery and Data Mining. AAAI /MIT Press

Di Fonzo T., Lisi F. (2005): Serie storiche economiche, Roma: Carocci

Dulli S., Furini S., E. Peron (2009) Data Mining. Metodi e strategie, Milano: Springer

Fayyad U. M., Piatetsky-Shapiro G., Smyth P. (1996), From Data Mining to Knowledge Discovery in Databases, AI Magazine Volume 17 Number 3

Fayyad, U. M.; Piatetsky-Shapiro, G.; Smyth, P. (1996): The KDD Process for Extracting Useful Knowledge from Volumes of Data. In Communications of the ACM, November 1996, Vol. 39, No. 11, pp. 27-34

Rezzani A. (2012), Business intelligence. Processi, metodi, utilizzo in azienda, Apogeo

http://www.academia.edu/2715996/Data_mining_e_statistica - Statistica e società: rivista quadriennale per la cultura statistica, 09.2004, anno III, n.1

http://ai.stanford.edu/users/gjohn/ftp/papers/thesis-large.ps

CAPITOLO TERZO

Burley H. (2012), Cases on Institutional Research Systems, Hershey (USA) Information Science Reference

http://crisp-dm.eu/home/crisp-dm-methodology/

http://crisp-dm.eu/reference-model/

ftp://public.dhe.ibm.com/software/analytics/spss/documentation/modeler/15.0/it/CRISP-DM.pdf

CRISP-DM: Towards a Standard Process Model for Data Mining Rüdiger Wirth: DaimlerChrysler Research & Technology FT3/KL PO BOX 2360 89013 Ulm, Germany ruediger.wirth@daimlerchrysler.com; Jochen Hipp: Wilhelm-Schickard-Institute, University of Tübingen Sand 13, 72076 Tübingen, Germany jochen.hipp@informatik.uni-tuebingen.de

CAPITOLO QUARTO

http://www.dis.uniroma1.it/~statistica/dispensaR.pdf

https://cran.r-project.org/doc/contrib/nozioniR.pdf

https://cran.r-project.org/doc/contrib/Ricci-ts-italian.pdf

https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/456405/CommentaryDocument.pdf

ABSTRACT

Since the mid-nineties the Web has progressively established itself and expanded as an all-inclusive and detailed platform, towards which a multitude of data, at times lacking of exact, intrinsic significance, has flown at a striking and often chaotic rate.

In virtue of this continued proliferation of data on the Internet, embedded in the assertion of the digital era, the global academic community felt the need of scientific methods for managing and interpreting this ever growing mass of data so as to make it available for social, statistical and economic purposes. As well as proposing innovative business intelligence solutions beneficial to both communities and businesses, including non profit organizations (Tertiary sector).

Comparative data analysis relative to the Tertiary sector carried out by John Hopkins University demonstrate how "practically everywhere, the non profit sector has taken on...important economic and occupational dimensions. In the countries under examination, the total costs of the sector amount to approximately 1.4 billion Euros, an amount that would make it the world's 8th largest economy".

In this context the fundamental importance of the statistical methodology of Data Mining - the driving force of the KDD (Knowledge Discovery in Database) process - needs to be underlined. It has gradually made 'raw' data, present in non profit organizations databases accessible; extracted information, patterns and relationships not immediately identifiable or known but which are useful for analyzing and understanding social and economic business aspects with the aim of rationalizing and optimizing the collection and management of resources.

The field of analysis of this research is delimited to the collection and management of data of English non profit organizations, specifically Charities, present throughout Great Britain (par.1.2); the illustration of their distinctive characteristics (per.1.2.1), different institutional forms (par.1.2.2) as well as economic and financial regimes (par.1.3), and a specific focus on the main sources of finance (par.1.4).

Successively, attention was drawn to the study of the different and multiform methodologies/techniques of Data Mining (DM) (Ch. 2), the link in the data analysis process, aimed at the discovery of information for understanding and foreseeing the trend of specific quantitative variables.

DM is a creative process which requires specialized competences and knowledge. It is not currently possible to trace a theoretical outline that allows data mining projects to be realised. Data mining needs a

procedural model that helps translate business problems into data mining solutions, suggesting appropriate interpretations of the data and providing the necessary instruments for evaluating the efficacy of the results, and for documenting the ongoing project.

Attention is strictly focused on the CRISP-DM (Cross Industry Standard Process for Data Mining) project, which proposes a comprehensive and intuitive procedural model for realising data mining projects. CRISP-DM, the reference model for data mining, provides an overview of the life cycle of a DM project. It includes project stages, their respective tasks, and results.

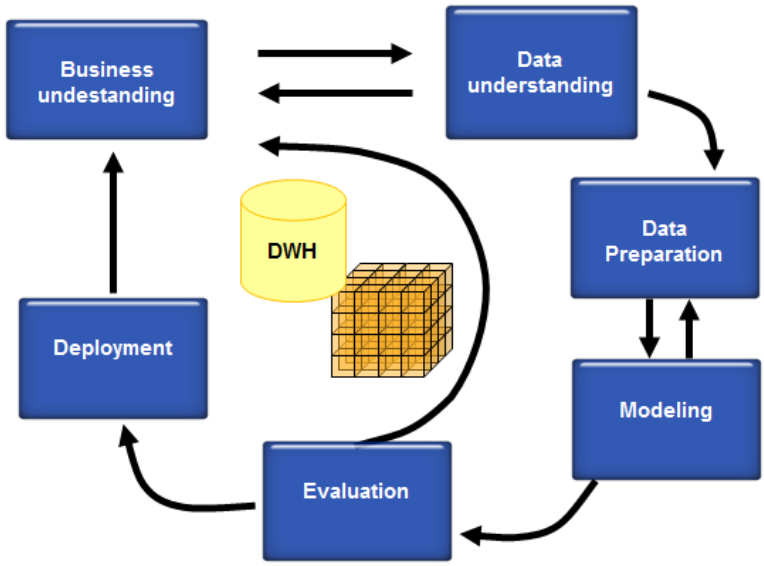


Fig.1 The Cross Industry Process for Data Mining (CRISP-DM) methodology

Finally, after the CRISP-DM illustration, an explanatory and predictive analysis of the trend of the number of donors and the sum of donations by privates and companies to non profit organizations in England from 1990 to 2014, through the statistical analysis of regression was formulated.

Regression analysis is used to explain the relationship between a variable Y, also called a dependent variable; and one or more independent variables (X1, X2,... Xk). The function is as follows:

$$Y=f(X1, X2,... Xk)+\varepsilon$$

this formula shows a functional bond, in average, between the dependent variable and the regressors, represented by the "f(X1, X2,... Xk)" component, which is usually called "systematic component". Another component denominated casual is added to the systematic component. The second component represents the part of result variability which cannot be accounted for by systematic factors or which is not easily identifiable but is due to chance, or (more generally) to causes not taken into consideration in the regression model.

In theory, the functional bond can be of any type. However, in the statistical procedure, it is preferable to use a linear function therefore defining "linear or multiple linear regressions", therefore:

$$Y= \beta 0 + \beta 1x + \varepsilon$$

or,

$$E(Y|x) = \beta 0 + \beta 1x$$

where β0 e β1 are numerical constants respectively denominated angular coefficient and intercept: the first represents the value E(Y|x) when x = 0; the second indicates the variation that E(Y|x) for a unitary increase in x.

Finally, in the case in which Y is dependent on a number of explicative variables ("X1, X2,... Xk"), you have "multiple linear regression", where:

$$Y= \beta 0+\beta 1x1+ \beta 2X2+...+\beta kxk$$

At this point, the degree of relationship between the total sum of the donations and the number of donors, needed to be verified. It was therefore necessary to construct two distinct linear regression models to identify the relationship between the dependent variables (the total of the donations and the numbers of donors) and the independent variables (the year). Once the regression parameters had been estimated, the goodness of fit of each model, through the determination index R^2-adjusted was measured. And finally with the use of suitable statistical tests, the hypotheses were verified, so as to be able to able to use the models for predictive aims.

Based on the results of the regression model for the total sum of donations, the hypothesized value from the determination index (R^2-adjusted = 0.9464) denotes a good trend of the regression line to the

observed points. In fact, approximately 95% of the variability of Y – the total sum of the donations - is explained by the regression line.

The goodness of fit is represented graphically (Fig.2).

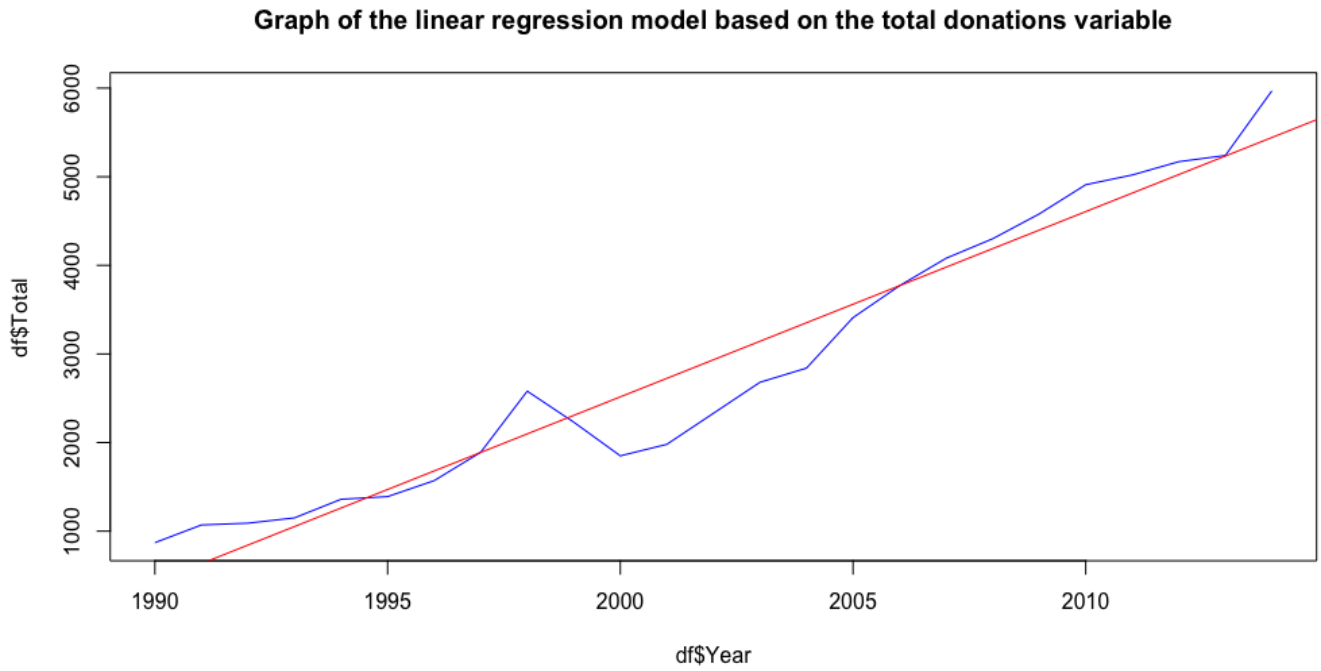


Fig.2 Graph of the linear regression model based on the total donations variable

According to this model, a relationship between the total donations received by the English non profit organizations and the arc of time taken into consideration exists. Therefore, during the years, the donations, expressed in millions of Pounds Sterling, will increase stably. This result demonstrates that the Tertiary sector in England is thriving and in constant growth. Furthermore, as all the model specification tests were positive, we can declare that the hypotheses of the regression model are valid.

To conclude, consequently to the discrete value obtained from the trend index it was possible to use the linear regression model for forecasts, so as to estimate the total sum of donations for the following five years (2015-2020) and to represent it graphically (Fig.3). The results confirm the meaningfulness of the linear regression model with respect to the available data.

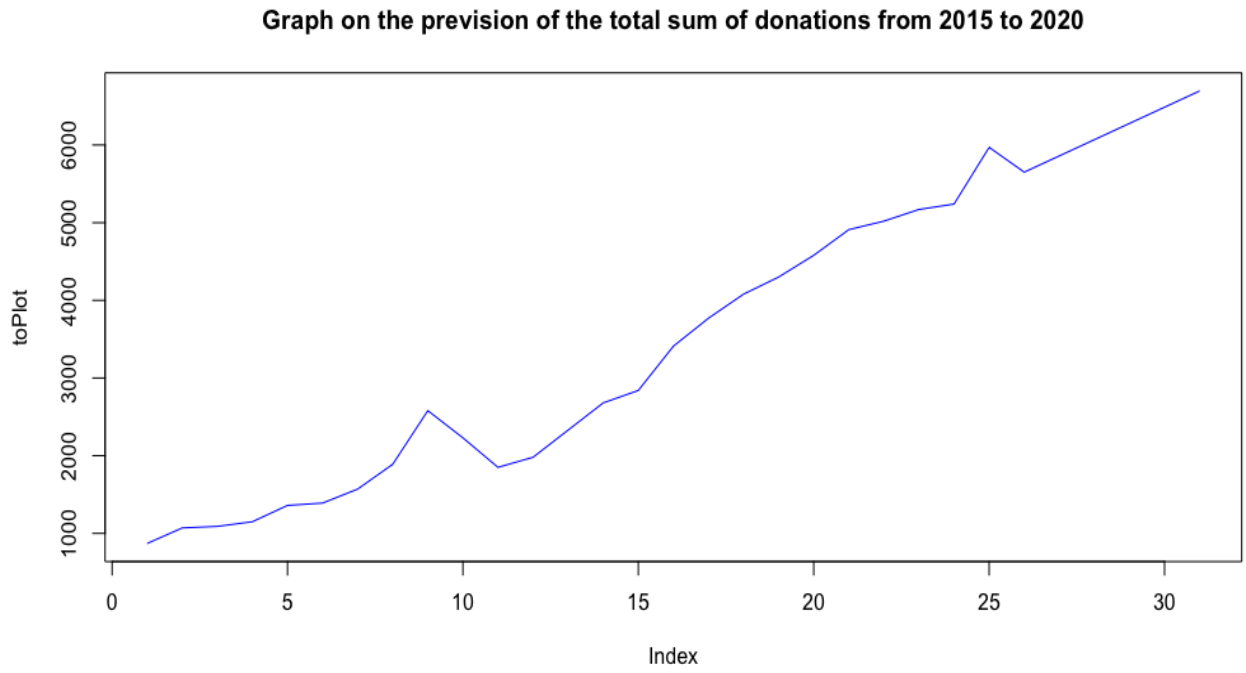


Fig.3 Graph on the prevision of the total sum of donations from 2015 to 2020

Regarding the results of the regression model for the number of donors, the hypothesized value from the determination index (R^2-adjusted = 0.8394) denotes a discrete trend of the regression line to the observed points. In fact, approximately 84% of the variability of Y – the number of donors - is explained by the regression line.

The goodness of fit is represented graphically (Fig.4).

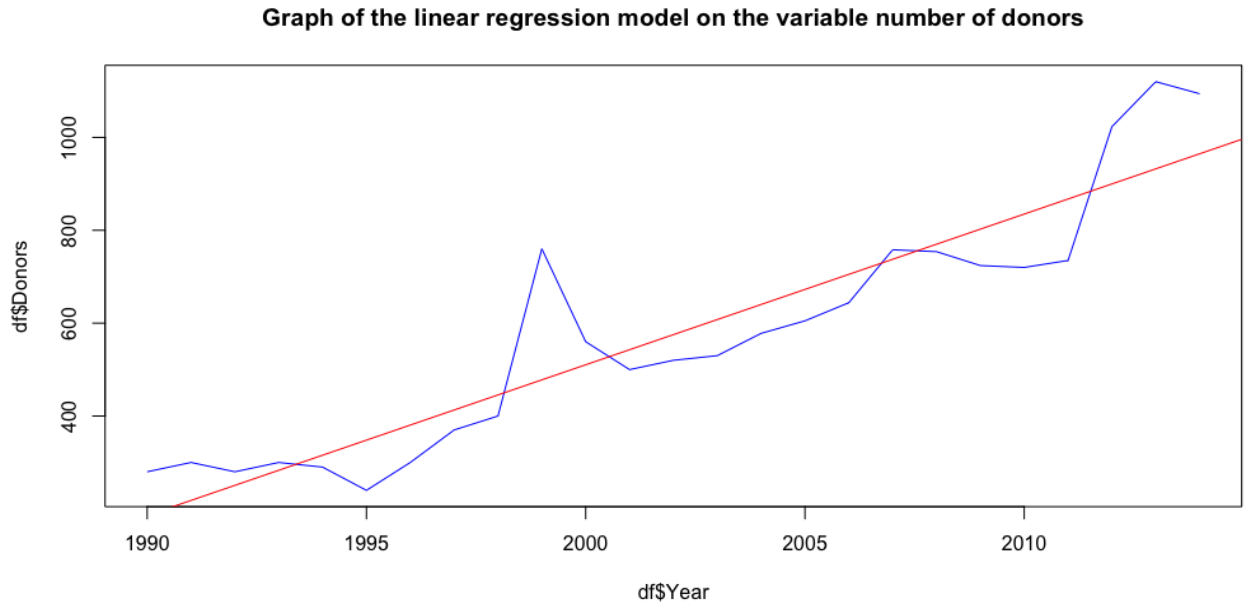


Fig.4 Graph of the linear regression model on the variable number of donors

As the model has a sinusoid trend and tends to over- or under-estimate data, a perfect linear relationship between the number of donors and the arc of time taken into consideration does not exist. For example, in 2000, English institutions reported problems with the registration of the number of donors who participated in joint funds initiated by non profit organizations. This demonstrates that further steps need to be taken towards the collaboration between the English Tertiary sector organizations, in order to monitor the flux of donors and consequently identify fitting strategies for facilitating fundraising.

Moreover, as all the model specification tests gave positive results, we can confirm the validity of the hypotheses of the model. Considering the discrete value of the adjustment index, it was possible to use the linear regression model for predictive purposes, using an estimate of the number of donors in the following five years 2015-2020 and successively representing it (Fig.5). The results confirm the modest importance of the model or linear regression line with respect to the available data.

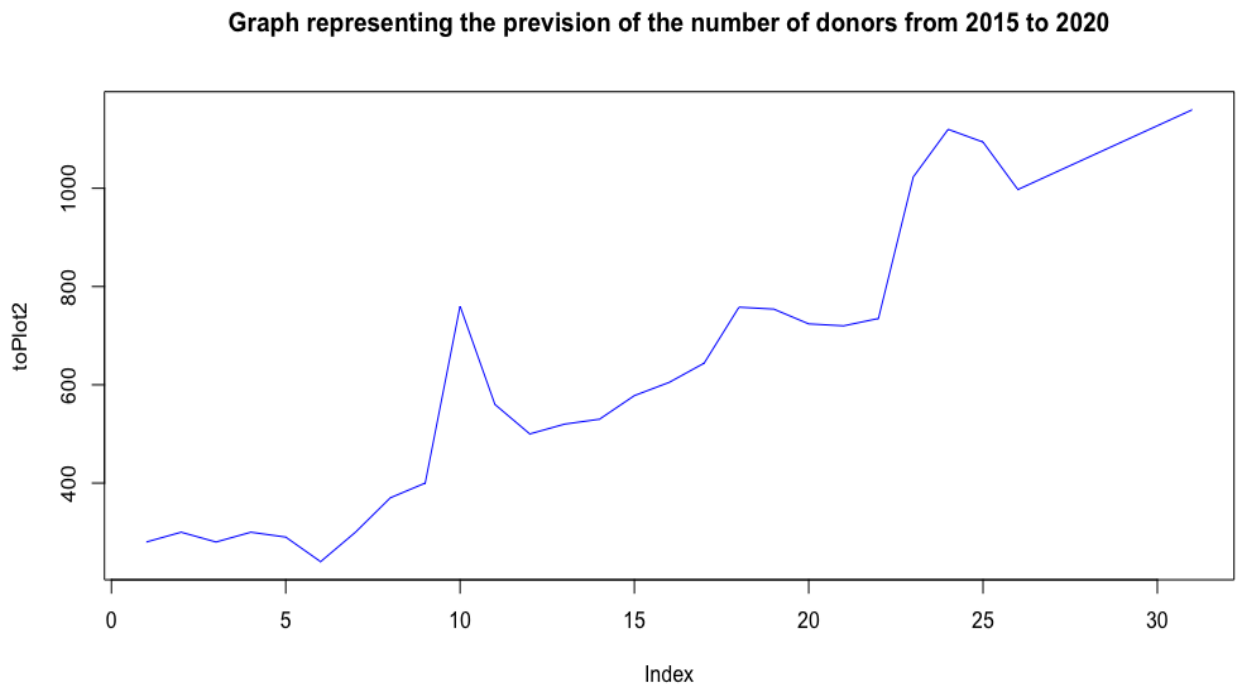


Fig.5 Graph representing the prevision of the number of donors from 2015 to 2020