UNIVERSITA' DI PISA

Facoltà di Economia
Facoltà di Scienze Matematiche, Fisiche e Naturali
Corso di Laurea Magistrale in Informatica per l'economia e l'azienda
(Business Informatics)

TESI DI LAUREA

IMPIEGO DI METODOLOGIE DI CLASSIFICAZIONE NELLA PIANIFICAZIONE DEI CONTROLLI FISCALI

RELATORI:
Prof. Dino PEDRESCHI
Prof.ssa Fosca GIANNOTTI
Dott. Diego PENNACCHIOLI

Candidato
Mauro BARONE

ANNO ACCADEMICO 2012-2013

A Valentina e Diego

Riassunto

L'evasione fiscale in Italia rappresenta un fenomeno diffuso, imponente nei numeri (oltre 120 miliardi di euro di imposte evase ogni anno) e foriero di gravi conseguenze sia dal punto di vista economico che dal punto di vista sociale.

Nel nostro Paese è l'Agenzia delle Entrate la struttura organizzata per far fronte a tutte le fasi che concernono la lotta all'evasione e la *tax compliance*, ma, nonostante gli sforzi profusi, ad oggi, il *gap* esistente tra gettito evaso e recupero delle imposte sottratte a tassazione è ancora molto forte.

Proprio da questa osservazione discendono molti problemi interessanti, tra cui la valutazione della possibilità che tecniche di *data mining* e, in particolare, di classificazione, possano essere applicate con successo nell'ambito della lotta all'evasione.

La stessa Agenzia delle Entrate ha quindi messo a disposizione per questa tesi uno specifico *dataset* perché vi fossero condotte analisi dei dati volte a dare una risposta al quesito su posto.

Le esperienze condotte sembrano poter autorizzare ad una risposta positiva.

Modelli che fanno uso di alberi di classificazione e regole decisionali, scelti per la loro spiccata espressività ed interpretabilità, opportunamente allenati hanno mostrato di ottenere buone *performances*, ottenendo buoni risultati anche in termini di gettito recuperato.

La novità introdotta da questo lavoro è l'idea che i *lift chart* associati ai modelli possano essere utilizzati come guida flessibile per la pianificazione fiscale, coniugando, da una parte, le esigenze ed obiettivi dell'utilizzatore e dall'altro l'*output* di buoni modelli di classificazione.

INDICE

Capitolo 1

Introduzione

1.1 Presentazione del problema

L'evasione fiscale in Italia è un fenomeno diffuso, che produce gravi conseguenze sia dal punto di vista economico che dal punto di vista sociale. Anche se manca una quantificazione "ufficiale" del fenomeno, per sua natura sfuggente, con riferimento all'evasione di IVA e IRAP, la Corte dei Conti stima per il solo 2011 un mancato introito di circa 50 miliardi di euro. A ciò deve aggiungersi l'evasione delle altre imposte (tra cui IRE – ex IRPEF – ed IRES) e dei contributi sociali. Recenti stime disponibili – [Conf10], riferite al 2009, propongono una evasione totale in Italia attestata su valori superiori ai 125 miliardi di euro.

Si tratta di un problema rilevante per diversi motivi: riduce il gettito fiscale, generando problemi nel bilancio dello Stato, indirizza il prelievo sulle basi imponibili che è meno agevole sottrarre alla tassazione (ad esempio, il lavoro dipendente), tende a produrre disparità di trattamento tra soggetti con uguale capacità contributiva (c.d. iniquità orizzontale) e introduce distorsioni tra gli operatori economici, alterando le condizioni di concorrenza sui mercati, con riflessi negativi sull'efficienza del sistema, per citarne alcuni.

L'obiettivo di ridurre, quindi, in maniera significativa e strutturale, le aree di evasione e di agevolare la scelta del contribuente verso l'adempimento spontaneo dei propri obblighi fiscali (c.d. *tax compliance*) diviene, per il Paese, strategico.

A fronte di una evasione stimata di oltre 120 miliardi di euro l'anno, le somme recuperate a seguito di specifica attività di

accertamento da parte dell'Agenzia delle Entrate[1], intese come maggiori imposte e sanzioni a fini IIDD, IRAP e IVA, nel 2011, si sono attestate sui 7,2 miliardi di euro.

E' evidente come l'Agenzia delle Entrate già disponga di strumenti e tecniche con cui correntemente seleziona i soggetti da sottoporre a controllo. Nonostante gli sforzi profusi, ad oggi, il *gap* esistente tra gettito evaso e recupero delle imposte sottratte a tassazione è ancora molto forte.

Dal quadro sopra descritto, discendono molte questioni di rilievo per la stessa Agenzia delle Entrate, che il presente lavoro, peraltro reso effettivamente possibile grazie alla collaborazione con il Servizio Studi della stessa Agenzia (coordinato dal Dott. Stefano Pisani) e con la SOGEI, suo partner tecnologico, che hanno fornito la base dati oggetto di studio, tenta di affrontare. In particolare, esse si incentrano sulla valutazione della possibilità che tecniche di *data mining* e, in particolare, di classificazione, possano essere applicate con successo in questo ambito.

Le esperienze condotte sembrano poter autorizzare ad una risposta positiva.

1.2 Rassegna della letteratura

Il tema dell'evasione fiscale, vista come come sottoinsieme del più vasto insieme delle frodi finanziarie, definite come "*un deliberato atto contrario alla legge, a una regola, a una politica, con l'intento di ottenere un beneficio finanziario non autorizzato*" è stato più volte affrontato in letteratura.

In particolare, spesso l'ottica seguita è stata quella di valutare l'efficacia di tecniche di classificazione nella fase della selezione dei soggetti da controllare (*tax fraud detection*). I modelli predittivi utilizzati, per la loro accuratezza, espressività e comprensibilità, sono stati alberi di decisione ([BGMP99], [YQJ03], [AT12]), applicati agli scenari più vari (Italia, Cina, Marocco, rispettivamente), a testimonianza del fatto che il problema trattato, lungi dall'essere solo italiano, è presente in molti altri Paesi. Altri lavori, come [WOLCY12] o [BGMPS09] hanno analizzato invece i benefici derivanti dall'impiego di regole di classificazione, applicate in special modo in

1 L'Agenzia delle Entrate è un ente pubblico non economico italiano che svolge le funzioni relative alla gestione, all'accertamento e al contenzioso dei tributi con l'obiettivo di perseguire il massimo livello di adempimento degli obblighi fiscali. È' sottoposta alla vigilanza del Ministero dell'economia e delle finanze, che ha la responsabilità dell'indirizzo politico, ed è dotata di autonomia regolamentare, amministrativa, patrimoniale, organizzativa, contabile e finanziaria. I rapporti tra il Ministero e l'Agenzia sono regolati dalla Convenzione triennale in cui sono indicati i servizi da assicurare, gli obiettivi da raggiungere e le risorse destinate a queste finalità. L'Agenzia delle Entrate, operativa dal 1° gennaio 2001, nasce dalla riorganizzazione dell'Amministrazione finanziaria a seguito del D. Lgs. 300 del 1999. Dal 1° dicembre 2012 l'Agenzia dell'Entrate ha incorporato l'Agenzia del Territorio (articolo 23-quater del Dl 95/2012).

ambito IVA (i casi di studio hanno riguardato Taiwan e Italia, rispettivamente). Tutti i lavori presentati concludono mostrando i potenziali benefici per gli enti impositori coinvolti.

1.3 Contenuto della tesi

La tesi, partendo dalla descrizione del fenomeno dell'evasione fiscale in Italia, si pone il problema di valutare se tecniche di *data mining* e, in particolare, di classificazione, possano essere applicate con successo in questo ambito.

Nel Capitolo 2 introdurremo alcuni fatti relativi al fenomeno dell'evasione fiscale in Italia, analizzandone le principali determinanti e conseguenze, concludendo con l'indicazione dei risultati raggiunti dall'Agenzia delle Entrate per mezzo della sua azione di prevenzione e contrasto (relativamente all'anno 2011).

Il Capitolo 3 contiene una prima analisi, multidimensionale, del *dataset* su cui saranno successivamente impiegate tecniche di *data mining*.

Nel Capitolo 4 presenteremo lo stato dell'arte in merito al tema della *tax fraud detection*, osservando come spesso, in questo ambito, vengano utilizzate tecniche di classificazione e, in particolare, alberi e regole decisionali, a causa della loro elevata espressività e comprensibilità.

Nel Capitolo 5 viene presentato un approccio al *dataset* in ottica di *mining*, attraverso l'utilizzo di tecniche di classificazione. L'analisi dei dati è stata svolta mediante le due famiglie di algoritmi, alberi e regole di decisione. Tale analisi ha evidenziato come classificatori opportunamente allenati possano ottenere buone *performances*, anche in termini di gettito recuperato. Non solo, ma l'introduzione di adeguati *lift chart* associati ai modelli, fornisce all'utilizzatore uno strumento per selezionare, sulla base dei propri obiettivi, un numero desiderato di soggetti da sottoporre a controllo. L'implementazione *software* del processo descritto in questo capitolo – preparazione dei dati (compresa una fase di selezione dei soggetti per attività svolta), generazione e confronto di modelli, generazione dei *lift chart* – può senz'altro costituire un nuovo strumento di lotta all'evasione per questo Paese, efficace e flessibile.

Nel Capitolo 6 alcune brevi conclusioni personali chiuderanno il presente lavoro.

Capitolo 2

Evasione fiscale in Italia: un quadro generale

In questo capitolo introdurremo alcuni fatti relativi al fenomeno dell'evasione fiscale in Italia, a cominciare da una sua stima in termini monetari. Analizzeremo poi le principali determinanti del fenomeno evasivo, astraendoci dal contesto italiano e facendo riferimento ad uno studio dell'Ocse del 2010 che ci indicherà le linee guida per l'ottenimento, in generale, della massima "tax compliance" possibile. Tali linee guida saranno successivamente calate nel contesto italiano e, poiché la presenza di norme sociali favorevoli alla compliance è fondamentale, porremo l'accento sull'atteggiamento degli italiani nei confronti del fisco, sulla scorta di due indagini effettuate nel 2007 e 2010, rispettivamente, da Banca d'Italia e Censis. Infine, cifre alla mano, vedremo quanto dichiarano gli italiani (facendo riferimento agli ultimi dati resi pubblici dal Ministero delle Finanze) e che risultati – in termini di recupero dell'evasione – sono raggiunti dall'Agenzia delle Entrate per mezzo della sua azione di prevenzione e contrasto.

2.1 Entità dell'evasione

L'evasione fiscale in Italia è un fenomeno molto diffuso, che produce gravi conseguenze sia dal punto di vista economico che dal punto di vista sociale.

Manca, tuttavia, una quantificazione "ufficiale" del fenomeno, per sua natura sfuggente, che trova nell'occultamento dell'imponibile la sua ragione ed il suo stesso presupposto. Si aggiunga che l'evasione fiscale può assumere diverse forme, a seconda delle imposte, dei soggetti interessati e delle modalità con le quali vengono effettuate le transazioni. La teoria economica ha elaborato diverse metodologie di stima dell'evasione fiscale e dell'economia sommersa, raggruppabili in

due grandi famiglie: metodi diretti ed indiretti, descritti in [Pal04]. I metodi diretti si basano su dati prettamente microeconomici rilevati, tramite indagini campionarie, su famiglie e imprese oppure tramite il confronto tra i dati desunti dall'attività di vigilanza tributaria (assunti come "veri") e quelli dichiarati dai contribuenti. I metodi indiretti, al contrario, ricavano l'entità dell'economia sommersa sia attraverso modelli economico-statistici, che hanno l'obiettivo di fornire una quantificazione dell'economia sommersa, sia attraverso approcci macroeconomici, che mirano al raggiungimento di una stima verosimile del PIL. Va rilevato tuttavia come i primi presentino limiti riconducibili all'incerto grado di rappresentatività dei campioni utilizzati e per questo motivo sono meno impiegati rispetto ai metodi indiretti.

I numeri dell'evasione fiscale nel nostro Paese sono comunque imponenti. Con riferimento all'evasione di IVA e IRAP, la Corte dei Conti in [Cdc13] stima per il solo 2011 un mancato introito di circa 50 miliardi di euro. A ciò deve aggiungersi l'evasione delle altre imposte (tra cui IRE – ex IRPEF ed IRES) e dei contributi sociali. Recenti stime disponibili, riportate in [Conf10], riferite al 2009, propongono una evasione totale in Italia attestata su valori superiori ai 125 miliardi di euro.

Ad analoghe quantificazioni pervengono anche elaborazioni effettuate dal Gruppo di lavoro "Economia non osservata e flussi finanziari" istituito dal Ministero dell'Economia e delle Finanze nel 2010 (e presieduto dal Presidente dell'Istat) in [Mef11], le quali riportano, per il 2008, il valore aggiunto del sommerso economico tra il 16,3% e il 17,5% del Pil (pari a 255-275 miliardi di euro). Occorre precisare però che il *sommerso* non coincide con l'imponibile fiscale non dichiarato, ma costituisce, ciò non di meno, la base di partenza "naturale" per la stima dell'evasione, che si aggira intorno ai 120 miliardi di euro annui.

L'evasione fiscale costituisce un problema rilevante per diversi motivi, come evidenziato in [CaDal07]: riduce il gettito fiscale, generando problemi nel bilancio dello Stato (con sottrazione di risorse fondamentali per lo sviluppo ed il *welfare*) e indirizza il prelievo sulle basi imponibili che è meno agevole sottrarre alla tassazione (ad esempio, il lavoro dipendente). Inoltre, tende a produrre disparità di trattamento tra soggetti con uguale capacità contributiva (c.d. iniquità orizzontale), alterando il principio di progressività cui il sistema tributario si informa (art. 53 Costituzione) e minando gli elementi di coesione all'interno della collettività (non va infatti dimenticata l'ingiustizia sociale generata dalla fruizione di beni e servizi pubblici da parte di soggetti che non concorrono al loro mantenimento, ma anzi, sfruttano il concorso altrui). Infine, introduce distorsioni tra gli operatori economici, alterando le condizioni di concorrenza sui mercati, con riflessi negativi sull'efficienza del sistema.

Una conferma di dette distorsioni si ha considerando i dati relativi alle varie tipologie di contribuenti. Facendo riferimento a quanto riportato in [Zam12], si scopre che sul totale delle imposte riscosse, le imprese concorrono con il 5% (dato del 2007; nel 1993, il medesimo dato fu del 13%), mentre i liberi professionisti sono passati dal 7,6% del 1993 al 4,2% del 2007. Di contro, salariati e pensionati sono le due categorie che hanno visto aumentare, negli ultimi 15 anni, la loro percentuale di incidenza sul totale riscosso. Ancora, su 42 milioni di contribuenti residenti in Italia, il reddito medio dichiarato è di circa 19.000 euro. Di questi, solo 400.000 soggetti dichiarano più di 100.000 euro l'anno e, all'interno di tale gruppo, il 70% è costituito da lavoratori dipendenti; il 20% da lavoratori autonomi e il 5% da imprenditori. E' di tutta evidenza che i dati citati evidenzino delle forti anomalie nel "sistema Italia".

L'obiettivo di ridurre, quindi, in maniera significativa e strutturale, le aree di evasione e di agevolare la scelta del contribuente verso l'adempimento spontaneo dei propri obblighi fiscali (c.d. *tax compliance)* diviene strategico per il Paese e richiede, per il suo raggiungimento, una visione coordinata e complessiva delle diverse variabili che influenzano il livello di fedeltà fiscale in una nazione.

Ci si può allora domandare quali siano i fattori alla base dell'evasione e quali possano essere le leve da manovrare nell'ambito di politiche di *tax compliance.* Va da sé che una migliore comprensione del comportamento dei contribuenti può sperabilmente mettere l'amministrazione finanziaria in una posizione più forte per progettare e attuare efficaci strategie di *compliance*, che contribuiscano alla sostenibilità dei sistemi di imposizione.

Va tuttavia osservato che se il fenomeno dell'evasione in Italia è strutturale e ormai di lungo corso, le scelte legislative sul suo contrasto hanno subito, nel nostro Paese, vicende alterne, con impostazioni a volte divergenti. Si pensi, in proposito, al ricorso che è stato fatto alle politiche dei condoni e degli "scudi fiscali"; all'introduzione e successiva abolizione di strumenti di controllo come l'obbligo di allegare alla dichiarazione IVA l'elenco clienti e fornitori (peraltro reintrodotto dal D.L. 31.05.2010 n. 78, significativamente modificato successivamente dal D.L. 2 marzo 2012, n. 16, convertito con la legge 26 aprile 2012, n. 44), ai requisiti fissati per la tracciabilità dei pagamenti e la limitazione dell'uso del contante, più volte modificati; alle norme in costante evoluzione sugli studi di settore, e così via. Tali andamenti "ondivaghi e contraddittori" sono peraltro stati stigmatizzati anche dalla Corte dei Conti, [Cdc13].

2.2 Le principali determinanti dell'evasione

In un interessante studio proposto da [CaDAl07], vengono descritti i principali fattori che possono influenzare l'entità del fenomeno evasivo.

L'evasione è un fenomeno complesso che nasce dall'interazione dei comportamenti di diversi attori: il legislatore che fissa le norme fiscali e contributive; l'amministrazione che ne cura gli aspetti applicativi, promuovendo e imponendo il rispetto delle regole; i cittadini, che da un lato modificano i loro comportamenti in risposta alla legislazione e alle modalità pratiche con cui viene attuata, e dall'altro, in quanto elettori, determinano, seppure indirettamente, la legislazione vigente e condizionano le prassi dell'amministrazione.

L'evasione deriva in primo luogo da *comportamenti opportunistici*: di fronte all'obbligo del pagamento delle tasse, l'individuo valuta la strategia ottimale da tenere sulla base dell'ammontare dell'imposta dovuta – a sua volta dipendente dal livello del reddito e delle aliquote –, della sua propensione al rischio, della probabilità di subire un controllo e dell'ammontare delle sanzioni previste, come osservato in [AS72]. L'evasione risente dunque anche dell'efficienza, della capacità di accertamento dell'Amministrazione pubblica e delle modalità con cui l'accertamento viene realizzato.

Fenomeni di evasione possono però essere determinati anche da *norme che non riguardano solo aspetti fiscali*. La regolamentazione in generale impone costi e vincoli, che in talune situazioni possono indurre non solo le imprese, ma anche i lavoratori, a ricorrere a forme di attività non ufficiali (o sommerse), determinando per quella via fenomeni di evasione fiscale, come spiegato in [SK04]. Per esempio, vincoli all'assunzione di lavoratori non dotati di permesso di soggiorno possono riflettersi in attività sommerse, che a loro volta implicano forme di evasione fiscale; simili effetti possono derivare anche dal divieto di svolgere secondi lavori o da norme particolarmente restrittive in materia di tutela dei lavoratori e dell'ambiente. Inoltre, dove le attività irregolari sono diffuse, gli amministratori pubblici possono essere restii ad applicare misure realmente efficaci di contrasto, o perché la riduzione del sommerso e dell'evasione fiscale potrebbe tradursi in un aumento della disoccupazione e quindi in una perdita di consenso, oppure perché gli amministratori stessi riconoscono l'inadeguatezza delle norme alla realtà locale. In questi casi i fenomeni di evasione sono il risultato di un'incoerenza tra le norme formali, il contesto in cui si applicano e i comportamenti dell'Amministrazione.

L'evasione è un fenomeno correlato anche con le *caratteristiche della struttura produttiva* di un Paese. In Italia il sistema produttivo è particolarmente frammentato e l'incidenza dei lavoratori indipendenti (il c.d. "popolo delle partite IVA") sul totale dell'occupazione è assai più elevata che in altri paesi europei. Nel confronto internazionale, la correlazione tra la quota di occupati indipendenti da un lato, e l'evasione e il sommerso economico dall'altro, è positiva e significativa. Come osservato in [CCD95], la piccola dimensione d'impresa e l'elevata diffusione del lavoro indipendente accentuano le difficoltà dell'Amministrazione finanziaria nell'esercitare i controlli. L'elevato numero di soggetti da controllare richiederebbe un'azione di

accertamento diffusa, più capillare e più costosa[1]. La maggiore facilità con cui una piccola impresa, scarsamente trasparente, può evadere, finisce poi col divenire uno dei fattori di disincentivo alla crescita dimensionale delle imprese e alla adozione di forme giuridiche che impongono una maggiore trasparenza nei confronti del mercato.

L'evasione dipende anche *dall'interazione del comportamento dei contribuenti e dell'Amministrazione fiscale*. In talune situazioni, l'Amministrazione, nella consapevolezza che alcuni contribuenti tenderanno ad evadere le imposte, può essere indotta a fissare aliquote elevate, che di fatto saranno applicate a valori largamente sottostimati rispetto a quelli reali. In altri casi, può essere disposta a tollerare fenomeni di evasione, subendo talvolta le pressioni di gruppi politicamente importanti. Forme di scarsa trasparenza dell'Amministrazione possono a loro volta riflettersi sul grado di fiducia nelle istituzioni e sull'equità del sistema fiscale, generando per questa via atteggiamenti più favorevoli all'evasione. Questi comportamenti possono essere rafforzati nelle situazioni di forte eterogeneità delle preferenze in merito al livello del prelievo e della spesa pubblica[2].

Le decisioni degli individui in tema di evasione sono inoltre condizionate dalle loro *convinzioni in merito ai doveri fiscali*. Come si rilevava già in [Cro44], l'obbligo etico del pagamento delle imposte si fonda sul concetto di tassa "giusta". Una tassa è giusta – e la sua evasione è eticamente condannabile – quando è imposta da un'autorità legittima, per una giusta causa ed è ripartita in modo equo. Tali criteri, applicati al sistema di tassazione delle moderne società occidentali, sono suscettibili di valutazioni di vario genere. In primo luogo la legittimità dell'autorità fiscale in un sistema democratico non è ragionevolmente messa in discussione. La destinazione delle risorse acquisite tramite l'imposizione è invece ovviamente più discutibile; il modo in cui la spesa pubblica viene allocata è espressione della maggioranza politica di volta in volta al governo. *Se la destinazione della spesa trova scarso consenso in ampie fasce della popolazione, si riduce il movente etico del pagamento delle imposte*[3]. Affinché la tassazione sia percepita come "giusta" è inoltre essenziale che sia *ripartita equamente tra la popolazione*. In questo senso, la diffusa convinzione che altri evadano è già di per sé uno stimolo ad ulteriore evasione. Il frequente ricorso dello Stato ai condoni, ad esempio, non

[1] Sarebbe errato, d'altra parte, ritenere che la piccola dimensione comporti una maggiore facilità di controllo. La minore presenza di controlli amministrativi interni all'azienda e la maggiore possibilità di comportamenti collusivi con i dipendenti, i fornitori e i clienti fanno sì che, nelle imprese di piccole dimensioni, risultanze contabili formalmente corrette possano frequentemente celare situazioni di fatto molto diverse.

[2] Sull'importanza di fattori quali la percezione di equità del sistema fiscale, il giudizio del contribuente in merito alla qualità della spesa pubblica e alla complessità del sistema fiscale, si vedano [AEF98] e la letteratura ivi citata.

[3] Vi è evidenza che un comportamento della pubblica Amministrazione orientato all'equità, alla correttezza e al rispetto nei confronti dei cittadini e un utilizzo delle risorse pubbliche per scopi socialmente desiderabili accrescono il grado di tax compliance. Cfr. per tutti, [CMMT04].

può che peggiorare questa situazione, causando aspettative di impunità e minando il consenso alla base della tassazione. La *compliance* dei contribuenti può essere inoltre intaccata, per questa via, da aliquote troppo alte, da modalità di esazione complicate, da adempimenti irragionevolmente costosi.

Le decisioni degli individui sono infine condizionate dalle *norme sociali* che si affermano all'interno di una comunità: un ambiente sociale che abbia fatto proprie norme di onestà e di "buon comportamento" tenderà a sanzionare gli individui che non rispettano le norme stesse. Con un rischio di "sanzione sociale" sufficientemente forte e un costo dell'esclusione elevato, il comportamento del contribuente potrebbe essere corretto, in linea teorica, anche in assenza di controlli da parte dell'Amministrazione finanziaria.

2.3 Fattori che influenzano la *tax compliance*

Analizzate le principali cause dell'evasione fiscale, ci si può chiedere quali azioni possano essere messe in campo dalle amministrazioni ai fini di contenerne gli effetti sull'economia.

Come accennato in precedenza, in [Oecd10], sono proposti sei fattori in grado di influenzare la *compliance*: deterrenza, norme, opportunità, imparzialità e fiducia, fattori economici. Analizziamoli ora in dettaglio.

2.3.1 Deterrenza

Abbiamo visto come l'evasione può derivare da comportamenti opportunistici dei contribuenti. Sotto questa ipotesi, si può ragionevolmente ritenere che l'effettuazione di molti controlli fiscali, con conseguente aumento della probabilità di subire un accertamento, unitamente a gravi sanzioni comminate agli evasori, costituiscano i fattori principali con cui le agenzie fiscali dei vari Paesi possono promuovere la *compliance* dei propri cittadini–contribuenti, inducendoli a limitare i propri comportamenti opportunistici. Non solo, ma la sola presenza di una maggiore deterrenza può scoraggiare, in generale, i contribuenti dal tenere comportamenti fraudolenti (c.d. esternalità positiva dei controlli).

Il modello standard sugli effetti della deterrenza è stato presentato in [AS72] e si basa sull'ipotesi che i contribuenti siano agenti economici razionali che agiscono ottimizzando il proprio interesse. Tuttavia, questo modello è stato criticato in letteratura perché, empiricamente, è stato osservato che i contribuenti sono, in generale, più *compliant* di quanto detto modello non predica: in [AS72] viene spiegato che normalmente le probabilità di essere scoperti sono talmente basse che la scelta razionale per il singolo sarebbe quella di evadere. Tuttavia, successivamente, diversi autori hanno evidenziato come anche altri fattori guidino le decisioni economiche – e in particolare quelle in condizioni di incertezza – quali elementi

intuitivi, la percezione soggettiva dei fenomeni e il comportamento degli altri membri della comunità. Il ruolo delle convinzioni, della cultura e delle credenze come fattori esplicativi del comportamento degli agenti è stato oggetto di rinnovata attenzione nel filone della cosiddetta *cultural economics* [GSZ06].

In questo quadro, a partire dall'evidenza che in numerosi Paesi il livello di fedeltà fiscale non era spiegato adeguatamente dai modelli basati sul solo principio di deterrenza, gli studi sull'evasione fiscale hanno iniziato a considerare più ampi schemi interpretativi.

Invero, diversi studi forniscono evidenze contraddittorie sul fatto che le strategie di dissuasione, da sole, possano avere un'influenza effettiva sul comportamento desiderato dei contribuenti.

La deterrenza può rafforzare l'obbligo morale di pagare le tasse perché può indicare ciò che è *giusto* fare. Ma la deterrenza può anche creare delle resistenze nel contribuente, generando in esso sentimenti di oppressione che lo portano alla devianza. Ne deriva che la deterrenza può avere effetti positivi o negativi sulla *compliance*.

La questione, a ben vedere, non è se le agenzie fiscali debbano o meno porre in essere strategie di deterrenza, ma il modo in cui lo fanno, che deve essere il più efficace possibile, tenuto conto del contesto nel quale la stessa viene esercitata. Di particolare importanza per il successo di una strategia di deterrenza sono le *norme personali* (valori, regole di condotta, etica di ciascuno) e *sociali* (corrispondenti ai comportamenti diffusi) esistenti nel tessuto economico-sociale di riferimento, come osservato in [Wen04]. In particolare, quando le norme personali in favore della *compliance* sono forti, la deterrenza può non esplicare effetti di rilievo, perché i contribuenti rispettano le leggi (fiscali) in quanto ritengono che sia la cosa giusta da fare, non tanto per timore delle sanzioni[4]. Ma quando le norme personali sono deboli, allora la deterrenza può divenire un fattore importante per favorire la *compliance*: se il contribuente non è guidato da una obbligazione morale a pagare le tasse, la minaccia della sanzione può avere un impatto positivo sul suo comportamento. Anche i comportamenti e valori diffusi sono importanti, perché se "in generale" il clima non è favorevole alla *compliance,* allora la deterrenza avrà un effetto minore, considerato che le sanzioni formali sono più efficaci se ad esse viene riconosciuto anche un valore sociale. Esiste tuttavia un'influenza anche nella direzione opposta, ovvero la deterrenza può supportare e modificare le norme sociali, segnalando quali siano i comportamenti eticamente desiderabili, rassicurando chi vi aderisce sul fatto che sta agendo bene ed esponendo allo stesso tempo l'evasore al rischio dell'isolamento o della stigmatizzazione. E' stato dimostrato che le persone sono più desiderose di comportarsi correttamente se percepiscono che anche altri si comportano allo

4 Pur senza arrivare a sostenere che "le tasse sono una cosa bellissima, un modo civilissimo di contribuire tutti insieme a beni indispensabili quali istruzione, sicurezza, ambiente e salute", per citare il ministro dell'economia Tommaso Padoa Schioppa, scomparso nel 2010.

stesso modo. Ma perché ciò avvenga, occorre che la strategia di deterrenza adottata sia percepita come "giusta" dai contribuenti e non vessatoria e che si eviti di ingenerare l'idea che l'evasione fiscale sia un fenomeno pervasivo e una prassi comune. In merito a questo rischio, compito delle Amministrazioni fiscali, e in particolare delle strutture di comunicazione, è di evitare queste generalizzazioni e chiarire che l'evasione fiscale è un comportamento limitato a una certa fetta di contribuenti, debitamente monitorati.

2.3.2 Norme

L'influenza delle norme sul comportamento del contribuente si esplica a un doppio livello, individuale (convinzioni personali su ciò che è giusto/sbagliato) e sociale (convinzioni e credenze diffuse all'interno dell'ambiente sociale di riferimento). L'aspetto decisivo in questo caso è l'effetto-moltiplicatore, vale a dire il condizionamento esercitato sull'individuo dai comportamenti, positivi o negativi, mostrati dai consociati. In ambito fiscale, la percezione che le persone, soprattutto quelle più vicine, siano più o meno inclini a pagare le tasse incide pesantemente sull'atteggiamento del singolo.

Lo studio Ocse del 2010 fornisce alcune indicazioni pratiche per gestire questi aspetti: in primo luogo, può essere efficace incorporare in ogni attività di impatto esterno (campagne informative, lettere ai contribuenti, sito internet, contatti faccia a faccia) un messaggio normativo di rinforzo alla *compliance*, tenendo presente che le norme, una volta interiorizzate, sono difficili da modificare. Può allora mostrarsi conveniente concentrarsi su chi non ha ancora un *set* di valori e comportamenti strutturato, come fa ad esempio l'Agenzia delle Entrate con il progetto "Fisco e Scuola"[5]. Meglio ancora se si scelgono come linee-guida della comunicazione dei valori-cardine (es. diritto all'istruzione o alle cure mediche), universalmente riconosciuti, legandoli al pagamento delle tasse, oppure degli esempi positivi in cui i destinatari possano immedesimarsi.

2.3.3 Opportunità

Altro fattore cruciale che influenza il comportamento del contribuente è legato alle possibilità di adempiere facilmente ai propri obblighi fiscali oppure di evadere. Tradizionalmente, le amministrazioni fiscali si sono concentrate su questo secondo versante, mentre sarebbe più proficuo, secondo il gruppo di lavoro dell'Ocse, agevolare il pagamento delle tasse riducendo gli ostacoli che il contribuente incontra, in termini di complessità delle procedure e di

5 Il progetto è nato nel 2004 con l'intento di diffondere nelle nuove generazioni la cultura della legalità fiscale, rendendole consapevoli del loro ruolo di futuri contribuenti. Tra le attività proposte agli studenti delle scuole inferiori e superiori, sono compresi cicli di lezioni con funzionari dell'Agenzia delle Entrate, visite guidate presso gli Uffici, convegni e seminari.

tempo da impiegare. Ad esempio, è possibile incrementare la *compliance* con un linguaggio comprensibile anche ai non addetti ai lavori o attraverso siti internet accessibili o ancora ricorrendo a modelli di dichiarazione di facile compilazione, infine - ma su questo si riconosce che il margine di intervento è molto ridotto - riducendo la complessità dell'impianto normativo. Viceversa, in concomitanza con altri fattori, "*leggi complesse, modelli incomprensibili, siti poco accessibili, call center sempre occupati, impiegati sgarbati*" possono spingere il contribuente verso comportamenti devianti.

2.3.4 Equità

Secondo un ampio *corpus* di ricerche, la *tax compliance* è influenzata anche dalla percezione di equità (che può essere distributiva, nella gestione del denaro pubblico; procedurale, nella gestione delle attività di controllo e di assistenza da parte dell'amministrazione fiscale; sanzionatoria, riferita alle punizioni previste per chi evade). Sanzioni poco trasparenti e atteggiamenti percepiti come aggressivi o non equi da parte dell'amministrazione fiscale possono condizionare negativamente il contribuente. Il compito delle burocrazie del fisco è in particolare quello di garantire l'equità procedurale: neutralità, precisione, coerenza, empatia e rispetto nei confronti del contribuente, in particolare nelle situazioni in cui è sottoposto a controllo, diventano fondamentali.

2.3.5 Fattori economici

Le relazioni tra dinamiche economiche ed evasione fiscale non sono univoche. In linea di tendenza, il comportamento deviante si associa alla mancanza di lavoro e ai fenomeni di economia sommersa, o anche a condizioni di lavoro particolarmente disagiate, mentre i fattori che determinano la crescita economica in genere favoriscono una maggiore *compliance*.

2.3.6 Interazioni fra fattori

Non deve essere, infine, trascurata la necessità di comprendere meglio come i vari fattori che guidano i comportamenti dei contribuenti interagiscono tra loro. In particolare, le agenzie fiscali devono essere prudenti quando implementano le loro politiche per indirizzare il comportamento dei contribuenti. Ad esempio, politiche troppo invasive possono indurre il contribuente a ritenere che lo Stato non si fidi di lui. La ricerca dell'Ocse mostra che quando questo accade, il contribuente adotta lo stesso atteggiamento verso il fisco e ciò può ridurre la *compliance*. Un'agenzia fiscale deve inviare un segnale chiaro al pubblico che comportamenti non conformi sono visti dalla società come sbagliati. Suggerendo che l'intera società (e non solo l'erario) considera determinati comportamenti sbagliati, rafforza le norme personali.

Data la domanda di una maggiore efficienza ed efficacia dell'azione amministrativa, diventa sempre più importante l'attività di ricerca volta a comprendere il comportamento dei contribuenti. Attraverso una migliore comprensione del comportamento dei contribuenti, le agenzie fiscali possono utilizzare in modo più efficace le loro limitate risorse per sviluppare strategie che avranno un impatto reale e sostenibile sul comportamento dei contribuenti. Senza dimenticare che un aumento della *compliance* permette ai governi di disporre di maggiori risorse da destinare alla spesa.

2.4 La *tax compliance* in Italia

Rivisitiamo ora in chiave italiana i principali fattori che influenzano la *compliance* visti nel paragrafo precedente.

In primo luogo, la *deterrenza*. La relazione tra deterrenza e *tax compliance* abbiamo visto essere molto complessa. I controlli, per essere efficaci e non generare ostilità devono essere mirati verso i contribuenti più a rischio e puntare alla qualità, tenendo allo stesso tempo ben presente i diritti dei contribuenti.

È proprio questo uno degli obiettivi strategici dell'Agenzia delle Entrate. Nella circolare che detta gli indirizzi annuali 2013 sull'attività di controllo (Circolare 25/E del 2013), l'Agenzia sottolinea come

> *la qualità e l'efficacia dell'attività di controllo dipendono, infatti, da una selezione accurata delle posizioni soggettive da sottoporre a controllo tra quelle individuate a seguito dell'analisi del rischio, nonché da un'adeguata attività istruttoria*

e

> *l'attività di controllo, oltre al recupero delle somme evase e all'irrogazione delle relative sanzioni, è mirata altresì a dissuadere i contribuenti da comportamenti fiscalmente non corretti e ad interrompere condotte illecite di frode fiscale messe in atto, in molti casi, sin dall'avvio dell'attività economica. In altri termini, i risultati efficaci derivanti dalla complessiva azione di prevenzione e contrasto all'evasione, oltre al necessario recupero dell'evasione pregressa, dovranno produrre un progressivo incremento dell'adempimento spontaneo (c.d. "compliance").*

L'efficacia dei controlli presuppone una mappatura del territorio attraverso gli applicativi informatici a disposizione, il censimento dei rischi di evasione e/o elusione e la individuazione delle posizioni di rischio da selezionare, graduate in funzione di un *risk score* attribuito: questi temi accompagnano l'intera circolare di indirizzo con riferimento alle singole categorie di contribuenti (grandi, medi e piccoli).

Per quanto riguarda la *riduzione delle opportunità di inadempimento*, sono state introdotte, negli ultimi anni, nuove norme anti evasione. Di seguito le più significative:

- *Indagini sui conti correnti*: il D.L. 06.12.2011 n. 201 (decreto "salva Italia", convertito nella Legge 214/2011) ha rafforzato lo strumento delle indagini finanziarie a disposizione degli Uffici finanziari, prevedendo, all'art. 11 (tra le misure "*per l'emersione della base imponibile e la trasparenza fiscale*") che a decorrere dal 01.01.2012, le banche e gli operatori finanziari comunicassero periodicamente (a regime, entro il 20 aprile dell'anno successivo a quello cui la comunicazione si riferisce) all'anagrafe tributaria le informazioni attinenti ai rapporti finanziari e delle operazioni fuori conto poste in essere dai contribuenti loro clienti. L'attuazione della norma è stata tuttavia in attesa di un apposito Provvedimento dell'Agenzia delle Entrate, che stabilisse le modalità di trasmissione dei dati. Detto provvedimento è stato approvato dal Direttore dell'Agenzia delle Entrate in data 25 marzo 2013; successivamente, dal 24 giugno, gli operatori hanno cominciato ad inviare le informazioni relative ai rapporti attivi nel 2011. In particolare gli intermediari finanziari dovranno segnalare:
  - ✓ i dati identificativi del rapporto, compreso il codice univoco, riferito al soggetto persona fisica o non fisica che ne ha la disponibilità e a tutti i cointestatari (nel caso di intestazione a più soggetti);
  - ✓ i dati relativi al saldo iniziale al 1° gennaio e al saldo finale al 31 dicembre. Per i rapporti avviati nel corso dell'anno il saldo iniziale dovrà tener conto della data di apertura, mentre per quelli chiusi nel corso dell'anno il saldo andrà contabilizzato al momento della data di chiusura;
  - ✓ i dati relativi agli importi totali delle movimentazioni distinte tra dare e avere per ogni tipologia di rapporto conteggiati su base annua.

  Dovranno essere segnalati oltre ai conti correnti, anche i conti deposito di risparmio, le gestioni individuali e collettive (i fondi), i rapporti fiduciari, le cassette di sicurezza, gli acquisti e le vendite di oro e di metalli preziosi, i finanziamenti, le garanzie e anche le polizze assicurative. Sono esclusi, invece, i fondi pensione. La fondamentale differenza esistente tra la "nuova" anagrafe e la "vecchia" anagrafe dei conti correnti è rappresentata dal fatto di aver introdotto l'obbligo, per gli istituti di credito, di inviare all'Agenzia delle Entrate non più solo dati sul l'esistenza del rapporto bancario o finanziario ma anche informazioni sulla sua consistenza. Ciò al fine sia di agevolare i compiti di accertamento fiscale da parte dell'amministrazione finanziaria sia di evidenziare la ricchezza finanziaria complessiva del contribuente anche in vista della possibile introduzione di una patrimoniale. Inoltre, dette comunicazioni avranno un immediato utilizzo, nella

misura in cui, come affermato dal Direttore dell'Agenzia nell'audizione del 31 ottobre 2012 alla Commissione parlamentare di vigilanza sull'anagrafe tributaria, "*le informazioni contribuiranno alla selezione delle posizioni da controllare*" nella lotta all'evasione.

- *Tracciabilità dei pagamenti*: per garantire maggiore tracciabilità ai mezzi di pagamento e contrastare l'evasione, nelle operazioni tra privati o nelle transazioni tra consumatori e imprese, non può essere utilizzato denaro contante se i pagamenti sono relativi ad importi pari o superiori a 1.000 euro. La stessa limitazione si applica agli assegni, bancari o circolari, privi della clausola di "non trasferibilità" e senza indicazione del beneficiario (norma introdotta dal citato decreto "salva Italia", art. 12, che ha modificato l'art. 49 del D.Lgs. 231/2007, c.d. normativa antiriciclaggio).
- *Spesometro ed elenco clienti-fornitori*: l'art. 21 del D.L. 31.05.2010 n. 78 aveva previsto l'obbligo, per i soggetti passivi IVA, della trasmissione telematica all'Anagrafe tributaria della "*comunicazione telematica delle operazioni rilevanti ai fini dell'imposta sul valore aggiunto, di importo non inferiore a euro tremila" (c.d. spesometro)*. Tale articolo di legge è stato successivamente più volte oggetto di modifiche, fino alla versione attuale, dettata dalle previsioni contenute nell'art. 2 D.L. 16/2012, convertito nella L. 44/2012.

  In virtù di detto intervento legislativo, "*l'obbligo di comunicazione delle operazioni rilevanti ai fini dell'imposta sul valore aggiunto per le quali è previsto l'obbligo di emissione della fattura è assolto con la trasmissione, per ciascun cliente e fornitore, dell'importo di tutte le operazioni attive e passive effettuate. Per le sole operazioni per le quali non è previsto l'obbligo di emissione della fattura (ad esempio: prestazioni alberghiere e di ristorazione, commercio al dettaglio) la comunicazione telematica deve essere effettuata qualora le operazioni stesse siano di importo non inferiore ad euro 3.600, comprensivo dell'imposta sul valore aggiunto […]"* (comma 1). Inoltre, *"al fine di semplificare gli adempimenti dei contribuenti, l'obbligo di comunicazione delle operazioni di cui al comma 1, effettuate nei confronti di contribuenti non soggetti passivi ai fini dell'imposta sul valore aggiunto, è escluso qualora il pagamento dei corrispettivi avvenga mediante carte di credito, di debito o prepagate emesse da operatori finanziari soggetti all'obbligo di comunicazione previsto dall'articolo 7, sesto comma, del decreto del Presidente della Repubblica 29 settembre 1973, n. 605"* (comma 1 bis). Pertanto, "*gli operatori finanziari soggetti all'obbligo di comunicazione previsto dall'articolo 7, sesto comma del decreto del Presidente della Repubblica 29 settembre 1973, n. 605 che emettono carte di credito, di debito o prepagate, comunicano all'Agenzia delle entrate le operazioni di cui al comma 1-bis in relazione alle quali il pagamento dei corrispettivi sia avvenuto mediante carte di credito, di debito o prepagate emesse dagli operatori finanziari stessi*" (comma 1 ter). Il termine per la presentazione della comunicazione è fissato per il 30 aprile dell'anno successivo a quello di riferimento.

In sostanza, con le modifiche introdotte, lo *spesometro* diventa una versione aggiornata del vecchio elenco clienti e fornitori, introdotto dal D.L. 223/2006 (conv. Legge 248/2006), art. 37, commi 8 e 9. In altre parole, se prima la comunicazione delle operazioni con obbligo di fattura riguardava la singola operazione e quindi l'adempimento era oggettivo, ora diventa soggettivo poiché racchiude in un'unica comunicazione tutti gli importi delle operazioni intercorse fra impresa e cliente-fornitore nel corso dell'anno. La comunicazione soggettiva deve essere fatta anche per le operazioni con emissione di fattura che coinvolgono i privati. Eliminando la soglia dei 3.000 euro, è importante sottolineare come la fattura vada emessa anche per quelle operazioni al di sotto di tale cifra, poiché la nuova versione dello *spesometro* obbliga alla comunicazione di qualsiasi operazione con fattura. Fanno eccezione le operazioni di "commercio al minuto e attività assimilate", come indicato dall'art. 22 del D.P.R. n. 633 del 26 ottobre 1972, che non prevedono fattura se non su esplicita richiesta del cliente.

- *Contrasto ai paradisi fiscali:* l'articolo 12, comma 2 del D.L. 78/2009, convertito nella L. 102/2009 prevede che, in caso di omessa dichiarazione del quadro RW, gli investimenti e le attività finanziarie detenute negli Stati a fiscalità privilegiata (di cui al decreto del Ministro delle finanze 4 maggio 1999, pubblicato nella Gazzetta Ufficiale della Repubblica italiana del 10 maggio 1999, n. 110, e al decreto del Ministro dell'economia e delle finanze 21 novembre 2001, pubblicato nella Gazzetta Ufficiale della Repubblica italiana del 23 novembre 2001, n. 273, senza tener conto delle limitazioni ivi previste) sono considerati costituiti (salvo prova contraria) mediante redditi sottratti a tassazione. Sono previsti, al contempo, inasprimenti delle sanzioni (duplicate) ed il raddoppio dei termini di decadenza per l'accertamento (quest'ultimo introdotto dal D.L. 194/2009).

Per quanto riguarda, infine, gli aspetti concernenti l'*equità*, due temi si intrecciano a questo proposito: da un lato imparzialità del sistema tributario e dall'altro l'imparzialità del trattamento, che si traduce nel senso di equità percepito da ogni contribuente nei rapporti col fisco, relativamente agli altri contribuenti.

Si segnala quindi la L. 27 luglio 2000 n. 212, c.d. statuto dei diritti del contribuente. In particolare, l'art. 11 di tale legge prevede il diritto di interpello (secondo cui "*ciascun contribuente può inoltrare per iscritto all'amministrazione finanziaria, che risponde entro centoventi giorni, circostanziate e specifiche istanze di interpello concernenti l'applicazione delle disposizioni tributarie a casi concreti e personali, qualora vi siano obiettive condizioni di incertezza sulla corretta interpretazione delle disposizioni stesse*") e l'art. 12 prevede una serie di tutele per il soggetto sottoposto ad accessi, ispezioni e verifiche, stabilendo anche che "*nel rispetto del*

principio di cooperazione tra amministrazione e contribuente, dopo il rilascio della copia del processo verbale di chiusura delle operazioni da parte degli organi di controllo, il contribuente può comunicare entro sessanta giorni osservazioni e richieste che sono valutate dagli uffici impositori. L'avviso di accertamento non può essere emanato prima della scadenza del predetto termine, salvo casi di particolare e motivata urgenza".

La tendenza alla definizione condivisa del maggiore imponibile accertabile si ravvisa poi in più fasi del procedimento di accertamento, tra cui l'accertamento con adesione, disciplinato dal D. Lgs. 218/97.

Negli accertamenti da studi di settore è prevista poi l'obbligatorietà del preventivo invito al contraddittorio (art. 10, comma 3 bis L. 146/98).

Lo stesso dicasi per l'accertamento sintetico, previsto dall'art. 38 DPR 600/73, così come novellato dal D.L. 78/2010, art. 22 comma 19: "*L'ufficio che procede alla determinazione sintetica del reddito complessivo ha l'obbligo di invitare il contribuente a comparire di persona o per mezzo di rappresentanti per fornire dati e notizie rilevanti ai fini dell'accertamento e, successivamente, di avviare il procedimento di accertamento con adesione ai sensi dell'articolo 5 del D. Lgs. 218/97*".

Infine, considerato che la fiducia nell'amministrazione fiscale deriva anche dal riconoscimento di un trattamento rispondente alla tenuta di comportamenti contributivi congrui e trasparenti, si segnala come per i soggetti che risultano "congrui" al relativo studio di settore, siano stati introdotti, in forme diverse nel corso del tempo, dei "premi", a partire dalla L. 296/06 (Finanziaria 2007), che ha aggiunto il comma 4 bis all'art. 10 L. 146/98 (inibendo l'accertamento analitico induttivo, ex art. 39 comma 1 lett. d) DPR 600/73 per chi si trovasse in particolari condizioni), fino ad arrivare al D.L. 201/2011 che ha introdotto un nuovo regime premiale collegato alla congruità e coerenza agli studi di settore (abrogando, peraltro, il citato comma 4 bis dell'art. 10 L. 146/98). Il nuovo regime premiale, che si applica a partire dall'annualità 2011, concede ai contribuenti soggetti agli studi di settore congrui e coerenti:

- preclusione de gli accertamenti basati sulle presunzioni semplici di cui all'articolo 39, primo comma, lettera d), secondo periodo, del DPR 600/73 e all'articolo 54, secondo comma, ultimo periodo, del DPR 633/72;
- riduzione di un anno dei termini di decadenza per l'attività di accertamento previsti dall'articolo 43, primo comma, del DPR 600/73 e dall'articolo 57, primo comma, del DPR 633/72; la disposizione non si applica in caso di violazione che comporta obbligo di denuncia ai sensi dell'articolo 331 del codice di procedura penale per uno dei reati previsti dal D. Lgs. 74/2000;
- la determinazione sintetica del reddito complessivo di cui all'articolo 38 del DPR 600/73 è ammessa a condizione che il reddito complessivo accertabile ecceda di almeno un terzo quello dichiarato.

2.5 Gli italiani e l'atteggiamento nei confronti del fisco

2.5.1 L'indagine della Banca d'Italia (2007)

In [CaDAl07] vengono esaminate le opinioni degli italiani sull'evasione fiscale. Alcuni dei risultati trovati confermano quanto esposto alla precedente sezione 1.2 circa le possibili determinanti del fenomeno evasivo. L'analisi viene condotta costruendo un indicatore di propensione all'evasione che assume valori mediamente più elevati per i lavoratori indipendenti (professionisti, imprenditori) che per quelli dipendenti; per questi ultimi poi la propensione a evadere risulta maggiore per gli operai e minore per i dirigenti e direttivi.

L'atteggiamento nei confronti dell'evasione fiscale tende inoltre a divenire via via meno favorevole al crescere del livello di istruzione. Per età, i giovani sono la classe che risulta più favorevole all'evasione fiscale.

Il legame tra l'età, l'istruzione, la condizione professionale e la propensione a evadere risulta essere significativo anche nei modelli di regressione multipla adottati.

Come ci si poteva aspettare, la propensione ad evadere risulta superiore nelle province caratterizzate da più elevati livelli di disoccupazione e criminalità e nelle aree dove è più bassa la qualità della pubblica amministrazione e minori le dotazioni di capitale sociale. In queste aree il lavoro irregolare è molto diffuso ed è quindi possibile che la riprovazione sociale nei confronti dell'evasione fiscale sia più contenuta che in altre aree.

Laddove la qualità della pubblica amministrazione è più elevata, la propensione ad evadere risulta più bassa. La percezione di un cattivo funzionamento della pubblica amministrazione è dunque correlata con un atteggiamento dei contribuenti meno orientato al rispetto delle regole fiscali. Anche questi risultati appaiono relativamente solidi.

Il comportamento dei singoli contribuenti, inoltre, appare significativamente influenzato da quello degli altri membri della collettività (di qui l'importanza delle *norme sociali,* sopra ricordata). A tal proposito, è stato riscontrato che la propensione ad evadere risulta influenzata sia dal comportamento delle famiglie residenti nella stessa località dell'intervistato, sia da quello delle famiglie della provincia di origine; questo risultato da un lato conferma l'importanza del contesto culturale e ambientale nella formazione dei valori, dall'altro è indice di persistenza dei medesimi valori nel tempo. Ciò è tanto più significativo se si considera che con un elevato grado di inerzia non è agevole, per le politiche volte alla diffusione di buone regole di comportamento sociale, conseguire il successo. Ciò nondimeno, osservano gli Autori, poiché alla base dell'evasione fiscale vi sono anche considerazioni etiche e influenze di contesto sociale, il recupero degli imponibili sottratti alla tassazione non può non

fondarsi, oltre che su elementi coercitivi, anche sulla rimozione dei fattori che vengono utilizzati come giustificazione dell'evasione, nonché sulla diffusione di una cultura della legalità e sull'applicazione di sanzioni anche sociali per chi viola le regole.

2.5.2 L'indagine Censis (2010)

Una recente indagine, realizzata nel 2010 dal Censis per il Consiglio Nazionale dei Dottori Commercialisti e degli Esperti Contabili, ha analizzato il rapporto degli italiani con il sistema fiscale. Pur con tutte le cautele con cui devono essere valutati i risultati che rilevano opinioni facendo scegliere agli intervistati fra risposte alternative proposte dall'intervistatore, ne emergono alcune convinzioni molto diffuse e radicate.

In questa sede si riportano solo i risultati più interessanti del rapporto.

In primo luogo, emerge che oltre la metà, il 58%, degli intervistati dal Censis ritiene che l'evasione sia aumentata negli ultimi tre anni (solo il 13,1 % ritiene che sia diminuita), ovvero dal 2007 al 2010. E questo nonostante il 45,6 % pensi che nello stesso periodo siano aumentati anche la numerosità e l'efficienza dei controlli fiscali da parte delle amministrazioni.

Gli estensori del rapporto si preoccupano di sottolineare che questo atteggiamento apparentemente illogico (pensare che l'evasione sia cresciuta pur a fronte di accertamenti più numerosi ed efficienti) sia da ricondurre al fatto che si stanno rilevando delle mere percezioni. È invece opportuno sottolineare che i due fenomeni non sono necessariamente contraddittori, ma al contrario possono in buona misura coesistere. La battaglia all'evasione si combatte infatti su più fronti: da un lato, occorre cercare di prevenirla, attraverso misure di deterrenza e altre misure che spingano al miglioramento negli adempimenti spontanei da parte del contribuente, anche grazie a una più efficace azione di supporto da parte dell'amministrazione fiscale, dall'altro si deve cercare di reprimerla, attraverso i controlli. Se si punta esclusivamente sulla repressione, può accadere che l'evasione aumenti, specie in periodi di crisi, e si può determinare il risultato paradossale che, proprio poiché l'evasione aumenta, è più facile scoprirla e, dunque, i controlli diventano più efficienti.

Inoltre, secondo il rapporto del Censis, gli italiani considerano l'evasione il fattore più critico nel rapporto tra fisco e contribuenti. Lo indicano come tale il 44,4 % degli intervistati, una percentuale più che doppia rispetto a quella di chi considera come fattore maggiormente critico l'eccessivo livello di tassazione (22 %).

L'evasione rappresenta un problema particolarmente sentito al Nord Est e al Centro mentre lo è meno al Nord Ovest e al Sud. È un problema che riscuote particolare attenzione tra i lavoratori

dipendenti, mentre tra i liberi professionisti si registra la percentuale più bassa dell'intero campione (30,7 %).

Chiamati ad esprimere un giudizio sull'evasione fiscale, la ritengono inaccettabile prioritariamente sotto il profilo morale il 43,4 % degli intervistati, mentre il 38,3 % la condanna principalmente perché arreca un danno ai cittadini onesti e alle imprese che subiscono concorrenza sleale.

Vi è però anche un 18,3 % di intervistati che ritiene l'evasione una condotta almeno in parte giustificabile. In alcuni casi, è vista come unica via di uscita per mantenere in piedi una piccola attività e mettere da parte qualche risparmio. La percentuale di intervistati che sottoscrive questa idea di un'"evasione di necessità" come causa principale dell'evasione fiscale è pari all'11 % del totale, ma raggiunge il 19 % tra i lavoratori autonomi e il 16,1 % fra i disoccupati. L'altra causa che potrebbe giustificare l'evasione fiscale, e cioè la non corrispondenza fra l'elevata pressione fiscale e la quantità e qualità dei servizi erogati dallo Stato, è considerata prioritaria solo dal 7,3% degli intervistati dal Censis (percentuale che sale al 10,3% al Nord Est).

Anche se non è considerata un fattore sufficiente a giustificare l'evasione, la mancata corrispondenza fra tasse pagate e livello dei servizi ottenuti emerge come elemento critico del nostro sistema fiscale anche da altre sezioni dell'indagine del Censis. Se è vero che l'81,1% degli italiani ritiene la pressione fiscale troppo alta, è anche vero che solo il 23% la ritiene troppo alta in assoluto, mentre il 58,1% la ritiene troppo alta solo in termini relativi e cioè in relazione ai servizi che si ottengono in cambio. La maggior parte degli intervistati (55,7%) sarebbe infatti disponibile addirittura a pagare più tasse a fronte di un aumento della qualità e quantità dei servizi.

Le convinzioni espresse dagli intervistati, anche se non sono sufficienti ad indurli a tenere un comportamento coerente in un contesto, come quello del nostro Paese, in cui l'evasione è così diffusa (più di un terzo ammette di non chiedere ricevute o fatture in nessun caso, o almeno in tutti i casi in cui ciò si traduce in un risparmio sul prezzo di acquisto di beni o servizi), forniscono tuttavia un'informazione importante a sostegno della necessità di combattere l'evasione.

Infine, le opinioni raccolte suggeriscono anche che il gettito recuperato con l'evasione fiscale deve essere utilizzato per ridurre le imposte oggi pagate e che il contrasto all'evasione ha tante più probabilità di risultare condiviso se si accompagna ad un impegno credibile a migliorare la capacità della spesa pubblica di rispondere ai bisogni dei cittadini.

2.6 Quanto dichiarano gli italiani

Visto cosa pensano gli italiani del loro rapporto col fisco, è interessante vedere anche quanto dichiarano effettivamente.

Il Ministero dell'Economia e delle Finanze - Dipartimento delle Finanze pubblica annualmente le statistiche sulle dichiarazioni fiscali presentate dai contribuenti[6].

Facciamo riferimento ai dati relativi all'anno d'imposta 2011, limitatamente alle sole persone fisiche.

Da questi dati risulta come i contribuenti che hanno assolto all'obbligo di presentazione della dichiarazione dei redditi Irpef per l'anno d'imposta 2011 - o in via diretta, attraverso i modelli Unico e 730, o come soggetti sottoposti a trattenute per opera del soggetto che eroga loro i redditi (Mod. 770) - sono stati più di 41,3 milioni.

L'analisi per ventili del numero di contribuenti, ordinati in base a valori crescenti di reddito complessivo, evidenzia che il 90% dei soggetti dichiara un reddito complessivo fino a 35.601 euro (+1,2% rispetto al 2010 in cui la stessa percentuale di soggetti dichiarava 35.166 euro) mentre l'ultimo ventile (ossia il 5% dei contribuenti con redditi maggiori) produce il 22,9% del reddito complessivo, in linea con l'anno precedente.

A livello nazionale, il reddito complessivo totale dichiarato è pari 805 miliardi di euro, mentre il reddito medio è pari a 19.655 euro. Entrambi i valori sono in aumento rispetto all'anno precedente (rispettivamente +1,5% e +2,1%), in linea con l'andamento del PIL nominale.

Se si sposta l'attenzione sul reddito complessivo del contribuente mediano, non influenzato, come quello medio, da valori *outlier* (ossia particolarmente elevati o bassi), il valore scende a 15.723 euro. Ciò significa che la metà dei contribuenti non supera il reddito annuo di 15.723 euro.

Più interessanti sono però i dati relativi a soggetti che svolgono attività di impresa o professionale, tenuti alla compilazione dello studio di settore, sempre curati dal Dipartimento delle finanze. In tale ambito i dati più recenti si riferiscono al 2010 e sono di seguito riportati. I commenti sono lasciati al lettore.

[6] Tali dati sono disponibili al sito: http://www.finanze.gov.it/export/finanze/Per_conoscere_il_fisco/studi_statistiche/dichiarazioni.html

Studi di Settore in vigore nel periodo d'imposta 2010

Analisi della congruità e della normalità economica

TOTALE CONTRIBUENTI

GRUPPO DI SETTORE	TOTALE CONTRIBUENTI			RICAVI / COMPENSI DICHIARATI OLTRE 30.000 EURO					RICAVI / COMPENSI DICHIARATI FINO A 30.000 EURO				
					CONGRUI NATURALI O PER ADEGUAMENTO		NON CONGRUI E NON ADEGUATI			CONGRUI NATURALI O PER ADEGUAMENTO		NON CONGRUI E NON ADEGUATI	
	Numero	Ricavi o Compensi medi dichiarati	Reddito medio d'impresa o di lavoro autonomo	Numero	Ricavi o Compensi medi dichiarati	Reddito medio d'impresa o di lavoro autonomo	Ricavi o Compensi medi dichiarati	Reddito medio d'impresa o di lavoro autonomo	Numero	Ricavi o Compensi medi dichiarati	Reddito medio d'impresa o di lavoro autonomo	Ricavi o Compensi medi dichiarati	Reddito medio d'impresa o di lavoro autonomo
Estrazione di minerali	7.948	387,5	16,0	7.214	453,3	29,4	334,2	-22,6	734	19,8	10,0	12,3	-3,7
Industrie alimentari, delle bevande e del tabacco	51.590	278,6	22,9	47.642	295,0	29,8	325,7	-2,2	3.948	20,1	7,2	14,6	-1,3
Industrie tessili e dell'abbigliamento	35.700	362,0	18,3	29.028	463,9	35,4	363,1	-27,3	6.672	18,1	8,6	11,9	-3,2
Industrie conciarie, fabbricazione di prodotti in cuoio, pelle e similari	12.959	533,4	31,7	11.565	618,6	46,2	492,0	-18,7	1.394	20,1	12,4	13,1	1,4
Industria del legno e dei prodotti in legno; fabbricazione di mobili	46.419	325,1	18,9	37.748	399,0	31,8	382,1	-19,5	8.671	19,8	9,3	13,2	0,7
Fabbricazione della carta e dei prodotti di carta, stampa ed editoria	21.711	465,5	24,4	19.570	592,0	43,5	329,9	-14,0	2.141	18,0	8,1	11,8	0,4
Fabbricazione di prodotti chimici, di fibre sintetiche e artificiali	3.440	1.057,0	55,7	3.337	1.204,9	88,1	752,8	-33,1	103	19,7	6,6	11,7	-0,4
Fabbricazione di articoli in gomma e materie plastiche	8.098	974,7	45,5	7.817	1.067,3	71,5	850,5	-20,5	281	19,4	9,9	14,3	1,6
Fabbricazione di prodotti della lavorazione dei minerali non metalliferi	10.587	627,1	18,8	9.138	766,9	45,1	599,2	-48,8	1.449	17,1	7,0	11,9	-3,0
Produzione di metalli e fabbricazione di prodotti in metallo	59.943	492,8	31,0	53.835	579,1	45,2	386,5	-22,5	6.108	19,7	9,6	14,4	1,3
Fabbricazione di macchine e apparecchi meccanici	49.253	660,7	45,2	46.240	752,4	62,2	434,5	-31,4	3.013	20,6	13,8	15,6	5,7

Fabbricazione di macchine elettriche e di apparecchiature elettriche ed ottiche	36.870	367,1	34,7	30.742	449,7	49,7	376,7	-7,0	6.128	19,7	12,4	15,1	5,0
Fabbricazione di mezzi di trasporto	4.285	535,4	21,4	3.802	644,0	45,5	471,2	-45,4	483	19,2	10,7	13,0	1,8
Altre industrie manifatturiere	5.324	799,4	63,5	4.945	908,7	88,0	713,7	8,2	379	17,6	8,9	11,7	1,6
Pesca, piscicoltura e servizi connessi	5.018	104,9	-1,5	3.297	155,9	2,2	146,8	-26,4	1.721	12,1	4,0	7,0	-2,5
Costruzioni	476.553	252,5	27,3	383.131	334,3	39,7	224,7	2,0	93.422	19,5	13,5	15,7	6,9
Manutenzione e riparazione di autoveicoli, motocicli, trattori agricoli	71.158	197,4	23,9	59.620	242,4	32,2	187,5	5,7	11.538	19,6	9,3	14,6	1,9
Intermediari del commercio	189.415	76,2	35,9	135.481	103,0	48,9	68,8	23,6	53.934	18,6	10,8	15,4	6,4
Strutture ricettive	33.541	375,7	16,5	29.630	417,9	28,1	452,3	-33,9	3.911	14,8	3,4	12,3	-2,4
Pubblici esercizi	199.142	159,9	15,7	184.890	180,8	22,3	140,7	-0,1	14.252	20,2	6,3	16,0	-1,9
Trasporti, magazzinaggio e comunicazioni	101.835	294,0	19,1	82.181	356,4	26,9	371,8	-3,9	19.654	22,9	11,8	17,3	6,2
Attività immobiliari	212.169	146,6	36,6	202.775	160,9	47,3	124,7	5,8	9.394	16,4	9,1	11,4	1,5
Servizi di consulenza	122.226	153,3	38,3	89.420	211,6	54,8	140,6	-1,3	32.806	16,5	11,4	11,5	4,4
Attività ricreative, culturali e sportive	33.521	240,0	16,0	27.221	317,0	25,7	195,3	-8,2	6.300	17,8	7,2	12,6	-0,5
Servizi alla persona	112.341	88,4	14,7	74.128	124,3	22,6	124,3	1,8	38.213	20,9	9,8	14,9	1,9
Altre attività di servizi	205.405	197,0	27,1	162.691	250,2	40,4	220,9	-4,2	42.714	16,8	10,5	12,2	3,3
Attività degli studi legali e notarili	111.746	115,1	67,7	71.620	171,3	101,1	147,1	57,9	40.126	16,4	10,4	10,7	5,6
Attività professionali di consulenza	112.761	101,5	54,9	81.331	136,6	73,8	78,9	35,9	31.430	17,0	10,4	12,6	6,4
Attività in materia di architettura, ingegneria ed altre attività tecniche	221.363	70,5	34,3	129.096	109,6	53,7	114,4	16,5	92.267	16,0	10,4	11,0	5,3
Attività professionali sanitarie	199.546	91,6	54,1	151.781	115,9	70,4	110,0	36,3	47.765	16,4	10,8	13,3	6,0
Altre attività professionali	16.799	89,0	38,1	11.113	124,7	58,6	140,9	17,5	5.686	15,2	11,2	12,3	5,6
Commercio all'ingrosso di materie prime agricole e di animali vivi	6.529	804,7	28,3	6.259	905,3	38,4	557,9	-7,7	270	17,3	4,5	12,4	-10,9
Commercio all'ingrosso di prodotti alimentari, bevande e	24.375	872,9	24,8	23.471	994,5	36,7	628,8	-9,2	904	15,4	3,6	11,1	0,0

tabacco													
Commercio all'ingrosso di altri beni di consumo finale	35.106	659,4	26,4	32.817	762,0	45,1	535,5	-21,5	2.289	15,9	4,5	12,1	-5,9
Commercio all'ingrosso di prodotti intermedi non agricoli, di rottami e cascami	1.964	1.307,5	45,4	1.908	1.481,7	62,6	844,5	-12,7	56	16,4	5,5	8,4	2,2
Commercio all'ingrosso di macchinari ed attrezzature	18.159	725,8	35,3	17.412	801,4	49,7	606,7	-6,4	747	16,7	5,9	13,3	-2,5
Commercio all'ingrosso di altri prodotti	22.587	720,7	36,1	21.237	853,2	53,1	504,9	-6,3	1.350	18,2	7,9	12,9	0,0
Commercio al dettaglio di prodotti alimentari, bevande e tabacco	134.482	234,4	22,2	123.515	250,0	26,5	269,9	11,3	10.967	17,1	6,8	16,0	-0,4
Commercio al dettaglio di prodotti per la persona	139.137	334,8	21,7	119.341	419,2	34,3	272,4	-7,4	19.796	18,4	4,4	14,5	-5,8
Commercio al dettaglio di prodotti per la casa	128.605	348,1	17,5	112.986	410,6	26,7	330,6	-6,7	15.619	18,1	4,5	14,7	-4,3
Commercio al dettaglio di prodotti per il tempo libero	54.250	157,9	16,0	42.469	206,5	23,1	158,9	1,7	11.781	18,4	8,5	16,3	-1,5
Commercio al dettaglio di altri prodotti	71.584	363,4	16,1	60.836	425,4	25,3	421,8	-6,7	10.748	17,6	5,3	13,7	-3,8
Commercio al dettaglio ambulante	67.418	68,0	12,3	43.233	96,9	17,6	97,8	6,8	24.185	18,7	8,0	11,6	0,9
TOTALE	***3.482.862***	***229,2***	***30,1***	***2.797.513***	***287,7***	***43,2***	***251,1***	***-0,9***	***685.349***	***17,8***	***10,2***	***13,9***	***2,8***

Tabella 2.1: dati relativi a studi di settore, anno 2010

2.7 I risultati dell'azione di controllo dell'Agenzia delle Entrate

L'Agenzia delle Entrate comunica annualmente i risultati della propria attività di prevenzione e contrasto all'evasione. I dati più recenti disponibili sono quelli riferiti all'anno 2011 e sono raccolti nel "*book sul recupero dell'evasione*" disponibile *on line* sul sito internet della stessa Agenzia.

In particolare, sul fronte del recupero dell'evasione sono stati registrati, nel 2011, incassi per 12,7 miliardi (+15,5% rispetto al 2010).

Di questi 12,7 miliardi, i versamenti diretti ammontano a 8,2 miliardi (erano 6,6 nel 2010), mentre il totale riscosso mediante ruolo dagli agenti del gruppo Equitalia risulta pari a 4,5 miliardi (contro i 4,4 miliardi del 2010). Queste cifre comprendono le entrate erariali e

non erariali (imposte, sanzioni e interessi) derivanti dalla complessiva azione di contrasto degli inadempimenti tributari. L'attività di controllo dell'Agenzia si sostanzia sia nell'accertamento vero e proprio (dal quale sono stati incassati 7,2 miliardi di euro) sia nella liquidazione delle dichiarazioni (dalla quale sono giunti 5,5 miliardi).

Le statistiche, come riportato nel «*book sul recupero dell'evasione*» testimoniano un graduale e costante incremento degli incassi da attività di controllo (dai 2,9 miliardi nel 2007 ai 7,2 miliardi del 2011). E anche per quanto attiene ai versamenti diretti riferiti all'attività di controllo, la crescita a partire dal 2009 è stata importante: nel 2007 erano pari a 1,9 miliardi mentre nel 2011 hanno superato quota 5,5 miliardi, accrescendo peraltro l'incidenza sul totale del riscosso (dal 67% del 2007 al 77% del 2011).

Occorre tuttavia precisare che l'evasione fiscale, una volta accertata (attraverso lo svolgimento di un'attività propria dell'Agenzia delle Entrate), solo in un momento successivo darà i frutti sperati, ovvero i recuperi in termini monetari (attraverso un'attività spontanea del contribuente accertato oppure tramite l'attività dell'esattore). Tra la fase dell'accertamento e quella della riscossione, possono verificarsi diverse ipotesi che possono trasformare l'accertato in incassato con una misura percentuale variabile tra il 100 e addirittura lo 0%: adesione, contenzioso, autotutela, conciliazione, iscrizione a ruolo di quote inesigibili sono tutti possibili eventi successivi alla notifica di un avviso di accertamento che incidono grandemente sul recupero monetario derivante dallo stesso; inoltre, il versamento a seguito di un accertamento può avvenire anche in più soluzioni. Pertanto, l'ammontare incassato e/o riscosso nel 2011 (12,7 miliardi di euro) quale risultato di recupero dell'evasione fiscale non riguarda interamente l'attività svolta nel 2011, ma può essere in parte il frutto di un'attività risalente addirittura al 2001. Del resto, i frutti dell'attività svolta nel 2011 saranno visibili anche in anni successivi al 2011.

Inoltre, tra i tipi di controllo effettuati vi sono quelli espletati ai sensi degli articoli 36-bis e 36-ter DPR 600/73. Nello specifico, i controlli automatici (ex articoli 36-bis del DPR 600/1973 e 54-bis del DPR 633/1972) verificano la correttezza del modello o l'eventuale presenza di errori. I controlli formali (ex articolo 36-ter del DPR 600/1973) appurano la rispondenza tra i dati indicati e la documentazione conservata dal contribuente o le informazioni presenti nelle dichiarazioni trasmesse da altri soggetti (datori di lavoro, enti previdenziali ecc…).

Negli ultimi cinque anni si è verificata un'inversione di tendenza nell'incidenza percentuale degli incassi derivanti da queste attività ordinarie rispetto all'accertamento vero e proprio. Nel 2007 esse pesavano per il 55% (3,5 miliardi sui 6,4 totali riscossi), nel 2011 per il 43% (5,5 miliardi su 12,7 incassati). Ora, nei casi di 36 bis e 36 ter, non si può, a rigore, parlare di recupero di evasione fiscale vero e

proprio, in quanto il reddito o l'imponibile Iva sono dichiarati e non vengono rettificati, sebbene siano stati commessi errori materiali (e infatti, per questo motivo rientrano tra i controlli c.d. *formali*). In sintesi, l'evasione recuperata a seguito di accertamenti, intesa come maggiori imposte e sanzioni a fini IIDD, IRAP e IVA, nel 2011, è stata di 7,2 miliardi.

Dal quadro sopra descritto, discendono molte questioni di rilievo per la stessa Agenzia delle Entrate, che il presente lavoro, peraltro reso effettivamente possibile grazie alla collaborazione con il Servizio Studi della stessa Agenzia (coordinato dal Dott. Stefano Pisani) e con la SOGEI, suo *partner* tecnologico, tenta di affrontare.

Un primo problema che ci si deve porre è dato da come individuare un numero "adeguato" di casi su cui indirizzare gli accertamenti, così da ottimizzare l'utilizzo delle (relativamente scarse) risorse umane disponibili. E' chiaro come l'Agenzia delle Entrate già disponga di strumenti e tecniche con cui correntemente seleziona i soggetti da sottoporre a controllo. Tuttavia, considerato che la materia imponibile che ogni anno sfugge all'erario è molto vasta, dal problema generale citato, ne scaturiscono altri di ordine più strettamente tecnico, ai quali siamo qui interessati. In particolare, essi si incentrano sulla valutazione della possibilità che tecniche di *data mining* possano essere applicate o meno con successo in questo ambito.

Nei prossimi capitoli cercheremo di dare una risposta a questo particolare aspetto del problema.

Capitolo 3

Tax fraud detection: una prima analisi multidimensionale dei dati

L'entità del fenomeno dell'evasione fiscale in Italia, descritto nel capitolo precedente nei suoi tratti essenziali, pone questioni interessanti in capo all'Agenzia delle Entrate, che esigono soluzioni rapide ed efficaci. Quelle che qui ci interessano sono chiaramente legate all'impiego di tecniche di data mining, che però non possono prescindere dalla conoscenza del dominio specifico in cui si dovessero trovare ad essere applicate. Analizziamo quindi in questo capitolo un dataset relativo ad una platea di soggetti accertati nel periodo d'imposta 2007 (che sarà successivamente impiegato nelle analisi di mining). Tale dataset contiene informazioni relative a quasi 2000 tra imprenditori e professionisti, tutti persone fisiche residenti in Toscana e appartenenti alla classe dei soggetti di "piccole dimensioni" – ovvero con volume d'affari inferiore a € 5.164.000. In particolare, di ciascun contribuente presente nel dataset sono noti i dati delle dichiarazioni presentate (mod. Unico) per l'anno 2007 e quelli dell'accertamento subito per lo stesso anno (alcuni attributi sono stati tuttavia calcolati nel presente lavoro). In questo capitolo le analisi condotte saranno fondamentalmente OLAP ma vedremo che queste, pur essendo utilissime per la conoscenza del dominio, da sole, non sono in grado di fornire una caratterizzazione ("profilo") degli evasori esistenti, per cui è necessario un passo di analisi successivo, da effettuare mediante utilizzo di tecniche di data mining, che saranno oggetto dei prossimi capitoli.

3.1 I dati oggetto di analisi

La base dati a disposizione, fornita dal Servizio Studi dell'Agenzia delle Entrate in collaborazione con SOGEI, suo partner

tecnologico, contiene informazioni di carattere fiscale relative a 1843 contribuenti persone fisiche, imprenditori o professionisti, residenti in Toscana, riguardanti il periodo d'imposta 2007, che in quell'anno hanno subito un controllo (accertamento) fiscale. Tale *dataset* costituisce la base di partenza per lo svolgimento di analisi volte all'implementazione di strategie di selezione dei soggetti da sottoporre a controllo (limitatamente a persone fisiche, di piccole dimensioni) che siano guidate dalle informazioni ritraibili dalle basi dati in possesso della stessa Agenzia delle Entrate. Il *dataset* in questione viene in questa sede studiato ed analizzato per la prima volta.

In particolare, il *dataset* contiene un elenco di soggetti per i quali sono riportati:

- una serie di informazioni riguardanti le dichiarazioni IRPEF, IRAP e IVA presentate,
- i dati inerenti il relativo accertamento subito,
- vari attributi derivati da altre fonti informative (per lo più di carattere anagrafico).

Le informazioni sopra indicate sono state fornite in un'unica tabella, in cui ogni *record* si riferisce ad un contribuente di cui, per motivi di tutela dei dati, non sono noti né il codice fiscale né il codice di partita IVA.

Come detto, i dati sono relativi al periodo d'imposta 2007: esso costituisce il più recente anno per il quale sono ad oggi scaduti i termini per l'accertamento (salvo situazioni particolari che in questa sede non sono importanti)[1].

Le uniche caratteristiche che accomunano i soggetti presenti nel *dataset* sono quelle di essere titolari di ditte individuali (persone fisiche) e di essere residenti in Toscana. Sotto altri punti di vista, i dati sono alquanto eterogenei: per dimensioni (misurata ad esempio dal volume d'affari dichiarato), per il tipo di dati indicati in dichiarazione (es. indicazione di utile o perdita fiscale, quadri compilati, presenza o assenza di crediti IVA, ecc...), per entità dell'accertamento, le posizioni dei singoli contribuenti sono le più diverse tra loro.

Per ogni soggetto (*record*) abbiamo in totale 134 attributi e in Appendice A ne è riportato il significato.

[1] L'art. 43 DPR 600/73, rubricato "*Termine per l'accertamento*", stabilisce infatti, al comma 1, che "*Gli avvisi di accertamento devono essere notificati, a pena di decadenza, entro il 31 dicembre del quarto anno successivo a quello in cui è stata presentata la dichiarazione*". Possibili eccezioni a detta regola sono stabilite ai commi successivi: in particolare, in caso di omessa presentazione della dichiarazione, il termine di decadenza si allunga di un anno, mentre se il contribuente commette anche un reato penale di cui al D.Lgs. 74/2000, i termini raddoppiano.

3.2 Richiami di OLAP[2]

Il primo tipo di analisi che può essere condotto sul *dataset* a disposizione è quello noto come *on line analytical processing* (OLAP)[3]. Tale tipo di analisi si può caratterizzare, sulla base di [Pen04] come *veloce* (la "O" in OLAP richiede che tali sistemi, usati di solito in modo interattivo, debbano fornire rapidamente i risultati), *analitico* (la "A" in OLAP richiede che tali sistemi debbano fornire un ampio repertorio di funzioni analitiche riducendo al minimo la necessità di doverle definire con opportuni programmi), *condiviso*, in quanto di solito un sistema OLAP è una risorsa condivisa e quindi devono essere previsti opportuni meccanismi di controllo degli accessi ai dati, *multidimensionale*, in quanto deve fornire una visione multidimensionale dei dati con la possibilità di cambiare rapidamente le prospettive di analisi ed i livelli di dettagli, sfruttando la presenza di gerarchie, *informativo*, tenuto conto che i sistemi OLAP sono progettati per gestire grandi quantità di dati e per consentire analisi di varia natura per produrre informazioni utili e sintetiche rappresentate in modi diversi (tabellare o grafica).

Le analisi OLAP vengono solitamente condotte su particolari basi di dati, chiamate *data warehouse*, organizzate per facilitare le analisi di grandi quantità di dati al fine di produrre delle loro opportune sintesi di supporto ai processi decisionali delle organizzazioni. Un *data warehouse* possiede le seguenti principali caratteristiche, descritte in [Kim96] e [KR02]:

- *Tempificato:* i dati hanno un interesse storico e quindi contengono un'informazione sul tempo in cui si verificano certi eventi per consentire analisi di tendenza storicizzate;
- *Integrato:* i dati memorizzati nel *data warehouse* non provengono in genere da un'unica sorgente (una base di dati dell'organizzazione), ma sono il risultato di un lungo e costoso processo di integrazione di dati eterogenei;
- *Statico:* i dati vengono usati interattivamente per operazioni di ricerca e non di modifica. Periodicamente ai dati disponibili se ne aggiungono di nuovi o si rimuovono quelli ritenuti obsoleti;
- *Organizzato per soggetti (fatti):* nei sistemi operazionali i dati sono organizzati per eseguire le attività aziendali quotidiane, mentre nei sistemi direzionali come appunto i *data warehouse* i dati sono organizzati per analizzare dei soggetti di interesse che influenzano l'andamento complessivo dell'azienda. Quando i dati riguardano un solo soggetto di interesse, si parla di *data*

2 La trattazione che segue prende le mosse da [Alb10].

3 Il termine OLAP fu proposto da E. F. Codd, l'autore del modello relazionale dei dati, in [Codd93], in cui descrive il concetto usando 12 regole. Il termine fu proposto come variante dell'espressione *On Line Transaction Processing* per caratterizzare un nuovo approccio all'analisi dei dati di supporto alle decisioni che consentisse ai dirigenti di passare dall'uso dei tradizionali e numerosi rapporti statici stampati periodicamente su carta, a rapporti in formato elettronico modificabili interattivamente per ottenere rapidamente risposte a sempre nuove richieste di analisi dei dati.

mart e possono essere un sottoinsieme di un *data warehouse* più generale;

- *Di supporto alle decisioni:* i *data warehouse* sono progettati per valutare le prestazioni dei processi aziendali e per identificare possibili aree di intervento.

3.3 Prime considerazioni sul *datamart* di analisi

La base dati a disposizione costituisce un vero e proprio *datamart* in cui l'oggetto di interesse (*fatto*) è dato dalla "situazione fiscale" generale del singolo contribuente relativamente al periodo d'imposta 2007.

Lo schema concettuale di tale *datamart* è riportato in Figura 3.1. Il *fatto*, dato dalla situazione fiscale di ciascun soggetto, nel nostro caso, viene descritto da una serie di *dimensioni*, che ne precisano il contesto (ad esempio, *residenza, sesso, età, codice attività svolta, stato del controllo*) e da una serie di *misure* che lo rappresentano da un punto di vista quantitativo (esse sono costituite dai dati relativi ai singoli quadri delle dichiarazioni presentate e i dati relativi agli accertamenti):

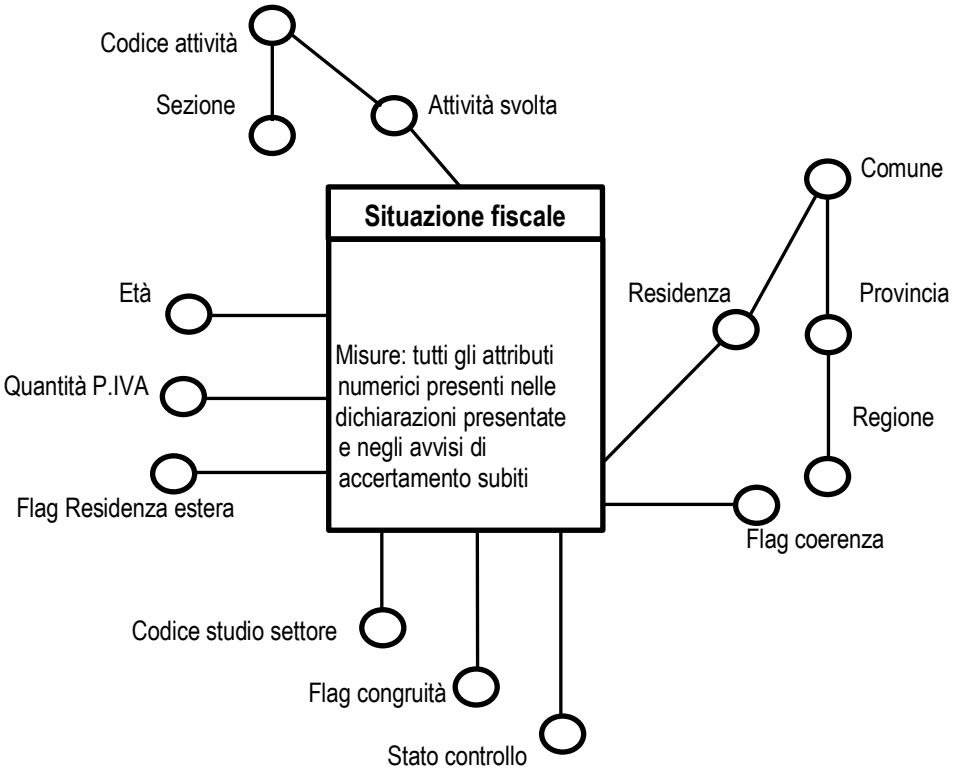

Figura 3.1: schema concettuale data mart "Situazione fiscale" di ogni contribuente

In generale, una *dimensione* è caratterizzata da un insieme di *attributi dimensionali*. In presenza di attributi dimensionali, un aspetto interessante da considerare, ai fini delle operazioni di analisi dei dati,

è dato da particolari relazioni gerarchiche (relazioni N:1) fra i loro valori, dette *gerarchie*. Per esempio, la dimensione "*residenza*" è caratterizzata da regione, provincia e città (identificata dal proprio codice catastale). Si osserva allora che i valori dell'attributo [CODICE_CATASTALE] sono in gerarchia con quelli di [SIGLA_PROVINCIA], ovvero, ([CODICE_CATASTALE] →[SIGLA_PROVINCIA]) nel senso che, ad un valore di una provincia, corrispondono più città e ad una città corrisponde una sola provincia. La stessa relazione vale tra [SIGLA_PROVINCIA] e [COD_REG]. In particolare, la gerarchia citata si dice che è *bilanciata* (*balanced*), in quanto i possibili livelli della stessa sono in numero predefinito ed i valori degli attributi che ne fanno parte sono sempre definiti. La presenza di gerarchie tra attributi dimensionali aumenta la possibilità di analisi da prospettive diverse (*analisi multidimensionali*).

Le tipiche operazioni OLAP su dati multidimensionali sono di tre tipi:

- restrizioni, per isolare particolari sottoinsiemi;
- aggregazioni, per calcolare il valore di una o più funzioni di aggregazione (classiche sono le COUNT, SUM, AVG, . . .) applicate a delle misure con i dati raggruppati secondo certe dimensioni;
- analisi multidimensionali con aggregazioni a diversi livelli di dettaglio, che sfruttano le gerarchie definite sugli attributi dimensionali.

Ciò premesso, con i dati a disposizione è possibile condurre una serie di analisi sotto molteplici profili, a seconda degli interessi e degli obbiettivi conoscitivi che si intendono soddisfare.

Si possono perciò individuare alcune linee di analisi di base che possono essere utili ai fini della comprensione e della valutazione della *consistenza* dei dati.

La base dati a disposizione, lo ricordiamo, consiste in una sola tabella riportante, per ogni *record*, gli attributi di natura fiscale relativi a un determinato soggetto. Per semplicità, quindi, si ritiene di impiegare direttamente tale tabella nelle analisi che seguono, con l'aggiunta di una sola altra tabella, denominata ATECO, nella quale è riportata la gerarchia presente nei codici attività, di cui si dirà in seguito (chiave primaria di tale tabella è il codice attività stesso).

In Appendice B vengono fornite, in SQL, le *query* di seguito citate.

3.4 Misure calcolate

Rispetto alla tabella originaria fornita dall'Agenzia delle Entrate, la base dati è stata arricchita di una serie di misure calcolate, per meglio precisare la situazione del singolo contribuente.

- *Importo valore aggiunto imponibile*
- *Importo valore aggiunto totale*
- *Maggiore imposta IRPEF accertata / definita*
- *Maggiore IVA cessioni accertata / definita*

- *Minori costi accertati / definiti*
- *Indebita detrazione IVA accertata / definita*
- *Maggiore IVA accertata / definita*
- *Maggiore IRAP accertata / definita*

La modalità di calcolo di dette misure viene riportata in Appendice A.

3.5 Analisi preliminari del *datamart*

Considerando in un primo momento solo i dati di natura anagrafica e prescindendo dai dati dichiarati e da quelli dell'attività di controllo, è già possibile caratterizzare la popolazione dei contribuenti presenti nel *datamart*, per esempio estraendone la distribuzione per *provincia*, come evidenziato in Figura 3.2.

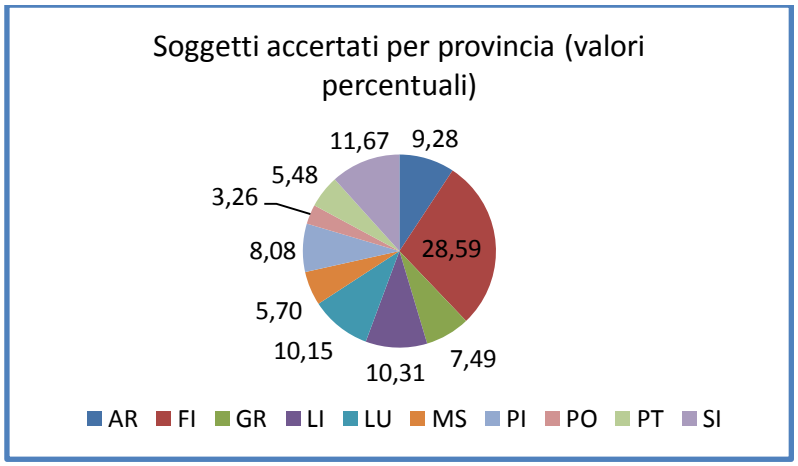

Figura 3.2: ripartizione su base provinciale dei contribuenti

Questo risultato si ottiene raggruppando i dati secondo l'attributo [SIGLA_PROVINCIA] inserito nella gerarchia [CODICE_CATASTALE]→[SIGLA_PROVINCIA]→[COD_REG] in base alla quale è strutturata la dimensione *residenza* e calcolando le funzioni COUNT(*) e COUNT(*)/(SELECT COUNT(*) FROM DATI07) su ciascun gruppo ottenuto. Detta operazione viene formalizzata nella [Query 1] riportata in Appendice B.

La distribuzione per provincia dei soggetti accertati presenti nel *dataset* ricalca abbastanza fedelmente quella della popolazione residente e pertanto i dati non sono "distorti", ovvero non sono sbilanciati a favore di una provincia piuttosto che un'altra. Si può infatti confrontare il grafico riportato in Figura 3.2 con la tabella di seguito riportata, contenente i dati della popolazione residente in Toscana al 31.12.2007[4].

4 Fonte: http://www.tuttitalia.it/toscana/statistiche/

Statistiche delle province della Toscana

Popolazione residente nelle province della Toscana degli ultimi anni (valori in migliaia).

Provincia	*2011 (1)*	*2010*	*2009*	*2008*	*2007*	*2006*	*2005*	*2004*
Arezzo	343	350	348	346	**342**	337	336	333
Firenze	972	998	992	985	**977**	970	967	965
Grosseto	220	228	227	226	**223**	221	219	218
Livorno	335	343	341	341	**339**	337	336	331
Lucca	388	394	392	390	**387**	383	380	379
Massa-Carrara	199	204	204	204	**202**	201	201	201
Pisa	411	418	414	410	**406**	400	397	394
Pistoia	288	293	292	291	**287**	281	279	277
Prato	245	250	248	246	**246**	245	242	239
Siena	267	273	271	269	**266**	263	262	261
Totale Regione	**3.668**	**3.750**	**3.730**	**3.708**	**3.677**	**3.638**	**3.620**	**3.598**

(1) popolazione al 31 dicembre 2011, calcolata a partire dalla popolazione censita il 9 ottobre 2011

Figura 3.3: ripartizione su base provinciale della popolazione toscana, anni 2004-2011

Partendo dai dati raggruppati per provincia, con un'operazione di *drill down* è possibile verificare in che modo la popolazione (per provincia) si suddivida secondo un altro attributo, ad esempio il [SESSO], e scoprire che la maggioranza dei titolari di partita IVA accertati in Toscana per il periodo d'imposta 2007 è costituita da persone di sesso maschile: [Query 2]).

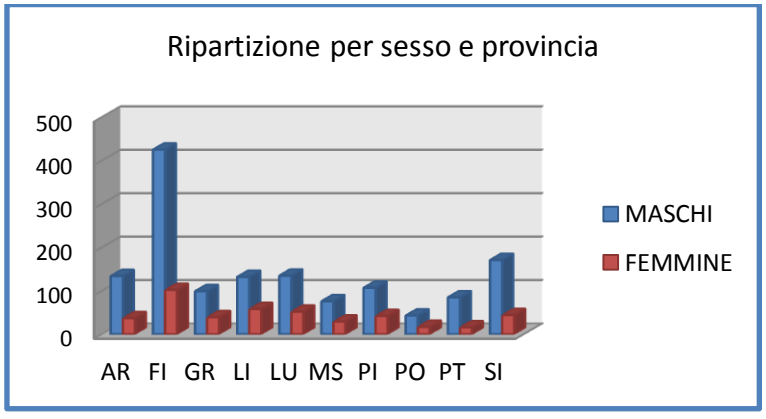

Figura 3.4: ripartizione dei soggetti accertati per sesso e provincia

Ancora, si può indagare sulla distribuzione per età di imprenditori e lavoratori autonomi [Query 3]: si scopre allora che solo il 20% ha un'età inferiore ai 40 anni (età minima 23 anni) e che il 75% ha un'età inferiore ai 57 anni, con media intorno ai 49 anni. I soggetti con più di 80 anni arrivano comunque ad essere oltre la decina.

Strettamente collegata all'età anagrafica del soggetto risulta essere il periodo di attività, ovviamente crescente con l'età stessa. La [Query 3bis] fornisce il numero di soggetti per età e durata attività. La rappresentazione Il risultato di tale *query* è stato poi rappresentato in Figura 3.5 b), apportando le seguenti modifiche (pe renderne più leggibile l'*output*): sulle ascisse, l'attributo età è stato discretizzato in classi di ampiezza 5 (ottenendo quindi un grafico per fasce di età: 23-27, 28-32 e così via) sulle ordinate è riportata la media di durata attività per ogni classe di età considerato. Si può osservare come la relazione sia "abbastanza" lineare, testimoniato anche dal fatto che il coefficiente di Pearson tra le due variabili risulta essere pari a 0,701.

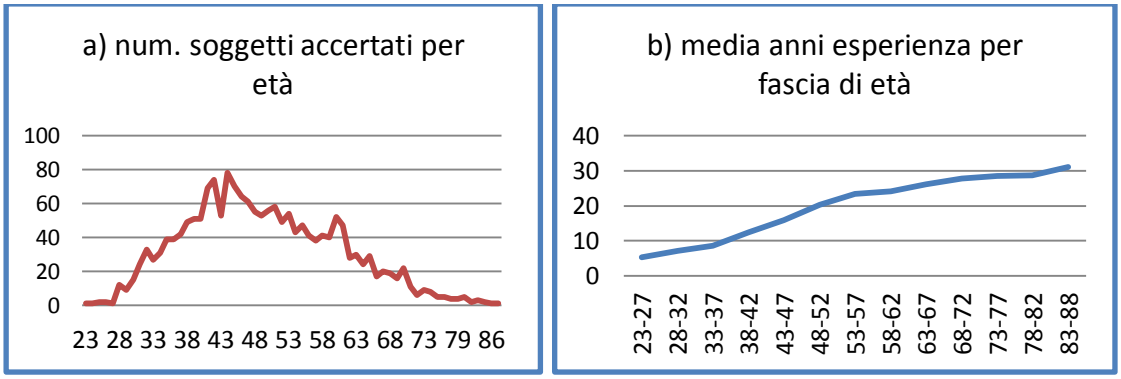

Figura 3.5: operatori economici per età e per esperienza

Volendo indagare sulle attività economiche svolte dagli imprenditori e professionisti che sono stati oggetto di controllo, si può vedere in che modo esse (meglio, *sezioni* di attività) si distribuiscono nel *dataset*[5]. La situazione che ne risulta è la seguente [Query 4]:

[5] Tecnicamente la classificazione delle attività economiche svolta dall'ISTAT (classificazione c.d. ATECO) si articola nei seguenti livelli, comprendenti, rispettivamente, le voci identificate da un codice:

1. alfabetico (sezioni);
2. numerico a due cifre (divisioni);
3. numerico a tre cifre (gruppi);
4. numerico a quattro cifre (classi);
5. numerico a cinque cifre (categorie);
6. numerico a sei cifre (sotto categorie).

La struttura di classificazione parte dal livello 1, più aggregato, distinto in 21 sezioni, fino a giungere al livello massimo di dettaglio, punto 6, comprendente 1.226 sotto categorie.

Tale gerarchia è stata riportata nella tabella ATECO, contenente un record per ogni codice attività (identificato dall'attributo "codice", e tanti attributi quanti gli elementi della gerarchia, considerato che la stessa è "bilanciata".

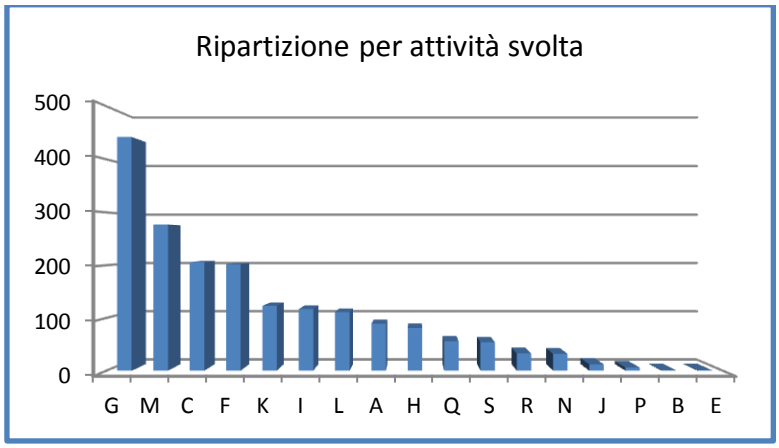

Figura 3.6: ripartizione soggetti accertati in base all'attività svolta

dove i codici indicati stanno a significare:

A AGRICOLTURA, SILVICOLTURA E PESCA
B ESTRAZIONE DI MINERALI DA CAVE E MINIERE
C ATTIVITÁ MANIFATTURIERE
D FORNITURA DI ENERGIA ELETTRICA, GAS, VAPORE E ARIA CONDIZIONATA
E FORNITURA DI ACQUA; RETI FOGNARIE, ATTIVITÀ DI GESTIONE DEI RIFIUTI E RISANAMENTO
F COSTRUZIONI
G COMMERCIO ALL'INGROSSO E AL DETTAGLIO; RIPARAZIONE DI AUTOVEICOLI E MOTOCICLI
H TRASPORTO E MAGAZZINAGGIO
I ATTIVITÀ DEI SERVIZI DI ALLOGGIO E DI RISTORAZIONE
J SERVIZI DI INFORMAZIONE E COMUNICAZIONE
K ATTIVITÀ FINANZIARIE E ASSICURATIVE
L ATTIVITA' IMMOBILIARI
M ATTIVITÀ PROFESSIONALI, SCIENTIFICHE E TECNICHE
N NOLEGGIO, AGENZIE DI VIAGGIO, SERVIZI DI SUPPORTO ALLE IMPRESE
O AMMINISTRAZIONE PUBBLICA E DIFESA; ASSICURAZIONE SOCIALE OBBLIGATORIA
P ISTRUZIONE
Q SANITA' E ASSISTENZA SOCIALE
R ATTIVITÀ ARTISTICHE, SPORTIVE, DI INTRATTENIMENTO E DIVERTIMENTO
S ALTRE ATTIVITÀ DI SERVIZI
T ATTIVITÀ DI FAMIGLIE E CONVIVENZE COME DATORI DI LAVORO PER PERSONALE DOMESTICO; PRODUZIONE DI BENI E SERVIZI INDIFFERENZIATI PER USO PROPRIO DA PARTE DI FAMIGLIE E CONVIVENZE
U ORGANIZZAZIONI ED ORGANISMI EXTRATERRITORIALI

Data la natura giuridica dei soggetti accertati, persone fisiche, le attività che compaiono con maggiore frequenza sono quelle che con più facilità possono essere svolte nella forma di ditta individuale (commercianti e professionisti) rispetto ad altre per le quali la forma societaria si addice maggiormente (ad es. trasporto, immagazzinaggio, attività finanziarie, attività immobiliari) o sono oggettivamente esercitate da un ristretto numero di operatori (fornitura di energia elettrica, gas, estrazione di minerali da cave).

3.6 Analisi delle dichiarazioni presentate

Dopo aver svolto le considerazioni di tipo descrittivo di cui sopra, al fine di caratterizzare la popolazione di imprenditori e lavoratori autonomi della Toscana, occorre entrare nel dettaglio dei dati da loro dichiarati.

Si tratta di soggetti che, ai fini delle imposte dei redditi, hanno compilato (quasi) sicuramente almeno uno tra i quadri RD, RE, RF e RG[6]; vale la pena osservare come, in genere, la compilazione di detti quadri sia esclusiva, a meno che il contribuente non gestisca contemporaneamente più attività che richiedono la compilazione di quadri distinti, ma questa tipologia di soggetti costituisce comunque una percentuale minima del *dataset*. Peraltro, si osserva come vi sia una porzione di soggetti, seppur minima, che non compila alcuno dei quattro quadri citati (tali soggetti rappresentano circa l'1% della popolazione).

Nella nostra base dati abbiamo la situazione riportata in Figura 3.7, ottenuta a partire dalla [Query5], che, variando di volta in volta le condizioni di WHERE, restituisce i soggetti che hanno presentato i vari quadri (o combinazioni di essi). Nello specifico, la [Query5] restituisce i codici univoci dei 73 soggetti che hanno presentato solo il quadro RD.

Quadri compilati				Count(*)
RD	**RE**	**RF**	**RG**	
🟢	🔴	🔴	🔴	**73**
🟢	🟢	🔴	🔴	3
🟢	🟢	🟢	🔴	0
🟢	🟢	🟢	🟢	0
🟢	🔴	🟢	🔴	1
🟢	🔴	🟢	🟢	0
🟢	🔴	🔴	🟢	15
🟢	🟢	🔴	🟢	2
🔴	🟢	🔴	🔴	**371**
🔴	🟢	🟢	🔴	2
🔴	🟢	🔴	🟢	10
🔴	🟢	🟢	🟢	0
🔴	🔴	🟢	🔴	**382**
🔴	🔴	🟢	🟢	0
🔴	🔴	🔴	🟢	**967**
🔴	🔴	🔴	🔴	17
				1843

Figura 3.7: ripartizione contribuenti in base ai quadri dichiarati (verde = presentato, rosso = non presentato)

Nel prosieguo, verranno presi in considerazione solo i soggetti che presentano un solo quadro, per evitare di dover trattare dati impuri.

[6] Per conoscere quale tipologia di contribuente è tenuta a compilare i vari quadri, si veda l'Appendice A.

E' possibile naturalmente indagare se vi siano delle differenze apprezzabili nei costi, ricavi (o compensi) e redditi medi dichiarati dai soggetti che svolgono attività professionale (quadro RE), dagli imprenditori (con dettaglio se in contabilità ordinaria o semplificata) e addetti al settore agricolo.

Considerando i dati *medi*, si ottengono le situazioni riportate in Figura 3.8 [Query6a, 6b, 6c, 6d]:

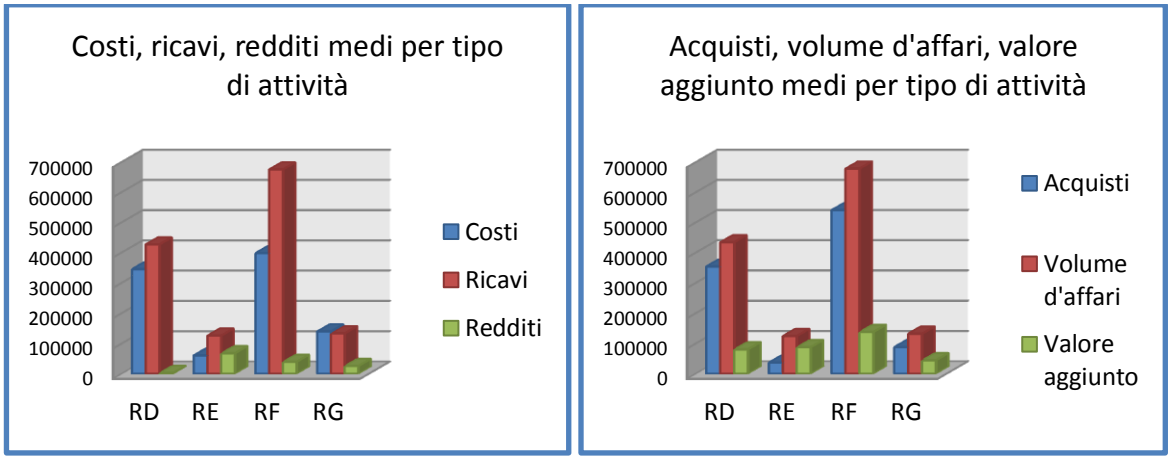

Figura 3.8: Costi, ricavi, redditi, valore aggiunto, volume d'affari, acquisti medi per tipo di attività (valori in euro)

Oltre ai dati reddituali, sono stati presi in considerazione anche quelli dichiarati ai fini IVA: volume d'affari (sostanzialmente il dato coincide con il valore delle vendite effettuate e, quindi, esprime generalmente un dato molto vicino a quelli che sono i ricavi[7]), acquisti e, per differenza, il valore aggiunto prodotto dall'impresa o dal professionista.

Sebbene questi grafici mettano bene in evidenza con pochi dati le tendenze di fondo delle varie tipologie di contribuenti, è comunque possibile condurre analisi statistiche con un maggior grado di dettaglio.

Tali analisi sono condotte importando la base dati in *Clementine* e sfruttandone gli strumenti messi a disposizione.

Prendiamo ad esempio in considerazione i dati indicati nel quadro RD, presentato dai 73 soggetti accertati che presentano *solo* tale quadro. Un'analisi statistica delle medesime grandezze di Figura

[7] Volume d'affari e ricavi possono anche non coincidere, in quanto il primo registra il volume di vendite effettuate secondo il *criterio della cassa*, mentre i ricavi, nell'ambito delle attività d'impresa, sono registrati secondo il *criterio della competenza*. In sintesi, il criterio di cassa consiste nel registrare le operazioni attive e passive quando queste effettivamente sono avvenute e quindi tenendo conto esclusivamente delle somme incassate e delle spese pagate. Viceversa, un costo è di competenza dell'esercizio se, nell'esercizio stesso, è maturato o ha dato la sua utilità o ha trovato copertura in un relativo ricavo. Allo stesso modo un ricavo può essere considerato di competenza dell'esercizio se è maturato nell'esercizio o se ha trovato in esso il suo correlativo costo Per eventuali approfondimenti, si rimanda a un qualsiasi testo di ragioneria o di diritto tributario.

3.8 condotta prendendo in considerazione altre metriche (ad esempio, mediana, *range*, deviazione standard), permette di osservare quanto riportato nella tabella che segue:

Grandezza	REDDITO	RICAVI	COSTI	VAL AGG	VOL AFF	ACQUISTI
Count	73	73	73	73	73	73
Mean	**2.436**	**426.363**	**344.508**	**80.222**	**433.719**	**353.497**
Sum	177.831	31.124.518	25.149.132	5.856.248	31.661.550	25.805.302
Min	0	0	0	-173.949	0	0
Max	56.492	15.529.552	13.276.800	2.274.929	15.626.270	13.351.341
Range	56.492	15.529.552	13.276.800	2.448.878	15.626.270	13.351.341
Standard Deviation	7.775	2.134.620	1.808.700	351.105	2.143.243	1.817.862
Median	**0**	**19.886**	**13.790**	**7.795**	**26.036**	**20.757**
Mode	0	0	0	-173949	0	0

Tabella 3.1: analisi statistica grandezze economiche quadro RD

Si nota immediatamente come vi sia una forte differenza tra il valore assunto dalla media e quello assunto dalla mediana di tutte le grandezze esaminate. Poiché il primo è, per sua natura, influenzato da elementi c.d. *outliers*, risulta interessante indagare quanti ve ne siano, considerato l'effetto distorsivo che essi producono sui dati: si possono allora individuare quei valori che distano più di un certo numero di deviazioni standard dalla media (valori generalmente utilizzati sono 3 o 5). Da questa analisi si individua un soggetto anomalo, in corrispondenza di [CODEC_CF] pari a 0000000000123836, che in effetti presenta ricavi per oltre € 15 milioni, pari a circa la metà dei ricavi dichiarati complessivamente dai 73 soggetti in argomento.

L'osservazione dei valori mediani rispetto a quelli medi consente di prescindere, in buona misura, da detto valore estremo e di valutare, quindi, in maniera più veritiera e corretta le grandezze economiche che caratterizzano i soggetti del settore agricolo.

Analoghe considerazioni relativamente ai soggetti esercenti le altre attività portano alle conclusioni di seguito indicate.

Cominciamo dai 371 soggetti che presentano il quadro RE, per i quali si hanno i dati di seguito riportati:

Grandezza	REDDITO	RICAVI	COSTI	VAL AGG	VOL AFF	ACQUISTI
Count	371	371	371	371	371	371
Mean	**67.542**	**125.625**	**60.352**	**87.433**	**124.097**	**36.665**
Sum	25.058.261	46.606.751	22.390.593	32.437.520	46.040.124	13.602.604
Min	0	0	0	-214.024	0	0
Max	2.351.531	3.460.011	1.304.972	2.757.156	3.548.858	997.304
Range	2.351.531	3.460.011	1.304.972	2.971.180	3.548.858	997.304
Standard Deviation	143.185	244.606	129.745	182.373	240.597	79.538
Median	**40.252**	**75.534**	**27.016**	**50.939**	**76.097**	**19.034**

Tabella 3.2: analisi statistica grandezze economiche quadro RE

Nel gruppo dei professionisti non emergono *outliers* da dover escludere dall'analisi.

Per quanto riguarda i 382 imprenditori in contabilità ordinaria, che presentano cioè il quadro RF, abbiamo la situazione di seguito riportata:

Grandezza	REDDITO	RICAVI	COSTI	VAL AGG	VOL AFF	ACQUISTI
Count	382	382	382	382	382	382
Mean	**39.381**	**674.774**	**397.649**	**137.476**	**677.789**	**540.313**
Sum	15.043.702	257.763.858	151.901.923	52.516.048	258.915.692	206.399.644
Min	-4.290.258	0.000	0.000	-598.495	0.000	0.000
Max	857.279	15.303.108.000	7.758.317	1.359.422	15.370.760	14.805.381
Range	5.147.537	15.303.108	7.758.317	1.957.917	15.370.760	14.805.381
Standard Deviation	238.115	1.294.499	820.241	205.739	1.279.472	1.205.060
Median	**29.835**	**296.260**	**126.103**	**96.362**	**315.253**	**201.281**

Tabella 3.3: analisi statistica grandezze economiche quadro RF

Sei soggetti presentano un volume di ricavi superiore a € 5.164.000 e costituiscono degli *outliers*: più precisamente quelli con [CODEC_CF] pari a 0000000001402088, 0000000000088113, 0000000000729101, 0000000000274970, 0000000001453383, 0000000000393022.

Infine, i 967 imprenditori in contabilità semplificata presentano i seguenti dati statistici, che non evidenziano la presenza di *outliers*.

Grandezza	REDDITO	RICAVI	COSTI	VAL AGG	VOL AFF	ACQUISTI
Count	967	967	967	967	967	967
Mean	**27.080**	**133.011**	**139.190**	**44.165**	**131.746**	**87.580**
Sum	26.186.819	128.622.478	134.597.033	42.708.235	127.398.734	84.690.499
Min	-511.681	0.000	0.000	-868.635	0.000	0.000
Max	572.712	4.696.669	6.743.446	2.204.718	4.696.669	2.491.951
Range	1.084.393	4.696.669	6.743.446	3.073.353	4.696.669	2.491.951
Standard Deviation	48.113	212.627	291.718	111.510	215.245	161.260
Median	**18.053**	**81.488**	**66.262**	**25.547**	**80.755**	**39.576**

Tabella 3.4: analisi statistica grandezze economiche quadro RG

Lo stesso grafico riportato in Figura 3.8, viene riportato di seguito, in Figura 3.9, con i valori mediani anziché quelli medi:

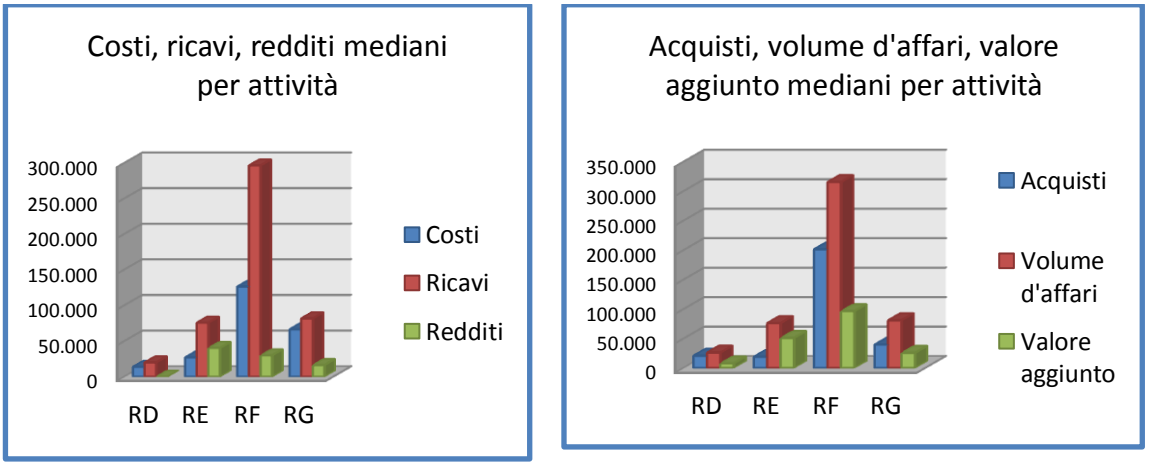

Figura 3.9: Costi, ricavi, redditi, valore aggiunto, volume d'affari, acquisti mediani per tipo di attività

Dal confronto tra le figure 3.8 e 3.9, emerge immediatamente che i valori mediani sono in generale più bassi di quelli medi: tale situazione è lampante per il settore agricolo, dove abbiamo visto essere presente un *outlier* che sconvolge le medie del settore e ben marcata anche nel settore delle imprese in contabilità ordinaria, per lo stesso motivo. Ciò che pare essere una costante è invece la proporzione in cui stanno tra loro costi, ricavi e reddito medi e mediani nelle varie tipologie di contribuenti.

E' possibile poi ottenere il dettaglio di costi, redditi e ricavi dichiarati di ogni singola tipologia di contribuente per singola provincia; ad esempio, nel caso del *reddito di impresa ordinario*, abbiamo la situazione riportata nelle figure 3.10 e 3.11.

Le statistiche prese in considerazione sono la media e la mediana di costi, ricavi e reddito dichiarati:

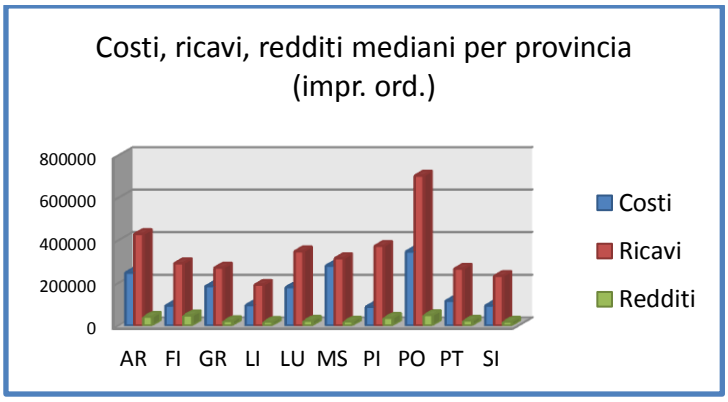

Figura 3.10: mediana grandezze economiche quadro RF, per provincia

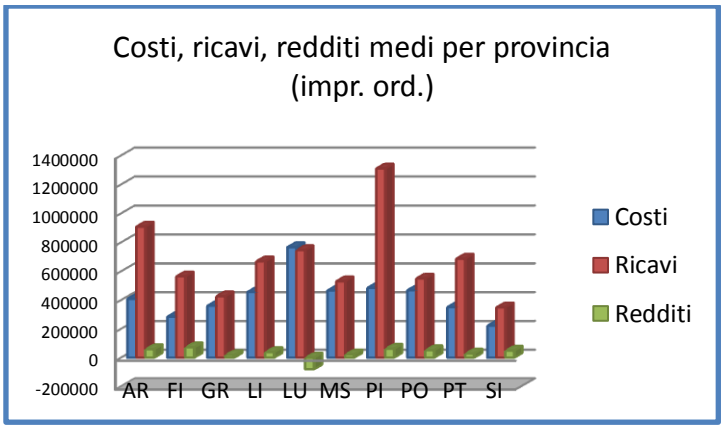

Figura 3.11: media grandezze economiche quadro RF, per provincia

Lo stesso tipo di analisi può essere condotta con riguardo all'attività svolta (sezione):

	Ricavi	Acquisti	Reddito	Volume affari	Acquisti IVA	Valore aggiunto	Count(*)
A	**230.569**	**183.293**	**1.563**	**240.737**	**193.021**	**47.716**	**89**
B	0	0	0	0	0	0	1
C	330.792	205.871	37.083	333.869	236.002	97.866	206
E	787.002	668.570	121.364	788.341	567.485	220.856	1
F	302.498	267.360	32.395	300.467	195.141	105.326	202
G	388.840	304.513	13.232	400.032	362.875	37.157	442
H	154.839	114.187	19.522	161.573	102.920	58.653	81
I	311.859	248.711	12.849	301.680	229.116	72.564	117
J	89.596	73.604	20.041	86.583	46.440	40.143	12
K	185.232	76.303	105.297	189.279	51.749	137.530	123
L	128.768	59.422	58.061	127.157	49.455	77.703	111
M	132.505	63.809	69.550	131.228	41.023	90.205	277
N	196.574	145.788	16.482	177.468	105.699	71.768	32
P	49.100	19.465	21.102	47.995	10.253	37.742	6
Q	120.572	63.667	57.676	113.106	43.710	69.395	56
R	227.830	199.908	28.293	103.803	73.603	30.200	33
S	75.297	62.854	5.029	75.369	44.685	30.684	54

Tabella 3.5: media grandezze economiche per famiglia di attività

	Ricavi	Acquisti	Reddito	Volume affari	Acquisti IVA	Valore aggiunto	Count()
A	19.886	13.351	0	24.844	21.564	6.734	89
B	0	0	0	0	0	0	1
C	153.233	95.384	21.246	152.764	81.607	42.316	206
E	787.002	668.570	121.364	788.341	567.485	220.856	1
F	136.486	105.302	23.125	144.971	74.082	41.944	202
G	115.214	94.825	14.691	115.929	84.367	16.941	442
H	49.425	25.922	17.592	49.847	25.619	25.434	81
I	111.526	93.009	12.965	109.446	69.397	31.215	117
J	35.128	27.621	18.427	41.113	15.433	25.680	12
K	158.485	53.015	84.469	166.546	34.069	115.785	123
L	106.500	42.166	58.046	101.603	23.587	71.768	111
M	62.920	24.040	34.307	66.921	20.319	42.490	277
N	55.047	47.081	14.957	55.047	38.998	30.932	32
P	58.745	5.454	15.383	53.274	6.828	43.024	6
Q	96.702	25.809	40.668	82.977	14.946	47.021	56
R	96.830	58.761	18.582	100.348	45.588	30.302	33
S	50.221	51.940	8.352	50.221	23.785	19.128	54

Tabella 3.6: : mediana grandezze economiche per famiglia di attività

Ancora è poi possibile indagare sulle correlazioni (lineari) esistenti tra i vari valori dichiarati[8].

Alcune di esse sono ovvie, come per esempio quelle esistenti tra importo della produzione netta, reddito e valore aggiunto: sebbene siano grandezze calcolate su basi imponibili diverse, ci si aspetta, nella normalità dei casi, che vi sia una relazione positiva più o meno lineare tra le tre. E di fatti, i dati mostrano questa forte correlazione:

- produzione netta [IMP_PROD_NETTA] – reddito [REDD_IMP_2007] = 0,677
- reddito [REDD_IMP_2007] – valore aggiunto [IMP_V_AGG_IVA] = 0,569
- valore aggiunto [IMP_V_AGG_IVA] – produzione netta [IMP_PROD_NETTA]=0,821

E' ragionevole supporre una forte correlazione anche tra altre grandezze, quali ad esempio ricavi [RICAVI_ATT_2007] e volume d'affari [IMP_VE_VOLAFF_2007], per le quali il coefficiente di Pearson vale 0,983 ed eventuali disallineamenti possono essere dovuti, ad esempio, al fatto che i ricavi, nell'attività di impresa, sono registrati secondo il principio della competenza, mentre l'IVA è un'imposta che sottostà al regime di cassa (vedi nota 6 per maggiori dettagli).

Ancora, una forte correlazione vale, in generale, per costi [TOT_PASS_2007] e ricavi [RICAVI_ATT_2007] , indipendentemente dal settore o tipo di attività svolta: più alti gli uni, più alti anche gli altri, per cui la correlazione è fortemente positiva: di fatti l'indice di correlazione è pari a 0,747.

Altri risultati, al contrario, appaiono meno intuitivi. Ad esempio, considerando gli esercenti attività agricola, si nota che corrispettivi [PR_AGR_CORR] e costi [PR_AGR_ACQ] sono, come ci si poteva aspettare, fortemente correlati tra loro (coefficiente di Pearson pari a 0.997), ma sorprendentemente entrambi sono scarsamente correlati col reddito (0,163 e 0,165 i rispettivi coefficienti). Ora, poiché in prima approssimazione $Reddito = Ricavi - Costi\ d'acquisto$, se la relazione tra *Ricavi* e *Costi d'acquisto* è lineare, ci si dovrebbe aspettare una relazione altrettanto lineare anche tra *Reddito* e *Costi* e *Reddito* e *Ricavi*. Ma tale relazione lineare non si riscontra nei dati. Questo comportamento anomalo si può spiegare in virtù del particolare regime di tassazione di detti soggetti, largamente forfettario, per cui, in definitiva, sganciato dalle dinamiche di costi e ricavi d'esercizio.

Altrettanto avviene nel caso di imprenditori (sia in contabilità ordinaria che in semplificata): il reddito ([IMP_REDD_IMP_ORD] o [IMP_REDD_IMP_SMPL]) risulta essere scarsamente correlato sia con i costi di acquisto ([SPESE_ORD] – coefficiente di Pearson pari a -0,184 o [IMP_TOT_CMPN_NEG] – coefficiente pari a 0,161) che con i ricavi ([IMP_RICAVI_ORDIN] – coefficiente di Pearson pari a 0,130 o [IMP_RICAVI_SMPL] – coefficiente pari a 0,375). Ciò suggerisce che la relazione $Reddito = Ricavi - Costi\ d'acquisto$ appare essere troppo semplificatrice, non tenendo conto di altri componenti, positivi e negativi, che giocano un ruolo importantissimo nella determinazione

[8] Si considera sempre la base dati con 1843 soggetti.

del reddito: ad esempio, la variazione delle rimanenze di magazzino, gli ammortamenti, gli oneri e proventi della gestione finanziaria.

Nel caso dei lavoratori autonomi, invece, la relazione lineare positiva esistente tra ricavi [IMP_CMPNS_ATTIV_2007] e costi [IMP_TOT_SPESE], misurata dal coefficiente di Pearson in 0,900, si riflette anche su quelle tra reddito [IMP-REDD_LAV_AUT] e ricavi (0,925) e, seppure in misura minore, tra reddito e costi (0,674): difatti, per i professionisti, la normativa non contempla le rimanenze di magazzino (essendo soggetti che sottostanno al regime per cassa e non per competenza come chi produce reddito di impresa) e in generale le voci di conto economico diverse da ricavi e spese correnti (di ogni tipo) sono di entità modesta.

Vi sono altre relazioni interessanti, peraltro di buon senso, che emergono dai dati. Ad esempio, la correlazione tra la [IMP_VAR_RIM_MP] e [IMP_REDD_IMP_ORD] è forte e pari a -0793,: difatti un aumento di rimanenze di materie prime può voler sottintendere un rallentamento del processo produttivo ed una conseguente riduzione degli utili. Ancora, una relazione ragionevole è quella esistente tra [DEB_FORN_ORD] e [IMP_ONERI_DIV] pari a 0,645: tenuto conto che le normali modalità di pagamento delle forniture prevedono tempi di pagamento dilazionati, più alti sono i costi sostenuti per acquisti di vario genere, maggiori sono le probabilità che il debito che ne deriva non sia stato ancora saldato al 31.12, con conseguente iscrizione in bilancio del debito verso il fornitore.

Un'altra interessante relazione è quella tra reddito imponibile [REDD_IMP] e la relativa imposta lorda [IMPST_LRD] pari a 0,998: nella sostanza si tratta di una relazione lineare perfetta. Ora, considerato che il sistema di tassazione IRPEF è progressivo per scaglioni d'imposta, se i redditi medi dichiarati fossero stati più alti, la relazione lineare sarebbe stata meno evidente: la relazione lineare indica infatti che i redditi dichiarati si concentrano nel primo scaglione (e in effetti, guardando alle tabelle riportate in precedenza, così è), mentre se la quota di redditi rientrante negli scaglioni successivi fosse stata maggiormente significativa, la relazione tra de due grandezze sarebbe stata, probabilmente, meglio descritta da una funzione quadratica.

Infine, a livello descrittivo, i dati sopra riportati consentono di "scoprire" interessanti caratteristiche di tipo "sociale" presenti nella base dati.

A titolo meramente esemplificativo, si può osservare come:

- mediamente le donne producano redditi inferiori rispetto agli uomini [Query 7]:

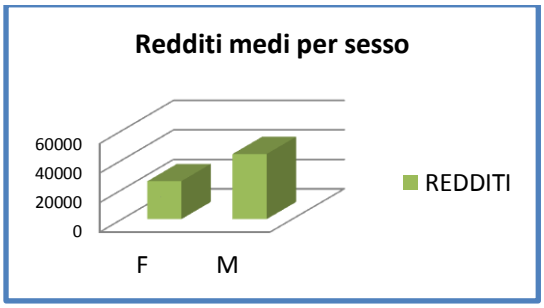

Figura 3.12: redditi in base al sesso

- il reddito prodotto vari in funzione dell'età [Query 8]: la linea di tendenza è mediamente crescente, anche se dopo una certa soglia i redditi appaiono leggermente decrescenti. Ciò può spiegarsi con l'abbandono progressivo dell'attività e del tempo ad essa dedicato con l'avanzare dell'età (peraltro ciò costituisce un fatto fisiologico di tutte le attività umane):

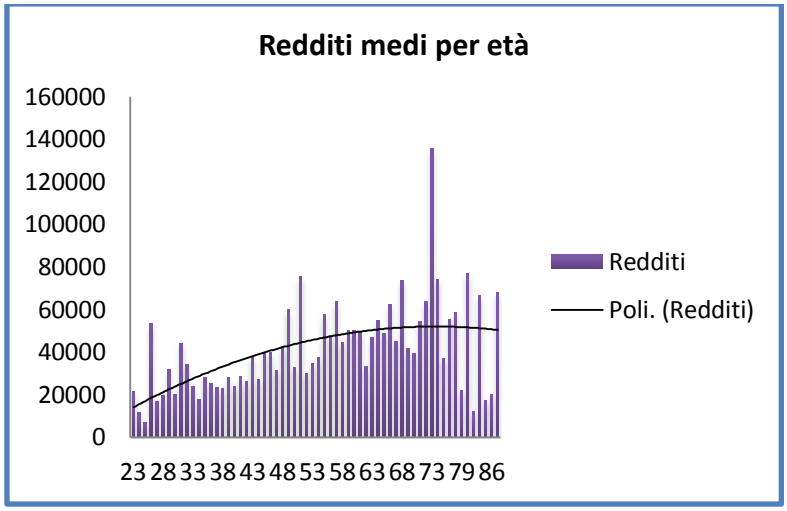

Figura 3.13: reddito ed età

- ancora, è possibile valutare il peso dell'esperienza, misurata dall'attributo [DURATA_ATTIV_SOGG], nella produzione di reddito [Query 9]. I dati sembrano premiare l'esperienza, fino a quando essa non si scontra, verosimilmente, con l'età troppo avanzata o comunque con il pensionamento, per cui i risultati tendono a peggiorare oltre una certa soglia di età:

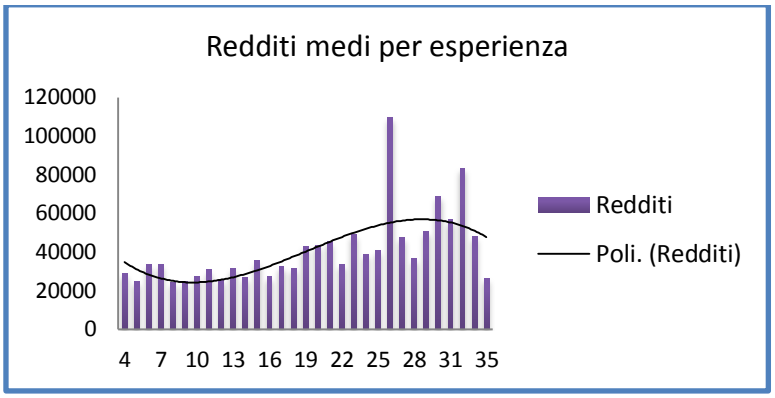

Figura 3.14: reddito ed esperienza

3.7 Analisi dei dati relativi all'attività di accertamento

Considerando, ora, gli attributi relativi agli accertamenti, si può dare risposta a moltissime esigenze conoscitive.

Abbiamo visto in che modo gli accertamenti si suddividano, per numerosità, tra imprese in semplificata, imprese in ordinaria e lavoratori autonomi [Query 5]. Ora, per ciascuna categoria, è possibile confrontare reddito medio dichiarato e maggiore imponibile IRPEF medio accertato, [Query 10]:

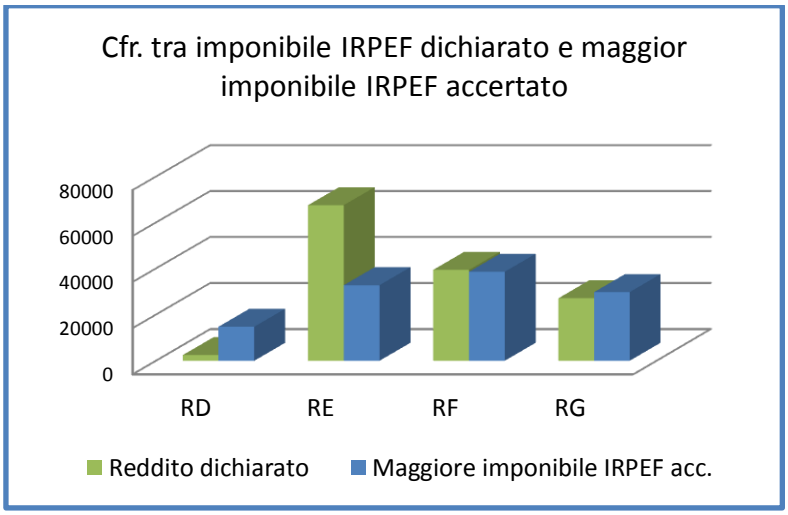

Figura 3.15: imponibili dichiarati e maggiori imponibili accertati

Le stesse informazioni possono essere, al solito, suddivise per provincia. Avendo in precedenza osservato che le diverse province della Toscana registrano livelli di reddito anche sensibilmente diversi tra loro, è interessante osservare se tali differenze siano correlate in qualche misura all'entità dei recuperi in accertamento. Nel proseguo, per non appesantire troppo l'analisi, non si tiene conto della categoria del soggetto, ovvero si prescinde dal quadro dichiarativo presentato [Query 11]:

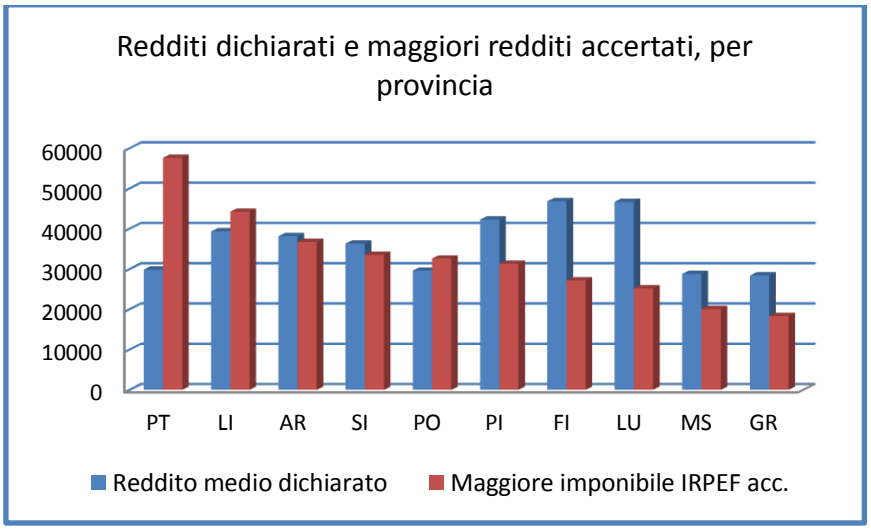

Figura 3.16: imponibili dichiarati e maggiori imponibili accertati, per provincia

Lo stesso ragionamento può essere fatto per il settore attività al fine di individuare in quale di essi viene accertato mediamente il più alto imponibile IRPEF e quale sia lo scostamento con i redditi mediamente dichiarati nel medesimo settore ([Query 12]). Per avere dati maggiormente significativi, si restringe l'analisi ai 10 settori, che hanno almeno 20 accertamenti, con il maggior imponibile IRPEF accertato:

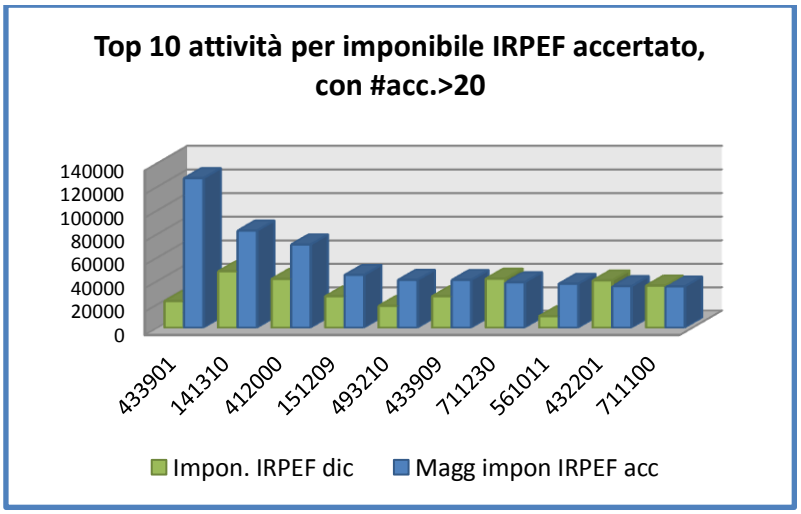

Figura 3.17: top 10 settori per imponibile IRPEF accertato, con # acc.>20

dove:

Sezione	Cod.Attività	Descrizione	# accertamenti
F	433901	Attività non specializzate di lavori edili (muratori)	43
C	141310	Confezione in serie di abbigliamento esterno	24
F	412000	Costruzione di edifici residenziali e non residenziali	49
C	151209	Fabbricazione di altri articoli da viaggio, borse e simili, pelletteria e selleria	48
H	493210	Trasporto con taxi	39
F	433909	Altri lavori di completamento e di finitura degli edifici nca	25
M	711230	Attività tecniche svolte da geometri	43
I	561011	Ristorazione con somministrazione	34
F	432201	Installazione di impianti idraulici, di riscaldamento e di condizionamento dell'aria	28
M	711100	Attività degli studi di architettura	34

Figura 3.18: Sezione, codice attività e descrizione dei top 10 per accertamenti

I dati evidenziano quali sono i settori in cui l'evasione si annida con maggiore forza: in particolare, risultano più a rischio le attività connesse all'edilizia e alcune figure professionali. L'evidenziazione di settori particolarmente a rischio può suggerire specifiche politiche di contrasto (nel caso dell'edilizia, ad esempio, sono previste speciali detrazioni d'imposta per chi ristruttura la propria abitazione sottostando a determinate condizioni, connesse alla tracciabilità dei pagamenti).

In generale, in un accertamento, possono venir contestati o un maggior ricavo non dichiarato, o un costo non inerente e, per questo, indebitamente dedotto, o entrambi. Il maggior ricavo è tipicamente una vendita a nero, mentre la contestazione del costo non inerente presenta una casistica più variegata: si va dalle fatture false al portare in contabilità (e quindi detrarre dai ricavi) costi che attengono alla sfera personale dell'imprenditore e non dell'attività, al caricare costi non inerenti l'attività per i motivi più disparati. La stima dei minori costi accertati è stata effettuata, per ogni record, in base all'espressione di seguito riportata, che esprime la normalità dei casi quando il soggetto accertato sia di piccole dimensioni. Abbiamo quindi la seguente regola per individuare i minori costi accertati:

if (MAG_IMPON_IRF_ACC=MAG_VOL_AFF_ACC) then 0

else if (MAG_IMPON_IRF_ACC>MAG_VOL_AFF_ACC and MAG_VOL_AFF_ACC>0)

then MAG_IMPON_IRF_ACC-MAG_VOL_AFF_ACC

else if (MAG_IMPON_IRF_ACC>MAG_VOL_AFF_ACC and MAG_VOL_AFF_ACC=0 and MAG_IMPON_IRAP_ACC=0)

then 0

else if (MAG_IMPON_IRF_ACC>MAG_VOL_AFF_ACC and MAG_VOL_AFF_ACC=0)

then MAG_IMPON_IRF_ACC;

else if (MAG_IMPON_IRF_ACC< MAG_VOL_AFF_ACC)

then 0

else 0

E' possibile, quindi, indagare se i due tipi di recuperi sopra descritti, maggiori ricavi o minori costi, si presentino in modo diverso a seconda della tipologia di soggetto accertato [Query 15]:

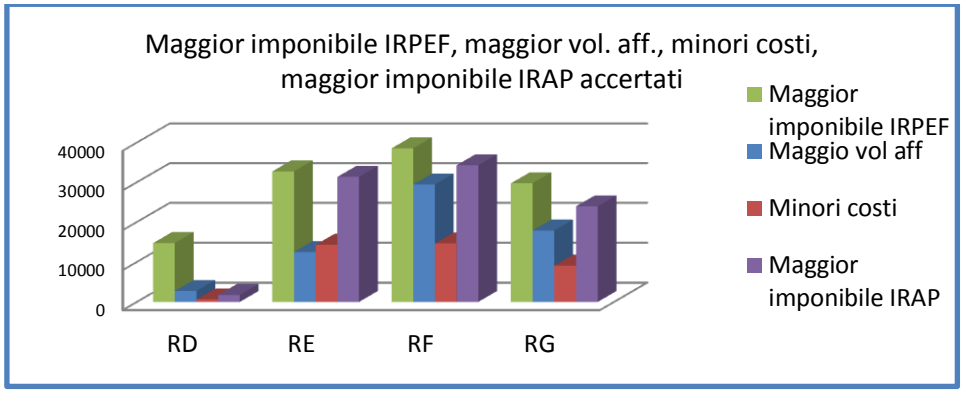

Figura 3.19: Maggior impon. IRPEF, maggior vol. aff., minori costi, maggior impon. IRAP accertati

In generale, si osserva come i recuperi derivanti dall'accertamento di minori costi siano di entità inferiore a quelli derivanti dall'accertamento di maggiori ricavi. Tale constatazione è ragionevole, dato il tipo di contribuenti inserito nel *dataset* (soggetti di piccole dimensioni), che presentano, normalmente, una struttura di costi ed una gestione aziendale semplici, per i quali la forma di evasione più immediata da adottare è sostanzialmente l'occultamento di materia imponibile derivante da vendite (senza l'emissione, cioè, di scontrino o ricevuta o fattura).

Nel caso di accertamento definito in sede di adesione (individuato dall'attributo [STATO CONTROLLO] con valore 19), è possibile osservare gli scostamenti tra accertato e definito. Si prendono in considerazione solo gli accertamenti per i quali si è poi addivenuti ad una definizione dell'atto con l'istituto dell'accertamento con adesione, previsto dal D.Lgs. 218/97 [Query 14]:

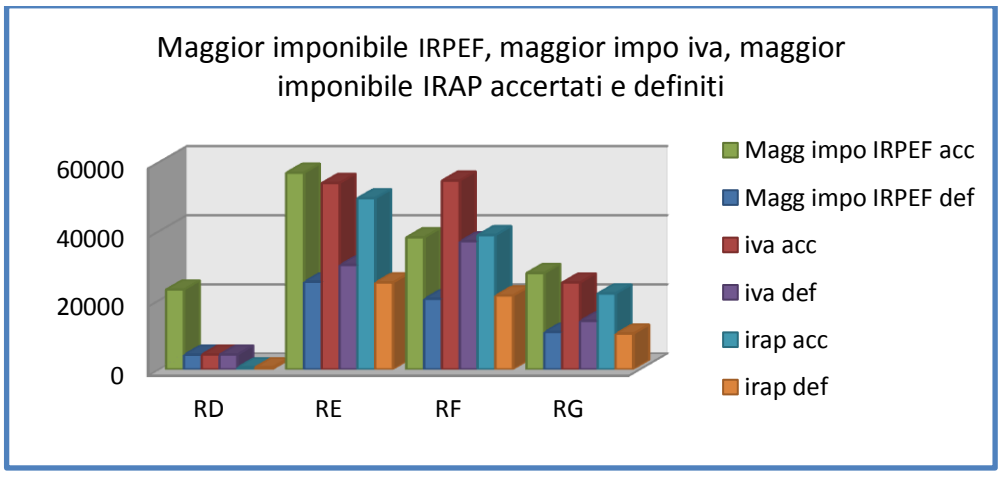

Figura 3.20: maggiori imponibili IRPEF, imposta IVA, imponibile IRAP, accertati e definiti

In generale, si può osservare come gli abbattimenti concessi in sede di adesione siano abbastanza consistenti. Questo fatto può essere spiegato da molti fattori, che andrebbero tuttavia verificati in maniera più approfondita avendo a disposizione ulteriori informazioni non presenti nel *dataset*: ad esempio, gli abbattimenti possono essere dovuti effettivamente ad accertamenti contenenti pretese erariali eccessive; possono essere sintomo di una "contrattazione" accesa col contribuente al fine di chiudere positivamente la procedura, ecc…

Oltre al tipo di analisi fin qui svolto, risulta interessante osservare se vi siano correlazioni significative tra le variabili dell'accertamento e qualche dato della corrispondente dichiarazione rettificata, in quanto tali correlazioni potrebbero fornire delle prime indicazioni per indurre, dalla sola lettura di certi dati dichiarati, un certo grado di confidenza sulla fraudolenza del soggetto esaminato.

Poiché a seconda della tipologia di soggetto accertato il relativo tipo di accertamento può essere diverso (si veda ad esempio la diversa ripartizione dei rilievi tra maggiori ricavi e minori costi tra professionisti, imprenditori e addetti del settore agricolo – figura 3.18), conviene ricercare eventuali correlazioni tra le variabili dell'accertamento e quelle delle dichiarazioni restringendo di volta in volta l'attenzione solo sui soggetti appartenenti alla medesima tipologia.

Peraltro, se prendessimo ad esempio i ricavi di impresa ordinaria e ne calcolassimo l'indice di correlazione con il maggior imponibile IRPEF accertato, avremmo un dato distorto dal fatto che la prima variabile assume valore zero per tutti i soggetti che non sono imprenditori in contabilità ordinaria, mentre la seconda assume i valori scaturiti dagli accertamenti a tutti i soggetti presenti nel *dataset*, indipendentemente dal tipo di attività svolta. E' chiaro che procedendo in tal modo otterremmo dei valori di correlazioni tra attributi non significativi.

Si procede quindi di seguito suddividendo la popolazione dei soggetti accertati in base al quadro (unico) presentato:

Soggetti che presentano solo il quadro RD.

Le correlazioni si basano su 72 soggetti (escluso l'*outlier* identificato in precedenza).

Maggior imponibile IRPEF: scarsamente correlato con i dati delle dichiarazioni

Maggior volume d'affari: mediamente correlato con molti attributi, tra cui: i dati del quadro VA, in particolar modo con l'importo dei beni strumentali non ammortizzabili, dei beni destinati alla rivendita e degli altri acquisti e importazioni; con le cessioni INTRA, le importazioni ed esportazioni e le cessioni non imponibili. Alcune delle relazioni evidenziate appaiono singolari, come quella che vede il maggior volume d'affari accertato correlato positivamente con le cessioni di beni intracomunitarie (coeff. Pearson 0,741): dato il tipo di attività svolto dai soggetti di cui si tratta, ci si aspetterebbe uno

scarso peso, sugli accertamenti, delle cessioni intracomunitarie di beni. L'alto valore esibito dai dati si spiega, al solito, tenendo conto che il coefficiente di Pearson "lavora" sulle medie delle distribuzioni X e Y su cui viene calcolato, essendo pari, come noto, a:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

dove $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ e $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ sono le medie delle due distribuzioni X e Y e pertanto soffre degli stessi difetti della media, ovvero può essere influenzato dagli *outliers*. Nel caso specifico, solo tre soggetti hanno dichiarato di aver effettuato delle cessioni intracomunitarie, ma uno di essi le ha effettuate per € 2.708.056,00. Tale valore ha completamente soverchiato gli altri, tanto da indurre a credere in una correlazione positiva importante tra [MAG_VOLAFF_ACC] e [CESS_BENI_INTRA]. Se si esclude tale valore anomalo dal computo del coefficiente di Pearson, esso scende a -0,047, ovvero pressoché nullo.

Minori costi IVA e Maggior imponibile IRAP: entrambi detti attributi sono fortemente correlati con il reddito e mediamente con la presenza di lavoratori dipendenti. Questo risultato è plausibile, in quanto è ragionevole supporre, in linea teorica, che maggiore la dimensione del soggetto, maggiore il rilievo. Poiché la dimensione del soggetto può essere misurata anche in base al reddito e al fatto di avere o meno dipendenti, si spiegano le correlazioni che emergono dai dati. Tuttavia, anche in questo caso, un *caveat* è d'obbligo: contando i soggetti cui sono stati contestati minori costi, si scopre che essi sono solo 2, mentre quelli con maggiori recuperi ai fini IRAP sono 8: pertanto, data lo scarso numero di soggetti coinvolti in detti tipi di accertamento, la significatività delle relazioni trovate ne rimane inficiata.

Soggetti che presentano solo il quadro RE:

Sorprendentemente, in questo caso, tutti i possibili tipi di rilievi presenti in un accertamento (*Maggior imponibile IRPEF, Maggior volume d'affari, Minori costi IVA e Maggior imponibile IRAP*) risultano essere scarsamente correlati con i dati delle dichiarazioni.

Soggetti che presentano solo il quadro RF:

Anche per questa classe di soggetti, tutti i possibili tipi di rilievi presenti in un accertamento (*Maggior imponibile IRPEF, Maggior volume d'affari, Minori costi IVA e Maggior imponibile IRAP*) risultano essere scarsamente correlati con i dati delle dichiarazioni.

Soggetti che presentano solo il quadro RG:

Maggior imponibile IRPEF: questo attributo è mediamente correlato con alcuni dati della dichiarazione ai fini IRPEF, tra cui le rimanenze, iniziali e finali ([RIM_INI_SMPL] e [RIM_FIN_SMPL]), i componenti positivi, [IMP_TOT_CMPN_POS], i c.d. costi residuali, [COSTI_RSDL] mentre con il reddito [IMP_REDD_PERD_2007], la correlazione è molto più debole.

Maggior volume d'affari: scarsamente correlato con i dati delle dichiarazioni.

Minori costi IVA: fortemente correlato con alcuni dati della dichiarazione ai fini IRPEF, tra cui i componenti positivi, [IMP_TOT_CMPN_POS], i c.d. costi residuali, [COSTI_RSDL] mentre con il reddito [IMP_REDD_PERD_2007], ancora una volta, la correlazione è molto più debole.

Maggior imponibile IRAP: anche questo attributo risulta mediamente correlato con alcuni dati della dichiarazione ai fini IRPEF, tra cui le rimanenze, iniziali e finali ([RIM_INI_SMPL] e [RIM_FIN_SMPL]), i componenti positivi, [IMP_TOT_CMPN_POS], i c.d. costi residuali, [COSTI_RSDL] mentre con il reddito [IMP_REDD_PERD_2007], la correlazione è molto più debole.

3.8 Osservazioni conclusive sull'analisi OLAP

Dall'analisi dei dati sin qui condotta, sono emerse molte informazioni inerenti i soggetti persone fisiche sottoposte ad accertamento per l'anno 2007. Sostanzialmente, ogni contribuente è stato osservato secondo molteplici dimensioni, avendo avuto, nel corso della trattazione precedente, riguardo alle loro caratteristiche anagrafiche, alla provincia di residenza, al tipo di attività svolta e al tipo di accertamento subito.

Per ogni dimensione di analisi, in definitiva, sono stati effettuati dei raggruppamenti e per ciascun gruppo sono state calcolate delle misure di interesse. Le *query* riportate in Appendice B hanno infatti tutte la medesima struttura di seguito riportata:

$SELECT\ [Attr_1, Attr_2, \dots, Attr_n], [fun_1, fun_2, \dots, fun_m]$
$FROM\ Basedati$
$WHERE\ [exp]$
$GROUP\ BY\ [Attr_1, Attr_2, \dots, Attr_n]$
$ORDER\ BY\ [Attr_1, Attr_2, \dots, Attr_n]$

Inoltre, per alcuni attributi, siamo andati alla ricerca di possibili correlazioni (lineari) esistenti.

Gli strumenti OLAP di analisi multidimensionale di dati sono molti utili per produrre rapporti di sintesi secondo criteri ben definiti.

Per esempio, disponendo dei dati sugli accertamenti relativi all'anno 2007, è semplice produrre un rapporto per stabilire "quanti sono i contribuenti accertati per provincia": la [Query1] ci fornisce la risposta a questo quesito. Ma se fossimo invece interessati a stabilire "qual è il profilo generale dei soggetti accertati in contabilità semplificata residenti nella provincia di Pisa" le tecniche di analisi multidimensionali sarebbero poco utili.

In un caso simile, infatti, si vuole estrarre dai dati un'informazione utile che però è nascosta nei dati stessi e non può

essere ricavata con interrogazioni e tecniche OLAP, ma sono richieste altre tecniche di analisi note con il nome di *data mining*.

Ecco allora, tra le tante che si possono ritrovare in letteratura, una definizione di *data mining* che bene ne indica la natura:

- *il data mining è l'esplorazione e l'analisi, attraverso mezzi automatici e semiautomatici, di grosse quantità di dati allo scopo di scoprire modelli e regole significative [BL97];*

Questa definizione caratterizza la natura del *data mining* e le differenze dalle analisi OLAP:

- il *data mining* è un'attività complessa, e non solo un insieme di tecniche, di solito semi-automatica, che richiede capacità di analisi sofisticate e competenze scientifiche specifiche;
- il *data mining* opera su dati raccolti in precedenza e indipendentemente dagli obiettivi dell'analisi che è di tipo esplorativa e non confermativa, tipica dei metodi statistici di verifica di ipotesi. I dati vengono esplorati senza un'idea chiara di ciò che si cerca;
- il *data mining* utilizza l'apprendimento automatico basato sull'induzione per generare definizioni di concetti generali a partire da esempi specifici;
- il *data mining* cerca di estrarre informazione dai dati rappresentata in una forma opportuna, detta modello o pattern. Gli esempi più comuni di modelli sono alberi, regole, reti ed equazioni. Un modello può essere descrittivo o predittivo. Nel primo caso il modello fornisce solo informazione sui dati usati per costruirlo, mentre nel secondo caso il modello può essere usato anche fare previsioni sul valore sconosciuto di un attributo di nuovi dati.
- il *data mining* scopre modelli di natura probabilistica che per essere utili devono essere attendibili, inattesi, rilevanti ai fini decisionali e traducibili in azioni concrete a fini pratici.

Una volta stabilito le analisi OLAP non sono sufficienti a delineare profili di evasori fiscali e che quindi il problema di *tax fraud detection* rientra tra quelli risolubili con il *data mining,* la prima decisione da prendere è il tipo di strumento da usare. Nei prossimi capitoli si presentano le caratteristiche degli strumenti che si prenderanno in considerazione, come saranno applicati, nonché gli algoritmi ed i modelli generati.

Capitolo 4

Tax fraud detection: approcci con il data mining

Considerata l'entità del fenomeno dell'evasione fiscale in Italia, descritto nel primo capitolo nei suoi tratti essenziali, nonché la consapevolezza che il fenomeno è in realtà presente in molti Paesi (tanto che la stessa Ocse dedica al problema parte delle proprie risorse), l'attività di rilevazione ed individuazione dei soggetti "evasori" (c.d. tax fraud detection) sta divenendo sempre più un'importantissima area di studio e ricerca in cui possono trovare applicazione diverse tecniche di data mining. In particolare, saremo interessati alle tecniche di classificazione e in questo capitolo, dopo averne richiamato brevemente i concetti fondamentali, esporremo i risultati più interessanti presenti in letteratura relativamente al tema della tax fraud detection.

4.1 Introduzione

Come osservato in precedenza, la lotta contro l'evasione fiscale in Italia risulta essere particolarmente complessa, essendo la stessa evasione divenuta, col tempo, un fenomeno di massa molto costoso da combattere, sia in termini economici che di risorse umane.

Di conseguenza, sempre più sentita è la necessità di sviluppare strumenti sempre più efficaci in grado per lo meno di contenere il fenomeno, sfruttando anche l'enorme mole di dati ormai presenti nei *database* dell'Agenzia delle Entrate.

Diviene quindi naturale cercare di applicare a questa mole di dati, in prima battuta, tecniche di analisi OLAP, come quelle viste nel capitolo precedente, ma anche tecniche di *data mining,* al fine di determinare il rischio di frode di uno specifico soggetto d'imposta e permettere all'amministrazione finanziaria di concentrarsi solo su quelli che presentano un rischio maggiore. In particolare, occorre costruire modelli, o profili, di comportamenti fraudolenti, che possano

servire da supporto per le decisioni nell'ambito della pianificazione di efficaci strategie di *audit*.

Per quanto riguarda il campo fiscale, che in questa sede interessa, un enorme gettito evaso potrebbe essere recuperato, in linea di principio, da controlli più efficaci, se si pensa che l'evasione fiscale vale, ogni anno, oltre 120 miliardi di euro, come visto nel primo capitolo.

Il fenomeno tuttavia non è solo italiano, ma riguarda molti paesi, tanto è vero che la stessa OCSE ha creato, a partire dal 2000, il *Global Forum* sulla trasparenza e lo scambio di informazioni a fini fiscali, entità multilaterale entro il quale operano i 120 stati membri nel campo della cooperazione a fini fiscali. Le questioni affrontate nel *forum* includono il riciclaggio di denaro sporco, l'evasione fiscale, i paradisi fiscali, gli accordi di scambio di informazioni fiscali e Convenzioni sulla doppia imposizione[1].

L'evasione fiscale è quindi un fenomeno presente, in maniera variabile, in molti (se non tutti) Paesi del mondo e questo spiega il crescente interesse e gli investimenti dei governi in sistemi intelligenti per la pianificazione degli *audit*. Negli USA, ad esempio, il Texas è stato uno dei primi stati ad applicare sul campo tecniche di *data mining* per individuare dichiarazioni dei redditi sospette e recuperare gettito [Hoo09].

Nel proseguo, quindi, dopo aver richiamato brevemente alcuni degli aspetti teorici delle tecniche di *classificazione*, vedremo in che modo le stesse siano state oggetto di lavori scientifici con specifica applicazione al tema della lotta all'evasione fiscale (non solo in Italia) e con quali risultati pratici.

4.2 Classificatori: generalità

Uno dei più importanti *task* nel campo del *data mining* e, in particolare, dell'area *machine learning*, è dato dalla classificazione. L'obbiettivo generale di una procedura di classificazione è, come dice il nome, suddividere i dati in classi (altri nomi per *classe* sono *categoria* o *attributo target*). Un pò più precisamente, si vuole, in presenza di oggetti (*istanze, esempi, record*) di cui si conoscono determinate caratteristiche (*variabili, attributi*), costruire dei modelli, secondo determinati criteri, per assegnare ciascun oggetto a una classe predefinita, sulla base delle caratteristiche osservate.

E' bene precisare subito che in alcuni casi, gli individui sono suddivisibili in gruppi noti a priori, mentre in altri l'esistenza di una 'sensata' partizione degli individui in gruppi non è scontata, ma anzi è uno degli scopi dell'indagine decidere se tali gruppi esistano e quali e quanti siano. Nel primo caso il problema viene anche detto di *classificazione supervisionata*, nel secondo prende il nome di *classificazione non supervisionata* o raggruppamento (*clustering*). Nel

[1]Per maggiori dettagli si consulti il sito http://www.oecd.org/tax/transparency/

proseguo si farà riferimento solo al caso di *classificazione supervisionata*, rimandando, per l'approfondimento delle tecniche di *classificazione non supervisionata* a testi di *data mining* quali [TSK06].

Nel caso dell'apprendimento supervisionato disponiamo, quindi, oltre che delle variabili predittive, anche della classe di appartenenza di ciascuno degli individui. Si suppone cioè che la classe di appartenenza di un individuo sia osservabile direttamente. Formalmente l'informazione a disposizione potrà essere rappresentata in forma matriciale:

$$(y; X) = \begin{bmatrix} y_1 & x_{11} & \dots & x_{1p} \\ \vdots & & \ddots & \vdots \\ y_n & x_{n1} & \dots & x_{np} \end{bmatrix} = (\mathbf{y}, \quad \mathbf{x_1}, \quad \dots, \quad \mathbf{x_p})$$

Ivi, y_i rappresenta la classe di appartenenza dell'individuo i-mo e assume un valore nell'insieme $\Gamma = (G_1, \dots, G_g)$, essendo g il numero totale di classi. Inoltre, x_{ij} rappresenta il valore assunto dalla variabile j-ma per il soggetto i-mo.

Date le informazioni sulle n istanze di cui si conoscono sia gli attributi predittivi, sia la classe di appartenenza, tutti i modelli di classificazione prendono in *input* un c.d. *training set* di tuple di cui è nota la classe e restituiscono in *output* un modello, il classificatore appunto, che è in grado di assegnare una classe a ciascuna tupla che gli viene data da "riconoscere", basandosi su tutti i suoi attributi, eccetto, ovviamente, la classe stessa. Il modello può quindi essere utilizzato per predire la classe di una nuova tupla, per la quale il valore di classe non è noto a priori. Si tratta quindi di una tecnica di apprendimento induttivo, il cui obiettivo è "imparare" la definizione di una funzione di classificazione.

In generale, allora, tutte le osservazioni (tuple, esempi) sono partizionate in un *training set* ed un *test set*, che insieme compongono, ovviamente, il *dataset*.

L'intero processo di classificazione viene eseguito in due passaggi:

i. *training* – costruzione del modello sulla base dei *training data*: il modello si "allena" e impara, osservando le caratteristiche di ogni tupla, delle quali conosce anche la classe di appartenenza.
ii. *test* – controllo e verifica dell'accuratezza del modello utilizzando il *test set*: in questa seconda fase, ciò che il modello ha appreso nella prima viene validato, su dati che il modello stesso non aveva considerato nella precedente fase. Tuttavia, poiché lo sperimentatore conosce il valore della classe anche delle tuple del *test set*, è possibile confrontare i valori predetti dal modello con quelli reali e quindi è possibile valutare la bontà dello stesso, data, in definitiva, dalla sua capacità di predire correttamente la classe di un soggetto (anche se le possibili metriche di valutazione di un modello utilizzabili sono molte e alquanto diverse tra loro).

Il risultato di un'analisi di classificazione si sostanzia, quindi, nella individuazione di una legge (modello) in grado di associare ad ogni istanza presa in *input*, una classe. E' poi possibile complicare lo scenario, ad esempio decidendo di fornire non solo l'informazione sull'appartenenza o meno ad una classe ma anche il grado (probabilità) di appartenenza alla stessa.

E' spesso difficile o quasi impossibile costruire un modello di classificazione perfetto, capace di classificare correttamente tutti i record del *test set* (mentre sarebbe anche troppo facile – ma alquanto inutile – costruire un classificatore capace di classificare correttamente tutti i record del *training set*), pertanto spesso non si può che scegliere un modello di classificazione subottimale, che sia il più adatto alle esigenze dello sperimentatore e funzioni meglio di altri nell'ambito del dominio specifico in cui si opera.

Quindi, dovendo confrontare modelli ottenuti attraverso l'impiego di diverse tecniche che utilizzano concetti e paradigmi diversi tra loro, un aspetto non secondario del problema è dato dalla valutazione della bontà di un certo modello di classificazione che si intende utilizzare, tenuto conto che sovente la scelta dipenderà anche dallo specifico dominio in cui un modello si troverà ad "operare".

Ad esempio, limitandoci ai soli alberi decisionali, la loro costruzione non è univoca, ma dipende da come vengono affrontate le seguenti fasi:

- la scelta, per ciascun nodo, di come effettuare lo *split*, ovvero (a) scelta della variabile da considerare; (b) scelta del criterio di suddivisione. L'intenzione, ad ogni nodo, è di suddividere in due (o più) gruppi le osservazioni in base a una regola che coinvolga una delle variabili. Il criterio di massima è di scegliere la variabile che faccia in modo che i nodi figli realizzino la maggiore separazione possibile tra le classi secondo un qualche indice di (im)purezza – ad esempio indice di Gini o entropia;
- la decisione se un nodo debba essere considerato terminale (foglia). La questione rientra tra i problemi di bilanciamento tra complessità e precisione di un modello: al massimo della complessità otteniamo un albero che ha per ciascun nodo terminale una sola osservazione e che quindi comporta zero errori di classificazione riferendosi al *training set*, ma, presumibilmente, un elevato errore di classificazione su una generica unità non compresa in esso; pertanto, normalmente, un modello non si presenta mai con solo foglie pure (specifiche tecniche di *pruning* – *pre* o *post* – evitano che ciò possa accadere);
- l'assegnazione di una classe a ciascuna foglia (nodo terminale): in generale il criterio seguito è quello della classe modale, ovvero si assegna a ciascuna foglia l'etichetta della maggioranza delle osservazioni in essa contenute (sempreché il

campione di apprendimento sia un campione casuale semplice e gli errori di classificazione siano tutti egualmente gravi).

Ciascuno degli aspetti citati è ben noto in letteratura e ad essa si rimanda per approfondimenti.

Inoltre, nel caso dei *dataset* in cui si "nascondono" frodi, non infrequente è il problema che la classe "frodatore" sia sbilanciata, ovvero che nel *dataset* vi siano "pochi" *record* riguardanti una frode rispetto al totale; ancora, sempre più spesso, le analisi non vengono svolte sulla base di classificatori "semplici", ma "composti" (c.d. *ensemble methods*), in quanto questi ultimi, per i motivi che vedremo, danno spesso risultati migliori rispetto ai primi. L'utilizzo di c.d. *ensemble methods* pone un ulteriore problema da affrontare, ovvero quello di trovare una sintesi opportuna dei risultati prodotti dai c.d. classificatori di base.

Nel proseguo ci occuperemo quindi di richiamare i seguenti tre aspetti generali, che saranno trattati nelle analisi svolte nel capitolo che segue, decisivi per ogni problema di classificazione:

- metriche di valutazione del classificatore (matrice di confusione, curva ROC e *lift chart*)
- problema delle classi sbilanciate e relative tecniche utilizzate per risolverlo (*cost sensitive learning* e *resampling*).
- aggregazione di modelli per ottimizzare la qualità della classificazione (*ensemble methods* – con particolare riferimento a due delle tecniche più utilizzate in tale ambito*, bagging* e *boosting).*

4.2.1 Metriche di valutazione

Supponiamo, per semplicità, di trattare problemi di classificazione a due classi. In questi casi, possiamo sempre etichettare una classe come positiva e l'altra come negativa. Il *test set* consiste di P esempi positivi e N esempi negativi (le misure che seguono hanno senso solo una volta deciso quale delle due possibili classi è considerata positiva e quale negativa). Un classificatore assegna una classe a ciascuno di loro, ma alcuni degli assegnamenti saranno, inevitabilmente, errati. Per valutare i risultati della classificazione, contiamo il numero dei *true positive* (TP), *true negative* (TN), *false positive* (FP) - effettivamente negativi, ma classificati come *positive)* e *false negative* (FN) – effettivamente positivi, ma classificati come *negative*.

Si può definire allora una matrice di confusione nel seguente modo:

$$Classe\ prevista$$

$$Classe\ reale \begin{bmatrix} & C_1 & C_2 & \\ C_1 & TP(t) & FN(t) & P \\ C_2 & FP(t) & TN(t) & N \end{bmatrix}$$

Valgono le seguenti identità:

$$TP + FN = P$$

e

$$TN + FP = N$$

Il classificatore assegna $TP + FP$ esempi alla classe positiva e $TN + FN$ esempi a quella negativa.

Definiamo ora alcune note e molto usate metriche di valutazione:

- $FP\ rate = FP/N$
- $TP\ rate = TP/P = recall = sensitivity$
- $TN\ rate = TN/(TN + FP) = specificity$
- $Y\ rate = (TP + FP)/(P + N)$
- $Precision = TP/(TP + FP)$
- $Accuracy = (TP + TN)/(P + N)$
- $F - measure = 2pr/(r + p)$

Precision e *accuracy* sono spesso utilizzati per misurare la qualità della predizione di un classificatore binario.

La *F-measure* è una metrica che riassume *precision* e *recall*, rappresentandone la media armonica. Poiché la media armonica tra due numeri x e y tende ad essere vicina al più piccolo dei due numeri, se è elevata significa che sia *precision*, sia *recall* sono tali e quindi non si sono prodotti né (tanti) falsi negativi né (tanti) falsi positivi.

Un classificatore può però essere definito in modo più fine, se gli si consente di assegnare una classe a ciascun esempio, con un certo livello di confidenza (probabilità). Dato $X = test\ set$ e $\mathbb{C}$ l'insieme dei valori di classe, un classificatore probabilistico può essere definito da una funzione $f: X \rightarrow \{[0,1], \mathbb{C}\}$ che mappa ogni esempio $x \in X$ in una coppia [classe, $c \in \mathbb{C}$ e probabilità che x appartenga a quella classe, $f(x|c)$].

Sempre nell'ambito della classificazione binaria, normalmente, viene selezionata una soglia t per cui gli esempi x con $f(x) \geq t$ sono considerati positivi e gli altri negativi (normalmente $t = 0{,}5$). Questo implica che ogni coppia di classificatore probabilistico e soglia t definisce un classificatore binario. Le misure definite in precedenza possono essere utilizzate anche per classificatori probabilistici, ma diventano funzioni parametriche in t. Si noti che i $TP(t)$ e $FP(t)$ sono sempre funzioni decrescenti monotone di t. Al variare di t, otteniamo una famiglia di classificatori binari.

Ciò premesso, introduciamo due noti strumenti ampiamente utilizzati per la valutazione di un classificatore, la *curva ROC* e il *lift chart*. In particolare, quest'ultimo, opportunamente adattato, sarà utilizzato in questo lavoro per la valutazione dei modelli.

4.2.1.1 Curva ROC

Le curve ROC – *receiver operating characteristic* – [Faw04] rappresentano una descrizione grafica dell'accuratezza di una procedura di classificazione. Esse nascono per valutare l'*output* di una classificazione binaria. Per estendere l'uso delle curve ROC al caso di più di due classi sono state fatte diverse proposte. La più ovvia consiste nell'uso della classe di riferimento: si disegnano tante curve ROC quante sono le classi, considerando una classe contro l'unione delle altre.

Ragioniamo dunque nell'ambito di una classificazione binaria, in cui la curva ROC può essere rappresentata su un sistema di assi cartesiani, X,Y. Il grafico della curva ROC è definito in modo parametrico:

$$x = FPrate(t),\ \ y = TPrate(t)$$

Ciascun classificatore binario, dato un *test set* di esempi, è rappresentato da un particolare punto nello spazio $(FPrate,\ TPrate)$. Variando la soglia del classificatore probabilistico, t, otteniamo un insieme di classificatori binari, rappresentati da un insieme di punti che insieme disegnano la curva.

In sostanza, questa curva ci dice, una volta fissato il $FPrate$, quanto ci possiamo attendere di $TPrate$. Ma fissare il $FPrate$ significa, in ultima analisi, fissare la soglia t, ovvero il livello di confidenza che si vuole avere nel classificare un certo esempio come positivo: ad esempio, per $t = 0$, ogni esempio viene classificato come P: quindi $TPrate = 1$, ma anche $FPrate = 1$: in sostanza, questa estrema permissività nell'attribuire la classe P, rende il classificatore completamente inutile. All'opposto, se $t \to 1$, $FPrate = 0$, ma anche $TPrate = 0$ (ovvero, si predice sempre N). Il classificatore ottimale è quello individuato dal punto (0,1), che classifica sempre correttamente le istanze, per cui si hanno solo TP e TN.

L'idea quindi è che più vicini si è, da subito, al vertice alto sinistro dello spazio di coordinate $(FPrate,\ TPrate)$, meglio è. In particolare, la misura che si suole calcolare è data dalla c.d. *area sotto la curva*. Un classificatore casuale ha un'area sotto la curva di 0,5, mentre un classificatore perfetto ha un'area di 1. I classificatori usati nella pratica dovrebbero quindi avere un'area compresa tra 0,5 e 1, preferibilmente vicino a 1.

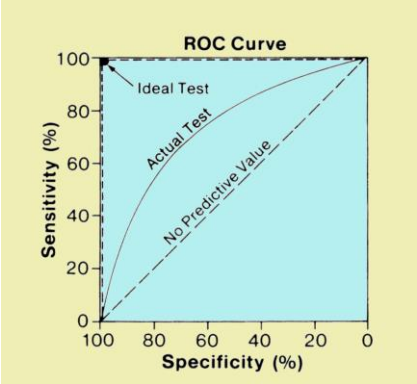

Figura 4.1: esempio di curva ROC

4.2.1.2 Lift chart

Il *lift chart* viene di solito introdotto con un esempio in cui, tipicamente, un'agenzia di *marketing* deve pianificare l'invio di promozioni ad un selezionato numero di famiglie, con l'obiettivo di incrementare le vendite di un certo prodotto. L'agenzia dispone di una serie di informazioni su questi potenziali clienti e di un classificatore in grado di effettuare predizioni circa le loro intenzioni di acquisto. Ogni promozione inviata ha un costo *c*, ma porterà un guadagno *g* in caso di acquisto. Quindi, l'agenzia vuol massimizzare il profitto atteso, spedendo la promozione solo a quei soggetti che con probabilità maggiore effettueranno l'acquisto.

Per verificare l'efficienza del classificatore usato e la relazione tra costo e guadagno atteso della postalizzazione, può essere utilizzato un *lift chart*.

Anche il grafico del *lift chart*, rappresentato su un sistema di assi cartesiani, X,Y, viene definito in modo parametrico:

$$x = Yrate(t) = \frac{TP(t) + FP(t)}{P + N}, \quad y = TP(t)$$

Così come per la curva ROC, ogni classificatore binario corrisponde a un punto del grafico in questo spazio parametrico. Variando la soglia del classificatore probabilistico otteniamo un insieme di punti (i.e. insieme di classificatori binari) che, nello spazio individuato, descrivono una funzione non decrescente.

Anche in questo caso si può definire l'area sotto la curva, che, facendo riferimento all'esempio dell'agenzia di *marketing* introdotto in precedenza, va interpretata nel seguente modo: fissata una soglia *t*, l'area sotto la curva partendo dal punto (0,0) fino al valore sulle ascisse corrispondente a detto valore soglia rappresenta la porzione del campione che viene postalizzata, mentre il corrispondente valore sulle ordinate rappresenta la proporzione "attesa" dei soggetti che risponderanno positivamente. Minore *t*, maggiore la quota del campione che viene postalizzata. Anche in questo caso, più

velocemente sale la curva, meglio è, ovvero, maggiore l'area sotto la curva, migliore il classificatore.

4.2.2 Problema delle classi sbilanciate

Datasets con distribuzioni delle classi sbilanciate sono piuttosto frequenti in scenari reali. In particolare, nell'ambito dell'individuazione di frodi, spesso i frodatori sono una esigua minoranza dei soggetti analizzati. In questi casi, metriche di valutazione come ad esempio l'accuratezza possono non essere adatte a valutare la bontà di un classificatore: esempio tipico sono le frodi nell'ambito dell'utilizzo delle carte di credito. Se l'1% delle transazioni fosse fraudolento, un classificatore a maggioranza (ovvero che esprime il proprio responso in base a com'è la maggioranza delle osservazioni che osserva) non predirebbe mai che una certa transazione è in realtà fraudolenta; ciò non di meno, avrebbe un'accuratezza del 99%, pur essendo del tutto inutile. Sono state quindi studiate tecniche che "aiutano" un classificatore in situazioni in cui occorre predire una classe rara, tra cui il *cost sensitive learning* e approcci c.d. *sampling-based.*

4.2.2.1 *Cost sensitive learning*

L'apprendimento guidato da costi presuppone una matrice (di costi) in cui vengono codificate le penalità nel classificare i records del *test set* in maniera errata. Se $\mathbb{C}(i,j)$ denota il costo nel predire un record di classe i come se fosse di classe j, $\mathbb{C}(+,-)$[2] rappresenta il costo di commettere un errore di falso negativo, mentre $\mathbb{C}(-,+)$ rappresenta un falso positivo. Un valore negativo in una qualche cella della matrice può rappresentare la ricompensa per aver effettuato una corretta classificazione.

Dato un insieme di N *test records*, il costo complessivo di un modello M è dato da:

$$\mathbb{C}_t\,(M) = TP \times \mathbb{C}(+,+) + FP \times \mathbb{C}(-,+) + FN \times \mathbb{C}(+,-)\; TN \times \mathbb{C}(-,-)$$

Con una matrice 0/1, i.e. $\mathbb{C}(+,+) = \mathbb{C}(-,-) = 0\; et\; \mathbb{C}(+,-) = \mathbb{C}(-,+) = 1$, il costo del modello è pari al numero degli errori di misclassificazione.

Una tecnica *cost sensitive* tiene conto della matrice di costi nella fase di costruzione del modello al fine di generare quello che ha il costo complessivo minore.

Ad esempio, se gli errori di tipo *falso negativo* fossero più costosi rispetto a quelli di *falso positivo*, l'algoritmo cercherebbe di ridurre il tipo di errori più costoso, scegliendo la classe negativa solo se abbastanza "sicuro": in questo modo tenderebbe ad etichettare in

[2] La notazione utilizzata nel seguito è ripresa da [TSK06].

modo positivo più frequentemente rispetto al caso in cui i due tipi di errore avessero lo stesso peso. Ciò porterebbe ad una riduzione dei falsi negativi FN, ma, al contempo, verosimilmente, ad un aumento dei falsi positivi, FP.

Ci sono molti modi per incorporare in un algoritmo di classificazione l'informazione sui costi degli errori. Nel contesto degli alberi di decisione, ad esempio, il costo di classificazione può essere utilizzato in vario modo e in diverse fasi della generazione dell'albero: nella selezione dell'attributo migliore per *splittare* i dati, nel determinare se un sotto albero debba essere potato o modificando la regola di decisione ad ogni nodo foglia, tra gli altri.

In particolare, per quest'ultimo caso, sia $p(i|t)$ la frazione dei *training records* di classe i che appartengono al nodo foglia t. Una tipica regola di decisione per un problema di classificazione binaria è quella di assegnare la classe i al nodo t se $p(i|t) > 0{,}5$, ovvero in base alla classe di maggioranza dei *record* che raggiungono una determinata foglia. Allora, al posto di decidere in base alla maggioranza, un algoritmo *cost sensitive* può assegnare la classe i al nodo t se minimizza la seguente espressione:

$$\mathbb{C}(i|t) = \sum_j p(j|t)\mathbb{C}(j,i)$$

Nel caso in cui $\mathbb{C}(+,+) = \mathbb{C}(-,-) = 0$, un nodo viene assegnato alla classe positiva se:

$$p(+|t)\mathbb{C}(+,-) > p(-|t)\mathbb{C}(-,+) \Rightarrow$$
$$p(+|t)\mathbb{C}(+,-) > \big(1 - p(+|t)\big)\mathbb{C}(-,+) \Rightarrow\ p(+|t) > \frac{\mathbb{C}(-,+)}{\mathbb{C}(-,+) + \mathbb{C}(+,-)}$$

Questa espressione suggerisce di modificare la soglia per la regola di decisione da 0,5 a $\mathbb{C}(-,+)/(\mathbb{C}(-,+) + \mathbb{C}(+,-))$ per ottenere un classificatore *cost sensitive*. Se $\mathbb{C}(-,+) < \mathbb{C}(+,-)$ allora la soglia sarà inferiore a 0,5. Questo risultato è bene interpretabile perché il costo di commettere un errore di tipo *falso negativo* è più alto rispetto a uno di tipo *falso positivo*; quindi, abbassando la soglia, si è più "generosi" verso la classe positiva.

4.2.2.2 Approcci *sampling-based*

Il *sampling* costituisce un altro approccio largamente utilizzato per far fronte al problema delle classi sbilanciate. L'idea, in questo caso, è quella di modificare la distribuzione delle istanze in modo tale che la classe di minoranza sia meglio rappresentata nel *training set*. Le tecniche di *sampling* principali sono l'*undersampling*, l'*oversampling* e combinazioni dei due.

Un esempio potrà aiutarci a comprendere meglio come funzionano dette tecniche. Immaginando di avere a disposizione un *dataset* che contenga 100 esempi positivi e 1.000 negativi, nel caso

dell'*undersampling*, un sottoinsieme casuale di 100 elementi viene estratto dai 1000 negativi e, assieme ai 100 positivi, costituirà il *training set* per i classificatori di apprendimento utilizzati. Un problema che si pone utilizzando tale tecnica è dato dalla perdita di informazioni che questo metodo comporta, perché tra le istanza negative perse ve ne potrebbero essere di interessanti ai fini dell'analisi. La loro esclusione potrebbe quindi dar luogo alla creazione di modelli non ottimali.

Un possibile rimedio potrebbe essere quello di ripetere gli esperimenti più volte, ogni volta modificando il *training set* e di indurre quindi molti classificatori (spirito analogo all'*ensemble learning – vedi infra* par. 4.2.3).

Con l'*oversampling*, invece, si replicano gli esempi della classe minoritaria, fintanto che nel *training set* le istanze delle due classi si trovino nella proporzione desiderata. In presenza di dati con rumore, questa tecnica potrebbe amplificarne gli effetti, replicando, appunto, il rumore. L'*oversampling* non aggiunge informazione al modello, ma evita che il classificatore poti parte dello spazio delle osservazioni positive per scarsa frequenza delle stesse e classifichi come negative le poche istanze positive sparse nel *dataset*. Sostanzialmente, nel caso degli alberi di classificazione, l'*oversampling* crea artificiosamente un numero sufficientemente elevato di istanze positive da poter essere inserite in una foglia; senza l'utilizzo di una tecnica del genere, le poche istanze positive presenti avrebbero, con ogni probabilità, costituito la parte "impura" di una foglia etichettata come negativa.

L'approccio ibrido combina in vario modo le due tecniche brevemente descritte sopra.

4.2.3 *Ensemble methods*

Le tecniche di classificazione esistenti possono predire il valore di una classe ignota usando un singolo classificatore, indotto dall'analisi dei dati di *training*.

Esistono tuttavia tecniche che possono migliorare l'accuratezza della classificazione mediante l'aggregazione delle predizioni effettuate da più classificatori. Dette tecniche sono note come *ensemble methods*, mediante le quali viene costruito un insieme di classificatori di base a partire dal *training set*, e si arriva alla classificazione finale combinando in vario modo le predizioni di ciascuno di essi. Gli *ensemble methods* tendono, sotto determinate condizioni, ad avere *performances* migliori rispetto ai singoli classificatori.

Un esempio può chiarire questa affermazione[3]. Supponiamo di utilizzare $N = 25$ classificatori binari, ciascuno con un *error rate* ϵ di 0.35 e ipotizziamo che la classificazione sia effettuata a maggioranza, (come nel *bagging*). Allora, se i classificatori di base sono indipendenti – ovvero se i loro errori sono tra loro incorrelati – allora viene

[3] Esempio tratto da [TSK06]

predetta la classe sbagliata solo se la maggioranza dei classificatori di base predice la classe sbagliata. La probabilità che ciò avvenga, considerando i dati appena citati, è pari a:

$$\sum_{13}^{25} \binom{25}{i} \epsilon^i (1-\epsilon)^{25-i} = 0.06 = \epsilon_{ensemble}$$

sensibilmente inferiore al valore di ϵ originario.

Vi sono però delle condizioni che i singoli classificatori di base devono soddisfare affinché poi l'*ensemble* funzioni meglio di un singolo classificatore: la prima è che essi siano realmente indipendenti (tale condizione in realtà è molto stringente e nella viene spesso rilassata in vario modo), la seconda è che i classificatori di base devono ottenere risultati migliori di un classificatore *random* [TSK06].

Un *ensemble classifier* può essere costruito mediante l'impiego di varie tecniche, tra cui si ricordano:

Manipolazione del training set: secondo questo approccio, si derivano molti *training sets* per campionamento di quello originario, secondo un qualche schema (casuale, stratificato, con o senza reimmissione, ecc…), che può variare da un *training set* derivato all'altro. I singoli classificatori (normalmente basati sullo stesso algoritmo di apprendimento) vengono poi costruiti sulla base dei singoli *training sets* derivati. *Bagging*, *boosting* e *stacking* (*stacked generalization*) sono esempi di tecniche di questo tipo.

Manipolazione degli attributi di *input* del modello: i singoli classificatori vengono addestrati su sottoinsiemi diversi (non necessariamente disgiunti) delle *features* originali. Anche in questo caso si usa di solito lo stesso algoritmo di classificazione per ogni *training set*. Questa tecnica funziona bene in caso di attributi ridondanti sul *training set* originario. *Random forests* è un metodo che sfrutta il principio appena descritto, manipolando gli attributi di *input* e utilizzando alberi di classificazione quali classificatori di base. Nella versione originale descritta in [Bre01], la tecnica in esame prevedeva la combinazione di *bagging* e selezione *random* degli attributi. Gli algoritmi di base erano alberi decisionali (CART), di qui il nome "foresta" dato al metodo.

Manipolazione dell'*output* dei classificatori: è un metodo che può essere utilizzato quando le classi sono in numero sufficientemente grande. Il *training set* viene trasformato più volte in classi binarie partizionando in maniera *random* le etichette in due insiemi, A_0 e A_1. Gli esempi che appartengono ad A_0 sono etichettati con 0 e quelli che appartengono ad A_1 sono etichettati con 1. Gli esempi rietichettati sono utilizzati per allenare un classificatore di base. Ripetendo la rietichettatura delle classi e creando classificatori ogni volta, si ottiene un *ensemble* di classificatori di base. Quando si presenta un caso di test, ogni classificatore di base C_i è utilizzato per predirne la classe. Se il test di esempio è predetto di classe 0, allora tutte le classi che appartengono ad A_0 ricevono un voto, altrimenti lo riceveranno

quelle appartenenti ad A_1. I voti vengono quindi conteggiati e alla classe che avrà ricevuto più voti sarà assegnato l' esempio di test. Un esempio di questo approccio è dato dall'*error correcting output coding* proposto in origine in [DB95].

Questi approcci, di carattere generale, possono essere applicati a qualsiasi tipo di classificatore. I classificatori di base possono essere generati sia in sequenza (uno dopo l'altro), sia in parallelo (tutti in una volta). L'algoritmo 3.1 mostra la procedura generale per la costruzione di un *ensemble classifier.* In seguito approfondiremo due tecniche in particolare, su citate, *bagging* e *boosting*, che verranno utilizzate "sul campo" nel prossimo capitolo.

Algoritmo 4.1 Procedura generale per *ensemble classifier*

1: Sia D il training set originario, k il numero di classificatori di base e T il *test set*.
2: **for** i = 0 to k **do**
3: Crea il *training set* D_i a partire da D
4: Costruisci un classificatore di base C_i a partire da D_i
5: **end for**
6: **for each** *test record* $x \in$ T **do**
7: $C^*(x) = Combine\,(C_1(x), C_2(x), \dots, C_k(x))$
8: **end for**

Figura 4.2: procedura generale per *ensemble classifiers* (adattato da [TSK06])

4.2.3.1 Bagging

Il *bagging* (*bootstrap aggregation*) consiste semplicemente nell'addestrare i singoli classificatori su repliche del *training set* originale ottenute con il metodo *bootstrap,* ovvero con tante estrazioni con reimmissioni delle varie istanze [Bre96]. Generalmente, i *training set* generati hanno tutti la stessa cardinalità. I singoli classificatori vengono poi combinati con la regola di maggioranza o con la media semplice delle loro uscite. Ci si potrebbe aspettare che i vari alberi allenati sui diversi *training set* diano risultati simili. In realtà ciò in genere non avviene, soprattutto se i *training sets* sono relativamente piccoli, perché l'induzione mediante alberi di decisione è un processo instabile: piccoli cambiamenti nel *training set* possono provocare la scelta di diversi attributi di *split*, con conseguenze notevoli per i sottoalberi che ne derivano.

Algoritmo 4.2 Bagging (generazione del modello e classificazione)

1: Sia n il numero di istanze del *training set* originario D, t il numero dei modelli, T il *test set*
2: **for each** iterazione **t do**
3: Campiona n istanze con reimmissione da D
4: Applica l'algoritmo di apprendimento sul campione ottenuto
5: Memorizza il modello, m_i
5: **end for**
6: **for each** modello m **do**
7: Predici la classe di ogni istanza r di T
8: **end for**
9: **for each** istanza r di T **do**
10: **return** la classe predetta più spesso dai modelli m_i
11: **end for**

Figura 4.3: *bagging* (adattato da [WFH11])

Questa tecnica riduce la varianza della predizione con il voto, ovvero, sostanzialmente, facendo una media delle predizioni, riducendo in tal modo l'errore atteso complessivo.

4.2.3.2 Boosting

Gli algoritmi di *boosting* (tra cui il noto *Adaboost*) consistono nell'addestrare in sequenza i singoli classificatori dello stesso tipo (non quindi in parallelo come avveniva per il *bagging)*, facendo in modo che ciascuno si concentri sulle istanze del *training set* classificate in modo errato dai classificatori che lo hanno preceduto [FS97][4]. A questo scopo viene associato un peso ad ogni istanza del *training set*, il cui valore viene modificato durante l'addestramento dei classificatori in modo che le istanze classificate in modo errato abbiano sempre un peso maggiore di quelle classificate correttamente. Approfondiamo il metodo in questione spiegando la variante nota come *Adaptive Boost – Adaboost*, implementata specificatamente per la classificazione. I classificatori di base devono saper trattare istanze pesate. La presenza di istanze pesate incide sul modo in cui viene calcolato l'errore del classificatore stesso, pari alla somma dei pesi delle istanze classificate erroneamente fratto la somma complessiva dei pesi, anziché essere la frazione delle istanze misclassificate. Assegnando un peso alle varie istanze, il classificatore viene indotto a concentrarsi su un particolare sottoinsieme delle stesse, ovvero su quelle che pesano di più, perché commettere errori su di esse provoca un deterioramento della *performance* del modello stesso.

Una traccia del metodo *Adaboost* viene di seguito riportata:

[4] Per il loro lavoro, Freund e Schapire vinsero il Premio Gödel nel 2003.

Algoritmo 4.3 Boosting (generazione del modello)

```
1:  Si assegni un peso uguale a tutte le istanze del training set, D
2:  for each iterazione t do
3:      Applica l'algoritmo di apprendimento su D pesato e memorizza il
        modello che ne risulta
4:      Computa l'errore e del modello su D pesato e memorizzalo
5:      if (or((e == 0); (e ≥ 0.5))) then
6:          termina la generazione del modello
5:      else for each istanza di D
6:          if (istanza correttamente classificata) then
7:              Moltiplica il peso dell'istanza per e/(1 − e)
8:      end for
9:      Normalizza i pesi di tutte le istanze
10: end for
```

Figura 4.4: *boosting* – generazione modello (adattato da [WFH11])

Algoritmo 4.4 Boosting (classificazione)

```
1:  Si assegni un peso uguale a 0 a tutte le classi
2:  for each modello t (o s < t) do
3:      Aggiungi −log(e/(1 − e)) al peso della classe predetta dal modello
4:  return la classe col peso maggiore
```

Figura 4.5: *boosting* - classificazione (adattato da [WFH11])

Osserviamo, per concludere, che i metodi *ensemble* funzionano meglio con classificatori instabili, ovvero classificatori sensibili a perturbazioni anche piccole del *training set*. Alberi di classificazione e regole di decisione sono esempi di classificatori di questo tipo. Tuttavia, se da un lato questi metodi riducono generalmente l'errore di classificazione, spesso presentano in *output* in forme scarsamente intellegibili, al contrario di alberi e regole, che hanno, come caratteristica positiva, quella di fornire risultati in un formato ben comprensibile ed intuitivo.

4.3 Rassegna della letteratura

Dopo aver richiamato gli aspetti fondamentali delle tecniche di classificazione, vediamo ora in che modo l'attenzione della comunità scientifica – economisti, matematici, *computer scientists* – si sia posta nei confronti dello specifico problema della lotta all'evasione fiscale e della ricerca di possibili metodologie applicabili in detto campo, al fine di consentire agli organi preposti di individuare con sempre maggiore efficacia i soggetti fraudolenti.

Iniziamo col dare una giustificazione economica al comportamento del contribuente, seguendo il modello proposto da [AS72]: la scelta del contribuente se evadere o meno dipende dal

confronto tra un profitto certo nel caso di non evasione e un altro profitto, atteso, in caso di evasione.

Nell'ipotesi semplicistica di neutralità al rischio, il contribuente guarda semplicemente all'ammontare atteso delle imposte da pagare in caso di evasione, $E[t]$,pari a:

$$E[t] = t(y_d) + (1+\theta)p[t(y) - t(y_d)]$$

dove

- p = probabilità di essere controllati e sanzionati
- y = reddito vero
- t(y) = imposta da pagare su reddito vero
- y_d = reddito dichiarato,
- t(y_d) = imposta da pagare su reddito dichiarato
- $y_d \leq y$,
- t(y_d)≤t(y),
- θ = sanzione proporzionale all'imposta evasa.

In caso di non evasione (onestà) il contribuente dovrebbe pagare t(y) e quindi il contribuente sceglie di evadere solo se:

$$\mathrm{E}[t] \leq t(y) \xrightarrow{el.alg.}$$

$$[t(y_d) - t(y) + (1+\theta)p[t(y) - t(y_d)] < 0 \xrightarrow{el.alg.}$$

$$[t(y_d) - t(y)] * [1 - (1+\theta)p] < 0$$

dato che $[t(y_d) - t(y)] < 0$

la diseguaglianza è verificata se e solo se

$$[1 - (1+\theta)p] > 0 \xrightarrow{el.alg.} (1+\theta)p < 1 \xrightarrow{el.alg.}$$

$$p < 1/(1+\theta)$$

In questo caso il contribuente evade se la probabilità e/o la sanzione sono "troppo basse". Se consideriamo anche la possibilità di non neutralità al rischio, la scelta se e quanto evadere dipende da 4 fattori:

- probabilità del controllo;
- entità della sanzione;
- attitudine del contribuente verso il rischio;
- reddito reale del contribuente e aliquote di imposta.

Secondo le previsioni del modello:

- l'evasione dovrebbe diminuire all'aumentare della probabilità del controllo e dell'entità della sanzione;
- l'evasione dovrebbe diminuire se il contribuente è avverso al rischio;
- l'evasione potrebbe diminuire o aumentare al variare dell'aliquota.

E' evidente che alcune delle variabili di cui sopra sono esogene rispetto all'operato dell'Agenzia delle Entrate (e gli analoghi enti degli

altri Paesi), come l'entità della sanzione, il reddito prodotto o le aliquote. Ma avendo a disposizione una metodologia con cui effettuare controlli efficaci ed efficienti, la probabilità stessa di controllare un evasore (magari "interessante") aumenta (per la riduzione dei controlli a soggetti non frodatori, a parità di numero di controlli) e una conseguenza di ciò potrebbe essere l'induzione di un aumento di avversione al rischio nella collettività dei contribuenti (effetto deterrenza).

Una "metodologia" che supporti la selezione dei soggetti da controllare non può non passare da un'attenta analisi dei dati a disposizione, che deve portare a trovare (estrarre) informazione non banale, implicita nei dati, non nota in precedenza e utile: in definitiva, un'analisi condotta con strumenti di *data mining*.

Va però sottolineato come una maggiore efficacia dell'attività di controllo non sia sufficiente per l'azzeramento dell'evasione (per i motivi visti nel primo capitolo), ma ciò non di meno, può senz'altro costituire un valido freno al fenomeno.

Il settore dell'evasione fiscale non è certo l'unico in cui possono prendere vita e forma le frodi: a ben vedere esso è un sottoinsieme del più vasto insieme delle frodi finanziarie. Anche se non vi è una definizione universalmente accettata di frode finanziaria, [WLTH06] propongono la seguente, che ci sembra condivisibile: "*un deliberato atto contrario alla legge, a una regola, a una politica, con l'intento di ottenere un beneficio finanziario non autorizzato*".

In [NHWCS11] viene proposto un modello grafico di classificazione della letteratura esistente in materia di *frodi finanziarie*, nel quale, per i possibili ambiti fraudolenti individuati (bancario, assicurativo, del mercato mobiliare e merci, oltre ad una categoria residuale – *altro*), sono individuate le sei classi di tecniche di *data mining* più frequentemente impiegate (*classificazione, regressione, clustering, predizione, outlier detection e visualizzazione*):

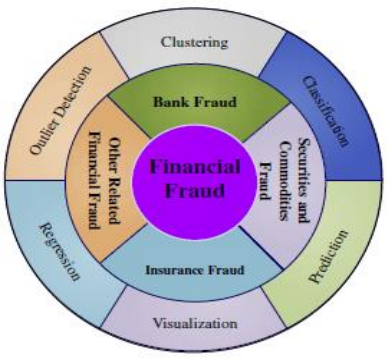

Figura 4.6: classificazione delle frodi finanziarie e tecniche di mining utilizzate.

Gli autori sottolineano come, dall'esame degli articoli pubblicati in questo ambito, la maggior parte (30 articoli su 49) impieghi tecniche di classificazione, con diverse scelte di algoritmi, dalle *reti neurali artificiali* ([FCS95], [CC99], [DGSC97]) ai modelli di regressione logistica ([Spa02]), dai modelli *bayesiani naive* ([VDD04])

agli alberi decisionali ([KSM07], [KKTT06]), ai *support vector machines* ([KPHKB03]) tra gli altri.

In un lavoro simile, [PLSG05], sono analizzati, confrontati e riassunti articoli scientifici in cui vengono proposti modelli e soluzioni di *data mining* in tema di rilevamento automatizzato delle frodi in molteplici settori[5]. Anche in questo caso, la rassegna della letteratura presentata, mostra chiaramente che le tecniche di classificazione sono di gran lunga le più utilizzate: in particolare, si ricordano modelli quali alberi decisionali, regole induttive e *case-based reasoning*.

Infine, su un *dataset* di riguardante carte di credito, [WFYH03] costruiscono un classificatore "composto", con classificatori di base dati da alberi decisionali derivati con l'algoritmo C4.5 e in cui la regola decisionale finale è data dalla media pesata dei benefici attesi indotti da ciascun albero di base. Gli autori mostrano come l'*ensemble classifier*, applicato a dati reali, presenti buone probabilità di dare risultati migliori rispetto a un singolo classificatore.

Ci si può chiedere che cosa abbiano in comune i modelli elaborati in letteratura, al di là delle specifiche tecniche utilizzate.

In primo luogo, si nota come lo scopo principale dei sistemi o modelli di rilevamento di frodi, indipendentemente dal dominio di applicazione, sia sempre quello di individuare tendenze (*patterns*) generali dei comportamenti sospetti e/o fraudolenti. Nel caso delle assicurazioni, ad esempio, i truffatori richiedono all'assicuratore di beneficiare di trattamenti per i quali non hanno titolo, attraverso l'utilizzo di informazioni falsificate, oppure richiedono beni e servizi (ad es. finanziamenti) utilizzando dati di persone inesistenti oppure tramite "furti di identità". Nel caso di frodi transazionali, il truffatore utilizza illegalmente un *account* (es. carta di credito) legittimo. Nel caso delle imposte, il reddito dichiarato non corrisponde con il reddito realmente conseguito in un dato anno. Si tratta quindi di andare a vedere se il comportamento dei frodatori in qualche modo possa essere riconosciuto da un qualche algoritmo di classificazione.

In secondo luogo, negli ambiti sopra citati, chi predispone modelli di *mining*, usualmente associa un valore monetario alle predizioni dagli stessi generate in *output*, al fine di massimizzare risparmi di spesa o profitti, in base a criteri e politiche prefissate; in questo ambito si possono definire modelli con costi espliciti (come in [PAL04]) o con benefici attesi.

---

[5] In questo lavoro vengono individuati i seguenti settori in cui si possono manifestare frodi: *internal fraud detection,* che si occupa di scoprire le false comunicazioni sociali da parte degli amministratori delle aziende o le anomale transazioni *retail* effettuate dai dipendenti; *insurance fraud detection* in cui vengono riconosciuti quattro sottogruppi principali: assicurazione casa, assicurazione in agricoltura, assicurazione RCA e assicurazione medica; *credit fraud detection*, che si riferisce allo *screening* delle domande per l'ottenimento di un finanziamento e/o delle transazioni con carte di credito; *telecommunications fraud detection*; *e-business ed e-commerce* (che presentano un duplice *task* per l'analista di *data mining*, perché al confine tra sistemi di rilevamento delle frodi e sistemi di rilevamento delle intrusioni di rete) e infine il settore del rilevamento delle frodi da parte di organizzazioni governative, tra cui le frodi fiscali.

Ancora, la maggior parte dei lavori basati su algoritmi supervisionati, da lungo tempo ha abbandonato misure di *performance* quali *true positive rate* (frodi correttamente individuate rispetto al reale numero di frodi) e *accuracy* (numero di istanze correttamente classificate sul numero totale di istanze). Infatti, nell'ambito del rilevamento delle frodi, i costi di misclassificazione (costi dei falsi positivi e dei falsi negativi) non sono in genere uguali tra loro e quindi altre metriche sono prese in considerazione. Ad esempio, [VDD04], nel loro studio con algoritmi supervisionati hanno cercato di massimizzare l'area sotto la *receiver operating curve* (AUC) e minimizzare la *cross entropy* (CXE)[6]. Oppure, [FS04] cercano di minimizzare lo *score di Brier* descritto in [Bri50].

4.3.1 *Tax fraud detection* in letteratura

Alcuni lavori presenti in letteratura riguardano specificatamente il tema della *tax fraud detection*.

In uno dei primi lavori in questo ambito, [BGMP99], il modello predittivo utilizzato è un classificatore binario basato su alberi di decisione, generati secondo l'algoritmo C5.0 di J. R. Quinlan.

Il *dataset* di riferimento utilizzato per la sperimentazione conteneva informazioni raccolte in 175 attributi relative a 80.643 società medio grandi, di cui 4.103 sottoposte a controllo fiscale. I dati degli accertamenti, relativi a questi ultimi soggetti, sono stati registrati in un *dataset* separato, con 7 attributi, di cui uno, *recovery*, rappresentava la pretesa erariale.

L'identificazione di un *modello dei costi* associato alle verifiche fiscali ha costituito una parte fondamentale del caso di studio ed è stato incluso nel modello, pertanto il recupero di ogni controllo (*actual recovery*) è stato assunto al netto dei costi dovuti dal controllo stesso.

La fase di *pre-processing* ha dato luogo ad un ridimensionamento del *dataset* di analisi, che alla fine contava 3.880 *record*, di cui 3.514 impiegati in fase di *training* e 366 in fase di *test*. In particolare, delle 366 tuple del *test set*, 118 erano a soggetti meritevoli di controllo e 248 a soggetti da non controllare.

E' interessante osservare come, oltre alla accuratezza del classificatore, fosse stato calcolato anche il valore complessivo di *actual recovery* ottenuto applicando il classificatore stesso al *test set* e sottoponendo a verifica tutti e soli i soggetti suggeriti dallo stesso.

Tale valore, quindi, è stato confrontato con il recupero di tasse evase realmente alle 366 tuple del *test set*:

- *audit # real = # test set = 366*
- *actual recovery(real)* = $\sum_{i \in test\ set} actual\_recovery(i)$= *159,6*
- *audit cost(real)* = $\sum_{i \in test\ set} audit\_costs(i)$ = *24,9*

[6] Come noto, la CXE costituisce una possibile misura dello scarto tra *score* predetto e *target score*.

dove costi e recupero sono espressi in milioni di euro. Poiché il *test set* era composto da soggetti controllati, confrontando i valori del caso reale con quelli dei soggetti classificati come positivi dai vari classificatori, è stato possibile valutare il potenziale miglioramento della pianificazione degli *audit* ottenibile grazie all'impiego di tecniche di *data mining*.

Nella costruzione del modello predittivo sono stati seguiti due approcci complementari, guidati da due possibili politiche di pianificazione dei controlli. Nel primo ci si poneva l'obiettivo di minimizzare i falsi positivi (FP), allo scopo di minimizzare le spese in verifiche fiscali inutili. Nel secondo, l'obiettivo era quello di minimizzare i falsi negativi (FN), allo scopo di perdere meno evasori possibili e quindi massimizzare il recupero di imposte evase.

In situazioni reali questi due obiettivi sono in conflitto tra loro, dato che il primo tende a minimizzare l'impiego di risorse anche a scapito di qualche evasore mancato, mentre il secondo tende a far grande uso di risorse, col rischio di controllare anche soggetti non evasori.

La soluzione ottimale si è rivelata consistere nel trovare un giusto compromesso tra le due. I seguenti parametri di *tuning* della costruzione di modelli hanno fornito il mezzo per raggiungere il suddetto *trade-off*: *livello di pruning*, *pesi di misclassifcazione* (pesi assegnati ai due tipi di errori di misclassificazione possibili, FP e FN), *replicazione della classe minoritaria nel training set* e l'*adozione della tecnica adaptive boosting*.

Tra gli esperimenti effettuati nel corso del caso di studio, quattro si sono rivelati particolarmente significativi, dando luogo ad altrettanti classificatori. Nei primi due casi è stata seguita la politica di minimizzare i FP (e quindi le risorse impiegate), negli ultimi due, invece, è stata seguita la politica di minimizzare i FN (più dispendiosa in termini di risorse, ma che individua più evasori).

Il primo classificatore è stato costruito secondo la politica di minimizzare i FP senza utilizzare pesi di misclassificazione, sfruttando il fatto che il *training set* originale, contenendo molti casi negativi (non evasori), avrebbe indotto il modello a classificare spesso come tali anche le nuove tuple, riducendo così indirettamente il rischio di assegnare erroneamente etichette positive. Per ridurre ulteriormente gli errori è stato applicato un *adaptive boosting* con 10 classificatori (numero scelto dopo diversi tentativi), ottenendo la seguente matrice di confusione:

classificato **negative**	classificato **positive**	
237	11	realmente **negative**
70	48	realmente **positive**

Il classificatore, quindi, suggeriva 59 verifiche di cui 11 inutili, ed ottiene i seguenti valori degli indicatori:

- *indice di misclassifcazione* = 22% (81 errori)
- *actual recovery* = 141.7 MEuro
- *audit cost* = 4 MEuro
- *profitabilty* = 2.401 MEuro/verifica

Se confrontato con i valori dell'intero *test set*, si osserva un recupero dell'88% delle imposte evase con solo il 16% delle verifiche, coerentemente con la politica al risparmio che si voleva ottenere.

Nel secondo classificatore la politica di minimizzare i *FP* è stata portata all'estremo, mediante utilizzo di pesi di misclassifcazione che fanno pesare di più gli errori di tipo *FP*. Più esattamente, agli errori *FP* è stato dato peso doppio rispetto a quelli *FN*. Inoltre, anche qui è stato applicato un *adaptive boosting*, questa volta con 3 classificatori (numero sempre scelto dopo diversi tentativi), ottenendo la seguente matrice di confusione:

classificato **negative**	classificato **positive**	
246	2	realmente **negative**
108	10	realmente **positive**

Il classificatore risultante suggeriva solo 12 verifiche di cui 2 inutili, ed ottiene i seguenti valori degli indicatori:

- *indice di misclassifcazione* = 30% (110 errori)
- *actual recovery* = 15.5 MEuro
- *audit cost* = 1.1 MEuro
- *profitabilty* = 1.291 MEuro/verifca

Si osserva che le pochissime verifiche suggerite portano anche a mancare molti evasori, rendendo il modello utile in quei casi in cui le risorse a disposizione siano limitatissime.

Il terzo classificatore, a differenza dei precedenti, adottava la politica di minimizzare i *FN*. A tal fine è stato manipolato il *training set* e le tuple positive sono state duplicate fino a raggiungere un bilanciamento tra le due classi; inoltre, sono stati adottati dei pesi di misclassifcazione che assegnavano agli errori *FN* un peso triplice rispetto a quelli *FP*. Anche qui è stato applicato un *adaptive boosting* con 3 classificatori ottenendo la seguente matrice di confusione:

classificato **negative**	classificato **positive**	
150	98	realmente **negative**
28	90	realmente **positive**

Con i parametri così specificati, il classificatore ha suggerito 188 verifiche, di cui, tuttavia, più della metà inutili, ottenendo i seguenti valori degli indicatori:

- *indice di misclassifcazione* = 34% (126 errori)
- *actual recovery* = 165.2 MEuro

- *audit cost* = 12.9 MEuro
- *proftabilty* = 0.878 MEuro/verifca

Sorprendentemente, il recupero netto del modello è superiore al caso reale con solo il 50% delle verifiche effettuate. Questo è dovuto naturalmente ai tanti casi *TP* (soggetti fraudolenti scoperti, 90 su 118) ottenuti. Si noti però come la *profitability* sia comunque molto più bassa dei modelli precedenti (per via dei tanti *FP* ottenuti) che si concentravano su pochi soggetti altamente proficui. Anche l'ultimo classificatore generato adottava la politica di minimizzare i *FN*, bilanciando le due classi nel *training set*, questa volta adottando dei pesi di misclassifcazione che assegnavano agli errori *FN* un peso quadruplo (anziché triplice come nel classificatore precedente) rispetto a quelli *FP*. In questo caso non è stato applicato alcun *boosting*.

La matrice di confusione ottenuta è la seguente:

classificato **negative**	classificato **positive**	
135	113	realmente **negative**
21	97	realmente **positive**

Il classificatore suggerisce 210 verifiche di cui circa metà inutili, ed ottiene i seguenti valori degli indicatori:

- *indice di misclassifcazione* = 36.6% (126 errori)
- *actual recovery* = 163.5 MEuro
- *audit cost* = 14.4 MEuro
- *profitabilty* = 0.778 MEuro/verifca

Rispetto al modello precedente, diminuendo i *FN*, venivano catturati più evasori, ma il maggior numero di *FP* ottenuti (verifiche effettuate a vuoto), compensando il guadagno conquistato, portava il recupero globale ad una piccola riduzione.

Come mostrato dagli esperimenti, l'uso di classificatori nella selezione di soggetti da sottoporre a verifica porta ad ottenere generalmente buoni recuperi economici con un ridotto numero di verifiche, mentre ponendo l'accento sul numero di evasori rintracciati o, inversamente, sulle limitate risorse a disposizione, si possono ottenere diversi risultati che forniscono un compromesso tra copertura degli evasori e risparmio nell'eseguire i controlli.

Gli alberi di classificazione sono di immediata lettura e comprensione e per questo sono spesso utilizzati in contesti in cui la predizione di una classe va "spiegata" perché il lavoro abbia una qualche utilità pratica[7].

Difatti, anche [YQJ03] presentano un caso di studio su come poter predisporre un sistema di rilevamento di dichiarazioni dei

[7] Ad esempio, le *reti neurali artificiali* non godono di questa proprietà.

redditi fraudolente, che fosse di supporto all'attività delle autorità governative.

Il sistema viene descritto, oltre che negli aspetti architetturali, anche in quelli algoritmici. Viene quindi specificato che la tecnica prescelta è quella della classificazione, implementata nella forma degli alberi decisionali, in particolare secondo l'algoritmo C5.0 di Quinlan.

Più recentemente, [AT12], nel proporre un approccio di contrasto all'evasione fiscale mediante tecniche di *data mining*, individuano nella tecnica della classificazione basata su alberi lo strumento migliore da utilizzare per la scoperta di profili di comportamenti fraudolenti. Detto studio fa riferimento al Marocco, con i dati relativi a 3.500 soggetti di piccole dimensioni (volume d'affari fino a 50.000.000 *dirhams* – i.e. circa € 4,5M) di cui 500 sottoposte a controllo. Le informazioni circa le società hanno riguardato volume d'affari, imposte dovute, imposte pagate e altri dati relativi all'attività svolta da ciascuna. L'algoritmo utilizzato è stato CART, nella sua implementazione sul *software* Tanagra[8]. I risultati degli esperimenti condotti sembrano essere molto incoraggianti e gli algoritmi impiegati si sono rivelati molto precisi, dato che l'*error rate* sul *test set* è stato di solo l'1,7%.

Altri lavori, anziché gli alberi, utilizzano le *regole associative* quale strumento di *mining*. In particolare, in [WOLCY12] si applica tale strumento al fine di migliorare la *performance* dell'attività di controllo delle autorità di Taiwan, con specifico riguardo all'evasione IVA. I dati sono stati raccolti dal *dataset* in uso presso la locale amministrazione finanziaria e contenevano informazioni relative alle dichiarazioni bimestrali presentate (*tax reports*)[9], alle attività di controllo eventualmente subite (*tax evasion control file*)[10] e ad alcuni dati anagrafici (*tax registration file*). I dati sono stati, come di consueto, suddivisi in *training* e *validation set* e le regole associative sono state generate sulla base del primo e testate sul secondo. Inoltre, sono stati raccolti i dati relativi al 2005 di alcuni soggetti e messi in un insieme a parte; da quest'ultimo insieme, sulla base delle regole associative trovate dal modello, sono stati selezionati alcuni soggetti da sottoporre effettivamente a controllo. La validazione del modello è stata effettuata tramite *(threefold) cross-validation*. Il *tool* utilizzato è stato DBMiner, *software* sviluppato dalla Simon Fraser University – Canada. Per fini pratici, sono stati creati due *training sets*, ciascuno con un sottoinsieme degli attributi di cui sopra, per allenare il

[8] Tanagra è un software gratuito di Data Mining per la didattica e la ricerca scritto dal Prof. R. Rakotomalala, dell'Università di Lione, Francia. Il pacchetto di installazione e le guide all'utilizzo del software, sono scaricabili dal sito Tanagra (http://eric.univ-lyon2.fr/)

[9] Nella fase di *pre-processing*, i sei *report* che vengono presentati in capo a un anno sono stati sintetizzati in modo da ottenere dati validi su scala annuale.

[10] Considerato l'utilizzo di dati sensibili, sono stati resi disponibili i dati relativi ad anni anteriori al 2006. Pertanto il *dataset* utilizzato contiene dati degli anni 2003 e 2004.

modello a regole e inoltre molti degli attributi numerici presenti sono stati discretizzati.

Supporto e confidenza delle regole sono stati settati in un intervallo di 4-15% e 80-90% rispettivamente, in modo da ottenere un *set* di regole ritenuto ottimale (si ricorda che maggiore il supporto, minore il numero di regole scoperte e che a parità di supporto, maggiore la confidenza, minore è il numero di regole).

Le regole ottenute sono del tipo:

$$BSCD38 = |03 \mid AND\ sales\_cap\_class\ =\ |01|\ AND\ sales\_class\ =\ |01| \rightarrow EVD\_TAX\_CLASS\ =\ |02|$$

che si interpreta come: *se la categoria di business di un soggetto è commerciale (classe 03), il rapporto tra vendite e capitale è compreso tra 0 e 10 (Classe 01) e il volume d'affari è inferiore al milione (Classe 01), allora il soggetto è classificato come soggetto la cui evasione è quantificabile tra 0 e 100.000 $ (classe 02).*

Dal primo *training set* sono state ottenute 16 regole, mentre dal secondo 10. Dette regole sono state successivamente, ove possibile, accorpate, per arrivare a 13 regole in totale. Due regole con lo stesso *pattern* possono essere accorpate, come ad esempio:

$$BSCD38 = |01 \mid AND\ return\_sales\ =\ |00|\ AND\ xmp\_sales\ =\ |00| \rightarrow EVD\_TAX\_CLASS\ =\ |01|$$

e

$$BSCD38 =\ |01 \mid AND\ return\_sales\ =\ |00|\ AND\ EVD\_TAX\_CLASS\ =\ |01| \rightarrow xmp\_sales =\ |00|$$

Ancora, due regole come le seguenti:

$$sales\_cap\_class = |01|\ AND\ valued\_rate = |09|\ \rightarrow EVD\_TAX\_CLASS\ =\ |01|$$

e

$$sales\_cap\_class = |01|\ AND\ valued\_rate = |08|\ \rightarrow EVD\_TAX\_CLASS\ =\ |01|$$

possono essere accorpate in:

$$sales\_cap\_class = |01|\ AND\ valued\_rate = |08{\sim}09|\ \rightarrow EVD\_TAX\_CLASS\ =\ |01|$$

In questo modo, può accadere che due regole che individualmente non raggiungono la soglia minima di supporto o confidenza, dopo l'unione possano superarle ed essere incluse nel modello.

L'aspetto interessante del lavoro appena descritto è che è stato confrontato il numero di controlli che sarebbero scaturiti secondo le modalità tradizionali (1.017) con quelli suggeriti dalle regole (746 con quelle derivate dal primo *training set* e 856 dal secondo). Pur essendovi molti FP, i TN sono anch'essi molti, per cui gran parte dei soggetti non selezionati dalle regole, in effetti non andavano selezionati.

Ancora, classificatori basati su regole sono stati utilizzati in [BGMPS09], nell'ambito di un progetto condotto congiuntamente dall'Università di Pisa e dall'Agenzia delle Entrate (c.d. progetto DIVA), il cui obiettivo era quello di fornire una metodologia utile per

l'identificazione di richieste di rimborso IVA fraudolenti, che fosse di supporto per l'attività di pianificazione e implementazione di controlli fiscali efficaci. Sono stati scelti classificatori basati su regole per le loro proprietà di alta espressività e comprensibilità rispetto ad altri modelli.

La costruzione del modello si è basata su dati storici concernenti le dichiarazioni IVA di 45.442 soggetti, ciascuna etichettata con un valore ottenuto in funzione dell'esito del controllo effettuato dall'Agenzia delle Entrate, attraverso un sistema di *scoring* che suggerisse i soggetti con una maggior probabilità di essere evasori e al contempo permettesse di minimizzare i falsi positivi.

Poiché il *dataset* di *training* del modello era costituito solo da soggetti accertati (e non poteva essere diversamente, in quanto per sapere se un soggetto è frodatore o meno, è necessario prima controllarlo), la funzione di *scoring* doveva servire a distinguere evasori "*interessanti*" da evasori "*non interessanti*", sotto il profilo congiunto di proficuità (ovvero entità della frode in valore assoluto), efficienza (ovvero entità del rimborso richiesto rispetto a quello dovuto) ed equità (ovvero entità della frode rispetto al volume d'affari del frodatore); l'ideazione di detta funzione si è resa quindi necessaria proprio perché nel *training set* non erano presenti soggetti non accertati e quindi era necessario, in qualche modo, trovare un metodo per distinguerli in base alla "gravità" dell'evasione commessa. In tal modo il frodatore "non interessante" veniva "approssimato" al soggetto onesto, o per lo meno, al soggetto verso cui possibilmente non indirizzare, in prima battuta, controlli.

Lo *scoring* dei soggetti è stato ottenuto grazie all'impiego di funzioni obiettivo binarie di primo livello molto semplici, successivamente aggregate da una funzione di secondo livello per ottenere lo *score* di ciascun soggetto.

Le prime, calcolate dopo una fase di *preprocessing* dei dati, sono del tipo:

$$Ob_1 = if\ \left(Valore_{frode} > S_1\right)\ then\ 1\ else\ 0$$
$$Ob_2 = if\ \left(Valore_{frode\_vol\_aff} > S_2\right)\ then\ 1\ else\ 0$$
$$Ob_3 = if\ \left(Valore_{frode\_crediti} > S_3\right)\ then\ 1\ else\ 0$$

Ovvero, per ogni dimensione della frode (valore assoluto, valore rispetto al volume d'affari e valore rispetto al rimborso richiesto), e fissati dei valori soglia per ciascuna di essa (di concerto con gli esperti del dominio), ad ogni *record* sono stati assegnati 3 valori. Detti valori sono stati aggregati da una funzione di secondo livello, anch'essa binaria, del tipo:

$$\begin{aligned} AND_{EXT}\left(Valore_{Ob_i}\right) = AND\ &\left(Valore_{Ob_i}\right) \\ &OR\ \left(Valore_{frode} > S_1'\right) \\ &OR\ \left(Valore_{frode\_vol\_aff} > S_2'\right) \\ &OR\ \left(Valore_{frode\_crediti} > S_3'\right) \end{aligned}$$

Ovvero, la funzione di secondo livello individua come soggetto interessante quello che soddisfa tutte e tre le funzioni di primo livello oppure presenta valori particolarmente alti in una o più di dette funzioni (le soglie impiegate da questa funzione sono, naturalmente, più alte rispetto a quelle di primo livello).

La funzione di secondo livello, da binaria, è stata trasformata in continua, ottenendo finalmente lo *score*:[11]

$$Score\ (x) = \begin{cases} 0, & se\ per\ x\ si\ ha\ \ AND_{EXT}(OB_i) = 0 \\ \prod_{i \in [1,k]} \left(\mathcal{N}(f_i(x))\right)^{p_i} & altrimenti \end{cases}$$

dove f_i e p_i sono rispettivamente trasformazioni delle funzioni di primo livello e pesi ad esse associate (per differenziare il contributo allo score di ciascuna funzione di base) e k=3.

Ottenuta la funzione di *score*, con essa è stato ottenuto il valore di classe dei singoli *records*.

A questo punto, la tecnica utilizzata per l'estrazione di un numero desiderato X di potenziali evasori più interessanti ha avuto il suo *core* in *Sniper*, tecnica che permette la modellazione di un classificatore binario, basato sul *training* di un *set* di classificatori di base, il cui *output* è dato dalla fusione in un singolo insieme di regole dei risultati ottenuti dai classificatori di base. L'approccio è simile, nello spirito, al *bagging*. Tuttavia, anziché arrivare ad un sistema di voto finale, *Sniper* considera tutti i *ruleset* generati dai classificatori di base, per poi fonderli in un insieme contenente le regole giudicate migliori, sotto il profilo della confidenza.

Una traccia dell'algoritmo di *Sniper* viene proposta di seguito:

Algoritmo 4.5 Sniper: selezione delle regole migliori

1: Sia M l'insieme delle regole estratte dal modello e R l'insieme delle regole estratte dai classificatori di base
2: $M = \emptyset$
3: $R = \{r \in R \mid \gamma(r) \geq \gamma_{min}\}$
4: **while** $R \neq \emptyset$ **do** *// prima condizione di stop*
5: $r^* = argmax_{r \in R}\{\gamma(r)\}$ *// sceglie la regola migliore*
6: $M = M \cup \{r^*\}$ *// aggiorna il modello corrente*
7: **if** $\#(M) \geq X$ **then** *// seconda condizione di stop*
8: **return** M
9: $R = \{r' = r \barwedge r^* | r \in R \backslash r^* \wedge (\gamma(r') \geq \gamma_{min}\}$ *// aggiorna il set di regole*
10: **return** M

Figura 4.7: Sniper (adattato da [GPS09])

[11] Alternativamente, la funzione di score può essere ideata aggregando per somma:

$$\mathcal{F}_{\Sigma}(x) = \sum_{i \in [1,k]} p_i \left(\mathcal{N}(f_i(x))\right)$$

L'operatore $\bar{\wedge}$ applicato a r e r' produce una regola che attiva tutti gli oggetti attivati da r e non da r'.

Con l'algoritmo sopra descritto, *Sniper* ha estratto 14 regole. I soggetti che ne soddisfano almeno una sono ritenuti "*frodatori interessanti*":

R1	$FLG_{VOLAFF} =' 1{,}0' AND$ $IMP_{IVA\_CREDITO} > 25966\ AND$ $IMP_{BENI\_ALTRI} \leq 348\ AND$ $COD_{DIR\_REG}\ in \begin{Bmatrix} 902, 903, 905, 906, \\ 908, 909, 910, 912, \\ 915, 916, 917, 918 \end{Bmatrix}$	• *Volume d'affari* < € 12.500 • *IVA a credito di competenza* > €25.966 • *Totale acquisti e importazioni (ALTRI):* bassi o assenti • *Domicilio* in (Valle d'Aosta, Liguria, Trentino-Alto Adige, Friuli-Venezia Giulia, Emilia Romagna, Marche, Umbria, Abruzzo, Molise, Puglia, Basilicata)
R2	$FLG_{VOLAFF} =' 1{,}0' AND$ $IMP_{IVA\_CREDITO} > 25966\ AND$ $IMP_{BENI\_ALTRI} \leq 348$	*Inclusa in R1*
R3	$FLG_{VOLAFF} =' 1{,}0' AND$ $IMP_{IVA\_CREDITO} > 25966\ AND$ $IMP_{BENI\_ALTRI} > 348\ AND$ $IMP_{TOT\_ACQ} > 135.028\ AND$ $FLG_{VJ} =' 0{,}0' AND$ $FLG_{DICHIAR\_SOST} =' 0{,}0' AND$ $VAL_{ALIQ\_MEDIA\_ACQ} > 16{,}66$	• *Volume d'affari* < € 12.500 • *IVA a credito di competenza* > €25.966 • *Totale acquisti e importazioni (ALTRI):* maggiori di € 348 • *Totale acquisti imponibili consistenti* • *Assenza quadro VJ* • *Assenza mod. 770* • *Aliquota media acquisti alta, ma non massima*
R4	$FLG_{VE} =' 0{,}0' AND$ $IMP_{ACQ\_IMPON} > 25.688\ AND$ $FLG_{DICHIAR\_SOST} =' 0{,}0' AND$ $FLG_{MUTUI} =' 0{,}0'$	• *Assenza quadro VE* • *Totale imposta su acquisti e importazioni maggiore di € 25.688* • *Assenza mod. 770* • *Assenza di mutui*
R5	$IMP_{BENI\_RIV} \leq 139.787\ AND$ $IMP_{ACQ\_IMPON} > 25.688\ AND$ $FLG_{VE} =' 0{,}0' AND$ $FLG_{DICHIAR\_SOST} =' 0{,}0'$	• *Totale acquisti e imp.(BENI DEST. ALLA RIV.) fino a circa € 140.000* • *Totale imposta su acquisti e importazioni maggiore di € 25.688* • *Assenza quadro VE* • *Assenza mod. 770*
R6	$FLG_{VOLAFF} =' 1{,}0' AND$ $IMP_{IVA\_CREDITO} > 25.681\ AND$ $FLG_{DICHIAR\_SOST} =' 0{,}0' AND$ $VAL_{TOT\_MUTUI} \leq 121.884\ AND$ $IMP_{CESS\_BENI} \leq 2.974\ AND$ $IMP_{IMPST\_DEB} \leq 184\ AND$ $IMP_{BENI\_RIV} \leq 139.787$	• *Volume d'affari* < € 12.500 • *IVA a credito di competenza* > €25.681 • *Assenza mod. 770* • *Presenza atti mutui* • *Cessioni beni ammortizzabili fino a € 2.974* • *Imposta IIDD molto bassa* • *Totale acquisti e imp.(BENI DEST. ALLA RIV.) fino a circa € 140.000*
R7	$FLG_{VOLAFF} =' 0{,}0' AND$ $IMP_{REDD\_IMP\_SMPL} > -56\ AND$ $FLG_{PIVA\_CESS\_ANNO} =' 0{,}0' AND$	• *Volume d'affari* > € 12.500 • *Presenza di reddito di impresa sempl.* • *Soggetto con P.IVA non cessata nel 2005* • *Soggetto con P.IVA attiva da almeno due*

	$FLG_{PIVA\_BREVE\_DUR} =' 0{,}0' AND$ $VAL_{ALIQ\_MEDIA\_ACQ} > 19{,}94\ AND$ $IMP_{TOT\_IMP} > 85.079\ AND$ $FLG_{VH} =' 1{,}0' AND$ $FLG_{LIQUIDAZIONE}\ in\ (0{,}000) AND$ $FLG_{DICHIAR\_SOST} =' 0{,}0' AND$ $IMP_{ECC\_PREC} < 7.640\ AND$ $VAL_{ALIQ\_M\_ACQ\_IMP} > 19{,}99\ AND$ $IMP_{BENI\_RIV} \leq 4$	*anni* • *Aliquota media sul totale acquisti molto alta* • *Totale imposta su acquisti imponibili consistente* • *Quadro VH compilato* • *Soggetto non in liquidazione* • *Assenza mod. 770* • *Eccedenza imposta risultante dalla precedente dichiarazione < 7.640* • *Aliquota media sugli acquisti imponibili massima* • *Totale acquisti e imp. (BENI DEST. ALLA RIV.) inesistenti*
R8	$IMP_{BENI\_RIV} \leq 4$ $IMP_{TOT\_IMP} > 85.079\ AND$ $VAL_{ALIQ\_M\_ACQ\_IMP} > 19{,}99\ AND$ $VAL_{ALIQ\_MEDIA\_ACQ} > 19{,}94\ AND$ $FLG_{DICHIAR\_SOST} =' 0{,}0' AND$ $FLG_{LIQUIDAZIONE}\ in\ (0{,}000) AND$ $IMP_{ECC\_PREC} < 7.640\ AND$ $FLG_{VOLAFF} =' 0{,}0'$	*Inclusa in R7*
R9	$VAL_{ALIQ\_M\_VOL\_IMP} \leq 0\ AND$ $IMP_{V\_AGG\_IMP} \leq -128.493\ 9\ AND$ $IMP_{BENI\_ALTRI} \leq 345$	• *Aliquota media su operazioni imponibili uguale a zero* • *Valore aggiunto imponibile molto negativo* • *Totale acquisti e imp. (ALTRI) molto bassi*
R10	$IMP_{V\_AGG\_IVA} \leq -8.251\ AND$ $VAL_{ALIQ\_MEDIA\_ACQ} > 19{,}995\ AND$ $IMP_{V\_AGG\_IMP} \leq -127.540\ AND$ $IMP_{BENI\_ALTRI} \leq 1.023$	• *Valore aggiunto molto negativo* • *Aliquota media sul totale acquisti molto alta* • *Valore aggiunto imponibile molto negativo* • *Totale acquisti e imp. (ALTRI) molto bassi*
R11	$IMP_{VE\_VOLAFF} \leq -1.225\ AND$ $IMP_{BEN\_AMM} \geq 110.000\ AND$ $DUR_{P\_PIVA\_MM} \leq 14$	• *Volume d'affari molto basso* • *Totale acquisti e imp. (BENI AMMORTIZZABILI) piuttosto alti* • *P.IVA in vita al max da 14 mesi*
R12	$VAL_{ALIQ\_MEDIA\_ACQ} > 19{,}97\ AND$ $IMP_{V\_AGG\_IVA} \leq -722.181\ AND$ $IMP_{VE\_VOLAFF} \leq 35.119\ AND$ $\notin \{R11\}$	• *Aliquota media sul totale acquisti molto alta* • *Valore aggiunto molto negativo* • *Volume d'affari molto basso* • *Non rientrano tra i soggetti individuati da R11*

R13	$VAL_{ALIQ\_MEDIA\_ACQ} > 19{,}98\ AND$ $IMP_{V\_AGG\_IMP} \leq -132.194\ AND$ $IMP_{V\_AGG\_IVA} \leq -972.049\ AND$ $IMP_{CRED\_AP} \leq 2.335\ AND$ $IMP_{ACQ\_NODETR} \leq 800\ AND$ $IMP_{IMPST\_CRED} \leq 180\ AND$ $\notin \{R11\}\ AND$ $\notin \{R12\}$	• *Aliquota media sul totale acquisti molto alta* • *Valore aggiunto imponibile fortemente negativo* • *Valore aggiunto fortemente negativo* • *Credito risultante dalla dichiarazione anno precedente basso o assente* • *Acquisti e importazioni per i quali non è ammessa la detrazione di imposta bassi* • *Imposta IIDD a credito bassa o assente* • *Non rientrano tra i soggetti individuati da R11 né da R12*
R14	$IMP_{V\_AGG\_IMP} \leq -104.834\ AND$ $DUR_{P\_PIVA\_MM} \leq 38\ AND$ $IMP_{BEN\_AMM} \leq 2.917\ AND$ $IMP_{VE\_VOLAFF} \geq 1.000.171\ AND$ $VAL_{ALIQ\_MEDIA\_ACQ} \geq 20\ AND$ $\notin \{R11\}\ AND$ $\notin \{R12\}\ AND$ $\notin \{R13\}$	• *Valore aggiunto imponibile fortemente negativo* • *P.IVA in vita al max da 38 mesi* • *Totale acquisti e imp. (BENI AMMORTIZZABILI) piuttosto bassi* • *Volume d'affari molto alto* • *Aliquota media sul totale acquisti massima* • *Non rientrano tra i soggetti individuati da né da R11, né da R12, né da R13*

Non sorprende che i profili individuati dalle regole presentino in generale un'IVA sugli acquisti alta in relazione al volume d'affari dichiarato (R1, R2, R3, R4, R5), corredata da imposte ai fini imposte dirette molto basse (R6), da un valore aggiunto (pari alla differenza tra l'imponibile delle vendite e quello degli acquisti) fortemente negativo (R12, R13, R14), e da aliquote medie di acquisto elevate, vicine alla massima (20%). L'insieme di regole estratte appare quindi ragionevole, in quanto dette caratteristiche si confanno a soggetti che formulano richieste di rimborso IVA.

Capitolo 5

Tax fraud detection: utilizzo di tecniche di classificazione

Dopo aver esposto i risultati più interessanti presenti in letteratura relativamente al tema della tax fraud detection nel capitolo precedente e dopo aver constatato che le analisi OLAP ci aiutano nella comprensione dei dati, ma non a disegnare profili di evasori che possono emergere dai dati stessi, ci approcceremo ora al dataset in ottica di mining, attraverso l'utilizzo di tecniche di classificazione, ponendo particolare attenzione ad alberi e regole di classificazione, con l'obbiettivo di individuare pattern interessanti e, soprattutto, utili, che possano fornire una guida flessibile ed efficiente per la pianificazione dei controlli fiscali, ad uso e vantaggio delle autorità a ciò preposte (segnatamente, l'Agenzia delle Entrate).

5.1 Costruzione del *dataset* di analisi

A partire dalla base dati descritta nell'ambito delle analisi OLAP svolte in precedenza, valuteremo, da un lato, valutare se e come procedere ad una fase di *features selection* e, dall'altro, se introdurre nuove variabili, utili al fine di costruire e poi testare i modelli predittivi, oltre che eliminare dall'analisi quei *record* ritenuti troppo devianti dagli altri e tali da essere poco rappresentativi della popolazione dei soggetti accertati sotto analisi. Difatti, il *dataset* finale sul quale andremo a costruire modelli di *mining* avrà bisogno, eventualmente, di vedere eliminati *record* e attributi di minore qualità e di avere nuovi campi calcolati a partire da quelli di origine, oltre a discretizzazioni e altre operazioni di preparazione all'analisi.

5.1.1 Eliminazione *record* devianti

Come evidenziato in sede di analisi OLAP, alcuni *record* relativi ai contribuenti accertati presentano valori anomali in quanto a ricavi e/o volume d'affari dichiarati e pertanto non saranno considerati nelle successive analisi (valori in euro). Saranno quindi esclusi dall'analisi i seguenti *record*, con le seguenti caratteristiche:

Codec_CF	Ricavi_Att.
0000000000123836	15.529.552
0000000001402088	15.303.108
0000000000274970	11.363.614
0000000002047247	9.471.714
0000000000088113	7.426.823
0000000000393022	5.439.730
0000000001453383	5.427.637
0000000000729101	5.421.131

Tabella 5.1: record eliminati perché devianti

La platea dei soggetti accertati si restringe quindi a 1835.

5.1.2 *Attribute selection*

In questa sezione illustriamo le scelte effettuate per quanto riguarda la selezione degli attributi più promettenti e la rimozione di quelli che invece sembrano apportare uno scarso contributo informativo, oppure sono già usati in altri attributi calcolati.

La tabella seguente riporta gli attributi eliminati perché rappresentano codici identificativi del soggetto (ad esempio [CODEC_CF]) oppure date (non utilizzabili dai modelli), come ad esempio [MIN_DATA_INI_ATTIV], oppure non valorizzati o valorizzati tutti allo stesso modo (ad esempio [COD_REG]) o ritenuti inutili (ad esempio [COD_CATASTALE], in quanto, considerata la non eccessiva numerosità dei dati, si ritiene che un dettaglio dei contribuenti al livello di comune di residenza non sia opportuno).

DIZIONE_LOC	AREA	FLG_RES_ESTERO
COD_REG	CODEC_CF	COD_STATO_EST_RES
DESC_REG	MIN_DATA_INI_ATTIV	CF_PROFESSIONISTA_2007
AREA_AGENZIA	COM_ESER_ATT	COD_CATASTALE

Tabella 5.2: attributi eliminati preliminarmente

Un altro tipo di analisi preliminare consiste nel calcolare e in seguito eliminare gli attributi che presentano, tra loro, una forte correlazione (a meno che tali attributi non rappresentino concetti completamente diversi tra loro e che solo accidentalmente risultino essere correlati nel *dataset* analizzato, nel qual caso vengono conservati entrambi). A tale scopo è stata costruita una matrice di correlazione di tutti gli attributi presenti nel *dataset* (ad esclusione di quelli riguardanti gli accertamenti, che saranno successivamente opportunamente eliminati in fase di *mining*) ed è stata poi scelta una

soglia di correlazione pari a 0,9 (tale valore naturalmente può essere calibrato diversamente in base agli obbiettivi di analisi): nelle coppie di attributi il cui coefficiente di correlazione superava tale soglia, è stato mantenuto quello ritenuto più significativo (per semantica, valorizzazione ecc…).

Di seguito si riporta la tabella degli attributi eliminati.

Attributo correlato mantenuto	Attributo correlato eliminato	Coefficiente di correlazione
IMP_PROD_NETTA	IMP_TOT_IMPST_IRAP	0,985
IMP_CMPNS_ATTIV_2007	IMP_COMPNS_SZ1_2007	0,998
IMP_REDD_LAV_AUT_2007	IMP_REDD_AUT_SZ1_2007	1
IMP_REDD_LORD_ORD	UTILE_PERD_CIV_ORD	0,992
IMP_REDD_LORD_ORD	IMP_REDD_IMP_ORD	1
RIM_FIN_SMPL	RIM_INI_SMPL	0,948
IMP_RICAVI_SMPL	IMP_TOT_CMPN_POS	0,949
REDD_IMP	IMPST_LRD	0,998
IMP_OPER_IMPVE23_2007	IMP_OPER_IMPVE23_IMPST_2007	0,951
IMP_VE_VOLAFF_2007	IMP_OPER_IMPVE23_2007	0,934
IMP_TOT_ACQ_2007	IMP_TOT_IMP_2007	0,982
TOT_ACQ	IMP_TOT_ACQ_2007	0,980
TOT_ACQ	IMP_ACQ_IMPON_2007	0,965
TOT_ACQ	IMP_IVA_DETR_2007	0,965
RICAVI_ATT_2007	TOT_ATT_2007	0,980

Tabella 5.3: attributi eliminati per correlazione

Un altro tipo di analisi effettuato riguarda le proprietà intrinseche di ogni singolo attributo: in particolare, sono stati esclusi dall'analisi i campi *fortemente valorizzati su un singolo valore*. Anche in questo caso è stato scelto di eliminare tutti i campi che presentavano una percentuale di occorrenza di un singolo valore (in tutti i casi, detto valore era lo zero) maggiore del 90%. Tali campi sono riportati nella tabella che segue:

RF_REDD_IMPR	PREST_SERV_INTRA	IMP_VE_SEZ1
PR_AGR_CORR	ACQ_BENI_INTRA	IMPST_VE_SEZ1
PR_AGR_ACQ	ACQ_BENI_INTRA_IMPST	ACQ_PLAF
IMP_LAV_DIP_2007	IMPORT_IMPO	NUM_DIP
AGR_REDD	IMPORT-IMPST	EXP_IMPO
CESS_BENI_INTRA		

Tabella 5.4: attributi eliminati per uniformità di valori

Infine, sono stati considerati i due attributi [COD_ATT_2007] e [COD_STUDIO_07]. Entrambi questi attributi forniscono informazioni sulla specifica attività svolta da un soggetto.

Tuttavia, osservando come il primo presenti ben 344 valori distinti, di cui 163 assunti da un solo *record*, e il secondo presenta 153

valori distinti, di cui 47 assunti da un solo *record*, tali attributi appaiono parcellizzare oltre modo i dati a disposizione e potrebbero portare i modelli di apprendimento a risultati non significativi, in quanto non pare ragionevole trarre conclusioni sui soggetti che svolgono una specifica attività osservando pochissimi suoi rappresentanti (al limite anche uno solo) tenuto conto del fatto che situazioni del genere possono provocare facilmente fenomeni di *overfitting* dei modelli. Codice attività e codice studio di settore sono poi molto legati tra loro, tant'è che uno studio di settore è specifico per attività o gruppi di attività omogenei tra loro. Pertanto, decidiamo di eliminare i due campi citati e di introdurre, quale indicatore del tipo di attività svolta da un soggetto, la famiglia di appartenenza dell'attività del soggetto stesso, che permette una maggiore generalizzazione dei risultati ottenuti.

5.2 Definizione degli obiettivi

In questa sezione illustreremo i passi ed i metodi con i quali siamo arrivati a costruire le *funzioni obiettivo* utilizzate successivamente per la generazione dei modelli predittivi. Prima di iniziare chiariamo il concetto di *funzione obiettivo.*

Nel nostro caso, una *funzione obiettivo* (come tutte le funzioni) è semplicemente una "regola", che chiamiamo $f(x)$, che ad ogni elemento x del *dataset* associa un valore, continuo o discreto, che indica il *grado di fraudolenza* dello stesso elemento x. Tali valori restituiti da $f(x)$ servono ai modelli predittivi di seguito utilizzati perché costituiranno, opportunamente discretizzati (addirittura, vedremo, binarizzati), i valori della *classe* che dovrà essere, dagli stessi modelli, predetta.

5.2.1 *Dataset* sbilanciato e nozione di *frodatore interessante*

Una considerazione di per sé banale, che tuttavia rappresenta una criticità per i modelli previsionali che andremo a costruire, è data dal fatto che tutti i soggetti del *dataset* sono soggetti in qualche misura fraudolenti o che per lo meno apparivano essere tali, essendo tutti stati selezionati per un controllo per l'anno d'imposta 2007. Quindi, il *dataset* di partenza è sicuramente *biased* rispetto alla reale popolazione dei contribuenti. Tale distorsione del dataset è tuttavia ineliminabile, perché abbiamo bisogno dei dati relativi agli accertamenti e i soggetti accertati sono già stati oggetto di selezione da parte dell'Agenzia delle Entrate, secondo un qualche criterio, non certo in maniera casuale.

Ciò premesso, avere una funzione obiettivo che etichetti come frodatore *tout court* chiunque avesse subito un accertamento, non sarebbe utile ai fini dei nostri obiettivi di classificazione. Infatti, avendo una base dati formata solo da soggetti fraudolenti, qualunque algoritmo concluderebbe la propria analisi proponendo un criterio "a

maggioranza", predicendo, quindi, sempre "*fraudolento*". E' evidente che tale risultato non può essere generalizzato su dati "nuovi" rispetto a quelli di *training*: in primo luogo perché la reale popolazione è composta anche di soggetti che fraudolenti non sono; in secondo luogo perché avere un modello che suggerisce di controllare chiunque, si rivelerebbe, ai fini pratici, assolutamente inutile.

Abbiamo quindi bisogno di una metodologia che permetta di differenziare tra loro i soggetti presenti nel *dataset*. Anche sulla scorta dell'esperienza riportata in [BGMPS09], scegliamo di focalizzare l'attenzione solo su un sottoinsieme dei soggetti accertati, che definiamo "*frodatori interessanti*", sulla base di due criteri guida:

- *Proficuità*: intesa come capacità di individuare soggetti con un alto quantitativo di frode, indipendentemente da tutti gli altri fattori.
- *Equità*: rappresenta la relazione fra la frode e altri fattori come la capacità contributiva, il volume d'affari, la grandezza in generale del soggetto.

I modelli che presenteremo dovranno quindi fornire indicazioni per predire se un soggetto appartiene alla categoria dei *frodatori interessanti* oppure no (per approssimazione, assumeremo che i soggetti effettivamente non frodatori possano essere trattati come *frodatori non interessanti*). Tale assunzione rappresenta un modo per ovviare, o comunque limitare, il *bias* originario presente nel *dataset*.

L'impostazione seguita, nella quale si ricercano solo i frodatori più gravi, senza avere la pretesa di individuare tutti i possibili evasori, ben si concilia con il mondo reale, in cui operano agenti (tipicamente, l'Agenzia delle Entrate) dalle risorse limitate (i funzionari impiegati nelle attività di controllo non sono infiniti) che si pongono l'obiettivo di ottimizzarne l'impiego (i controlli devono essere proficui e bene indirizzati).

Definiti questi concetti generali, abbiamo provveduto a trasporli nei dati a disposizione. A tale scopo sono stati progettati e costruiti ulteriori campi, ottenuti a partire dalle informazioni presenti nel *dataset*.

In primo luogo, per tenere conto del concetto di *proficuità* degli accertamenti, è stata introdotta la variabile [PRETESA_TRIB_IMPST], che rappresenta le maggiori imposte accertate o definite che costituiscono la pretesa erariale (al netto delle sanzioni, che però in genere sono commisurate alle maggiori imposte accertate, per cui non tenerne conto non inficia l'analisi). Tale valore è funzione dello stato degli accertamenti. In particolare, si deve tener conto della presenza di autotutela[1] e verificare se si è in presenza di attivazione di procedura di accertamento con adesione[2].

[1] L'autotutela nel settore tributario è il potere che ha l'Amministrazione finanziaria di intervenire quando la stessa si rende conto di aver commesso un errore che può danneggiare illegittimamente un contribuente; in sostanza, quando l'Amministrazione rileva che in un atto da essa emanato è contenuto un errore, in presenza del quale lo stesso atto non sarebbe stato emanato o avrebbe assunto un contenuto diverso, ha la

Pertanto, tale campo è dato dalla seguente espressione:

$if\ (StatoControllo = 18)$ // stato 18 = presenza di autotutela
$then\ 0$
$else\ if\ (StatoControllo = 19)$ // stato 19 = presenza di adesione
$then\ \sum(Valori\ definiti)$
$else\ \sum(Valori\ accertati)$

Per tener conto dell'*equità*, ovvero per guidare l'attività di controllo anche verso i soggetti più piccoli, nei confronti dei quali normalmente vengono emessi accertamenti con maggiori imposte accertate di più lieve entità rispetto ai soggetti più grossi, considerato sia che appare equo non discriminare i soggetti sulla base del loro volume d'affari, sia l'effetto deterrenza indotto sulla platea dei soggetti di piccole dimensioni da tali controlli, è stato creato il campo [PRETESA_SU_VA], espresso come:

$if\ (IMP\_VE\_VOLAFF\_2007 \neq 0)$
$then\ [PRETESA\_TRIB\_IMPST]\ /\ [IMP\_VE\_VOLAFF\_2007]$
$else\ 0$

A questo punto, la definizione di obiettivi specifici per i concetti chiave illustrati in precedenza è abbastanza agevole; di seguito, illustriamo due funzioni obiettivo che denominiamo di "1° livello" che traducono, rispettivamente, le definizioni di frode derivate dall'applicazione dei concetti sopra descritti di *proficuità* ed *equità*.

Tali funzioni obiettivo sono presentate nella versione binaria in base al confronto rispetto ad un valore di soglia.

Si evidenzia che i valori soglia devono essere calati nella realtà in cui i modelli vanno eseguiti, tenendo conto sia delle caratteristiche della platea dei soggetti potenzialmente controllabili (per un *dataset* riportante i dati di società di grandi dimensioni probabilmente le soglie, in particolare S_1, dovrebbero essere impostate su valori più elevati rispetto a quelli ritenuti validi per le ditte individuali ed i professionisti), sia del numero di controlli che si intendono eseguire (compatibilmente con le risorse disponibili).

possibilità non solo di emendarlo ma anche di ritirarlo, evitando in tal modo di danneggiare ingiustamente il contribuente nei cui confronti è stato emesso

[2] Tale istituto è disciplinato dal D.Lgs. 218/97, e consente al contribuente di definire le imposte dovute ed evitare, in tal modo, l'insorgere di una lite tributaria. Si tratta, sostanzialmente, di un "accordo" tra contribuente e ufficio che può essere raggiunto sia prima dell'emissione di un avviso di accertamento, che dopo, sempre che il contribuente non presenti ricorso davanti al giudice tributario.

Funzione	Descrizione	Espressione di calcolo
OB_1	Individua le evasioni superiori ad un valore di soglia (per caratterizzare il requisito della proficuità dell'accertamento)	$if\ (PRETESA\_TRIB\_IMPST > S_1)$ $then\ 1$ $else\ 0$
OB_2	Individua le evasioni percentualmente rilevanti rispetto al volume d'affari (per caratterizzare l'equità dell'accertamento)	$if\ (PRETESA\_SU\_VA > S_2)$ $then\ 1$ $else\ 0$

Tabella 5.5: funzioni obiettivo di primo livello

Riguardo al dimensionamento delle soglie, in questo lavoro sono stati utilizzati dei valori che fossero adeguati rispetto alla platea dei contribuenti considerati (persone fisiche e in genere attività di piccole dimensioni). In generale, calati nella realtà, tali valori soglia possono essere anche utili al fine di ottenere un partizionamento del *dataset* in soggetti *frodatori interessanti* e *frodatori non interessanti* secondo percentuali desiderate. Nel caso specifico, sono stati scelti i valori di soglia di seguito riportati:

Parametro	Valore	Effetto sulla funzione obiettivo
S_1	2.000	OB_1 rileva come "evasori" i soggetti la cui evasione supera i € 2.000,00.
S_2	0,0025	OB_2 rileva come "evasori" i soggetti la cui evasione supera il 2,5‰ del volume d'affari reale

Tabella 5.6: valori soglia

Una volta definite le funzioni che caratterizzano nel modo visto l'evasione dei singoli soggetti, occorre trovare il miglior modo o i migliori modi possibili per comporle in un'unica formulazione, fornendo un unico concetto di *frodatore interessante* che tenga conto contemporaneamente dei singoli aspetti su citati dell'evasione, ovvero ampiezza dell'evasione, in valore assoluto e rispetto al proprio volume d'affari.

La realizzazione di funzioni obiettivo complesse con la capacità di fornire il grado di interesse di un soggetto è stata derivata da [BGMPS09]. Qui di seguito riportiamo il primo *step*, dove abbiamo combinato i due obiettivi binari fra loro, sia attraverso l'operatore AND, sia attraverso l'OR, ottenendo due candidati a funzione (binaria) di secondo livello:

Funzione	Descrizione	Espressione di calcolo
$AND(OB_i)$	Individua gli evasori che siano tali per ciascuna delle funzioni obiettivo di primo livello	$OB_1\ and\ OB_2$
$AND_{EXT}(OB_i)$	Individua gli evasori per ciascuna delle funzioni obiettivo di primo livello o che abbiano dei valori di punta rispetto alle singole variabili su cui sono basate quelle di primo livello	$AND(OB_i)\ or$ $OB_1 > S'_1\ or$ $OB_2 > S'_2$

Tabella 5.7: composizione delle funzioni di primo livello

Anche nel caso della funzione $AND_{EXT}(OB_i)$, abbiamo usato delle soglie per rendere la funzione binaria, mostrate nella seguente tabella. Valgono anche per questi parametri le stesse considerazioni svolte in precedenza.

Parametro	Valore	Effetto sulla funzione obiettivo $AND_{EXT}(OB_i)$
S'_1	20.000	Sono considerati "evasori" tutti i soggetti la cui evasione supera i 20.000 euro
S'_2	0,025	Sono considerati "evasori" tutti i soggetti la cui evasione supera il 2,5% del volume d'affari reale

Tabella 5.8: valori soglia funzione AND estesa

5.2.2 Trasformazione di una funzione obiettivo da binaria a multi valore

Descriviamo ora una possibile metodologia per la trasformazione di una funzione obiettivo (sia di primo che di secondo livello) dalla sua versione binaria in una versione multi valore.

La definizione di una funzione obiettivo multi valore è interessante, in quanto dà la possibilità di catturare in modo più fine il grado di evasione di un soggetto e ciò può agevolare la costruzione di modelli previsionali più precisi in termini di accuratezza e ancor più efficaci nel recupero di gettito evaso.

Una tale trasformazione può avvenire tipicamente secondo due modalità:

- tramite l'impiego di più soglie per determinare il valore della funzione obiettivo (ad esempio, utilizzando una serie di valori S_{1i} per generare più valori della funzione OB_2 associati ciascuno ad un livello di "evasione");
- tramite la definizione di un operatore di *scoring* dei soggetti e la successiva discretizzazione dei valori (continui) dello *score*.

La modalità adottata in questo lavoro è la seconda e la funzione obiettivo oggetto della trasformazione è la $AND_{EXT}(OB_i)$.

Dato che la funzione che si intende trasformare tiene conto contemporaneamente dei due concetti principali su esposti di proficuità ed equità, un operatore di *scoring* coerente con tali concetti può essere definito come una funzione $Score\ (x)$ che associa a ciascun soggetto (*record*) x un punteggio pari a zero se si tratta di un soggetto "*non evasore interessante*" (in tal caso, il soggetto è *non evasore interessante* secondo la funzione obiettivo $AND_{EXT}(OB_i)$) e pari al prodotto pesato dei valori normalizzati di [PRETESA_TRIB_IMPST] e [PRETESA_SU_VA] altrimenti.

Più formalmente, la formulazione di un siffatto operatore è la seguente:

$$Score\ (x) = \begin{cases} 0, & se\ per\ x\ si\ ha\ \ AND_{EXT}(OB_i) = 0 \\ f(x)^{\alpha_f} * v(x)^{\alpha_v}, & altrimenti \end{cases}$$

dove

$$f(x) = PRETESA\_TRIB\_IMPST_{norm}(x) = \frac{PRETESA\_TRIB\_IMPST(x)}{max(PRETESA\_TRIB\_IMPST)}$$

$$v(x) = PRETESA\_SU\_VA_{norm}(x) = \frac{PRETESA\_SU\_VA(x)}{max(PRETESA\_SU\_VA)}$$

con $max(A)$ pari al massimo valore che l'attributo A assume nei soggetti del *dataset* e con coefficienti α_i positivi o nulli che rappresentano dei fattori di compressione dei singoli attributi dello *score* come segue:

$$\alpha_i = \begin{cases} 0\ se\ p_i = 0 \\ 1\ se\ p_i = 1 \\ 1 - p_i altrimenti \end{cases}$$

dove con p_i si rappresenta il peso che si intende assegnare a ciascun singolo contributo allo *score* e per i quali si ha $0 < p_f + p_v \leq 1$.

Considerati dei pesi uniformi $(p_i = 1/2)$, l'operatore $Score(x)$ così definito realizza una funzione obiettivo continua, che associa ad ogni soggetto un valore fra 0 e 1, nella quale ognuno dei due concetti fondamentali, *proficuità* ed *equità*, è considerato in maniera egualmente importante.

5.3 Estrazione dei modelli e valutazione dei risultati

Dopo aver definito, tramite la funzione $Score\ (x)$, la *classe* da predire, in questa sezione illustreremo la fase di estrazione di modelli predittivi a partire dal *dataset* costruito in base alle indicazioni descritte nella sezione precedente, con lo scopo di identificare, sulla base dei dati a disposizione, interessanti regolarità (*patterns*).

La costruzione di modelli di classificazione presuppone, come ricordato nel capitolo precedente, l'apprendimento di una funzione in grado di associare singoli dati ad una specifica *classe* appartenente ad un insieme di valori predefinito.

L'obbiettivo è stato quindi quello di realizzare un sistema di classificazione dei dati, da utilizzare per la predizione delle caratteristiche di "evasore" (*non interessante* / *interessante*) dei soggetti non ancora accertati e per i quali si hanno a disposizione i soli dati provenienti dalle dichiarazioni dei redditi presentate.

I modelli predittivi estratti in seguito sono ottenuti mediante due note tecniche di classificazione, ovvero *alberi di classificazione* e *regole di classificazione*.

L'estrazione dei modelli è avvenuta mediante utilizzo della *suite open source Weka*. In particolare, per assicurare uniformità a tutta l'analisi, è stata utilizzata l'interfaccia *knowledgeflow* che permette di fissare una volta per tutte il *training set* ed il *test set* consentendo ai vari "esperimenti" di essere in qualche modo "ripetibili" (a meno di componenti *random* (es. *resample*) sulle quali non si ha pieno controllo).

L'attributo da predire è una discretizzazione del campo *score* calcolato in precedenza secondo la funzione:

$$Score\ (x) = \begin{cases} 0, & se\ per\ x\ si\ ha\ \ AND_{EXT}(OB_i) = 0 \\ f(x)^{\alpha_f} * v(x)^{\alpha_v}, & altrimenti \end{cases}$$

Ad esempio, un possibile criterio di discretizzazione potrebbe essere quello di suddividere la popolazione in 4 classi con etichette "0", "1", "2", "3", che individuano rispettivamente i "non evasori", gli "evasori non gravi", gli "evasori gravi" e gli "evasori molto gravi" (dove la scala di gravità dipende dagli obiettivi dell'utilizzatore del modello). Ad esempio, se non si hanno particolari restrizioni sul numero di controlli da effettuare, si può banalmente supporre una equiripartizione delle classi all'interno del *dataset*.

Gli attributi a disposizione del classificatore sono *solo* quelli relativi alle dichiarazioni (*attributi predittivi*) e la classe (discretizzata) *score*; non interessano, al classificatore, tutti i campi relativi all'accertamento e tutti gli altri campi aggiunti che non sono osservabili su un soggetto da selezionare per una eventuale attività di controllo.

E' stato utilizzato, come *dataset* di analisi, quello composto da tutti i 1.835 soggetti accertati.

Si evidenzia che la metodologia su descritta di preparazione dei dati per le analisi di *mining* è di carattere generale e può applicare anche a *dataset* di grosse dimensioni.

Nel nostro caso, tuttavia, il *dataset* non contiene moltissimi record, per cui, in un primo momento, la classe è stata discretizzata in quattro valori, come nell'esempio su riportato. Successivamente, al fine di evitare una eccessiva parcellizzazione dei risultati, lo stesso attributo *classe* è stato binarizzato, con i soli valori "*interessante*"

(corrispondente al valore "molto grave" visto in precedenza) e "non interessante" (che raggruppa i tre valori "non evasori", "evasori non gravi" ed "evasori gravi"). In tal modo, sostanzialmente, un quarto dei soggetti del *dataset* sono stati ritenuti interessanti e tre quarti non interessanti. Vale la pena di ribadire che nulla vieta all'utilizzatore del modello di definire, nella pratica, in base alle proprie esigenze, risorse ed obiettivi, i criteri per ritenere un determinato soggetto "*interessante*" e quanto.

Nel seguito si presenteranno una serie di modelli ritenuti più significativi per evidenziare le potenzialità insite nell'uso di alberi e regole di classificazione nell'ambito del problema in esame di *tax fraud detection*. La trattazione non pretende di presentare tutti i possibili modelli implementabili sul *dataset* a disposizione, ma intende piuttosto descrivere le logiche che guidano il "settaggio" di determinati parametri, sia dei singoli classificatori che dei metaclassificatori forniti dalla *suite Weka* e le modalità per utilizzare nel modo più proficuo possibile l'*output* di detti modelli, fornendo quindi, in definitiva, una guida per l'impiego delle descritte tecniche di *mining* nell'ambito della pianificazione del controlli fiscali.

Del resto, le possibili combinazioni dei vari tecnicismi utilizzabili sono moltissime, potendo l'analista/utilizzatore operare scelte di *tuning* a livello di preparazione dei dati (*resample, feature selection ecc...*), di metaclassificatori (*bagging, boosting, randomforest, metacost*,…), di classificatori (alberi *J48, Cart, decision stump, ID3*, ecc.. e relativi parametri, regole ottenute con algoritmo *RIPPER, PART*, ecc…), di ripartizione delle istanze tra *training set* e *test set*, di criteri di valutazione dei modelli e, infine, di selezione dei soggetti da sottoporre a controllo, sulla base delle indicazioni fornite dai modelli stessi. Tutto ciò rende peraltro gli strumenti di *mining* estremamente flessibili ed adattabili alle esigenze di chi le dovesse impiegare.

Il *dataset* di analisi, a seguito della binarizzazione dell'attributo *classe* contiene 1.378 istanze relative a soggetti ritenuti "*non interessanti*" (pari al 75,1% della popolazione totale) e 457 istanze di soggetti "*interessanti*" (24,9% del totale), ovvero, circa un quarto della popolazione è da ritenersi meritevole di accertamento.

Nonostante i dati non siano molto abbondanti, si è proceduto nel partizionare una volta per tutte il *dataset* in *training* e *test set* (anziché operare con tecniche alternative, ad esempio *k-fold cross validation*), in modo da poter agevolmente confrontare sullo stesso *test set* i vari classificatori costruiti. Nello stabilire la ripartizione dei *records* in *training set* e *test set* è stato preso in considerazione il problema dell'*overfitting* dei modelli predittivi, che può emergere quando il *training set* è "troppo" grande (nel qual caso i dati tendono ad uniformarsi al *training set* perdendo la loro capacità di essere *generalizzabili* a dati diversi da quelli sulla cui base sono stati costruiti) o "troppo" piccolo (immaginando i singoli *record* descritti da *n* attributi come vettori di uno spazio n-dimensionale, la mancanza di dati in talune regioni di detto spazio ne rende difficoltosa la predizione

dell'etichetta di classe e può causare la generazione di regole o rami sulla base di dati irrilevanti). Nel caso specifico, tale pericolo è stato valutato eseguendo una serie di esperimenti nei quali sono stati utilizzati *training sets* dalle dimensioni crescenti, dal 10% al 90% dell'intero *dataset* di *input*. I risultati di tali esperimenti hanno portato ad escludere il rischio di *overfitting* (le *performances* dei classificatori usati restavano grosso modo costanti al variare della dimensione dei *training sets* di volta in volta provati). Quindi, data la cardinalità del *dataset* a disposizione, è stata scelta una proporzione 80:20 tra *training set* e *test set*. Attraverso il *filtro supervised Resample senza replacement* all'80% è stato ottenuto un *training set* con 1468 istanze distinte. Per differenza, è stato ottenuto il *test set* (367 istanze).

Il *test set* ottenuto presenta le seguenti caratteristiche:

- *# test set* = 367
- *# frodatori interessanti* = 86
- *# frodatori non_interessanti* = 281
- *recupero (test set)* = $\sum_{i \in test\ set} recupero(i)$ = € 5.516.901,00
- *recupero _medio=recupero (test set)/#test set* = € 15.032,00
- *recupero _mediana(test set)* = € 2.810,00
- *recupero (interessanti)* = $\sum_{i \in test\ set | classe=interessante} recupero(i)$ = € 4.707.616,00
- *recupero _media (interessanti)* = € 4.707.616,00/86 = € 54.740,00
- *recupero _mediana (interessanti)* = € 21.037,00
- *recupero (non_interessanti)* = $\sum_{i \in test\ set | classe=non\_interessante} recupero(i)$ = € 809.285,00
- *recupero _media (non_interessanti)* = € 809.285,00/281 = € 2.880,00
- *recupero _mediana (non_interessanti)* = € 1.536,00

Questi primi dati già evidenziano come una selezione accurata dei soggetti da sottoporre a controllo possa portare ad un notevole incremento della produttività del lavoro svolto dall'Agenzia delle Entrate. Infatti, i soggetti interessanti presenti sul *test set* rappresentano il 23,4% della popolazione (dato peraltro in linea con quello riscontrato sull'intero *dataset*), ma la pretesa erariale, e quindi, il potenziale recupero di gettito, nei loro confronti, rappresenta oltre l'85% del totale: è quindi evidente che avere a disposizione uno strumento che sia in grado di riconoscere tale tipologia di contribuenti tra tutti quelli che pure non sono in regola con il fisco per qualche motivo, non può che aiutare a migliorare le *performances* dell'attività dell'accertamento.

Tuttavia, osservando che il valore mediano dei recuperi nel *data set* è molto basso se confrontato con il corrispondente valore medio, si intuisce che vi devono essere alcuni soggetti eccezionalmente fraudolenti nel *dataset* stesso.

Poiché la differenza tra media e mediana del recupero nei soggetti *non interessanti* non è marcata, mentre nei soggetti *interessanti* detti due valori divergono in maniera considerevole, si intuisce che i

valori eccezionali appartengono alla classe *interessanti*. E' naturale a questo punto chiedersi quanti siano.

Presentiamo quindi due grafici che descrivono la distribuzione di frequenza del recupero da accertamento sul *test set*.

Il primo mostra la distribuzione di frequenza del valore delle pretese tributarie classica, riportante, sull'asse delle ordinate, la frequenza assoluta dei valori di pretesa tributaria discretizzati. Poiché il valore minimo della pretesa tributaria è 0 e quello massimo è € 1.268.993,00 e tenuto conto che la distribuzione è molto dispersa, i valori di pretesa tributaria sono stati discretizzati in classi di ampiezza crescente: la prima 100, la seconda $100 + 100 * 1{,}2$, la terza $100 + 100 * 1{,}2 + 100 * 1{,}2^2$ e così via.

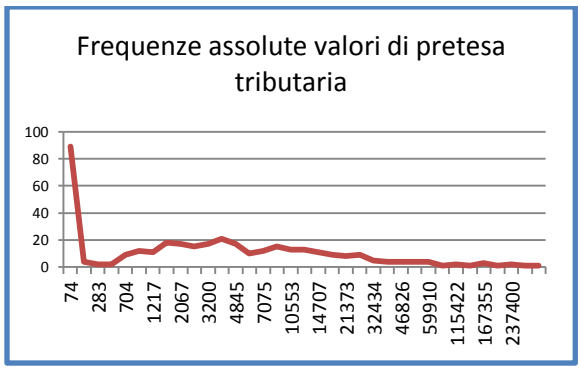

Figura 5.1: frequenze assolute pretesa tributaria

La distribuzione su riportata evidenzia, da un lato, come vi siano molti accertamenti con pretese erariali modeste (in particolare, la classe modale risulta essere quella con pretesa erariale da 0 a 100), dall'altro, che accertamenti con valori monetari più consistenti sono via via meno frequenti, fino ad arrivare alla classe estrema (contenente il valore € 1.2 milioni), molto distanziata dalle altre, con frequenza pari a 1.

Un modo alternativo di vedere la distribuzione dei dati del *test set* è quello di indicare sulle ordinate il valore della pretesa tributaria e sulle ascisse il *rank* (posizione nell'ordinamento) di ciascuna pretesa (discretizzata come sopra descritto), come nella figura che segue:

Figura 5.2: pretesa tributaria vs *ranking*

Anche da questo grafico emerge come vi sia un valore di pretesa erariale particolarmente elevato (il solito valore di € 1.2 milioni, che da solo vale il 23% della pretesa dell'intero campione di contribuenti appartenenti al *test set*) e molto più elevato rispetto al valore immediatamente inferiore.

Ci si chiede ora se tale valore estremo possa essere ragionevolmente classificato come *interessante* da un qualsiasi modello di classificazione.

Per rispondere, diviene importante capire se vi siano, nel *training set*, altri soggetti che danno luogo ad una pretesa erariale simile, analizzando i quali, i modelli che presenteremo in seguito, possano derivare una qualche loro caratteristica comune utile per poter riconoscere tale tipologia di soggetti su dati nuovi, non precedentemente osservati in fase di *training*.

Al fine di verificare la situazione sul *training set*, viene presentato, in figura 3, il grafico della pretesa erariale *vs rank* sull'intero *data set*:

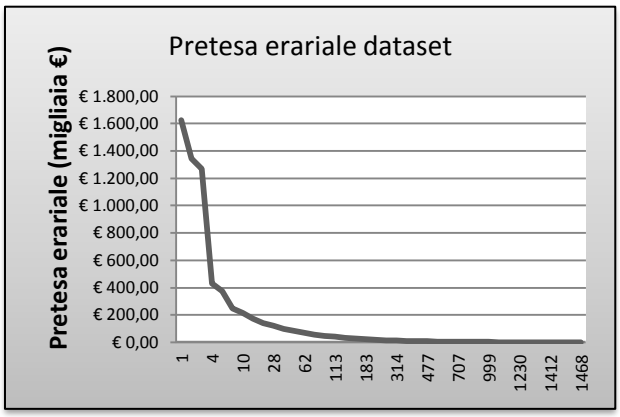

Figura 5.3:pretesa erariale vs *ranking* pretesa

Si nota che nell'intero *data set* appaiono esservi 3 soggetti in totale per i quali la pretesa derivante dal controllo effettuato supera il milione di euro: due sono presenti nel *training set* e l'altro, come visto, nel *test set*.

Analizzando gli attributi che caratterizzano tali soggetti, si osserva come essi presentino caratteristiche completamente diverse tra loro. I due soggetti del *training set*, ad esempio, sono residenti in due province diverse (diverse, a loro volta da quella del soggetto del *test set*), uno è professionista, l'altro imprenditore in contabilità semplificata (il soggetto del *test set* è in contabilità ordinaria), presentano volumi d'affare non paragonabili.

Il fatto quindi che l'*outlier* presente nel *test set* sia ritenuto o meno interessante da un certo modello (che non ha avuto modo, per i motivi visti, di comprenderne la "pericolosità" nella fase di *training*), non deve influenzarne la bontà nel suo complesso, perché essa non può dipendere dall'individuazione *casuale* di soggetti nei confronti dei quali siano stati emessi accertamenti particolarmente elevati (usiamo

il termine *casuale* proprio per indicare situazioni in cui tali soggetti evasori particolarmente "importanti" vengano individuati non perché nel *training set* vi fossero altri soggetti con le loro stesse caratteristiche né perché, se anche ve ne fossero stati, risultavano aver evaso a livelli molto alti, ma semplicemente perché "somiglianti" ad altri evasori meno gravi e quindi, il fatto di scoprire una c.d. "maxievasione" appare essere, appunto, un evento *casuale*).

Ne deriva che l'impiego di misure quali la "*media del recupero per controllo*" non appare significativo ai fini della valutazione dei modelli perché, essendo tale misura influenzata da valori *outlier,* penalizza i modelli che non li dovessero ritenere interessanti rispetto ad altri che, al contrario, li etichettassero come tali, stante la sostanziale casualità con cui un *outlier*, nel caso specifico, può essere classificato come *interessante* o *non interessante*. Ciò non toglie che un *dataset* più ampio possa consentire ad un dato modello di trattare anche gli *outlier* con maggior cura rispetto al *dataset* di analisi impiegato nel presente lavoro.

Ad ogni buon conto, nel proseguo, si riporterà sempre la media dei recuperi, ma verrà data importanza alla relativa mediana (indice robusto rispetto alla presenza di *outliers*) o ad altre metriche che saranno in seguito definite (area sotto la curva *lift chart)*.

Inoltre, presenteremo i risultati dei modelli calcolati sull'intero *test set*, ma anche quelli ottenuti senza considerare l'*outlier*.

Tutto ciò premesso, iniziamo l'analisi dei dati mediante l'applicazione di singoli classificatori per poi introdurre i metodi *ensemble*.

5.3.1 Alberi di classificazione

In questo ambito analizziamo i dati con due modelli presenti in *Weka*: *J48*, l'implementazione *java* del modello C4.5 sviluppato da R. Quinlan in [Qui93] e *SimpleCart*, trasposizione del modello CART sviluppato da Breiman in [BFOS84].

Nel corso della descrizione del primo, *J48*, vedremo che il *resample* della base dati di apprendimento gioca un ruolo cruciale nel determinare l'*output* del modello; descrivendo il secondo, *SimpleCart*, proporremo un semplice metodo per scegliere il livello ottimale di *resample* del *training test*.

5.3.1.1 Modello C4.5

L'algoritmo C4.5[3] produce alberi di decisione utilizzando una strategia *divide-et-impera* (ovvero seguendo un processo detto *partizionamento ricorsivo* dello spazio χ delle osservazioni mediante

[3] C4.5 appartiene ad una lunga serie di algoritmi di apprendimento che affonda le sue origini nei lavori di Hunt degli anni '50 e '60 [Hun62], passando per ID3 ideato sempre da Quinlan [Qui79] e C4 del medesimo autore [Qui87].

successive suddivisioni dello spazio stesso; la crescita degli alberi decisionali si fonda cioè sull'idea di esplorare le relazioni tra le variabili mediante suddivisione progressiva del campione iniziale in gruppi sempre più omogenei al loro interno rispetto alla variabile classe), impiegando quale criterio di *split*, il *gain ratio*, derivato dal concetto di entropia. C4.5 utilizza *test* di tre tipi (a seconda del tipo dell'attributo A di riferimento: se è discreto, il *test* è del tipo "$A =?$", con tanti *outcome* quanti i valori assunti da A, oppure "$A \in G?$" dove l'insieme G rappresenta una partizione dei valori di A; se A è continuo, il *test* è del tipo "$A \leq \theta$", con *outcome vero* o *falso*), ognuno dei quali coinvolge un solo attributo. Le regioni di decisione nello spazio delle istanze vengono così delimitate da iperpiani, ciascuno ortogonale ad uno degli assi di attributo. La strategia *divide-et-impera* partiziona i dati fino a quando ogni foglia contiene casi di una singola classe, o fino a quando un ulteriore partizionamento è reso impossibile perché due casi hanno gli stessi valori per ogni attributo, ma appartengono a classi differenti. Di conseguenza, se non ci sono casi confliggenti, l'albero decisionale classificherà correttamente tutti gli elementi del *training set*. Questo fenomeno, detto di *overfitting*, generalmente porta ad una perdita di potere predittivo da parte del modello nella maggior parte delle applicazioni e può essere evitato imponendo un criterio di arresto che impedisca ad alcuni sottoinsiemi del *training set* di essere ulteriormente suddivisi (*pre-pruning*), oppure rimuovendo, dall'albero finale, parte della struttura prodotta in fase di costruzione (*post-pruning*). C4.5 impiega un meccanismo di quest'ultimo tipo.

L'algoritmo viene eseguito con i seguenti parametri, manipolabili in *Weka*:

binarySplits -- Whether to use binary splits on nominal attributes when building the trees	*False*
confidenceFactor -- The confidence factor used for pruning (smaller values incur more pruning).	*0,25*
minNumObj -- The minimum number of instances per leaf.	*15*
numFolds -- Determines the amount of data used for reduced-error pruning. One fold is used for pruning, the rest for growing the tree.	*3*
reducedErrorPruning -- Whether reduced-error pruning is used instead of C.4.5 pruning.	*False*
saveInstanceData -- Whether to save the training data for visualization.	*False*
seed -- The seed used for randomizing the data when reduced-error pruning is used	*1*
subtreeRaising -- Whether to consider the subtree raising operation when pruning.	*True*
unpruned -- Whether pruning is performed.	*False*
useLaplace -- Whether counts at leaves are smoothed based on Laplace.	*False*

L'algoritmo in argomento, impiegato con i parametri sopra specificati, senza ulteriori manipolazioni del *training set*, produce un classificatore *abbastanza* prudente, che suggerisce di effettuare 85 controlli, come risulta dalla matrice di confusione di seguito riportata:

```
=== Confusion Matrix ===

  a   b   <-- classified as
 228  53 |   a = non_interessante
  54  32 |   b = interessante
```

Valutando il modello secondo le metriche *standard* (vedi par. 3.2.1), si può rilevare come esso sia anche *abbastanza accurato*, predicendo correttamente il 70% delle istanze del *test set*:

```
Correctly Classified Instances      260               70.8447 %
Incorrectly Classified Instances    107               29.1553 %
```

e che anche altre misure siano discrete:

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.811	0.628	0.809	0.811	0.81	0.662	Non_interessante
	0.372	0.189	0.376	0.372	0.374	0.662	Interessante
Weig. Avg.	0.708	0.525	0.707	0.708	0.708	0.662	

Come osservato in precedenza, il *training set*, per come è stato costruito il valore di classe, si presenta sbilanciato verso la classe *non interessante*. Nel caso in cui "positivo" venga interpretato come *interessante*, l'albero presenta *performances* peggiori, sia come precisione che come richiamo, che si attestano entrambe sull'ordine del 37%. Questo fenomeno può essere spiegato ipotizzando che, nel suo sforzo di essere preciso, il modello si lasci scappare molti *evasori interessanti* (54 su 86), che vengono classificati come *non interessanti* e allo stesso tempo non riesca comunque a distinguere bene i soggetti *interessanti* dai *non interessanti* (su 85 soggetti stimati come *interessanti*, solo 32 effettivamente lo sono).

La situazione descritta può essere dovuta all'esistenza di una fascia di contribuenti non chiaramente classificabile né come *interessante* né come *non interessante*. Tuttavia, poiché i soggetti interessanti effettivi sono meno numerosi dei non interessanti, questa area "grigia" di confine pesa molto di più, in termini percentuali, sui primi rispetto ai secondi. La cardinalità di tale area grigia potrebbe essere stimata attraverso il numero di errori commessi dal modello (falsi positivi e falsi negativi), pari a circa 100 unità. E' chiaro che una condizione di incertezza su circa 100 contribuenti incide relativamente poco sulla popolazione dei *non interessanti*, che in totale conta circa 280 unità, mentre la stessa incertezza pesa assai di più sulla popolazione dei soggetti *interessanti*, che conta, complessivamente, meno di 100 soggetti.

L'albero che origina dall'applicazione dell'algoritmo al *training set* è di seguito riportato:

```
REDD_LORDO_2007 <= 55444
|  IMP_VE_VOLAFF_2007 <= 28941
|  |  IMP_VE_VOLAFF_2007 <= 0: non_interessante (28.0)
|  |  IMP_VE_VOLAFF_2007 > 0
|  |  |  IMP_VE_VOLAFF_2007 <= 3892: interessante (22.0/2.0)
|  |  |  IMP_VE_VOLAFF_2007 > 3892
|  |  |  |  FLG_PRES_DIP_07 <= 0
|  |  |  |  |  ACQ_NO_DETR <= 0
|  |  |  |  |  |  BENI_DEST_RIV <= 2668
|  |  |  |  |  |  |  FLG_NO_CONGRUENZA_07 <= 0
|  |  |  |  |  |  |  |  SESSO = M: interessante (25.0/8.0)
|  |  |  |  |  |  |  |  SESSO = F: non_interessante (20.0/6.0)
|  |  |  |  |  |  |  FLG_NO_CONGRUENZA_07 > 0: interessante (45.0/8.0)
|  |  |  |  |  |  BENI_DEST_RIV > 2668: non_interessante (17.0/4.0)
|  |  |  |  |  ACQ_NO_DETR > 0
|  |  |  |  |  |  IMP_BEN_AMM <= 2402
|  |  |  |  |  |  |  TOT_IVA_OPE_IMPO <= 2760: non_interessante (37.0/1.0)
|  |  |  |  |  |  |  TOT_IVA_OPE_IMPO > 2760
|  |  |  |  |  |  |  |  IMP_VE_VOLAFF_2007 <= 21291: interessante (15.0/6.0)
|  |  |  |  |  |  |  |  IMP_VE_VOLAFF_2007 > 21291: non_interessante (17.0/3.0)
|  |  |  |  |  |  IMP_BEN_AMM > 2402: interessante (16.0/5.0)
|  |  |  |  FLG_PRES_DIP_07 > 0: interessante (21.0/5.0)
|  IMP_VE_VOLAFF_2007 > 28941
|  |  IMP_SPS_DIPEND_SMP <= 1404
|  |  |  IMP_V_AGG_IVA <= 188241
|  |  |  |  PROF_COSTI <= 29306
|  |  |  |  |  RIM_FIN_ORD <= 146730
|  |  |  |  |  |  CRED_ANNO_PREC <= 3345: non_interessante (457.0/57.0)
|  |  |  |  |  |  CRED_ANNO_PREC > 3345
|  |  |  |  |  |  |  RIM_FIN_SMPL <= 1820
|  |  |  |  |  |  |  |  IMP_V_AGG_IVA <= 29447: non_interessante (25.0)
|  |  |  |  |  |  |  |  IMP_V_AGG_IVA > 29447
|  |  |  |  |  |  |  |  |  ETA <= 48: interessante (18.0/6.0)
|  |  |  |  |  |  |  |  |  ETA > 48: non_interessante (22.0/3.0)
|  |  |  |  |  |  |  RIM_FIN_SMPL > 1820: interessante (22.0/7.0)
|  |  |  |  |  RIM_FIN_ORD > 146730
|  |  |  |  |  |  CRED_CLI_ORD <= 5544: interessante (17.0/3.0)
|  |  |  |  |  |  CRED_CLI_ORD > 5544: non_interessante (16.0/2.0)
|  |  |  |  PROF_COSTI > 29306: interessante (26.0/12.0)
|  |  |  IMP_V_AGG_IVA > 188241
|  |  |  |  REDD_IMP_2007 <= 23365: non_interessante (15.0/6.0)
|  |  |  |  REDD_IMP_2007 > 23365: interessante (15.0/2.0)
|  |  IMP_SPS_DIPEND_SMP > 1404
|  |  |  RIM_FIN_SMPL <= 175539
|  |  |  |  TOT_IMPST_CRED <= 1786
|  |  |  |  |  TOT_IMPST_CRED <= 1297
|  |  |  |  |  |  ALTRI_ACQ_IMP <= 82365
|  |  |  |  |  |  |  IMP_BEN_STRUM_NA <= 307
|  |  |  |  |  |  |  |  ALIQ_MEDIA_ACQ <= 17.63: non_interessante (34.0/2.0)
|  |  |  |  |  |  |  |  ALIQ_MEDIA_ACQ > 17.63
|  |  |  |  |  |  |  |  |  DURATA_ATTIV_SOGG <= 8: interessante (21.0/4.0)
|  |  |  |  |  |  |  |  |  DURATA_ATTIV_SOGG > 8
|  |  |  |  |  |  |  |  |  |  ALIQ_MEDIA_CESS <= 19.92: interessante (17.0/4.0)
|  |  |  |  |  |  |  |  |  |  ALIQ_MEDIA_CESS > 19.92: non_interessante (26.0/2.0)
|  |  |  |  |  |  |  IMP_BEN_STRUM_NA > 307: interessante (35.0/8.0)
|  |  |  |  |  |  ALTRI_ACQ_IMP > 82365: interessante (18.0/1.0)
|  |  |  |  |  TOT_IMPST_CRED > 1297: interessante (16.0)
|  |  |  |  TOT_IMPST_CRED > 1786
|  |  |  |  |  REDD_IMP_2007 <= 8098: interessante (26.0/11.0)
|  |  |  |  |  REDD_IMP_2007 > 8098: non_interessante (52.0/4.0)
|  |  |  RIM_FIN_SMPL > 175539: non_interessante (16.0)
REDD_LORDO_2007 > 55444: non_interessante (311.0/25.0)
```

Figura 5.4: sviluppo dell'albero, modello C4.5

La struttura dell'albero ottenuta si presta immediatamente ad alcune considerazioni:

i. nell'albero possono essere presenti nodi etichettati con attributi presi da tutti i quadri (IVA, RE, RG, RF): i *pattern* che ne derivano possono quindi essere "sporchi", perché possono fare riferimento ad attributi che sono del tutto estranei ad un determinato soggetto, soprattutto nei casi in cui nodi consecutivi che portano alla stessa foglia sono del tipo:

$$Attr_x \in A = \{xor(RE, RG, RF)\} < x$$
$$Attr_y \notin A > y: \{xor(interessante, non\_interessante)\}$$

nel qual caso la prima condizione può essere soddisfatta solo per $x = 0$ e quindi tra gli attributi del soggetto appartenente alla foglia risultante, in realtà l'attributo x non è significativo. La seguente porzione di albero presenta tale problema (si mescolano dati delle imprese in contabilità ordinaria ([RIM_FIN_ORD]) semplificata [RIM_FIN_SMPL] e dei professionisti [PROF_COSTI]):

```
| | | | PROF_COSTI <= 29306
| | | | | RIM_FIN_ORD <= 146730
| | | | | | CRED_ANNO_PREC <= 3345: non_interessante (457.0/57.0)
| | | | | | CRED_ANNO_PREC > 3345
| | | | | | | RIM_FIN_SMPL <= 1820
| | | | | | | | IMP_V_AGG_IVA <= 29447: non_interessante (25.0)
| | | | | | | | IMP_V_AGG_IVA > 29447
| | | | | | | | | ETA <= 48: interessante (18.0/6.0)
| | | | | | | | | ETA > 48: non_interessante (22.0/3.0)
| | | | | | | RIM_FIN_SMPL > 1820: interessante (22.0/7.0)
| | | | | RIM_FIN_ORD > 146730
| | | | | | CRED_CLI_ORD <= 5544: interessante (17.0/3.0)
| | | | | | CRED_CLI_ORD > 5544: non_interessante (16.0/2.0)
| | | | PROF_COSTI > 29306: interessante (26.0/12.0)
```

Figura 5.5: particolare dell'albero C4.5

Ne deriva che un'analisi che prenda in considerazione solo i contribuenti appartenenti ad una specifica tipologia tra lavoratori autonomi (che compilano il quadro RE, ma non i quadri RF ed RG), imprese in contabilità semplificata (che compilano solo il quadro RG) e imprese in contabilità ordinaria (che compilano solo il quadro RF) potrebbe portare ad una migliore caratterizzazione dei soggetti, non potendosi presentare situazioni come quella appena descritta[4];

ii. il fatto che nella sequenza dei nodi *non* vi siano le province, suggerirebbe che il grado di interesse di un soggetto non sia influenzato da situazioni dipendenti dal territorio. Pertanto, i profili scoperti potrebbero essere validi in tutta la Toscana, che risulterebbe, da questo punto di vista, un'area geografica omogenea;

[4] Avendo a disposizione un dataset non troppo numeroso, tale prova non è stata effettuata in questo lavoro in quanto i *datasets* risultanti avrebbero avuto cardinalità effettivamente troppo ridotta per ottenere risultati significativi.

iii. volendo tracciare un profilo dell'evasore *interessante*, esso potrebbe essere caratterizzato dalla presenza di redditi relativamente bassi (visto che la radice [REDD_LORDO_2007] produce già una foglia etichettata *non interessante* se >55.444), di un basso volume d'affari, inferiore a € 28.941,00 e di non congruità allo studio di settore; per volumi d'affare superiori a € 28.941,00 vi sono diversi *pattern* che portano a soggetti interessanti, che tuttavia per essere tali devono soddisfare più condizioni

Una possibile tecnica utilizzabile per migliorare la *performance* del classificatore, ovviando al problema delle classi sbilanciate, prevede l'applicazione di un filtro *resample* sul *training set*. In particolare, la configurazione che segue fa sì che le istanze del *training set* vengono manipolate in modo tale da garantire una equa distribuzione delle classi. Lo scopo è di fornire all'algoritmo di apprendimento un numero maggiore di esempi della classe minoritaria, nello spirito e per i motivi indicati nella sezione 3.2.2. del capitolo precedente.

biasToUniformClass -- Whether to use bias towards a uniform class. A value of 0 leaves the class distribution as-is, a value of 1 ensures the class distribution is uniform in the output data.	1
invertSelection -- Inverts the selection (only if instances are drawn WITHOUT replacement).	False
noReplacement -- Disables the replacement of instances.	False
randomSeed -- Sets the random number seed for subsampling.	1
sampleSizePercent -- The subsample size as a percentage of the original set.	100

I parametri del modello restano quelli indicati in precedenza.

I risultati ottenuti sul *test set* sono però alquanto deludenti:

```
Correctly Classified Instances         194               52.861 %
Incorrectly Classified Instances       173               47.139 %
```

Detti valori appaiono infatti peggiori rispetto a quelli del classificatore ottenuto senza manipolazione delle istanze del *training set*.

Andando a valutare il modello secondo la *matrice di confusione*, ci si rende conto che lo stesso è molto meno selettivo del caso precedente, suggerendo di effettuare ben 183 controlli, di cui però solo 48 si rivelerebbero bene indirizzati, come si può vedere dalla matrice sotto riportata:

```
=== Confusion Matrix ===
   A         B      ← classified as
 146       135     |   a = non_interessante
  38        48     |   b = interessante
```

L'applicazione del filtro *resample* è responsabile dell'eccessivo numero di controlli proposto dal modello. Passando da un valore di *biasToUniformClass* 0 (*training set* originario) a 1 (*training set* con egual numero di casi positivi e negativi), il modello sembra aver "*overshot*": ciò suggerisce di rimodulare il valore di tale parametro. Empiricamente, un settaggio che modifica leggermente il *training set* si dimostra proficuo ai fini del miglioramento della qualità del modello stesso.

Difatti, ponendo il parametro *biasToUniformClass* a 0.1, si ottiene il seguente albero:

```
REDD_LORDO_2007 <= 56492
|  IMP_REDD_LRD_ORD <= -2340: non_interessante (34.0/1.0)
|  IMP_REDD_LRD_ORD > -2340
|  |  REDD_IMP_2007 <= 934
|  |  |  COSTO_LAV_2007 <= 57730
|  |  |  |  GRP_ATTIV_2007 = F: interessante (5.0)
|  |  |  |  GRP_ATTIV_2007 = Q: interessante (0.0)
|  |  |  |  GRP_ATTIV_2007 = I: non_interessante (9.0/3.0)
|  |  |  |  GRP_ATTIV_2007 = G
|  |  |  |  |  IMP_V_AGG_IVA <= -3876: interessante (16.0/2.0)
|  |  |  |  |  IMP_V_AGG_IVA > -3876: non_interessante (19.0/5.0)
|  |  |  |  GRP_ATTIV_2007 = L: interessante (4.0)
|  |  |  |  GRP_ATTIV_2007 = H: interessante (2.0)
|  |  |  |  GRP_ATTIV_2007 = A: interessante (16.0/5.0)
|  |  |  |  GRP_ATTIV_2007 = S: interessante (11.0/4.0)
|  |  |  |  GRP_ATTIV_2007 = M: interessante (3.0)
|  |  |  |  GRP_ATTIV_2007 = C: non_interessante (2.0/1.0)
|  |  |  |  GRP_ATTIV_2007 = K: interessante (2.0)
|  |  |  |  GRP_ATTIV_2007 = N: interessante (0.0)
|  |  |  |  GRP_ATTIV_2007 = R: interessante (0.0)
|  |  |  |  GRP_ATTIV_2007 = J: interessante (0.0)
|  |  |  |  GRP_ATTIV_2007 = P: interessante (0.0)
|  |  |  |  GRP_ATTIV_2007 = E: interessante (0.0)
|  |  |  |  GRP_ATTIV_2007 = B: interessante (0.0)
|  |  |  COSTO_LAV_2007 > 57730: non_interessante (15.0/1.0)
|  |  REDD_IMP_2007 > 934
|  |  |  GRP_ATTIV_2007 = F
|  |  |  |  DEB_FORN_ORD <= 118897
|  |  |  |  |  QTA_PART_IVA <= 1: non_interessante (75.0/14.0)
|  |  |  |  |  QTA_PART_IVA > 1
|  |  |  |  |  |  TOT_PASS_2007 <= 81178: non_interessante (20.0/4.0)
|  |  |  |  |  |  TOT_PASS_2007 > 81178: interessante (26.0/9.0)
|  |  |  |  DEB_FORN_ORD > 118897: interessante (17.0/4.0)
|  |  |  GRP_ATTIV_2007 = Q: non_interessante (24.0/9.0)
|  |  |  GRP_ATTIV_2007 = I

|  |  |  |  IMPST_CRED <= 1638: non_interessante (51.0/6.0)
|  |  |  |  IMPST_CRED > 1638: interessante (20.0/9.0)
|  |  |  GRP_ATTIV_2007 = G
|  |  |  |  ALTRI_ACQ_IMP <= 104653
|  |  |  |  |  REDD_LORDO_2007 <= 1195: interessante (19.0/6.0)
|  |  |  |  |  REDD_LORDO_2007 > 1195: non_interessante (252.0/42.0)
|  |  |  |  ALTRI_ACQ_IMP > 104653: interessante (15.0/4.0)
|  |  |  GRP_ATTIV_2007 = L
|  |  |  |  RICAVI_ATT_2007 <= 79759
|  |  |  |  |  IMP_REDD_IMP_SMPL_2007 <= 15333: non_interessante (16.0/4.0)
|  |  |  |  |  IMP_REDD_IMP_SMPL_2007 > 15333: interessante (16.0/4.0)
|  |  |  |  RICAVI_ATT_2007 > 79759: non_interessante (17.0)
|  |  |  GRP_ATTIV_2007 = H: non_interessante (47.0/12.0)
|  |  |  GRP_ATTIV_2007 = A: non_interessante (41.0/11.0)
|  |  |  GRP_ATTIV_2007 = S: non_interessante (24.0/2.0)
|  |  |  GRP_ATTIV_2007 = M
|  |  |  |  ETA <= 53
|  |  |  |  |  QTA_PART_IVA <= 1: non_interessante (79.0/9.0)
```

```
| | | | | QTA_PART_IVA > 1
| | | | | | PROF_COMP <= 36520: non_interessante (17.0/1.0)
| | | | | | PROF_COMP > 36520: interessante (16.0/5.0)
| | | | ETA > 53
| | | | | REDD_LORDO_2007 <= 30636: interessante (29.0/5.0)
| | | | | REDD_LORDO_2007 > 30636: non_interessante (18.0/4.0)
| | | GRP_ATTIV_2007 = C
| | | | ALIQ_MEDIA_CESS <= 17.38
| | | | | ALIQ_MEDIA_ACQ <= 13.48: non_interessante (22.0)
| | | | | ALIQ_MEDIA_ACQ > 13.48
| | | | | | ACQ_NO_DETR <= 679: interessante (15.0/4.0)
| | | | | | ACQ_NO_DETR > 679: non_interessante (15.0)
| | | | ALIQ_MEDIA_CESS > 17.38
| | | | | TOT_PASS_2007 <= 23369: interessante (25.0)
| | | | | TOT_PASS_2007 > 23369
| | | | | | IMP_TOT_COMP_NEG_2007 <= 135228: non_interessante (29.0/11.0)
| | | | | | IMP_TOT_COMP_NEG_2007 > 135228: interessante (19.0)
| | | GRP_ATTIV_2007 = K: non_interessante (9.0)
| | | GRP_ATTIV_2007 = N: non_interessante (15.0/3.0)
| | | GRP_ATTIV_2007 = R: non_interessante (17.0/8.0)
| | | GRP_ATTIV_2007 = J: non_interessante (7.0/3.0)
| | | GRP_ATTIV_2007 = P: interessante (6.0/2.0)
| | | GRP_ATTIV_2007 = E: non_interessante (0.0)
| | | GRP_ATTIV_2007 = B: non_interessante (2.0)
REDD_LORDO_2007 > 56492: non_interessante (310.0/32.0)
```

Figura 5.6: Sviluppo albero C4.5 – *resample* 0.1

Si può osservare come questa volta i risultati che il modello ottiene sul *test set* siano ben migliori del precedente:

```
Correctly Classified Instances         261         71.1172 %
Incorrectly Classified Instances       106         28.8828 %
```

Andando a valutare il modello secondo altre metriche di valutazione, si osserva la seguente matrice di confusione:

```
=== Confusion Matrix ===
   a         B      ← classified as
 240        41      |   a = non_interessante
  65        21      |   b = interessante
```

Il modello suggerisce quindi di effettuare 62 controlli (circa un terzo del modello precedente), di cui 21 (un terzo del totale dei controlli suggeriti) rivolti nei confronti di soggetti realmente interessanti.

Altre metriche sono di seguito riportate (tutte nettamente migliori rispetto al caso precedente):

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.854	0.756	0.787	0.858	0.819	0.61	Non_interessante
	0.244	0.146	0.339	0.244	0.284	0.61	Interessante
Weig. Avg.	0.711	0.613	0.682	0.711	0.694	0.61	

Andando ad osservare la struttura dell'albero derivato dall'algoritmo C4.5 previo *resample* del *training set* con parametro *biasToUniformClass* pari a 0.1, si nota come emerga quello che può essere considerato uno svantaggio di questo tipo di classificatori, ben noto in letteratura: ovvero la loro *instabilità*. Ciò vuol dire che leggere

variazioni del *training set* sono capaci di modificare pesantemente l'*output* del modello, come si può vedere dalle Figure 5.4 (no *resample* del *training set*) e 5.6 (leggero *resampling* del *training set*). Ad esempio, uno dei percorsi che partono dalla radice del primo albero è il seguente:

```
REDD_LORDO_2007 <= 55444
|  IMP_VE_VOLAFF_2007 <= 28941
|  |  IMP_VE_VOLAFF_2007 <= 0: non_interessante (28.0)
```

mentre il secondo prevede la seguente situazione:

```
REDD_LORDO_2007 <= 56492
|  IMP_REDD_LRD_ORD <= -2340: non_interessante (34.0/1.0)
```

Come visto nel capitolo precedente, tale caratteristica degli alberi decisionali li rende però molto adatti ad essere impiegati in *ensemble methods*.

Non solo, ma nell'albero di Figura 5.6 gioca un ruolo determinante l'attributo relativo all'*attività svolta*, che invece era completamente assente nell'albero rappresentato in Figura 5.4. Ciò suggerisce che *datasets* specifici per attività possano anch'essi essere utili ai fini dell'individuazione di caratteristiche peculiari dei vari settori economici, importanti ai fini della individuazione di comportamenti fraudolenti.

Si evidenzia ancora come gli attributi sui singoli campi dei quadri presentati vengono raramente contemplati, in quanto gli algoritmi portano all'effettuazione di *split* su quegli attributi per i quali molte istanze presentano valori diversi da zero e possibilmente diversi tra loro. Ciò implica che saranno soprattutto i campi IVA e quelli riepilogativi di costi, ricavi e redditi ad essere presenti, mentre i campi specifici dei singoli quadri RE, RF ed RG vengono per lo più ignorati. Questo però farà sì che le specificità delle varie tipologie di contribuenti (professionisti e imprenditori) si perdano e non possono essere adeguatamente sfruttate. Anche per questo motivo, *datasets* specifici per tipo di contribuente non possono che portare a miglioramenti nella qualità di modelli che li dovessero analizzare.

Infine, si riportano le seguenti misure di carattere "monetario", in quanto la bontà di un modello non può prescindere dal gettito che permette di recuperare. Il modello generato ritiene che il soggetto *outlier* indicato in precedenza sia un "*frodatore interessante*". Sono perciò riportati i dati del modello con e senza l'*outlier*:

recupero	**€ 2.310.992,00** (su 62)	**€ 1.041.999,00** (su 61)
recupero (interessanti)	**€ 2.118.216,00** (su 21)	**€ 919.223,00** (su 20)
recupero (non interessanti)	**€ 122.776,00** (su 41)	**€ 122.776,00** (su 41)
media recupero	**€ 37.274,00** (su 62)	**€ 17.081,00** (su 61)
mediana recupero	**€ 4.006,00** (su 62)	**€ 3.474,00** (su 61)

Non va tuttavia dimenticato che il modello si è allenato in un *dataset* in cui un soggetto era ritenuto interessante (e quindi meritevole di controllo) se aveva evaso oltre una certa soglia o se aveva evaso in maniera considerevole in relazione al proprio volume d'affari – il grado di interesse di un soggetto è infatti dato dal valore della funzione $Score(x)$. Pertanto, i modelli elaborati non necessariamente massimizzano il recupero della pretesa erariale, dovendo soddisfare i due obiettivi di *proficuità* ed *equità*. Peraltro, gli effetti del parametro equità non sono immediatamente misurabili in termini monetari, in quanto tale parametro è studiato anche per garantire all'attività di accertamento un certo effetto deterrenza.

Tanto premesso, il modello individua 21 (20) contribuenti *interessanti* sui 62 (61) segnalati ma le *performance* delle verifiche suggerite appaiono tuttavia buone: sia nel caso in cui si consideri l'*outlier*, sia nel caso in cui non lo si consideri, il recupero di gettito è più che proporzionale rispetto ai controlli eseguiti: suggerendo di controllare 62 (61) soggetti, il modello si concentra su circa il 16% della popolazione del *test set*, consentendo allo stesso tempo un recupero del 38% del gettito evaso nel primo caso, e il 24% nel secondo.

Un'analisi condotta attraverso il *lift chart* del modello permette di comprendere meglio la capacità di recupero di gettito dello stesso. Nel caso in cui si consideri l'*outlier*, tale grafico, di seguito riportato, presenta, sulle ascisse, le varie istanze ordinate secondo la probabilità che il modello attribuisce loro di essere *interessanti* e sulle ordinate, la cumulata del gettito recuperato grazie al modello stesso[5]. Chiaramente, in corrispondenza dell'ultima unità del *test set* la cumulata vale il 100% del gettito. Fino alla 62^ unità, la curva mostra il gettito recuperato con le unità che il modello reputa meritevoli di controllo, mentre le successive sono invece classificate come *non interessanti* e pertanto non sarebbero selezionate per eventuali controlli. In corrispondenza della 60^ unità il grafico mostri una forte discontinuità: essa è provocata dal contribuente *outlier* che il modello ha intercettato e correttamente classificato come *interessante*:

[5] A parità di probabilità, i record sono ordinati in base al recupero, in ordine decrescente.

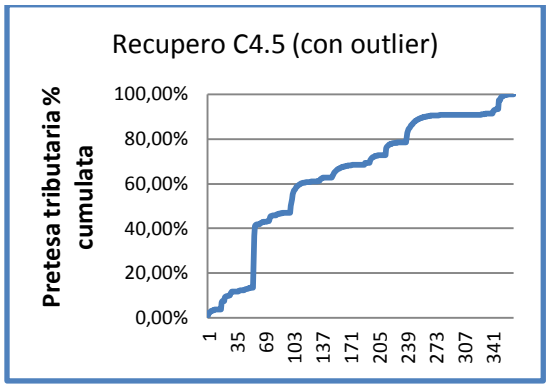

Figura 5.7: *lift chart* modello J48-resample 0.1

Può essere interessante confrontare il *lift chart* di figura 4.7 con quello ottenuto senza considerare l'*outlier* di cui sopra:

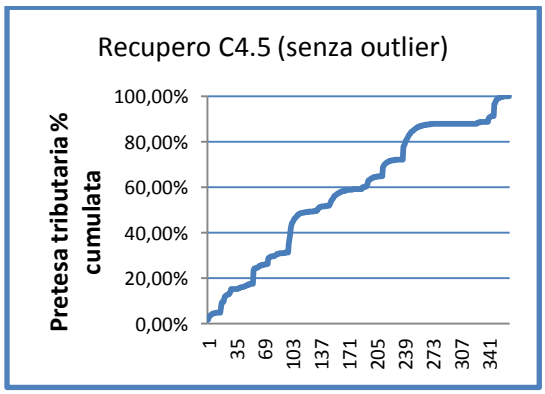

Figura 5.8: *lift chart* senza *outlier*

E' facile osservare come il grafico che se ne ottiene sia molto più regolare.

L'introduzione del *lift chart* fornisce all'analista un ulteriore elemento di flessibilità: nonostante il modello indichi 62 (61) soggetti meritevoli di controllo, è tuttavia possibile, sulla base delle proprie esigenze, essere più o meno "severi" (o, se vogliamo, "prudenti") del modello e procedere, di conseguenza, ad un maggiore o minore numero di controlli (nel primo caso, ciò significa effettuare controlli anche a soggetti con un livello di confidenza inferiore di 0.5, nel secondo, ciò significa esaminare solo soggetti con un livello di confidenza superiore a una certa soglia, comunque maggiore di 0.5). In entrambi i casi, il *lift chart* dà un'utile indicazione circa i possibili recuperi, fornendo una visione globale, sull'intero *test set,* dell'andamento degli stessi. Ad esempio, se la curva dovesse essere molto ripida all'inizio per poi proseguire con pendenza meno marcata, potrebbe essere accettabile effettuare anche meno controlli di quelli suggeriti dal modello.

Il *lift chart*, o meglio, *l'area sotto la curva*, è una nota misura di valutazione di un modello predittivo. Tale area può essere calcolata in valore assoluto o come rapporto rispetto all'area massima ottenibile da un classificatore ottimo "perfetto". Si possono allora confrontare,

ad esempio, i due *lift chart* precedenti con quelli di un ipotetico modello "*indovino*", che seleziona per primi i soggetti dal cui controllo deriva la maggiore pretesa tributaria e via via tutti gli altri.

Tale confronto può essere preso a base della valutazione complessiva dei modelli, tenendo comunque presente, come in precedenza già accennato, che gli stessi sono allenati secondo criteri di *proficuità* ed *equità* e pertanto sono disposti a "rinunciare" a una parte del gettito e a selezionare soggetti meno proficui in ossequio al criterio di equità.

Di seguito si riportano i *lift charts* del modello ottenuto, nei due casi, con e senza *outlier*.

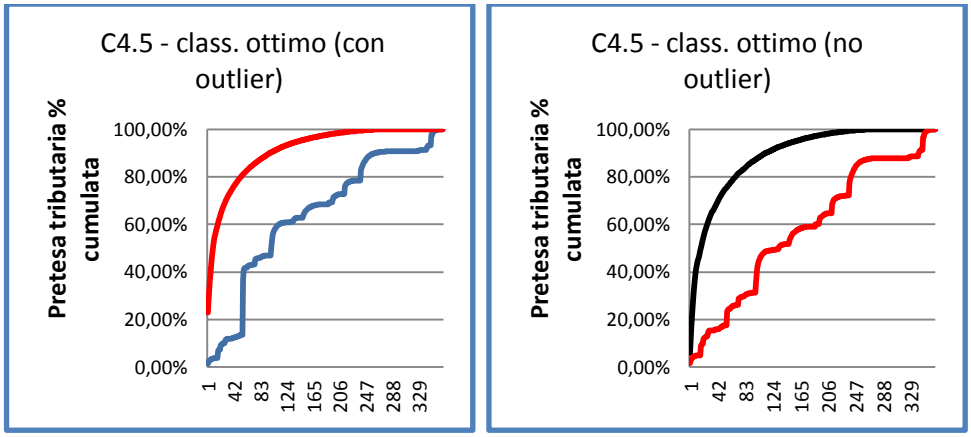

Figura 5.9: confronto C4.5 – classificatore ottimo

Oltre ad un'analisi per così dire "visiva" del *lift chart* estrapolato sulla base delle predizioni del modello C4.5, è possibile calcolarne *l'area sotto la curva (AUC)* e confrontarla con quella massima possibile. Se si normalizza a 1 l'area del classificatore ottimo (nel senso indicato in precedenza), l'area sotto la curva del modello C4.5 su riportata viene ad essere pari a 0.7041 nel caso in cui si consideri l'*outlier* e 0.6546 senza. Naturalmente, maggiore l'area sotto la curva, migliore il modello e tale misura viene ad essere presa in considerazione ai fini della valutazione della bontà dello stesso.

L'idea che i *lift chart* associati ai modelli possano essere utilizzati come guida flessibile per la pianificazione fiscale, coniugando, da una parte, le esigenze ed obiettivi dell'utilizzatore e dall'altro l'*output* di buoni modelli di classificazione costituisce una novità rispetto ai lavori presenti in letteratura e citati nel capitolo precedente.

5.3.1.2 Modello CART

Rimaniamo nell'ambito dei classificatori semplici, per vedere se altri algoritmi producono risultati paragonabili a C4.5. Utilizziamo quindi un altro algoritmo fornito dalla *suite Weka, SimpleCart*, derivato

dal ben noto algoritmo CART, descritto in [BFOS84], previo *resample* del *training set*.

Anche l'algoritmo CART utilizza una strategia *divide-et-impera*, ma si differenzia dal C4.5 per quanto riguarda la struttura dell'albero (sono previsti solo *split* binari), i criteri di *split* (l'algoritmo utilizza il *Gini index* anziché misure derivate dal concetto di entropia), il metodo di *pruning* (chiamato *minimal cost complexity pruning*) ed il modo in cui sono trattati i valori mancanti. Il modello viene qui utilizzato con i seguenti parametri:

heuristic -- If heuristic search is used for binary split for nominal attributes in multi-class problems (default yes).	*True*
minNumObj -- The minimal number of observations at the terminal nodes (default 2).	*15*
numFoldsPruning -- The number of folds in the internal cross-validation (default 5).	*5*
seed -- The random number seed to be used.	*1*
sizePer -- The percentage of the training set size (0-1, 0 not included).	*1*
useOneSE -- Use the 1SE rule to make pruning decision.	*False*
usePrune -- Use minimal cost-complexity pruning (default yes).	*True*

Anche in questo caso, il *resample* del *training set* gioca un ruolo fondamentale per i risultati proposti dall'algoritmo.

Vale allora la pena andare ad indagare, dati alla mano, in che modo cambia, al variare del parametro *biasToUniformClass,* il numero di controlli suggeriti dal modello e in che modo si ripartiscano tra loro i soggetti *interessanti* e *non interessanti* nell'ambito dell'insieme dei soggetti selezionati per un controllo.

Si ottiene il seguente risultato:

Resample	Non interessanti (A)	Interessanti (B)	Rapporto C=B/(A+B)	Precisione complessiva
0	26	14	35,00%	73,297%
0,1	33	9	21,43%	70,027%
0,2	46	19	29,23%	69,209%
0,3	52	32	38,10%	71,117
0,4	40	21	34,43%	70,389
0,5	73	29	28,43%	64,577
0,6	83	34	29,06%	63,215
0,7	141	51	26,56%	52, 043
0,8	82	37	31,09%	64,305
0,9	116	36	23,68%	55,041
1	159	48	23,19%	46,321

Tabella 5.9: *resampling* del training set e controlli suggeriti dal modello CART

Pur notando che al crescere del fattore *biasToUniformClass* aumenta mediamente il numero di controlli proposto dal modello (come del resto ci si poteva aspettare), tuttavia la ripartizione dei controlli "buoni" rispetto a quelli scarsamente proficui non sembra seguire un andamento lineare, né monotono, ma sembra essere, piuttosto, *random*.

Pertanto, empiricamente, adottando il criterio di massimizzazione del rapporto tra soggetti *interessanti* e totale dei soggetti selezionati per il controllo, si ritiene che il miglior classificatore si ottenga con un *resample* che utilizzi il valore di *biasToUniformClass* pari a 0,3.

Con detto settaggio, i risultati che il modello ottiene sul *test set* sono i seguenti:

```
Correctly Classified Instances         261               71.117 %
Incorrectly Classified Instances       106               28.883 %
```

I risultati sopra riportati sono sorprendentemente identici a quelli del modello visto in precedenza. Andando a valutare il modello secondo altre metriche, si scopre però come esso sia meno "prudente" nel selezionare i soggetti da accertare rispetto a C4.5, in quanto suggerisce di controllare 84 soggetti su 367, probabilmente a causa del diverso valore del parametro *biasToUniformClass*. La relativa matrice di confusione viene di seguito riportata:

```
=== Confusion Matrix ===

   a   b   <-- classified as
 229  52  |   a = non_interessante
  54  32  |   b = interessante
```

Altre metriche sono di seguito riportate:

```
=== Detailed Accuracy By Class ===
```

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.815	0.628	0.809	0.815	0.812	0.61	Non_interessante
	0.372	0.185	0.381	0.372	0.376	0.61	Interessante
Weig. Avg.	0.711	0.524	0.709	0.711	0.71	0.61	

Inoltre, anche per questo modello, si riportano le misure di carattere "monetario", già viste in precedenza. In questo caso, il modello non seleziona il soggetto *outlier*.

- *recupero (84)* = **€ 1.487.773,00**
- *recupero interessanti (32)* = **€ 1.290,989,00**
- *recupero non interessanti (52)* = **€ 196.784,00**
- *recupero (media 84)* = **€ 17.711,00**
- *recupero (mediana 84)* = **€ 4.956,00**

Il modello, pur suggerendo relativamente pochi controlli, riesce, selezionando poco meno del 22% dei soggetti accertabili, a

recuperare il 27% del gettito. Se però non consideriamo l'*outlier*, il gettito recuperato dal modello, coi soli controlli suggeriti, sale al 35%.

Valori medio e mediano dei recuperi appaiono essere leggermente superiori rispetto a quelli del modello C4.5.

Analizzando gli attributi di *split* impiegati, diversamente dal modello C4.5, si osserva come questo algoritmo utilizzi pesantemente, per le proprie previsioni, la *provincia* ed i *gruppi di attività*, oltre ad altri attributi per lo più inerenti le dichiarazioni IVA, suggerendo che i contribuenti toscani presentino effettivamente differenze legate al territorio (cosa che peraltro era emersa anche in sede di analisi OLAP dei dati) e all'attività svolta.

Sulla base di quanto emerso dal modello in esame, analisi specifiche per provincia, gruppo attività, distinte per imprenditori in contabilità ordinaria e semplificata e professionisti potrebbero quindi portare ad individuare *pattern* di evasori più circostanziati.

La struttura ad albero individuata dall'algoritmo è di seguito riportata:

```
IMP_VE_VOLAFF_2007 < 29190.5
| IMP_VE_VOLAFF_2007 < 0.5: non_interessante(16.0/0.0)
| IMP_VE_VOLAFF_2007 >= 0.5
| | SIGLA_PROVINCIA=(FI)|(PT)|(PI)|(PO)|(AR)|(SI)
| | | ALIQ_MEDIA_CESS < 19.35
| | | | TOT_PASS_2007 < 14058.5: interessante(29.0/15.0)
| | | | TOT_PASS_2007 >= 14058.5: non_interessante(18.0/2.0)
| | | ALIQ_MEDIA_CESS >= 19.35: interessante(95.0/18.0)
| | SIGLA_PROVINCIA!=(FI)|(PT)|(PI)|(PO)|(AR)|(SI)
| | | ALIQ_MEDIA_CESS < 11.67: interessante(18.0/3.0)
| | | ALIQ_MEDIA_CESS >= 11.67
| | | | IMP_REDD_PERD_2007 < 11829.0: interessante(13.0/9.0)
| | | | IMP_REDD_PERD_2007 >= 11829.0: non_interessante(37.0/2.0)
IMP_VE_VOLAFF_2007 >= 29190.5
| GRP_ATTIV_2007=(R)|(C)|(N)|(S)|(F)|(I)|(H)
| | IMP_SPS_DIPEND_SMP < 6648.0
| | | RIM_FIN_ORD < 230203.5
| | | | SIGLA_PROVINCIA=(PO)|(PT)|(FI)
| | | | | ACQ_NO_DETR < 104.0: interessante(26.0/6.0)
| | | | | ACQ_NO_DETR >= 104.0
| | | | | | CRED_ANNO_PREC < 14697.5: non_interessante(46.0/12.0)
| | | | | | CRED_ANNO_PREC >= 14697.5: interessante(13.0/2.0)
| | | | SIGLA_PROVINCIA!=(PO)|(PT)|(FI): non_interessante(159.0/35.0)
| | | RIM_FIN_ORD >= 230203.5: interessante(17.0/3.0)
| | IMP_SPS_DIPEND_SMP >= 6648.0
| | | RIM_FIN_SMPL < 2156.5
| | | | SIGLA_PROVINCIA=(SI)|(GR)|(PO)|(FI)|(AR): interessante(59.0/5.0)
| | | | SIGLA_PROVINCIA!=(SI)|(GR)|(PO)|(FI)|(AR)
| | | | | IMP_V_AGG_IVA < 70861.5: non_interessante(24.0/7.0)
| | | | | IMP_V_AGG_IVA >= 70861.5: interessante(20.0/2.0)
| | | RIM_FIN_SMPL >= 2156.5
| | | | REDD_IMP_2007 < 12307.0: interessante(19.0/11.0)
| | | | REDD_IMP_2007 >= 12307.0: non_interessante(32.0/5.0)
| GRP_ATTIV_2007!=(R)|(C)|(N)|(S)|(F)|(I)|(H): non_interessante(559.0/131.0)
```

Figura 5.10: sviluppo albero modello SimpleCart

Anche questo modello produce un albero con cammini composti da attributi appartenenti simultaneamente ai quadri RE, RF e/o RG.

Ad esempio, abbiamo la seguente ramificazione:

```
| | IMP_SPS_DIPEND_SMP < 6648.0
| | | RIM_FIN_ORD < 230203.5
| | | | SIGLA_PROVINCIA=(PO)|(PT)|(FI)
| | | | | ACQ_NO_DETR < 104.0: interessante(26.0/6.0)
| | | | | ACQ_NO_DETR >= 104.0
| | | | | | CRED_ANNO_PREC < 14697.5: non_interessante(46.0/12.0)
| | | | | | CRED_ANNO_PREC >= 14697.5: interessante(13.0/2.0)
| | | | SIGLA_PROVINCIA!=(PO)|(PT)|(FI): non_interessante(159.0/35.0)
| | | RIM_FIN_ORD >= 230203.5: interessante(17.0/3.0)
```

Ovvero, partendo da uno *split* su un attributo che si riferisce a un campo del quadro RG ([IMP_SPS_DIPEND_SMP]), il sottoalbero successivo prende in considerazione un campo del quadro RF ([RIM_FIN_ORD]), pertanto tale cammino produrrà un *pattern* non pulito.

Anche in questo caso, l'analisi con *lift chart* può rivelare alcuni fatti importanti circa il modello. Innanzitutto, si osserva che all'*outlier* viene data una bassa probabilità di essere evasore *interessante* (la discontinuità nel grafico è molto spostata verso destra) e i recuperi del modello sono abbastanza lineari proprio fino al momento di incontrare questo soggetto, che, pur dando luogo ad un controllo molto proficuo, tuttavia è riuscito a ben mimetizzarsi tra gli altri contribuenti.

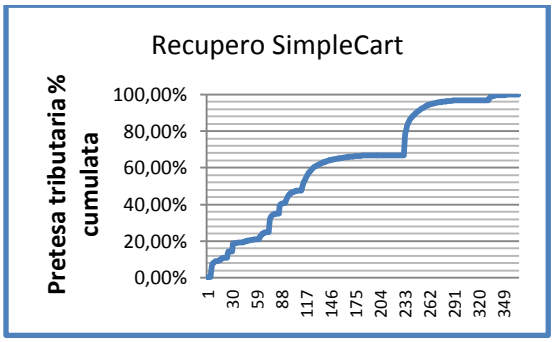

Figura 5.11: *lift chart* SimpleCart

Ai fini della valutazione del modello, presentiamo i due *lift charts*, nei due casi, con e senza *outlier*.

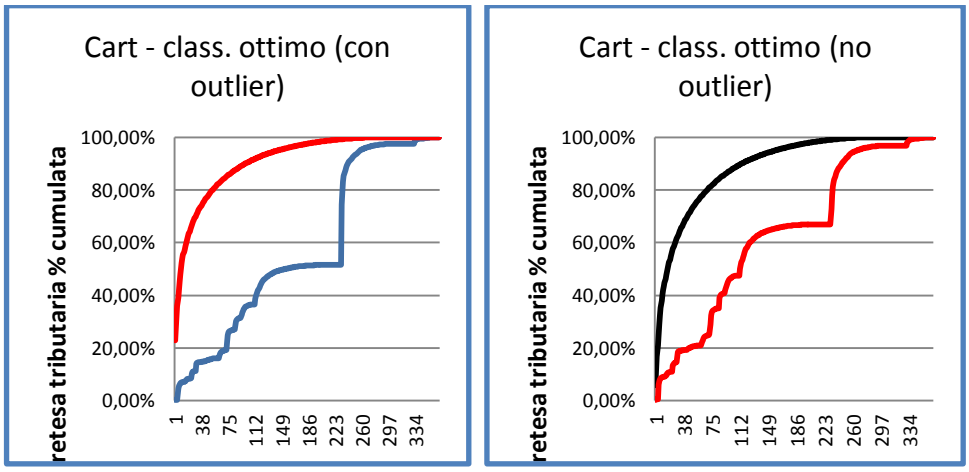

Figura 5.12: confronto CART – classificatore ottimo

L'area sotto la curva è pari al 63,11% del massimo possibile se si considera l'*outlier* nel *test set* e al 71,65% in caso contrario. Confrontando tali risultati con quelli ottenuti dal modello precedente,

emergono, da un lato, l'effetto premiante (o penalizzante) della corretta (o errata) classificazione dell'*outlier*, e dall'altro, un metro di valutazione più neutro laddove si decida di non considerare il soggetto deviante.

Per quanto espresso in precedenza, si valuta che la metrica più adatta per la valutazione del modello sia l'*AUC* senza l'inclusione dell'*outlier*.

5.3.1.3 *Ensemble methods*: *boosting, bagging*

Dopo aver visto i risultati di due classificatori classici, C4.5 e CART, iniziamo ora ad utilizzare tecniche *ensemble* per cercare di migliorare le *performance* di detti algoritmi.

Utilizziamo in prima battuta l'algoritmo *AdaBoost,* presentato in [FS96] con classificatore *SimpleCart* (con i parametri visti in precedenza), previo *resample* del *training set* (*biasToUniformClass* 0.2)[6].

Abbiamo quindi:

classifier -- The base classifier to be used.	*SimpleCart*
debug -- If set to true, classifier may output additional info to the console.	*False*
numIterations -- The number of iterations to be performed.	*10*
seed -- The random number seed to be used.	*1*
useResampling -- Whether resampling is used instead of reweighting.	*False*
weightThreshold -- Weight threshold for weight pruning.	*100*

Con detto settaggio, i risultati che il modello ottiene sul *test set* sono i seguenti:

```
Correctly Classified Instances         250               68.1199 %
Incorrectly Classified Instances       117               31.8801 %
```

Osservando solo la precisione del modello, non sembra ci sia stato un miglioramento delle *performances*. Andando a valutare il modello secondo altre metriche, osserviamo la seguente matrice di confusione:

```
=== Confusion Matrix ===

   a   b   <-- classified as
 217  64 |   a = non_interessante
  53  33 |   b = interessante
```

e:

[6] Si presentano i soli risultati relativi all'impiego, come classificatore di base, di CART, in quanto, da test effettuati, è risultato essere migliore rispetto al C4.5.

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.772	0.616	0.804	0.772	0.788	0.591	Non interessante
	0.384	0.228	0.34	0.384	0.361	0.591	Interessante
Weig. Avg.	0.681	0.525	0.695	0.681	0.688	0.591	

Come per i modelli già visti, si riportano le seguenti misure di carattere "monetario". Anche in questo caso, il modello non seleziona il soggetto *outlier*.

- *recupero (97)* = **€ 1.859.721,00**
- *recupero interessanti (33)* = **€ 1.643.646,00**
- *recupero non interessanti (64)* = **€ 216.075,00**
- *recupero (media 97)* = **€ 19.172,00**
- *recupero (mediana 97)* = **€ 4.463,00**

Il modello, pur suggerendo relativamente pochi controlli, riesce, selezionando circa il 26% dei soggetti accertabili, a recuperare il 44% del gettito (senza contare l'*outlier*).

L'*output* del modello è dato da tanti alberi quante sono le iterazioni impostate:

```
=== Classifier model ===

Scheme:  AdaBoostM1
AdaBoostM1: Base classifiers and their weights:

CART Decision Tree

REDD_LORDO_2007 < 31582.0
| SIGLA_PROVINCIA=(FI)|(AR)|(SI)|(LI)|(PT)
| | ALIQ_MEDIA_ACQ < 19.880000000000003
| | | GRP_ATTIV_2007=(J)|(M)|(F)|(A)|(S)|(C)|(I)|(G)|(H)|(K)|(R)|(E)|(B)
| | | | DURATA_ATTIV_SOGG < 17.5
| | | | | COSTI_RSDL < 10340.0
| | | | | | TOT_ACQ < 18049.0
| | | | | | | IMP_V_AGG_IMPON < 1761.0: non_interessante(16.0/0.0)
| | | | | | | IMP_V_AGG_IMPON >= 1761.0: interessante(25.0/10.0)
| | | | | | TOT_ACQ >= 18049.0: non_interessante(63.0/7.0)
| | | | | COSTI_RSDL >= 10340.0
| | | | | | GRP_ATTIV_2007=(S)|(K)|(I)|(F): interessante(24.0/4.0)
| | | | | | GRP_ATTIV_2007!=(S)|(K)|(I)|(F)
| | | | | | | IMP_V_AGG_IVA < 25135.0: interessante(15.0/10.0)
| | | | | | | IMP_V_AGG_IVA >= 25135.0: non_interessante(19.0/1.0)
| | | | DURATA_ATTIV_SOGG >= 17.5
| | | | | ETA < 56.0
| | | | | | COSTI_RSDL < 14957.5: interessante(91.0/21.0)
| | | | | | COSTI_RSDL >= 14957.5: non_interessante(23.0/10.0)
| | | | | ETA >= 56.0
| | | | | | GRP_ATTIV_2007=(H)|(J)|(F)|(M)|(C): interessante(23.0/6.0)
| | | | | | GRP_ATTIV_2007!=(H)|(J)|(F)|(M)|(C): non_interessante(40.0/4.0)
| | | GRP_ATTIV_2007!=(J)|(M)|(F)|(A)|(S)|(C)|(I)|(G)|(H)|(K)|(R)|(E)|(B):non_interessante(46.0/9.)
| | ALIQ_MEDIA_ACQ >= 19.880000000000003
| | | IMP_VE_VOLAFF_2007 < 35862.0: interessante(56.0/4.0)
| | | IMP_VE_VOLAFF_2007 >= 35862.0
| | | | RICAVI_ATT_2007 < 54995.0: non_interessante(17.0/0.0)
| | | | RICAVI_ATT_2007 >= 54995.0: interessante(52.0/26.0)
| SIGLA_PROVINCIA!=(FI)|(AR)|(SI)|(LI)|(PT): non_interessante(265.0/77.0)
REDD_LORDO_2007 >= 31582.0
| GRP_ATTIV_2007=(R)|(E)|(H)|(C)
| | BENI_DEST_RIV < 43910.0
| | | SIGLA_PROVINCIA=(PT)|(LU)|(FI)|(LI)|(MS)|(PO): interessante(25.0/4.0)
| | | SIGLA_PROVINCIA!=(PT)|(LU)|(FI)|(LI)|(MS)|(PO): non_interessante(12.0/3.0)
```

```
| | BENI_DEST_RIV >= 43910.0: non_interessante(33.0/7.0)
| GRP_ATTIV_2007!=(R)|(E)|(H)|(C): non_interessante(370.0/50.0)

Number of Leaf Nodes: 19

Size of the Tree: 37

Weight: 1.18
CART Decision Tree
CRED_ANNO_PREC < 9670.0
| SIGLA_PROVINCIA=(PO)|(MS)|(LU)|(PI)|(PT)
| | TOT_ACQ < 92501.5
| | | ETA < 38.5: interessante(65.0/12.0)
| | | ETA >= 38.5
| | | | IMP_V_AGG_IVA < 7030.0: interessante(59.0/17.0)
| | | | IMP_V_AGG_IVA >= 7030.0
| | | | | TOT_IVA_OPE_IMPO < 14226.5
| | | | | | TOT_IMPST_DOV < 2608.0: non_interessante(99.0/16.0)
| | | | | | TOT_IMPST_DOV >= 2608.0: interessante(17.0/2.0)
| | | | | TOT_IVA_OPE_IMPO >= 14226.5
| | | | | | RICAVI_ATT_2007 < 110671.0: non_interessante(14.0/8.0)
| | | | | | RICAVI_ATT_2007 >= 110671.0: interessante(50.0/7.0)
| | TOT_ACQ >= 92501.5: non_interessante(84.0/17.0)
| SIGLA_PROVINCIA!=(PO)|(MS)|(LU)|(PI)|(PT)
| | REDD_LORDO_2007 < 42425.5
| | | TOT_IVA_OPE_IMPO < 59696.0
| | | | COSTI_ACQ_MP < 175756.5
| | | | | IMP_V_AGG_IVA < 119916.5
| | | | | | GRP_ATTIV_2007=(J)|(Q)|(L)|(C)|(A)|(R)|(M)
| | | | | | | TOT_IMPST_DOV < 2037.0
| | | | | | | | IMP_REDD_PERD_2007 < 10674.5: interessante(45.0/24.0)
| | | | | | | | IMP_REDD_PERD_2007 >= 10674.5
| | | | | | | | | GRP_ATTIV_2007=(J)|(Q)|(R)|(L)
| | | | | | | | | | IMP_V_AGG_IVA < 20263.5: interessante(23.0/3.0)
| | | | | | | | | | IMP_V_AGG_IVA >= 20263.5: non_interessante(15.0/0.0)
| | | | | | | | | GRP_ATTIV_2007!=(J)|(Q)|(R)|(L): non_interessante(74.0/11.0)
| | | | | | | TOT_IMPST_DOV >= 2037.0: interessante(45.0/12.0)
| | | | | | GRP_ATTIV_2007!=(J)|(Q)|(L)|(C)|(A)|(R)|(M)
| | | | | | | IMP_SPS_DIPEND_SMP < 6441.0
| | | | | | | | DURATA_ATTIV_SOGG < 25.5: non_interessante(168.0/18.0)
| | | | | | | | DURATA_ATTIV_SOGG >= 25.5
| | | | | | | | | IMP_V_AGG_IMPON < 17735.0: interessante(27.0/6.0)
| | | | | | | | | IMP_V_AGG_IMPON >= 17735.0: non_interessante(25.0/0.0)
| | | | | | | IMP_SPS_DIPEND_SMP >= 6441.0: non_interessante(56.0/47.0)
| | | | | IMP_V_AGG_IVA >= 119916.5: interessante(21.0/3.0)
| | | | COSTI_ACQ_MP >= 175756.5: interessante(20.0/2.0)
| | | TOT_IVA_OPE_IMPO >= 59696.0: non_interessante(47.0/0.0)
| | REDD_LORDO_2007 >= 42425.5: non_interessante(165.0/23.0)
CRED_ANNO_PREC >= 9670.0: interessante(82.0/39.0)

Number of Leaf Nodes: 21

Size of the Tree: 41

Weight: 1.12

CART Decision Tree

IMP_CESS_BENI_AMM < 2400.0
| COSTO_LAV_2007 < 8802.0
| | CRED_ANNO_PREC < 1962.0: non_interessante(423.0/169.0)
| | CRED_ANNO_PREC >= 1962.0
| | | CRED_ANNO_PREC < 6641.0: interessante(87.0/30.0)
| | | CRED_ANNO_PREC >= 6641.0: non_interessante(73.0/18.0)
| COSTO_LAV_2007 >= 8802.0
| | ALIQ_MEDIA_CESS < 11.670000000000002
| | | TOT_IMPST_CRED < 0.5: non_interessante(47.0/4.0)
| | | TOT_IMPST_CRED >= 0.5
| | | | TOT_IMPST_CRED < 4626.0: interessante(38.0/18.0)
| | | | TOT_IMPST_CRED >= 4626.0: non_interessante(39.0/11.0)
| | ALIQ_MEDIA_CESS >= 11.670000000000002
| | | RICAVI_ATT_2007 < 947872.5
| | | | GRP_ATTIV_2007=(Q)|(H)|(E)|(N)|(F)|(S)|(C)|(R)|(G)|(M)|(K)|(P)|(B): interessante (210.0/98.0)
| | | | GRP_ATTIV_2007!=(Q)|(H)|(E)|(N)|(F)|(S)|(C)|(R)|(G)|(M)|(K)|(P)|(B): non_interessante (28.0/4.0)
| | | RICAVI_ATT_2007 >= 947872.5: non_interessante(21.0/0.0)
IMP_CESS_BENI_AMM >= 2400.0
| IMP_TOT_COMP_NEG_2007 < 36842.0
| | DURATA_ATTIV_SOGG < 13.0: interessante(17.0/1.0)
| | DURATA_ATTIV_SOGG >= 13.0: non_interessante(40.0/13.0)
| IMP_TOT_COMP_NEG_2007 >= 36842.0: interessante(72.0/7.0)

Number of Leaf Nodes: 12
```

Size of the Tree: 23

Weight: 0.79

CART Decision Tree

```
IMP_V_AGG_IVA < 4796.5
| IMP_VE_VOLAFF_2007 < 0.5: non_interessante(23.0/0.0)
| IMP_VE_VOLAFF_2007 >= 0.5
| | IMP_SPS_DIPEND_SMP < 41987.5
| | | GRP_ATTIV_2007=(Q)|(S)|(P)|(M)|(F)|(I)|(G)|(H)|(N)|(R)|(J)|(E)|(B)
| | | | IMP_REDD_IMP_SMPL_2007 < 25550.5: interessante(181.0/32.0)
| | | | IMP_REDD_IMP_SMPL_2007 >= 25550.5: non_interessante(12.0/5.0)
| | | GRP_ATTIV_2007!=(Q)|(S)|(P)|(M)|(F)|(I)|(G)|(H)|(N)|(R)|(J)|(E)|(B)
| | | | SIGLA_PROVINCIA=(AR)|(MS)|(FI)|(SI)|(GR): interessante(23.0/8.0)
| | | | SIGLA_PROVINCIA!=(AR)|(MS)|(FI)|(SI)|(GR): non_interessante(19.0/3.0)
| | IMP_SPS_DIPEND_SMP >= 41987.5: non_interessante(19.0/0.0)
IMP_V_AGG_IVA >= 4796.5
| GRP_ATTIV_2007=(P)|(E)|(H)|(N)|(R)|(Q)|(L)|(F)|(A)|(I)|(C)|(B)
| | RIM_FIN_SMPL < 6707.0
| | | COSTI_RSDL < 16506.5
| | | | IMP_VE_VOLAFF_2007 < 27264.5
| | | | | IMP_V_AGG_IVA < 11454.5
| | | | | | GRP_ATTIV_2007=(H)|(A)|(C)|(G)|(S)|(M)|(K)|(J)|(P)|(E)|(B): interessante(21.0/9.0)
| | | | | | GRP_ATTIV_2007!=(H)|(A)|(C)|(G)|(S)|(M)|(K)|(J)|(P)|(E)|(B): non_interessante (20.0/1.0)
| | | | | IMP_V_AGG_IVA >= 11454.5: interessante(53.0/4.0)
| | | | IMP_VE_VOLAFF_2007 >= 27264.5
| | | | | ACQ_NO_DETR < 2142.5
| | | | | | IMPST_DOV < 4624.0
| | | | | | | TOT_IVA_OPE_IMPO < 3620.5
| | | | | | | | GRP_ATTIV_2007=(L)|(C)|(A)|(F): interessante(16.0/4.0)
| | | | | | | | GRP_ATTIV_2007!=(L)|(C)|(A)|(F): non_interessante(22.0/0.0)
| | | | | | | TOT_IVA_OPE_IMPO >= 3620.5: non_interessante(93.0/3.0)
| | | | | | IMPST_DOV >= 4624.0
| | | | | | | IMP_RICAVI_SMPL_2007 < 110520.0
| | | | | | | | GRP_ATTIV_2007=(Q)|(H)|(N)|(R)|(A): interessante(19.0/2.0)
| | | | | | | | GRP_ATTIV_2007!=(Q)|(H)|(N)|(R)|(A): non_interessante(57.0/17.0)
| | | | | | | IMP_RICAVI_SMPL_2007 >= 110520.0: interessante(18.0/0.0)
| | | | | ACQ_NO_DETR >= 2142.5
| | | | | | IMP_COSTI_MP < 19010.5: non_interessante(26.0/13.0)
| | | | | | IMP_COSTI_MP >= 19010.5: interessante(51.0/6.0)
| | | COSTI_RSDL >= 16506.5
| | | | IMP_REDD_IMP_SMPL_2007 < 31665.5: interessante(96.0/15.0)
| | | | IMP_REDD_IMP_SMPL_2007 >= 31665.5: non_interessante(20.0/5.0)
| | RIM_FIN_SMPL >= 6707.0
| | | IMP_SPS_DIPEND_SMP < 536.5: interessante(16.0/8.0)
| | | IMP_SPS_DIPEND_SMP >= 536.5: non_interessante(64.0/4.0)
| GRP_ATTIV_2007!=(P)|(E)|(H)|(N)|(R)|(Q)|(L)|(F)|(A)|(I)|(C)|(B)
| | ACQ_ESENTI < 10.0
| | | IMP_CMPNS_TERZI_2007 < 2810.0: non_interessante(207.0/38.0)
| | | IMP_CMPNS_TERZI_2007 >= 2810.0
| | | | IMP_PROD_NETTA < 403.0: non_interessante(15.0/0.0)
| | | | IMP_PROD_NETTA >= 403.0
| | | | | IMP_PROD_NETTA < 70134.5: interessante(26.0/1.0)
| | | | | IMP_PROD_NETTA >= 70134.5: non_interessante(11.0/4.0)
| | ACQ_ESENTI >= 10.0
| | | SIGLA_PROVINCIA=(PO)|(LI)|(GR)|(AR)|(FI)|(PT)|(LU)|(MS)
| | | | TOT_IVA_OPE_IMPO < 16822.5
| | | | | SIGLA_PROVINCIA=(AR)|(GR)|(LI)|(PO)|(PT)|(LU)|(MS)|(SI)|(PI): interessante(21.0/9.0)
| | | | | SIGLA_PROVINCIA!=(AR)|(GR)|(LI)|(PO)|(PT)|(LU)|(MS)|(SI)|(PI): non_interessante (24.0/1.0)
| | | | TOT_IVA_OPE_IMPO >= 16822.5
| | | | | CRED_ANNO_PREC < 868.5: interessante(54.0/6.0)
| | | | | CRED_ANNO_PREC >= 868.5: non_interessante(9.0/6.0)
| | | SIGLA_PROVINCIA!=(PO)|(LI)|(GR)|(AR)|(FI)|(PT)|(LU)|(MS): non_interessante(27.0/1.0)
```

Number of Leaf Nodes: 30

Size of the Tree: 59

Weight: 1.28

CART Decision Tree

```
ALIQ_MEDIA_CESS < 18.09
| COSTI_ACQ_MP < 3389.0
| | GRP_ATTIV_2007=(R)|(E)|(N)|(F)|(Q)|(H)|(I)|(S)|(J)|(B)
| | | ETA < 40.5: interessante(73.0/10.0)
| | | ETA >= 40.5
| | | | IMP_REDD_PERD_2007 < 23157.5: non_interessante(39.0/10.0)
| | | | IMP_REDD_PERD_2007 >= 23157.5
```

```
| | | | | SIGLA_PROVINCIA=(FI)|(LI)|(SI)|(PI): interessante(42.0/10.0)
| | | | | SIGLA_PROVINCIA!=(FI)|(LI)|(SI)|(PI): non_interessante(20.0/11.0)
| | GRP_ATTIV_2007!=(R)|(E)|(N)|(F)|(Q)|(H)|(I)|(S)|(J)|(B)
| | | COSTI_ACQ_MP < 338.5: non_interessante(145.0/43.0)
| | | COSTI_ACQ_MP >= 338.5: interessante(19.0/4.0)
| COSTI_ACQ_MP >= 3389.0: non_interessante(162.0/38.0)
ALIQ_MEDIA_CESS >= 18.09
| GRP_ATTIV_2007=(P)|(Q)|(H)|(C)|(F)|(G)|(S)|(L)|(E)|(B)
| | REDD_LORDO_2007 < -13410.5: non_interessante(30.0/7.0)
| | REDD_LORDO_2007 >= -13410.5
| | | ALIQ_MEDIA_CESS < 21.384999999999998
| | | | SIGLA_PROVINCIA=(PT)|(PO)|(MS)|(AR)|(FI)|(LU)|(SI)
| | | | | RIM_FIN_SMPL < 90125.0: interessante(313.0/75.0)
| | | | | RIM_FIN_SMPL >= 90125.0: non_interessante(18.0/11.0)
| | | | SIGLA_PROVINCIA!=(PT)|(PO)|(MS)|(AR)|(FI)|(LU)|(SI)
| | | | | REDD_LORDO_2007 < 17618.5
| | | | | | ACQ_ESENTI < 0.5
| | | | | | | TOT_ACQ < 78124.0: interessante(18.0/7.0)
| | | | | | | TOT_ACQ >= 78124.0: non_interessante(17.0/0.0)
| | | | | | ACQ_ESENTI >= 0.5: interessante(36.0/3.0)
| | | | | REDD_LORDO_2007 >= 17618.5: non_interessante(34.0/2.0)
| | | ALIQ_MEDIA_CESS >= 21.384999999999998: non_interessante(23.0/4.0)
| GRP_ATTIV_2007!=(P)|(Q)|(H)|(C)|(F)|(G)|(S)|(L)|(E)|(B)
| | ALTRE_SPS_DOC < 28412.5
| | | SIGLA_PROVINCIA=(PO)|(PI)|(FI)|(LU)|(GR)
| | | | IMP_BEN_AMM < 255.0
| | | | | FLG_NO_COERENZA_07 < 0.5: interessante(41.0/6.0)
| | | | | FLG_NO_COERENZA_07 >= 0.5
| | | | | | ACQ_NO_DETR < 731.0: non_interessante(21.0/2.0)
| | | | | | ACQ_NO_DETR >= 731.0: interessante(16.0/3.0)
| | | | IMP_BEN_AMM >= 255.0: non_interessante(42.0/10.0)
| | | SIGLA_PROVINCIA!=(PO)|(PI)|(FI)|(LU)|(GR): non_interessante(71.0/6.0)
| | ALTRE_SPS_DOC >= 28412.5: interessante(23.0/3.0)

Number of Leaf Nodes: 21

Size of the Tree: 41

Weight: 1.13

CART Decision Tree

TOT_IVA_OPE_IMPO < 1662.0
| SIGLA_PROVINCIA=(LI)|(AR)|(LU)|(PT)|(PI)|(MS)|(GR): interessante(152.0/22.0)
| SIGLA_PROVINCIA!=(LI)|(AR)|(LU)|(PT)|(PI)|(MS)|(GR)
| | ALIQ_MEDIA_ACQ < 19.525: non_interessante(55.0/21.0)
| | ALIQ_MEDIA_ACQ >= 19.525: interessante(16.0/1.0)
TOT_IVA_OPE_IMPO >= 1662.0
| REDD_IMP_2007 < 8118.0
| | SIGLA_PROVINCIA=(PO)|(MS)|(GR)|(FI)|(SI)|(PT)|(LI)
| | | IMP_RICAVI_ORDIN_2007 < 142352.5: interessante(154.0/42.0)
| | | IMP_RICAVI_ORDIN_2007 >= 142352.5: non_interessante(16.0/3.0)
| | SIGLA_PROVINCIA!=(PO)|(MS)|(GR)|(FI)|(SI)|(PT)|(LI): non_interessante(32.0/13.0)
| REDD_IMP_2007 >= 8118.0
| | TOT_ACQ < 70122.5
| | | IMP_RICAVI_SMPL_2007 < 61068.0
| | | | IMP_V_AGG_IMPON < 57467.0
| | | | | TOT_IMPST_DOV < 1777.5
| | | | | | TOT_IMPST_CRED < 1372.0
| | | | | | | SIGLA_PROVINCIA=(PO)|(AR)|(FI)|(MS)|(SI)
| | | | | | | | DURATA_ATTIV_SOGG < 21.5: interessante(43.0/21.0)
| | | | | | | | DURATA_ATTIV_SOGG >= 21.5: non_interessante(54.0/13.0)
| | | | | | | SIGLA_PROVINCIA!=(PO)|(AR)|(FI)|(MS)|(SI): non_interessante(55.0/1.0)
| | | | | | TOT_IMPST_CRED >= 1372.0: interessante(30.0/4.0)
| | | | | TOT_IMPST_DOV >= 1777.5: interessante(70.0/15.0)
| | | | IMP_V_AGG_IMPON >= 57467.0: non_interessante(64.0/11.0)
| | | IMP_RICAVI_SMPL_2007 >= 61068.0: non_interessante(122.0/19.0)
| | TOT_ACQ >= 70122.5
| | | REDD_IMP_2007 < 14641.5: non_interessante(57.0/14.0)
| | | REDD_IMP_2007 >= 14641.5
| | | | IMPST_CRED < 4558.5
| | | | | SIGLA_PROVINCIA=(GR)|(PT)|(LI)|(AR)|(FI): interessante(160.0/41.0)
| | | | | SIGLA_PROVINCIA!=(GR)|(PT)|(LI)|(AR)|(FI)
| | | | | | IMP_V_AGG_IMPON < 70021.5: non_interessante(27.0/1.0)
| | | | | | IMP_V_AGG_IMPON >= 70021.5
| | | | | | | GRP_ATTIV_2007=(F)|(H)|(R)|(G)|(Q)|(L)|(A)|(S)|(K)|(N)|(J)|(P)|(E)|(B): interessante (22.0/3.0)
||||||| GRP_ATTIV_2007!=(F)|(H)|(R)|(G)|(Q)|(L)|(A)|(S)|(K)|(N)|(J)|(P)|(E)|(B): non_interessante (14.0/4.0)
| | | | IMPST_CRED >= 4558.5
| | | | | TOT_IMPST_CRED < 25250.0: non_interessante(42.0/4.0)
| | | | | TOT_IMPST_CRED >= 25250.0: interessante(23.0/7.0)

Number of Leaf Nodes: 20
```

Size of the Tree: 39

Weight: 1.25

CART Decision Tree

```
ALIQ_MEDIA_CESS < 10.015
| IMP_BEN_AMM < 30259.0
| | SIGLA_PROVINCIA=(MS)|(LI)|(PO)|(SI)|(FI)|(LU): interessante(164.0/62.0)
| | SIGLA_PROVINCIA!=(MS)|(LI)|(PO)|(SI)|(FI)|(LU)
| | | GRP_ATTIV_2007=(H)|(N)|(C)|(A)|(F)|(K)|(I)|(L)|(S)|(M)|(J)|(P)|(E)|(B)
| | | | TOT_IVA_OPE_IMPO < 12203.0
| | | | | TOT_IVA_OPE_IMPO < 2093.5
| | | | | | DURATA_ATTIV_SOGG < 19.0: non_interessante(18.0/9.0)
| | | | | | DURATA_ATTIV_SOGG >= 19.0: interessante(26.0/1.0)
| | | | | TOT_IVA_OPE_IMPO >= 2093.5: non_interessante(26.0/2.0)
| | | | TOT_IVA_OPE_IMPO >= 12203.0: interessante(25.0/2.0)
| | | GRP_ATTIV_2007!=(H)|(N)|(C)|(A)|(F)|(K)|(I)|(L)|(S)|(M)|(J)|(P)|(E)|(B): non_interessante (28.0/0.0)
| IMP_BEN_AMM >= 30259.0: non_interessante(27.0/0.0)
ALIQ_MEDIA_CESS >= 10.015
| CRED_ANNO_PREC < 222.0
| | GRP_ATTIV_2007=(E)|(Q)|(M)|(S)|(F)|(C)|(G)|(N)|(H)|(P)|(B)
| | | ALIQ_MEDIA_CESS < 20.07
| | | | IMP_REDD_IMP_SMPL_2007 < 41368.0
| | | | | IMP_VE_VOLAFF_2007 < 34567.0
| | | | | | IMP_PROD_NETTA < 58.5
| | | | | | | ETA < 61.5: non_interessante(58.0/12.0)
| | | | | | | ETA >= 61.5: interessante(14.0/3.0)
| | | | | | IMP_PROD_NETTA >= 58.5: interessante(66.0/20.0)
| | | | | IMP_VE_VOLAFF_2007 >= 34567.0
| | | | | | GRP_ATTIV_2007=(Q)|(N)|(E)|(S): interessante(18.0/1.0)
| | | | | | GRP_ATTIV_2007!=(Q)|(N)|(E)|(S)
| | | | | | | RIM_FIN_ORD < 98938.5
| | | | | | | | ALTRI_ACQ_IMP < 36190.5: non_interessante(178.0/39.0)
| | | | | | | | ALTRI_ACQ_IMP >= 36190.5
| | | | | | | | | IMP_BEN_AMM < 3435.0: interessante(31.0/6.0)
| | | | | | | | | IMP_BEN_AMM >= 3435.0: non_interessante(29.0/2.0)
| | | | | | | RIM_FIN_ORD >= 98938.5: interessante(21.0/10.0)
| | | | IMP_REDD_IMP_SMPL_2007 >= 41368.0: interessante(21.0/3.0)
| | | ALIQ_MEDIA_CESS >= 20.07: non_interessante(51.0/5.0)
| | GRP_ATTIV_2007!=(E)|(Q)|(M)|(S)|(F)|(C)|(G)|(N)|(H)|(P)|(B): non_interessante(90.0/10.0)
| CRED_ANNO_PREC >= 222.0
| | CRED_ANNO_PREC < 906.5: interessante(80.0/21.0)
| | CRED_ANNO_PREC >= 906.5
| | | IMP_V_AGG_IMPON < 58353.0
| | | | ACQ_NO_DETR < 892.5
| | | | | GRP_ATTIV_2007=(L)|(H)|(R)|(P)|(G)|(I)
| | | | | | SIGLA_PROVINCIA=(PT)|(PI)|(LI)|(SI)|(FI)|(GR)|(PO): interessante(25.0/5.0)
| | | | | | SIGLA_PROVINCIA!=(PT)|(PI)|(LI)|(SI)|(FI)|(GR)|(PO): non_interessante(17.0/4.0)
| | | | | GRP_ATTIV_2007!=(L)|(H)|(R)|(P)|(G)|(I): non_interessante(59.0/5.0)
| | | | ACQ_NO_DETR >= 892.5: interessante(76.0/15.0)
| | | IMP_V_AGG_IMPON >= 58353.0: non_interessante(71.0/12.0)
```

Number of Leaf Nodes: 24

Size of the Tree: 47

Weight: 1.06

CART Decision Tree

```
GRP_ATTIV_2007=(R)|(L)|(H)|(M)|(F)|(A)|(J)|(Q)
| DURATA_ATTIV_SOGG < 18.5
| | TOT_IVA_OPE_IMPO < 3255.0
| | | IMP_V_AGG_IVA < 7325.0: interessante(58.0/12.0)
| | | IMP_V_AGG_IVA >= 7325.0
| | | | IMP_REDD_LRD_SMPL_2007 < 14966.0: non_interessante(29.0/0.0)
| | | | IMP_REDD_LRD_SMPL_2007 >= 14966.0: interessante(21.0/4.0)
| | TOT_IVA_OPE_IMPO >= 3255.0
| | | ACQ_ESENTI < 713.5: non_interessante(134.0/24.0)
| | | ACQ_ESENTI >= 713.5: interessante(28.0/11.0)
| DURATA_ATTIV_SOGG >= 18.5
| | IMP_BEN_AMM < 3192.5
| | | IMP_V_AGG_IMPON < 90620.0
| | | | OP_ESENTI < 1169.0
| | | | | IMP_PROD_NETTA < 212.0
| | | | | | SIGLA_PROVINCIA=(SI)|(MS): interessante(15.0/2.0)
| | | | | | SIGLA_PROVINCIA!=(SI)|(MS): non_interessante(32.0/6.0)
| | | | | IMP_PROD_NETTA >= 212.0: interessante(108.0/26.0)
| | | | OP_ESENTI >= 1169.0: non_interessante(24.0/5.0)
```

```
| | | IMP_V_AGG_IMPON >= 90620.0: non_interessante(31.0/7.0)
| | IMP_BEN_AMM >= 3192.5: interessante(96.0/17.0)
GRP_ATTIV_2007!=(R)|(L)|(H)|(M)|(F)|(A)|(J)|(Q)
| COSTI_RSDL < 6010.0
| | IMP_VAR_RIM_PF < 3770.5
| | | IMP_REDD_PERD_2007 < 71177.0: non_interessante(273.0/39.0)
| | | IMP_REDD_PERD_2007 >= 71177.0
| | | | IMP_V_AGG_IVA < 109451.0: interessante(14.0/2.0)
| | | | IMP_V_AGG_IVA >= 109451.0: non_interessante(18.0/3.0)
| | IMP_VAR_RIM_PF >= 3770.5: interessante(20.0/5.0)
| COSTI_RSDL >= 6010.0
| | IMP_REDD_LRD_SMPL_2007 < 14539.0
| | | TOT_IMPST_DOV < 992.0: non_interessante(126.0/37.0)
| | | TOT_IMPST_DOV >= 992.0: interessante(27.0/10.0)
| | IMP_REDD_LRD_SMPL_2007 >= 14539.0
| | | TOT_IMPST_DOV < 17.0
| | | | ALIQ_MEDIA_ACQ < 17.895
| | | | | COSTI_RSDL < 23084.5: non_interessante(22.0/2.0)
| | | | | COSTI_RSDL >= 23084.5: interessante(17.0/5.0)
| | | | ALIQ_MEDIA_ACQ >= 17.895: interessante(83.0/5.0)
| | | TOT_IMPST_DOV >= 17.0: non_interessante(52.0/18.0)

Number of Leaf Nodes: 21

Size of the Tree: 41

Weight: 1.12

CART Decision Tree

IMP_V_AGG_IVA < -1951.0
| IMP_RICAVI_SMPL_2007 < 2995.0
| | IMP_V_AGG_IVA < -9943.0: non_interessante(43.0/15.0)
| | IMP_V_AGG_IVA >= -9943.0: interessante(19.0/2.0)
| IMP_RICAVI_SMPL_2007 >= 2995.0: interessante(99.0/25.0)
IMP_V_AGG_IVA >= -1951.0
| PROF_COMP < 106864.0
| | SIGLA_PROVINCIA=(PT)|(AR)|(PI)|(FI)|(MS)
| | | TOT_IMPST_DOV < 1378.0
| | | | TOT_PASS_2007 < 692519.0
| | | | | GRP_ATTIV_2007=(N)|(H)|(M)|(K)|(P)|(G)|(C)|(F)|(A)|(R)|(Q)|(J)|(E)|(B)
| | | | | | TOT_ACQ < 525.0: non_interessante(21.0/2.0)
| | | | | | TOT_ACQ >= 525.0
| | | | | | | IMP_TOT_COMP_NEG_2007 < 227099.0
| | | | | | | | IMP_ONERI_DIV < 6741.5
| | | | | | | | | IMP_SPS_DIPEND_ORD_2007 < 250.0
| | | | | | | | | | SIGLA_PROVINCIA=(AR)|(PI)|(FI)|(PT)|(SI)|(LU)|(GR)|(LI)|(PO): interessante (214.0/106.0)
| | | | | | | | | SIGLA_PROVINCIA!=(AR)|(PI)|(FI)|(PT)|(SI)|(LU)|(GR)|(LI)|(PO): non_interessante (20.0/7.0)
| | | | | | | | IMP_SPS_DIPEND_ORD_2007 >= 250.0: non_interessante(22.0/1.0)
| | | | | | | IMP_ONERI_DIV >= 6741.5: interessante(61.0/10.0)
| | | | | | IMP_TOT_COMP_NEG_2007 >= 227099.0: non_interessante(16.0/0.0)
| | | | GRP_ATTIV_2007!=(N)|(H)|(M)|(K)|(P)|(G)|(C)|(F)|(A)|(R)|(Q)|(J)|(E)|(B): non_interessante (34.0/4.0)
| | | | TOT_PASS_2007 >= 692519.0: non_interessante(27.0/0.0)
| | | TOT_IMPST_DOV >= 1378.0: non_interessante(102.0/46.0)
| | SIGLA_PROVINCIA!=(PT)|(AR)|(PI)|(FI)|(MS)
| | | GRP_ATTIV_2007=(I)|(L)|(R)|(M)|(G)|(J)|(E)|(B)
| | | | RIM_FIN_SMPL < 21579.0
| | | | | COSTI_RSDL < 6306.5
| | | | | | ETA < 49.5: non_interessante(60.0/10.0)
| | | | | | ETA >= 49.5
| | | | | | | TOT_IMPST_DOV < 12.0: non_interessante(22.0/6.0)
| | | | | | | TOT_IMPST_DOV >= 12.0: interessante(32.0/5.0)
| | | | | COSTI_RSDL >= 6306.5: interessante(61.0/23.0)
| | | | RIM_FIN_SMPL >= 21579.0
| | | | | IMP_PROD_NETTA < 5742.5: non_interessante(40.0/0.0)
| | | | | IMP_PROD_NETTA >= 5742.5: interessante(9.0/7.0)
| | | GRP_ATTIV_2007!=(I)|(L)|(R)|(M)|(G)|(J)|(E)|(B)
| | | | REDD_IMP_2007 < 4171.5: interessante(19.0/12.0)
| | | | REDD_IMP_2007 >= 4171.5: non_interessante(176.0/15.0)
| PROF_COMP >= 106864.0: interessante(57.0/18.0)

Number of Leaf Nodes: 21

Size of the Tree: 41

Weight: 0.97

CART Decision Tree

TOT_IMPST_DOV < 1964.5
| RIM_FIN_ORD < 223073.0
| | ALIQ_MEDIA_ACQ < 19.125
```

```
| | | ALIQ_MEDIA_ACQ < 17.994999999999997
| | | | IMP_COSTI_SERV < 69052.0
| | | | | IMP_V_AGG_IVA < 5827.0
| | | | | | DURATA_ATTIV_SOGG < 18.0
| | | | | | | IMP_REDD_PERD_2007 < 4700.5: non_interessante(45.0/1.0)
| | | | | | | IMP_REDD_PERD_2007 >= 4700.5
| | | | | | | | IMP_PROD_NETTA < 882.0: interessante(23.0/0.0)
| | | | | | | | IMP_PROD_NETTA >= 882.0: non_interessante(15.0/1.0)
| | | | | | DURATA_ATTIV_SOGG >= 18.0: interessante(60.0/19.0)
| | | | | IMP_V_AGG_IVA >= 5827.0
| | | | | | TOT_ACQ < 9174.5: non_interessante(106.0/4.0)
| | | | | | TOT_ACQ >= 9174.5
| | | | | | | GRP_ATTIV_2007=(R)|(N)|(H): interessante(28.0/9.0)
| | | | | | | GRP_ATTIV_2007!=(R)|(N)|(H)
| | | | | | | | ETA < 40.5
| | | | | | | | | IMPST_DOV < 1635.0: non_interessante(23.0/3.0)
| | | | | | | | | IMPST_DOV >= 1635.0: interessante(32.0/5.0)
| | | | | | | | ETA >= 40.5: non_interessante(142.0/43.0)
| | | | IMP_COSTI_SERV >= 69052.0: interessante(26.0/7.0)
| | | ALIQ_MEDIA_ACQ >= 17.994999999999997: non_interessante(143.0/22.0)
| | ALIQ_MEDIA_ACQ >= 19.125
| | | TOT_PASS_2007 < 668634.0
| | | | COSTI_ACQ_MP < 152188.5
| | | | | SESSO=(M)
| | | | | | TOT_PASS_2007 < 46116.0
| | | | | | | IMP_REDD_PERD_2007 < 11938.0: interessante(19.0/5.0)
| | | | | | | IMP_REDD_PERD_2007 >= 11938.0
| | | | | | | | GRP_ATTIV_2007=(Q)|(I)|(C)|(L): interessante(14.0/10.0)
| | | | | | | | GRP_ATTIV_2007!=(Q)|(I)|(C)|(L): non_interessante(51.0/3.0)
| | | | | | TOT_PASS_2007 >= 46116.0
| | | | | | | REDD_LORDO_2007 < 7539.0: non_interessante(24.0/8.0)
| | | | | | | REDD_LORDO_2007 >= 7539.0
| | | | | | | | COSTO_LAV_2007 < 80639.0: interessante(118.0/35.0)
| | | | | | | | COSTO_LAV_2007 >= 80639.0: non_interessante(18.0/9.0)
| | | | | SESSO!=(M)
| | | | | | TOT_ACQ < 25366.5: interessante(10.0/5.0)
| | | | | | TOT_ACQ >= 25366.5: non_interessante(57.0/10.0)
| | | | COSTI_ACQ_MP >= 152188.5: interessante(29.0/4.0)
| | | TOT_PASS_2007 >= 668634.0: non_interessante(23.0/0.0)
| RIM_FIN_ORD >= 223073.0: interessante(39.0/5.0)
TOT_IMPST_DOV >= 1964.5
| SIGLA_PROVINCIA=(LI)|(PO)|(MS)|(FI)|(PT)|(AR)|(PI)|(LU)
| | IMP_PROD_NETTA < 29121.0: interessante(97.0/20.0)
| | IMP_PROD_NETTA >= 29121.0
| | | REDD_LORDO_2007 < 86377.0: non_interessante(37.0/9.0)
| | | REDD_LORDO_2007 >= 86377.0: interessante(30.0/2.0)
| SIGLA_PROVINCIA!=(LI)|(PO)|(MS)|(FI)|(PT)|(AR)|(PI)|(LU): non_interessante(18.0/2.0)

Number of Leaf Nodes: 26

Size of the Tree: 51

Weight: 1.25

Number of performed Iterations: 10
```

Anche in questo caso, l'analisi con *lift chart* può rivelare alcuni fatti importanti circa il modello. Dalla discontinuità presente nel grafico, si nota immediatamente che all'*outlier* viene data una bassa probabilità di essere evasore interessante, per cui non viene proposto per un controllo.

Anche in questo caso, si propongono i *lift charts* con e senza l'inclusione dell'*outlier*, ai fini della valutazione del modello.

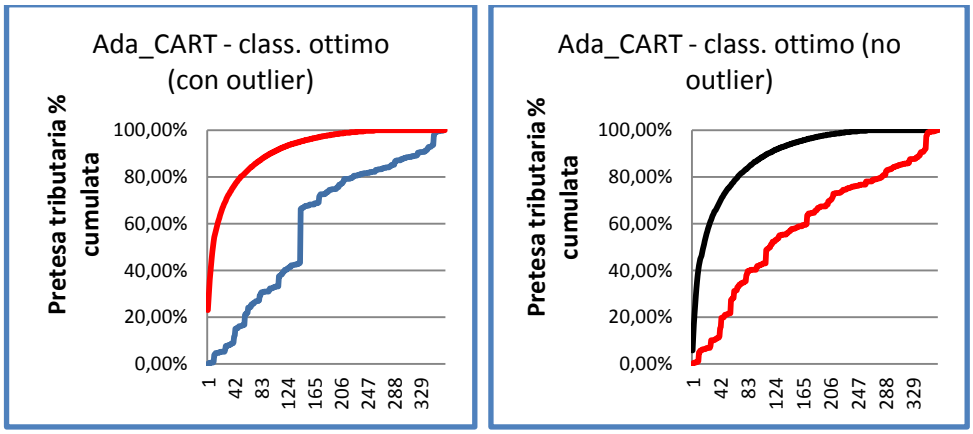

Figura 5.13: confronto Adaboost(CART) – classificatore ottimo

L'area sotto la curva è pari al 65,68% del massimo possibile se si considera l'*outlier* nel *test set*, mentre è pari al 67,07% in caso contrario.

Per quanto concerne le variabili che entrano in gioco nella costruzione degli alberi delle varie iterazioni, vale la pena di osservare che gli alberi stessi mediamente hanno molti nodi e foglie, per cui la loro interpretabilità non è immediata. Inoltre, essendo la predizione della classe di ogni istanza una media pesata ottenuta con le risultanze dei diversi alberi via via generati, l'individuazione di *pattern* che possano rivelare la presenza di evasori interessanti non appare intuitiva e immediata.

Una tecnica alternativa per combinare le predizioni di più classificatori è data dal *bagging*, come descritto in [Bre96]. In questo caso, presenteremo i risultati ottenuti utilizzando, come classificatore, un albero *J48* (con i parametri visti in precedenza) previo *resample* del *training set* (*biasToUniformClass* 0.1), in quanto, da *test* effettuati, ha fornito risultati migliori rispetto a CART.

Sono stati utilizzati i seguenti parametri:

bagSizePercent -- Size of each bag, as a percentage of the training set size.	*0,85*
calcOutOfBag -- Whether the out-of-bag error is calculated.	*False*
classifier -- The base classifier to be used.	*J48*
debug -- If set to true, classifier may output additional info to the console.	*False*
numIterations -- The number of iterations to be performed.	*10*
seed -- The random number seed to be used.	*1*

Con detto settaggio, i risultati che il modello ottiene sul *test set* sono i seguenti:

```
Correctly Classified Instances         281               76.5668 %
Incorrectly Classified Instances        86               23.4332 %
```

Questa metodologia *ensemble* evidenzia un numero di istanze correttamente classificate maggiore rispetto ai casi precedenti.

Andando a valutare il modello secondo le ormai usuali metriche, si scopre poi come esso sia estremamente "prudente" nel selezionare i soggetti da accertare, in quanto suggerisce solo 22 controlli su 367. La relativa matrice di confusione viene di seguito riportata:

```
=== Confusion Matrix ===

   a   b   <-- classified as
 270  11 |   a = non_interessante
  75  11 |   b = interessante
```

Altre metriche sono di seguito riportate:

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.961	0.872	0.783	0.961	0.863	0.652	Non_interessante
	0.128	0.039	0.5	0.128	0.204	0.652	Interessante
Weig. Avg.	0.766	0.677	0.716	0.766	0.708	0.652	

Inoltre, come per i modelli già visti, si riportano le usuali misure di carattere "monetario", con l'avvertenza che anche in questo caso, il modello non seleziona il soggetto *outlier*.

- *recupero (22)* = **€ 543.336,00**
- *recupero interessanti (11)* = **€ 514.220,00**
- *recupero non interessanti (11)* = **€ 29.116,00**
- *recupero (media_22)* = **€ 24.697,00**
- *recupero (mediana_22)* = **€ 3.200,00**

Il modello, pur suggerendo relativamente pochi controlli, riesce, dopo aver selezionato circa il 6% dei soggetti accertabili, a recuperare il 10% del gettito (o il 13, se non si considera l'*outlier*).

I soggetti selezionati sono quelli che producono il recupero medio più alto tra i modelli visti sin ora e la percentuale di soggetti *interessanti* sul totale di quelli selezionati è la massima finora riscontrata, pur non essendo, in valore assoluto, elevatissima (50%). Risultato che appare comunque pregevole, considerato che i soggetti non interessanti occupano circa ¾ del *dataset*.

L'*output* del modello è dato da:

```
Scheme:   Bagging

All the base classifiers:

J48 pruned tree
------------------

J48 pruned tree (1)
------------------

FLG_STUDIO_SETTORE_07 <= 0
|  IMP_VE_VOLAFF_2007 <= 3016: non_interessante (16.0)
|  IMP_VE_VOLAFF_2007 > 3016
|  |  IMP_VE_VOLAFF_2007 <= 21593: interessante (34.0/2.0)
|  |  IMP_VE_VOLAFF_2007 > 21593
|  |  |  REDD_IMP_2007 <= 14862: non_interessante (21.0/2.0)
|  |  |  REDD_IMP_2007 > 14862: interessante (22.0/7.0)
FLG_STUDIO_SETTORE_07 > 0
|  REDD_LORDO_2007 <= 52400
|  |  IMP_V_AGG_IVA <= 187103
|  |  |  QTA_PART_IVA <= 2
|  |  |  |  ALTRI_ACQ_IMP <= 66467
|  |  |  |  |  ETA <= 70
|  |  |  |  |  |  CRED_ANNO_PREC <= 2226
|  |  |  |  |  |  |  ACQ_ESENTI <= 1100
|  |  |  |  |  |  |  |  SIGLA_PROVINCIA = SI: non_interessante (91.0/11.0)
|  |  |  |  |  |  |  |  SIGLA_PROVINCIA = PT: non_interessante (17.0/1.0)
|  |  |  |  |  |  |  |  SIGLA_PROVINCIA = LU: non_interessante (82.0/11.0)
|  |  |  |  |  |  |  |  SIGLA_PROVINCIA = AR: non_interessante (18.0/3.0)
|  |  |  |  |  |  |  |  SIGLA_PROVINCIA = GR: non_interessante (48.0)
|  |  |  |  |  |  |  |  SIGLA_PROVINCIA = MS: non_interessante (27.0/5.0)
|  |  |  |  |  |  |  |  SIGLA_PROVINCIA = PI: non_interessante (47.0/4.0)
|  |  |  |  |  |  |  |  SIGLA_PROVINCIA = FI
|  |  |  |  |  |  |  |  |  ALIQ_MEDIA_ACQ <= 19.93: non_interessante (88.0/17.0)
|  |  |  |  |  |  |  |  |  ALIQ_MEDIA_ACQ > 19.93
|  |  |  |  |  |  |  |  |  |  COSTI_RSDL <= 2347: non_interessante (16.0/5.0)
|  |  |  |  |  |  |  |  |  |  COSTI_RSDL > 2347: interessante (25.0/3.0)
|  |  |  |  |  |  |  |  SIGLA_PROVINCIA = LI: non_interessante (46.0/9.0)
|  |  |  |  |  |  |  |  SIGLA_PROVINCIA = PO: non_interessante (13.0/1.0)
|  |  |  |  |  |  |  ACQ_ESENTI > 1100
|  |  |  |  |  |  |  |  DURATA_ATTIV_SOGG <= 21: interessante (21.0/6.0)
|  |  |  |  |  |  |  |  DURATA_ATTIV_SOGG > 21: non_interessante (22.0/4.0)
|  |  |  |  |  |  CRED_ANNO_PREC > 2226
|  |  |  |  |  |  |  IMP_VAR_RIM_MP <= 3358
|  |  |  |  |  |  |  |  TOT_IMPST_DOV <= 556
|  |  |  |  |  |  |  |  |  RICAVI_ATT_2007 <= 48490: interessante (27.0/3.0)
|  |  |  |  |  |  |  |  |  RICAVI_ATT_2007 > 48490
|  |  |  |  |  |  |  |  |  |  TOT_PASS_2007 <= 121172: non_interessante (22.0/3.0)
|  |  |  |  |  |  |  |  |  |  TOT_PASS_2007 > 121172: interessante (22.0/6.0)
|  |  |  |  |  |  |  |  TOT_IMPST_DOV > 556: non_interessante (18.0/2.0)
|  |  |  |  |  |  |  IMP_VAR_RIM_MP > 3358: non_interessante (16.0)
|  |  |  |  |  ETA > 70
|  |  |  |  |  |  ALIQ_MEDIA_ACQ <= 19.87: interessante (23.0/2.0)
|  |  |  |  |  |  ALIQ_MEDIA_ACQ > 19.87: non_interessante (16.0)
|  |  |  |  ALTRI_ACQ_IMP > 66467
|  |  |  |  |  SESSO = M
|  |  |  |  |  |  ALIQ_MEDIA_ACQ <= 19.65: non_interessante (31.0/2.0)
|  |  |  |  |  |  ALIQ_MEDIA_ACQ > 19.65: interessante (30.0/9.0)
|  |  |  |  |  SESSO = F: interessante (18.0)
|  |  |  QTA_PART_IVA > 2
|  |  |  |  IMP_TOT_COMP_NEG_2007 <= 47207: non_interessante (19.0/3.0)
|  |  |  |  IMP_TOT_COMP_NEG_2007 > 47207: interessante (17.0/1.0)
|  |  IMP_V_AGG_IVA > 187103: interessante (32.0/7.0)
|  REDD_LORDO_2007 > 52400: non_interessante (302.0/28.0)

Number of Leaves  :     32

Size of the tree :      55


J48 pruned tree (2)
------------------

IMP_REDD_LRD_ORD <= 121364
|  IMP_VE_VOLAFF_2007 <= 17730
|  |  IMP_VE_VOLAFF_2007 <= 0: non_interessante (22.0)
|  |  IMP_VE_VOLAFF_2007 > 0
|  |  |  ACQ_NO_DETR <= 151
|  |  |  |  ALTRI_ACQ_IMP <= 264: non_interessante (16.0/4.0)
|  |  |  |  ALTRI_ACQ_IMP > 264: interessante (64.0/8.0)
|  |  |  ACQ_NO_DETR > 151
|  |  |  |  IMP_V_AGG_IVA <= 2025: interessante (18.0/8.0)
```

```
| | | | IMP_V_AGG_IVA > 2025: non_interessante (20.0/2.0)
| IMP_VE_VOLAFF_2007 > 17730
| | IMP_VAR_RIM_PF <= 1200
| | | BENI_DEST_RIV <= 476477
| | | | FLG_STUDIO_SETTORE_07 <= 0
| | | | | IMP_TOT_COMP_NEG_2007 <= 40515: non_interessante (47.0/12.0)
| | | | | IMP_TOT_COMP_NEG_2007 > 40515: interessante (17.0/3.0)
| | | | FLG_STUDIO_SETTORE_07 > 0
| | | | | QTA_PART_IVA <= 2
| | | | | | TOT_ACQ <= 246562
| | | | | | | RIM_FIN_ORD <= 0
| | | | | | | | IMP_REDD_LAV_AUT_2007 <= 65258
| | | | | | | | | PROF_COSTI <= 29306
| | | | | | | | | | IMP_CMPNS_ATTIV_2007 <= 55000
| | | | | | | | | | | TOT_IMPST_DOV <= 2016
| | | | | | | | | | | | ALTRE_SPS_DOC <= 2364
| | | | | | | | | | | | | IMP_BEN_AMM <= 104: non_interessante (211.0/20.0)
| | | | | | | | | | | | | IMP_BEN_AMM > 104
| | | | | | | | | | | | | | REDD_IMP_2007 <= 2544: interessante (25.0/7.0)
| | | | | | | | | | | | | | REDD_IMP_2007 > 2544
| | | | | | | | | | | | | | | ETA <= 65
| | | | | | | | | | | | | | | REDD_LORDO_2007 <= 75012
| | | | | | | | | | | | | | | | TOT_PASS_2007 <= 30209
| | | | | | | | | | | | | | | | | COSTI_RSDL <= 5315: non_interessante (16.0/3.0)
| | | | | | | | | | | | | | | | | COSTI_RSDL > 5315: interessante (29.0/6.0)
| | | | | | | | | | | | | | | | TOT_PASS_2007 > 30209
| | | | | | | | | | | | | | | | | DURATA_ATTIV_SOGG <= 7: non_interessante (36.0)
| | | | | | | | | | | | | | | | | DURATA_ATTIV_SOGG > 7
| | | | | | | | | | | | | | | | | | IMPST_CRED <= 35
| | | | | | | | | | | | | | | | | | | RICAVI_ATT_2007 <= 96365: non_interessante (44.0)
| | | | | | | | | | | | | | | | | | | RICAVI_ATT_2007 > 96365
| | | | | | | | | | | | | | | | | | | | RIM_FIN_SMPL <= 3320
| | | | | | | | | | | | | | | | | | | | | REDD_IMP_2007 <= 27236: interessante (15.0/5.0)
| | | | | | | | | | | | | | | | | | | | | REDD_IMP_2007 > 27236: non_interessante (16.0/4.0)
| | | | | | | | | | | | | | | | | | | | RIM_FIN_SMPL > 3320: non_interessante (15.0)
| | | | | | | | | | | | | | | | | | | IMPST_CRED > 35: interessante (26.0/11.0)
| | | | | | | | | | | | | | | | | | REDD_LORDO_2007 > 75012: non_interessante (30.0)
| | | | | | | | | | | | | | | ETA > 65: interessante (23.0/8.0)
| | | | | | | | | | | | ALTRE_SPS_DOC > 2364: non_interessante (35.0)
| | | | | | | | | | | TOT_IMPST_DOV > 2016
| | | | | | | | | | | | PROF_COMP <= 12338
| | | | | | | | | | | | | DURATA_ATTIV_SOGG <= 13
| | | | | | | | | | | | | DURATA_ATTIV_SOGG <= 6: non_interessante (22.0/3.0)
| | | | | | | | | | | | | DURATA_ATTIV_SOGG > 6: interessante (23.0/3.0)
| | | | | | | | | | | | DURATA_ATTIV_SOGG > 13: non_interessante (38.0/4.0)
| | | | | | | | | | | PROF_COMP > 12338: interessante (19.0/3.0)
| | | | | | | | | IMP_CMPNS_ATTIV_2007 > 55000: non_interessante (44.0)
| | | | | | | | PROF_COSTI > 29306
| | | | | | | | | TOT_ACQ <= 38271: non_interessante (20.0/6.0)
| | | | | | | | | TOT_ACQ > 38271: interessante (17.0/3.0)
| | | | | | | IMP_REDD_LAV_AUT_2007 > 65258: non_interessante (48.0/2.0)
| | | | | | RIM_FIN_ORD > 0: non_interessante (67.0)
| | | | | TOT_ACQ > 246562
| | | | | | TOT_IVA_OPE_IMPO <= 134447
| | | | | | | DURATA_ATTIV_SOGG <= 8: interessante (15.0)
| | | | | | | DURATA_ATTIV_SOGG > 8
| | | | | | | | ACQ_NO_DETR <= 274: non_interessante (20.0/2.0)
| | | | | | | | ACQ_NO_DETR > 274
| | | | | | | | | DEB_FORN_ORD <= 87256: non_interessante (16.0/5.0)
| | | | | | | | | DEB_FORN_ORD > 87256: interessante (27.0/4.0)
| | | | | | TOT_IVA_OPE_IMPO > 134447: non_interessante (19.0)
| | | | | QTA_PART_IVA > 2
| | | | | | IMP_V_AGG_IVA <= 27769: interessante (16.0/2.0)
| | | | | | IMP_V_AGG_IVA > 27769: non_interessante (25.0/4.0)
| | | BENI_DEST_RIV > 476477: non_interessante (34.0)
| | IMP_VAR_RIM_PF > 1200: interessante (26.0/12.0)
IMP_REDD_LRD_ORD > 121364: non_interessante (26.0)

Number of Leaves :      39

Size of the tree :      77


J48 pruned tree (3)
------------------

FLG_STUDIO_SETTORE_07 <= 0
| IMP_V_AGG_IMPON <= 39317
| | ALIQ_MEDIA_ACQ <= 19.38
| | | IMP_BEN_AMM <= 1123
| | | | IMP_PROD_NETTA <= 98: non_interessante (33.0/10.0)
| | | | IMP_PROD_NETTA > 98: interessante (15.0/6.0)
```

```
| | | IMP_BEN_AMM > 1123: interessante (16.0/2.0)
| | ALIQ_MEDIA_ACQ > 19.38: interessante (21.0)
| IMP_V_AGG_IMPON > 39317: non_interessante (32.0/6.0)
FLG_STUDIO_SETTORE_07 > 0
| IMP_VE_VOLAFF_2007 <= 7013: interessante (31.0/10.0)
| IMP_VE_VOLAFF_2007 > 7013
| | REDD_LORDO_2007 <= 55476
| | | IMPST_DOV <= 30802
| | | | TOT_PASS_2007 <= 3141: non_interessante (52.0/1.0)
| | | | TOT_PASS_2007 > 3141
| | | | | BENI_DEST_RIV <= 1000
| | | | | | COSTI_RSDL <= 53281
| | | | | | | TOT_PASS_2007 <= 130528
| | | | | | | | CESS_NON_IMPO <= 1144
| | | | | | | | | GRP_ATTIV_2007 = F: interessante (7.0/3.0)
| | | | | | | | | GRP_ATTIV_2007 = Q: non_interessante (14.0/6.0)
| | | | | | | | | GRP_ATTIV_2007 = I: non_interessante (1.0)
| | | | | | | | | GRP_ATTIV_2007 = G: non_interessante (33.0/4.0)
| | | | | | | | | GRP_ATTIV_2007 = L
| | | | | | | | | | IMPST_DOV <= 7562: interessante (17.0/7.0)
| | | | | | | | | | IMPST_DOV > 7562: non_interessante (15.0)
| | | | | | | | | GRP_ATTIV_2007 = H: non_interessante (25.0/1.0)
| | | | | | | | | GRP_ATTIV_2007 = A: non_interessante (0.0)
| | | | | | | | | GRP_ATTIV_2007 = S: non_interessante (7.0/2.0)
| | | | | | | | | GRP_ATTIV_2007 = M
| | | | | | | | | | IMP_PROD_NETTA <= 9238: non_interessante (39.0/2.0)
| | | | | | | | | | IMP_PROD_NETTA > 9238
| | | | | | | | | | | IMP_VE_VOLAFF_2007 <= 56394: interessante (19.0/2.0)
| | | | | | | | | | | IMP_VE_VOLAFF_2007 > 56394
| | | | | | | | | | | | PROF_COSTI <= 21237: non_interessante (20.0)
| | | | | | | | | | | | PROF_COSTI > 21237: interessante (16.0/7.0)
| | | | | | | | | GRP_ATTIV_2007 = C: interessante (32.0/2.0)
| | | | | | | | | GRP_ATTIV_2007 = K: non_interessante (4.0)
| | | | | | | | | GRP_ATTIV_2007 = N: non_interessante (2.0)
| | | | | | | | | GRP_ATTIV_2007 = R: non_interessante (3.0)
| | | | | | | | | GRP_ATTIV_2007 = J: non_interessante (0.0)
| | | | | | | | | GRP_ATTIV_2007 = P: non_interessante (3.0)
| | | | | | | | | GRP_ATTIV_2007 = E: non_interessante (0.0)
| | | | | | | | | GRP_ATTIV_2007 = B: non_interessante (0.0)
| | | | | | | | CESS_NON_IMPO > 1144: interessante (19.0/5.0)
| | | | | | | TOT_PASS_2007 > 130528: interessante (18.0/2.0)
| | | | | | COSTI_RSDL > 53281: non_interessante (18.0)
| | | | | BENI_DEST_RIV > 1000
| | | | | | ACQ_NO_DETR <= 1332
| | | | | | | FLG_PRES_FAM_07 <= 0
| | | | | | | | TOT_PASS_2007 <= 82248: non_interessante (151.0/6.0)
| | | | | | | | TOT_PASS_2007 > 82248
| | | | | | | | | IMPST_CRED <= 4651
| | | | | | | | | | TOT_PASS_2007 <= 108836: interessante (16.0/4.0)
| | | | | | | | | | TOT_PASS_2007 > 108836
| | | | | | | | | | | TOT_IMPST_DOV <= 17
| | | | | | | | | | | | RIM_FIN_SMPL <= 12000: interessante (29.0/6.0)
| | | | | | | | | | | | RIM_FIN_SMPL > 12000: non_interessante (38.0/7.0)
| | | | | | | | | | | TOT_IMPST_DOV > 17: non_interessante (51.0/2.0)
| | | | | | | | | IMPST_CRED > 4651: non_interessante (36.0)
| | | | | | | FLG_PRES_FAM_07 > 0: non_interessante (22.0)
| | | | | | ACQ_NO_DETR > 1332
| | | | | | | OP_ESENTI <= 346
| | | | | | | | DURATA_ATTIV_SOGG <= 26
| | | | | | | | | ALIQ_MEDIA_CESS <= 19.41: non_interessante (50.0/10.0)
| | | | | | | | | ALIQ_MEDIA_CESS > 19.41: interessante (18.0/4.0)
| | | | | | | | DURATA_ATTIV_SOGG > 26: interessante (20.0/3.0)
| | | | | | | OP_ESENTI > 346: non_interessante (20.0)
| | | IMPST_DOV > 30802
| | | | IMP_PROD_NETTA <= 148685: interessante (15.0/1.0)
| | | | IMP_PROD_NETTA > 148685: non_interessante (15.0/5.0)
| | REDD_LORDO_2007 > 55476: non_interessante (254.0/23.0)

Number of Leaves :      45

Size of the tree :      74


J48 pruned tree (4)
------------------

IMP_REDD_LRD_SMPL_2007 <= 59489
| RICAVI_ATT_2007 <= 15283
| | COSTI_RSDL <= 1643: non_interessante (77.0/28.0)
| | COSTI_RSDL > 1643: interessante (40.0/4.0)
| RICAVI_ATT_2007 > 15283
| | TOT_IVA_OPE_IMPO <= 140798
```

```
|  |  |  TOT_IVA_OPE_IMPO <= 93117
|  |  |  |  TOT_IMPST_CRED <= 21799
|  |  |  |  |  OP_NON_IMPO_DI <= 5383
|  |  |  |  |  |  CRED_CLI_ORD <= 2340
|  |  |  |  |  |  |  IMP_BEN_STRUM_NA <= 1906
|  |  |  |  |  |  |  |  QTA_PART_IVA <= 2
|  |  |  |  |  |  |  |  |  SIGLA_PROVINCIA = SI: non_interessante (112.0/13.0)
|  |  |  |  |  |  |  |  |  SIGLA_PROVINCIA = PT: non_interessante (32.0/9.0)
|  |  |  |  |  |  |  |  |  SIGLA_PROVINCIA = LU: non_interessante (92.0/11.0)
|  |  |  |  |  |  |  |  |  SIGLA_PROVINCIA = AR
|  |  |  |  |  |  |  |  |  |  IMP_PROD_NETTA <= 23714: interessante (19.0/6.0)
|  |  |  |  |  |  |  |  |  |  IMP_PROD_NETTA > 23714: non_interessante (21.0)
|  |  |  |  |  |  |  |  |  SIGLA_PROVINCIA = GR: non_interessante (33.0/2.0)
|  |  |  |  |  |  |  |  |  SIGLA_PROVINCIA = MS
|  |  |  |  |  |  |  |  |  |  COSTO_LAV_2007 <= 13567: non_interessante (24.0/1.0)
|  |  |  |  |  |  |  |  |  |  COSTO_LAV_2007 > 13567: interessante (15.0/4.0)
|  |  |  |  |  |  |  |  |  SIGLA_PROVINCIA = PI: non_interessante (41.0/2.0)
|  |  |  |  |  |  |  |  |  SIGLA_PROVINCIA = FI
|  |  |  |  |  |  |  |  |  |  ALIQ_MEDIA_ACQ <= 19.98
|  |  |  |  |  |  |  |  |  |  |  REDD_LORDO_2007 <= 55476
|  |  |  |  |  |  |  |  |  |  |  |  DURATA_ATTIV_SOGG <= 6: non_interessante (19.0/1.0)
|  |  |  |  |  |  |  |  |  |  |  |  DURATA_ATTIV_SOGG > 6
|  |  |  |  |  |  |  |  |  |  |  |  |  SESSO = M
|  |  |  |  |  |  |  |  |  |  |  |  |  |  IMP_PROD_NETTA <= 25618: non_interessante (48.0/10.0)
|  |  |  |  |  |  |  |  |  |  |  |  |  |  IMP_PROD_NETTA > 25618: interessante (22.0/5.0)
|  |  |  |  |  |  |  |  |  |  |  |  |  SESSO = F: non_interessante (19.0/4.0)
|  |  |  |  |  |  |  |  |  |  |  REDD_LORDO_2007 > 55476: non_interessante (42.0)
|  |  |  |  |  |  |  |  |  |  ALIQ_MEDIA_ACQ > 19.98: interessante (21.0/2.0)
|  |  |  |  |  |  |  |  |  SIGLA_PROVINCIA = LI
|  |  |  |  |  |  |  |  |  |  PROF_COSTI <= 2517
|  |  |  |  |  |  |  |  |  |  |  IMP_CESS_BENI_AMM <= 72
|  |  |  |  |  |  |  |  |  |  |  |  TOT_IVA_OPE_IMPO <= 6158: non_interessante (15.0)
|  |  |  |  |  |  |  |  |  |  |  |  TOT_IVA_OPE_IMPO > 6158
|  |  |  |  |  |  |  |  |  |  |  |  |  DURATA_ATTIV_SOGG <= 13: non_interessante (19.0/4.0)
|  |  |  |  |  |  |  |  |  |  |  |  |  DURATA_ATTIV_SOGG > 13: interessante (15.0/4.0)
|  |  |  |  |  |  |  |  |  |  |  IMP_CESS_BENI_AMM > 72: non_interessante (16.0)
|  |  |  |  |  |  |  |  |  |  PROF_COSTI > 2517: non_interessante (15.0)
|  |  |  |  |  |  |  |  |  SIGLA_PROVINCIA = PO: non_interessante (22.0/4.0)
|  |  |  |  |  |  |  |  QTA_PART_IVA > 2: non_interessante (27.0/11.0)
|  |  |  |  |  |  |  IMP_BEN_STRUM_NA > 1906
|  |  |  |  |  |  |  |  ACQ_NO_DETR <= 372: interessante (16.0/1.0)
|  |  |  |  |  |  |  |  ACQ_NO_DETR > 372
|  |  |  |  |  |  |  |  |  IMP_BEN_AMM <= 95: non_interessante (19.0)
|  |  |  |  |  |  |  |  |  IMP_BEN_AMM > 95
|  |  |  |  |  |  |  |  |  |  TOT_IMPST_DOV <= 1014
|  |  |  |  |  |  |  |  |  |  |  IMPST_DOV <= 998: non_interessante (17.0/4.0)
|  |  |  |  |  |  |  |  |  |  |  IMPST_DOV > 998: interessante (25.0/6.0)
|  |  |  |  |  |  |  |  |  |  TOT_IMPST_DOV > 1014: non_interessante (16.0/2.0)
|  |  |  |  |  |  CRED_CLI_ORD > 2340: non_interessante (105.0/7.0)
|  |  |  |  |  OP_NON_IMPO_DI > 5383
|  |  |  |  |  |  TOT_IMPST_CRED <= 46: interessante (22.0)
|  |  |  |  |  |  TOT_IMPST_CRED > 46: non_interessante (24.0/3.0)
|  |  |  |  TOT_IMPST_CRED > 21799: interessante (34.0/11.0)
|  |  |  TOT_IVA_OPE_IMPO > 93117
|  |  |  |  BENI_DEST_RIV <= 99274: interessante (18.0)
|  |  |  |  BENI_DEST_RIV > 99274: non_interessante (26.0/6.0)
|  |  TOT_IVA_OPE_IMPO > 140798: non_interessante (48.0/2.0)
IMP_REDD_LRD_SMPL_2007 > 59489: non_interessante (71.0/1.0)

Number of Leaves  :     37

Size of the tree :      65


J48 pruned tree (5)
------------------

IMP_VE_VOLAFF_2007 <= 17730
|  IMP_VE_VOLAFF_2007 <= 0: non_interessante (17.0)
|  IMP_VE_VOLAFF_2007 > 0
|  |  IMP_TOT_COMP_NEG_2007 <= 1643
|  |  |  GRP_ATTIV_2007 = F: non_interessante (2.0)
|  |  |  GRP_ATTIV_2007 = Q: interessante (5.0/1.0)
|  |  |  GRP_ATTIV_2007 = I: non_interessante (7.0)
|  |  |  GRP_ATTIV_2007 = G: non_interessante (4.0)
|  |  |  GRP_ATTIV_2007 = L: non_interessante (4.0)
|  |  |  GRP_ATTIV_2007 = H: non_interessante (0.0)
|  |  |  GRP_ATTIV_2007 = A: interessante (18.0/6.0)
|  |  |  GRP_ATTIV_2007 = S: non_interessante (0.0)
|  |  |  GRP_ATTIV_2007 = M: non_interessante (24.0/10.0)
|  |  |  GRP_ATTIV_2007 = C: non_interessante (0.0)
|  |  |  GRP_ATTIV_2007 = K: non_interessante (0.0)
```

```
|  |  |  GRP_ATTIV_2007 = N: non_interessante (0.0)
|  |  |  GRP_ATTIV_2007 = R: non_interessante (0.0)
|  |  |  GRP_ATTIV_2007 = J: non_interessante (0.0)
|  |  |  GRP_ATTIV_2007 = P: non_interessante (0.0)
|  |  |  GRP_ATTIV_2007 = E: non_interessante (0.0)
|  |  |  GRP_ATTIV_2007 = B: non_interessante (0.0)
|  |  IMP_TOT_COMP_NEG_2007 > 1643: interessante (64.0/12.0)
IMP_VE_VOLAFF_2007 > 17730
|  COSTI_ACQ_MP <= 255305
|  |  IMP_REDD_LRD_ORD <= -2340: non_interessante (28.0)
|  |  IMP_REDD_LRD_ORD > -2340
|  |  |  REDD_LORDO_2007 <= 56492
|  |  |  |  TOT_IVA_OPE_IMPO <= 88451
|  |  |  |  |  PROF_COSTI <= 29306
|  |  |  |  |  |  IMP_BEN_AMM <= 31067
|  |  |  |  |  |  |  IMP_BEN_AMM <= 16324
|  |  |  |  |  |  |  |  |  QTA_PART_IVA <= 2
|  |  |  |  |  |  |  |  |  |  IMP_V_AGG_IVA <= -1262
|  |  |  |  |  |  |  |  |  |  |  CRED_ANNO_PREC <= 3281: non_interessante (29.0/11.0)
|  |  |  |  |  |  |  |  |  |  |  CRED_ANNO_PREC > 3281: interessante (17.0/2.0)
|  |  |  |  |  |  |  |  |  |  IMP_V_AGG_IVA > -1262
|  |  |  |  |  |  |  |  |  |  |  IMPST_CRED <= 1944: non_interessante (545.0/95.0)
|  |  |  |  |  |  |  |  |  |  |  IMPST_CRED > 1944
|  |  |  |  |  |  |  |  |  |  |  |  RIM_FIN_SMPL <= 1142: interessante (29.0/8.0)
|  |  |  |  |  |  |  |  |  |  |  |  RIM_FIN_SMPL > 1142: non_interessante (18.0/2.0)
|  |  |  |  |  |  |  |  |  QTA_PART_IVA > 2: interessante (27.0/11.0)
|  |  |  |  |  |  |  IMP_BEN_AMM > 16324
|  |  |  |  |  |  |  |  IMP_REDD_LRD_SMPL_2007 <= 18216: non_interessante (18.0/7.0)
|  |  |  |  |  |  |  |  IMP_REDD_LRD_SMPL_2007 > 18216: interessante (17.0/2.0)
|  |  |  |  |  |  IMP_BEN_AMM > 31067: non_interessante (35.0)
|  |  |  |  |  PROF_COSTI > 29306: interessante (25.0/9.0)
|  |  |  |  TOT_IVA_OPE_IMPO > 88451: interessante (30.0/7.0)
|  |  |  REDD_LORDO_2007 > 56492: non_interessante (257.0/25.0)
|  COSTI_ACQ_MP > 255305: interessante (27.0/9.0)

Number of Leaves  :     33

Size of the tree :      50


J48 pruned tree (6)
------------------

IMP_REDD_LRD_SMPL_2007 <= -7397
|  IMP_V_AGG_IVA <= 1830: interessante (43.0/8.0)
|  IMP_V_AGG_IVA > 1830: non_interessante (18.0/3.0)
IMP_REDD_LRD_SMPL_2007 > -7397
|  IMP_VAR_RIM_PF <= 1578
|  |  REDD_LORDO_2007 <= 52400
|  |  |  OP_NON_IMPO_DI <= 5981
|  |  |  |  IMP_REDD_LRD_ORD <= -2340: non_interessante (31.0)
|  |  |  |  IMP_REDD_LRD_ORD > -2340
|  |  |  |  |  QTA_PART_IVA <= 2
|  |  |  |  |  |  TOT_IVA_OPE_IMPO <= 1240
|  |  |  |  |  |  |  ALIQ_MEDIA_ACQ <= 19.01
|  |  |  |  |  |  |  |  SIGLA_PROVINCIA = SI: non_interessante (16.0)
|  |  |  |  |  |  |  |  SIGLA_PROVINCIA = PT: non_interessante (1.0)
|  |  |  |  |  |  |  |  SIGLA_PROVINCIA = LU: non_interessante (0.0)
|  |  |  |  |  |  |  |  SIGLA_PROVINCIA = AR: interessante (13.0/1.0)
|  |  |  |  |  |  |  |  SIGLA_PROVINCIA = GR: non_interessante (3.0)
|  |  |  |  |  |  |  |  SIGLA_PROVINCIA = MS: interessante (1.0)
|  |  |  |  |  |  |  |  SIGLA_PROVINCIA = PI: interessante (7.0/3.0)
|  |  |  |  |  |  |  |  SIGLA_PROVINCIA = FI: non_interessante (19.0/2.0)
|  |  |  |  |  |  |  |  SIGLA_PROVINCIA = LI: non_interessante (3.0)
|  |  |  |  |  |  |  |  SIGLA_PROVINCIA = PO: non_interessante (1.0)
|  |  |  |  |  |  |  ALIQ_MEDIA_ACQ > 19.01: interessante (26.0)
|  |  |  |  |  |  TOT_IVA_OPE_IMPO > 1240
|  |  |  |  |  |  |  IMP_V_AGG_IVA <= 187103
|  |  |  |  |  |  |  |  IMP_REDD_LRD_ORD <= 7324
|  |  |  |  |  |  |  |  |  SIGLA_PROVINCIA = SI
|  |  |  |  |  |  |  |  |  |  IMP_V_AGG_IMPON <= 26454
|  |  |  |  |  |  |  |  |  |  |  ALIQ_MEDIA_ACQ <= 19.42: interessante (24.0/8.0)
|  |  |  |  |  |  |  |  |  |  |  ALIQ_MEDIA_ACQ > 19.42: non_interessante (18.0/4.0)
|  |  |  |  |  |  |  |  |  |  IMP_V_AGG_IMPON > 26454: non_interessante (21.0)
|  |  |  |  |  |  |  |  |  SIGLA_PROVINCIA = PT: non_interessante (19.0/5.0)
|  |  |  |  |  |  |  |  |  SIGLA_PROVINCIA = LU: non_interessante (80.0/8.0)
|  |  |  |  |  |  |  |  |  SIGLA_PROVINCIA = AR: non_interessante (19.0/6.0)
|  |  |  |  |  |  |  |  |  SIGLA_PROVINCIA = GR: non_interessante (47.0/4.0)
|  |  |  |  |  |  |  |  |  SIGLA_PROVINCIA = MS: non_interessante (28.0/5.0)
|  |  |  |  |  |  |  |  |  SIGLA_PROVINCIA = PI: non_interessante (52.0/5.0)
|  |  |  |  |  |  |  |  |  SIGLA_PROVINCIA = FI
|  |  |  |  |  |  |  |  |  |  TOT_IMPST_CRED <= 3921
```

```
| | | | | | | | | | | TOT_ACQ <= 115541
| | | | | | | | | | | | IMP_TOT_COMP_NEG_2007 <= 37267
| | | | | | | | | | | | | IMP_V_AGG_IMPON <= 13901: non_interessante (37.0/7.0)
| | | | | | | | | | | | | IMP_V_AGG_IMPON > 13901: interessante (47.0/7.0)
| | | | | | | | | | | | IMP_TOT_COMP_NEG_2007 > 37267: non_interessante (33.0)
| | | | | | | | | | | TOT_ACQ > 115541: interessante (20.0/3.0)
| | | | | | | | | | TOT_IMPST_CRED > 3921: non_interessante (28.0)
| | | | | | | | | SIGLA_PROVINCIA = LI: non_interessante (76.0/9.0)
| | | | | | | | | SIGLA_PROVINCIA = PO: non_interessante (25.0/8.0)
| | | | | | | | IMP_REDD_LRD_ORD > 7324: non_interessante (99.0/5.0)
| | | | | | | IMP_V_AGG_IVA > 187103: interessante (25.0/8.0)
| | | | | QTA_PART_IVA > 2
| | | | | | ALIQ_MEDIA_CESS <= 9.61: non_interessante (15.0)
| | | | | | ALIQ_MEDIA_CESS > 9.61: interessante (29.0/7.0)
| | | OP_NON_IMPO_DI > 5981
| | | | TOT_IMPST_DOV <= 1: non_interessante (15.0/5.0)
| | | | TOT_IMPST_DOV > 1: interessante (15.0)
| | REDD_LORDO_2007 > 52400: non_interessante (266.0/20.0)
| IMP_VAR_RIM_PF > 1578: interessante (27.0/11.0)

Number of Leaves  :     38

Size of the tree :      59


J48 pruned tree (7)
------------------

IMP_REDD_LRD_ORD <= 121364
| OP_ESENTI <= 68395
| | IMP_REDD_LAV_AUT_2007 <= 29070
| | | PROF_COMP <= 38911
| | | | IMP_VAR_RIM_PF <= 1430
| | | | | RICAVI_ATT_2007 <= 14900
| | | | | | IMP_VE_VOLAFF_2007 <= 1200: non_interessante (15.0)
| | | | | | IMP_VE_VOLAFF_2007 > 1200
| | | | | | | IMP_V_AGG_IVA <= 4076: interessante (53.0/5.0)
| | | | | | | IMP_V_AGG_IVA > 4076
| | | | | | | | IMP_VE_VOLAFF_2007 <= 10387: non_interessante (23.0/2.0)
| | | | | | | | IMP_VE_VOLAFF_2007 > 10387: interessante (27.0/11.0)
| | | | | RICAVI_ATT_2007 > 14900
| | | | | | IMP_SPS_DIPEND_SMP <= 1404
| | | | | | | IMPST_CRED <= 1404
| | | | | | | | ETA <= 38
| | | | | | | | | REDD_LORDO_2007 <= 32509
| | | | | | | | | | IMP_REDD_LRD_SMPL_2007 <= 19519: non_interessante (54.0/15.0)
| | | | | | | | | | IMP_REDD_LRD_SMPL_2007 > 19519: interessante (15.0/4.0)
| | | | | | | | | REDD_LORDO_2007 > 32509: non_interessante (19.0/1.0)
| | | | | | | | ETA > 38: non_interessante (374.0/42.0)
| | | | | | | IMPST_CRED > 1404
| | | | | | | | REDD_IMP_2007 <= 17952: interessante (24.0/7.0)
| | | | | | | | REDD_IMP_2007 > 17952: non_interessante (49.0/11.0)
| | | | | | IMP_SPS_DIPEND_SMP > 1404
| | | | | | | OP_ESENTI <= 640
| | | | | | | | TOT_PASS_2007 <= 19636: interessante (29.0/4.0)
| | | | | | | | TOT_PASS_2007 > 19636
| | | | | | | | | OP_NON_IMPO_DI <= 3301
| | | | | | | | | | FLG_PRES_DIP_07 <= 0: interessante (31.0/10.0)
| | | | | | | | | | FLG_PRES_DIP_07 > 0
| | | | | | | | | | | ALIQ_MEDIA_ACQ <= 13.27: non_interessante (24.0)
| | | | | | | | | | | ALIQ_MEDIA_ACQ > 13.27
| | | | | | | | | | | | RIM_FIN_SMPL <= 54027
| | | | | | | | | | | | | TOT_ACQ <= 17851: non_interessante (20.0)
| | | | | | | | | | | | | TOT_ACQ > 17851
| | | | | | | | | | | | | | IMP_SPS_DIPEND_SMP <= 7951: non_interessante (23.0/2.0)
| | | | | | | | | | | | | | IMP_SPS_DIPEND_SMP > 7951
| | | | | | | | | | | | | | | ACQ_NO_DETR <= 37: interessante (34.0/5.0)
| | | | | | | | | | | | | | | ACQ_NO_DETR > 37
| | | | | | | | | | | | | | | | BENI_DEST_RIV <= 110965
| | | | | | | | | | | | | | | | | ACQ_ESENTI <= 7: non_interessante (23.0/7.0)
| | | | | | | | | | | | | | | | | ACQ_ESENTI > 7: interessante (29.0/6.0)
| | | | | | | | | | | | | | | | BENI_DEST_RIV > 110965: non_interessante (15.0)
| | | | | | | | | | | | | | | RIM_FIN_SMPL > 54027: non_interessante (19.0)
| | | | | | | | | | | | OP_NON_IMPO_DI > 3301: interessante (21.0/4.0)
| | | | | | | OP_ESENTI > 640: non_interessante (16.0/1.0)
| | | | IMP_VAR_RIM_PF > 1430
| | | | | ALIQ_MEDIA_CESS <= 12.82: interessante (17.0/1.0)
| | | | | ALIQ_MEDIA_CESS > 12.82: non_interessante (20.0/7.0)
| | | PROF_COMP > 38911: interessante (28.0/8.0)
| | IMP_REDD_LAV_AUT_2007 > 29070: non_interessante (133.0/11.0)
| OP_ESENTI > 68395: non_interessante (78.0/4.0)
IMP_REDD_LRD_ORD > 121364: non_interessante (34.0)
```

```
Number of Leaves  :     28

Size of the tree :      55


J48 pruned tree (8)
------------------

REDD_LORDO_2007 <= 56492
|  IMP_REDD_IMP_SMPL_2007 <= 40013
|  |  ALTRI_ACQ_IMP <= 182925
|  |  |  FLG_PRES_FAM_07 <= 0
|  |  |  |  REDD_IMP_2007 <= 934
|  |  |  |  |  IMP_RICAVI_ORDIN_2007 <= 0
|  |  |  |  |  |  IMP_SPS_DIPEND_SMP <= 17613: interessante (65.0/13.0)
|  |  |  |  |  |  IMP_SPS_DIPEND_SMP > 17613: non_interessante (23.0/6.0)
|  |  |  |  |  IMP_RICAVI_ORDIN_2007 > 0: non_interessante (19.0)
|  |  |  |  REDD_IMP_2007 > 934
|  |  |  |  |  QTA_PART_IVA <= 2
|  |  |  |  |  |  IMP_REDD_LRD_SMPL_2007 <= 34947
|  |  |  |  |  |  |  ALIQ_MEDIA_ACQ <= 19.99
|  |  |  |  |  |  |  |  CRED_ANNO_PREC <= 13486: non_interessante (620.0/129.0)
|  |  |  |  |  |  |  |  CRED_ANNO_PREC > 13486: interessante (26.0/11.0)
|  |  |  |  |  |  |  ALIQ_MEDIA_ACQ > 19.99
|  |  |  |  |  |  |  |  IMP_V_AGG_IVA <= 9675: interessante (18.0/1.0)
|  |  |  |  |  |  |  |  IMP_V_AGG_IVA > 9675
|  |  |  |  |  |  |  |  |  FLG_NO_CONGRUENZA_07 <= 0: non_interessante (22.0/1.0)
|  |  |  |  |  |  |  |  |  FLG_NO_CONGRUENZA_07 > 0: interessante (23.0/9.0)
|  |  |  |  |  |  IMP_REDD_LRD_SMPL_2007 > 34947: interessante (26.0/9.0)
|  |  |  |  |  QTA_PART_IVA > 2
|  |  |  |  |  |  ALIQ_MEDIA_ACQ <= 18.82: non_interessante (15.0/4.0)
|  |  |  |  |  |  ALIQ_MEDIA_ACQ > 18.82: interessante (20.0/5.0)
|  |  |  FLG_PRES_FAM_07 > 0: non_interessante (40.0/1.0)
|  |  ALTRI_ACQ_IMP > 182925
|  |  |  IMP_REDD_LRD_SMPL_2007 <= 0
|  |  |  |  ETA <= 52: non_interessante (20.0/3.0)
|  |  |  |  ETA > 52: interessante (18.0/5.0)
|  |  |  IMP_REDD_LRD_SMPL_2007 > 0: interessante (19.0/1.0)
|  IMP_REDD_IMP_SMPL_2007 > 40013: non_interessante (26.0)
REDD_LORDO_2007 > 56492: non_interessante (247.0/22.0)

Number of Leaves  :     17

Size of the tree :      33


J48 pruned tree (9)
------------------

OP_ESENTI <= 68395
|  IMP_V_AGG_IVA <= -2320
|  |  IMP_REDD_IMP_SMPL_2007 <= 20607
|  |  |  IMP_BEN_STRUM_NA <= 3700
|  |  |  |  ALIQ_MEDIA_ACQ <= 12.93: interessante (23.0)
|  |  |  |  ALIQ_MEDIA_ACQ > 12.93
|  |  |  |  |  ACQ_ESENTI <= 0
|  |  |  |  |  |  TOT_ACQ <= 93239: interessante (30.0/9.0)
|  |  |  |  |  |  TOT_ACQ > 93239: non_interessante (16.0/1.0)
|  |  |  |  |  ACQ_ESENTI > 0: interessante (21.0/3.0)
|  |  |  IMP_BEN_STRUM_NA > 3700: non_interessante (16.0/3.0)
|  |  IMP_REDD_IMP_SMPL_2007 > 20607: non_interessante (26.0/2.0)
|  IMP_V_AGG_IVA > -2320
|  |  IMPST_CRED <= 1973
|  |  |  IMPST_CRED <= 206
|  |  |  |  TOT_IMPST_CRED <= 4739
|  |  |  |  |  ACQ_ESENTI <= 3102
|  |  |  |  |  |  IMP_REDD_LAV_AUT_2007 <= 28553
|  |  |  |  |  |  |  SIGLA_PROVINCIA = SI
|  |  |  |  |  |  |  |  ACQ_ESENTI <= 4: non_interessante (65.0/4.0)
|  |  |  |  |  |  |  |  ACQ_ESENTI > 4: interessante (18.0/5.0)
|  |  |  |  |  |  |  SIGLA_PROVINCIA = PT: non_interessante (26.0/13.0)
|  |  |  |  |  |  |  SIGLA_PROVINCIA = LU: non_interessante (94.0/11.0)
|  |  |  |  |  |  |  SIGLA_PROVINCIA = AR
|  |  |  |  |  |  |  |  FLG_PRES_FAM_07 <= 0
|  |  |  |  |  |  |  |  |  DURATA_ATTIV_SOGG <= 10: non_interessante (22.0/4.0)
|  |  |  |  |  |  |  |  |  DURATA_ATTIV_SOGG > 10: interessante (24.0/10.0)
|  |  |  |  |  |  |  |  FLG_PRES_FAM_07 > 0: non_interessante (20.0)
|  |  |  |  |  |  |  SIGLA_PROVINCIA = GR: non_interessante (50.0/8.0)
|  |  |  |  |  |  |  SIGLA_PROVINCIA = MS: non_interessante (26.0/9.0)
|  |  |  |  |  |  |  SIGLA_PROVINCIA = PI: non_interessante (54.0/9.0)
|  |  |  |  |  |  |  SIGLA_PROVINCIA = FI
|  |  |  |  |  |  |  |  RIM_FIN_SMPL <= 4488
```

```
| | | | | | | | | GRP_ATTIV_2007 = F: non_interessante (2.0/1.0)
| | | | | | | | | GRP_ATTIV_2007 = Q: non_interessante (0.0)
| | | | | | | | | GRP_ATTIV_2007 = I: non_interessante (5.0)
| | | | | | | | | GRP_ATTIV_2007 = G: interessante (22.0/9.0)
| | | | | | | | | GRP_ATTIV_2007 = L: non_interessante (32.0/3.0)
| | | | | | | | | GRP_ATTIV_2007 = H: interessante (6.0/1.0)
| | | | | | | | | GRP_ATTIV_2007 = A: non_interessante (8.0/4.0)
| | | | | | | | | GRP_ATTIV_2007 = S: non_interessante (2.0)
| | | | | | | | | GRP_ATTIV_2007 = M: interessante (26.0/7.0)
| | | | | | | | | GRP_ATTIV_2007 = C: interessante (44.0/12.0)
| | | | | | | | | GRP_ATTIV_2007 = K: non_interessante (8.0)
| | | | | | | | | GRP_ATTIV_2007 = N: non_interessante (0.0)
| | | | | | | | | GRP_ATTIV_2007 = R: non_interessante (0.0)
| | | | | | | | | GRP_ATTIV_2007 = J: interessante (1.0)
| | | | | | | | | GRP_ATTIV_2007 = P: non_interessante (0.0)
| | | | | | | | | GRP_ATTIV_2007 = E: non_interessante (0.0)
| | | | | | | | | GRP_ATTIV_2007 = B: non_interessante (0.0)
| | | | | | | | RIM_FIN_SMPL > 4488: non_interessante (36.0/1.0)
| | | | | | | SIGLA_PROVINCIA = LI: non_interessante (56.0/8.0)
| | | | | | | SIGLA_PROVINCIA = PO: non_interessante (25.0/11.0)
| | | | | | IMP_REDD_LAV_AUT_2007 > 28553: non_interessante (102.0/8.0)
| | | | | ACQ_ESENTI > 3102: interessante (28.0/11.0)
| | | | TOT_IMPST_CRED > 4739: non_interessante (79.0/2.0)
| | | IMPST_CRED > 206: non_interessante (44.0)
| | IMPST_CRED > 1973
| | | IMPST_CRED <= 3701: interessante (19.0)
| | | IMPST_CRED > 3701
| | | | ACQ_NO_DETR <= 2614
| | | | | IMP_SPS_DIPEND_ORD_2007 <= 12552
| | | | | | IMP_V_AGG_IVA <= 65845: non_interessante (24.0/2.0)
| | | | | | IMP_V_AGG_IVA > 65845: interessante (16.0/6.0)
| | | | | IMP_SPS_DIPEND_ORD_2007 > 12552: non_interessante (16.0)
| | | | ACQ_NO_DETR > 2614: interessante (20.0/4.0)
OP_ESENTI > 68395: non_interessante (95.0/4.0)

Number of Leaves  :     46

Size of the tree :      68


J48 pruned tree (10)
------------------

IMP_VE_VOLAFF_2007 <= 17730
|  COSTI_RSDL <= 1643
|  |  ACQ_NO_DETR <= 1
|  |  |  TOT_IMPST_DOV <= 2
|  |  |  |  IMP_VE_VOLAFF_2007 <= 11016: interessante (21.0/9.0)
|  |  |  |  IMP_VE_VOLAFF_2007 > 11016: non_interessante (18.0/1.0)
|  |  |  TOT_IMPST_DOV > 2: interessante (23.0/3.0)
|  |  ACQ_NO_DETR > 1: non_interessante (27.0/2.0)
|  COSTI_RSDL > 1643: interessante (53.0/7.0)
IMP_VE_VOLAFF_2007 > 17730
|  IMP_REDD_LAV_AUT_2007 <= 51360
|  |  IMP_VAR_RIM_MP <= -4859: non_interessante (52.0/2.0)
|  |  IMP_VAR_RIM_MP > -4859
|  |  |  TOT_PASS_2007 <= 83361
|  |  |  |  OP_ESENTI <= 68395
|  |  |  |  |  IMP_COSTI_BEN_TRZ <= 923
|  |  |  |  |  |  COSTI_ACQ_MP <= 32265
|  |  |  |  |  |  |  SIGLA_PROVINCIA = SI: non_interessante (63.0/11.0)
|  |  |  |  |  |  |  SIGLA_PROVINCIA = PT: non_interessante (31.0/2.0)
|  |  |  |  |  |  |  SIGLA_PROVINCIA = LU: non_interessante (60.0/12.0)
|  |  |  |  |  |  |  SIGLA_PROVINCIA = AR: non_interessante (16.0/6.0)
|  |  |  |  |  |  |  SIGLA_PROVINCIA = GR: non_interessante (30.0/2.0)
|  |  |  |  |  |  |  SIGLA_PROVINCIA = MS: non_interessante (14.0/5.0)
|  |  |  |  |  |  |  SIGLA_PROVINCIA = PI: non_interessante (43.0/6.0)
|  |  |  |  |  |  |  SIGLA_PROVINCIA = FI
|  |  |  |  |  |  |  |  COSTI_RSDL <= 10943
|  |  |  |  |  |  |  |  |  ALIQ_MEDIA_ACQ <= 16.39: interessante (29.0/4.0)
|  |  |  |  |  |  |  |  |  ALIQ_MEDIA_ACQ > 16.39
|  |  |  |  |  |  |  |  |  |  IMP_CMPNS_ATTIV_2007 <= 8795: interessante (34.0/9.0)
|  |  |  |  |  |  |  |  |  |  IMP_CMPNS_ATTIV_2007 > 8795: non_interessante (15.0)
|  |  |  |  |  |  |  |  COSTI_RSDL > 10943: non_interessante (40.0/4.0)
|  |  |  |  |  |  |  SIGLA_PROVINCIA = LI: non_interessante (35.0/7.0)
|  |  |  |  |  |  |  SIGLA_PROVINCIA = PO: non_interessante (23.0/10.0)
|  |  |  |  |  |  COSTI_ACQ_MP > 32265: non_interessante (28.0)
|  |  |  |  |  IMP_COSTI_BEN_TRZ > 923: non_interessante (46.0/1.0)
|  |  |  |  OP_ESENTI > 68395: non_interessante (41.0)
|  |  |  TOT_PASS_2007 > 83361
|  |  |  |  PRES_FAM_SEMPL_07 <= 0
|  |  |  |  |  CESS_NON_IMPO <= 0
```

```
| | | | | | IMP_REDD_LRD_SMPL_2007 <= 28772
| | | | | | | IMP_REDD_LRD_ORD <= -20584: non_interessante (15.0)
| | | | | | | IMP_REDD_LRD_ORD > -20584
| | | | | | | | ACQ_NO_DETR <= 5819
| | | | | | | | | IMP_SPS_DIPEND_SMP <= 34624
| | | | | | | | | | IMP_SPS_DIPEND_SMP <= 17943
| | | | | | | | | | | ACQ_NO_DETR <= 1652
| | | | | | | | | | | | IMPST_DOV <= 15475
| | | | | | | | | | | | | IMP_REDD_PERD_2007 <= 2074: interessante (21.0/5.0)
| | | | | | | | | | | | | IMP_REDD_PERD_2007 > 2074: non_interessante (75.0/13.0)
| | | | | | | | | | | | IMPST_DOV > 15475: interessante (19.0/1.0)
| | | | | | | | | | | ACQ_NO_DETR > 1652
| | | | | | | | | | | | IMP_COSTI_BEN_TRZ <= 23028: interessante (38.0/1.0)
| | | | | | | | | | | | IMP_COSTI_BEN_TRZ > 23028: non_interessante (18.0/8.0)
| | | | | | | | | | IMP_SPS_DIPEND_SMP > 17943: interessante (23.0)
| | | | | | | | | IMP_SPS_DIPEND_SMP > 34624: non_interessante (26.0/4.0)
| | | | | | | | ACQ_NO_DETR > 5819: non_interessante (28.0/3.0)
| | | | | | IMP_REDD_LRD_SMPL_2007 > 28772: non_interessante (67.0/7.0)
| | | | | CESS_NON_IMPO > 0: non_interessante (25.0/2.0)
| | | | PRES_FAM_SEMPL_07 > 0
| | | | | IMPST_DOV <= 1000: interessante (24.0)
| | | | | IMPST_DOV > 1000
| | | | | | RIM_FIN_SMPL <= 5350: interessante (26.0/8.0)
| | | | | | RIM_FIN_SMPL > 5350: non_interessante (16.0)
| IMP_REDD_LAV_AUT_2007 > 51360: non_interessante (84.0/3.0)

Number of Leaves  :     37

Size of the tree :     65
```

Mostriamo ora gli usuali *lift charts* relativi al modello appena descritto. All'*outlier* correttamente classificato fino ad ora solo dal modello C4.5 utilizzato singolarmente viene assegnata una bassa probabilità di essere evasore *interessante*, per cui non viene proposto per un controllo. Nel caso specifico, in cui il modello propone, effettivamente, un esiguo numero di controlli, il *lift chart* può fornire una ulteriore guida per la pianificazione degli *audit*, suggerendo se e in che misura assoggettare a controllo un maggior numero di soggetti. Infatti, sulla base del *lift chart,* è possibile stimare il gettito recuperabile da un numero dato di controlli e valutarne la fattibilità tenuto conto delle risorse disponibili e degli obbiettivi prefissati dall'ente controllante.

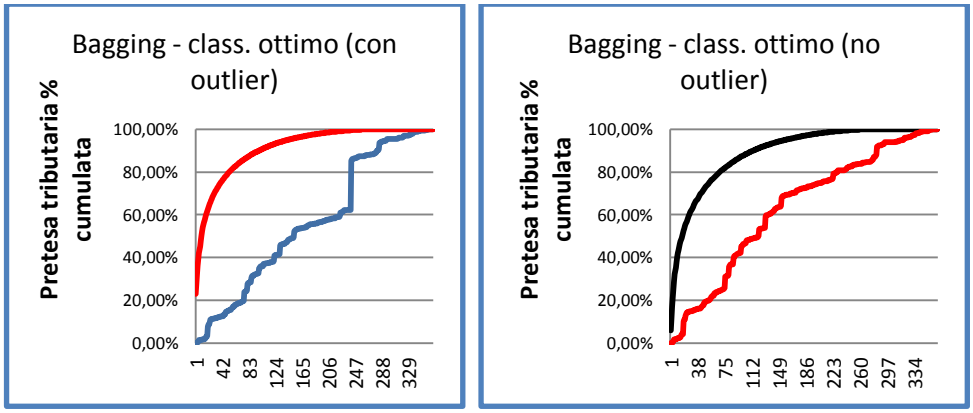

Figura 5.14 confronto *bagging* – classificatore ottimo

L'area sotto la curva è pari al 61,72% del massimo possibile se si considera l'*outlier* nel *test set*, mentre è pari al 70,57% in caso contrario.

Per quanto concerne le variabili che entrano in gioco nella costruzione degli alberi delle varie iterazioni, vale la pena di osservare, come nel caso precedente di *boosting*, che gli alberi stessi mediamente hanno molti nodi e foglie, per cui la loro interpretabilità non è immediata. Inoltre, essendo la predizione della classe di ogni istanza una media ottenuta con le risultanze dei diversi alberi generati, l'individuazione di *pattern* che possano rivelare la presenza di evasori interessanti non appare intuitiva e immediata.

5.3.1.4 Costi nella generazione dei modelli

Dopo aver visto i risultati di due classificatori classici (C4.5 e CART) e di alcuni *ensemble methods* noti in letteratura[7], introduciamo un'altra tecnica ampiamente utilizzata per cercare di migliorare le *performance* dei classificatori.

Utilizzeremo in questo caso il *metaclassificatore Metacost* implementato in *Weka*, combinato con *bagging* di alberi *J48* (quest'ultimo con i parametri visti in precedenza), previo *resample* del *training set* (*biasToUniformClass* 0.1). L'implementazione impiegata è quella di [Dom99] e vengono utilizzati i seguenti parametri:

bagSizePercent -- The size of each bag, as a percentage of the training set size.	*100*
classifier -- The base classifier to be used.	*Bagging – J48*
costMatrix -- A misclassification cost matrix.	*0* | *x* *y* | *0*|*-z*

In questo contesto, l'attribuzione dei costi ai due diversi tipi di errore che il modello può commettere risulta cruciale ai fini dei risultati attesi del classificatore.

Infatti, sulla base della differenza più o meno marcata tra errori di tipo *FP* o *FN*, si ottengono classificatori più o meno "prudenti" nel predire una classe piuttosto che l'altra, in quanto a seconda del settaggio sul peso degli errori il modello creato tenderà a minimizzare i *FP* o i *FN*.

Un controllo preliminare da effettuare è vedere empiricamente in che modo variano le predizioni del modello al variare dei pesi. I risultati ottenuti sono facilmente interpretabili:

[7] Non sono state presentate le tecniche *random forest* o *rotation forest* perché, da un lato non presentano *performances* migliori sui dati di *test*, dall'altro, i risultati ottenuti sono difficilmente interpretabili.

Matrici dei costi	***Matrici di confusione***

$\begin{bmatrix} pred \rightarrow & N & P \\ actual \downarrow & & \\ N & 0 & 1 \\ P & 5 & 0 \end{bmatrix}$ $\rightarrow$	`  N   P   <-- classified as`
	`121 160 |   N = non_interessante`
	`19  67 |   P = interessante`

$\begin{bmatrix} pred \rightarrow & N & P \\ actual \downarrow & & \\ N & 0 & 2 \\ P & 4 & 0 \end{bmatrix}$ $\rightarrow$	`  N   P   <-- classified as`
	`231  50 |   N = non_interessante`
	`54  32 |   P = interessante`

$\begin{bmatrix} pred \rightarrow & N & P \\ actual \downarrow & & \\ N & 0 & 3 \\ P & 3 & 0 \end{bmatrix}$ $\rightarrow$	`  N   P   <-- classified as`
	`268  13 |   N = non_interessante`
	`75  11 |   P = interessante`

$\begin{bmatrix} pred \rightarrow & N & P \\ actual \downarrow & & \\ N & 0 & 4 \\ P & 2 & 0 \end{bmatrix}$ $\rightarrow$	`  N   P   <-- classified as`
	`278   3 |   N = non_interessante`
	`85   1 |   P = interessante`

$\begin{bmatrix} pred \rightarrow & N & P \\ actual \downarrow & & \\ N & 0 & 5 \\ P & 1 & 0 \end{bmatrix}$ $\rightarrow$	`  N   P   <-- classified as`
	`281   0 |   N = non_interessante`
	`86   0 |   P = interessante`

Il valore più opportuno da impostare per la generazione dei vari modelli di costo naturalmente dipende dal contesto e, in particolare, dalle risorse a disposizione dell'ente accertatore: nel primo caso:

$$\begin{bmatrix} pred \rightarrow & N & P \\ actual \downarrow & & \\ N & 0 & 1 \\ P & 5 & 0 \end{bmatrix}$$

in cui si minimizza il peso dei falsi positivi rispetto ai falsi negativi, il modello tenderà a classificare le istanze come positive, il che vuol dire che proporrà molti controlli, anche nei confronti di soggetti che in realtà non sarebbero meritevoli di controllo. Allo stesso tempo, però, il modello sarà molto preciso quando proporrà di *non* controllare un determinato soggetto. La strategia di attribuzione dei pesi appena descritta può essere adeguata nel caso in cui si ritenga che perdere anche un solo soggetto interessante sia da considerarsi un errore grave: una tale situazione presuppone allora un numero complessivo di soggetti fraudolenti sul territorio alquanto basso, che devono essere accertati quanto più è possibile. Tale scenario appare tuttavia poco plausibile nella realtà. Non solo, ma seguire i suggerimenti del modello, dati i pesi così impostati, presupporrebbe anche un ente accertatore con ampi mezzi a disposizione, che possa permettersi di effettuare molti controlli, accettando il rischio che molti di essi possano rivelarsi rivolti a evasori *non interessanti*.

Nel caso opposto, rappresentato dalla seguente matrice di costi:

$$\begin{bmatrix} pred \rightarrow \\ actual \downarrow \end{bmatrix} \begin{matrix} N & P \end{matrix}$$
$$\begin{bmatrix} N \\ P \end{bmatrix} \begin{bmatrix} 0 & 5 \\ 1 & 0 \end{bmatrix}$$

l'errore di falso positivo è molto penalizzato, per cui il modello cercherà di essere ben sicuro prima di selezionare un soggetto per un controllo: di conseguenza, ne saranno proposti pochi (al limite, anche nessuno). Lo scenario in cui una tale strategia appare ragionevole è quello in cui le risorse dell'ente accertatore siano limitate e vadano impiegate in modo ottimale. Inevitabilmente, molti soggetti non saranno oggetto di controllo e si registreranno molti falsi negativi, ma ciò non costituisce un problema nella misura in cui le risorse dell'ente siano comunque pienamente impiegate. Tale scenario appare essere molto più aderente alla realtà rispetto al precedente.

Qualora si intendesse premiare un *true positive* rispetto ad un *true negative* (immaginando che, essendo più difficile individuare un vero evasore rispetto a un non evasore, la predizione corretta di un evasore sia più "preziosa"), si può immaginare di mettere un valore negativo in corrispondenza del TP. Ciò indurrà il modello ad aumentare il numero dei soggetti ritenuti meritevoli di attenzione rispetto al caso in cui il peso di un TP fosse stato 0.

Ad esempio, si ottiene il seguente risultato:

$$\begin{bmatrix} pred \rightarrow \\ actual \downarrow \end{bmatrix} \begin{matrix} N & P \end{matrix}$$
$$\begin{bmatrix} N \\ P \end{bmatrix} \begin{bmatrix} 0 & 5 \\ 1 & -3 \end{bmatrix} \rightarrow$$

```
   N    P   <-- classified as
 275    6 |   N = non_interessante
  77    9 |   P = interessante
```

che, confrontato con il corrispondente modello con errore TP pari a zero, che non avrebbe individuato alcun soggetto interessante, ne individua 9, suggerendo 15 controlli in totale. In questo modello, per la prima volta, si osserva come la maggioranza dei soggetti ritenuti interessanti, effettivamente si dimostra effettivamente tale. Altre metriche sono:

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.979	0.895	0.781	0.979	0.869	0.578	Non interessante
	0.105	0.021	0.60	0.105	0.178	0.578	Interessante
Weig. Avg.	0.774	0.691	0.739	0.774	0.707	0.578	

Inoltre, come per i modelli già visti, si riportano le usuali misure di carattere "monetario", con l'avvertenza che anche in questo caso, il modello non seleziona il soggetto *outlier*.

- *recupero (15)* = **€ 425.554,00**
- *recupero interessanti (9)* = **€ 406.178,00**
- *recupero non interessanti (6)* = **€ 19.376,00**
- *recupero (media_15)* = **€ 28.370,00**

- *recupero (mediana_15)* = **€ 4.690,00**

Il modello, pur suggerendo relativamente pochi controlli, riesce, dopo aver selezionato circa il 4% dei soggetti accertabili, a recuperare l'8% del gettito (o il 10, se non si considera l'*outlier*).

I soggetti selezionati sono quelli che producono il recupero medio più alto tra i modelli visti sin ora e la percentuale di soggetti *interessanti* sul totale di quelli selezionati è la massima finora riscontrata, pur non essendo, in valore assoluto, elevatissima (60%). Risultato che appare comunque pregevole, considerato che i soggetti non interessanti occupano circa ¾ del *dataset*.

Il modello, come visto, risente fortemente della struttura dei pesi degli errori (o meglio, del rapporto dei pesi delle varie tipologie di errore). L'introduzione del premio per un TP apporta un miglioramento nel modello, che riesce ad essere più selettivo rispetto al caso in cui si utilizzi una matrice di costo classica, in cui il costo di $TP = TN = 0$. Ad esempio, per ottenere almeno 9 successi, in quest'ultimo caso, il numero di controlli da eseguire sarebbe stato superiore:

$$\begin{bmatrix} pred \rightarrow & N & P \\ actual \downarrow & & \\ N & 0 & 3 \\ P & 3 & 0 \end{bmatrix} \rightarrow$$

```
   N    P   <-- classified as
 268   13 |   N = non_interessante
  75   11 |   P = interessante
```

La combinazione di utilizzo di pesi diversi a seconda della tipologia di errore e tecniche di *bagging* sembra dare risultati buoni, e la precisione dimostrata potrebbe rilevarsi particolarmente utile in contesti reali, in considerazione del fatto che normalmente la platea dei soggetti da controllare è molto ampia, per cui avere una metodologia che permette di selezionarne solo il 4/5%, con un successo del 60% appare un ottimo risultato (si consideri che le selezioni reali effettuate sul *test set* hanno portato a controllare 367 soggetti, di cui solo un quarto proficui).

Mostriamo ora gli usuali *lift charts* relativi al modello appena descritto. Anche nel caso specifico, in cui il modello propone, effettivamente, un esiguo numero di controlli, il *lift chart* può fornire una ulteriore guida per la pianificazione degli *audit*, suggerendo se e in che misura assoggettare a controllo un maggior numero di soggetti. Infatti, sulla base del *lift chart,* è possibile stimare il gettito recuperabile da un numero dato di controlli e valutarne la fattibilità tenuto conto delle risorse disponibili e degli obbiettivi prefissati dall'ente controllante.

L'area sotto la curva è pari al 66,99% del massimo possibile se si considera l'*outlier* nel *test set*, mentre è pari al 64,03% in caso contrario. Nonostante quindi la precisione dimostrata sui casi ritenuti accertabili, peraltro caratterizzati da un buon recupero medio, l'area sotto la curva pare non essere migliore rispetto agli altri modelli fin qui considerati.

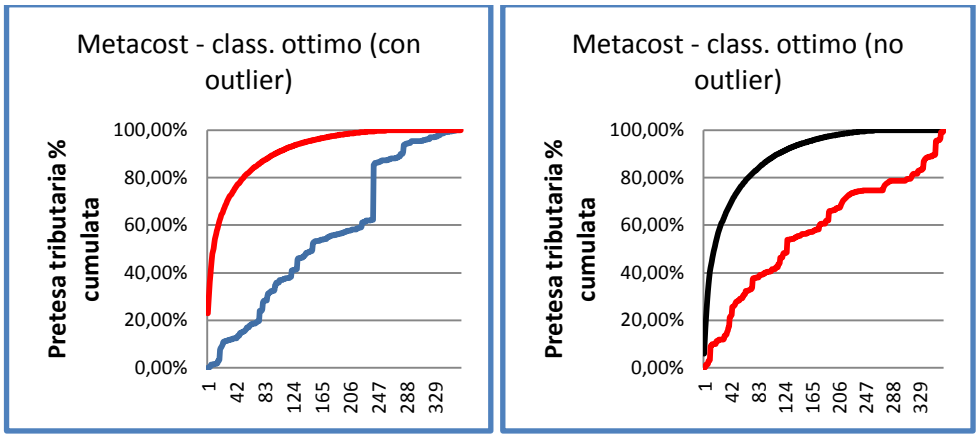

Figura 5.15*: lift chart metacost*

5.3.2 Regole di classificazione

I classificatori a regole classificano i *records* di un *data set* utilizzando insiemi di regole del tipo "*if... then...*"

Ogni regola di classificazione può essere espressa nel seguente modo:

$$r_i = (Condizione_i) \rightarrow y_i$$

dove $(Condizione_i)$ è chiamata anche antecedente o precondizione della regola e rappresenta una congiunzione di predicati su attributi, ovvero:

$$Condizione_i = (A_1 op\ v_1) \wedge (A_2 op\ v_2) \wedge \ldots \wedge (A_k op\ v_k)$$

dove $(A_1 op\ v_1)$ è una coppia attributo-valore e op è un operatore logico appartenente all'insieme $\{=, \neq, <, >, \leq, \geq\}$. La parte destra della regola, y_i, è l'etichetta della classe predetta.

Si richiamano alcuni concetti di base:

Una regola r copre una istanza x se i valori di x soddisfano l'antecedente di r. In questo caso, si dice che x attiva la regola r.

Dato un *dataset* D e una regola di classificazione $A \rightarrow y$, definiamo:

Copertura (*coverage*): Frazione dei record che soddisfano l'antecedente della regola:

$$Coverage\ (r) = \frac{|A|}{|D|}$$

Accuratezza (*accuracy*): Frazione dei record che, soddisfacendo l'antecedente, soddisfano anche il conseguente della regola:

$$Accuracy\ (r) = \frac{|A \cap y|}{|A|}$$

Regole mutuamente esclusive: un insieme di regole R è detto mutuamente esclusivo se nessuna coppia di regole può essere attivata dallo stesso record e quindi ogni record è coperto da al più una regola;

Regole esaustive: un insieme di regole *R* ha una copertura esaustiva se esiste una regola per ogni combinazione di valori degli attributi, quindi ogni record è coperto da almeno una regola.

La sussistenza contemporanea di queste proprietà assicura che ogni istanza sia coperta esattamente da una sola regola. Non sempre però è possibile determinare un insieme di regole esaustive e mutualmente esclusive. In caso di mancanza di mutua esclusività, un record può attivare più regole dando vita a classificazioni discordanti. Possibili soluzioni di questo problema sono definire un ordine di attivazione delle regole (*decision list*) oppure assegnare il record alla classe per la quale vengono attivate più regole (voto). La mancanza di esaustività porta alla situazione in cui un record può non attivare nessuna regola. Una soluzione è quella di utilizzare una classe di *default* ("altro") a cui il *record* viene associato in caso di non attivazione delle regole.

Quali sono le modalità di ordinamento di un insieme di regole?

Ordinamento Rule-based: le singole regole sono ordinate in base a una qualche misura della loro qualità

Ordinamento Class-based: l'ordinamento rilevante diventa quello tra le classi, che può dipendere dall'importanza della classe o dalla gravità di commettere un errore di classificazione per quella classe. Le regole che predicono la stessa classe non sono ordinate (proprio perché, predicendo la stessa classe, non entrano in conflitto tra loro).

Il rischio di questa soluzione è che una regola di buona qualità sia superata da una regola di qualità inferiore ma che predice una classe ritenuta più "importante". Quest'ultima soluzione è tuttavia utilizzata dai principali algoritmi di costruzione delle regole (tra cui RIPPER).

Costruire un modello vuol dire estrarre dai dati un *set* di regole che identifichi le relazioni chiave tra gli attributi di un data set e la classe. Vi sono due classi di metodi per l'estrazione di regole di classificazione: *metodi diretti*, che estraggono regole direttamente dai dati e *metodi indiretti* che le estraggono da altri modelli predittivi, quali alberi decisionali o reti neurali.

Ciò premesso, vediamo i risultati da due ben noti algoritmi di apprendimento: RIPPER e PART.

5.3.2.1 Modello RIPPER

Utilizziamo un algoritmo di apprendimento proposizionale chiamato *Repeated Incremental Pruning to Produce Error Reduction* (RIPPER), descritto in [Coh95]. L'algoritmo RIPPER si sviluppa in due fasi: nella prima viene indotto un primo *set* di regole, che viene poi raffinato tramite aggiustamenti delle singole regole al fine di renderle più efficaci nel loro complesso, attraverso un processo di ottimizzazione globale[8].

[8] Come si ritrova in [Coh95]: "*..RIPPER starts with an initial model and iteratively improves it using heuristic techniques..*".

RIPPER implementa una strategia *divide-et-impera,* generando una regola alla volta, rimuovendo le istanze coperte da quella regola e iterativamente inducendo ulteriori regole a partire dalle rimanenti istanze. Diversi metodi di *pruning* sono stati implementati per gli algoritmi *divide-et-impera*: in particolare RIPPER utilizza come criterio il *reduced error pruning* presentato in [Qui87], che lascia alcuni dati di *training* da parte per determinare quando lasciar cadere la coda di una regola, incorporando una euristica basata sul principio di *minimum description length* come criterio di stop.

L'algoritmo viene eseguito con i parametri di seguito evidenziati:

checkErrorRate -- Whether check for error rate >= 1/2 is included in stopping criterion.	*True*
folds -- Determines the amount of data used for pruning. One fold is used for pruning, the rest for growing the rules.	*3*
minNo -- The minimum total weight of the instances in a rule.	*5*
optimizations -- The number of optimization runs.	*2*
seed -- The seed used for randomizing the data.	*1*
usePruning -- Whether pruning is performed.	*True*

Sulla scorta dell'esperienza ricavata dagli esperimenti condotti con gli alberi di classificazione, abbiniamo l'algoritmo a un metaclassificatore ed, in particolare, al *bagging*, che ha mostrato buoni risultati.

In particolare, procedendo ad un preventivo *resampling* del *training set* (*biasToUniformClass* = 0,1), otteniamo i risultati di seguito esposti.

Andando a valutare il modello secondo le ormai usuali metriche, osserviamo, in primo luogo, una buona precisione:

```
Correctly Classified Instances         280               76.2943 %
Incorrectly Classified Instances        87               23.7057 %
```

In secondo luogo, si scopre poi come esso sia estremamente "prudente" nel selezionare i soggetti da accertare, in quanto suggerisce solo 17 controlli su 367 (tale dato peraltro replica il risultato ottenuto in precedenza dallo stesso *ensemble method* applicato agli alberi decisionali). La relativa matrice di confusione viene di seguito riportata:

```
=== Confusion Matrix ===

   a   b   <-- classified as
 272   9 |   a = non_interessante
  78   8 |   b = interessante
```

Altre metriche sono di seguito evidenziate:

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.968	0.907	0.777	0.968	0.862	0.603	Non interessante
	0.093	0.032	0.471	0.093	0.155	0.603	Interessante
Weig. Avg.	0.763	0.702	0.705	0.763	0.697	0.603	

Inoltre, come per i modelli già visti, si riportano le misure di carattere "monetario", con l'avvertenza che anche in questo caso, il modello non seleziona il soggetto *outlier*.

- *recupero (17)* = **€ 493.977,00**
- *recupero interessanti (8)* = **€ 469.143,00**
- *recupero non interessanti (9)* = **€ 24.834,00**
- *recupero (media_17)* = **€ 29.057,00**
- *recupero (mediana_17)* = **€ 4.690,00**

L'output del modello è il seguente:

=== Classifier model ===

Scheme: Bagging

All the base classifiers:

JRIP rules:
===========

(CRED_ANNO_PREC >= 2385) and (REDD_LORDO_2007 <= 27721) and (COSTO_LAV_2007 >= 1577) and (IMP_BEN_AMM <= 5600) and (ACQ_NO_DETR >= 632) => CLASSE=interessante (27.0/1.0)
(ACQ_NO_DETR <= 6) and (TOT_ACQ <= 7737) and (COSTI_RSDL >= 700) and (ALTRI_ACQ_IMP >= 2675) => CLASSE=interessante (31.0/0.0)
(IMP_SPS_DIPEND_SMP >= 6898) and (ACQ_NO_DETR <= 371) and (IMP_TOT_COMP_NEG_2007 <= 144340) and (FLG_NO_COERENZA_07 <= 0) and (CESS_NON_IMPO <= 0) => CLASSE=interessante (28.0/4.0)
(IMP_V_AGG_IVA <= 6734) and (CRED_ANNO_PREC >= 2385) and (ALTRI_ACQ_IMP >= 402) => CLASSE=interessante (40.0/6.0)
(REDD_LORDO_2007 <= 52400) and (SIGLA_PROVINCIA = FI) and (ALIQ_MEDIA_CESS >= 20) and (COSTO_LAV_2007 >= 7343) => CLASSE=interessante (39.0/4.0)
(REDD_LORDO_2007 <= 29135) and (BENI_DEST_RIV <= 23672) and (TOT_IVA_OPE_IMPO >= 9391) and (BENI_DEST_RIV >= 4901) and (IMP_BEN_AMM <= 4829) => CLASSE=interessante (20.0/0.0)
(TOT_PASS_2007 <= 16632) and (CRED_ANNO_PREC <= 0) and (PROF_COMP >= 28707) and (PROF_COSTI <= 8906) => CLASSE=interessante (15.0/0.0)
(IMPST_CRED >= 1456) and (FLG_NO_CONGRUENZA_07 >= 1) and (DURATA_ATTIV_SOGG >= 20) => CLASSE=interessante (22.0/2.0)
(REDD_LORDO_2007 <= 7573) and (IMPST_DOV <= 1270) and (IMP_BEN_AMM >= 4545) and (IMP_CESS_BENI_AMM <= 0) => CLASSE=interessante (12.0/0.0)
(TOT_IVA_OPE_IMPO <= 5036) and (IMP_V_AGG_IMPON >= 2759) and (TOT_IVA_OPE_IMPO <= 1176) and (IMP_PROD_NETTA <= 7719) => CLASSE=interessante (12.0/1.0)
(COSTI_RSDL >= 899) and (CRED_ANNO_PREC >= 1993) and (CRED_ANNO_PREC <= 4881) and (TOT_IVA_OPE_IMPO >= 10859) => CLASSE=interessante (13.0/2.0)
(REDD_LORDO_2007 <= 29135) and (COSTI_RSDL >= 6867) and (ALTRI_ACQ_IMP >= 37124) and (ALIQ_MEDIA_ACQ >= 18.26) and (SESSO = M) => CLASSE=interessante (16.0/0.0)
(TOT_ACQ <= 14927) and (ALIQ_MEDIA_ACQ <= 13.47) and (TOT_ACQ >= 5483) => CLASSE=interessante (11.0/2.0)
(DURATA_ATTIV_SOGG <= 8) and (TOT_IVA_OPE_IMPO >= 22700) and (ACQ_ESENTI >= 500) => CLASSE=interessante (15.0/4.0)
(ALTRI_ACQ_IMP <= 10341) and (GRP_ATTIV_2007 = C) and (ALIQ_MEDIA_ACQ >= 19.31) => CLASSE=interessante (10.0/1.0)
=> CLASSE=non_interessante (936.0/52.0)

Number of Rules : 16

JRIP rules:

==========

(REDD_LORDO_2007 <= 30023) and (BENI_DEST_RIV <= 1000) and (TOT_IMPST_DOV >= 3) and (COSTO_LAV_2007 >= 4151) and (TOT_IVA_OPE_IMPO >= 11036) => CLASSE=interessante (28.0/0.0)
(IMP_V_AGG_IMPON <= 26847) and (BENI_DEST_RIV <= 5300) and (ALIQ_MEDIA_ACQ >= 18.53) and (COSTI_RSDL >= 2772) and (ALTRI_ACQ_IMP <= 5240) => CLASSE=interessante (34.0/0.0)
(IMP_V_AGG_IVA <= 7006) and (IMP_PROD_NETTA <= 221) and (TOT_PASS_2007 >= 855) and (TOT_ACQ <= 10100) and (IMP_REDD_LAV_AUT_2007 <= 0) => CLASSE=interessante (29.0/0.0)
(CRED_ANNO_PREC >= 2377) and (IMPST_CRED >= 1062) and (ACQ_ESENTI <= 90) and (COSTI_RSDL >= 9824) and (REDD_IMP_2007 <= 20502) => CLASSE=interessante (34.0/0.0)
(ACQ_ESENTI >= 624) and (IMP_REDD_PERD_2007 <= 11953) and (IMP_BEN_STRUM_NA <= 1958) => CLASSE=interessante (20.0/2.0)
(ACQ_NO_DETR >= 2614) and (IMP_VE_VOLAFF_2007 <= 56080) and (ACQ_NO_DETR >= 9318) and (ETA >= 40) => CLASSE=interessante (12.0/0.0)
(DEB_FORN_ORD >= 50587) and (IMP_REDD_PERD_2007 >= 26962) and (REDD_IMP_2007 <= 29616) and (ACQ_NO_DETR >= 1788) => CLASSE=interessante (23.0/2.0)
(IMP_TOT_COMP_NEG_2007 >= 49756) and (TOT_IVA_OPE_IMPO >= 16347) and (BENI_DEST_RIV <= 70005) and (BENI_DEST_RIV >= 1055) and (TOT_ACQ >= 61372) => CLASSE=interessante (35.0/10.0)
(IMP_VE_VOLAFF_2007 <= 56824) and (TOT_IMPST_DOV >= 2030) and (COSTI_RSDL <= 975) and (IMP_CMPNS_TERZI_2007 <= 5080) and (IMP_RICAVI_SMPL_2007 <= 21593) => CLASSE=interessante (20.0/0.0)
(ETA <= 32) and (COSTI_RSDL <= 10933) and (IMP_V_AGG_IVA >= 114692) => CLASSE=interessante (11.0/1.0)
(GRP_ATTIV_2007 = Q) and (ALIQ_MEDIA_CESS >= 1.88) and (TOT_IMPST_DOV <= 327) => CLASSE=interessante (16.0/1.0)
(IMP_VAR_RIM_PF >= 231454) => CLASSE=interessante (6.0/0.0)
(TOT_IMPST_CRED >= 1320) and (IMP_RICAVI_SMPL_2007 >= 36095) and (IMP_VE_VOLAFF_2007 <= 56824) and (DURATA_ATTIV_SOGG <= 12) => CLASSE=interessante (14.0/0.0)
(ETA >= 66) and (RICAVI_ATT_2007 >= 435415) => CLASSE=interessante (14.0/3.0)
(TOT_PASS_2007 <= 13840) and (IMP_SPS_DIPEND_SMP >= 1358) => CLASSE=interessante (7.0/0.0)
(FLG_STUDIO_SETTORE_07 <= 0) and (IMP_PROD_NETTA >= 6670) and (ALTRI_ACQ_IMP <= 14735) and (DURATA_ATTIV_SOGG >= 15) => CLASSE=interessante (10.0/1.0)
 => CLASSE=non_interessante (934.0/55.0)

Number of Rules : 17

JRIP rules:
==========

(IMP_V_AGG_IVA <= 20680) and (BENI_DEST_RIV <= 1055) and (COSTI_RSDL >= 1741) and (ALTRI_ACQ_IMP <= 9309) and (OP_ESENTI <= 0) => CLASSE=interessante (50.0/1.0)
(IMP_V_AGG_IVA <= 6193) and (COSTI_ACQ_MP >= 98198) and (REDD_IMP_2007 <= 17500) and (ALTRI_ACQ_IMP >= 10492) and (IMP_BEN_STRUM_NA <= 0) => CLASSE=interessante (30.0/0.0)
(GRP_ATTIV_2007 = C) and (ALIQ_MEDIA_CESS >= 20) and (BENI_DEST_RIV <= 2699) => CLASSE=interessante (36.0/4.0)
(REDD_LORDO_2007 <= 57966) and (COSTI_RSDL <= 966) and (DURATA_ATTIV_SOGG >= 18) and (REDD_LORDO_2007 >= 18088) and (TOT_IMPST_CRED >= 9) and (IMP_REDD_LRD_ORD <= 39766) and (IMP_PROD_NETTA >= 23714) => CLASSE=interessante (28.0/1.0)
(ALTRI_ACQ_IMP <= 1396) and (ALTRI_ACQ_IMP >= 612) and (QTA_PART_IVA <= 1) and (ACQ_NO_DETR <= 6) => CLASSE=interessante (18.0/0.0)
(REDD_LORDO_2007 <= 29135) and (ACQ_NO_DETR >= 1588) and (ACQ_NO_DETR <= 1962) and (IMP_COSTI_SERV <= 16931) => CLASSE=interessante (15.0/0.0)
(IMP_V_AGG_IVA <= 18040) and (REDD_IMP_2007 >= 16613) and (REDD_IMP_2007 <= 20725) => CLASSE=interessante (26.0/6.0)
(ACQ_NO_DETR >= 2614) and (IMP_REDD_PERD_2007 <= 59565) and (IMP_V_AGG_IVA <= 3441) and (REDD_LORDO_2007 >= 2909) => CLASSE=interessante (18.0/0.0)
(CRED_ANNO_PREC >= 2266) and (IMP_REDD_PERD_2007 >= 28379) and (IMP_V_AGG_IMPON <= 25714) and (IMP_V_AGG_IMPON >= -18501) => CLASSE=interessante (10.0/0.0)
(IMP_REDD_PERD_2007 <= 4153) and (IMP_V_AGG_IVA >= 4076) and (IMP_COSTI_MP <= 21135) and (DURATA_ATTIV_SOGG >= 19) => CLASSE=interessante (16.0/1.0)
(ALTRI_ACQ_IMP >= 19392) and (ALIQ_MEDIA_ACQ >= 19.14) and (IMPST_DOV <= 4447) and (IMP_V_AGG_IMPON >= 8101) and (IMP_PROD_NETTA <= 90792) => CLASSE=interessante (19.0/3.0)
(TOT_PASS_2007 >= 90053) and (TOT_PASS_2007 <= 123778) and (IMP_BEN_AMM >= 2896) and (CRED_ANNO_PREC <= 644) and (ACQ_NO_DETR <= 1871) => CLASSE=interessante (13.0/0.0)
(IMP_BEN_STRUM_NA >= 3261) and (IMP_BEN_AMM <= 1534) and (ALIQ_MEDIA_ACQ <= 15.22) and (ACQ_ESENTI <= 5386) => CLASSE=interessante (13.0/1.0)
(GRP_ATTIV_2007 = M) and (ETA >= 55) and (REDD_LORDO_2007 <= 24004) and (ALIQ_MEDIA_ACQ <= 19.19) => CLASSE=interessante (13.0/0.0)
 => CLASSE=non_interessante (942.0/42.0)

Number of Rules : 15

JRIP rules:
==========

(IMP_V_AGG_IVA <= 5808) and (SIGLA_PROVINCIA = FI) and (PROF_COMP <= 8795) => CLASSE=interessante (37.0/7.0)
(REDD_LORDO_2007 <= 56492) and (BENI_DEST_RIV <= 17683) and (COSTO_LAV_2007 >= 2345) and (IMP_REDD_IMP_SMPL_2007 >= 14110) and (TOT_IMPST_CRED <= 0) and (ALIQ_MEDIA_ACQ >= 19.55) => CLASSE=interessante (29.0/0.0)

(IMP_V_AGG_IVA <= 7006) and (TOT_IVA_OPE_IMPO <= 1430) and (TOT_PASS_2007 >= 2392) and (ALIQ_MEDIA_CESS >= 10.01) and (IMP_CMPNS_TERZI_2007 <= 0) => CLASSE=interessante (17.0/0.0)
(REDD_LORDO_2007 <= 55444) and (CRED_ANNO_PREC >= 5941) and (TOT_IMPST_CRED >= 14676) and (ALTRI_ACQ_IMP <= 37194) and (PROF_COMP <= 0) => CLASSE=interessante (17.0/1.0)
(REDD_LORDO_2007 <= 59489) and (IMP_RICAVI_SMPL_2007 >= 136159) and (REDD_IMP_2007 <= 13183) and (IMP_TOT_COMP_NEG_2007 <= 274284) => CLASSE=interessante (21.0/0.0)
(REDD_LORDO_2007 <= 59489) and (IMP_V_AGG_IVA >= 123895) and (ACQ_NO_DETR >= 1871) and (IMP_SPS_DIPEND_ORD_2007 <= 92236) => CLASSE=interessante (23.0/1.0)
(IMP_VE_VOLAFF_2007 <= 98691) and (DURATA_ATTIV_SOGG >= 23) and (ETA <= 54) and (IMP_BEN_STRUM_NA >= 10) => CLASSE=interessante (21.0/2.0)
(IMP_VE_VOLAFF_2007 <= 56824) and (IMP_REDD_PERD_2007 >= 19487) and (IMP_BEN_AMM >= 250) and (OP_ESENTI >= 16400) => CLASSE=interessante (12.0/0.0)
(RICAVI_ATT_2007 <= 42158) and (TOT_IMPST_DOV >= 2030) and (IMP_VAR_RIM_MP <= 0) and (IMP_REDD_LRD_SMPL_2007 <= 9305) => CLASSE=interessante (17.0/0.0)
(REDD_LORDO_2007 <= 29135) and (TOT_ACQ >= 251803) and (ALIQ_MEDIA_CESS <= 10) => CLASSE=interessante (21.0/2.0)
(IMP_VE_VOLAFF_2007 <= 36055) and (CRED_ANNO_PREC >= 965) and (REDD_IMP_2007 >= 15170) and (IMP_PROD_NETTA <= 13148) => CLASSE=interessante (14.0/1.0)
(IMP_V_AGG_IMPON <= 18442) and (IMP_V_AGG_IMPON >= 13996) and (IMP_REDD_LRD_SMPL_2007 >= 18740) and (ETA >= 47) => CLASSE=interessante (11.0/0.0)
(ALIQ_MEDIA_CESS >= 15.42) and (REDD_IMP_2007 <= 20989) and (IMP_REDD_IMP_SMPL_2007 >= 20190) and (ACQ_NO_DETR <= 203) => CLASSE=interessante (10.0/0.0)
(PROF_COMP >= 50000) and (TOT_IVA_OPE_IMPO <= 11216) and (IMP_BEN_AMM <= 192) and (BENI_DEST_RIV <= 0) => CLASSE=interessante (11.0/0.0)
(ALIQ_MEDIA_CESS >= 15.42) and (CRED_ANNO_PREC >= 6379) and (CRED_ANNO_PREC <= 6978) and (DURATA_ATTIV_SOGG >= 20) and (QTA_PART_IVA <= 2) => CLASSE=interessante (10.0/0.0)
(RIM_FIN_ORD >= 162500) and (CRED_ANNO_PREC <= 0) and (DEB_FORN_ORD <= 131029) => CLASSE=interessante (8.0/0.0)
(REDD_LORDO_2007 <= 2563) and (IMP_TOT_COMP_NEG_2007 >= 95194) and (IMP_TOT_COMP_NEG_2007 <= 181869) => CLASSE=interessante (9.0/0.0)
 => CLASSE=non_interessante (959.0/55.0)

Number of Rules : 18

JRIP rules:
===========

(IMP_V_AGG_IVA <= 14135) and (CRED_ANNO_PREC >= 2016) and (TOT_IMPST_CRED <= 394) and (QTA_PART_IVA <= 1) => CLASSE=interessante (20.0/0.0)
(IMP_V_AGG_IVA <= 14371) and (IMP_V_AGG_IVA <= -5200) and (IMP_PROD_NETTA <= 17555) and (COSTI_ACQ_MP >= 8293) => CLASSE=interessante (35.0/3.0)
(ALIQ_MEDIA_ACQ >= 18.68) and (SIGLA_PROVINCIA = FI) and (IMP_VE_VOLAFF_2007 <= 35937) and (ALTRI_ACQ_IMP >= 2675) => CLASSE=interessante (29.0/0.0)
(ALIQ_MEDIA_ACQ >= 18.82) and (ACQ_NO_DETR >= 1400) and (BENI_DEST_RIV <= 119866) and (IMP_REDD_LRD_ORD >= 26962) => CLASSE=interessante (25.0/1.0)
(REDD_LORDO_2007 <= 30708) and (COSTO_LAV_2007 >= 6484) and (BENI_DEST_RIV <= 742) and (ETA <= 52) and (IMP_REDD_LRD_SMPL_2007 >= 8748) => CLASSE=interessante (24.0/2.0)
(IMP_V_AGG_IVA <= 14371) and (RICAVI_ATT_2007 <= 17597) and (IMP_VE_VOLAFF_2007 >= 8619) and (REDD_IMP_2007 <= 10882) and (IMP_VE_VOLAFF_2007 >= 15133) => CLASSE=interessante (21.0/1.0)
(QTA_PART_IVA >= 2) and (IMP_PROD_NETTA >= 5102) and (IMP_PROD_NETTA <= 23714) and (IMPST_DOV >= 3340) and (ALIQ_MEDIA_ACQ >= 19.63) => CLASSE=interessante (22.0/1.0)
(CRED_ANNO_PREC >= 4881) and (IMP_REDD_PERD_2007 >= 26051) and (REDD_LORDO_2007 <= 56492) and (TOT_ACQ >= 289664) => CLASSE=interessante (18.0/2.0)
(IMP_V_AGG_IVA <= 73718) and (PROF_COMP >= 34070) and (ACQ_ESENTI >= 497) => CLASSE=interessante (15.0/2.0)
(REDD_IMP_2007 <= 17750) and (REDD_IMP_2007 >= 15495) and (ALIQ_MEDIA_ACQ <= 19.43) => CLASSE=interessante (28.0/10.0)
(IMP_V_AGG_IVA <= 18413) and (BENI_DEST_RIV <= 1055) and (IMP_TOT_COMP_NEG_2007 >= 3283) and (DURATA_ATTIV_SOGG <= 14) and (IMPST_DOV >= 1214) => CLASSE=interessante (15.0/0.0)
(IMP_TOT_COMP_NEG_2007 >= 49756) and (ALIQ_MEDIA_ACQ >= 18.65) and (REDD_IMP_2007 <= 4951) and (RIM_FIN_SMPL <= 102463) and (IMP_SPS_DIPEND_SMP <= 47398) => CLASSE=interessante (12.0/0.0)
(IMP_V_AGG_IVA <= 9123) and (FLG_NO_CONGRUENZA_07 <= 0) and (DURATA_ATTIV_SOGG >= 18) and (DURATA_ATTIV_SOGG <= 28) and (IMP_PROD_NETTA <= 17138) => CLASSE=interessante (19.0/3.0)
(IMP_SPS_DIPEND_SMP >= 10336) and (IMP_BEN_AMM >= 2900) and (TOT_ACQ <= 52954) => CLASSE=interessante (8.0/0.0)
(QTA_PART_IVA >= 2) and (DURATA_ATTIV_SOGG <= 11) and (COSTI_ACQ_MP >= 15462) => CLASSE=interessante (9.0/0.0)
(PROF_COSTI >= 855) and (CRED_ANNO_PREC >= 957) and (IMP_BEN_AMM <= 128) and (GRP_ATTIV_2007 = M) => CLASSE=interessante (10.0/1.0)
 => CLASSE=non_interessante (937.0/58.0)

Number of Rules : 17

JRIP rules:
===========

(IMP_V_AGG_IVA <= 6734) and (IMP_REDD_LRD_SMPL_2007 <= -105) and (COSTI_RSDL >= 8885) and (ACQ_ESENTI <= 0) => CLASSE=interessante (25.0/1.0)
(ALIQ_MEDIA_ACQ >= 19.23) and (COSTI_RSDL >= 700) and (REDD_IMP_2007 <= 19642) and (ALTRI_ACQ_IMP >= 37124) => CLASSE=interessante (37.0/3.0)

(IMP_V_AGG_IVA <= 6734) and (TOT_IVA_OPE_IMPO <= 1240) and (TOT_ACQ >= 2675) and (IMP_V_AGG_IMPON >= -7855) => CLASSE=interessante (28.0/0.0)
(IMP_V_AGG_IVA <= 6348) and (ALIQ_MEDIA_CESS >= 19.96) and (IMP_REDD_IMP_SMPL_2007 >= 449) and (IMP_BEN_AMM <= 2116) and (COSTI_RSDL >= 1660) => CLASSE=interessante (19.0/2.0)
(IMP_VE_VOLAFF_2007 <= 17730) and (COSTI_RSDL >= 1941) and (ACQ_NO_DETR <= 84) => CLASSE=interessante (14.0/1.0)
(ETA <= 45) and (RICAVI_ATT_2007 >= 610950) and (TOT_IVA_OPE_IMPO <= 140798) and (ALTRI_ACQ_IMP >= 80839) and (IMP_VAR_RIM_MP >= 0) => CLASSE=interessante (22.0/1.0)
(CRED_ANNO_PREC >= 2994) and (IMP_BEN_AMM <= 10968) and (ACQ_NO_DETR >= 821) and (BENI_DEST_RIV <= 119866) and (PROF_COMP <= 54992) and (RICAVI_ATT_2007 <= 240841) => CLASSE=interessante (17.0/0.0)
(COSTO_LAV_2007 >= 5056) and (RICAVI_ATT_2007 <= 71324) and (TOT_IVA_OPE_IMPO >= 8578) and (IMP_VE_VOLAFF_2007 <= 63032) => CLASSE=interessante (25.0/3.0)
(REDD_LORDO_2007 <= 57966) and (IMP_V_AGG_IVA >= 191373) and (IMPST_DOV >= 39839) => CLASSE=interessante (14.0/0.0)
(ACQ_NO_DETR <= 296) and (IMP_SPS_DIPEND_SMP >= 1553) and (ALIQ_MEDIA_CESS >= 14.92) and (ETA <= 56) and (RIM_FIN_SMPL <= 36300) => CLASSE=interessante (26.0/4.0)
(ALTRI_ACQ_IMP <= 4689) and (IMP_PROD_NETTA >= 19751) and (TOT_IVA_OPE_IMPO >= 11220) and (ALIQ_MEDIA_ACQ <= 19.97) => CLASSE=interessante (14.0/1.0)
(IMP_REDD_LRD_SMPL_2007 >= 14545) and (IMP_REDD_PERD_2007 <= 16303) and (TOT_IMPST_CRED >= 363) => CLASSE=interessante (12.0/1.0)
(PROF_COMP >= 2759) and (IMP_VE_VOLAFF_2007 <= 6045) => CLASSE=interessante (6.0/0.0)
(COSTO_LAV_2007 >= 28894) and (TOT_ACQ <= 79987) and (ACQ_NO_DETR >= 5714) => CLASSE=interessante (9.0/0.0)
 => CLASSE=non_interessante (979.0/66.0)

Number of Rules : 15

JRIP rules:
===========

(ACQ_NO_DETR <= 14) and (IMP_VE_VOLAFF_2007 <= 18487) and (TOT_PASS_2007 >= 2392) and (ALIQ_MEDIA_CESS >= 10) => CLASSE=interessante (47.0/3.0)
(REDD_LORDO_2007 <= 29135) and (BENI_DEST_RIV <= 1055) and (ALIQ_MEDIA_ACQ >= 18.55) and (IMP_PROD_NETTA >= 16988) => CLASSE=interessante (53.0/9.0)
(IMPST_CRED >= 1456) and (IMP_PROD_NETTA <= 13148) and (COSTI_RSDL >= 9625) and (IMPST_CRED <= 4479) => CLASSE=interessante (26.0/0.0)
(IMP_SPS_DIPEND_SMP >= 1553) and (ALIQ_MEDIA_CESS >= 14.92) and (IMP_REDD_PERD_2007 >= 15067) and (ACQ_NO_DETR <= 0) and (DURATA_ATTIV_SOGG <= 16) => CLASSE=interessante (28.0/0.0)
(CRED_ANNO_PREC >= 2266) and (IMPST_DOV <= 3150) and (TOT_IVA_OPE_IMPO >= 10859) and (DURATA_ATTIV_SOGG >= 31) => CLASSE=interessante (14.0/1.0)
(IMP_V_AGG_IVA <= 18521) and (IMP_TOTALE_SPESE_2007 >= 4498) and (ETA >= 42) => CLASSE=interessante (16.0/4.0)
(IMP_REDD_PERD_2007 <= 11784) and (ACQ_ESENTI >= 42) and (ALTRI_ACQ_IMP >= 31217) and (ETA <= 45) => CLASSE=interessante (18.0/0.0)
(IMP_SPS_DIPEND_SMP >= 14986) and (COSTI_RSDL <= 37232) and (IMP_BEN_STRUM_NA >= 153) and (ETA >= 35) => CLASSE=interessante (24.0/3.0)
(RICAVI_ATT_2007 <= 10698) and (TOT_ACQ >= 4050) and (ALTRI_ACQ_IMP <= 9772) => CLASSE=interessante (17.0/1.0)
(IMP_VE_VOLAFF_2007 >= 223809) and (ALIQ_MEDIA_CESS <= 7.1) and (CRED_CLI_ORD >= 37725) and (ETA <= 52) => CLASSE=interessante (15.0/0.0)
(ALTRI_ACQ_IMP >= 17927) and (IMP_TOT_COMP_NEG_2007 >= 108275) and (PRES_FAM_ORD_07 >= 1) and (BENI_DEST_RIV <= 107541) => CLASSE=interessante (18.0/2.0)
(IMP_V_AGG_IVA <= 18521) and (IMP_REDD_PERD_2007 >= 19487) and (IMP_REDD_PERD_2007 <= 22044) and (REDD_IMP_2007 <= 20637) => CLASSE=interessante (12.0/2.0)
(ALTRI_ACQ_IMP <= 835) and (PROF_COMP >= 2759) and (TOT_PASS_2007 <= 994) => CLASSE=interessante (7.0/0.0)
(CRED_ANNO_PREC >= 1337) and (TOT_ACQ <= 27301) and (BENI_DEST_RIV >= 1372) => CLASSE=interessante (14.0/4.0)
(DURATA_ATTIV_SOGG >= 26) and (TOT_IMPST_DOV >= 1108) and (FLG_NO_CONGRUENZA_07 >= 1) => CLASSE=interessante (16.0/5.0)
(ALTRI_ACQ_IMP >= 47722) and (ETA >= 62) and (CRED_ANNO_PREC >= 5107) => CLASSE=interessante (10.0/1.0)
 => CLASSE=non_interessante (912.0/46.0)

Number of Rules : 17

JRIP rules:
===========

(IMP_REDD_PERD_2007 <= 22080) and (SIGLA_PROVINCIA = FI) and (ALIQ_MEDIA_ACQ >= 17.66) and (RIM_FIN_SMPL <= 3502) => CLASSE=interessante (57.0/8.0)
(ACQ_NO_DETR <= 150) and (IMP_VE_VOLAFF_2007 <= 28941) and (TOT_IMPST_DOV >= 11) and (ALIQ_MEDIA_ACQ <= 19.86) => CLASSE=interessante (29.0/1.0)
(IMP_V_AGG_IVA <= 6193) and (BENI_DEST_RIV <= 4792) and (COSTI_RSDL >= 2392) and (ACQ_NO_DETR <= 1400) => CLASSE=interessante (31.0/1.0)
(DURATA_ATTIV_SOGG >= 18) and (IMP_SPS_DIPEND_SMP >= 12709) and (TOT_ACQ <= 93663) and (COSTI_ACQ_MP >= 8847) and (ETA >= 49) => CLASSE=interessante (31.0/1.0)
(IMPST_CRED >= 550) and (IMP_PROD_NETTA <= 17555) and (COSTO_LAV_2007 <= 145) and (FLG_NO_COERENZA_07 <= 0) and (IMPST_CRED >= 962) => CLASSE=interessante (26.0/1.0)

(CRED_ANNO_PREC >= 2266) and (SIGLA_PROVINCIA = FI) and (TOT_ACQ >= 303823) =>
CLASSE=interessante (14.0/1.0)
(DURATA_ATTIV_SOGG >= 18) and (IMP_VE_VOLAFF_2007 <= 91955) and (COSTO_LAV_2007 >= 11281)
and (IMP_BEN_AMM <= 117) and (ALTRI_ACQ_IMP >= 12742) => CLASSE=interessante (17.0/2.0)
(REDD_IMP_2007 <= 934) and (ALIQ_MEDIA_CESS <= 4) => CLASSE=interessante (16.0/3.0)
(QTA_PART_IVA >= 2) and (ALIQ_MEDIA_CESS <= 1.88) and (TOT_ACQ >= 25630) and
(IMP_REDD_LRD_SMPL_2007 <= 39857) => CLASSE=interessante (16.0/3.0)
(TOT_PASS_2007 >= 79976) and (IMP_BEN_AMM <= 2900) and (IMP_BEN_STRUM_NA >= 4604) and
(COSTO_LAV_2007 >= 76268) => CLASSE=interessante (13.0/1.0)
(SIGLA_PROVINCIA = PO) and (CRED_ANNO_PREC >= 2016) => CLASSE=interessante (8.0/1.0)
(IMP_REDD_LRD_ORD >= 23105) and (IMP_BEN_AMM <= 8769) and (RIM_FIN_ORD >= 22000) and
(FLG_PRES_FAM_07 <= 0) => CLASSE=interessante (10.0/0.0)
=> CLASSE=non_interessante (979.0/82.0)

Number of Rules : 13

JRIP rules:
===========

(IMPST_CRED >= 1456) and (ACQ_NO_DETR >= 2275) and (ETA >= 41) and (ALIQ_MEDIA_ACQ >= 11.28) =>
CLASSE=interessante (32.0/1.0)
(IMP_V_AGG_IVA <= 9675) and (BENI_DEST_RIV <= 4792) and (COSTI_RSDL >= 2392) and
(IMP_CESS_BENI_AMM <= 1000) and (ALIQ_MEDIA_ACQ >= 18.26) => CLASSE=interessante (29.0/0.0)
(RICAVI_ATT_2007 <= 52683) and (IMP_BEN_AMM >= 213) and (ALIQ_MEDIA_ACQ <= 15.22) =>
CLASSE=interessante (29.0/5.0)
(IMP_SPS_DIPEND_SMP >= 10336) and (ALIQ_MEDIA_ACQ >= 19.74) and (BENI_DEST_RIV <= 45991) and
(IMPST_CRED >= 1518) => CLASSE=interessante (19.0/0.0)
(RICAVI_ATT_2007 <= 51636) and (IMP_VE_VOLAFF_2007 >= 34180) and (TOT_IMPST_CRED >= 523) and
(ALIQ_MEDIA_ACQ <= 19.89) and (ALIQ_MEDIA_ACQ >= 17.99) => CLASSE=interessante (22.0/0.0)
(ALIQ_MEDIA_CESS >= 17.11) and (GRP_ATTIV_2007 = C) and (TOT_PASS_2007 <= 27794) =>
CLASSE=interessante (24.0/1.0)
(IMP_SPS_DIPEND_SMP >= 21541) and (ALIQ_MEDIA_CESS >= 19.44) and (ETA <= 41) =>
CLASSE=interessante (13.0/1.0)
(DURATA_ATTIV_SOGG >= 29) and (IMP_RICAVI_SMPL_2007 >= 130868) and (RIM_FIN_SMPL <= 4488)
=> CLASSE=interessante (14.0/0.0)
(IMP_REDD_IMP_SMPL_2007 <= 9619) and (TOT_ACQ >= 251803) and (TOT_IVA_OPE_IMPO <= 102896)
and (TOT_IVA_OPE_IMPO >= 85543) => CLASSE=interessante (20.0/1.0)
(IMP_V_AGG_IVA <= 18521) and (TOT_IVA_OPE_IMPO <= 1198) and (ALTRI_ACQ_IMP >= 55) and
(IMPST_DOV <= 342) => CLASSE=interessante (18.0/2.0)
(IMP_REDD_PERD_2007 <= 10476) and (TOT_IVA_OPE_IMPO >= 9981) and (BENI_DEST_RIV <= 26661)
and (CRED_ANNO_PREC <= 2377) => CLASSE=interessante (17.0/0.0)
(FLG_NO_COERENZA_07 >= 1) and (PROF_COSTI >= 73717) => CLASSE=interessante (12.0/4.0)
(IMP_V_AGG_IVA <= 53132) and (ALTRI_ACQ_IMP >= 20387) and (ALIQ_MEDIA_ACQ <= 15.88) and
(ALIQ_MEDIA_ACQ >= 11.57) => CLASSE=interessante (25.0/7.0)
(IMP_V_AGG_IVA <= 18521) and (IMP_V_AGG_IVA >= 18040) => CLASSE=interessante (9.0/1.0)
(RIM_FIN_ORD >= 99762) and (IMP_PROD_NETTA <= 51867) and (RIM_FIN_ORD >= 162500) =>
CLASSE=interessante (14.0/4.0)
(QTA_PART_IVA >= 2) and (ETA <= 46) and (ALIQ_MEDIA_ACQ >= 19.87) and (TOT_IMPST_DOV <= 823)
=> CLASSE=interessante (11.0/1.0)
(PROF_COSTI >= 855) and (TOT_ACQ <= 14927) and (ALTRE_SPS_DOC >= 8262) and (IMP_BEN_AMM <=
129) => CLASSE=interessante (8.0/0.0)
=> CLASSE=non_interessante (931.0/41.0)

Number of Rules : 18

JRIP rules:
===========

(IMP_V_AGG_IVA <= 7006) and (BENI_DEST_RIV <= 610) and (COSTI_RSDL >= 2392) and
(ALIQ_MEDIA_ACQ >= 19) => CLASSE=interessante (29.0/0.0)
(IMP_REDD_PERD_2007 <= 24290) and (RICAVI_ATT_2007 <= 17597) and (IMP_TOT_COMP_NEG_2007
>= 2363) => CLASSE=interessante (36.0/3.0)
(IMP_TOT_COMP_NEG_2007 >= 143341) and (BENI_DEST_RIV <= 107541) and (COSTI_ACQ_MP >= 67584)
and (IMP_V_AGG_IVA <= 84453) => CLASSE=interessante (32.0/1.0)
(ACQ_NO_DETR <= 0) and (IMP_VE_VOLAFF_2007 <= 10784) and (IMP_VE_VOLAFF_2007 >= 1100) =>
CLASSE=interessante (27.0/3.0)
(COSTO_LAV_2007 >= 3362) and (REDD_LORDO_2007 <= 28896) and (REDD_LORDO_2007 >= 14119) and
(IMP_VE_VOLAFF_2007 <= 92516) and (TOT_ACQ <= 19412) => CLASSE=interessante (36.0/3.0)
(TOT_PASS_2007 >= 83544) and (PRES_FAM_ORD_07 >= 1) and (IMP_TOT_COMP_NEG_2007 >= 233679)
and (RIM_FIN_SMPL <= 4796) => CLASSE=interessante (20.0/0.0)
(TOT_PASS_2007 >= 83499) and (TOT_ACQ <= 93462) and (REDD_IMP_2007 <= 21555) and (COSTI_RSDL
>= 11413) and (IMP_TOT_COMP_NEG_2007 <= 123778) => CLASSE=interessante (32.0/5.0)
(CRED_ANNO_PREC >= 1797) and (ACQ_NO_DETR >= 787) and (ACQ_NO_DETR <= 2783) and
(ACQ_NO_DETR >= 1514) => CLASSE=interessante (31.0/5.0)
(ALIQ_MEDIA_ACQ >= 18.59) and (IMP_REDD_LRD_SMPL_2007 >= 9305) and
(IMP_REDD_IMP_SMPL_2007 <= 10920) => CLASSE=interessante (19.0/3.0)
(FLG_STUDIO_SETTORE_07 <= 0) and (TOT_IMPST_DOV >= 60) and (ACQ_NO_DETR <= 138) =>
CLASSE=interessante (17.0/0.0)
(IMP_V_AGG_IVA >= 192168) and (IMP_PROD_NETTA <= 137563) and (IMP_COSTI_SERV >= 21081) =>
CLASSE=interessante (20.0/2.0)

```
(ALIQ_MEDIA_ACQ >= 19.78) and (SIGLA_PROVINCIA = FI) and (IMP_SPS_DIPEND_SMP >= 2345) =>
CLASSE=interessante (13.0/2.0)
(IMP_V_AGG_IMPON <= 212) and (ALIQ_MEDIA_ACQ <= 12.09) and (ALIQ_MEDIA_CESS >= 4) =>
CLASSE=interessante (17.0/2.0)
(COSTO_LAV_2007 >= 26944) and (IMP_V_AGG_IMPON <= 89141) and (IMP_V_AGG_IMPON >= 53279) and
(TOT_ACQ >= 73014) and (IMP_RICAVI_SMPL_2007 <= 151647) => CLASSE=interessante (16.0/2.0)
(RICAVI_ATT_2007 <= 60508) and (ACQ_NO_DETR >= 9318) => CLASSE=interessante (10.0/2.0)
(IMPST_CRED >= 1518) and (IMPST_CRED <= 3453) and (ALTRI_ACQ_IMP <= 17452) and
(IMP_RICAVI_SMPL_2007 >= 21526) => CLASSE=interessante (9.0/0.0)
 => CLASSE=non_interessante (883.0/44.0)

Number of Rules : 17
```

Quanto all'interpretabilità delle regole estratte dal modello, vale la pena di osservare come le varie iterazioni producano set di regole diversi, che suddividono lo spazio delle osservazioni in vario modo. Identificare le regole attivate da ciascuna istanza appare alquanto arduo senza disporre di un *software ad hoc*. Ancora, si osserva come a volte vengano mescolati attributi tratti da quadri diversi della dichiarazione, per cui si evidenziano le problematiche già descritte nell'ambito degli alberi di classificazione e come anche in questo caso l'algoritmo prediliga attributi tratti da quadri compilati dalla generalità dei contribuenti presenti nel *test set*.

Mostriamo ora gli usuali *lift charts* relativi al modello appena descritto. Anche nel caso specifico, in cui il modello propone, effettivamente, un esiguo numero di controlli, il *lift chart* può fornire una ulteriore guida per la pianificazione degli *audit*, nel senso espresso ad esempio nel caso del metaclassificatore che utilizzava matrici di costo.

L'area sotto la curva è pari al 70,97% del massimo possibile se si considera l'*outlier* nel *test set*, mentre è pari al 71,59% in caso contrario.

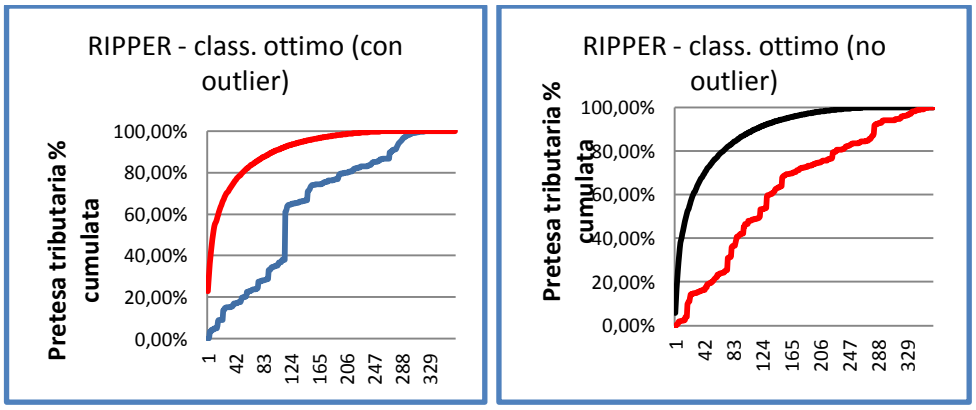

Figura 5.16: confronto bagging_RIPPER e ottimo

5.3.2.2 Modello PART

Sempre nell'ambito delle regole di classificazione, utilizziamo l'algoritmo PART, che genera una *decision list*, utilizzando anch'esso una tecnica *divide et impera*, nel senso che crea una regola, rimuove le istanze che questa copre e continua a creare regole ricorsivamente

dalle istanze via via rimaste, finché non ne rimane nessuna. Differisce dall'algoritmo precedente poiché ad ogni iterazione, per estrarre una regola, crea un albero di decisione C4.5 *parziale* e la foglia "migliore" (ovvero con la maggiore copertura), viene trasformata in regola. Contrariamente a RIPPER, non necessita di una fase di ottimizzazione globale per produrre il *set* di regole e questo è il suo principale vantaggio.

L'algoritmo viene utilizzato con i seguenti parametri.

binarySplits -- Whether to use binary splits on nominal attributes when building the partial trees.	*False*
confidenceFactor -- The confidence factor used for pruning (smaller values incur more pruning).	*0,15*
debug -- If set to true, classifier may output additional info to the console.	*False*
minNumObj -- The minimum number of instances per rule.	*20*
numFolds -- Determines the amount of data used for reduced-error pruning. One fold is used for pruning, the rest for growing the rules.	*3*
reducedErrorPruning -- Whether reduced-error pruning is used instead of C.4.5 pruning.	*False*
seed -- The seed used for randomizing the data when reduced-error pruning is used.	*1*
unpruned -- Whether pruning is performed.	*False*

Anche in questo caso abbiniamo l'algoritmo ad un metaclassificatore, *bagging*, previo *resample* del *training set* (*biasToUniformClass* 0.1). il modello prodotto viene di seguito presentato:

```
=== Classifier model ===

Scheme:   Bagging

All the base classifiers:

PART decision list
------------------

FLG_STUDIO_SETTORE_07 > 0 AND
REDD_LORDO_2007 > 52400: non_interessante (302.0/28.0)

IMP_V_AGG_IVA <= 187103 AND
IMP_REDD_LRD_ORD > -2340 AND
FLG_STUDIO_SETTORE_07 > 0 AND
QTA_PART_IVA <= 2 AND
ALTRI_ACQ_IMP <= 66467 AND
ETA <= 70 AND
CRED_ANNO_PREC <= 2226 AND
ACQ_ESENTI <= 1100 AND
CRED_CLI_ORD <= 1080 AND
```

SIGLA_PROVINCIA = FI AND
ALIQ_MEDIA_ACQ <= 19.93: non_interessante (84.0/17.0)

IMP_V_AGG_IVA <= 187103 AND
IMP_REDD_LRD_ORD > -2340 AND
IMP_PROD_NETTA <= 88740 AND
ALTRI_ACQ_IMP <= 66467 AND
IMP_V_AGG_IVA > -21344 AND
FLG_STUDIO_SETTORE_07 > 0 AND
IMP_SPS_DIPEND_SMP <= 4928 AND
RICAVI_ATT_2007 > 16454 AND
PROF_COMP <= 33276: non_interessante (340.0/42.0)

IMP_REDD_LRD_ORD > -2340 AND
ETA <= 69 AND
PROF_COMP <= 55000 AND
FLG_PRES_FAM_07 <= 0 AND
RICAVI_ATT_2007 <= 585481 AND
SPESE_ORD <= 115668 AND
REDD_IMP_2007 <= 36450 AND
ALTRI_ACQ_IMP > 6 AND
REDD_LORDO_2007 <= 27721 AND
SIGLA_PROVINCIA = FI: interessante (45.0/4.0)

IMP_REDD_LRD_ORD > -2340 AND
ETA <= 69 AND
TOT_IVA_OPE_IMPO <= 91711 AND
IMP_PROD_NETTA <= 87788 AND
PROF_COMP <= 55000 AND
REDD_IMP_2007 <= 38352 AND
SIGLA_PROVINCIA = LU: non_interessante (42.0/12.0)

IMP_REDD_LRD_ORD > -2340 AND
ETA <= 69 AND
TOT_IVA_OPE_IMPO <= 91711 AND
IMP_PROD_NETTA <= 87788 AND
PROF_COMP > 55000: non_interessante (35.0/6.0)

IMP_REDD_LRD_ORD > -2340 AND
ALIQ_MEDIA_CESS > 1.93 AND
ETA <= 69 AND
ALIQ_MEDIA_CESS <= 7.14: interessante (41.0/8.0)

IMP_REDD_LRD_ORD > -2340 AND
ETA <= 69 AND
REDD_IMP_2007 <= 38352 AND
ALIQ_MEDIA_CESS > 9.96 AND
CRED_ANNO_PREC > 1901: interessante (56.0/17.0)

IMP_REDD_LRD_ORD > -2340 AND
REDD_IMP_2007 > 38352: interessante (66.0/14.0)

CRED_CLI_ORD > 68: non_interessante (64.0/1.0)

IMPST_CRED > 35: non_interessante (45.0/6.0)

ALTRI_ACQ_IMP <= 35161 AND
FLG_PRES_DIP_07 <= 0 AND
IMP_TOT_COMP_NEG_2007 <= 2392: non_interessante (39.0/13.0)

FLG_PRES_DIP_07 > 0 AND
IMP_RICAVI_SMPL_2007 <= 118102: non_interessante (33.0/1.0)

: interessante (55.0/10.0)

Number of Rules : 14

PART decision list

IMP_REDD_LRD_ORD <= 121364 AND
IMP_VE_VOLAFF_2007 > 17730 AND
IMP_VAR_RIM_PF <= 1200 AND
BENI_DEST_RIV <= 476477 AND
FLG_STUDIO_SETTORE_07 > 0 AND
QTA_PART_IVA <= 2 AND
TOT_ACQ <= 246562 AND
RIM_FIN_ORD <= 0 AND
IMP_REDD_LAV_AUT_2007 <= 65258 AND
PROF_COSTI <= 29306 AND
IMP_CMPNS_ATTIV_2007 <= 55000 AND
TOT_IMPST_DOV <= 2016 AND
ALTRE_SPS_DOC <= 2364 AND
IMP_BEN_AMM <= 104: non_interessante (211.0/20.0)

IMP_REDD_LRD_SMPL_2007 > -83 AND
IMP_REDD_LRD_ORD <= 121364 AND
IMP_REDD_LAV_AUT_2007 <= 24004 AND
IMP_COSTI_MP > 7056: non_interessante (218.0/40.0)

IMP_REDD_LRD_ORD <= 102421 AND
IMP_REDD_LAV_AUT_2007 > 24004: non_interessante (151.0/22.0)

IMP_REDD_LRD_ORD <= 102421 AND
BENI_DEST_RIV <= 117072 AND
ALTRI_ACQ_IMP > 264 AND
RICAVI_ATT_2007 > 6786 AND
QTA_PART_IVA <= 2 AND
OP_ESENTI <= 75417 AND
ALIQ_MEDIA_CESS > 1.86 AND
OP_ESENTI <= 1150 AND
COSTO_LAV_2007 <= 2238 AND
PROF_COMP <= 15384 AND
IMP_V_AGG_IMPON > 18736: non_interessante (65.0/9.0)

IMP_REDD_LRD_ORD <= 102421 AND
BENI_DEST_RIV <= 117072 AND
ALTRI_ACQ_IMP > 264 AND
OP_ESENTI <= 75417 AND
TOT_IVA_OPE_IMPO <= 1248: interessante (79.0/16.0)

IMP_REDD_LRD_ORD <= 102421 AND
ALIQ_MEDIA_CESS <= 19.61 AND
DURATA_ATTIV_SOGG <= 28: non_interessante (228.0/48.0)

IMP_REDD_LRD_ORD <= 102421 AND
ALTRI_ACQ_IMP <= 66922 AND
RIM_FIN_SMPL <= 63240 AND
TOT_IVA_OPE_IMPO > 1689 AND

GRP_ATTIV_2007 = C: interessante (42.0/5.0)

IMP_REDD_LRD_ORD <= 102421 AND
ALTRI_ACQ_IMP <= 64837 AND
REDD_LORDO_2007 > -83 AND
ACQ_NO_DETR <= 999: non_interessante (111.0/24.0)

IMP_REDD_LRD_ORD <= 102421 AND
TOT_IVA_OPE_IMPO > 7633: interessante (83.0/9.0)

REDD_LORDO_2007 > 19370: non_interessante (38.0)

: interessante (21.0/10.0)

Number of Rules : 11

PART decision list

FLG_STUDIO_SETTORE_07 > 0 AND
IMP_VE_VOLAFF_2007 > 7013 AND
COSTI_ACQ_MP <= 198225 AND
REDD_LORDO_2007 <= 55476 AND
IMPST_DOV <= 30802 AND
TOT_PASS_2007 > 3141 AND
BENI_DEST_RIV > 1000 AND
ACQ_NO_DETR <= 1332: non_interessante (325.0/39.0)

REDD_LORDO_2007 > 56492: non_interessante (263.0/26.0)

OP_ESENTI <= 22468 AND
TOT_PASS_2007 <= 2625: non_interessante (87.0/14.0)

OP_ESENTI > 22468: non_interessante (40.0/2.0)

IMP_V_AGG_IMPON > 19863 AND
IMP_REDD_LRD_ORD <= 23014 AND
IMP_CESS_BENI_AMM <= 2700 AND
SIGLA_PROVINCIA = PI: non_interessante (38.0)

TOT_IVA_OPE_IMPO <= 1776: interessante (80.0/17.0)

OP_ESENTI <= 10 AND
CRED_ANNO_PREC > 8484: interessante (44.0/9.0)

IMP_CESS_BENI_AMM <= 10000 AND
ALTRI_ACQ_IMP > 1446 AND
GRP_ATTIV_2007 = L: non_interessante (37.0/9.0)

IMP_CESS_BENI_AMM <= 10000 AND
ALTRI_ACQ_IMP > 1446 AND
GRP_ATTIV_2007 = C AND
ALIQ_MEDIA_CESS > 19.99: interessante (34.0/4.0)

IMP_CESS_BENI_AMM <= 10000 AND
ALTRI_ACQ_IMP > 1446 AND
IMP_REDD_IMP_SMPL_2007 > 16930: non_interessante (54.0/7.0)

IMP_CESS_BENI_AMM <= 10000 AND

IMP_SPS_DIPEND_SMP > 5452: interessante (41.0/10.0)

IMP_CESS_BENI_AMM <= 10000 AND
ETA > 36 AND
GRP_ATTIV_2007 = G: non_interessante (38.0/17.0)

IMP_CESS_BENI_AMM <= 3050 AND
ETA > 37 AND
ALTRI_ACQ_IMP > 3825 AND
TOT_PASS_2007 <= 24046: non_interessante (37.0/5.0)

IMP_CESS_BENI_AMM > 3050: non_interessante (30.0)

ETA > 37 AND
IMP_VE_VOLAFF_2007 <= 56824: interessante (29.0/2.0)

ACQ_ESENTI <= 8: non_interessante (45.0/7.0)

: interessante (25.0/7.0)

Number of Rules : 17

PART decision list

IMP_REDD_LRD_SMPL_2007 <= 59489 AND
RICAVI_ATT_2007 > 15283 AND
IMP_VAR_RIM_PF <= 7120 AND
TOT_IVA_OPE_IMPO <= 140798 AND
TOT_IVA_OPE_IMPO <= 93117 AND
TOT_IMPST_CRED <= 21799 AND
OP_NON_IMPO_DI <= 5383 AND
CRED_CLI_ORD <= 2340 AND
IMP_BEN_STRUM_NA <= 1906 AND
QTA_PART_IVA <= 2 AND
SIGLA_PROVINCIA = SI: non_interessante (109.0/10.0)

IMP_REDD_LRD_SMPL_2007 <= 59489 AND
IMP_REDD_LRD_ORD <= 58949 AND
OP_NON_IMPO_DI <= 5383 AND
RICAVI_ATT_2007 > 17597 AND
TOT_IMPST_CRED <= 16026 AND
IMP_BEN_STRUM_NA <= 15953 AND
TOT_PASS_2007 <= 514483 AND
IMP_RICAVI_SMPL_2007 <= 141066 AND
IMP_COSTI_MP <= 33083 AND
IMP_REDD_LRD_ORD <= 3595 AND
IMPST_DOV <= 19054 AND
TOT_IMPST_DOV <= 3006 AND
GRP_ATTIV_2007 = G: non_interessante (111.0/16.0)

REDD_LORDO_2007 > 59489: non_interessante (220.0/16.0)

IMP_SPS_DIPEND_ORD_2007 > 92451: non_interessante (32.0/2.0)

OP_NON_IMPO_DI > 5383 AND
TOT_IMPST_DOV <= 1: non_interessante (23.0/9.0)

OP_NON_IMPO_DI <= 3357 AND

IMP_V_AGG_IVA > 6734 AND
TOT_ACQ > 3796 AND
IMP_BEN_STRUM_NA <= 13683 AND
SIGLA_PROVINCIA = LI: non_interessante (64.0/8.0)

OP_NON_IMPO_DI <= 3357 AND
RICAVI_ATT_2007 > 17598 AND
TOT_IMPST_CRED <= 16026 AND
IMP_COSTI_MP > 33083: non_interessante (85.0/6.0)

OP_NON_IMPO_DI <= 3357 AND
RIM_FIN_ORD <= 60622 AND
IMP_COSTI_BEN_TRZ <= 6114 AND
IMP_VE_VOLAFF_2007 > 63032 AND
REDD_IMP_2007 > 8784 AND
IMP_VE_VOLAFF_2007 > 75517 AND
TOT_IMPST_CRED <= 5130 AND
IMP_CESS_BENI_AMM <= 1 AND
ALIQ_MEDIA_ACQ > 11.75 AND
FLG_NO_CONGRUENZA_07 <= 0: non_interessante (55.0/14.0)

OP_NON_IMPO_DI <= 3357 AND
RIM_FIN_ORD <= 60622 AND
IMP_V_AGG_IVA <= 119696 AND
IMP_V_AGG_IMPON <= 69630 AND
DEB_FORN_ORD <= 0 AND
ALIQ_MEDIA_ACQ > 19.79 AND
REDD_LORDO_2007 > 7019 AND
BENI_DEST_RIV <= 12346 AND
IMP_BEN_AMM <= 917: interessante (35.0/6.0)

OP_NON_IMPO_DI <= 3357 AND
RIM_FIN_ORD <= 60622 AND
IMP_COSTI_SERV <= 66368 AND
IMP_VE_VOLAFF_2007 > 15283 AND
IMP_CMPNS_TERZI_2007 <= 4181 AND
TOT_IMPST_CRED <= 18206 AND
IMPST_CRED <= 5559 AND
IMP_BEN_AMM <= 14008 AND
IMP_CESS_BENI_AMM > 1: non_interessante (37.0)

OP_NON_IMPO_DI <= 3536 AND
RIM_FIN_ORD <= 50697 AND
IMP_SPS_DIPEND_ORD_2007 <= 2677 AND
IMP_V_AGG_IVA > 18570 AND
IMPST_CRED <= 269 AND
IMP_TOT_COMP_NEG_2007 <= 132043 AND
IMP_TOT_COMP_NEG_2007 > 16768: non_interessante (70.0/3.0)

OP_NON_IMPO_DI <= 5383 AND
IMP_RICAVI_ORDIN_2007 <= 424301 AND
ALTRI_ACQ_IMP <= 1: non_interessante (43.0/8.0)

OP_NON_IMPO_DI <= 5383 AND
IMP_VE_VOLAFF_2007 > 7013 AND
IMP_RICAVI_ORDIN_2007 <= 424301 AND
SIGLA_PROVINCIA = LU: non_interessante (35.0/10.0)

OP_NON_IMPO_DI <= 5383 AND
ETA <= 69 AND

IMP_REDD_LRD_SMPL_2007 <= 27660 AND
IMP_V_AGG_IVA <= 115725 AND
IMP_TOT_COMP_NEG_2007 > 1788 AND
ALTRI_ACQ_IMP <= 4880: interessante (40.0/1.0)

OP_NON_IMPO_DI <= 3536 AND
ETA <= 69 AND
RIM_FIN_ORD <= 14249 AND
COSTO_LAV_2007 <= 4242 AND
TOT_IVA_OPE_IMPO > 840 AND
RIM_FIN_SMPL <= 2140: non_interessante (92.0/14.0)

OP_NON_IMPO_DI <= 3536 AND
ETA <= 69 AND
IMP_ONERI_DIV <= 3171 AND
IMP_V_AGG_IVA <= 61428 AND
IMP_V_AGG_IVA <= 23604 AND
IMP_V_AGG_IVA <= 12771: interessante (57.0/15.0)

OP_ESENTI <= 1188 AND
OP_NON_IMPO_DI <= 1752 AND
IMP_ONERI_DIV <= 3171 AND
BENI_DEST_RIV <= 2775: interessante (34.0/5.0)

OP_NON_IMPO_DI <= 1752 AND
IMP_ONERI_DIV > 3171: interessante (37.0/2.0)

OP_NON_IMPO_DI <= 1752 AND
ETA > 44: interessante (25.0/11.0)

OP_NON_IMPO_DI > 1752: interessante (23.0)

: non_interessante (20.0/2.0)

Number of Rules : 21

PART decision list

IMP_VE_VOLAFF_2007 > 17730 AND
COSTI_ACQ_MP <= 255305 AND
OP_ESENTI <= 89425 AND
IMP_REDD_LRD_ORD > -2340 AND
IMP_VAR_RIM_PF <= 1430 AND
REDD_LORDO_2007 > 56492: non_interessante (181.0/16.0)

OP_ESENTI <= 23932 AND
IMP_REDD_LRD_ORD > -2340 AND
TOT_IVA_OPE_IMPO <= 93117 AND
IMP_REDD_LRD_SMPL_2007 > -17316 AND
FLG_PRES_FAM_07 <= 0 AND
IMP_VE_VOLAFF_2007 > 17730 AND
PROF_COSTI <= 29149 AND
QTA_PART_IVA <= 2 AND
ALIQ_MEDIA_CESS > 2.81 AND
IMP_VAR_RIM_MP > -230 AND
IMP_REDD_LRD_ORD <= 7324 AND
IMP_VE_VOLAFF_2007 > 21591 AND
PROF_COMP <= 55000 AND

PROF_COMP <= 31341 AND
ALTRE_SPS_DOC <= 651 AND
ALIQ_MEDIA_CESS <= 13.87: non_interessante (153.0/17.0)

OP_ESENTI > 23246: non_interessante (112.0/7.0)

IMP_REDD_LRD_ORD > -2340 AND
TOT_IVA_OPE_IMPO <= 93117 AND
IMP_VAR_RIM_MP <= 628 AND
IMP_ONERI_DIV <= 7178 AND
IMP_V_AGG_IMPON <= -4719 AND
ETA > 45: interessante (51.0/5.0)

IMP_REDD_LRD_ORD > -2340 AND
TOT_IVA_OPE_IMPO <= 93117 AND
CRED_CLI_ORD > 1590: non_interessante (75.0/3.0)

IMP_REDD_LRD_ORD > -1303 AND
ALTRI_ACQ_IMP > 71839: interessante (62.0/13.0)

IMP_SPS_DIPEND_ORD_2007 <= 19931 AND
IMP_VAR_RIM_MP <= 5 AND
FLG_STUDIO_SETTORE_07 > 0 AND
QTA_PART_IVA <= 2 AND
SIGLA_PROVINCIA = LU: non_interessante (63.0/14.0)

IMP_COSTI_MP <= 92742 AND
FLG_STUDIO_SETTORE_07 > 0 AND
RIM_FIN_SMPL > 19800: non_interessante (70.0/9.0)

IMP_COSTI_MP <= 92742 AND
IMPST_CRED <= 149 AND
TOT_IMPST_CRED <= 1645 AND
IMP_REDD_LAV_AUT_2007 > 29070: non_interessante (45.0/6.0)

IMP_COSTI_MP > 92742: non_interessante (45.0)

PROF_COMP <= 39045 AND
IMP_CMPNS_TERZI_2007 <= 38 AND
IMPST_CRED <= 149 AND
TOT_IMPST_CRED <= 1645 AND
GRP_ATTIV_2007 = C: interessante (42.0/9.0)

PROF_COMP <= 39045 AND
IMPST_CRED <= 149 AND
IMP_CMPNS_TERZI_2007 <= 38 AND
TOT_IMPST_CRED <= 1645 AND
FLG_STUDIO_SETTORE_07 > 0 AND
ALIQ_MEDIA_CESS > 19.46 AND
ALTRI_ACQ_IMP <= 19289 AND
TOT_IMPST_CRED <= 17: non_interessante (101.0/9.0)

TOT_IMPST_CRED > 1865: non_interessante (40.0/9.0)

GRP_ATTIV_2007 = M AND
ETA <= 54: non_interessante (38.0/13.0)

: interessante (169.0/58.0)

Number of Rules : 15

PART decision list

IMP_REDD_LRD_SMPL_2007 > -7397 AND
IMP_VAR_RIM_PF <= 1578 AND
BENI_DEST_RIV <= 476477 AND
REDD_LORDO_2007 > 52400: non_interessante (233.0/20.0)

TOT_IVA_OPE_IMPO <= 134447 AND
IMP_REDD_LRD_SMPL_2007 > -7397 AND
IMP_V_AGG_IVA > -26801 AND
OP_NON_IMPO_DI <= 5102 AND
SIGLA_PROVINCIA = LI: non_interessante (88.0/6.0)

TOT_IVA_OPE_IMPO <= 134447 AND
IMP_V_AGG_IVA > 4076 AND
IMP_VAR_RIM_MP > -4859 AND
OP_NON_IMPO_DI <= 5102 AND
IMP_V_AGG_IVA <= 187103 AND
QTA_PART_IVA <= 2 AND
FLG_STUDIO_SETTORE_07 > 0 AND
IMPST_CRED <= 1987 AND
IMP_RICAVI_ORDIN_2007 > 0: non_interessante (71.0/2.0)

TOT_IVA_OPE_IMPO <= 134447 AND
CRED_ANNO_PREC <= 5704 AND
OP_NON_IMPO_DI <= 5102 AND
REDD_IMP_2007 > 7019 AND
IMP_VE_VOLAFF_2007 > 15283 AND
TOT_IMPST_DOV <= 2370 AND
QTA_PART_IVA <= 1: non_interessante (288.0/34.0)

TOT_IVA_OPE_IMPO <= 134447 AND
ALIQ_MEDIA_CESS <= 20 AND
TOT_PASS_2007 > 1643 AND
IMP_V_AGG_IVA <= 4076 AND
IMPST_DOV <= 11 AND
TOT_PASS_2007 > 62500: interessante (51.0/14.0)

TOT_IVA_OPE_IMPO <= 134447 AND
IMP_CESS_BENI_AMM <= 917 AND
TOT_PASS_2007 > 1643 AND
TOT_IVA_OPE_IMPO > 998 AND
OP_ESENTI <= 10 AND
IMPST_DOV <= 11536 AND
QTA_PART_IVA <= 2 AND
IMP_COSTI_SERV <= 10157 AND
GRP_ATTIV_2007 = G: non_interessante (60.0/11.0)

TOT_IVA_OPE_IMPO > 134447: non_interessante (41.0)

ALIQ_MEDIA_ACQ > 19.18 AND
TOT_IVA_OPE_IMPO > 2080 AND
IMP_BEN_AMM <= 84 AND
IMP_VE_VOLAFF_2007 > 69670: interessante (33.0)

ALIQ_MEDIA_ACQ > 0.48 AND
TOT_IVA_OPE_IMPO > 998 AND

IMP_CESS_BENI_AMM <= 917 AND
OP_ESENTI <= 10 AND
TOT_PASS_2007 > 2772 AND
DURATA_ATTIV_SOGG <= 30 AND
COSTI_ACQ_MP > 3780: non_interessante (59.0/11.0)

ALIQ_MEDIA_ACQ > 0.48 AND
ALIQ_MEDIA_ACQ > 9.15 AND
TOT_PASS_2007 > 877 AND
ETA <= 63 AND
IMP_PROD_NETTA <= 72697 AND
COSTO_LAV_2007 > 5000 AND
ETA <= 51: interessante (61.0/6.0)

ETA <= 58 AND
ACQ_NO_DETR > 90 AND
ALIQ_MEDIA_ACQ > 13.99: non_interessante (83.0/5.0)

ALIQ_MEDIA_ACQ > 0.48 AND
TOT_PASS_2007 > 2733: interessante (130.0/35.0)

: non_interessante (49.0/8.0)

Number of Rules : 13

PART decision list

IMP_REDD_LRD_ORD <= 121364 AND
OP_ESENTI <= 68395 AND
IMP_REDD_LAV_AUT_2007 <= 29070 AND
PROF_COMP <= 38911 AND
IMP_VAR_RIM_PF <= 1430 AND
RICAVI_ATT_2007 > 14900 AND
IMP_SPS_DIPEND_SMP <= 1404 AND
ALTRI_ACQ_IMP <= 216559 AND
QTA_PART_IVA <= 2 AND
IMPST_CRED <= 1404 AND
ETA > 38: non_interessante (341.0/27.0)

REDD_LORDO_2007 > 59699 AND
ACQ_NO_DETR > 25: non_interessante (207.0/12.0)

IMP_REDD_LAV_AUT_2007 > 29070: non_interessante (69.0/10.0)

REDD_IMP_2007 > 41707 AND
DURATA_ATTIV_SOGG > 12 AND
IMP_REDD_PERD_2007 > 61857: non_interessante (22.0/10.0)

REDD_IMP_2007 <= 41707 AND
REDD_IMP_2007 <= 33282 AND
OP_ESENTI <= 10 AND
OP_NON_IMPO_DI <= 1752 AND
ALIQ_MEDIA_ACQ > 19.99: interessante (57.0/15.0)

REDD_IMP_2007 <= 41707 AND
REDD_IMP_2007 <= 33282 AND
ALIQ_MEDIA_ACQ > 0.95 AND
TOT_IVA_OPE_IMPO <= 960: interessante (47.0/11.0)

```
REDD_IMP_2007 <= 41707 AND
TOT_IVA_OPE_IMPO > 529 AND
OP_ESENTI <= 10 AND
OP_NON_IMPO_DI <= 1752 AND
IMP_REDD_PERD_2007 > 10635 AND
ACQ_NO_DETR <= 2503: non_interessante (199.0/44.0)

ALTRI_ACQ_IMP > 264 AND
REDD_IMP_2007 <= 44300 AND
OP_ESENTI <= 10 AND
PROF_COMP <= 5450 AND
IMP_REDD_PERD_2007 > 116 AND
FLG_NO_COERENZA_07 > 0: interessante (76.0/9.0)

REDD_IMP_2007 <= 41707 AND
TOT_IVA_OPE_IMPO > 1424 AND
IMP_COSTI_BEN_TRZ <= 1878 AND
FLG_NO_COERENZA_07 > 0: non_interessante (53.0/10.0)

TOT_ACQ > 2823 AND
IMP_COSTI_BEN_TRZ <= 1878 AND
TOT_IVA_OPE_IMPO <= 29435 AND
IMP_V_AGG_IVA > 4713 AND
IMP_REDD_PERD_2007 <= 37040: non_interessante (35.0/5.0)

RICAVI_ATT_2007 > 10698 AND
IMP_COSTI_BEN_TRZ <= 1878: interessante (88.0/8.0)

: non_interessante (53.0/3.0)

Number of Rules  :      12


PART decision list
------------------

REDD_LORDO_2007 <= 56492 AND
IMP_REDD_IMP_SMPL_2007 <= 40013 AND
ALTRI_ACQ_IMP <= 182925 AND
FLG_PRES_FAM_07 <= 0 AND
REDD_IMP_2007 > 934 AND
QTA_PART_IVA <= 2 AND
IMP_REDD_LRD_SMPL_2007 <= 34947 AND
ALIQ_MEDIA_ACQ <= 19.99 AND
CRED_ANNO_PREC <= 13486: non_interessante (620.0/129.0)

REDD_LORDO_2007 > 33708 AND
IMP_BEN_STRUM_NA <= 10293: non_interessante (313.0/32.0)

IMP_RICAVI_ORDIN_2007 > 0 AND
IMP_COSTI_SERV <= 143747: non_interessante (68.0/8.0)
GRP_ATTIV_2007 = F: interessante (51.0/14.0)

GRP_ATTIV_2007 = G AND
IMP_V_AGG_IMPON > -6023: non_interessante (39.0/12.0)

: interessante (156.0/47.0)

Number of Rules  :      6
```

PART decision list

OP_ESENTI <= 68395 AND
IMP_V_AGG_IVA > -2320 AND
IMPST_CRED <= 1973 AND
IMPST_CRED <= 206 AND
TOT_IMPST_CRED <= 4739 AND
ACQ_ESENTI <= 3102 AND
IMP_REDD_LAV_AUT_2007 > 28553: non_interessante (102.0/8.0)

OP_ESENTI > 68395: non_interessante (95.0/4.0)

PROF_COMP <= 33276 AND
IMP_CMPNS_ATTIV_2007 <= 15189 AND
IMP_V_AGG_IVA > -4841 AND
IMPST_CRED <= 1973 AND
CESS_NON_IMPO <= 3563 AND
IMPST_CRED <= 556 AND
TOT_IMPST_CRED > 4832: non_interessante (68.0/2.0)

PROF_COMP <= 33276 AND
IMP_CMPNS_ATTIV_2007 <= 15189 AND
CRED_ANNO_PREC <= 10911 AND
RICAVI_ATT_2007 > 18487 AND
IMP_SPS_DIPEND_SMP <= 3641 AND
TOT_IMPST_DOV > 1195: non_interessante (74.0/2.0)

PROF_COMP <= 33276 AND
IMP_CMPNS_ATTIV_2007 <= 15189 AND
TOT_ACQ > 444 AND
ALIQ_MEDIA_CESS > 0 AND
TOT_ACQ > 5289 AND
CRED_ANNO_PREC <= 9714 AND
CESS_NON_IMPO <= 2890 AND
QTA_PART_IVA <= 2 AND
RIM_FIN_SMPL <= 48752 AND
IMP_REDD_LRD_SMPL_2007 > -6490 AND
OP_NON_IMPO_DI <= 5383 AND
RIM_FIN_ORD <= 146730 AND
DEB_FORN_ORD <= 71307 AND
IMP_V_AGG_IMPON <= 142539 AND
TOT_PASS_2007 > 1300 AND
SIGLA_PROVINCIA = FI AND
IMP_TOT_COMP_NEG_2007 > 27794: non_interessante (55.0/5.0)

SIGLA_PROVINCIA = LU AND
TOT_IMPST_CRED <= 4165: non_interessante (83.0/8.0)

IMP_REDD_LRD_SMPL_2007 > -808 AND
IMP_BEN_AMM <= 50993 AND
TOT_ACQ > 444 AND
SIGLA_PROVINCIA = LI: non_interessante (81.0/19.0)

TOT_ACQ > 444 AND
IMP_BEN_AMM <= 50993 AND
SIGLA_PROVINCIA = SI AND
ACQ_ESENTI <= 19 AND

IMPST_CRED <= 472: non_interessante (65.0/5.0)

TOT_ACQ > 444 AND
IMP_BEN_AMM <= 50993 AND
RICAVI_ATT_2007 > 12865 AND
PRES_FAM_SEMPL_07 <= 0 AND
TOT_IMPST_DOV > 466 AND
BENI_DEST_RIV <= 742: interessante (52.0/8.0)

PRES_FAM_SEMPL_07 <= 0 AND
ALIQ_MEDIA_ACQ > 0.95 AND
IMP_VE_VOLAFF_2007 > 12567 AND
IMP_REDD_PERD_2007 > 7409 AND
IMP_BEN_AMM <= 60914 AND
DEB_FORN_ORD <= 246419 AND
CRED_ANNO_PREC <= 9714 AND
IMP_RICAVI_ORDIN_2007 <= 610950 AND
OP_ESENTI <= 346: non_interessante (245.0/60.0)

IMP_VE_VOLAFF_2007 > 0 AND
IMP_BEN_AMM <= 50993 AND
IMP_SPS_DIPEND_SMP <= 45486 AND
IMP_CESS_BENI_AMM <= 1300 AND
IMP_PROD_NETTA <= 148685 AND
IMP_COSTI_BEN_TRZ <= 12057 AND
ETA > 39: interessante (149.0/45.0)

IMP_SPS_DIPEND_SMP > 13513: interessante (41.0/9.0)

DEB_FORN_ORD <= 246419 AND
TOT_IVA_OPE_IMPO > 97541: non_interessante (36.0)

IMP_COSTI_BEN_TRZ <= 5412: non_interessante (69.0/12.0)

: interessante (32.0/8.0)

Number of Rules : 15

PART decision list

IMP_VE_VOLAFF_2007 > 17730 AND
IMP_REDD_LAV_AUT_2007 > 51360: non_interessante (84.0/3.0)

IMP_VAR_RIM_MP > -4859 AND
IMP_VE_VOLAFF_2007 > 17730 AND
TOT_PASS_2007 <= 83361 AND
OP_ESENTI <= 68395 AND
IMP_COSTI_BEN_TRZ <= 923 AND
COSTI_ACQ_MP <= 32265 AND
SIGLA_PROVINCIA = SI: non_interessante (63.0/11.0)

IMP_VAR_RIM_MP > -4859 AND
OP_ESENTI > 59692: non_interessante (72.0/7.0)

IMP_VAR_RIM_MP <= -4859: non_interessante (52.0/2.0)

TOT_IVA_OPE_IMPO > 1248 AND
CESS_NON_IMPO <= 855 AND

IMP_V_AGG_IVA <= 187103 AND
OP_NON_IMPO_DI <= 3536 AND
IMP_V_AGG_IMPON <= 129902 AND
IMP_VAR_RIM_MP > 628: non_interessante (51.0/5.0)

TOT_IVA_OPE_IMPO > 1248 AND
CESS_NON_IMPO <= 855 AND
IMP_V_AGG_IVA > 187103: interessante (43.0/13.0)

IMP_V_AGG_IMPON <= 129902 AND
SPESE_ORD <= 187079 AND
CRED_CLI_ORD <= 177 AND
RICAVI_ATT_2007 > 17597 AND
COSTO_LAV_2007 <= 8649 AND
REDD_IMP_2007 > 934 AND
TOT_IVA_OPE_IMPO <= 21443 AND
TOT_ACQ > 17168: non_interessante (226.0/23.0)

IMP_V_AGG_IMPON <= 129902 AND
PRES_FAM_SEMPL_07 <= 0 AND
CESS_NON_IMPO <= 0 AND
IMP_SPS_DIPEND_SMP > 34624: non_interessante (56.0/10.0)

IMP_V_AGG_IMPON <= 129902 AND
IMP_REDD_IMP_SMPL_2007 <= 29135 AND
CESS_NON_IMPO <= 0 AND
IMP_SPS_DIPEND_SMP <= 23391 AND
IMP_RICAVI_SMPL_2007 > 108550: interessante (54.0/11.0)

IMP_V_AGG_IMPON <= 129902 AND
PRES_FAM_SEMPL_07 <= 0 AND
COSTI_RSDL <= 32541 AND
CESS_NON_IMPO <= 0 AND
RIM_FIN_SMPL <= 11622 AND
TOT_PASS_2007 <= 187877 AND
CRED_CLI_ORD <= 177 AND
GRP_ATTIV_2007 = M AND
FLG_NO_CONGRUENZA_07 <= 0: non_interessante (73.0/21.0)

IMP_V_AGG_IMPON <= 129902 AND
IMP_CMPNS_TERZI_2007 <= 200 AND
IMP_REDD_IMP_SMPL_2007 <= 28472 AND
PRES_FAM_SEMPL_07 <= 0 AND
CESS_NON_IMPO <= 0 AND
TOT_PASS_2007 <= 187877 AND
CRED_CLI_ORD <= 177 AND
TOT_IVA_OPE_IMPO <= 17209 AND
ACQ_ESENTI <= 1 AND
REDD_LORDO_2007 > -2340 AND
TOT_IMPST_CRED <= 46: interessante (130.0/39.0)

IMP_CMPNS_TERZI_2007 <= 200 AND
IMP_V_AGG_IMPON <= 129902 AND
COSTO_LAV_2007 <= 50061 AND
REDD_LORDO_2007 > -1138 AND
PRES_FAM_SEMPL_07 <= 0 AND
DURATA_ATTIV_SOGG <= 30 AND
ALIQ_MEDIA_ACQ <= 19.76: non_interessante (116.0/18.0)

IMP_V_AGG_IMPON <= 129902 AND

RICAVI_ATT_2007 > 17597 AND
IMP_SPS_DIPEND_SMP <= 32683 AND
COSTI_RSDL <= 27983 AND
TOT_PASS_2007 > 86759: interessante (43.0/7.0)

IMP_VE_VOLAFF_2007 > 17730 AND
IMP_COSTI_BEN_TRZ <= 448 AND
ACQ_NO_DETR <= 1303 AND
ACQ_NO_DETR > 0: non_interessante (45.0/2.0)

IMP_COSTI_BEN_TRZ <= 448 AND
DURATA_ATTIV_SOGG > 8 AND
IMP_REDD_PERD_2007 > 11641 AND
ALTRI_ACQ_IMP > 8991: interessante (40.0/13.0)

IMP_COSTI_MP <= 0: interessante (63.0/17.0)

: non_interessante (36.0)

Number of Rules : 17

I risultati che il modello ottiene sul *test set* sono i seguenti:

```
Correctly Classified Instances         267               72.7520 %
Incorrectly Classified Instances       100               27.2480 %
```

La matrice di confusione indica che questo insieme di regole suggerisce un numero di controlli paragonabile a quello suggerito da RIPPER anche se presenta un maggior numero di istanze misclassificate:

```
=== Confusion Matrix ===

   a   b   <-- classified as
 258  23 |   a = non_interessante
  77   9 |   b = interessante
```

Altre metriche sono di seguito evidenziate:

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.918	0.895	0.77	0.918	0.838	0.621	Non_interessante
	0.105	0.082	0.281	0.105	0.153	0.621	Interessante
Weig. Avg.	0.728	0.705	0.656	0.728	0.677	0.621	

Inoltre, come per i modelli già visti, si riportano le misure di carattere "monetario", con l'avvertenza che anche in questo caso, il modello non seleziona il soggetto *outlier*.

- *recupero (32)* = **€ 237.813,00**
- *recupero interessanti (9)* = **€ 171.424,00**
- *recupero non interessanti (23)* = **€ 66.389,00**
- *recupero (media_32)* = **€ 7.431,00**

- *recupero (mediana_32)* = **€ 2.866,00**

Dai dati su riportati, si può osservare come in generale la *performance* del modello sia alquanto bassa. Il *lift chart* viene di seguito presentato (con la solita discontinuità in corrispondenza del *record* eccezionale visto in precedenza). Soprattutto se confrontato con il classificatore ottimo, si evidenzia come i soggetti selezionati non siano quelli che hanno dato origine alle maggiori pretese tributarie.

Tuttavia, il modello presenta delle *AUC* non nettamente peggiori rispetto agli altri: 66,69% con *outlier* e 70,57% in caso contrario. Probabilmente, l'utilizzatore del modello, dati questi risultati e considerato che il modello suggerisce, invero, pochi controlli, potrà decidere di selezionare per verifica fiscale un numero di soggetti maggiore rispetto ai 32 indicati dall'algoritmo, sfruttando il *lift chart* anche come guida per una pianificazione flessibile degli *audit*.

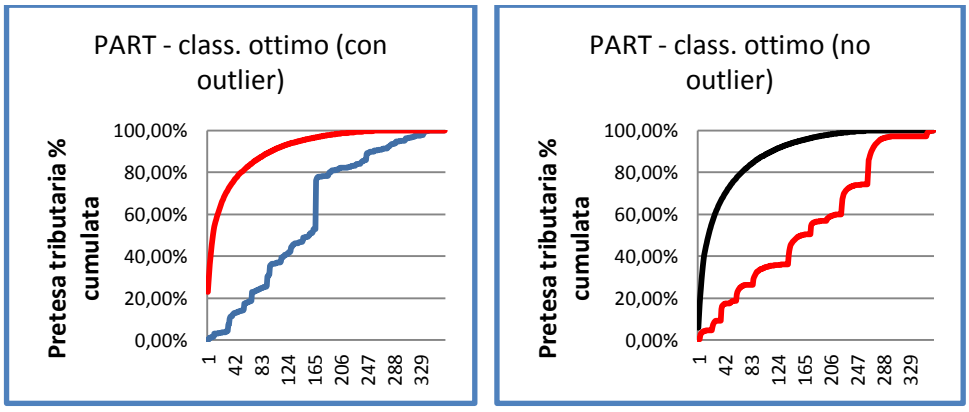

Figura 5.17: *lift chart* modello PART

Quanto all'interpretabilità delle singole regole estratte dal modello, vale la pena di osservare, ancora una volta, come a volte vengano mescolati attributi tratti da quadri diversi della dichiarazione, per cui si evidenziano le problematiche già descritte nell'ambito degli alberi di classificazione e come anche in questo caso l'algoritmo prediliga attributi tratti da quadri compilati dalla generalità dei contribuenti presenti nel *test set*. Infine, si evidenzia anche in questo caso come le varie regole non sembrino portare ad individuare (singolarmente o congiuntamente in gruppi) profili di evasione facilmente riconoscibili o interpretabili.

Conclusioni

Abbiamo introdotto questo lavoro descrivendo, per sommi capi, le caratteristiche dell'evasione fiscale in Italia, fenomeno diffuso, imponente nei numeri (oltre 120 miliardi di euro di imposte evase ogni anno) e foriero di gravi conseguenze sia dal punto di vista economico che dal punto di vista sociale.

Una delle leve con cui tutti i Paesi fanno fronte a tale problema è data dalla deterrenza, che si traduce, sostanzialmente, in controlli efficaci ed efficienti.

In Italia è l'Agenzia delle Entrate la struttura organizzata per far fronte a tutte le fasi che concernono la lotta all'evasione e la *tax compliance*, ma, nonostante gli sforzi profusi, ad oggi, il *gap* esistente tra gettito evaso e recupero delle imposte sottratte a tassazione è ancora molto forte.

Proprio da questa osservazione discendono molti problemi interessanti, cui il presente lavoro ha tentato di dare una risposta, o per lo meno delle indicazioni. In particolare, essi si incentrano sulla valutazione della possibilità che tecniche di *data mining* e, in particolare, di classificazione, possano essere applicate con successo in questo ambito.

Le esperienze condotte sembrano poter autorizzare ad una risposta positiva.

Abbiamo visto come utilizzare varie tecniche di classificazione, con particolare riferimento ad alberi di classificazione e regole decisionali, scelti per la loro spiccata espressività ed interpretabilità, implementando in vario modo gli algoritmi messi a disposizione dalla *suite Weka*, con la precisazione che la trattazione, non pretendendo di illustrare tutti i possibili modelli implementabili sul *dataset* a disposizione, ha inteso piuttosto descrivere le logiche che guidano il "settaggio" di determinati parametri, sia dei singoli classificatori che dei metaclassificatori, e le modalità per utilizzare nel modo più proficuo possibile l'*output* di detti modelli.

Classificatori opportunamente allenati (anche attraverso l'impiego di *metaclassificatori*) hanno mostrato di ottenere buone *performances*, suggerendo insiemi di soggetti da controllare in cui gli evasori *interessanti* erano in proporzione maggiore rispetto a quanti ne ve ne fossero nel *dataset* (si ricorda che tutti i soggetti del *dataset* sono stati oggetto di accertamento, ma solo una parte di essi, peraltro minoritaria, ha dato luogo a controlli realmente proficui per l'erario) e ottenendo buoni risultati anche in termini di gettito recuperato.

Non solo, ma attraverso l'introduzione di adeguati *lift chart* associati ai modelli, viene fornito all'utilizzatore uno strumento per selezionare, sulla base dei propri obiettivi, un numero desiderato di soggetti da sottoporre a controllo. L'implementazione *software* del processo descritto in questo capitolo – preparazione dei dati (compresa una fase di selezione dei soggetti per attività svolta), generazione e confronto di modelli, generazione dei *lift chart* – può senz'altro costituire un nuovo strumento di lotta all'evasione per questo Paese.

In particolare, la novità introdotta da questo lavoro è l'idea che i *lift chart* associati ai modelli, possano essere utilizzati come guida flessibile per la pianificazione fiscale, coniugando, da una parte, le esigenze ed obiettivi dell'utilizzatore e dall'altro l'*output* di buoni modelli di classificazione.

L'intervento, nel corso dello svolgimento di questa tesi, dell'Ufficio Studi della stessa Agenzia delle Entrate, senza cui peraltro questo lavoro non avrebbe mai potuto vedere la luce, mostra l'interesse della stessa Agenzia per le tematiche qui sviluppate e i risultati ottenuti saranno presi in considerazione al fine di implementare nuovi mezzi di lotta all'evasione fiscale.

Ringraziamenti

Desidero innanzitutto ringraziare la mia famiglia, per avermi trasmesso l'amore per lo studio e la tenacia nel coltivarlo.

Un ulteriore grazie alla mia famiglia "acquisita", per non avermi mai fatto mancare il suo supporto, con particolare "menzione" a Leda, cuoca sopraffina e Virginia, una suocera tutta speciale.

Ringrazio naturalmente il "team" di *data mining* dell'Università di Pisa e del CNR, in particolare Diego Pennacchioli per avermi più di tutti seguito in questo lavoro e i Proff. Dino Pedreschi e Fosca Giannotti per avermi fatto conoscere e appassionare a questa materia, così interessante, innovativa e dalle applicazioni più disparate.

Se l'Agenzia delle Entrate sarà in grado di affrontare con strumenti innovativi le sfide del futuro è merito di persone come Stefano Pisani e Andrea Spingola: a loro va il mio grazie per aver dato fiducia a un collega "perfetto sconosciuto".

Questa seconda laurea è stata una sfida e un sogno che si realizza ma, come tutti i sogni, ha richiesto molti sacrifici, sottraendo tempo e d energie agli affetti più cari; per questo ringrazio mia moglie Valentina per l'amore e la pazienza, e mio figlio Diego, arrivato nel mezzo di quest'avventura, per avermi insegnato a far fruttare al meglio le notti insonni.

Infine, un ricordo speciale va a mia madre: oggi come allora sei sempre nei miei pensieri e il tuo sorriso si riflette nel mio.

Bibliografia

[AEF98] Androni J., Erard B., Feinstein J., Tax Compliance, *Journal of Economic Literature*, vol. 36, n. 2, 1998, pp. 818-860.

[Alb10] Albano A., Basi di dati a supporto delle decisioni, Appunti delle lezioni del corso di Sistemi informatici direzionali, Università di Pisa, 2010.

[AS72] Allingham M.G., SandmoA., Income tax evasion: a theoretical analysis, *Journal of Public Economics,* n.1, 1972, pp. 323-338.

[AT12] Ameur F., Tkiouat M., Taxpayers Fraudulent Behavior Modeling. The Use of Datamining in Fiscal Fraud Detecting: the Moroccan Case*, Applied Mathematics,* vol. 3 n. 10, 2012, pp. 1207-1213.

[BFOS84] Breiman L., Friedman J.H., Olshen R.A., Stone C.J., *Classification and Regression Trees*, Wadsworth International Group, Belmont, California, 1984.

[BGMP99] Bonchi F., Giannotti F., Mainetto G., Pedreschi D., A Classification-based Methodology for Planning Auditing Strategies in Fraud Detection, in *Proc. of SIGKDD99,* 1999, pp. 175-184.

[BGMPS09] Basta S., Giannotti F., Manco G., Pedreschi D., Spisanti L., SNIPER: a data mining methodology for fiscal fraud detection, in *Mathematics for finance and economy. Special issue of ERCIM News,* n. 78, 2009, pp. 27-28.

[BL04] Berry M., Linoff G., *Data Mining Techniques for marketing, sales, and customer relationship management (2nd ed.)*, Wiley Publishing Inc., Indianapolis, 2004.

[BL97] Berry M., Linoff G., *Data mining techniques for marketing, sales and customer support*, Wiley Publishing Inc., New York, 1997.

[Bre01] Breiman L., Random Forests, *Machine Learning,* vol. 45, n. 1, 2001, pp. 5-32.

[Bre96] Breiman L., Bagging Predictors, *Machine Learning*, vol. 24, n. 2, 1996, pp. 123-140.

[Bri50] Brier G.W., Verification of Forecasts Expressed in Terms of Probability, *Monthly Weather Review,* n. 78, 1950, pp. 1-3.

[CaDal07] Cannari L., D'Alessio G., Le opinioni degli italiani sull'evasione fiscale. *Temi di discussione del Servizio Studi Banca d'Italia*, Num. 618 – febbraio 2007.

[CC99] Cerullo M.J., Cerullo V., Using neural networks to predict financial reporting fraud. *Computer Fraud & Security,* May/June 1999, pp. 14–17.

[CCD95] Cannari L., Ceriani V., D'Alessio G., Il recupero degli imponibili sottratti a tassazione, *Ricerche quantitative per la politica economica*, Banca d'Italia, vol. II, 1995.

[CdC13] Corte del Conti, *Elementi per l'audizione del Presidente della corte dei Conti presso le commissioni bilancio V e Finanze VI della Camera dei Deputati – Considerazioni in merito alle strategie e agli strumenti per il contrasto dell'evasione fiscale.* Roma, 19 giugno 2013.

[CMMT04] Cummings R.D., Martinez-Vazquez J., McKee M., Torgler B., Effects of Culture on Tax Compliance: A Cross Check of Environmental and Survey Evidence, *International Studies Program Working Paper,* n. 04-03, 2004.

[Codd93] Codd E.F., *Providing OLAP to User Analysts: An IT Mandate*, E. F.Codd & Associates, 1993.

[Coh95] Cohen W.W., Fast Effective Rule Induction, in: *Twelfth International Conference on Machine Learning,* 1995, pp. 115-123.

[Conf10] Confindustria – Centro Studi. *Scenari economici,* n. 9, Autunno 2010.

[Cro44] Crowe M.T., The moral obligation of paying just taxes, *The Catholic University of America, Studies in Sacred Theology*, n. 84, 1944.

[DB95] Dietterich T.G., Bakiri G., Solving multi-class learning problems via error-correcting output codes*, Journal of Artificial Intelligence Research,* n. 2, 1995, pp. 263-286.

[DGSC97] Dorronsoro J.R., Ginel F., Sánchez C., Cruz C.S., Neural fraud detection in credit card operations, *IEEE Transactions on Neural Networks,* vol. 8, n. 4, 1997, pp. 827–834.

[Dom99] Domingos P., MetaCost: A general method for making classifiers cost-sensitive, in: *Fifth International Conference on Knowledge Discovery and Data Mining,* 1999, pp. 155-164.

[Faw04] Fawcett T.. Roc Graphs: Notes and Practical Considerations for Researchers, march 2004.

[FCS95] Fanning K., Cogger K., Srivastava R., Detection of management fraud: A Neural Network Approach, *Journal of Intelligent Systems in Accounting, Finance and Management* vol. 4, 1995, pp. 113-126.

[FS96] Freund Y., Schapire R.E., Experiments with a new boosting algorithm, in *Thirteenth International Conference on Machine Learning, San Francisco,* 1996, pp. 148-156.

[FS97] Freund Y., Schapire R.E., A decision-theoretic generalization of on-line learning and an application to boosting, *Journal of Computer and System Sciences (Elsevier),* vol. 55, n. 1, 1997, pp. 119–139.

[FS04] Foster D., Stine R., Variable Selection in Data Mining: Building a Predictive Model for Bankruptcy, *Journal of American Statistical Association,* n. 99, 2004, pp. 303-313.

[FW98] Frank E., Witten I., Generating Accurate Rule Sets Without Global Optimization, in *Fifteenth International Conference on Machine Learning,* 1998, pp. 144-151.

[GSZ06] Guiso L., Sapienza P., Zingales L., Does Culture Affect Economic Outcomes? *Journal of Economic Perspectives,* Vol. 20, n. 2, 2006, pp. 23–48.

[HK06] Han J., Kamber M., *Data mining Concepts and techniques*, Morgan Kaufmann Publishers, second edition, 2006.

[Hoo09] Hoover J.N., States use BI, data warehousing to recoup unpaid taxes, *Intelligent Enterprise,* n. 12, 2009.

[Hun62] Hunt E.B., *Concept learning: an information processing problem,* Wiley Publishing Inc., New York, 1962.

[Kim96] Kimball R., *The Data Warehouse Toolkit*, J. Wiley & Sons, New York,1996.

[KKTT06] Kotsiantis S., Koumanakos E., Tzelepis D., Tampakas V., Forecasting fraudulent financial statements using data mining, *International Journal of Computational Intelligence,* vol. 3, n. 2, 2006, pp. 104–110.

[KPHKB03] Kim H., Pang S., Je H., Kim D., Bang S., Constructing Support Vector Machine Ensemble, *Pattern Recognition,* n. 36, 2003, pp. 2757-2767.

[KR02] Kimball R., Ross M., *Data warehouse. La guida completa*, Hoepli, Milano 2002.

[KSM07] Kirkos E., Spathis C., Manolopoulos Y., Data mining techniques for the detection of fraudulent financial statements, *Expert Systems with Applications,* vol. 32, n. 4, 2007, pp. 995–1003.

[Mef11] Ministero dell'Economia e delle finanze. Gruppo di lavoro *Economia non osservata e flussi finanziari*. Roma, 14 luglio 2011.

[NHWCS11] Ngai E.W.T., Hu Y., Wong Y.H., Chen Y., Sun X., The application of data mining techniques in financial fraud detection: a classification framework and an academic review of literature, *Journal Decision Support Systems,* vol. 50, n. 3, 2011, pp. 559-569.

[Oecd10] Organisation for economic co-operation and development. *Understanding and Influencing Taxpayers' Compliance Behaviour*, November 2010.

[Pal04] Palmieri I., Metodologie utilizzate per quantificare l'evasione fiscale con particolare riferimento ai metodi fondati sugli accertamenti. *Agenzia delle Entrate, Documenti di lavoro dell'Ufficio Studi*, 2004.

[PAL04] Phua C., Alahakoon D., Lee V., Minority Report in Fraud Detection: Classification of Skewed Data, in *SIGKDD Explorations,* 2004, pp. 50-59.

[Pen04] Pendse N., "What Is OLAP?", 2004 *http://barc-research.com/bi-verdict/ (www.olapreport.com)*

[PLSG05] Phua C., Lee V., Smith K., Gayler R., A comprehensive survey of data mining-based fraud detection research. *Artificial Intelligence Review,* 2005, pp. 1–14.

[Qui79] Quinlan R., Discovering rules from large collections of examples: a case study, in *D. Michie Ed. Expert systems in the micro electronic age*, Edinburgh University Press, 1979.

[Qui87] Quinlan R., Simplifying decision trees, *International journal of machine studies,* n. 27, 1897, pp. 221-234.

[Qui93] Quinlan R., *C4.5: Programs for Machine Learning*, Morgan Kaufmann Publishers, San Mateo, CA, 1993.

[SK04] Schneider F., Klinglmair R., Shadow economies around the world: what do we know?, *CESifo working paper,* n. 1167, 2004.

[Spa02] Spathis C.T., Detecting false financial statements using published data: some evidence from Greece, *Managerial Auditing Journal,* vol. 17, n. 4, 2002, pp. 179–191.

[TSK06] Tan P., Steinbach M., Kumar V., *Introduction to Data Mining*, Addison Wesley, 2006.

[VDD04] Viaene S., Derrig R.A., Dedene G., A case study of applying boosting naive Bayes to claim fraud diagnosis, *IEEE Transactions on Knowledge and Data Engineering*, vol. 16, n. 5, 2004, pp. 612–620.

[Wen04] Wenzel M., An analysis of norm processes in tax compliance, *Journal of Economic Psychology,* n. 25, 2004, pp. 213-228.

[WFH11] Witten I., Frank E., Hall M., *Data Mining Practical machine learning tools and techniques*, Morgan Kaufmann, 3rd edition, 2011.

[WFYH03] Wang H., Fan W., Yu P., Han J., Mining Concept-Drifting Data Streams Using Ensemble Classifiers, in *Proc. of SIGKDD03,* 2003, pp. 226-235.

[WLTH06] Wang J., Liao Y., Tsai T., Hung G., Technology-based financial frauds in Taiwan: issue and approaches, in *IEEE Conference on: Systems, Man and Cyberspace,* 2006, pp. 1120–1124.

[WOLCY12] Wu R., Ou C., Lin H., Chang S., Yen D., Using data mining technique to enhance tax evasion detection performance, *Expert Systems Applications,* vol. 39, n. 10, 2012, pp. 8769-8777.

[YQJ03] Yu F., Qin Z., Jia X., Data mining application issues in fraudulent tax declaration detection, in *International Conference on Machine Learning and Cybernetics,* vol. 4, 2003, pp. 2202-2206.

[Zam12] Zamagni S., Il contribuente virtuoso: come vincere la lotta all'evasione fiscale, AICCON c/o Università di Bologna - Facoltà di Economia, 2012.

Appendice A

Descrizione del *dataset* oggetto di analisi

Sono di seguito descritti, sinteticamente, i campi del dataset di riferimento per le analisi condotte in questo lavoro (per maggiori dettagli sul contenuto dei righi dei vari quadri, si rimanda alle istruzioni del modello UnicoPF2008, disponibili sul sito internet dell'Agenzia delle Entrate).

1.1 Attributi di natura "anagrafica"

Ogni soggetto, di cui per ovvie ragioni non vengono indicati né il codice fiscale né la partita IVA, viene descritto da una serie di dati anagrafici:

[CODICE_CATASTALE]	Codice catastale del comune di residenza del soggetto. Il codice catastale è un codice unico identificativo assegnato ad ogni comune italiano, composto da una lettera seguita da tre valori numerici.
[SIGLA_PROVINCIA]	Sigla della provincia cui appartiene il comune di residenza.
[COD_REG]	Codice regione. Nel caso di specie, alla Toscana è attribuito il valore "9", unico valore presente
[DESCR_REG]	Descrizione della regione (Toscana)
[AREA_AGENZIA]	Unico valore "Centro"
[AREA]	Unico valore "Centro"
[CODEC_CF]	Codice che identifica univocamente un soggetto (chiave primaria)

[SESSO]	M = maschio, F = femmina
[QTA_PART_IVA]	Numero totale di partite iva che il soggetto ha posseduto (in un dato momento un soggetto può essere titolare di una sola partita iva, ma nel corso degli anni può averne aperte e chiuse più d'una).
[MIN_DATA_INI_ATTIV]	Data di apertura della prima partita iva del soggetto.
[COM_ESER_ATT]	Codice del comune ove il soggetto esercita l'attività (se sono più d'uno, ne viene comunque riportato uno solo)
[COD_STATO_EST_RES]	Se [FLG_RES_ESTERO]=0 allora il codice dello stato estero è "ZZZZ", altrimenti è il codice dello stato di residenza.
[ETA]	Età anagrafica del soggetto, calcolata al 2007
[DURATA_ATTIV_SOGG]	Indica da quanti anni viene svolta l'attività, rispetto al 2007.

1.2 Attributi afferenti imposte dirette e IVA

Di seguito sono descritti i dati reperiti dalle dichiarazioni presentate per l'anno d'imposta 2007 da ciascun soggetto presente nel *dataset*.

Ciascun attributo preso in considerazione in questa sezione corrisponde ad un particolare quadro/rigo della dichiarazione dei redditi presentata dai vari contribuenti. Il rigo/quadro riportato fa riferimento al modello Unico08 Persone fisiche, modello nel quale andavano dichiarati i redditi per l'anno 2007.

In particolare abbiamo:

a) *dati delle dichiarazioni ai fini IRAP*.

La sezione I deve essere compilata dalle persone fisiche esercenti attività commerciali ai sensi dell'art. 55 del TUIR (DPR 917/86), a prescindere dal regime di contabilità adottato. Si tratta, in generale, degli stessi soggetti tenuti, a seconda dei casi, alla presentazione del quadro RF ovvero del quadro RG ai fini della dichiarazione dei redditi. Righi IQ1-IQ16.

La sezione III va compilata dai produttori agricoli titolari di reddito agrario di cui all'art. 32 del TUIR, dagli esercenti attività di allevamento di animali che determinano il reddito eccedente i limiti dello stesso art. 32 secondo le disposizioni del successivo art. 56, comma 5, del TUIR nonché dai soggetti che esercitano attività di agriturismo e che si avvalgono, ai fini delle imposte sui redditi, del regime forfetario di cui all'art. 5 della legge 30 dicembre 1991, n. 413. Righi IQ21-IQ23

La sezione IV va compilata dai soggetti esercenti arti e professioni di cui all'art. 53, comma 1, del TUIR, per i quali, ai

sensi dell'art. 8 del D. Lgs. n. 446, la base imponibile si determina sottraendo dall'ammontare dei compensi percepiti nel periodo d'imposta l'ammontare dei costi inerenti all'attività sostenuti nello stesso periodo d'imposta, compreso l'ammortamento dei beni materiali e immateriali ed esclusi gli interessi passivi e le spese per il personale dipendente. Righi IQ24-IQ26

[IMP_VAR_RIM_PF] IQ2 COL.1	Importo complessivo delle variazioni dei prodotti in corso di lavorazione, semilavorati e finiti.
[IMP_VAR_RIM_MP] IQ12 COL.1	Importo complessivo delle variazioni (di segno positivo o negativo) delle rimanenze di materie prime, sussidiarie e merci.
[IMP_COSTI_MP] IQ7 COL.1	Costi per materie prime, sussidiarie e merci.
[IMP_COSTI_SERV] IQ8 COL.1	Costi per servizi. Si riportano i valori contabili dei costi per servizi.
[IMP_COSTI_BEN_TRZ] IQ9 COL.1	Costi per il godimento di beni di terzi. Sono riportati i valori contabili.
[IMP_ONERI_DIV] IQ13 COL.1	Oneri diversi di gestione. Sono riportati i valori contabili.
[RF_REDD_IMPR] IQ17 COL.1	Importo del reddito d'impresa determinato forfetariamente per l'anno 2007.
[PR_AGR_CORR] IQ21 COL.1	Ammontare dei corrispettivi soggetti a registrazione ai fini dell'Iva, compresi i corrispettivi per le cessioni di beni strumentali.
[PR_AGR_ACQ] IQ22 COL.1	Ammontare degli acquisti inerenti l'attività agricola, soggetti a registrazione ai fini IVA.
[PROF_COMP] IQ24 COL.2	Compensi derivanti dall'attività professionale o artistica. Di fatto è pari alla somma degli importi indicati nei righi RE2, colonne 1 e 2, RE3, colonna 2 (con esclusione dell'importo indicato in colonna 1), RE4 e RE5, colonna 3 della dichiarazione dei redditi (con esclusione dei compensi di cui alla colonna 1 del rigo RE5 dichiarati per adeguamento ai parametri, ai sensi dell'art. 4, comma 2, del D.P.R. n. 195 del 1999.
[PROF_COSTI] IQ25 COL.1	Costi inerenti all'attività esercitata
[IMP_PROD_NETTA] IQ36 COL.1	Valore della produzione netta (nel caso in cui il risultato sia negativo va indicato 0)
[IMP_TOT_IMPST_IRAP] IQ90 COL.1	Totale imposta

b) *dati del quadro **RE**.*
Il quadro RE deve essere utilizzato per dichiarare i redditi derivanti dall'esercizio di arti e professioni indicati nel comma 1 dell'art. 53 del TUIR, rientranti nel regime analitico, i redditi

rientranti nel regime fiscale agevolato di cui agli artt. 13 e 14 della legge 23 dicembre 2000 n.388, nonché i proventi percepiti per prestazioni di volontariato o cooperazione rese ad organizzazioni non governative riconosciute idonee ai sensi dell'art. 28 della legge 26 febbraio 1987, n. 49, qualora dette prestazioni discendano dall'assunzione di obblighi riconducibili ad un rapporto di lavoro autonomo.
Gli altri redditi di lavoro autonomo indicati nel comma 2 dell'art. 53 del TUIR vanno dichiarati nel quadro RL.

[IMP_CMPNS_ATTIV] RE2 COL.2	Ammontare lordo complessivo dei compensi, in denaro e in natura, anche sotto forma di partecipazione agli utili, al netto dell'Iva, derivanti dall'attività professionale o artistica, percepiti nell'anno, compresi quelli derivanti da attività svolte all'estero.
[IMP_CMPNS_SZ1] RE6 COL.1	Somma dei compensi e proventi dei righi RE2, colonna 2, RE3, colonna 2, RE4, e RE5 colonna 3.
[IMP_LAV_DIPEND] RE11 COL.1	Ammontare complessivo di quanto corrisposto a titolo di retribuzione al lordo dei contributi assistenziali e previdenziali a lavoratori dipendenti ed assimilati.
[IMP_CMPNS_TERZI] RE12 COL.1	Ammontare complessivo dei compensi corrisposti a terzi per prestazioni professionali e servizi direttamente afferenti l'attività artistica o professionale del contribuente.
[ALTRE_SPS_DOC] RE19 COL.1	Spese sostenute, tra cui l'80 per cento delle spese di manutenzione relative a talune apparecchiature terminali per servizi di comunicazione elettronica; il 40 per cento delle spese sostenute nel periodo d'imposta, limitatamente a un solo veicolo, per l'acquisto di carburanti, lubrificanti e simili utilizzati esclusivamente per la trazione di ciclomotori e motocicli, nonché il 90 per cento di tali spese sostenute relativamente ai detti veicoli dati in uso promiscuo ai dipendenti; il 50 per cento delle spese di impiego dei beni mobili adibiti promiscuamente all'esercizio dell'arte o della professione e all'uso personale o familiare del contribuente e utilizzati in base a contratto di locazione finanziaria o di noleggio; l'ammontare delle altre spese inerenti l'attività professionale o artistica effettivamente sostenute e debitamente documentate.
[IMP_TOTALE_SPESE] RE20 COL.1	Totale delle spese, sommando gli importi da rigo RE7 a rigo RE19.
[IMP_REDD_AUT_SZ1] RE23 COL.1	Qualora non sia stato compilato il rigo RE22, va indicata la somma tra: – l'importo di rigo RE21, colonna 2, al netto di quello eventualmente indicato nella colonna 1

	del medesimo rigo; – il 10 per cento dell'importo di rigo RE21, colonna 1; – l'importo di rigo RE2, colonna 1. Nell'ipotesi in cui, invece, sia stato compilato il rigo RE22, nel presente rigo va indicato l'importo di rigo RE2, colonna 1.
[IMP_REDD_LAV_AUT] RE25 COL.1	Differenza tra l'importo di rigo RE23 (se positivo) ed il rigo RE24, colonna 2.Tale reddito va sommato agli altri redditi Irpef e riportato nel quadro RN.

c) *dati del quadro **RD***

Il quadro RD deve essere utilizzato per dichiarare il reddito derivante dall'attività di allevamento di animali e/o da quelle dirette alla produzione di vegetali eccedenti il limite di cui all'art. 32, comma 2, lett. b), del TUIR, qualora detto reddito sia determinato ai sensi del comma 5 dell'art. 56 (Sezione I) e/o del comma 1 dell'art. 56-bis (Sezione II). La Sezione III deve essere utilizzata per dichiarare i redditi derivanti dalle altre attività agricole di cui ai commi 2 e 3 dell'art. 56-bis, nonché quelli dei soggetti che esercitano attività di agriturismo, di cui alla legge n. 96 del 20 febbraio 2006, e che determinano il reddito secondo i criteri previsti dall'art. 5, comma 1, della legge n. 413 del 1991.

[AGR_REDD] RD18 COL.1	Differenza tra i righi RD16 e RD17, colonna 2. L'importo di rigo RD18 deve essere riportato, unitamente agli altri redditi, nel quadro RN.

d) *dati del quadro **RF**.*

Il quadro RF deve essere compilato dagli esercenti imprese commerciali in regime di contabilità ordinaria e da quelli che, pur potendosi avvalere della contabilità semplificata e determinare il reddito ai sensi dell'art. 66 del TUIR, hanno optato per il regime ordinario.

[UTILE_PERD_CIV_ORD] RF2 COL.1 xor -RF3 COL1	Il reddito d'impresa è determinato apportando all'utile (o alla perdita) risultante dal conto economico, da indicare rispettivamente nel rigo RF2 o RF3, le variazioni in aumento e in diminuzione conseguenti all'applicazione delle disposizioni contenute nel TUIR o in altre leggi.
[IMP_REDD_LRD_ORD] RF46 COL.2	Reddito d'impresa lordo (o perdita, se negativo).
[IMP_REDD_IMP_ORD]	Reddito d'impresa (o perdita), pari alla

RF49 COL.1	differenza tra l'importo di rigo RF46, colonna 2 e le erogazioni liberali di rigo RF47. Qualora nel rigo RF46, colonna 2 sia indicata una perdita, nel rigo RF49, deve essere esposta la perdita ridotta dell'importo del rigo RF48.
[RIM_FIN_ORD] RF58 COL.1	Valore iscritto in bilancio delle rimanenze finali relative a materie prime, sussidiarie, materiali di consumo (costituiti da materiali usati indirettamente nella produzione); prodotti in corso di lavorazione e semilavorati; lavori in corso su ordinazione; prodotti finiti e merci; acconti per forniture da ricevere.
[CRED_CLI_ORD] RF59 COL.1	Importo dei crediti iscritti in bilancio nei confronti dei clienti e derivanti dalla cessione di beni e dalla prestazione di servizi che rientrano nell'attività propria dell'azienda.
[DEB_FORN_ORD] RF70 COL.1	Importo iscritto in bilancio dei debiti verso i fornitori, derivanti dalla acquisizione di beni e servizi.
[IMP_RICAVI_ORDIN] RF74 COL.1	Ricavi delle vendite (va indicato l'ammontare dei ricavi di cui alle lett. a) e b) del comma 1 dell'art. 85 del DPR 917/86 cioè dei corrispettivi di cessioni di beni e delle prestazioni di servizi alla cui produzione o al cui scambio è diretta l'attività dell'impresa e dei corrispettivi delle cessioni di materie prime e sussidiarie, di semilavorati e di altri beni mobili, esclusi quelli strumentali, acquistati o prodotti per essere impiegati nella produzione)
[IMP_SPS_DIPEND_ORD] RF75 COL.1	Va indicata la spesa per lavoro dipendente e assimilato.
[SPESE_ORD] RF75 COL.2	Ammontare degli oneri di produzione e vendita.

e) *dati del quadro **RG**.*

Il quadro RG deve essere compilato dagli esercenti attività commerciali in contabilità semplificata di cui all'art. 18 del D.P.R. n. 600 del 1973. Gli esercenti attività commerciali in regime di contabilità semplificata (sempre che non abbiano optato per il regime di contabilità ordinaria) determinano il reddito ai sensi dell'art. 66 del DPR 917/86 se nel periodo d'imposta precedente hanno conseguito ricavi per un ammontare non superiore:

- a € 309.874,14, se trattasi di imprese aventi per oggetto prestazioni di servizi;
- a € 516.456,90, se trattasi di imprese aventi per oggetto altre attività;

[IMP_RICAVI_SMPL] RG2 COL.2	Ammontare dei ricavi di cui alle lettere a) e b) del comma 1 dell'art. 85 del TUIR, costituiti dai corrispettivi delle cessioni di beni e delle prestazioni di servizi alla cui produzione o al cui scambio è diretta l'attività dell'impresa e dai corrispettivi delle cessioni di materie prime e sussidiarie, di semilavorati e di altri beni mobili, esclusi quelli strumentali, acquistati o prodotti per essere impiegati nella produzione.
[RIM_FIN_SMPL] RG7 COL.1	Valore delle rimanenze finali relative a: • materie prime e sussidiarie, semilavorati, merci e prodotti finiti (art. 92, comma 1, del TUIR); • prodotti in corso di lavorazione e servizi non di durata ultrannuale (art. 92, comma 6, del TUIR).
[RIM_INI_SMPL] RG11 COL.1	Esistenze iniziali al 1° gennaio del periodo d'imposta oggetto della presente dichiarazione relative a materie prime e sussidiarie, semilavorati, merci e prodotti finiti nonché ai prodotti in corso di lavorazione e ai servizi di durata non ultrannuale.
[IMP_TOT_CMPN_PSV] RG10 COL.1	Totale dei componenti positivi, risultante dalla somma degli importi indicati nei righi da RG2 a RG9.
[COSTI_ACQ_MP] RG13 COL.1	Costo di acquisto di materie prime e sussidiarie, semilavorati e merci, incluse le spese sostenute per le lavorazioni effettuate da terzi esterni all'impresa.
[IMP_SPS_DIPEND_SMP] RG14 COL.1	Ammontare delle spese per prestazioni di lavoro dipendente, assimilato ed autonomo.
[COSTI_RSDL] RG20 COL.1	Altri componenti negativi deducibili non indicati nei precedenti righi.
[IMP_TOT_COMP_NEG] RG22 COL.1	Totale dei componenti negativi risultante dalla somma degli importi indicati nei righi da RG11 a RG21.
[IMP_REDD_LRD_SMPL] RG26 COL.2	Importo derivante dalla seguente somma algebrica : RG23 + RG24 col. 4 – RG25 col. 3
[IMP_REDD_IMP_SMPL] RG34 COL.1	Differenza positiva tra l'importo di rigo RG32, colonna 2 e quello di rigo RG33, colonna 2.

f) *dati del quadro **RN**:*

[IMP_REDD_PERD] RN1 COL.2	Reddito complessivo dato dalla somma dei singoli redditi indicati nei vari quadri.
[REDD_IMP] RN4 COL.1	Questo rigo serve per calcolare il reddito imponibile sul quale applicare l'imposta, pari a: RN1 col. 1 + RN1 col. 2 – RN2 – RN3
[IMPST_LOR]	Imposta lorda corrispondente al vostro reddito

RN5 COL.1	imponibile di rigo RN4.

g) *dati dei quadri **IVA:***

[COD_ATTIVITA] VA2 COL.1	Codice dell'attività svolta.
[IMP_BENI _AMM] VA3 COL.1	Acquisti beni ammortizzabili.
[IMP_BEN_STRUM_NA] VA3 COL.2	Acquisti beni strumentali non ammortizzabili.
[BENI_DEST_RIV] VA3 COL.3	Acquisti beni destinati alla rivendita.
[ALTRI_ACQ_IMP] VA3 COL.4	Altri acquisti e importazioni.
[CESS_BENI_INTRA] VA30 COL.1	Dato complessivo delle cessioni intra di beni.
[PREST_SERV_INTRA] VA30 COL.2	Dato complessivo delle prestazioni di servizi intra.
[ACQ_BENI_INTRA] VA31 COL.1	Dato complessivo degli acquisti intracomunitari di beni.
[ACQ_BENI_INTRA_IMPST] VA31 COL.2	Imposta relativa agli acquisti Intra.
[IMPORT_IMPO] VA32 COL.1	Dati complessivi relativi alle importazioni di beni risultanti dalle bollette doganali registrate nel periodo d'imposta.
[IMPORT_IMPST] VA32 COL.2	Imposta relativa alle operazioni imponibili anche se non detraibile ai sensi dell'art. 19-bis1 o di altre disposizioni.
[EXP_IMPO] VA33 COL.1	Ammontare complessivo delle esportazioni di beni effettuate nell'anno, risultanti dalle dichiarazioni doganali, di cui all'art. 8, primo comma, lettere a) e b).
[IMP_VE_SEZ1] VE10 COL.1	Totale degli imponibili e dell'imposta, determinato sommando gli importi riportati ai righi da VE1 a VE9, rispettivamente della colonna degli imponibili e della colonna dell'imposta.
[IMPST_VE_SEZ1] VE10 COL.2	
[IMP_OPER_IMPVE23] VE23 COL.1	Totale degli imponibili determinato sommando gli importi riportati ai righi da VE20 a VE22, nella colonna degli imponibili.
[IMP_OP_IMPVE23_IMPS] VE23 COL.2	Totale delle imposte, determinato sommando gli importi riportati ai righi da VE20 a VE22, nella colonna delle imposte.
[CESS_NON_IMPO] VE30 COL.1	Ammontare delle esportazioni e delle altre operazioni non imponibili che concorrono alla formazione del plafond di cui all'art. 2, comma 2, della legge 18 febbraio 1997, n. 28.
[OP_NON_IMPO_DI] VE31 COL.1	Ammontare delle operazioni non imponibili effettuate nei confronti di esportatori che abbiano rilasciato la dichiarazione di intento.
[ALTRE_NON_IMPO] VE32 COL.1	Ammontare delle altre operazioni qualificate non imponibili.
[OP_ESENTI] VE33 COL.1	Ammontare delle operazioni esenti di cui all'art. 10 e delle operazioni dichiarate

	esenti da altre disposizioni, come ad esempio quelle di cui all'art. 6, della legge n. 133 del 1999.
[IMP_CESS_BENI_AMM] VE38 COL.1	Operazioni (al netto dell'IVA) non rientranti nel volume d'affari.
[IMP_VE_VOLAFF] VE40 COL.1	Volume d'affari, determinato sommando gli importi indicati ai righi VE10 colonna 1 (volume d'affari agricoltori), VE23 colonna 1 ed ai righi da VE30 a VE37 (operazioni particolari, come esportazioni, cessioni intracomunitarie, operazioni esenti, cessioni di rottami, ecc..) e sottraendo l'importo indicato ai righi VE38 (operazioni effettuate in anni precedenti ma con imposta esigibile nel 2006) e VE39 (cessioni beni ammortizzabili).
[TOT_IVA_OPE_IMPO] VE41 COL.1	Totale dell'IVA sulle operazioni imponibili, ottenuto sommando gli importi indicati ai righi VE12 colonna 2 e VE25 colonna 2.
ALIQUOTA MEDIA CESS	Campo calcolato: $ARROTONDA(SE\left(A=0;0;\frac{B}{A}*100\right),2)$ A=VE40; B=VE41
[IMP_TOT_ACQ] VF12 COL.1	Somma degli acquisti all'interno, gli acquisti intracomunitari e le importazioni assoggettati ad imposta, per i quali si è verificata l'esigibilità ed è stato esercitato, nel 2006, il diritto alla detrazione, da riportare in corrispondenza delle aliquote o delle percentuali di compensazione prestampate.
[IMP_TOT_IMP] VF12 COL.2	Imposta detratta relativa al medesimo rigo, colonna 1.
[ACQ_PLAF] VF13	Acquisti all'interno, acquisti intracomunitari e importazioni effettuati senza pagamento dell'imposta, con utilizzo del plafond di cui all'art. 2, comma 2, della legge 18 febbraio 1997, n. 28.
[ACQ_ESENTI] VF15	Acquisti all'interno esenti (art. 10 e art. 6, legge n. 133 del 1999, vedi commento al rigo VE33), acquisti intracomunitari esenti (art. 42, comma 1, D.L. 331/93) e importazioni non soggette all'imposta (art. 68, esclusa la lettera a).
[ACQ_NO_DETR] VF18	Acquisti all'interno, acquisti intracomunitari e importazioni, al netto dell'IVA, per i quali, ai sensi dell'art. 19-bis1, o di altre disposizioni, non è ammessa la detrazione dell'imposta.
[TOT_ACQ] VF21	Somma dei righi da VF12 a VF18 (acquisti non imponibili, esenti, acquisti per i quali non è ammessa detrazione d'imposta, ecc…) meno VF19(Acquisti registrati negli anni precedenti ma con imposta esigibile nel 2006)
[IMP_ACQ_IMPON]	Imposta relativa al rigo VF21

VF23	
[ALIQ MEDIA ACQ]	Campo calcolato: $ARROTONDA(SE\left(A=0;0;\frac{B}{A}*100\right),2)$ A=VF21 B=VF23
[IMP_IVA_DETR] VG71	Il rigo deve essere sempre compilato da parte di tutti i contribuenti per l'indicazione dell'IVA ammessa in detrazione che, nel caso in cui non siano state compilate le prime cinque sezioni del quadro VG né il rigo VG70, corrisponde all'ammontare indicato nel rigo VF22.
[IMPST_DOV] VL7 COL.1	Imposta dovuta (da indicare nella colonna 1), determinata dalla differenza tra il rigo VL3 e il rigo VL6, ovvero imposta a credito (da indicare nella colonna 2), ricavata dalla differenza tra il rigo VL6 e il rigo VL3.
[IMPST_CRED] VL7 COL.2	
[CRED_ANNO_PREC] VL26	Credito risultante dalla dichiarazione relativa all'anno 2006 che non è stato chiesto a rimborso ma riportato in detrazione o in compensazione, risultante dal rigo VX5 ovvero dal corrispondente rigo del quadro RX per i soggetti che hanno presentato il modello unificato.
[TOT_IMPST_DOV] VL38	Totale dell'IVA dovuta che si ricava sottraendo al dato indicato al rigo VL33 i crediti eventualmente utilizzati (VL34 + VL35) e sommando gli interessi trimestrali dovuti (VL36).
[TOT_IMPST_CRED] VL39	Totale dell'IVA a credito risultante dal rigo VL32.

h) *Campi di riepilogo:*

[RICAVI_ATT_2007]	RE2 Col. 2 + RG2 col.2 + RF74
[TOT_ATT_2007]	RE6 + RG10 + RF74
[COSTO_LAV_2006]	RE11 + RE12 + RG14 + RF75 col.1
[TOT_PASS_2006]	RE20 + RG22 + RF75 col.2
[REDD_LORDO_2006]	RE23 + RG26 + RF49 + RD18
[IMP_V_AGG_IMPON]	VE10+VE23-VF12
[IMP_V_AGG_IVA]	VE40-VF21

1.3 Attributi afferenti gli accertamenti

Oltre ai dati relativi alle dichiarazioni presentate, è disponibile una serie di informazioni relative al controllo subito da **1.843** soggetti per l'annualità **2007**. Essi sono:

[TIPO_ACCERTAMENTO]	1=Accertamento unificato 4=Accertamento parziale 41 bis
[STATO_CONTROLLO]	Campo che descrive lo stato attuale del controllo, secondo la legenda che segue: 0 ASSENTE O NON DETERMINATO 1 NON ESITATI 2 NON NOTIFICATI 3 NOTIFICATI 4 NOTIFICATI CON PRESENZA DI VERSAMENTI 5 CON DEFINIZIONE SOLE SANZIONI (ART 17/472) 6 MANCATA IMPUGNAZIONE 7 MANCATA IMPUGNAZIONE CON VERSAMENTI 8 CON ADESIONE IN CORSO DI PERFEZIONAMENTO 9 CON ADESIONE NON PERFEZIONATA 10 IN CONTENZIOSO 11 DOPO RICORSO CON DECISIONE FAVOREVOLE 12 DOPO RICORSO CON DECISIONE PARZIALMENTE FAVOREVOLE ALL'A 13 DOPO RICORSO CON DECISIONE SFAVOREVOLE ALL'AE 14 CON DECISIONE DEFINITIVA IN ATTESA DI ISCRIZIONE A RUOLO 15 CONCILIAZIONE IN CORSO DI PERFEZIONAMENTO 16 DOMANDA DI CONDONO 17 CON ESITO NEGATIVO 18 ANNULLATI PER AUTOTUTELA 19 CON ADESIONE PERFEZIONATA 20 RINUNCIA ALL IMPUGNAZIONE (ART. 15/218) 21 RUOLO IN PRESENZA DI MANCATA IMPUGNAZIONE 22 RUOLO DOPO DECISIONE DEFINITVA FAVOREVOLE 23 CONTENZIOSO CON DECISIONE DEFINITIVA SFAVOREVOLE 24 CONCILIAZIONE GIUDIZIALE PERFEZIONATA 25 VERSAMENTI ART 16/472 PER ATTI DI CONTESTAZIONE 26 ALTRI TIPI DI DEFINIZIONE 27 AUTOTUTELA PARZIALE 28 DEFINIZIONE PARZIALE PER ART 16/472 29 DEFINIZIONE CONDONO(LEGGE N 28 DEL 2002)

	30 INEFFICACE PER DEFINIZIONE LEGGE N 289 DEL 2002 99 ASSENTE 34 DOPO CONTENZIOSO - ESTINZIONE 35 DOMANDA DI CHIUSURA LITI FISCALI PENDENTI (ART 16/289) 36 RUOLO IN PRESENZA DI DEFINIZIONE SANZIONI (ART 17/472) 37 MANCATA IMPUGNAZIONE SENZA ISCRIZIONE A RUOLO 38 RUOLO MANUALE 39 MANCATA DEFINIZIONE ART.5 BIS SOCIO/CONSOLIDANTE 41 INEFFICACE ART.13 BIS DL 78/2009 SCUDO FISCALE PARZIALE 42 INEFFICACE ART.13 BIS DL 78/2009 SCUDO FISCALE TOTALE
[MAG_IMPON_IRF_ACC]	Maggiore imponibile ai fini IRPEF accertato
[MAG_IMPST_IRF_ACC]	CAMPO CALCOLATO Calcolo dell'IRPEF dovuta su RN4+MAGG_IMPON_IRF_ACC al netto di quella dichiarata
[MAG_VOL_AFF_ACC]	Maggiore volume d'affari ai fini IVA accertato
[MAG_IVA_CESS_ACC]	CAMPO CALCOLATO *=ARROTONDA(SE(A>20;B*0,2;B*A/100);0)* A = ALIQ_MEDIA_CESS B = MAG_VOL_AFF_ACC
[MIN_COSTI_ACC]	CAMPO CALCOLATO *=SE(A=B;0;* *SE(E(A>B;B>0);A-B;* *SE(E(A>B;B=0;C=0);0;* *SE(E(A>B;B=0);A;* *SE(A<B;0)))))* A = MAG_IMPON_IRF_ACC B= MAG_VOL_AFF_ACC C= MAG_IMPON_IRAP_ACC
[INDEBITA_DETR_IVA_ACC]	CAMPO CALCOLATO *=ARROTONDA(SE(A>20;B*0,2;B*A/100);0)* A = ALIQ_MEDIA_ACQ B = [MIN_COSTI_ACC]
[MAG_IVA_ACC]	CAMPO CALCOLATO [MAG_IVA_CESS_ACC]+[INDEBITA_DETR_IVA_ACC]
[MAG_IMPON_IRAP_ACC]	Maggiore imponibile ai fini IRAP accertato
MAG_IMPST_IRAP_ACC	CAMPO CALCOLATO *=ARROTONDA(A*0,0425;0)* A=[MAG_IMPON_IRAP_ACC]
[MAG_IMPON_IRF_DEF]	CAMPO CALCOLATO *IF [STATO_CONTROLLO] ="19"* *THEN VAL_DEFINITO;*

	ELSE IF [STATO_CONTROLLO] "20" *THEN VAL_DEFINITO = VAL_ACCERTATO* *ELSE 0* Maggiore imponibile ai fini IRPEF definito (in sede di accertamento con adesione ex D.Lgs. 218/97)
MAG_IMPST_IRF_DEF	CAMPO CALCOLATO Calcolo dell'IRPEF dovuta su RN4+MAGG_IMPON_IRF_DEF al netto di quella dichiarata
[MAG_VOL_AFF_DEF]	CAMPO CALCOLATO *IF [STATO_CONTROLLO] ="19"* *THEN VAL_DEFINITO;* *ELSE IF [STATO_CONTROLLO] "20"* *THEN VAL_DEFINITO = VAL_ACCERTATO* *ELSE 0* Maggiore volume d'affari definito (in sede di accertamento con adesione ex D.Lgs. 218/97)
[MAG_IVA_CESS_DEF]	CAMPO CALCOLATO *=ARROTONDA(SE(A>20;B*0,2;B*A/100);0)* A = ALIQ_MEDIA_CESS B = MAG_VOL_AFF_DEF
[MIN_COSTI_DEF]	CAMPO CALCOLATO *=SE(A=0;0;B-C)* A=MIN_COSTI_ACC B=MAG_IMPON_IRF_DEF C=MAG_VOL_AFF_DEF
[INDEBITA_DETR_IVA_DEF]	CAMPO CALCOLATO *=ARROTONDA(SE(A>20;B*0,2;A*B/100);0)* A = ALIQ_MEDIA_ACQ B = MIN_COSTI_DEF
[MAG_IVA_DEF]	CAMPO CALCOLATO INDEBITA_DETR_IVA_DEF + MAG_IVA_CESS_DEF
[MAG_IMPON_IRAP_DEF]	CAMPO CALCOLATO *IF [STATO_CONTROLLO] ="19"* *THEN VAL_DEFINITO;* *ELSE IF [STATO_CONTROLLO] "20"* *THEN VAL_DEFINITO = VAL_ACCERTATO* *ELSE 0* Maggiore imponibile ai fini IRAP definito (in sede di accertamento con adesione ex D.Lgs. 218/97)
MAG_IMPST_IRAP_DEF	CAMPO CALCOLATO *=ARROTONDA(A*0,0425;0)* A = [MAG_IMPON_IRAP_DEF]

1.4 Attributi afferenti gli studi di settore presentati

Infine, una serie di dati relativi agli studi di settore presentati completano la base dati:

[FLG_STUDIO_SETTORE]	0 = NON PRESENTATO, 1 = PRESENTATO
[COD_STUDIO]	Codice dello studio di settore presentato
[FLG_NO_CONGRUENZA]	0 = CONGRUO, 1 = NON CONGRUO
[FLG_NO_COERENZA]	0 = COERENTE, 1 = NON COERENTE
[FLG_PRES_FAM]	0 = no presenza lavoro familiare, 1 = presenza lavoro familiare Dato desumibile dallo studio
[PRES_FAM_ORD]	Dato desumibile dallo studio
[PRES_FAM_SEMPL]	Dato desumibile dallo studio
[FLG_PRES_DIP]	0 = no presenza lavoratori dipendenti 1 = presenza lavoratori dipendenti Dato desumibile dallo studio
[NUM_DIPENDENTI]	Dato desumibile dallo studio

Appendice B

Le *query* dell'analisi OLAP.

Sono di seguito riportate le query utilizzate nel Capitolo 3. La sintassi utilizzata è quella di SQL Server 2012. I dati sono stati forniti in un'unica tabella e, per semplicità, le query che seguono fanno a riferimento a tale unica tabella contenente, in ogni record, gli attributi relativi ad un singolo soggetto. Unica tabella d'appoggio è quella contenente la gerarchia dei codici attività (tabella ATECO). La base dati con tutti i dati dei contribuenti è denominata DATI07 e il significato degli attributi ivi presenti è riportato in appendice A.

1.1 Elenco delle query

[Query 1]: trovare il numero di soggetti per provincia, in valore assoluto e percentuale

```
SELECT      SIGLA_PROVINCIA, COUNT (*) CONT,
            ROUND(100.0*(COUNT(*)*1.0/(SELECT   COUNT(*)
                                       FROM     DATI07
                                                ))),2) PERC
FROM        DATI07
GROUP BY    SIGLA_PROVINCIA
ORDER BY    SIGLA_PROVINCIA
```

[Query 2]: trovare il numero di soggetti per sesso e provincia

```
SELECT      SESSO, SIGLA_PROVINCIA, COUNT (*)
FROM        DATI07
GROUP BY    SIGLA_PROVINCIA, SESSO
ORDER BY    SIGLA_PROVINCIA
```

[Query 3] trovare il numero di soggetti per età

```
SELECT      ETA, COUNT(*)
FROM        DATI07
GROUP BY    ETA
ORDER BY    ETA
```

[Query 3bis] trovare il numero di soggetti per età e per durata (in anni) dell'attività

```
SELECT      ETA, DURATA_ATTIV_SOGG, COUNT(*)
FROM        DATI07
GROUP BY    ETA, DURATA_ATTIV_SOGG
ORDER BY    ETA
```

[Query 4] trovare il numero di soggetti per sezione

```
SELECT      SEZIONE, COUNT(*)
FROM        ATECO, DATI07
WHERE       DATI07.COD_ATTIVITA_2007=ATECO.CODICE
GROUP BY    SEZIONE
```

[Query 5] per trovare i soggetti che hanno presentato i singoli quadri delle dichiarazioni o combinazioni di essi

```
SELECT      CODEC_CF                                    //per il quadro RD
FROM        DATI07
WHERE       (AGR_REDD<>0
            OR PR_AGR_CORR<>0
            OR PR_AGR_ACQ<>0)
INTERSECT
SELECT      CODEC_CF                                    //per il quadro RE
FROM        DATI07
WHERE       NOT (PROF_COMP<>0
            OR PROF_COSTI<>0
            OR IMP_CMPNS_ATTIV_2007<>0
            OR IMP_CMPNS_SZ1_2007<>0
            OR IMP_LAV_DIPEND_2007<>0
            OR IMP_CMPNS_TERZI_2007<>0
            OR ALTRE_SPS_DOC<>0
            OR IMP_TOTALE_SPESE_2007<>0
            OR IMP_REDD_AUT_SZ1_2007<>0
            OR IMP_REDD_LAV_AUT_2007<>0)
INTERSECT
SELECT      CODEC_CF                                    //per il quadro RF
FROM        DATI07
WHERE       NOT (UTILE_PERD_CIV_ORD<>0
            OR IMP_REDD_LRD_ORD<>0
            OR IMP_REDD_IMP_ORD<>0
            OR RIM_FIN_ORD<>0
            OR CRED_CLI_ORD<>0
            OR DEB_FORN_ORD<>0
            OR IMP_RICAVI_ORDIN_2007<>0
            OR IMP_SPS_DIPEND_ORD_2007<>0
            OR SPESE_ORD<>0)
INTERSECT
SELECT      CODEC_CF                                    //per il quadro RG
FROM        DATI07
WHERE       NOT (IMP_RICAVI_SMPL_2007<>0
```

```
OR RIM_FIN_SMPL<>0
OR RIM_INI_SMPL<>0
OR IMP_TOT_CMPN_PSV<>0
OR COSTI_ACQ_MP<>0
OR IMP_SPS_DIPEND_SMP<>0
OR COSTI_RSDL<>0
OR IMP_TOT_COMP_NEG_2007<>0
OR IMP_REDD_LRD_SMPL_2007<>0
OR IMP_REDD_IMP_SMPL_2007<>0)
```

[Query 6a] per trovare la media di grandezze reddituali e IVA di chi presenta solo quadro RD

```
SELECT      ROUND (AVG(AGR_REDD),0) AS REDDITO_MEDIO,
            ROUND(AVG(PR_AGR_CORR),0) AS RICAVI,
            ROUND(AVG(PR_AGR_ACQ]),0) AS ACQUISTI,
            ROUND(AVG(IMP_V_AGG_IVA),0) AS VALAGG,
            ROUND(AVG(IMP_VE_VOLAFF_2007),0) AS VOLAFF,
            ROUND(AVG(TOT_ACQ),0) AS ACQUISTIIVA,
            ROUND(AVG(ALIQ_MEDIA_ACQ),0) AS ALIQ_MEDIA_ACQ,
            ROUND(AVG(ALIQ_MEDIA_CESS),0) AS ALIQ_MEDIA_CESS
FROM        DATI07
WHERE       CODEC_CF IN
                  (SELECT     CODEC_CF
                  FROM        DATI07
                  WHERE       (AGR_REDD<>0
                              OR PR_AGR_CORR<>0
                              OR PR_AGR_ACQ<>0)
                  INTERSECT
                  SELECT      CODEC_CF
                  FROM        DATI07
                  WHERE       NOT (PROF_COMP<>0
                              OR PROF_COSTI<>0
                              OR IMP_CMPNS_ATTIV_2007<>0
                              OR IMP_CMPNS_SZ1_2007<>0
                              OR IMP_LAV_DIPEND_2007<>0
                              OR IMP_CMPNS_TERZI_2007<>0
                              OR ALTRE_SPS_DOC<>0
                              OR IMP_TOT_SPESE_2007<>0
                              OR IMP_RED_AUT_SZ1_2007<>0
                              OR IMP_RED_LAV_AUT_2007<>0)
                  INTERSECT
                  SELECT      CODEC_CF
                  FROM        DATI07
                  WHERE       NOT (UTILE_PERD_CIV_ORD<>0
                              OR IMP_REDD_LRD_ORD<>0
                              OR IMP_REDD_IMP_ORD<>0
                              OR RIM_FIN_ORD<>0
                              OR CRED_CLI_ORD<>0
                              OR DEB_FORN_ORD<>0
                              OR IMP_RICAVI_ORDIN_2007<>0
                              OR IMP_SPS_DIP_ORD_2007<>0
                              OR SPESE_ORD<>0)
                  INTERSECT
                  SELECT      CODEC_CF
                  FROM        DATI07
                  WHERE       NOT (IMP_RICAVI_SMPL_2007<>0
                              OR RIM_FIN_SMPL<>0
                              OR RIM_INI_SMPL<>0
                              OR IMP_TOT_CMPN_PSV<>0
                              OR COSTI_ACQ_MP<>0
                              OR IMP_SPS_DIPEND_SMP<>0
                              OR COSTI_RSDL<>0
                              OR IMP_TOT_COMP_NEG_2007<>0
                              OR IMP_REDD_LRD_SMPL_2007<>0
                              OR IMP_REDD_IMP_SMPL_2007<>0))
```

[Query 6b] per trovare la media di grandezze reddituali e IVA di chi presenta solo RE. Cambiano solo primi 3 elementi della SELECT della Query 6a e le condizioni della WHERE:

```
SELECT      ROUND (AVG(IMP_REDD_LAV_AUT_2007),0) AS REDDITO_MEDIO,
            ROUND(AVG(IMP_CMPNS_ATTIV_2007),0) AS RICAVI,
            ROUND(AVG(IMP_TOTALE_SPESE_2007),0) AS ACQUISTI,
```

[Query 6c] per trovare la media di grandezze reddituali e IVA di chi presenta solo quadro RF. Cambiano solo primi 3 elementi della SELECT della Query 6a e le condizioni della WHERE:

```
SELECT      ROUND (AVG(IMP_REDD_IMP_ORD),0) AS REDDITO_MEDIO,
            ROUND(AVG(IMP_RICAVI_ORDIN_2007),0) AS RICAVI,
            ROUND(AVG(SPESE_ORD),0) AS ACQUISTI,
```

[Query 6d] per trovare la media di grandezze reddituali e IVA di chi presenta solo quadro RG. Cambiano solo primi 3 elementi della SELECT della Query 6a e le condizioni della WHERE:

```
SELECT      ROUND (AVG(IMP_REDD_LRD_SMPL_2007),0) AS REDDITO_MEDIO,
            ROUND(AVG(IMP_RICAVI_SMPL_2007),0) AS RICAVI,
            ROUND(AVG(IMP_TOT_COMP_NEG_2007),0) AS ACQUISTI,
```

[Query 07] per trovare il reddito medio dichiarato per sesso

```
SELECT      '2007', SESSO, AVG (REDD_IMP_2007)
FROM        DATI07
GROUP BY    SESSO
```

[Query 8] per trovare il reddito medio per età

```
SELECT      ETA, ROUND(AVG (REDD_IMP_2007),0)
FROM        DATI07
GROUP BY    ETA
ORDER BY    ETA
```

[Query 9] per trovare il reddoto medio per durata attività

```
SELECT      DURATA_ATTIV_SOGG, ROUND(AVG (REDD_IMP_2007),0)
FROM        DATI07
GROUP BY    DURATA_ATTIV_SOGG
ORDER BY    DURATA_ATTIV_SOGG
```

[Query 10] per trovare media maggior IRPEF accertata e media reddito dichiarato per tipologia di contribuente (modificando di volta in volta le condizioni di WHERE)

```
SELECT      ROUND(AVG(MAG_IMPON_IRF_ACC),0) AS MAGG_IRF_ACC,
            ROUND(AVG(REDD_IMP_2007),0) IRPEF_DIC
FROM        DATI07
WHERE       CODEC_CF IN
                  (SELECT       CODEC_CF
                  FROM          DATI07
                  WHERE (AGR_REDD<>0
                  OR PR_AGR_CORR<>0
                  OR PR_AGR_ACQ<>0)
                  INTERSECT
                  SELECT        CODEC_CF
                  FROM          DATI07
                  WHERE         NOT (PROF_COMP<>0
                                OR PROF_COSTI<>0
                                OR IMP_CMPNS_ATTIV_2007<>0
                                OR IMP_CMPNS_SZ1_2007<>0
                                OR IMP_LAV_DIPEND_2007<>0
                                OR IMP_CMPNS_TERZI_2007<>0
                                OR ALTRE_SPS_DOC<>0
                                OR IMP_TOTALE_SPESE_2007<>0
                                OR IMP_REDD_AUT_SZ1_2007<>0
                                OR IMP_REDD_LAV_AUT_2007<>0)
                  INTERSECT
                  SELECT        CODEC_CF
                  FROM          DATI07
                  WHERE         NOT (UTILE_PERD_CIV_ORD<>0
                                OR IMP_REDD_LRD_ORD<>0
                                OR IMP_REDD_IMP_ORD<>0
                                OR RIM_FIN_ORD<>0
                                OR CRED_CLI_ORD<>0
                                OR DEB_FORN_ORD<>0
                                OR IMP_RICAVI_ORDIN_2007<>0
                                OR IMP_SPS_DIPEND_ORD_2007<>0
                                OR SPESE_ORD<>0)
                  INTERSECT
                  SELECT        CODEC_CF
                  FROM          DATI07
                  WHERE         NOT (IMP_RICAVI_SMPL_2007<>0
                                OR RIM_FIN_SMPL<>0
                                OR RIM_INI_SMPL<>0
                                OR IMP_TOT_CMPN_PSV<>0
                                OR COSTI_ACQ_MP<>0
                                OR IMP_SPS_DIPEND_SMP<>0
                                OR COSTI_RSDL<>0
                                OR IMP_TOT_COMP_NEG_2007<>0
                                OR IMP_REDD_LRD_SMPL_2007<>0
                                OR IMP_REDD_IMP_SMPL_2007<>0))
```

[Query 11] per trovare media maggior IRPEF accertata e media reddito dichiarato per provincia

```
SELECT      SIGLA_PROVINCIA,
            ROUND (AVG(REDD_IMP_2007),0) AS REDDITO_MEDIO,
            ROUND(AVG(MAG_IMPON_IRF_ACC),0) MAGG_IRPEF_ACC
FROM        DATI07
GROUP BY    SIGLA_PROVINCIA
```

[Query 12] per trovare top 10 attività ordinate per maggior imponibile medio accertato

```
SELECT TOP 10   A.SEZIONE, COD_ATTIVITA_2007, A.DESCRIZIONE,
                COUNT(*) CONTA,
                ROUND(AVG(REDD_LORDO_2007),0) IRPEF_DIC,
                ROUND(AVG(MAG_IMPON_IRF_ACC),0) MAG_IRPEF_ACC
FROM            DATI07 D, ATECO A
WHERE           D.COD_ATTIVITA_2007 = A.CODICE
GROUP BY        COD_ATTIVITA_2007, A.DESCRIZIONE, A.SEZIONE
HAVING          COUNT(*)>20
ORDER BY        AVG(MAG_IMPON_IRF_ACC) DESC
```

[Query 13] per trovare valori medi degli accertamenti per tipologia di contribuente (modificando di volta in volta le clausole WHERE)

```
SELECT          ROUND(AVG(MAG_IMPON_IRF_ACC),0) MAG_IRPEF_ACC,
                ROUND(AVG(MAG_VOL_AFF_ACC),0) MAG_VOLAFF,
                ROUND(AVG(MIN_COSTI_ACC),0) MIN_COSTI,
                ROUND(AVG(MAG_IMPON_IRAP_ACC),0) MAG_IRAP
FROM            DATI07
WHERE           CODEC_CF IN
                        (SELECT         CODEC_CF
                        FROM            DATI07
                        WHERE           NOT(AGR_REDD<>0
                                        OR PR_AGR_CORR<>0
                                        OR PR_AGR_ACQ<>0)
                        INTERSECT
                        SELECT          CODEC_CF
                        FROM            DATI07
                        WHERE           NOT(PROF_COMP<>0
                                        OR PROF_COSTI<>0
                                        OR IMP_CMPNS_ATTIV_2007<>0
                                        OR IMP_CMPNS_SZ1_2007<>0
                                        OR IMP_LAV_DIPEND_2007<>0
                                        OR IMP_CMPNS_TERZI_2007<>0
                                        OR ALTRE_SPS_DOC<>0
                                        OR IMP_TOTALE_SPESE_2007<>0
                                        OR IMP_REDD_AUT_SZ1_2007<>0
                                        OR IMP_REDD_LAV_AUT_2007<>0)
                        INTERSECT
                        SELECT          CODEC_CF
                        FROM DATI07
                        WHERE           NOT (UTILE_PERD_CIV_ORD<>0
                                        OR IMP_REDD_LRD_ORD<>0
                                        OR IMP_REDD_IMP_ORD<>0
                                        OR RIM_FIN_ORD<>0
                                        OR CRED_CLI_ORD<>0
                                        OR DEB_FORN_ORD<>0
                                        OR IMP_RICAVI_ORDIN_2007<>0
                                        OR IMP_SPS_DIPEND_ORD_2007<>0
                                        OR SPESE_ORD<>0)
                        INTERSECT
                        SELECT          CODEC_CF
                        FROM            DATI07
                        WHERE           IMP_RICAVI_SMPL_2007<>0
                                        OR RIM_FIN_SMPL<>0
                                        OR RIM_INI_SMPL<>0
                                        OR IMP_TOT_CMPN_PSV<>0
                                        OR COSTI_ACQ_MP<>0
                                        OR IMP_SPS_DIPEND_SMP<>0
                                        OR COSTI_RSDL<>0
                                        OR IMP_TOT_COMP_NEG_2007<>0
                                        OR IMP_REDD_LRD_SMPL_2007<>0
                                        OR IMP_REDD_IMP_SMPL_2007<>0))
```

[Query 14] per trovare i valori medi accertati e definiti per tipologia di contribuente (modificando di volta in volta le clausole WHERE)

```
SELECT      ROUND(AVG(MAG_IMPON_IRF_ACC),0) MAG_IRPEF_ACC,
            ROUND(AVG(MAG_IMPON_IRF_DEF),0) MAG_IRPEF_DEF,
            ROUND(AVG(MAG_VOL_AFF_ACC+MIN_COSTI_ACC),0) M_IVA_AC,
            ROUND(AVG(MAG_VOL_AFF_DEF+MIN_COSTI_DEF),0) M_IVA_DEF,
            ROUND(AVG(MAG_IMPON_IRAP_ACC),0) MAG_IRAP_ACC,
            ROUND(AVG(MAG_IMPON_IRAP_DEF),0) MAG_IRAP_DEF
FROM        DATI07
WHERE       STATO_CONTROLLO=19 AND
            CODEC_CF IN
                  (SELECT     CODEC_CF
                  FROM        DATI07
                  WHERE       (AGR_REDD<>0
                              OR PR_AGR_CORR<>0
                              OR PR_AGR_ACQ<>0)
                  INTERSECT
                  SELECT      CODEC_CF
                  FROM        DATI07
                  WHERE       NOT (PROF_COMP<>0
                              OR PROF_COSTI<>0
                              OR IMP_CMPNS_ATTIV_2007<>0
                              OR IMP_CMPNS_SZ1_2007<>0
                              OR IMP_LAV_DIPEND_2007<>0
                              OR IMP_CMPNS_TERZI_2007<>0
                              OR ALTRE_SPS_DOC<>0
                              OR IMP_TOTALE_SPESE_2007<>0
                              OR IMP_REDD_AUT_SZ1_2007<>0
                              OR IMP_REDD_LAV_AUT_2007<>0)
                  INTERSECT
                  SELECT      CODEC_CF
                  FROM        DATI07
                  WHERE       NOT (UTILE_PERD_CIV_ORD<>0
                              OR IMP_REDD_LRD_ORD<>0
                              OR IMP_REDD_IMP_ORD<>0
                              OR RIM_FIN_ORD<>0
                              OR CRED_CLI_ORD<>0
                              OR DEB_FORN_ORD<>0
                              OR IMP_RICAVI_ORDIN_2007<>0
                              OR IMP_SPS_DIPEND_ORD_2007<>0
                              OR SPESE_ORD<>0)
                  INTERSECT
                  SELECT      CODEC_CF
                  FROM        DATI07
                  WHERE       NOT (IMP_RICAVI_SMPL_2007<>0
                              OR RIM_FIN_SMPL<>0
                              OR RIM_INI_SMPL<>0
                              OR IMP_TOT_CMPN_PSV<>0
                              OR COSTI_ACQ_MP<>0
                              OR IMP_SPS_DIPEND_SMP<>0
                              OR COSTI_RSDL<>0
                              OR IMP_TOT_COMP_NEG_2007<>0
                              OR IMP_REDD_LRD_SMPL_2007<>0
                              OR IMP_REDD_IMP_SMPL_2007<>0))
```