Ministero dell'Economia e delle Finanze

Documenti di lavoro dell'Ufficio Studi

2007/4

DATA MINING E ANALISI DEL RISCHIO DI FRODE FISCALE: IL CASO DEI CREDITI D'IMPOSTA

Paola De Sisti e Stefano Pisani

Documenti di lavoro dell'Ufficio Studi

Febbraio 2007

DATA MINING E ANALISI DEL RISCHIO DI FRODE FISCALE: IL CASO DEI CREDITI DI IMPOSTA

Paola De Sisti (*) e Stefano Pisani

(*) So.Ge.I.

Data mining e analisi del rischio di frode fiscale: il caso dei crediti di imposta

Le tecniche di data mining sono finalizzate all'individuazione delle frodi nella richiesta dei crediti d'imposta per "investimenti in aree svantaggiate". La scelta dello strumento è stata suggerita dai brillanti risultati ottenuti in ambiti analoghi (ad es. le frodi con carte di credito). Il modello associa a ciascun soggetto un punteggio che misura la probabilità che la richiesta sottintenda una frode. La sperimentazione condotta sul campo ha fornito risultati molto soddisfacenti (percentuale di successo dell'82%). Sarebbe, pertanto, indicato generalizzarne l'utilizzo, applicandolo, ad esempio,ai crediti IVA o alle compensazioni d'imposta (F24).

Indice

Sintesi direzionale[1]

L'amministrazione fiscale gestisce correntemente un enorme patrimonio informativo che, attraverso un importante opera di informatizzazione è stato trasferito su supporto informatico, creando una fonte informazione fruibile (tramite Data Mart o Data Warehouse) per le finalità operative dell'Agenzia delle Entrate.

Tale processo di informatizzazione rappresenta una pre condizione indispensabile, ma non è sufficiente per sfruttare a pieno le potenzialità che la vasta mole di informazioni disponibile renderebbe possibile. A tal fine è necessario sviluppare tecniche di indagine avanzate, che consentano di evidenziare quei segnali che possono essere di maggiore interesse nello svolgere le attività istituzionali a cui è preposta e, in particolare, per effettuare una corretta attività di prevenzione e contrasto dell'evasione.

Tali tecniche si sostanziano in modelli statistici che sono usualmente classificati con la dizione di data mining. Questi modelli sono finalizzati ad estrarre dai dati grezzi elementi di informazione utili, permettendo di acquisire una conoscenza di aspetti non evidenti nei dati stessi, con il fine di individuare comportamenti che si discostano dalla media del profilo di appartenenza.

Il lavoro illustra un'applicazione di tali strumenti nel campo delle prevenzione e contrasto dell'evasione, finalizzata all'individuazione delle frodi nella richiesta dei crediti d'imposta per "investimenti in aree svantaggiate"[2]. L'applicazione ha integrato più banche dati dell'anagrafe tributaria, consentendo di determinare il profilo del possibile frodatore e, quindi, di selezionare all'interno dell'universo dei contribuenti quelli a più alto rischio di commettere la frode.

Il ricorso al data mining nell'ambito dell'attività di prevenzione e contrasto all'evasione è suggerito anche in virtù dei brillanti risultati ottenuti in altri ambiti per prevenire fenomeni fraudolenti (come ad esempio, nel caso del mercato delle carte di credito). La procedura che si presenta si propone di costruire un modello orientato alla selezione dei controlli fiscali a posteriori. Basandosi sulle caratteristiche inerenti il soggetto che richiede il credito (desunte dalle banche dati dell'amministrazione finanziaria) il modello è in grado di associare a ciascun soggetto una probabilità, che misura la possibilità che la richiesta di credito ha di sottintendere una frode. La principale caratterista del procedimento proposto risiede nel fatto di poter considerare simultaneamente una vasta mole di informazioni (analisi multicriterio), e di poter offrire strumenti operativi che producono risposte in tempo reale.

La sperimentazione condotta sui crediti di imposta ha fornito risultati estremamente soddisfacenti, in quanto:

1. ha consentito di derivare degli indicatori sulla probabilità di essere frodatori scaturiti dall'osservazione delle informazioni disponibili per ciascun contribuente;

[1] La procedura esposta è il risultato di un gruppo di lavoro costituito presso la Direzione Accertamento dell'Agenzia delle Entrare costituito da: Vincenzo Errico, Fabio Celozzi, Pier Paolo Verna, Valerio D'Aiello, Fabiano Della Casa e da Stefano Pisani per l'Agenzia delle Entrate; da Barbara Pergameno e Paola De Sisti per la SOGEI.

[2] I crediti d'imposta per le aree svantaggiate sono stati introdotti dall'art. 10 del Decreto Legge 8 luglio 2002, n. 138, convertito dalla Legge 8 agosto 2002, n. 178 (successivamente modificato tramite **l'**art. 62 della legge 27 dicembre 2002, n.289).

2. la probabilità individuata al punto 1 è stata derivata da regole che sono state analizzate e possono fornire ulteriori spunti di approfondimento futuro o confortare ipotesi qualitative formulate precedentemente;
3. l'algoritmo adotta una logica multi criterio che ben si adatta al perseguimento di attività di deterrenza all'evasione (cioè coniuga la trasmissione di segnali forti con il perseguimento di più obiettivi simultaneamente);
4. è stata effettuata una ulteriore verifica, analizzando le verifiche svolte dagli uffici dopo l'elaborazione del modello, ed è risultata confermata la bontà dello strumento dato che ogni cento controlli effettuati ha individuato 82 frodatori sostanziali.

Di contro, la procedura richiede un sforzo di investimento iniziale molto oneroso, che però è adeguatamente remunerato elevando significativamente il tasso di positività dei controlli. Inoltre, dato che gran parte dello sforzo si concentra principalmente nella predisposizione e nell'analisi delle informazioni di base, una volta predisposto il modello le successive implementazioni consentono di realizzare significative economie di scala.

Alla luce dei risultati conseguiti e della flessibilità dello strumento, sarebbe auspicabile applicarlo ad altri campi di osservazione, che presentano un'elevata numerosità di soggetti da controllare ed una vasta mole di variabili da considerare congiuntamente, come, ad esempio: i crediti IVA o le compensazioni tramite F24.

1. Introduzione[3]

L'amministrazione fiscale gestisce correntemente un enorme patrimonio informativo, in grado di fornire una rappresentazione censuaria dei principali comportamenti economici dichiarati dai contribuenti. Attraverso un importante opera di informatizzazione, nel corso degli anni, tale patrimonio informativo è stato trasferito su supporto informatico, creando una fonte di informazioni fruibile (tramite Data Mart o Data Warehouse) per le finalità operative dell'Agenzia delle Entrate.

Tale processo di informatizzazione rappresenta una pre condizione indispensabile, ma non è sufficiente per sfruttare a pieno le potenzialità che la vasta mole di informazioni disponibile renderebbe possibile. A tal fine è necessario sviluppare tecniche di indagine avanzate, che consentano di evidenziare quei segnali che possono essere di maggiore interesse per l'amministrazione, nello svolgere le attività istituzionali a cui è preposta e, in particolare, per effettuare una corretta attività di prevenzione e contrasto dell'evasione[4].

Una possibile linea di sviluppo consiste nel promuovere un processo di ricerca di nuova conoscenza a partire dai dati già noti, ricorrendo a tecniche di data mining. Tali tecniche sono finalizzate ad estrarre dai dati grezzi elementi di informazione utili, permettendo di acquisire una conoscenza di aspetti non evidenti nei dati stessi, con il fine di individuare compartimenti che si discostano dalla media del profilo di appartenenza.

Per conferire un carattere di maggiore operatività alla trattazione, nelle pagine che seguono, si illustrerà un'applicazione nel campo delle prevenzione e contrasto dell'evasione, finalizzata all'individuazione delle frodi nella richiesta dei crediti d'imposta per "investimenti in aree svantaggiate"[5]. L'applicazione ha sottoposto più banche dati dell'anagrafe tributaria alle procedure di data mining, consentendo prima di tutto un'analisi esplorativa della popolazione sotto osservazione (tramite segmentazione o clustering), che ha fornito una prima chiave di lettura dei dati. Quindi, si è passati alla analisi delle correlazioni tra la presenza di frode e il comportamento dei soggetti, basate sulla costruzione di modelli statistici per la determinazione di anomalie di comportamento. Tale modello è stato poi proposto per fini previsivi, al fine di individuare i possibili frodatori.

Il ricorso al data mining nell'ambito dell'attività di prevenzione e contrasto all'evasione è suggerito anche in virtù dei brillanti risultati ottenuti in altri ambiti per prevenire fenomeni

---

[3] La procedura esposta è il risultato di un gruppo di lavoro costituito presso la Direzione Accertamento dell'Agenzia delle Entrare costituito da: Vincenzo Errico, Fabio Celozzi, Pier Paolo Verna, Valerio D'Aiello, Fabiano Della Casa e da Stefano Pisani per l'Agenzia delle Entrate; da Barbara Pergameno e Paola De Sisti per la SOGEI.

[4] Emblematica a tale proposito è una dichiarazione rilasciata dal prof. A. Fantozzi al Sole 34 ore (22 agosto 2006, p. 2), nella quale si afferma che "L'informatizzazione, e dunque le banche dati, è la frontiera su cui si combatte la battaglia del fisco. Però troppi dati equivalgono a pochi dati. L'eccessiva quantità di dati, anche poco rilevanti, può ingenerare confusioni o favorire usi distorti". Ed è proprio al fine di chiarire le "confusioni" e scongiurare la possibilità di "usi distorti", che si propone l'utilizzo di tecniche innovative per trattare vaste basi di dati.

[5] I crediti d'imposta per le aree svantaggiate sono stati introdotti dall'art. 10 del Decreto Legge 8 luglio 2002, n. 138, convertito dalla Legge 8 agosto 2002, n. 178 (successivamente modificato tramite **l'**art. 62 della legge 27 dicembre 2002, n.289). La norma stabilisce una procedura di ammissione al contributo tramite un'istanza preventiva da inoltrare al Centro operativo di Pescara, ha previsto, tra l'altro, per i soggetti che intendono effettuare investimenti nelle arre svantaggiate, l'obbligo di pianificare gli investimenti e gli utilizzi del credito.Il contributo comunque può essere fruito, in relazione all'investimento realizzato, successivamente all'atto di assenso espressamente adottato dall'Agenzia delle Entrate, esclusivamente in compensazione ai sensi del Decreto Legislativo 241/1997.

fraudolenti. Ad esempio, nel caso del mercato delle carte di credito[6] si identificano quali transazioni di acquisto possono essere state effettuate con carte rubate o falsificate e si interviene bloccandole; oppure nel caso dell'assegnazione di Credit Risk/Credit Scoring da parte delle banche si analizzano le richieste di credito assegnando un punteggio ai clienti che richiedono un fido o un prestito e si decide, in base alla classe di rischio cui appartengono, se concederli o meno, e quali strategie adottare.

La procedura che si presenta è assimilabile a quella seguita negli esempi appena citati: tuttavia si propone di costruire un modello non di tipo preventivo (come nel caso delle carte di credito), ma orientato ai controlli fiscali a posteriori.

Basandosi sulle caratteristiche inerenti il soggetto che richiede il credito (desunte dalle informazioni economiche, strutturali e demografiche desumibili dalle banche dati dell'amministrazione finanziaria) il modello è in grado di associare a ciascun soggetto una probabilità che misura la possibilità che la richiesta di credito ha di sottintendere una frode.

La principale caratterista del procedimento proposto risiede nel fatto di poter considerare simultaneamente una vasta mole di informazioni, e di poter offrire in prospettiva strumenti operativi che producono risposte in tempo reale.

L'agenzia delle Entrate si è già dotata di tecniche che consentono di effettuare un'analisi del rischio, indirizzata a prevenire sul nascere le attività illecite. Tale analisi consistono nell'esaminare le caratteristiche di tutti i soggetti che presentano la dichiarazione di inizio attività, per rilevarne eventuali indizi di pericolosità. L'identificazione di tale pericolosità avviene tramite una *check list* (composta di 16 voci), attribuendo un punteggio proporzionato alla gravità dei diversi indicatori di rischi contemplati e restituendo una graduatoria dei più "meritevoli" di attenzione.

L'impostazione di fondo dello strumento che si descrive nel presente lavoro è sostanzialmente analoga al metodo basato sulla check list, predeterminato in base a considerazioni logiche. L'utilizzo del modello fa si che l'assegnazione dei pesi per determinare la pericolosità relativa di ciascun fattore individuato è tratta da un'osservazione sistematica (e documentata) della realtà passata. L'impiego di un modello, inoltre, potrebbe porre le basi per la creazione di un sistema "esperto", cioè capace di aggiornarsi e di apprendere dalle nuove informazioni che via via si rendono disponibili. La descrizione dell'applicazione proposta è prevalentemente finalizzata all'illustrazione del carattere generale dello strumento, che risulta essere molto flessibile nelle applicazioni e, quindi, che potrebbe trovare ulteriori e proficui campi di applicazione in futuro.

Il lavoro è cosi strutturato: nel secondo paragrafo si illustra che cos'è il data mining nelle sue linee generali, la sezione seguente è dedicata alla descrizione del caso di studio, mentre i paragrafi 4, 5, 6 e 7 esaminano in sequenza le fasi salienti della costruzione del modello, a seguire si discutono i principali risultati conseguiti e le possibilità di utilizzo dello strumento, ed infine sono tratte alcune conclusioni.

[6] Dal Sole 24 del 2 agosto 2005, si legge che una nota marca ha approntato una soluzione che "permette un monitoraggio delle transazioni bancarie dei comportamenti sospetti, analizza dati per identificare nuovi modelli e ridefinire gli *engine* di allerta e gestisce i casi potenziali di frodi dalla scoperta alla notifica alle autorità. La soluzione di data mining … è in grado di assegnare una probabilità precisa ad ogni operazione effettuata con la carta di credito. Sulla base della storicità dei dati relativi, vengono prodotti degli indicatori che consentono di individuare gli eventi fraudolenti, consentendo ai gestori di bloccare le carte in tempo". Tratto dall'articolo "SAS e Hsbs alleate contro le frodi".

2. Che cos'è il data mining

Con il termine di data mining si intende il processo di ricerca e di estrazione di conoscenza da banche dati di grandi dimensioni, tramite l'applicazione di algoritmi che individuano le associazioni "nascoste" tra le informazioni e le rendono visibili[7].

Tali procedure hanno acquisito grande rilevanza con lo svilupparsi di grandi basi di dati, che sono difficilmente gestibili con tecniche descrittive di tipo tradizione e che, quindi, anziché arricchire il bagaglio di conoscenza, nascondono i segnali che più sono necessari all'utilizzatore.

Il data mining utilizza tecniche statistiche e di tipo ingegneristico, finalizzate all'individuazione delle informazioni più significative, rendendole disponibili e direttamente utilizzabili nell'ambito di un processo decisionale. In questo modo i dati si trasformano in "conoscenza" (informazioni significative), per il tramite dell'individuazione delle associazioni ("patterns"), o sequenze ripetute, o regolarità, nascoste nei dati. In questo contesto un "pattern" indica una struttura, un modello, o, in generale, una rappresentazione sintetica dei dati.

Il data mining rappresenta una evoluzione nei processi di analisi dei dati, che hanno iniziato a svilupparsi in modo organico negli anni 60. Si utilizzavano sistemi che producevano report standardizzati, che contenevano semplici informazioni riassuntive o predefinite.

Successivamente, negli anni 80, fu introdotta la possibilità di eseguire interrogazioni differenziate sui database, rendendo più facile l'identificazione di andamenti relativi, per esempio, a un certo prodotto o a una certa area geografica.

All'inizio degli anni 90 lo sviluppo del software di analisi ha puntato alla possibilità di "scavare" nei propri dati in tempo reale. Per esempio, guardando una tabella delle vendite ripartite per zona e prodotto, l'utente potrebbe selezionare una zona per vedere l'andamento a livello di singola regione o provincia.

Gli strumenti attuali sono finalizzati ad implementare la possibilità di passare al setaccio i dati per scoprire relazioni significative. Il processo cosi esemplificato può essere schematizzato tramite la figura 2.1.

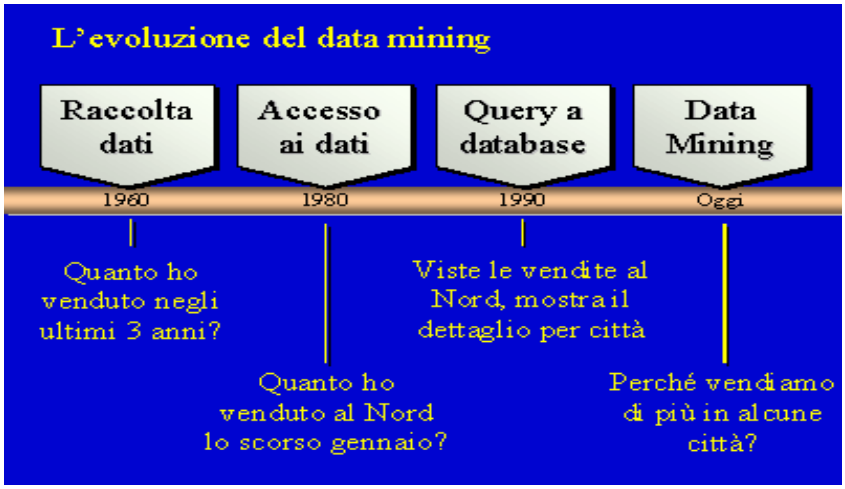


Figura 2.1 – Schematizzazione del processo di evoluzione del Data Mining[8]

[7] Per ulteriori approfondimenti sul data mining si rimanda a: Fayyad, e altri 1996; Hanand e. Kamber. 2000; Hand e altri 2001; Vapnik 1998; Hastie 2001; Piatetsky-Shapiroand e Frawley 1991.

[8] L'esempio è tratto da E-Business News (2001) pp. 66, 67, 68 e 69.

All'interno della schematizzazione proposta nella figura 2.1, si può collocare la posizione dell'Agenzia delle Entrate che, in virtù dell'elevato processo di informatizzazione condotto negli ultimi anni, ha prodotto un efficace sistema di "Query a database" e quindi, adesso è pronta a sviluppare tecniche che consentano di fornire risposte di carattere comportamentale.

Tra queste tecniche un ruolo importante lo gioca il data mining, che, nel caso specifico, può consentire di affrontare la domanda "Quando un contribuente pone in essere comportamenti fraudolenti con riferimento ai crediti di imposta". Conseguentemente, la risposta a tale domanda conduce anche a determinare la probabilità che un soggetto può avere nel porre in essere una frode (ovviamente con riferimento a specifiche azioni intraprese).

I principali punti di forza del data mining possono essere sintetizzati nel modo seguente:

- velocità degli algoritmi su grandi quantità di dati;
- risoluzione di problemi complessi;
- gestione di dati di diversa natura;
- semplicità e immediatezza della rappresentazione dei risultati;
- facilità di applicazione dei modelli a casi non osservati;

I vantaggi derivanti da queste applicazione si potranno apprezzare più compiutamente nel corso dell'esposizione del caso di studio.

3. Il caso di studio

L'applicazione delle tecniche previsive, proposte a titolo dimostrativo delle potenzialità del data mining, ha lo scopo di individuare le frodi fiscali relative ai "crediti di imposta per investimenti in aree svantaggiate". L'idea dello studio nasce dall'analisi degli "accessi mirati[9]" effettuati su tali fattispecie. All'epoca in cui è stato iniziato lo studio (fine 2004), l'Amministrazione aveva condotto circa 54.517 accessi mirati, di questi 21.511 (il 39,5% del totale dei controllati) hanno evidenziato che hanno utilizzato un importo a credito in eccedenza rispetto all'importo spettante.

L'obiettivo che si è posto il lavoro è consistito nell'approntare una metodologia che portasse ad individuare un numero di reali frodatori significativamente superiore a quello che si individuerebbe utilizzando la procedura di selezione standard (il 39,46%).

La metodologia si fonda sull'analisi delle correlazioni esistenti tra le caratteristiche economico-comportamentali dei contribuenti (desunti dagli archivi fiscali) e il realizzarsi dell'evento "credito di imposta utilizzato ma non spettante". Tale enunciazione mette il luce il fatto che, per realizzare la procedura sono necessari due insieme di dati:

1) il Data Base dei controlli (verifiche e accessi mirati) effettuati sui Crediti di Imposta;
2) l'insieme dei Data Base dai quali trarre le caratteristiche economico-comportamentali.

Il primo data base ci permette di individuare i soggetti che, avendo ricevuto un controllo fiscale, hanno frodato con certezza il fisco in relazione alla richiesta ed utilizzo dei crediti di

[9] Per "accesso mirato" si intende una istruttoria esterna, cioè svolta presso la sede del contribuente, per acquisire dati ed elementi necessari per controllare il rispetto degli obblighi previsti dalle norme tributarie e rilevare l'eventuale evasione relativamente ad uno specifico aspetto (da cui "mirato"). Nell'ambito dell'accesso il verificatore controlla sia la presenza dei presupposti legali per la richiesta del credito, sia la modalità di utilizzo del credito stesso in relazione a quanto previsto dalla normativa specifica vigente.

imposta. Tra i dati riportati nei verbali, presenti nel data base dei controlli, si è enucleata la variabile "credito di imposta utilizzato nell'anno non spettante", che indica il credito utilizzato in eccedenza rispetto all'importo spettante, così come valutato dal verificatore. I contribuenti che presentano tale grandezza maggiore di zero sono sicuramente soggetti frodatori.

L'informazione di presenza o assenza di un "credito di imposta utilizzato ma non spettante" è sintetizzata in una variabile obiettivo, detta "TARGET", che costituisce l'oggetto effettivo di apprendimento del modello.

Tramite questo apprendimento si analizzano e si classificano i soggetti frodatori al fine di verificare se esiste una tipizzazione che possa essere generalizzata.

Per realizzare questa fase è necessario utilizzare i data base indicati al precedente punto 2).

Durante la fase di apprendimento sono utilizzate tecniche automatiche che, analizzando casi noti, consentono di costruire dei modelli generalizzabili. Una volta terminata questa fase si passa a quella predittiva che ha lo scopo di classificare il soggetto non ancora controllato assegnandogli un punteggio.

Questo punteggio (score, calcolato tramite il modello generato nella fase di apprendimento) fornisce la propensione del soggetto al comportamento identificato come anomalo (nel nostro caso la frode fiscale).

E' possibile riassumere la procedura seguita in modo schematico:

a) è stato selezionato l'insieme di soggetti (indicato come Campione) che hanno subito verifiche o accessi mirati sui crediti di imposta richiesti per Investimenti in Aree Svantaggiate, dei quali è noto se sono o meno frodatori; dall'informazione relativa all'esito del controllo, viene costruita la Variabile Obiettivo (o Target);
b) è stato estratto, dalle banche dati dell'anagrafe tributaria, un set di informazioni di carattere economico – comportamentale, attraverso le quali è possibile caratterizzare il Campione;
c) il set di informazioni è stato associato alla variabile obiettivo;
d) è stata effettuata una analisi esplorativa del campione attraverso una segmentazione comportamentale;
e) è stato selezionato il 70% del campione per sviluppare la fase di apprendimento dell'algoritmo per la costruzione del modello revisionale (Fase di *Training*);
f) sulla base di indicatori statistici sono state *selezionate le variabili più rilevanti* (all'interno del set di informazioni estratte) per costruire un modello che consentisse di individuare il profilo del frodatore (fase di apprendimento, o *training*, vera e propria);
g) sul 30% del campione escluso dalla fase di *training*, si è verificata la capacità previsiva del modello, constatando quante volte il profilo di frodatore indicato dal modello corrispondesse ad un soggetto che ha effettivamente posto in essere una frode accertata (Fase di *Test*).
h) una volta ritenuta soddisfacente la fase di Test, il modello è stato applicato alla popolazione di riferimento, cioè all'insieme dei soggetti che hanno richiesto un credito di imposta per Investimenti in Aree Svantaggiate per i quali è possibile attribuire una indicazione se sono o meno frodatori, con una certa probabilità (Fase di *Apply* del modello)

4. La fase esplorativa

4.1 Le variabili utilizzate per definire il profilo del contribuente

Alla data in cui è stato iniziato il presente lavoro la popolazione di riferimento contava 230.000 contribuenti, comprendente l'insieme dei soggetti che avevano richiesto un credito per investimenti in aree svantaggiate. Da questo universo si è selezionato il campione dei soggetti che avevano subito un accesso mirato e/o una verifica relativamente alla medesima fattispecie contemplata per definire la popolazione di riferimento. La numerosità del campione è risultata pari a 54.517 contribuenti.

La prima analisi che si è resa necessaria è stata quella relativa all'associazione di specifiche variabili economico – comportamentali ai 54.517 soggetti che compongono il campione, finalizzate alla creazione di profili utente utilizzabili per individuare i possibili frodatori. Si è trattato, prevalentemente, di effettuare una selezione tra la vasta mole di informazioni per restringere la scelta unicamente a quelle maggiormente significative per le nostre finalità.

In questa fase è risultato molto utile applicare tecniche di analisi dei dati (indice di correlazione tra le variabili e clusterizzazione) che hanno consentito di individuare l'informazione necessaria per effettuare la segmentazione (raggruppamento in cluster omogenei) del campione.

La procedura ha condotto all'individuazione di due gruppi di variabili: quelle attive, cioè quelle sulle quali sono calcolati gli indici di similarità per effettuare la segmentazione, e quelle descrittive, presentate a corredo per illustrare le caratteristiche di ciascun gruppo di individui.

Il criterio iniziale che ha guidato la selezione delle variabili da utilizzare come attive nella segmentazione è frutto di una scelta ponderata scaturita dalle opinioni di un gruppo di esperti sia di tematiche fiscali (prevalentemente impegnati sul fronte dell'accertamento) che di economia aziendale. Successivamente si è effettuata un'ulteriore scrematura utilizzando una valutazione di carattere statistico basata sull'analisi delle correlazioni e sul comportamento della variabile stessa nel corso della segmentazione[10].

Sono state utilizzate diverse tipologie di variabili[11] :

- localizzazione geografica;
- settore di attività economica;
- natura giuridica;
- permanenza in vita dell'impresa;
- dimensione del soggetto (ricavi, numero addetti, etc.);
- i costi (costo del lavoro, valore aggiunto per addetto, importazioni, esportazioni, etc.);
- il reddito;
- posizione rispetto agli studi di settore;

[10] In questa fase sono state eliminate le variabili eccessivamente correlate tra loro, ciò che incorporavano un'informazione ridondante. In corso d'opera sono state selezionate con cura alcune variabili che, se pur importanti, non potevano essere utilizzate se non aggregate in modo adeguato. Ad esempio la variabile relativa alla collocazione geografica del soggetto non produceva risultati adeguati se considerata a livello provinciale. Si creavano delle problematiche di overfitting del modello, ed in fase di test il modello perdeva di efficacia. E' stata quindi utilizzata a livello di regione entrando bene nel modello.

[11] Tutte le variabili numeriche continue sono state discretizzate utilizzando come soglia i valori dei quartili (identificati con Q_1, Q_2, Q_3 e Q_4); dove gli importi a zero erano in numero rilevante, sono stati calcolati i quartili solo per i soggetti con gli importi maggiori di zero, isolando i soggetti con importi uguali a zero nella classe Q_0.

- informazioni relative al credito di imposta richiesto;
- altre fonti esterne all'Amministrazione.

Alcune variabili sono soltanto descrittive, cioè utilizzate unicamente per qualificare in modo più compiuto i risultati ottenuti dalle analisi statistiche.

Altre invece sono attive, vale a dire che contribuiscono alla creazione del modello.

E' importante sottolineare l'inclusione di variabili relative alla posizione dei contribuenti con riferimento agli studi di settore, anche se questa risulta assente per circa il 30% dei soggetti[12]. Si è ritenuto che questo dato dovesse comunque essere utilizzato in quanto fornisce delle informazioni qualitativamente importanti per quanto riguarda la posizione fiscale dei contribuenti nei confronti del fisco. Per rendere operative queste informazioni sono state costruite delle variabili di sintesi che danno un'indicazione sulla congruità e sulla non congruità in termini di distanza dal ricavo puntuale (rapporto tra ricavo dichiarato e ricavo puntuale). Inoltre è stata considerata anche la eventuale coerenza/incoerenza[13] del soggetto.

4.2 La segmentazione del campione

Per effettuare la segmentazione del campione è stato applicata una procedura di clusterizzazione[14], utilizzando alcune variabili attive. La procedura è stata tarata per ottenere segmenti molto omogenei al loro interno, che consentissero di connotare ciascun gruppo tramite poche caratteristiche economico – comportamentali salienti[15]. La "Segmentazione comportamentale" ha il duplice scopo di cercare di conoscere il gruppo di contribuenti oggetto di studio e di rendere più preciso il modello previsivo, costruendo un modello previsivo per ciascuno dei cluster individuati tramite la segmentazione.

Per valutare la bontà della segmentazione si utilizza l'indice di Condorcet, che varia tra zero e uno, e fornisce una misura sintetica dell'omogeneità interna dei gruppi ottenuti e, al tempo stesso, della diversità, o distanza, fra i segmenti stessi.

---

[12] Il motivo dell'assenza può essere imputato al fatto che il soggetto risulta fuori dal campo di applicazione dello strumento. Si è tenuto conto dell'assenza dell'informazione, per una parte dei soggetti, utilizzando un opzione dell'algoritmo di calcolo che permette di considerare la variabile solo nel caso di presenza di un valore significativo e di ignorarla altrimenti.

[13] Si definisce congruo un contribuente che dichiara un ammontare di ricavi e compensi non inferiore a quello stimato dall'amministrazione (ricavo puntuale) con riferimento ad alcune caratteristiche economiche strutturali dell'azienda. Contestualmente si identificano dei rapporti caratteristici (come ad es. la rotazione del magazzino o i margini di ricarico) che definiscono coerente la dichiarazione del contribuente.

[14] Il progetto è stato sviluppato con il prodotto Intelligent Miner. In particolare è stato utilizzato l'Algoritmo di Demographic Clustering. Coppie di record sono confrontate sulla base delle variabili che pilotano la segmentazione (*variabili attive*). Queste si differenziano dalle altre, le *descrittive*, le quali, una volta creati i segmenti, sono usate per descrivere i contribuenti appartenenti ai vari segmenti. Il numero di campi che hanno valori simili determinano il *grado* con cui i due record sono considerati simili. Il numero di campi che hanno valori dissimili determinano il *grado* in cui i due record sono diversi. Questi due numeri vengono considerati come voti a favore e contro la similitudine dei due record. Il concetto che sta alla base dell'algoritmo di clustering demografico è quello di costruire i segmenti confrontando ciascun record con quelli dei segmenti già creati e assegnarlo a un dato segmento in modo da rendere massimo il primo numero e minimo il secondo. Durante il processo (tipicamente iterativo) vengono creati nuovi segmenti.

[15] In termini tecnici è stato realizzato fissando un indice di similarità pari 0,7. La soglia di similarità controlla la probabilità che i record vengano assegnati a un determinato cluster. Per esempio se si specifica una soglia pari a 0.25 succederà che i record per i quali il 25% dei campi attivi ha valore identico saranno assegnati allo stesso cluster. Naturalmente se si vuole ottenere un elevato numero di cluster occorre aumentare tale parametro. La pratica consiglia di lavorare con valori da 0.5 a 0.75.

Ad ogni segmento viene associata una breve descrizione che ne evidenzia le caratteristiche distintive.

L'elaborazione ha avuto il seguente esito:

- Sono stati individuati 9 insiemi(segmenti o cluster)
- Il Condorcet globale è pari a 0,6149
- Il Condorcet per i singoli segmenti è sempre sopra lo 0,50

Nella tabella 4.3 si riportano i vari segmenti ottenuti dalla procedure, corredati da una breve descrizione, dalla numerosità (in termini assoluti e percentuali), nonché il corrispondente indice di Condorcet.

L'analisi sintetica dei risultati ottenuti porta a concludere che i principali elementi che distinguono tra loro i contribuenti sono:

a) la dimensione (piccole,. medie, grandi e grandissime);
b) la forma giuridica;
c) il tipo di contabilità;
d) la permanenza in vita dell'impresa;
e) il numero di addetti;
f) la posizione rispetto agli studi di settore.

Tabella 4.3 Cluster individuati dal processo di segmentazione all'interno del campione di controlli effettuati sui crediti di imposta in aree svantaggiate

N. Segm.	Descrizione Segmento	Numerosità	Composizione percentuale	Indice di Condorcet
2	Ditte individuali, di piccole dimensioni, in contabilità semplificata, monoaddetti	16.413	30,10	0,6510
0	Ditte individuali di medio piccole dimensioni, in contabilità semplificata, con limitato n. addetti, congrui e coerenti	8.658	15,88	0,5786
6	Ditte individuali di medio piccole dimensioni, monoaddetti, non congrui e non coerenti	7.568	13,88	0,6181
7	Società, in contabilità ordinaria, di grandi e grandissime dimensioni, n. elevato di addetti, congrui e coerenti	7.029	12,89	0,5567
4	Società, in contabilità ordinaria, di medio grandi dimensioni, n. elevato addetti, non congrui e non coerenti	4.392	8,05	0,5408
1	Società, in contabilità ordinaria, di medio piccole dimensioni, non congrui ma coerenti, spesso di recente costituzione	3.723	6,82	0,5279
5	Impresa ordinaria, di medie dimensioni, congrui e coerenti, prevalentemente di più vecchia costituzione	2.596	4,76	0,5112
3	Impresa ordinaria, di medio grandi dimensioni, congrui e coerenti, prevalentemente di più vecchia costituzione	2.521	4,62	0,5047
8	Impresa semplificata, di medie dimensioni, no congrui e non coerenti	1.617	2,97	0,5516
	TOTALE	**54.517**	**100,00**	**0,6149**

Il modello di segmentazione costruito sul campione è stato applicato a tutta la Popolazione, utilizzando quindi le medesime variabili anche sul totale dei contribuenti che hanno fatto richiesta di un credito per investimenti in aree svantaggiate[16].

I risultati sono stati confortanti. Il modello ha tenuto e i segmenti della popolazione identificati rispecchiano le caratteristiche individuate per il campione[17].

[16] Tecnicamente è stata applicata su tutta la popolazione la modalità apply del modello di segmentazione, individuato per il campione; poi sono stati analizzati i risultati.

Tale controllo ha mostrato sostanzialmente la tenuta del modello, in quanto i cluster si sono distribuiti in maniera abbastanza simile tra campione e il totale della popolazione (tabella 4.4).

Tabella 4.4 Composizione percentuale dei cluster ottenuti dalla segmentazione applicata al campione e di quella operata sulla popolazione di riferimento.

N. Segm.	Descrizione Segmento	Composizione Percentuale Campione	Composizione Percentuale Popolazione di riferimento
2	Ditte individuali, di piccole dimensioni, in contabilità semplificata, monoaddetti	30,10	36,44
0	Ditte individuali di medio piccole dimensioni, in contabilità semplificata, con limitato n. addetti, congrui e coerenti	15,88	14,69
6	Ditte individuali di medio piccole dimensioni, monoaddetti, non congrui e non coerenti	13,88	14,55
7	Società, in contabilità ordinaria, di grandi e grandissime dimensioni, n. elevato di addetti, congrui e coerenti	12,89	10,54
4	Società, in contabilità ordinaria, di medio grandi dimensioni, n. elevato addetti, non congrui e non coerenti	8,05	7,13
1	Società, in contabilità ordinaria, di medio piccole dimensioni, non congrui ma coerenti, spesso di recente costituzione	6,82	6,16
5	Impresa ordinaria, di medie dimensioni, congrui e coerenti, prevalentemente di più vecchia costituzione	4,76	3,98
3	Impresa ordinaria, di medio grandi dimensioni, congrui e coerenti, prevalentemente di più vecchia costituzione	4,62	3,88
8	Impresa semplificata, di medie dimensioni, no congrui e non coerenti	2,97	2,62
	TOTALE	100,00	100,00

5. La definizione di "frodatore"

La definizione del contribuente che può definirsi frodatore riveste un'importanza cruciale per la procedura, condizionando sia la fase di impostazione che quella di realizzazione del modello predittivo (cioè del modello che sarà realmente utilizzato per scoprire gli eventuali frodatori). Tale definizione, infatti, rappresenta la variabile obiettivo (target) del modello.

Per conferire una valenza oggettiva alla definizione si devono necessariamente utilizzare i dati riportati nei verbali relativi agli accessi mirati e/o verifiche sui crediti di imposta. In particolare l'attenzione si è soffermata sulle seguenti due informazioni:

1. "credito di imposta *maturato* nell'anno non spettante", che indica il credito, cosi come è calcolato dal contribuente, di cui è accertata l'assenza dei presupposti per usufruirne;

[17] Il gruppo di contribuenti che è stato utilizzato per costruire il modello di segmentazione demografica e successivamente il modello previsivo, ha tutte le caratteristiche per essere considerato un campione sufficientemente rappresentativo della popolazione di riferimento. Infatti è stato verificato che non esistevano delle liste di indirizzamento per gli uffici su questa tipologia di frode fiscale; inoltre il campione è pari a più del 20% della popolazione di riferimento, e quindi statisticamente valido. Infine sono state confrontate le statistiche descrittive del campione e della popolazione di riferimento sulle variabili più importanti, e non sono stati individuati scostamenti significati.

2. "credito di imposta *utilizzato* nell'anno non spettante", che indica l'importo utilizzato in eccedenza all'importo spettante, così come è accertato dal verificatore;

L'informazione riportata al punto 1 è prevalentemente centrata sulle pre-condizioni economico-tributarie che consentono di maturare un credito, mentre quella di punto 2 focalizza l'attenzione sul fatto che il credito di imposta sia stato realmente impiegato.

Dall'osservazione dei dati riportati nel campione dei controlli (tabella 5.1) risulta che al 48,58% dei soggetti non è stata effettuato alcun rilievo né di natura formale né di natura sostanziale (in quanto presentavano sia una importo maturato che un importo utilizzato non spettante uguale a zero). L' 11,95% presentava solamente irregolarità formali (importo maturato maggiore di zero ed importo utilizzato uguale a zero), mentre per il restante 39,46% (codici 2 e 4) sono state rilevate irregolarità sostanziali (sia associate che non associate ad irregolarità formali).

Tabella 5.1 Suddivisione del campione sulla base delle informazioni utilizzate per definire i soggetti frodatori

Codice	Descrizione	Numerosità	Composizione percentuale	Tipo di rilievo	Variabile obiettivo (Target)
1	Importo maturato non spettante = 0 e Importo utilizzato non spettante = 0	26.484	48,58%	1 – assenza di rilievo	0 – Non frodatore
2	Importo maturato non spettante = 0 e Importo utilizzato non spettante > 0	12.647	23,20%	2 – solo con rilievi sostanziali	**1 - Frodatore**
3	Importo maturato non spettante > 0 e Importo utilizzato non spettante = 0	6.514	11,95%	3 – solo con rilievi formali	0 – Non frodatore
4	Importo maturato non spettante > 0 e Importo utilizzato non spettante > 0	8.864	16,26%	4 - con rilievi sostanziali e formali	**1 - Frodatore**
	Totale soggetti del campione	54.517	100,00%		

Ancorché da un punto di vista giuridico si potrebbero definire frodatori tutte le tipologie della tabella, ad eccezione della prima, per le finalità del modello si è scelto di restringere la definizione unicamente alle tipologie 2 e 4. Tale scelta è stata dettata da una duplice motivazione:

I. i soggetti individuati sono coloro che effettivamente hanno utilizzato il credito, vale a dire con l'importo utilizzato non spettante maggiore di zero, e, quindi, sono i soggetti nei confronti dei quali l'Amministrazione può esigere una restituzione;
II. il modello predittivo si basa sull'osservazione dei comportamenti economici dei contribuenti ed è, quindi, più plausibile che differenze in questi comportamenti "spieghino" le frodi sostanziali, piuttosto che irregolarità negli adempimenti formali[18].

[18] Su questo aspetto è stata condotta una simulazione che ha portato a concludere che l'applicazione del modello ai rilievi solo formali non portava risultati statisticamente ammissibili.

6. Il modello predittivo

La seconda fase del progetto mira alla costruzione di un modello previsivo. Lo scopo è quello di classificare un soggetto non ancora verificato come potenziale frodatore assegnandogli uno score (ovvero un punteggio) che fornisce una indicazione sulla propensione del soggetto al comportamento identificato come anomalo (nel nostro caso la frode fiscale).

Dopo aver costruito il modello di segmentazione, sono stati compiuti diversi tentativi di costruzione del modello previsivo, seguendo varie strategie[19].

Sono stati costruiti modelli previsivi sull'intero campione. Inoltre sono stati selezionati alcuni segmenti che suscitavano maggior interesse, provando anche ad accorparne altri che si ritenevano simili, ed anche su questi sono stati costruiti dei modelli. Questa modalità di lavoro ha permesso di tarare il modello verificando quale fosse l'approccio migliore per procedere. Il modello finale che è stato selezionato, è quello costruito sull'intera popolazione, e può quindi essere considerato il più idoneo.

Nel corso dell'analisi della segmentazione comportamentale, si è verificato che la distribuzione della variabile target, indicante la presenza di frode, era sostanzialmente costante e simile a quella dell'intera popolazione. I segmenti non hanno quindi mostrato variazioni significative nella percentuale dei frodatori. Questo ha contribuito ad indirizzare la costruzione del modello non sui singoli segmenti ma sull'intera popolazione

6.1 La costruzione del modello

Dal data set iniziale (il campione iniziale 54.517 contribuenti) è stato estratto un sub-campione casuale di 38.497 unità (chiamato *training set* e pari al 70% dell'iniziale) che è stato utilizzato per stimare il modello, ovvero - adottando un terminologia tipica del *data mining* - che ha la funzione di istruire l'algoritmo per la costruzione del modello.

Nella fase di *training* sono state analizzate le variabili in funzione dell'obiettivo definito nel paragrafo 5. Dal risultato del *training* si ottiene il "profilo" del possibile frodatore, espresso attraverso le grandezze incluse nel modello, vale a dire attraverso il suo comportamento economico, cosi come appare al fisco.

Tra i possibili algoritmi utilizzabili per costruire il modello si è deciso di applicare quello basato sull'albero di decisione (*decision tree*). L'albero di decisione è un procedimento che

[19] Al fine di investigare l'associazione tra la variabile Target FRODE e ciascuna delle variabili esplicative (dicotomiche e continue) sono stati costruiti ed interpretati gli *odds ratio*.
Gli *odds ratio* testano il segno e la forza di associazione tra due variabili: nel nostro caso si tratta di valutare se le variabili sono associate alla variabile FRODE ovvero se il valore assunto da una variabile condiziona il valore assunto da FRODE.
L'odds ratio fra due variabili può assumere valore assoluto compreso tra 0 e +infinito; nel caso si verifichi il valore 1 allora c'è indipendenza tra le variabili. Più il valore è lontano da 1, sia sopra che sotto ad 1, maggiore è la forza della relazione. I valori superiori ad 1 indicano un'associazione positiva tra le variabili, mentre i valori inferiori ad 1 indicano un'associazione negativa.
Nel caso di variabili categoriche con più modalità è stato utilizzato il chi quadro per valutare la presenza di associazione
Per le variabili continue si è proceduto preliminarmente alla creazione di due classi utilizzando la mediana come separatore delle classi.
Si tratta di un'analisi iniziale volta a individuare tra tutte le variabili quelle maggiormente "promettenti" al fine di distinguere tra frodatori e non frodatori. Le informazioni ricavate dall'analisi univariata, anche se trascurano le eventuali sinergie che emergono dall'uso congiunto delle variabili, forniscono una prima indicazione per la corretta messa a punto di un modello multivariato.

consente di visualizzare una sequenza di decisioni ed i possibili eventi che da queste possono scaturire. Nella realizzazione proposta la sequenza di decisioni deve condurre ad identificare un profilo di frodatore; questo risultato si realizza effettuando delle partizioni della popolazione originaria, seguendo un criterio binario.

La figura 6.1 esemplifica il percorso logico dell'albero delle decisioni. Il cerchio da cui diparte tutta la serie di suddivisioni è indicato come "nodo iniziale" e contiene la totalità del popolazione inclusa nel *training set*. All'interno del nodo iniziale è indicata in verde la componente dei frodatori e in blu quella dei non frodatori. Il nodo è compreso all'interno di una corona che rappresenta la presenza di frodatori nel totale del campione. Dato che il nodo iniziale coincide con il totale del campione la ripartizione tra frodatori e non frodatori è identica sia nella parte interna che nella corona del nodo iniziale.

Dal nodo iniziale si dipartono due linee (i "rami" dell'albero), che indicano una partizione binaria della popolazione originaria. Tale bipartizione è stata effettuata utilizzando la variabile (attiva, inclusa nel modello): presenza di indebito utilizzo del credito, ponendo nel nodo di destra i contribuenti che hanno effettivamente utilizzato indebitamente il credito e nel nodo si sinistra gli altri. Confrontando la parte centrale del nodo con la corona si può osservare che quello di destra presenta una percentuale di frodatori molto maggiore di quella riscontrata all'interno del campione (l'area verde della parte interna del nodo è molto più ampia di quella riportata nella corona).

Scendendo sempre più in basso troveremo le foglie dell'albero che rappresentano gruppi individuati sulla base di più variabili di classificazione e, a ciascuna foglia, sarà associato un grado di purezza che rappresenta la quota di frodatori/non frodatori presenti all'interno del gruppo individuato.

Lo scopo della procedura è quello di massimizzare l'omogeneità finale delle foglie (*purezza*) in termini di valore della variabile Target. Tanto maggiore è tale purezza tanto più precisamente si potrà identificare il profilo tipico del frodatore.

Operativamente, l'algoritmo funziona massimizzando l'omogeneità dei gruppi (sottopopolazioni) in termini di valore della variabile obiettivo (la definizione di frodatore). A questo risultato si arriva per bipartizioni successive, effettuate scegliendo un criterio basato sulle variabili attive incluse nel modello a cui è associato un valore soglia che individua il grado di purezza raggiunto a ciascun livello di disaggregazione. Le soglie rappresentano le *regole di classificazione* che sono di facile interpretazione e potrebbero essere convertite in *query*, tramite operazioni algebriche.

Nelle figure 6.2 e 6.3 si riportano due esempi di regole che individuano, rispettivamente, frodatori e non frodatori.

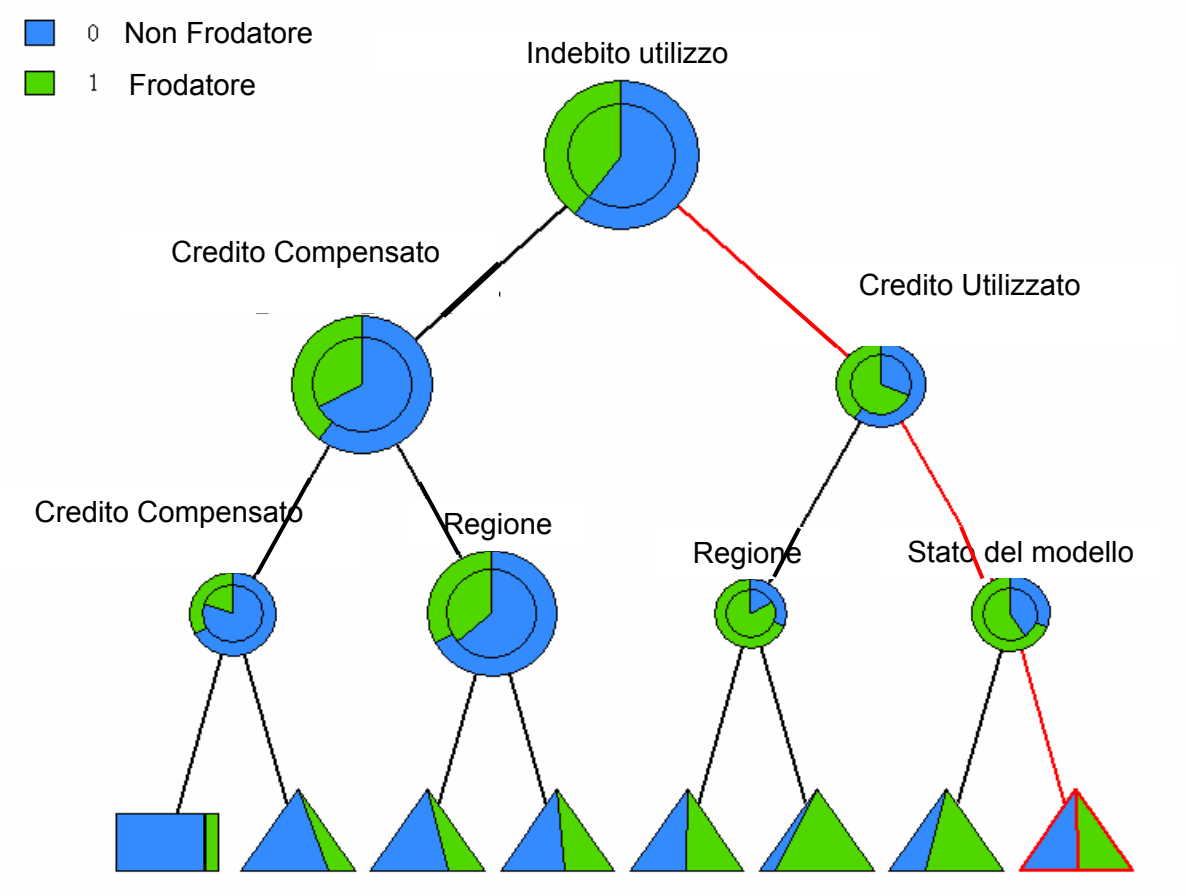
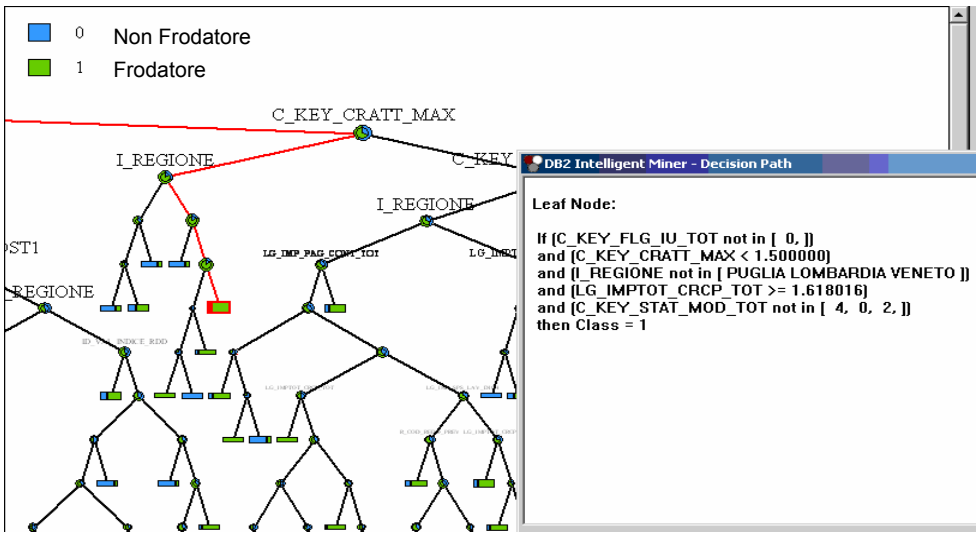


Figura 6.1 Rappresentazione grafica della parte superiore dell'albero delle decisioni

Figura6.2. Esempio 1 di Regola – individua frodatori

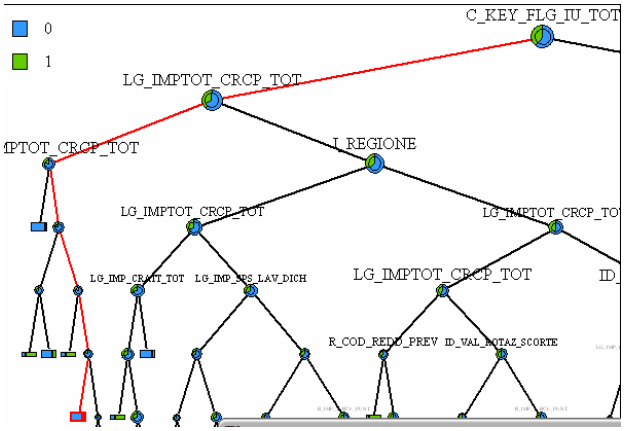


Soggetti=2.651 Frodatori=2.303 Non frodatori=348 Purezza=86,9%

Presenza di indebito utilizzo, assenza di credito attribuito, importo del credito compensato maggiore di 42 euro, regione prevalente Sicilia (42%), assenza del

➡ Frodatori nel 86,9% dei casi

Figura 6.3. Esempio 2 di Regola - individua non frodatori



Soggetti=2.056 Frodatori=331 Non Frodatori=1.725 Purezza=83,9%

Assenza di indebito utilizzo, importo compensato compreso tra 58 euro e 991 euro, regioni prevalenti Sicilia, Calabria, Puglia (Basilicata), reddito prevalente da impresa semplificata, importo relativo alle spese per i lavoratori minore di 19.201 euro,

➡ Non Frodatori nel 83,9% dei casi

Le indicazioni del modello potrebbero essere utilizzate in due modalità: selezionando i potenziali frodatori tra quelli con lo score più elevato, oppure eliminando dai potenziali frodatori quelli che hanno score più basso (scarsa probabilità di frode).

Poiché l'obiettivo del progetto è stilare delle liste per l'attività di controllo, il modello è stato utilizzato secondo la prima modalità, considerando i soggetti con lo score più elevato.

La facilità nell'interpretazione, unita alla velocità di elaborazione dell'algoritmo, fa si che gli alberi di decisione siano le tecniche maggiormente utilizzate per questo tipo di analisi.

6.2 La verifica del modello

In questa fase si applica il modello, definito nel paragrafo precedente, alla popolazione per la quale sono note sia le variabili indipendenti sia la variabile obiettivo; nel nostro caso è pari al 30% della popolazione (insieme che definiamo come *test set).*

La popolazione originaria contava 54.517 soggetti, e quindi il *test set* comprende 16.019 contribuenti (30%). E' importante per la verifica del modello, che si utilizzi una popolazione non interessata dalla fase di training, al fine di verificare la tenuta del modello stesso in fase predittiva; inoltre può contribuire ad ovviare il problema di impiegare modelli decisionali troppo dipendenti dai dati utilizzati per costruirli, definito in termine tecnico come *overfitting*.

In sintesi, in questa fase si effettua una simulazione di come potrebbe funzionare il modello nella realtà, cioè quando verrà applicato sull'intera popolazione di riferimento. Nella fase di test si può immediatamente effettuare tale verifica perché si dispone dell'informazione sulla variabile obiettivo (l'esito del controllo).

La valutazione della bontà in fase di test si effettua utilizzando la "Matrice di confusione" illustrata nella tavola 6.2. Letta per riga la tabella fornisce i valori realmente osservati nel test set, mentre per colonna si riporta la classificazione dei contribuenti effettuata tramite il modello predittivo.

Tavola 6.2 Schema teorico della matrice di confusione utilizzata per verificare l'efficacia del modello in fase predittiva

		Valori assegnati		
		Non frodatore	Frodatore	
Valori reali	Non frodatore	**TN**	**FP**	Realmente non frodatori
	Frodatore	**FN**	**TP**	Realmente frodatori
		Assegnati dal modello alla categoria non frodatori	Assegnati dal modello alla categoria frodatori	Totale

L'incrocio delle righe e delle colonne illustra tutte le possibili combinazioni che si possono ottenere considerando congiuntamente la distribuzione reale e quella teorica del modello; in particolare si possono verificare le seguenti possibilità:

1. Il soggetto è effettivamente non frodatore e cosi è identificato nel modello, il caso è indicato come **vero negativo** (sigla **TN**, cioè True Negative);

2. Il soggetto è effettivamente non frodatore ma è identificato nel modello come frodatore, il caso è indicato come **falso positivo** (sigla **FP**, cioè False Positive);
3. Il soggetto è effettivamente frodatore ma è identificato nel modello come non frodatore, il caso è indicato come **falso negativo** (sigla **FN**, cioè False Negative);
4. Il soggetto è effettivamente frodatore e cosi è identificato nel modello, il caso è indicato come **vero positivo** (sigla **TP**, cioè True Positive);

E' immediato concludere che i casi nei quali il modello da risultati soddisfacenti sono dati da **TN** (vero negativo) e **TP** (vero positivo), mentre negli altri due casi si commetterebbe un errore di assegnazione. Da questa deduzione elementare si derivano i due indicatori di sintesi utilizzati per valutare la bontà dell'adattamento del modello:

i. Correttezza (*Precision*) = **TP**/(**TP**+**FP**); percentuale di soggetti realmente frodatori correttamente classificati dal modello, rispetto al totale di soggetti classificati dal modello come frodatori;
ii. Completezza (*Recall*) = **TP**/(**TP**+**FN**); percentuale di soggetti realmente frodatori correttamente classificati dal modello, rispetto al totale di soggetti realmente frodatori;

Gli indicatori i) e ii) sono entrambi importanti e, da un punto di vista logico, il risultato migliore si otterrebbe se entrambi fossero uguali ad 1. Ciò vorrebbe dire che non esistono né i falsi positivi né i falsi negativi e, quindi, il modello rappresenta perfettamente la realtà. Purtroppo questa eventualità non si verifica mai e, quindi, occorre operare la scelta di massimizzare i) o ii).

Se si intende utilizzare l'analisi del rischio in modo preventivo, cioè, come nel caso delle carte di credito, bloccando all'origine la transazione che può dar luogo alla frode è necessario che il modello presenti un'elevata Completezza (*Recall* si rende minima la probabilità che non si intercetti un soggetto realmente frodatore). Nel caso di studio, si dovrebbe perseguire questo obiettivo qualora si intendesse porre in essere strumenti che limitano l'accesso al credito nel momento della formulazione della domanda (ad esempio bloccandone l'iter burocratico).

Dato che l'obiettivo della procedura che si sta testando è quello di elaborare un'analisi dei rischi utilizzabile per stilare delle liste per l'attività di controllo, l'indicatore che maggiormente interessa è quello relativo alla Correttezza (*Precision*) (indica quanto il modello è stato in grado di classificare bene i soggetti con riferimento alla platea di contribuenti definiti come frodatori dal modello stesso), al fine di minimizzare i costi connessi alla realizzazione degli accertamenti.

Nella tavola 6.3 si illustra la matrice di confusione relativa al test del modello individuato nel par. 6.1 fornita automaticamente dal prodotto, corredata dai rispettivi indici di Correttezza e Completezza.

Il prodotto fornisce un esito di default relativo al modello. In questo caso è stata fissata una soglia dello score pari a 0,5. Se i soggetti hanno uno score maggiore di 0,5, allora vengono considerati frodatori. In base agli indici si conclude che su cento soggetti identificati dal modello come frodatori circa 65 lo sono realmente (correttezza), mentre rispetto a 100 contribuenti realmente frodatori il modello ne identifica circa 40 (completezza).

Tavola 6.3 Matrice di confusione fornita in automatico dal prodotto per stimare l'efficacia del modello previsivo

		Valori assegnati		
		Non frodatore	Frodatore	
Valori reali	Non frodatore	**8.377**	**1.362**	9.739
	Frodatore	**3.778**	**2.502**	6.280
		12.155	**3.864**	16.019

Correttezza (precision) = 64,75. Completezza (recall) = 39,84.

7. La fase di applicazione del modello

In questa fase si applica il modello al totale della popolazione di riferimento, cioè a tutti i soggetti che hanno inoltrato richiesta per un credito di imposta (cioè a circa 230.000 soggetti). Come risultato a ciascun contribuente è associato un punteggio (*score*), che assume valore compreso tra zero e uno e 1; è zero qualora il soggetto rivesta la minima probabilità di essere frodatore, ed è uno per i soggetti a più alto rischio.

Il punteggio deriva dall'osservazione dei dati noti e sistematicamente rilevati all'interno delle banche dati dell'Agenzia, opportunamente ponderati tramite la fase di apprendimento del modello (si veda par. 6.1). In altre parole, tramite la costruzione dell'albero delle decisioni, sono state estratte, da un ampio data base di dati rappresentativo del normale comportamento economico dei soggetti, delle regole che permettono di ordinare i contribuenti in base al grado di rischio di frode.

L'approccio seguito si caratterizza, per di più, anche per essere multi criterio.

Il modello tiene conto di tutte le variabili utilizzate per la costruzione del modello, proponendo delle regole specifiche e articolate.

E' sempre opportuno ricordare che la definizione delle regole si trae dall'osservazione sistematica della realtà e, pertanto, il modello può variare nel corso del tempo in base al mutare del comportamenti dei contribuenti ma, aspetto ben più importante, può apprendere dall'esperienza a seguito dell'ampliamento del campione.

I contribuenti sono stati ordinati, dal più alto al più basso, in base al punteggio assegnato dal modello e poi suddivisi in decili[20]. Il primo decile conterrà quindi i soggetti ai quali è stato attribuito lo score più elevato.

Per la finalità del progetto, che mira ad identificare i frodatori con la massima precision ed evitando inutili accessi di controllo, si è fatto uno studio del comportamento del modello in base ai decili.

Come è stato appurato studiando il Test set, agli score più alti corrisponde una correttezza più alta, vale a dire maggiore è la probabilità che il soggetto risulti un frodatore, maggiore è la possibilità da parte dell'Amministrazione di incorrere in controlli con esito favorevole.

[20] Se ordiniamo un insieme di dati rispetto ad una certa variabile, i valori della variabile che divino l'insieme in dieci parti uguali si dicono decili.

Considerando approssimativamente il primo decile, ponendo la soglia dello score a 0,85 nel campione di test e considerando i soggetti al di sopra di tale valore, si è verificato che la correttezza è pari a 0,82, valore ritenuto accettabile. Quindi su 100 contribuenti controllati, il modello ci dice che 82 saranno frodatori e l'errore sarebbe ristretto a 18 soggetti.

Nel caso che la scelta fosse effettuata senza il modello, la percentuale di frodatori individuati è quella che è stata rilevata nella definizione della variabile target, pari a circa il 40% (vedi la percentuale di frodatori nella tabella 5.1)

La scelta è guidata anche da considerazioni relative alla capacità operativa ed ai costi che si intendono sostenere, limitando o ampliando il numero dei controlli di conseguenza.

7. Conclusioni

La sperimentazione, condotta applicando le tecniche di data mining al caso dei crediti di imposta, ha fornito risultati estremamente soddisfacenti, in quanto:

1. ha consentito di derivare degli indicatori sulla probabilità di essere frodatori scaturiti dall'osservazione delle informazioni disponibili per ciascun contribuente;
2. la probabilità individuata al punto 1 è stata derivata da regole che sono state analizzate e possono fornire ulteriori spunti di approfondimento futuro o confortare ipotesi qualitative formulate precedentemente;
3. l'algoritmo adotta una logica multi criterio che ben si adatta al perseguimento di attività di deterrenza all'evasione (cioè coniuga la trasmissione di segnali forti con il perseguimento di più obiettivi simultaneamente);
4. è stata effettuata una ulteriore verifica, analizzando le verifiche svolte dagli uffici dopo l'elaborazione del modello, ed è risultata confermata la bontà dello strumento dato che ogni cento controlli effettuati ha individuato 82 frodatori sostanziali (a fronte di un probabilità stimata dell'85%).

Di contro, la procedura richiede un sforzo di investimento iniziale molto oneroso, che però è adeguatamente remunerato elevando significativamente il tasso di positività dei controlli. Inoltre, dato che gran parte dello sforzo si concentra principalmente nella predisposizione e nell'analisi delle informazioni di base, una volta predisposto il modello le successive implementazioni consentono di realizzare significative economie di scala.

Alla luce dei risultati conseguiti e della flessibilità dello strumento, sarebbe auspicabile applicarlo ad altri campi di osservazione, che presentano un'elevata numerosità di soggetti da controllare ed una vasta mole di variabili da considerare congiuntamente, come, ad esempio: i crediti IVA o le compensazioni tramite F24.

Riferimenti bibliografici

AAAI/MIT Press, 1996; Hanand M. Kamber. Data Mining: Concepts and Techniques.

U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy. Advances in Knowledge Discovery and Data Mining MorganKaufmann, 2000.

E-Business News (2001) SOLUZIONI – Data Mining, una lente per l'e-business, n. 8, settembre.

Paolo Giudici Data Mining Mc Graw-Hill Companies, 2001.

D.Hand, H.Mannilaand P.Smyth Principles of Data Mining. The MIT Press, 2001.

T. Hastie, R. Tibshiraniand J. Friedman. The Elements of Statistical Learning. Springer 2001.

G. Piatetsky-Shapiroand W. J. Frawley. Knowledge Discoveryin Databases. AAAI/MIT Press, 1991.

V.N. Vapnik Statistical Learning Theory Wiley&Sons, 1998.