Progetto DIVA:
Data mining per l'individuazione delle frodi del credito IVA

Consiglio Nazionale della Ricerca (**CNR**): Istituti ISTI & ICAR

- Istituto di Calcolo e Reti ad Alte prestazioni (**ICAR**)
- Istituto di Scienza e Tecnologie dell'Informazione (**ISTI**)

Agenzia delle Entrate

Sogei Spa (Partner ICT dell'Amministrazione finanziaria, dal 1976)

Il problema dei crediti IVA

- Il credito IVA rappresenta:
  - una componente **fisiologica** nella gestione del tributo (esportaz; invest; apertura e chiusura)
  - una componente **strutturale** dell' evasione

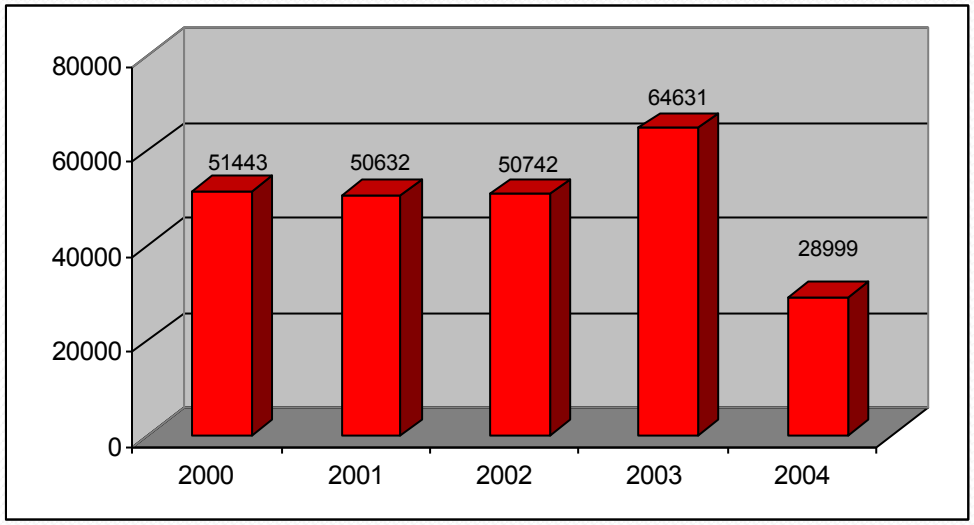
- Si dichiarano acquisti inesistenti di beni e servizi (falsi consumi intermedi) e si sotto-dichiara il ricavo delle vendite (omessa fatturazione di bar, ristoranti, ecc.).
- Si ottiene quindi un imponibile ridotto anche per le imposte sul reddito: l' IVA è il **moltiplicatore** dell' evasione fiscale sulla produzione
- Con il falso credito IVA si pagano le altre imposte grazie all' istituto della compensazione (Mod. F24)

Il problema per l'Agenzia delle Entrate

- La numerosità dei contribuenti a rischio (metà della platea) e la crescita del fenomeno nel lungo periodo
- Nel 2006 i crediti IVA rinviati all'anno successivo ammontano a circa 42 miliardi di euro: quasi il 5% della BIT contro il 2,1% degli anni ottanta;
- La difficoltà di separare gli evasori dai contribuenti corretti e da quelli che hanno commesso errori, a volte involontari, di lieve entità,
- La complessità delle informazioni necessarie all'Agenzia per avere successo nell'azione di controllo e poi nelle liti tributarie.

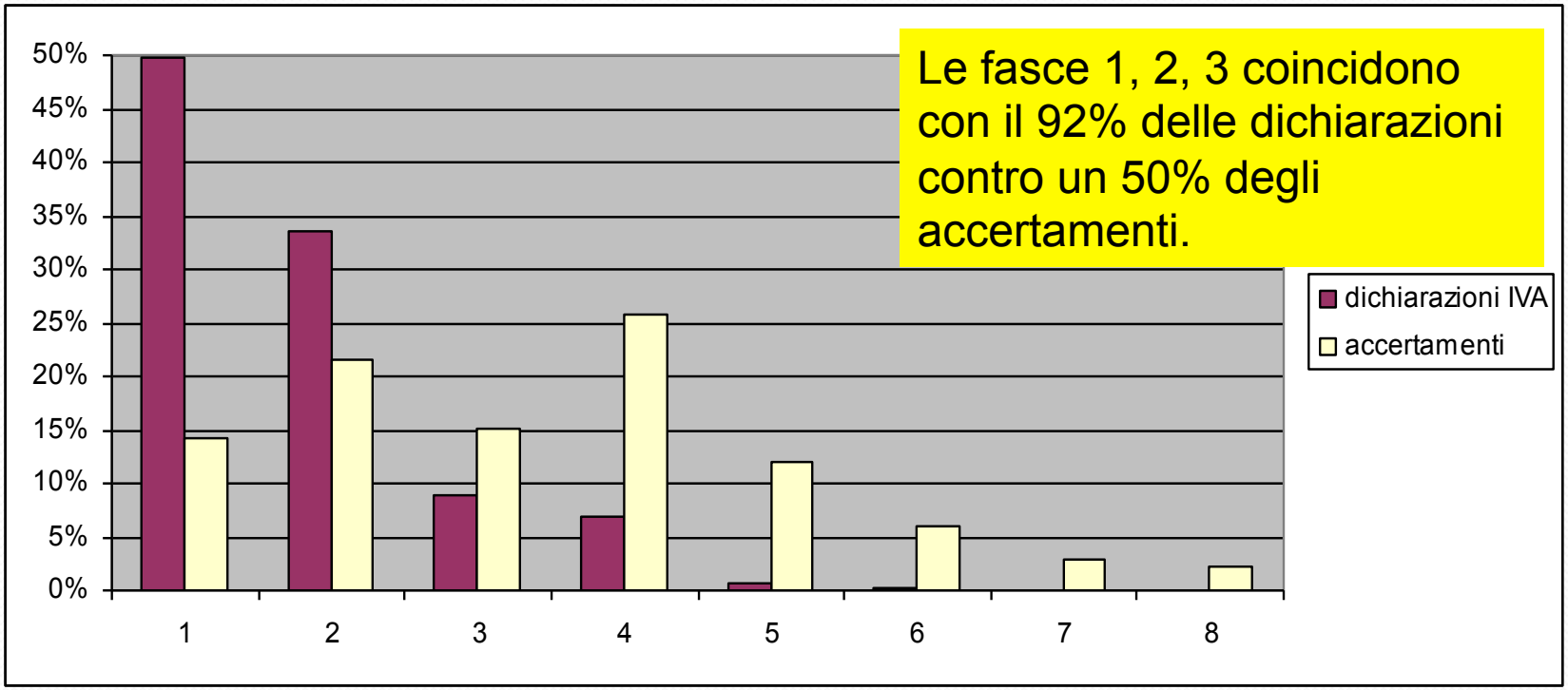
Progetto DIVA (1)

- studio, prototipazione e sperimentazione di tecniche di data mining per la costruzione di modelli previsionali **finalizzati all'analisi di rischio sui crediti IVA**
- utilizzare la conoscenza implicita contenuta negli accertamenti fiscali per far emergere l'esperienza e le buone pratiche
  - **38mila accertamenti IVA con credito iva di competenza dichiarato > 0**



Progetto DIVA (2)

Confronto tra la distribuzione delle dichiarazioni con IVA e i relativi accertamenti, distribuzione discretizzata sul volume d'affari dichiarato

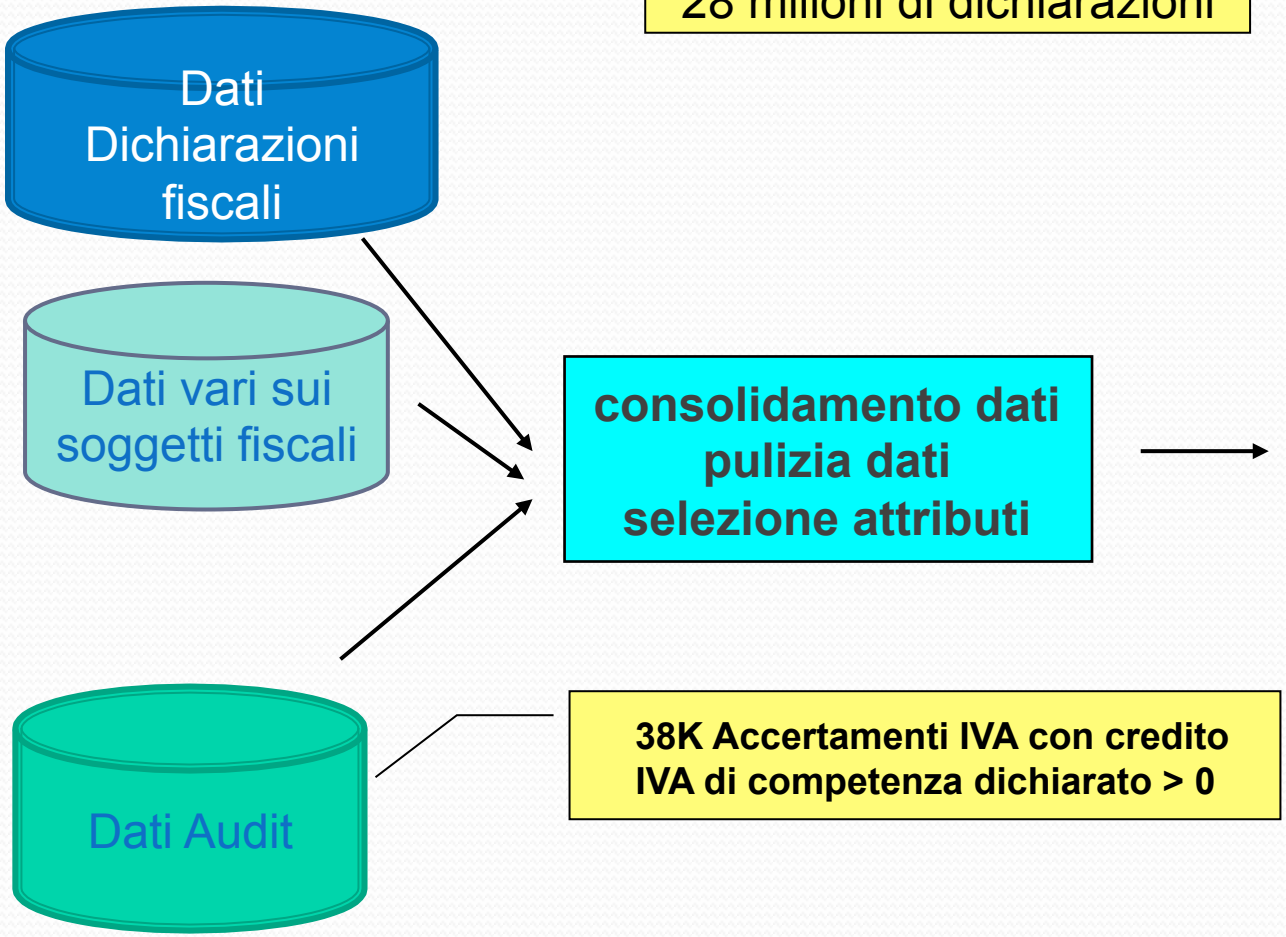


1 - VOL. AFF. DA 0 FINO A 30mila
2 - VOL. AFF. OLTRE 30mila FINO A 185mila
3 - VOL. AFF. OLTRE 185mila FINO A 516mila
4 - VOL. AFF. OLTRE 516mila FINO A 5milioni
5 - VOL. AFF. DA 5milioni FINO A 12milioni
6 - VOL. AFF. OLTRE 12milioni FINO A 25milioni
7 - VOL. AFF. DA 25milioni FINO A 51milioni
8 - VOL. AFF. OLTRE 51milioni

Dati a disposizione

Nome Tabella	Num. di Righe	Num. di Attributi	Attributi Chiave
IVA	33.929.801	110	Anno d'imposta, Identificativo soggetto, Tipo dichiarazione, Progressivo dichiarazione, Progressivo modello
RADAR	32.863.041	140	Anno d'imposta, Identificativo soggetto, Tipo dichiarazione, Progressivo dichiarazione
ACCERTAMENTI	406.443	56	Numero accertamento
ANAGRAFE	32.863.041	30	Anno d'imposta, Identificativo soggetto, Tipo dichiarazione, Progressivo dichiarazione
REDDITI	31.997.279	54	Anno d'imposta, Identificativo soggetto, Tipo dichiarazione, Progressivo dichiarazione

Dati



Accertamenti
dati dichiarati
dati accertati
dati definiti
Dichiarazione IVA
RADAR
prospetto di bilancio
movimentazioni di capitali da e per l'estero
crediti d'imposta
dati da fonti INPS
dichiarazioni dei sostituti d'imposta
dati forniti dalle banche
frontespizi
Dati anagrafici
informazioni sulla p.iva
informazioni su tutte le p.iva del soggetto
Redditi
dati sui redditi da lavoro autonomo
dati su redditi di impresa ordinaria
dati sui redditi di impresa semplificata
compensazioni e rimborsi IVA
determinazione dell'IRPEF/IRES

Operazioni di pre-processing (1)

- Si sono unite le varie fonti in un'unica tabella, con lo scopo di fornire un'informazione di sintesi riguardante i singoli contribuenti accertati in uno specifico anno d'imposta
- Sono stati costruiti attributi calcolati, per:
  - Semplificare alcune informazioni
  - Esplicitare i "significati nascosti" nei dati
  - Introdurre la conoscenza del dominio da parte degli esperti

Operazioni di pre-processing (2)

- Eliminazione di alcuni campi ritenuti non utili per le successive fasi d'analisi (ad esempio, perché non sufficientemente valorizzati o perchè rappresentanti informazioni di tipo 'data' o perché ancora generalizzati da attributi introdotti in precedenza)

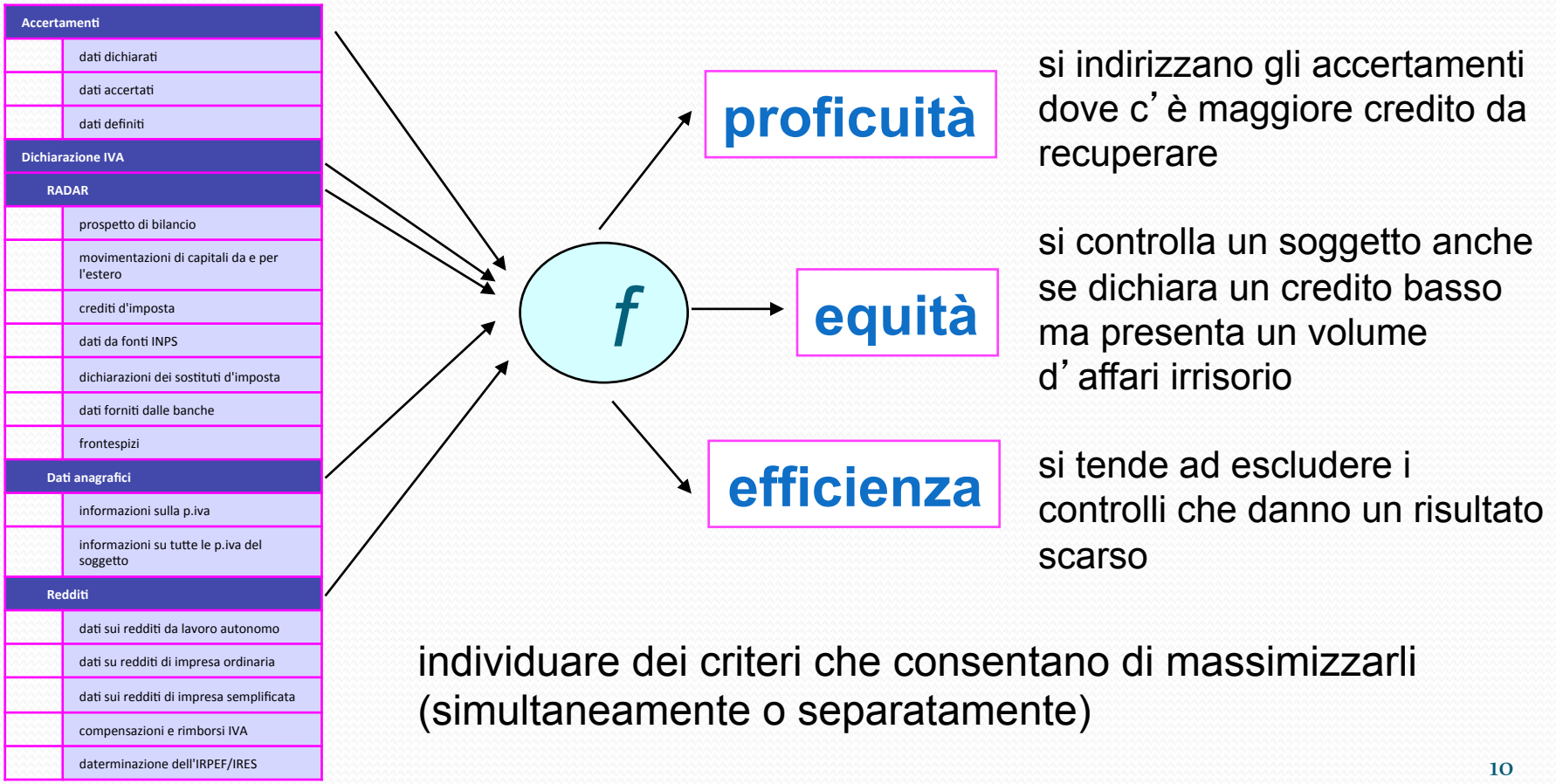
- Eliminazione di quegli attributi che più di tutti sono risultati correlati ad altri; a tal fine, si è considerata utile una soglia di correlazione limite di 0.9

- Eliminazione dei campi la cui valorizzazione è apparsa estremamente sbilanciata; anche qui come soglia, si è utilizzato il valore 0.9

La tabella risultante ha 45'442 record e circa 400 attributi, per un totale di 37'116 contribuenti

Soluzione proposta (1)

- Non è una situazione standard del tipo *buoni* e *cattivi*: si definiscono frodatori *interessanti* (che vale la pena accertare) e *non-interessanti* (che non vale la pena accertare)



si indirizzano gli accertamenti dove c'è maggiore credito da recuperare

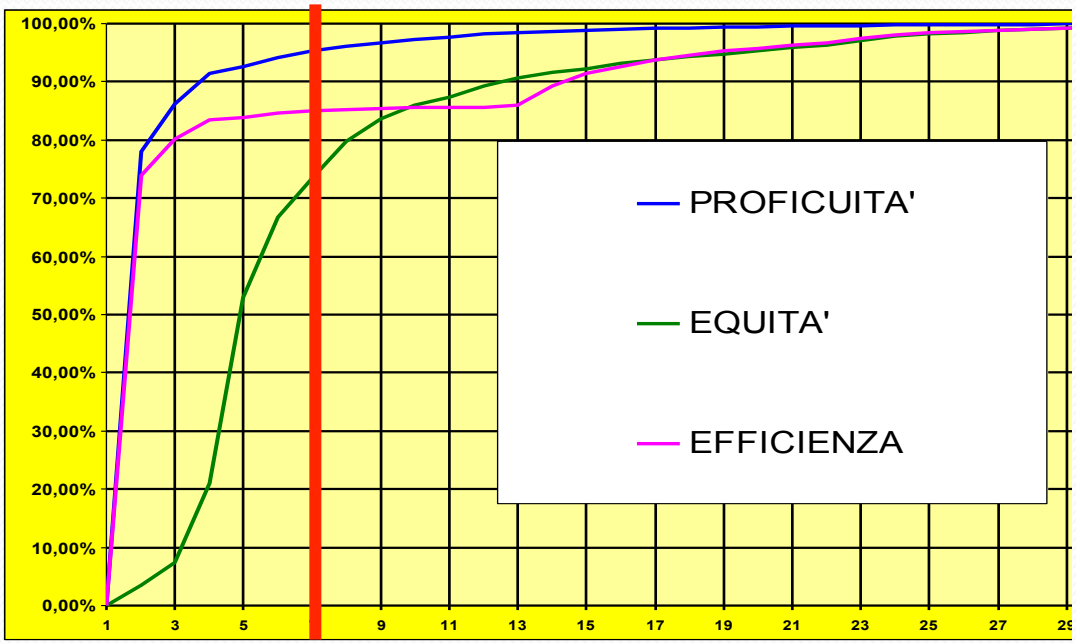
si controlla un soggetto anche se dichiara un credito basso ma presenta un volume d'affari irrisorio

si tende ad escludere i controlli che danno un risultato scarso

individuare dei criteri che consentano di massimizzarli (simultaneamente o separatamente)

Soluzione proposta (2)

- Sulle ascisse le dichiarazioni in ordine decrescente rispetto ai tre diversi criteri
- Sulle ordinate la percentuale di frode recuperata



Confronto tra l'andamento dei tre criteri

Esempio: prendendo i primi 7mila (linea rossa) frodatori più '*promettenti*', si ottiene un recupero del 95% della frode secondo il criterio della proficuità e circa l'85% secondo l'efficienza e 75% rispetto all'equità

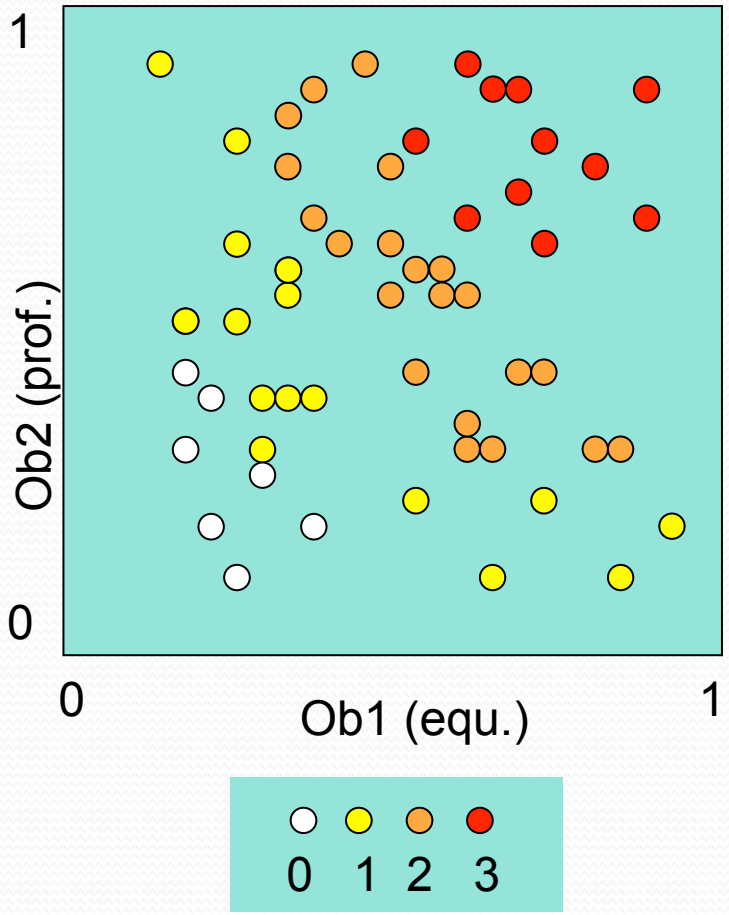
E' possibile usare i 3 criteri **simultaneamente** senza compromettere significativamente l'entità di frode recuperata

SCORE: assegnare un 'punteggio' ad ogni contribuente (1)

- per ogni contribuente si calcola un valore (SCORE) attraverso una formula che combina i tre obiettivi precedentemente descritti (ad esempio) :
  - efficienza * profiquità * equità
- questo valore viene utilizzato per assegnare un punteggio tra 0, 1, 2 e 3 (0 = frodatore non interessante, 3 = molto interessante)
- le soglie che determinano l'appartenenza ad una delle quattro classi suddividono la platea dei contribuenti nel seguente modo:
  - 0 → 50 %
  - 1 → 25 %
  - 2 → 17,5 %
  - 3 → 7,5 %

SCORE

Esempio con due funzioni obiettivo per semplicità rappresentativa



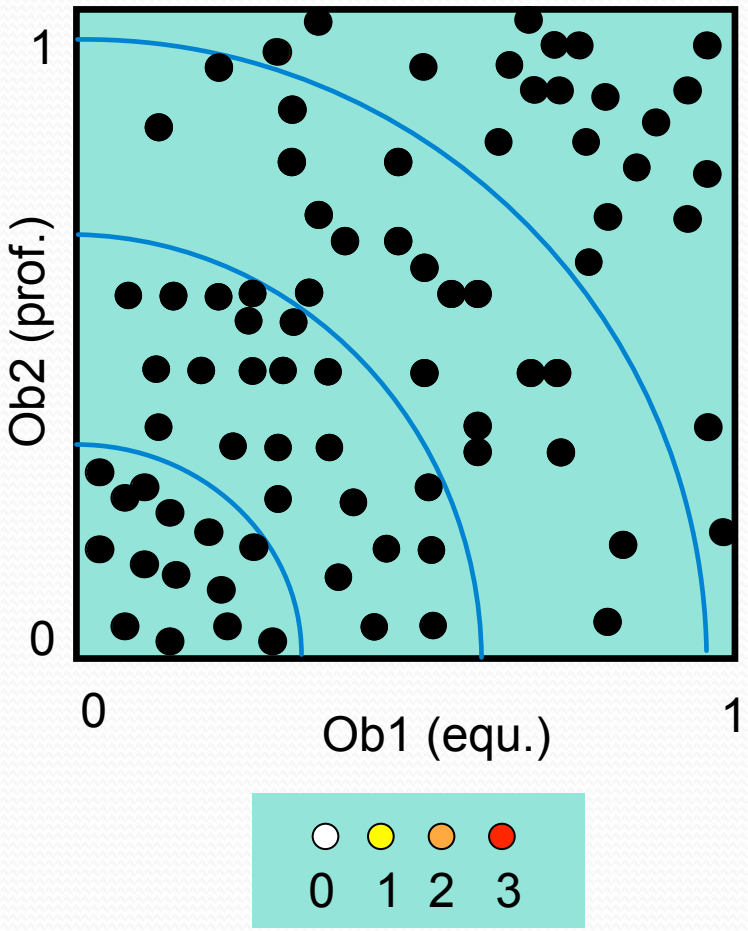
I frodatori che hanno un valore prossimo a 1 per entrambi gli obiettivi (tre nella realtà), vengono considerati più interessanti o a rischio e classificati con il valore 3 (rosso).

I frodatori che vengono classificati come 2 (arancioni) hanno o valori simili per entrambi gli obiettivi oppure uno alto e l'altro sotto la media per cui lo score che viene assegnato è più basso del precedente

Si procede nello stesso modo per interpretare i contribuenti classificati 1 e 0.

SCORE

Esempio (con 2 funzioni obiettivo per semplicità rappresentativa)



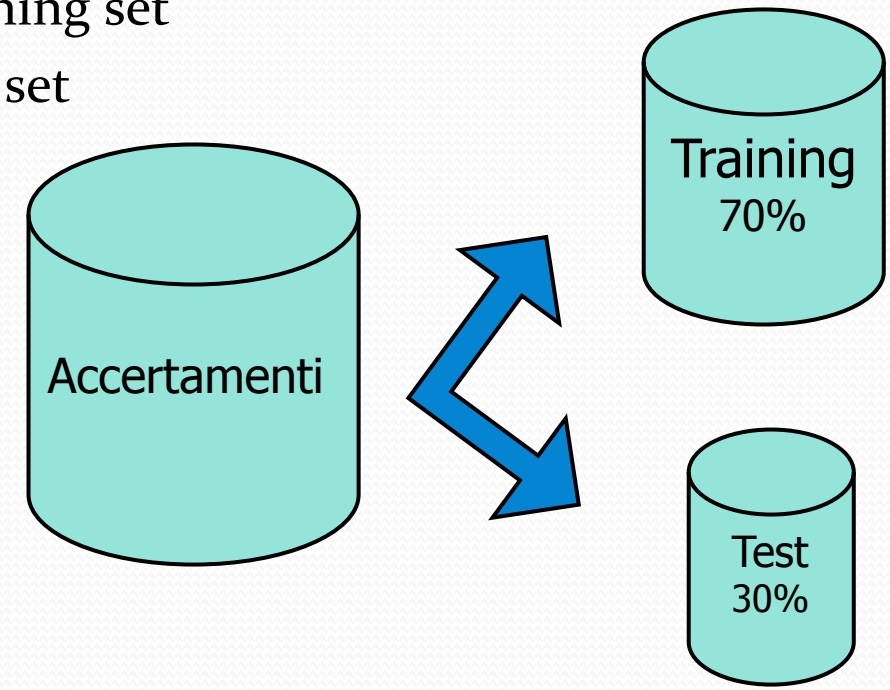
I frodatori che hanno un valore prossimo a 1 per entrambi gli obiettivi (tre nella realtà), vengono considerati più interessanti o a rischio e classificati con il valore 3 (rosso).

I frodatori che vengono classificati come 2 (arancioni) hanno o valori simili per entrambi gli obiettivi oppure uno alto e l'altro sotto la media per cui lo score che viene assegnato è più basso del precedente

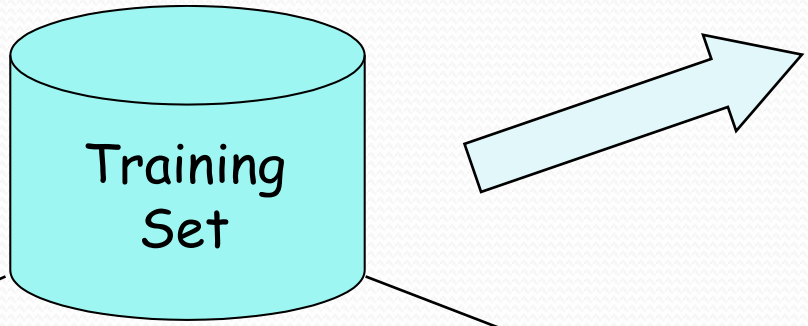
Si procede nello stesso modo per interpretare i contribuenti classificati 1 e 0.

Metodologia train & test

- I dati degli accertamenti sono stati divisi in due insiemi distinti
  - Training set
  - Test set

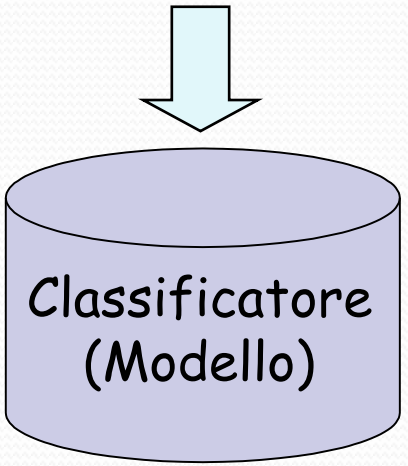


Training



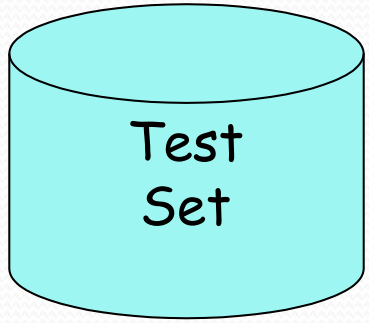
Anno	Credito	Vol.Aff	….	Classe
2004	€ 1.000	€ 10.000	….	2
2003	€ 20.000	€ 500.000	….	3
2000	€ 20.000	€ 20.000	….	3
2001	€ 2.000	€ 10.000	….	2
2001	€ 30.000	€ 50.000	….	3
2004	€ 5.000	€ 100.000	….	2
2005	€ 25.000	€ 50.000	….	3
2000	€ 1.000	€ 50.000	….	0

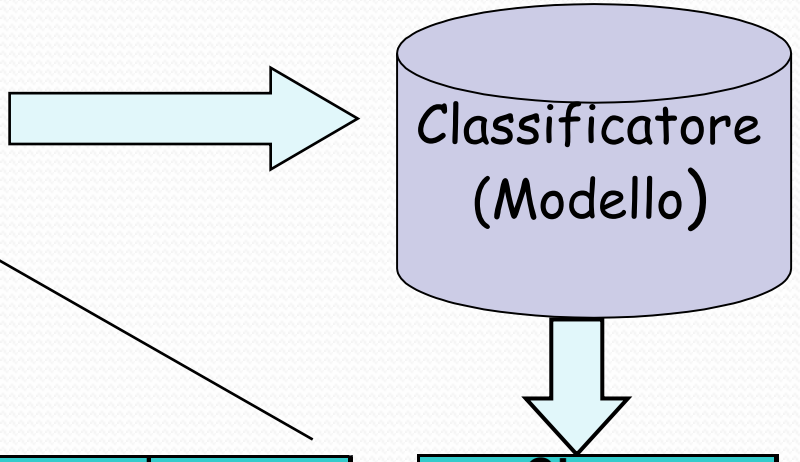
Algoritmi di Classificazione



IF Credito >= 20000
AND Vol.Aff <= 50000
THEN classe = 3

Test

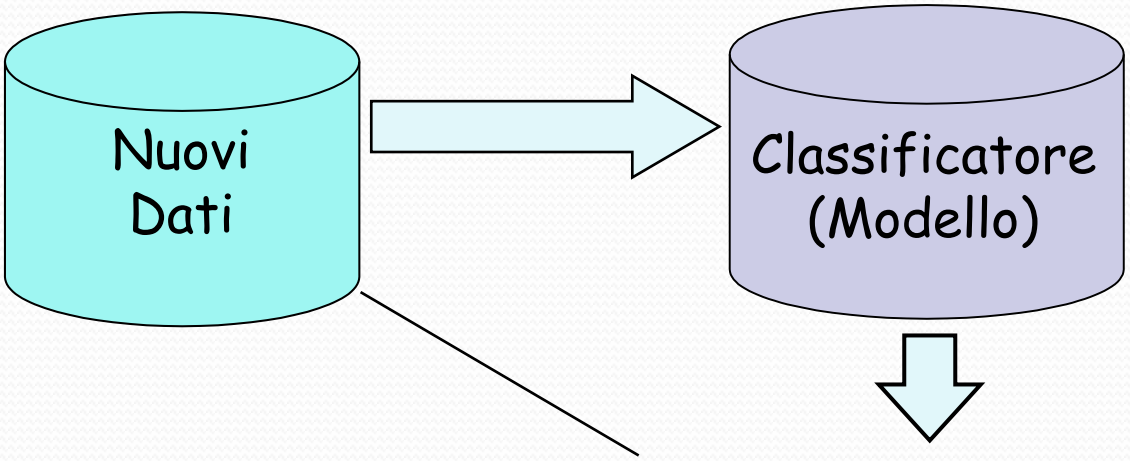




Anno	Credito	Vol.Aff	….	Classe
2004	€ 5,000	€ 15,000	….	2
2003	€ 50,000	€ 1,000,000	….	1
2000	€ 80,000	€ 500,000	….	3
2001	€ 5,000	€ 1,000	….	3
2001	€ 500	€ 100,000	….	0
2004	€ 2,000	€ 200,000	….	1
2005			….	…
2000			….	…

Classe
1
2
3
3
0
1

Previsione



Anno	Credito	Vol.Aff	….
2004	€ 5,000	€ 15,000	….
2003	€ 50,000	€ 1,000,000	….
2000	€ 80,000	€ 500,000	….
2001	€ 5,000	€ 1,000	….
2001	€ 500	€ 100,000	….
2004	€ 2,000	€ 200,000	….
2005			….
2000			….

Classe
1
2
3
3
0
1

Modello

- Un modello è costituito da un insieme di regole in base alle quali ad ogni contribuente viene assegnata una, e una sola, classe tra 0, 1, 2 e 3
- Ad ogni regola sono associati:
  - **supporto**: il numero di individui che vengono descritti dalla regola
  - **confidenza**: la percentuale di correttezza della regola
- Un modello può contenere decine (anche centinaia) di regole

Esempio di regola (1): descrizione

Campi della regola

IMP_VE_VOLAFF <= 1634
IMP_BENI_ALTRI <= 603
IMP_ECC_PREC <= 2
IMP_V_AGG_IMPON <= -105644
VAR_RIMBORSO <= 91929

Classe = 3

Confidenza = 93,93

Supporto = 494

Esempio di regola (1): interpretazione

La regola può essere interpretata nel seguente modo (devono valere contemporaneamente tutte le condizioni):

se

- Volume d'affari basso

e

- Totale acquisti e importazioni - altri (VA) < 603 €

e

- Eccedenza d'imposta (RN) risultante da precedente dichiarazione praticamente inesistente

e

- Valore aggiunto imponibile decisamente negativo (< -100.000 €)

e

- Rimborso IVA richiesto al di sotto dei 91.000 €

allora il contribuente è nella classe 3, con una precisione del 93,93% su un totale di 494 individui accertati per i quali valgono le suddette condizioni

Supporto sul totale della Popolazione delle dichiarazioni = 1.914

Esempio di regola (2)

VAL_INDICE_PCR <= 5,43	Indice di ricarico (vale 0 per PF e SP) inferiore al 5,5 %
IMP_REDD_IMP_ SMPL > -1	Reddito d'impresa semplificata positivo o assente
FLG_VOLAFF = '0,0'	Volume d'affari maggiore di 12.500 €
FLG_SOGG_CESS _DATA = '0,0'	Soggetto in attività alla fine dell'anno d'imposta
FLG_PIVA_BREVE _DUR = '1,0'	Soggetto con P.IVA di breve durata
COD_DIR_REG in ('901','908','913')	Residenza in una delle seguenti regioni: Piemonte, Friuli-Venezia Giulia, Lazio

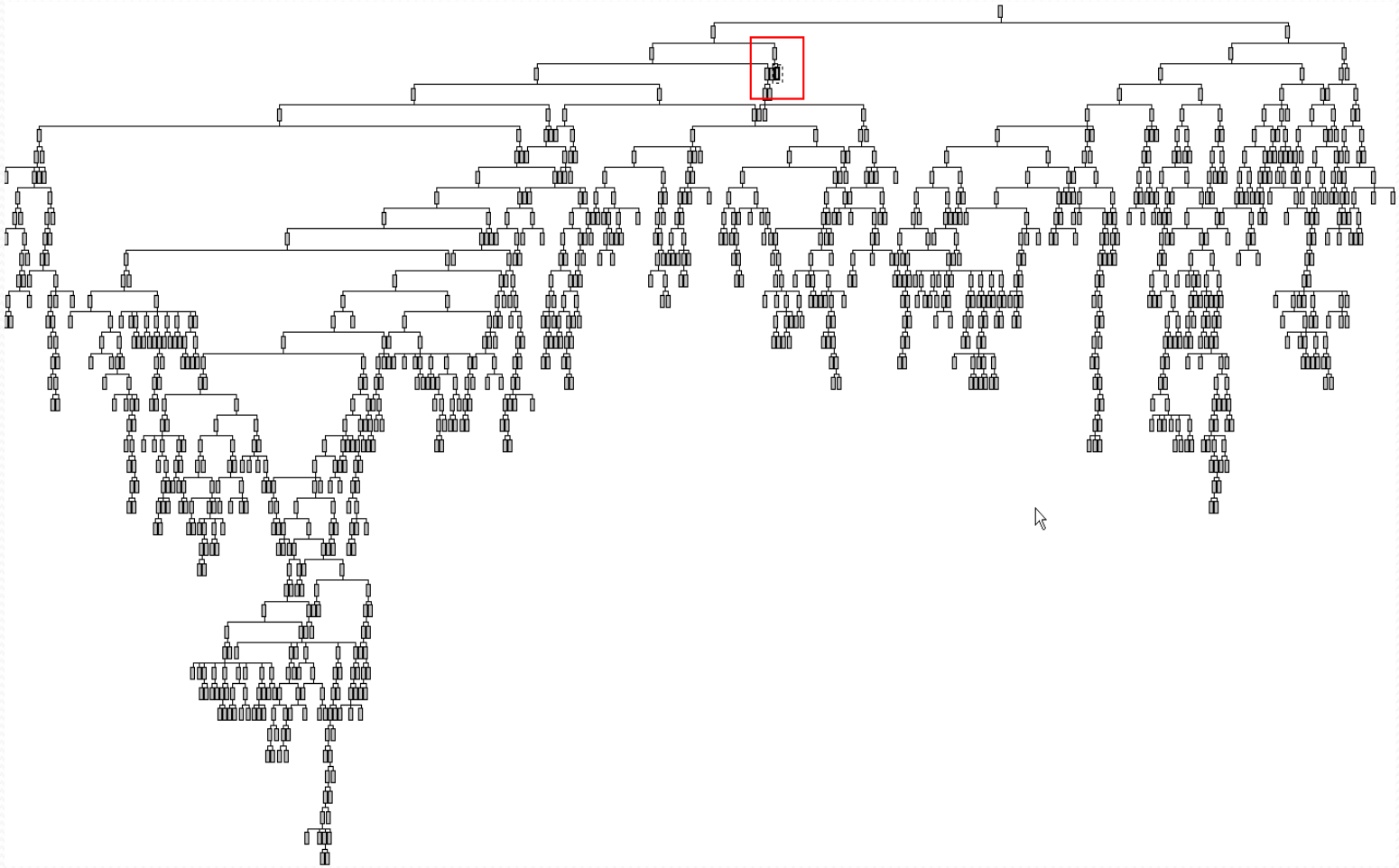
Classe = 3

Confidenza = 91,7 %

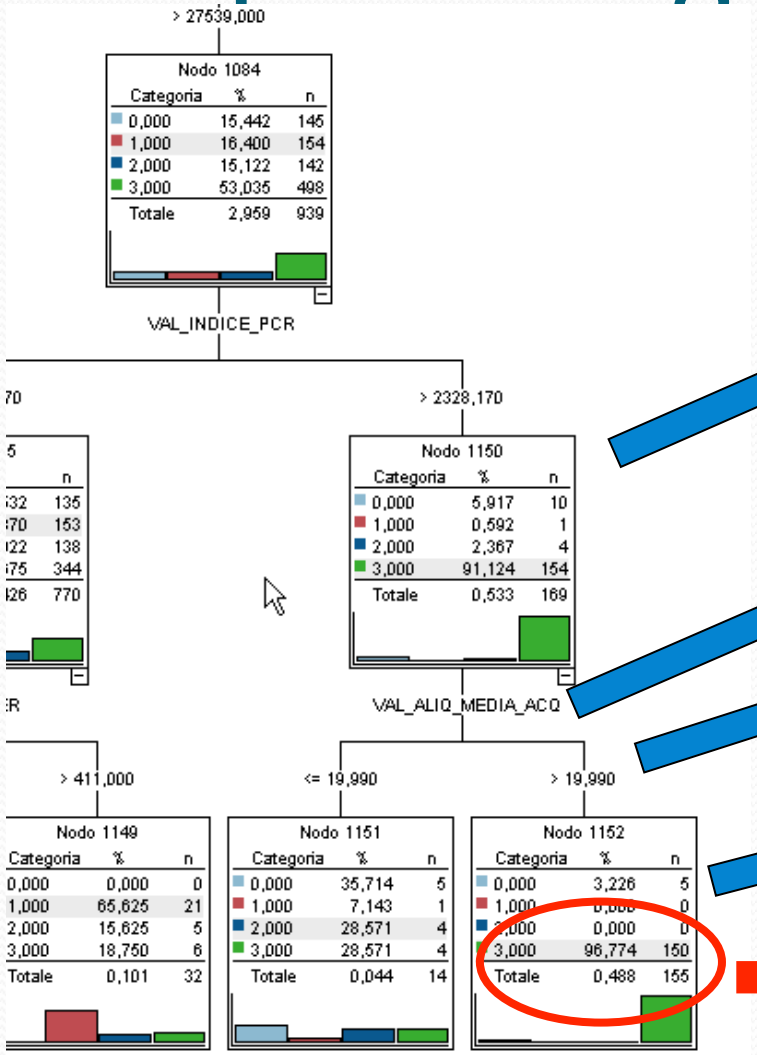
Supporto = 12

Supporto sul totale della popolazione di un anno = 499

Alcuni modelli possono essere rappresentati come un “albero” in cui i cammini dalla radice alle foglie rappresentano le regole generate



Esempio di regola estratta



NODO intermedio

Attributo su cui viene costruito il ramo

Valore della suddivisione

Foglia dell'albero

Classificazione: 3

Supporto: 155

Confidenza: 96,774%

Valutazione del modello: matrice di confusione

Istanze Classificate Correttamente	7.182	(52,48%)
Istanze Non Classificate Correttamente	6.504	(47,52%)
Totale	13.686	

		Valori Classificati			
		0	**1**	**2**	**3**
Valori Reali	**0**	**6408**	193	26	6
	1	2984	**558**	30	4
	2	1859	482	**112**	8
	3	723	91	98	**104**
Accuratezza		53.52%	42.15%	42.11%	85.25%

Valutazione del modello: valutazione economica

- Recupero del credito IVA (maggiore imposta):
  - Per tutti gli accertamenti nel test set (13.686):

    € 950.406.170
  - Per i soli accertamenti con score 3 nel test set (122):

    € 37.831.524
  - Quindi in proporzione si ha che a fronte del 0,89% degli accertamenti si recupera il 3,98% della frode quadruplicando l' efficacia degli accertamenti
  - La media della frode recuperata passa da 70K€ a 310K€ (440%)

Conclusioni

- I risultati ottenuti sono il frutto di applicazione di metodologie differenti e di molteplici esperimenti che convergono verso un modello che combina:
  - Una componente AND che privilegia la contemporanea presenza dei tre i criteri di interesse
  - Una componende OR che mette in evidenza la presenza di valori molto interessanti anche rispetto ad uno solo dei criteri

Risultati

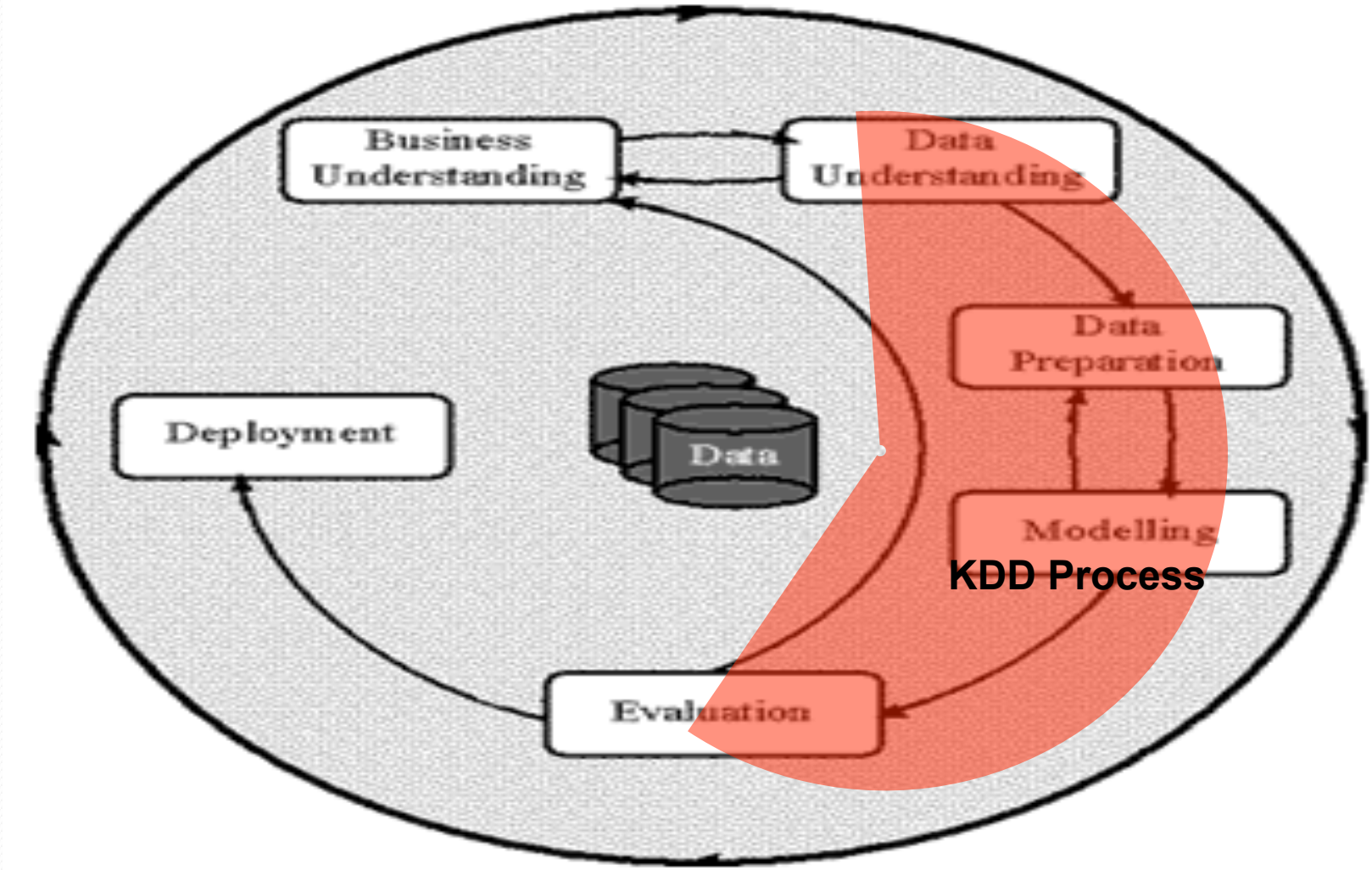
- Gli esperimenti sui dati delle verifiche fiscali mostrano **un'accuratezza del modello fino al 97%** nell'individuare le frodi di maggiore dimensione
- Possibilità di recupero frode maggiore con meno risorse

  (con il 2% degli accertamenti recupero 35% della frode, corrispondenti a € 1.724.810.890)
- E' stato deciso il passaggio alla fase della verifica sul campo estraendo un campione di dichiarazioni da accertare

Validazione

- La validazione di SNIPER verrà fatta su una platea di individui ancora da accertare selezionati con le seguenti caratteristiche:
  - Equa distribuzione sul territorio (al massimo 4 contribuenti per ufficio)
  - Utilizzo di tutte le regole (è scopo del committente scoprire più comportamenti fraudolenti possibili)

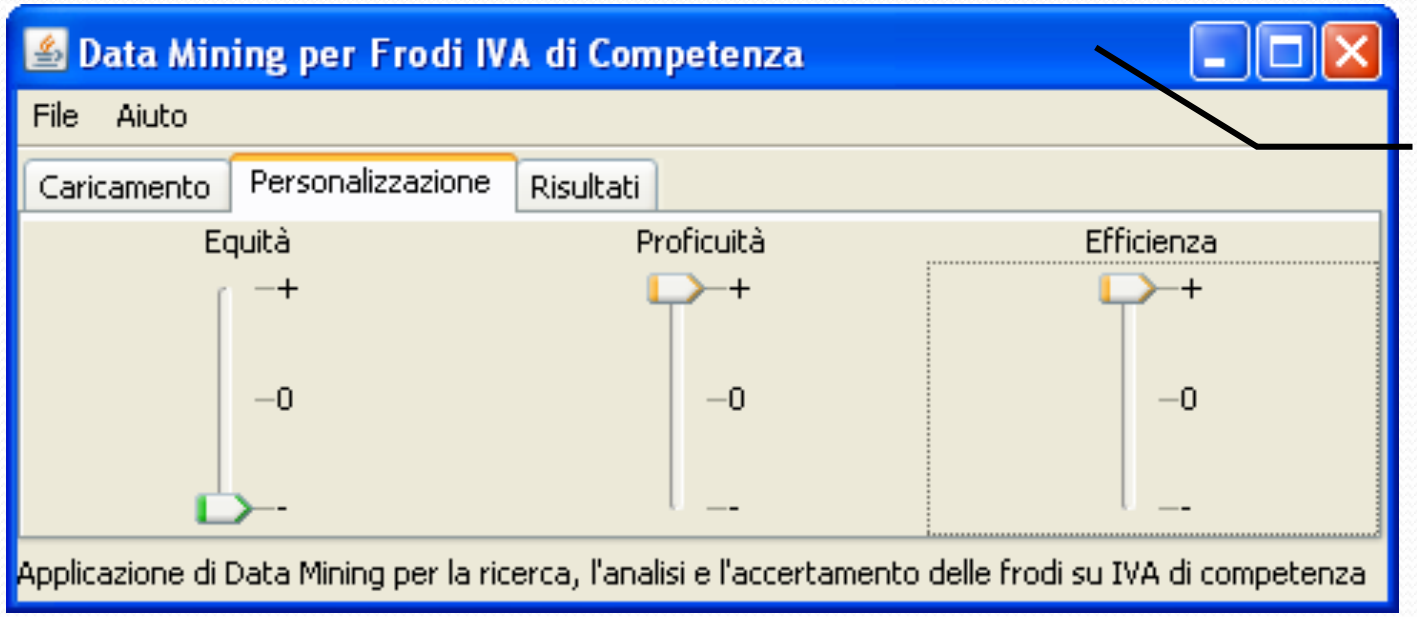
CRISP-DM: The life cicle of a data mining project



Quale deployment?

- I modelli sviluppati mantengono i tre obiettivi di efficienza, proficuità ed equità bilanciati tra loro.
- Che impatto hanno i singoli obiettivi sugli altri?
- Un possibile modo di rispondere prevede di poter dare pesi differenti alle tre variabili e rendere il modello ‘focalizzato’ su uno oppure due dei tre obiettivi.

Sviluppi conclusivi del progetto



Cruscotto dell'applicazione

- 'Focalizzare' il modello può prevedere diverse combinazioni possibili
  - Valutiamo insieme le combinazioni significative rispetto agli obiettivi operativi
- Dietro ogni scelta c'è un modello differente che è il migliore rispetto a quella combinazione