



# Rilevamento di frodi fiscali e pianificazione degli accertamenti

Sorgente: Ministero delle Finanze  
**Progetto Sogei, KDD Lab. Pisa**

# Lotta all'evasione – Min. Finanze/SOGEI ('98-'99)

- ◆ **Pianificazione di accertamenti fiscali**
- ◆ **Obiettivo:** costruire un modello predittivo che individui una porzione di contribuenti su cui risulti vantaggioso effettuare un controllo fiscale.
  - Estrazione di **alberi di decisione**
- ◆ **Dataset:**
  - dati storici provenienti da fonti diverse (mod. 760, mod. 770, INPS, ENEL, SIP, Camere del Commercio)
  - dati storici sui risultati degli accertamenti pregressi.
- ◆ **Variabile da predire:** imposta recuperata al netto delle spese di accertamento.
- ◆ **Valutazione dei modelli estratti rispetto ad **indici** generali (accuratezza) e specifici di dominio (redditività)**

# Rilevamento di frodi

## ◆ Obiettivo generale:

- Determinare *modelli* per la previsione del comportamento fraudolento per:
- **Prevenire frodi future** (rilevamento di frodi *on-line*)
- **Scoprire frodi passate** (rilevamento frodi *a posteriori*)

## ◆ Obiettivo specifico:

- **Analizzare i dati storici sulle verifiche per pianificare verifiche future più EFFICACI**

# Pianificazione di verifiche

◆ C'è un trade-off tra:

- *Massimizzare i benefici della verifica:*  
selezionare quei contribuenti che massimizzano il recupero di tasse evase.
- *Minimizzare il costo della verifica :*  
selezionare quei contribuenti che minimizzano le risorse necessarie alla verifica.

# Available data sources

- ◆ Dataset: **Dichiarazioni dei redditi**, su una classe selezionata di **aziende** italiane integrate con altre sorgenti:
- ◆ Contributi INPS per dipendenti, consumi ENEL e telefonici..
- ◆ Dimensione: **80 K** tuple, 175 numerici attribute.
- ◆ Un sottoinsieme di **4 K** tuples corrisponde ad aziende **verificate**:
  - I risultati delle verifiche sono memorizzati nell'attributo: **recovery** (= **amount of evaded tax ascertained**)

# Data preparation

originale  
dataset  
81 K

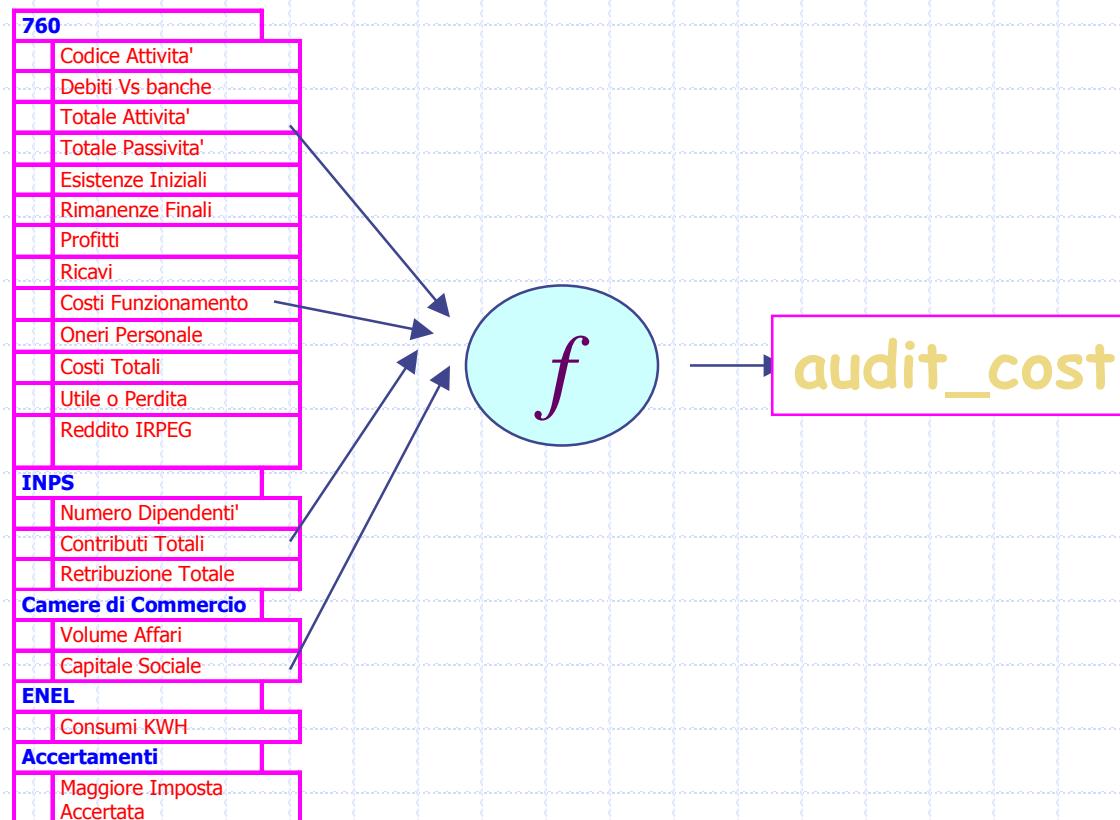
data consolidation  
data cleaning  
attribute selection

Risultati  
varifiche  
4 K

TAX DECLARATION	
	Codice Attivita'
	Debiti Vs banche
	Totale Attivita'
	Totale Passivita'
	Esistenze Iniziali
	Rimanenze Finali
	Profitti
	Ricavi
	Costi Funzionamento
	Oneri Personale
	Costi Totali
	Utile o Perdita
	Reddito IRPEG
SOCIAL BENEFITS	
	Numero Dipendenti'
	Contributi Totali
	Retribuzione Totale
OFFICIAL BUDGET	
	Volume Affari
	Capitale Sociale
ELECTRICITY BILLS	
	Consumi KWH
AUDIT	
	Recovery

# Modello di costo

◆ si definisce l'indicatore **audit\_cost** come funzione di altri attributi



# Modello dei costi e variabile target

- ◆ Recupero di una verifica

- $actual\_recovery = recovery - audit\_cost$

- ◆ La variabile target (class label) della nostra analisi: **Class of Actual Recovery (c.a.r.)**:

- ◆  $c.a.r. = \begin{matrix} negative & \text{if } actual\_recovery \leq 0 \\ positive & \text{if } actual\_recovery > 0. \end{matrix}$



# Indicatori di qualità

- ◆ Si costruiscono vari classificatori che sono valutati secondo diverse metriche:
- ◆ **Domain-independent** indicators
  - confusion matrix
  - misclassification rate
- ◆ **Domain-dependent** indicators
  - audit #
  - actual recovery
  - profitability
  - relevance

# Indicatori Domain-dependent

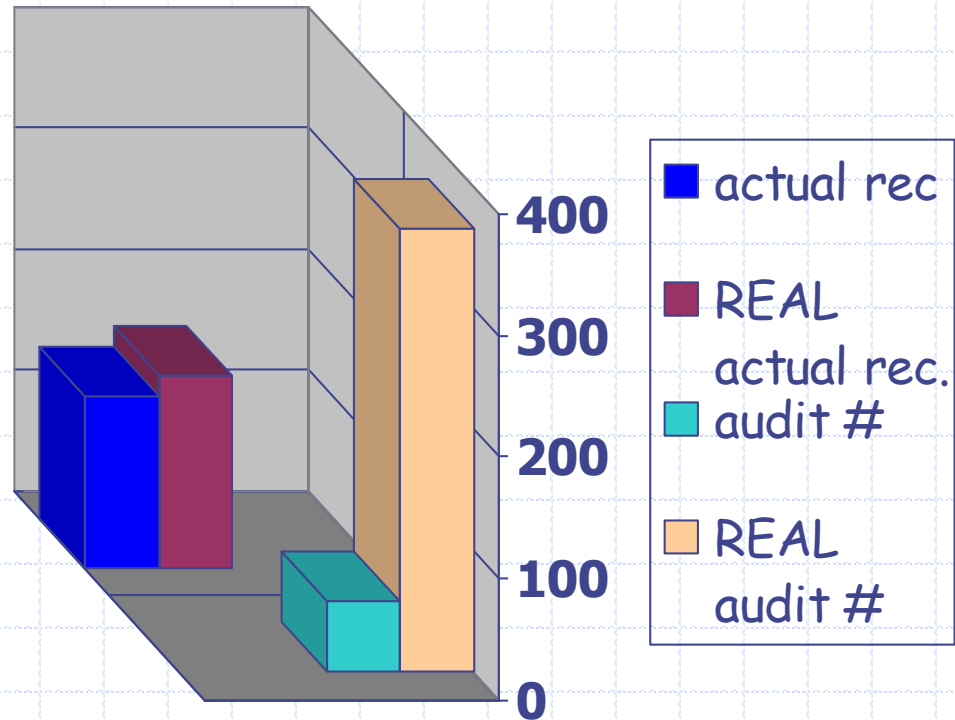
- ◆ **audit #** (di un dato classificatore): numero di tuple classificate come positive =  
 $\# (FP \cup TP)$
- ◆ **actual recovery**: ammontare totale del recupero effettivo per tutte le tuple classificate come positive
- ◆ **profitability**: recupero effettivo medio per verifica
- ◆ **relevance**: rapporto tra **profitability** e l'errore di classificazione

# Il caso REAL

- ◆ I Classificatori sono confrontati con l'intero test-set, cioè gli accertamenti veramente condotti.
- ◆ audit # (REAL) = 366
- ◆ actual recovery(REAL) = 159.6 M euro

# Classificatore 1 (min FP)

- *misc. rate* = 22%
- *audit #* = 59 (11 FP)
- *actual rec.* = 141.7 Meuro
- *profitability* = 2.401



# Classificatore 2 (min FN)

- *misc. rate* = 34%
- *audit #* = 188 (98 FP)
- *actual rec.* = 165.2 Meuro
- *profitability* = 0.878

