Università di Pisa	A.A. 2012-2013
Data Mining II	

Project assignments / Part 2

“Fiscal fraud detection”

General information

Objective of this project is the definition of interesting classes of fraudolent users in a fiscal auditing context, and to extract predictive models to identify them. The rules for this project are the same applied in the first part:

1. the project can be performed by single students or groups up to 3 persons each;
2. each group should perform the analyses indicated in the text, trying to answer to each request. Any spontaneous addition to that is welcome yet optional, and cannot replace the original TODO list;
3. each group should summarize the work done in a short report (indicatively 5-15 pages), loosely following the guidelines of the CRISP model;
4. each group is totally free to choose the tools and software it prefers (although, in this case the MAtlas software appears to be the simplest choice);
5. any question, suggestion or request related to the project can be addressed to Mirco Nanni (mirco.nanni@isti.cnr.it).

The dataset

The dataset is composed of tax declarations, some of them audited (i.e., we know the real values of the declaration, opposed to what the user presented), divided in the following files:

- **AuditFiscali**: audits, covering years from 2000 to 2005, for 45K users
- **DichiarazioneIva2000**: ~27K declarations for year 2000.
- **DichiarazioneIva2001**: ~27K declarations for year 2001.
- **DichiarazioneIva2002**: ~30K declarations for year 2002.
- **DichiarazioneIva2003**: ~31K declarations for year 2003.
- **DichiarazioneIva2004**: ~29K declarations for year 2004.
- **DichiarazioneIva2005**: ~27K declarations for year 2005.

A detailed description of the attribute values is provided in file “MetaDati.txt”.

Objectives

1. Target variables: starting from table “AuditFiscali”, define (and compute) 3 indicators to measure the level of fraud discovered in each audited declaration, following the three criteria listed below, deemed interesting by the domain experts:
   a) *Profiquity* : measure the volume of the fraud;
   b) *Equity* : measure the volume of the fraud relative to the business volume of the activity described in the declaration;
   c) *Efficiency* : measure the volume of the fraud relative to the credit claimed in the declaration.

   Study the distribution of these indicators.
2. Predictive variables: starting from all tables “DichiarazioneIva*”, define (and compute) several variables that you consider potentially useful for discriminating the users w.r.t. the fraud detection problem. In doing this, consider the fact that some users appear in the declarations of several different years, therefore it is needed to compute indicators that take their history into account.
3. Fraud classifier: use the indicators defined above as predictive variable in a classification task where the target variable Fraud/noFraud is defined based on the three indicators computed at the first step. Study the performances of the models obtained by changing the parameters of the construction algorithm. As optional step, more than one target variable can be defined and tested.
4. Unsupervised classification of users: use the predictive variables defined above to group users into homogeneous clusters. Describe the clusters obtained and test whether there are clusters where the percentage of frauds are particularly high or particularly low.
5. Evaluate the privacy issues that might arise in this application, and discuss possible counter-measures to adopt in order to remove or limit them.