

Università di Pisa	A.A. 2014-2015
<b>Data Mining II</b>	

## Project assignment

### *Customer segmentation based on user profiling*

### General information

Objective of this project is to perform a few analyses on a dataset of transactions involving the customers of a supermarket chain. The general guidelines for this assignment are the following:

1. the project can be performed by single students or groups up to 3 persons each;
2. each group should perform the processing and 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;
5. any question, suggestion or request related to the project can be addressed to Mirco Nanni ([mirco.nanni@isti.cnr.it](mailto:mirco.nanni@isti.cnr.it)) and Anna Monreale ([anna.monreale@unipi.it](mailto:anna.monreale@unipi.it)).

### The dataset

The project will be based on real data describing customers and transactions of a set of department stores. The data cover the purchases performed over 12 months, and include the details of each product sold in each transaction, together with the ID of the customer who performed the transaction (where available). The dataset consists of the following tables, provided as CSV files:

<b>articolo.csv</b>	textual description of the products (in Italian)
<b>cliente.csv</b>	basic information about customers (in Italian)
<b>data.csv</b>	translation table for date coding
<b>marketing.csv</b>	marketing hierarchy of products (in Italian)
<b>venduto.csv</b>	transactions, a line for each product sold

## Objectives

The following activities should be performed and reported:

- 1. Exploration:** a **short** data exploration phase, aimed at understanding what data can be useful and whether they present any issues or anomalies.
- 2. User Purchase Profiles:** for each customer, define and compute a *user purchase profile*, which should describe the set of products (s)he buys systematically, as well as **where** (which store) and **when** the customer systematically visits for purchasing those products. For instance, user 7999 might have a purchase profile of the form  $\{ (milk, store\_23, Monday), (bread, store\_23, Monday), (fish, store\_30, Friday) \}$ . The detail level of products (*milk, bread, ...*) and time slots (*Monday, ...*) should be chosen properly, exploiting the marketing and the temporal hierarchies.
- 3. Store Analysis:** select 2-3 significant stores, and understand which are the typical *user purchase profiles* that occur there. Suggested approach: find two or more features to represent each user profile, then perform a customer segmentation based on these features. Try to give an interpretation of the clusters found and compare the results obtained on the different stores.