First Classwork
data mining introduction
18/11/2025

DATASETS

The dataset is downloadable from the following google drive link:
https://drive.google.com/file/d/1p90YIpJ1NRO6pMlLeY92uEsR2o_xP9zX/view?usp=drive_link

The dataset contains products purchased in a supermarket about fidelity people.

Column name	Description
db_id	Database record id
scontrino_id	The id of the receipt
puntovendita_id	The supermarket id
data	The date of the receipt
cassa	The supermarket checkout id
cassiere	The employee id
nummero_scontrino	The receipt number
ora	The hour of the purchased
tessera	The supermarket card id of the user
num_riga	The row number
r_qta_pezzi	The number of a purchased product
r_peso	The weight of the product
r_importo_lordo	Gross price
r_iva	Tax in %
r_sconto	Discount
r_tipo_riga	The row type
cod_prod	The code of the product
descr_prod	The product description
cat_mer	Product category
liv1	First level category id
descr_liv1	First level category description
liv2	Second level category id
descr_liv2	Second level category description

liv3	Third level category id
descr_liv3	Third level category description
liv4	Fourth level category id
descr_liv4	Fourth level category description
tipologia	Typology
descr_tipologia	Descriptio of the typology
cod_rep	Position product id
descr_rep	Position product description

TASKS

1. The dataset is structured across four merchandising levels, all organized under a single umbrella hierarchy. The first level represents the macro-categories, while the subsequent levels correspond to increasingly specific sub-categories. The leaves of this hierarchy represent individual products. The goal of the first task is to compute, for each merchandising level, the frequency of every element. Then, for each level, you must create two bar plots: 1) one showing the five most frequent elements, and 2)another showing the five least frequent elements. Remember to exclude shoppers from the analysis.
2. The second task is similar to the first one, but requires stratifying the dataset. The first stratification divides the dataset into **three time periods** based on months:
   - **Range 1:** January to Mid-May
   - **Range 2:** Mid-May to September
   - **Range 3:** October to December

   For each of these ranges, you must create the same plots described in Task 1 for every merchandising level. The second stratification is based on **time slots**:
   - **Slot 1:** 08:30–12:30
   - **Slot 2:** 12:30–16:30
   - **Slot 3:** 16:30–20:30

   For each time slot and each merchandising level, you again need to generate the same plots as in Task 1.
3. Apply the APRIORI algorithm to create the association rules at level four.
4. Apply the FP-Growth algoritm to create the association rules at level four.
5. Consider only the rows where the tessera (card) field is not empty. From this filtered dataset, create a new table in which each row corresponds to a card and each column corresponds to a product. For any element (i,j) in this table, the

value should be:

- **0** if card *i* has never purchased product *j*,
- otherwise, the **frequency** with which card *i* has purchased product *j*.

After constructing this table, apply **PCA** to reduce its dimensionality, and then apply a **clustering technique** to group the cards based on their purchasing behavior.

NOTE

You have to do an object-oriented implementation. Therefore create a work directory, then put a README.md file in order to write the explanation of the implemented project, and finali creare a src directory (inside the work dir) containing the code. You need to create an open repository on github called IDM 2025, then you will create a sub-directory called first_classwork, and store the code inside such directory. You need to explain the results into a doc.

DEADLINE

18/12/2025
Send an email to antonio.dimaria1@unict.it when you complete the project, with the link to the repository