Data Mining Techniques

Spring Semester 2022-2023

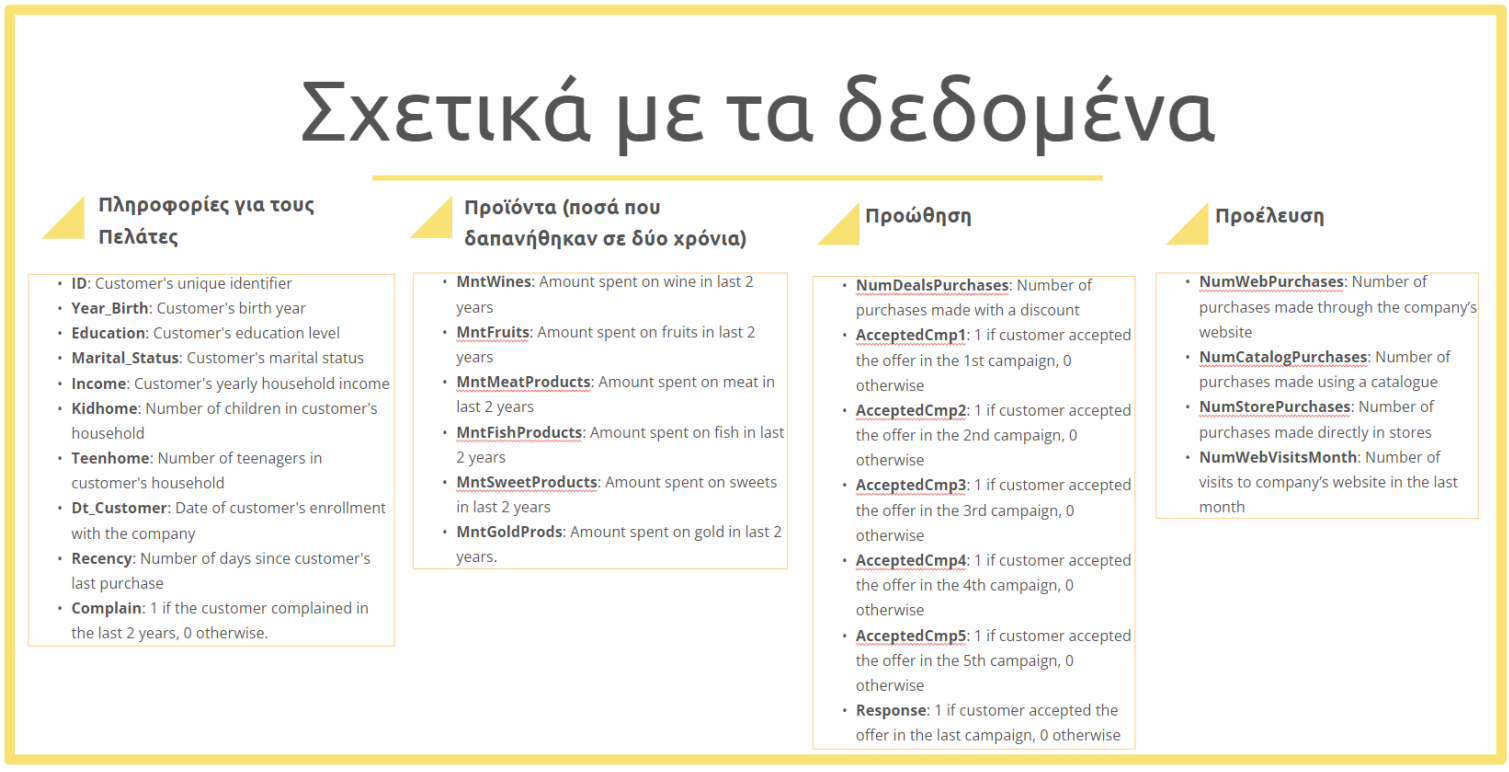
1st Exercise-Individual or group work (2 people)

The objective of the task

The purpose of the assignment is to familiarize you with the basic steps of the process followed for the application of data mining techniques, namely: pre-processing / cleaning, transformation, application of data mining techniques and evaluation. The implementation will be done in the Python programming language using the tools/libraries: jupyter notebook, pandas and SciKit Learn.

Description

THE**Customer Personality Analysis**is a detailed analysis of a company's ideal customers. It helps a business better understand its customers and makes it easier to modify products according to the specific needs, behaviors and concerns of different types of customers. For example, instead of spending money to promote a new product to every customer in the business' database , a business can analyze which customer category is most likely to buy the product and then promote the product only to that particular category.



Wanted

1. **Pretreatment/Cleaning**: Check for missing values in the data and handle them accordingly, convert date columns to DateTime objects and check for any dtype: object attributes that you can encode/convert to numeric values **(5%)**.

2. Print them out **unique prices** in the categorical characteristics **Marital_Status** and **Education** to get a clearer picture of the data. Change the values [Alone,Absurd,YOLO] of Marital_Status with value 'Single'. Use any type of graph you like to show the number of values in each category. **(5%)**

3. **Create new features**:**(10%)**

   A. Create an attribute ("Customer_For") that represents the number of days customers started shopping at the store relative to the last recorded date (Recency).

   B. Extracting the age "**Age**" of a customer based on "**Year_Birth**" indicating the respective person's year of birth.
   C. Create another attribute "**Spent**" indicating the total amount spent by the customer across all categories over a two-year period.
   D. Create a feature "**Children**" to indicate all children in a household, i.e. children and teenagers.
   E. To further clarify the household, create an attribute labeled "**Family_Size**" which shows the total number of people in a household

   St. Create a feature "**Is_Parent**" indicating whether a client is also a parent

   G. Create another feature "**Living_With**" using it "**Marital_Status**" to extract the living status of couples. Specifically this attribute must have two values, "Partner" and "Alone".
   H. Create the “Age Group” column using the “Age” column, which groups the ages into the following values "21-30", "31-40", "41-50", "51-60", " 61-70", "71-80", ">80".

4. Check if there are any **outliers** in the attributes and delete them from the data. **(5%)**

5. Then examine her **correlation between features** with a heatmap diagram. (Excluding categorical features at this point) **(5%)**

6. Questions to be answered with **graphs**-select 10 from the following. **(20%)**
    1. In which Marital_Status category does the largest percentage of the company's customers belong?
    2. How many customers have complained?
    3. Relationship between the number of purchases **Spent** and marital status.
    4. The relationship between the number of purchases **Spent** and of the number of children and of family size.
    5. What does age have to do with it? **Age Group** with the feature **Spent** of markets?
    6. What does income have to do with it? **Income** with the feature **Spent** of markets?
    7. What is the relationship between education and income?
    8. What is the relationship between income and family size?
    9. What is the relationship between income and number of children?
    10. What is the relationship between income and **Living_With**;
    11. What is the relationship between income and number **Spent** of markets?
    12. What is the relationship between the number of purchases from the website and the number of visits to the website?
    13. What is the percentage of customers who accept all offers from the store?

    14. Draw the histogram for the NumDealsPurchases column.
    15. Draw the histogram for the Income column.
    16. Draw the histogram for the Kidhome column.
    17. Draw the histogram for the Family_Size column.
    18. Do customers with graduate degrees spend more money on wine?

7. **Principal component analysis (PCA) (25%)**: In this problem, there are many factors on which a classification is made. These factors are key characteristics or traits. The greater the number of features, the more difficult the task. Many of these features are correlated and therefore redundant. This is why you will perform **reduction of dimension** in the selected features. Dimension reduction is the process of reduction of the number of random variables under consideration, and results in obtaining a set of main variables.

The variables in the dataset that are categorical rather than numeric, after the attribute additions made in the previous queries are the following ['Education','Marital_Status','Living_With']. For these variables you will use the **LabelEncoder()** to be converted into numeric data (the process is called **one hot encoding**).

Then create a copy of the dataframe that will contain all the numeric columns and delete the columns related to offers and promotions, i.e. the [*'AcceptedCmp3', 'AcceptedCmp4', 'AcceptedCmp5', 'AcceptedCmp1', 'AcceptedCmp2', 'Complain', 'Response'*].

Thus, the resulting data contain features of various dimensions and variances. Different variations of data characteristics adversely affect the modeling of a data set. The solution is to do what is called*Standardization*so that each column/attribute/variable has μ =0 and σ = 1.

Finally use the Principal Component Analysis (PCA) compression method to reduce the dimensions to n_components=3. Draw the (3D) view of the result.

8.**Implementation of Clustering (Clustering) (25%)**

Steps

- ELBOW method to determine the number of clusters to form
- Agglomerative and K-Means clustering
- Show the formed clusters through a diagram (eg scatter plot).

9. Profile of customers (bonus)

Try to sketch the profile of the clusters that are formed through diagrams in order to reach a conclusion about who is the "important" customer and who needs more attention from the store's marketing team.

To achieve this design some of the features which are indicative of the customer's personal characteristics in light of the cluster they are in (eg Age, Is_Parent, Family_Size etc.). Finally, for each of the clusters gather its main characteristics. E.g

Cluster 0: They spend the least They have the least income They have teenagers at home They are older	**Cluster 1:** They spend more They have more income. Most are not parents Actively participated in all 6 promotions

Deliverable:

The work can be done individually or in groups**2 persons**.

You will upload to eclass a folder of the format sdixxxx. (where sdi is the ID of one of the people in the group). The sdixxxx folder of the job is the folder in which you will ONLY have your code as described below (i.e.you will not resubmit the training/test data).

Your deliverable must contain**MANDATORY**one**Python notebook**with which it will can someone run your task step by step. In the notebook you can enter wherever you deem necessary**visualizations**the way we explained in the tutorials so that you can present your results in a nice way.**Thenotebook is also the complete reference**for your work (you will not deliver anything todoc, pdf) , design it carefully, remember to write a description at each step of what question is answered in each cell.