UNIVERSITÀ DI PISA

Data Mining Project on the customer_supermarket dataset

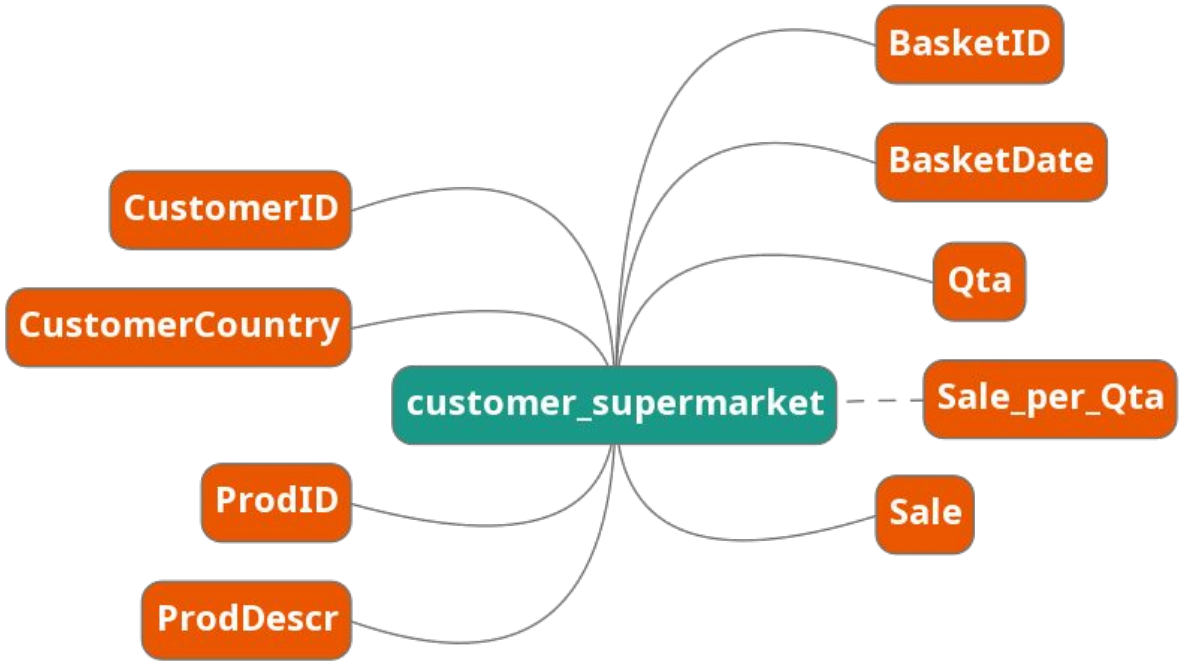
Data mining Course A.Y. 2020/2021
Diletta Goglia, Marco Petix

Task 1 - Data Understanding and Preparation

- **Analysing the customer_supermarket dataset:**
  - Data semantics and statistics
  - Assessing and improving Data Quality
  - Data Visualization

- **Representing the behaviour of the customers:**
  - Data semantics and statistics of the customers dataset
  - Assessing and improving Data Quality
  - Data Visualization

The customer_supermarket dataset



Data Semantics of the customer_supermarket dataset

Name (type)	Description
BasketID (string)	Identifies a shopping session initiated by a customer
BasketDate (datetime)	Identifies the date and time when the shopping session is conducted
CustomerID (string)	Identifies the customer taking part in a particular transaction
CustomerCountry (string)	Presumably it identifies the country from which the customer is conducting a transaction
ProdID (string)	Identifies the products purchased during a transaction
ProdDescr (string)	Contains a brief description of the products purchased during a transaction
Sale (float)	Represents the amount spent by the customer to purchase a single unit of the products bought during a transaction
Qta (integer)	Represents the amount of unity of a particular product purchased during a transaction

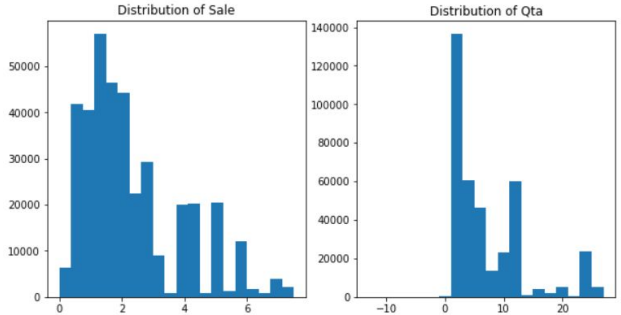
Assessing and Improving data quality

- **Duplicates** (- 5.232 rows)
- **Missing values** (65.073 missing CustomerIDs)
  - Segmentation by CustomerCountry and replacement with most frequent IDs
- **Canceled purchases** (- 16.265 rows)
  - Ad-hoc procedure to identify and drop the “purchase-refund” couples
- **Outliers** (- 68.848 rows)
  - Boxplots and Interquartile range

From 471.910 to 381.565 entries within the dataset, a 20% reduction

Assessing and Improving data quality

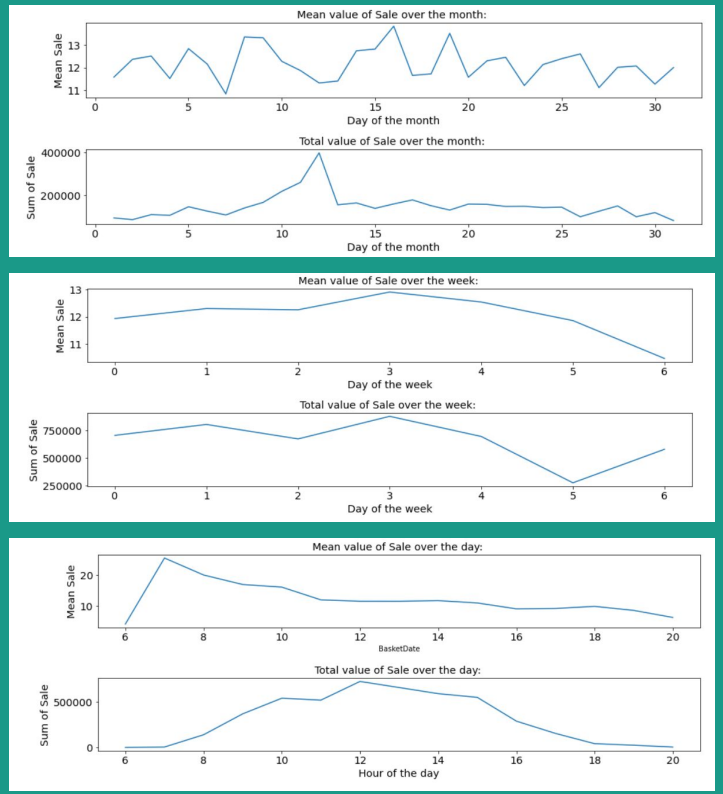
- **Duplicates** (- 5.232 rows)
- **Missing values** (65.073 missing CustomerIDs)
  - Segmentation by CustomerCountry and replacement with most frequent IDs
- **Canceled purchases** (- 16.265 rows)
  - Ad-hoc procedure to identify and drop the “purchase-refund” couples
- **Outliers** (- 68.848 rows)
  - Boxplots and Interquartile range



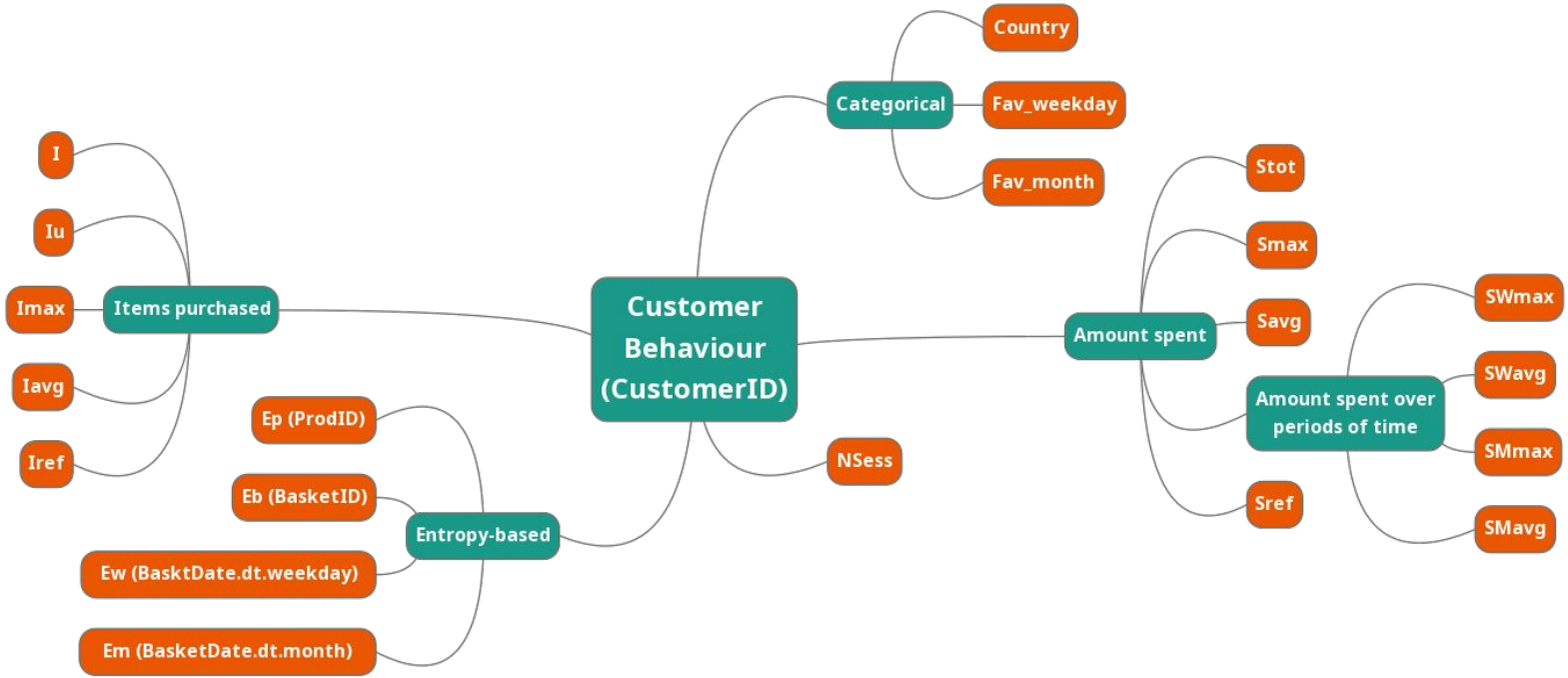
Data Visualization: Periodical Plots

Analysing the change in the **total and mean amount of sale** with respect to :

- the day of the month
- the day of the week
- the hour of the day



Data Semantics of the customer dataset



Data Semantics of the customer dataset

Name (type)	Description
I (integer)	The total number of items purchased by the customer
Iu (integer)	The number of distinct items purchased by the customer
Imax (integer)	The maximum number of items purchased by the customer within a single shopping session
Iavg (float)	The average number of items purchased by the customer within a single shopping session
Ir (integer)	The total number of items refunded to the customer

Data Semantics of the customer dataset

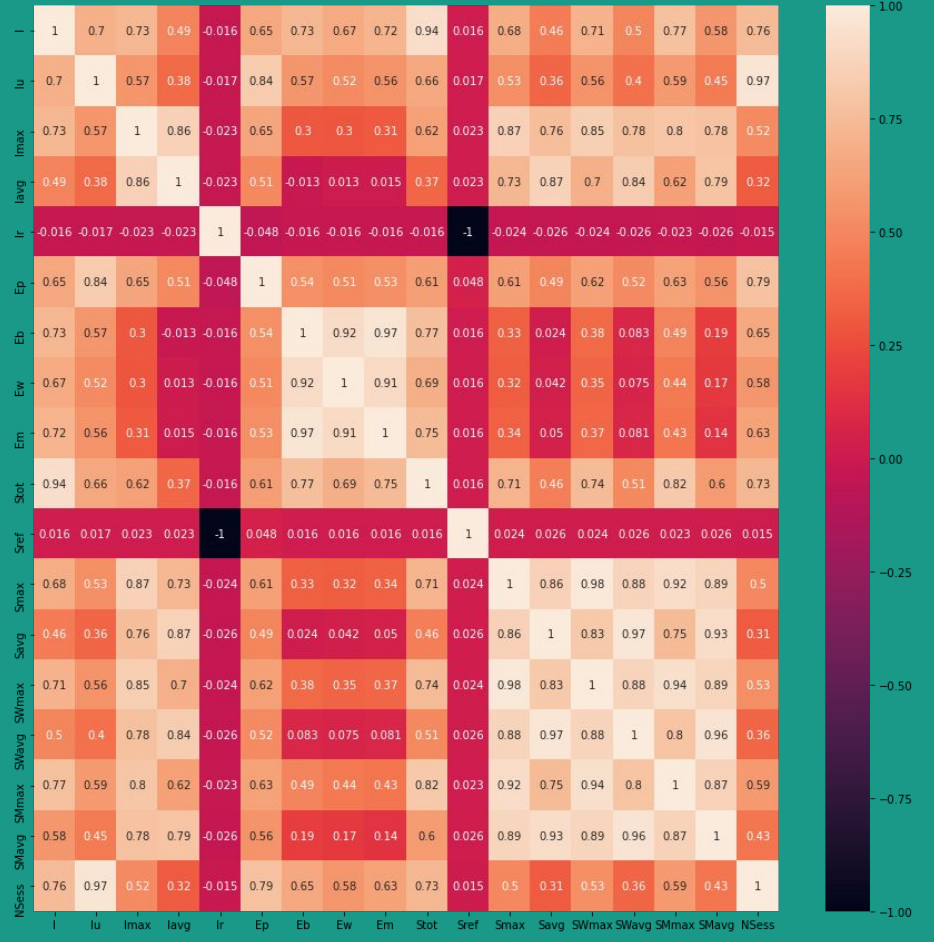
Name (type)	Description
Stot (float)	The total amount spent by the customer
Sref (float)	The total amount refunded to the customer
Smax (float)	The maximum amount spent by each customer within a single shopping session
Savg (float)	The average amount spent by each customer within a single shopping session
SWmax (float)	The maximum amount spent by each customer within a week
SWavg (float)	The average amount spent by each customer within a week
SMmax (float)	The maximum amount spent by each customer within a month
SMavg (float)	The average amount spent by each customer within a month

Data Semantics of the customer dataset

Name (type)	Description
Ep (float)	The Shannon's Entropy on the types of products purchased by the customer (ProdID)
Eb (float)	The Shannon's Entropy on the frequency and extent of the customer's shopping sessions (BasketID)
Ew (float)	The Shannon's Entropy on the weekday of the customer's purchases (BasketDate.dt.day_name)
Em (float)	The Shannon's Entropy on the month of the customer's purchases (BasketDate.dt.month_name)
NSess (integer)	The total number of shopping sessions initiated by the customer
Country (string)	The country associated with the majority of the customer's transactions
Fav_weekday (string)	The day of the week during which the customer tends to spend the most
Fav_month (string)	The month during which the customer tends to spend the most

Correlations

- High correlations between **I** and **Stot**, and **Imax** and **Smax**, refer to the increase in cost being obviously proportional to the number of products bought.
- Correlation between **Iu** and **Ep** is expected due to both attributes being an indicator of variety (and disorder) within the customer's choice of products .
- Very high correlation between **Iu** and **NSess**, as the one between **Ep** and **NSess**, highlights a type of non-routine spending, new sessions often lead to the purchase of new types of products.
- High correlations between the attributes **Eb**, **Ew** and **Em** is expected due to their very similar nature.
- **Smax**, **SWmax** and **SMmax**, just like **Savg**, **SWavg** and **SMavg**, all aim to measure the same kind of information while considering a specific frame of the whole period of observation, their very high correlation is not surprising.

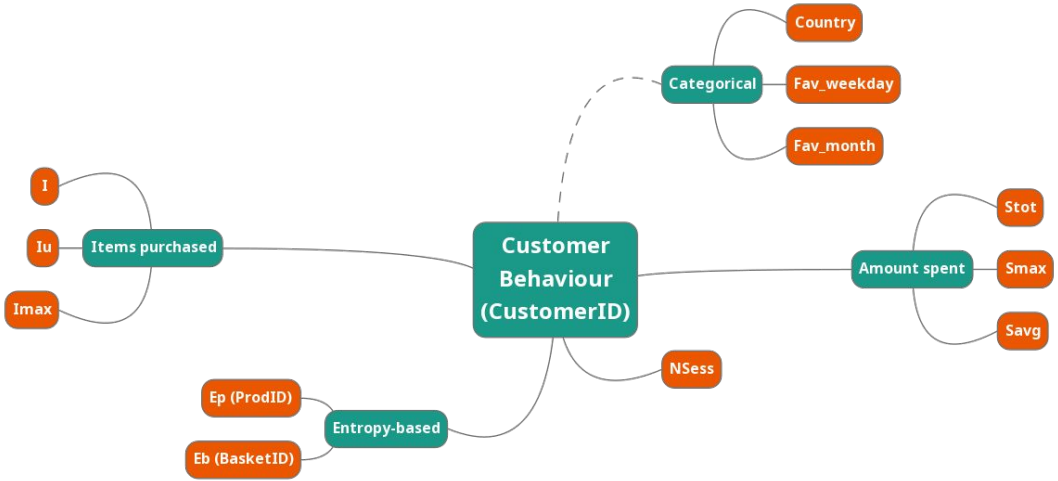


Task 2 - Data Clustering

- **Clustering by K-Means**
- **Clustering by DBSCAN**
- **Hierarchical Clustering**
- **Also clusterings by:**
  - CURE
  - OPTICS
  - Fuzzy C-means

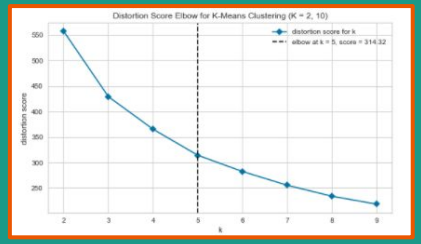
Task 2 - Preprocessing

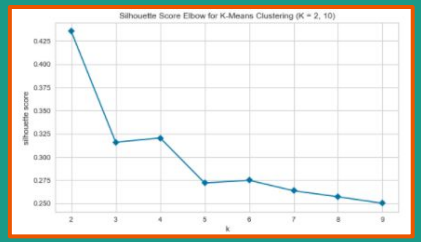
- Elimination of redundant attributes
- Min-max normalization
- Extraction of the categorical attributes for clusters characterization

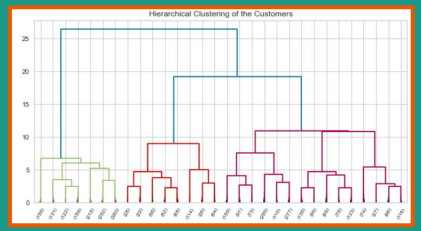


K-Means: Identifying the value of K

- Elbow method
- Average Silhouette method
- Insights from Hierarchical Clustering
  - Ward-linkage
- Evaluation via Internal Metrics
  - Sum of Squared Error (SSE)
  - Davies Bouldin Index
  - Silhouette Score
  - Calinski-Harabasz Index

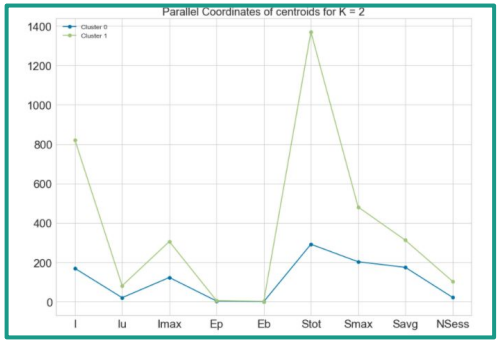


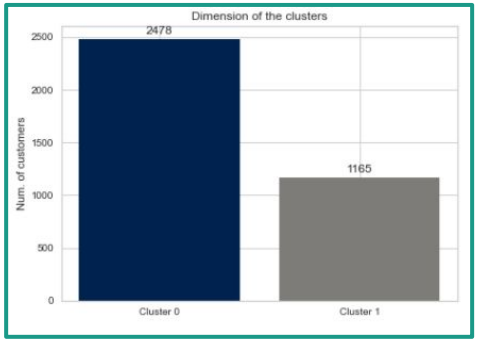


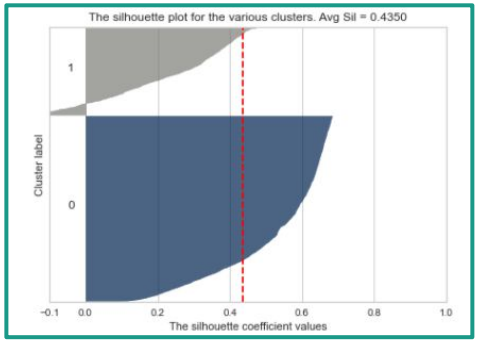


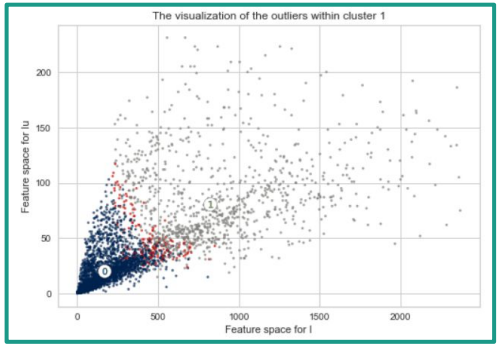
K-Means: Clustering for K = 2

- **Silhouette analysis**
  - Outliers from cluster 1
- **Visualization via scatterplots**
- **Identification of the influence of each attribute**
  - Parallel coordinates and Radar-plot
  - Low-spending and High-spending customers



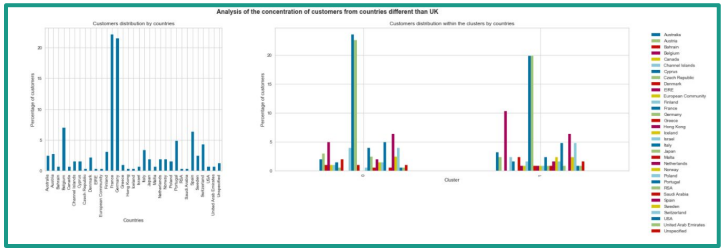


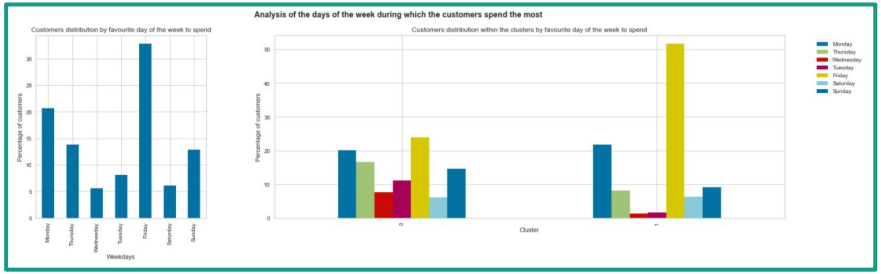


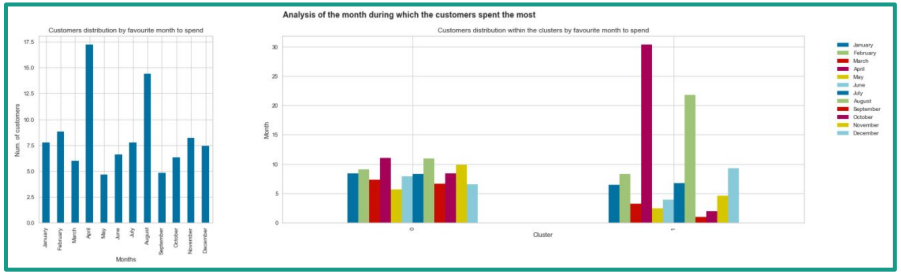


K-Means: Clusters characterization by categorical att.

- **CustomerCountry**
  - **Low-spending countries**: EIRE, European Community, Hong Kong, Iceland, RSA
  - **High-spending countries**: Saudi Arabia, Bahrain, Canada, Czech Republic
- **Fav_weekday & Fav_month**
  - **Low-spending customers**: mostly balanced shopping behaviour
  - **High-spending customers**: very unbalanced shopping behaviour





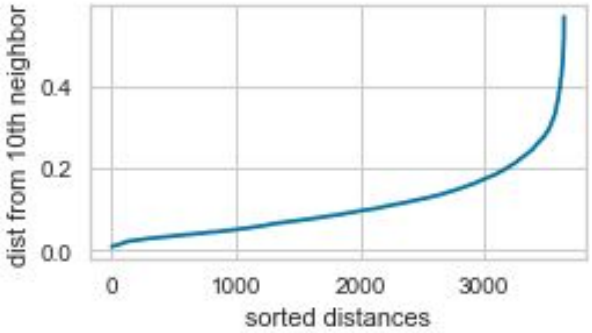


DBSCAN: Parameter Tuning with Elbow (or Knee) method

Double evaluation to ensure a more precise and stable combination of optimal values.

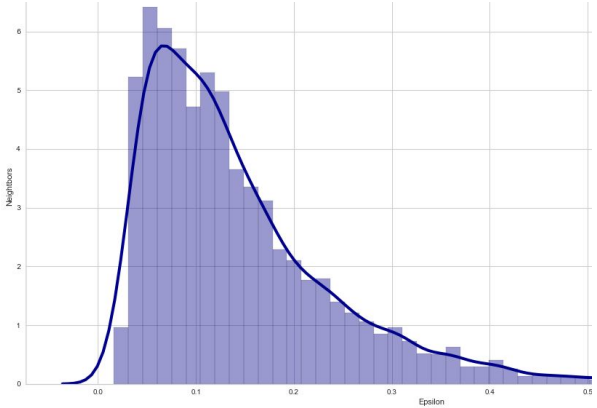
Fixing neighborhood

We tried different values of *min_sample* parameter, in a range between 5 and 35



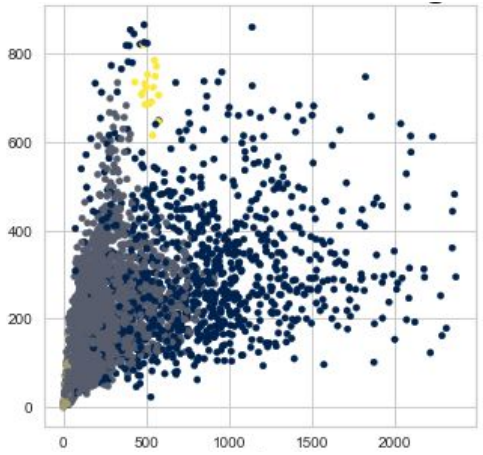
Fixing radius

We tried different values of the *epsilon* parameter in a range between 0 and 0.8



Best configuration result

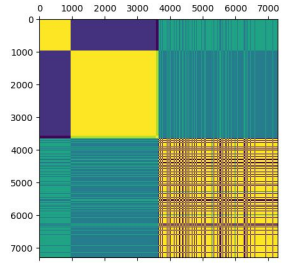
Four clusters detected



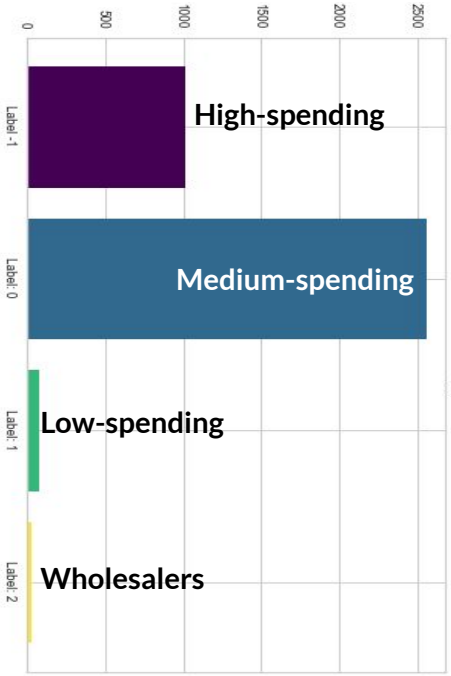
Points in low-density regions are classified as noise and omitted; thus, DBSCAN does not produce a complete clustering. Kumar Book, chapter 7, page 533

DBSCAN: Post-processing analysis

Similarity Matrix



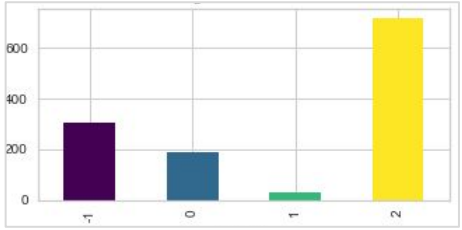
Population of clusters



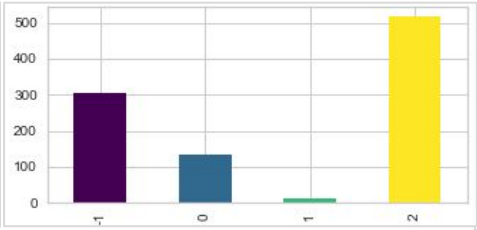
Since clusters are well-separated they show a very strong, block-diagonal pattern in the ordered similarity matrix.

Silhouette:	0.19
Separation:	0.90

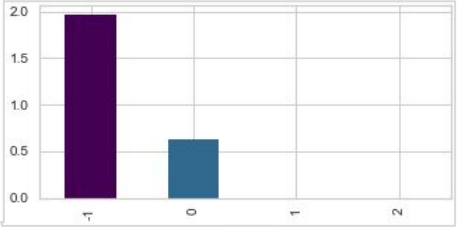
Wholesalers	High-spending
Bought the highest number of items and spent the most within a **single shopping session**.	Bought a lot of items and spent a lot, but in **many shopping sessions**.



Average spent



Maximum number of items purchased within a single shopping session

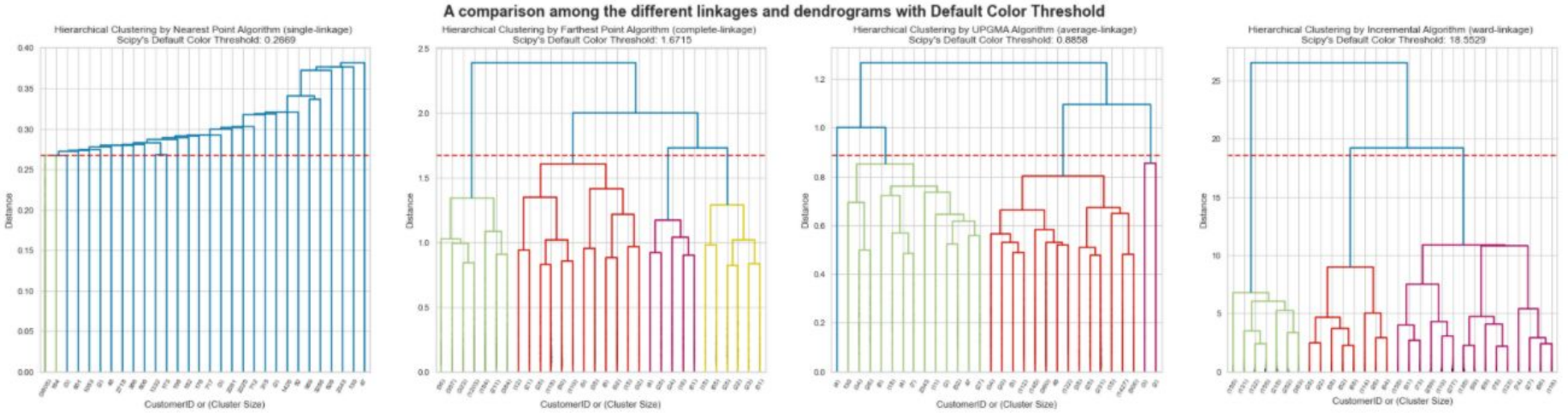


Shannon's Entropy on the number of baskets

Hierarchical Clustering

Comparison between the dendrograms for the four main linkages

- Visualization and cuts from the **default value for color_threshold**

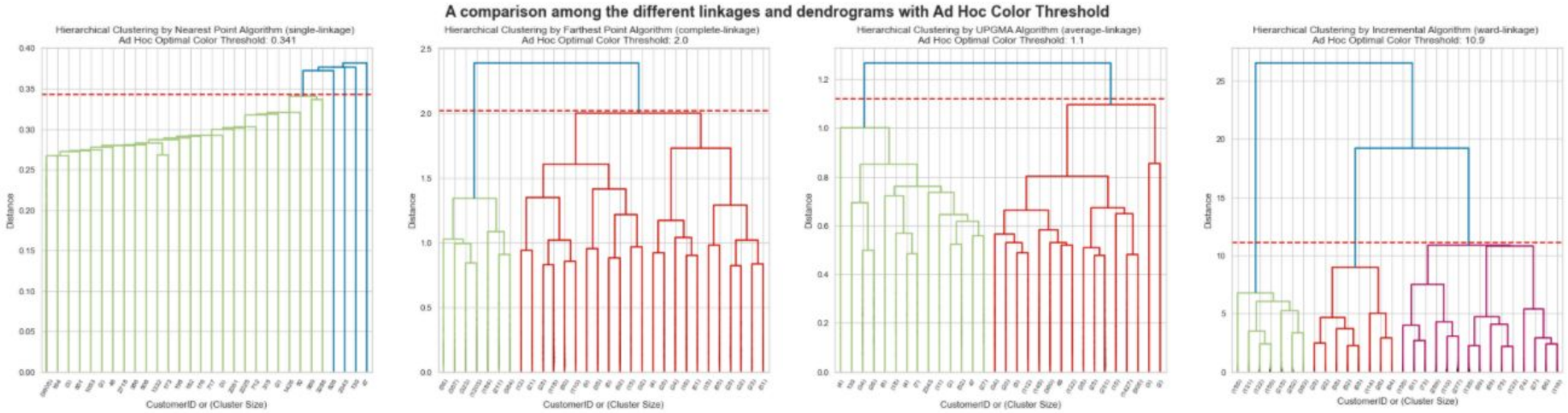


Single *Complete* *Average* *Ward*

Hierarchical Clustering: identifying the optimal cut

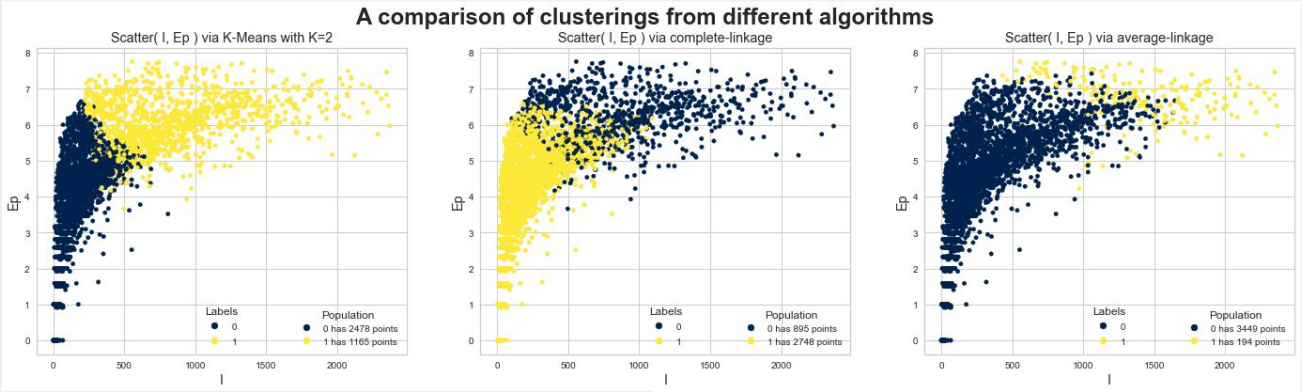
Comparison between the dendrograms for the four main linkages

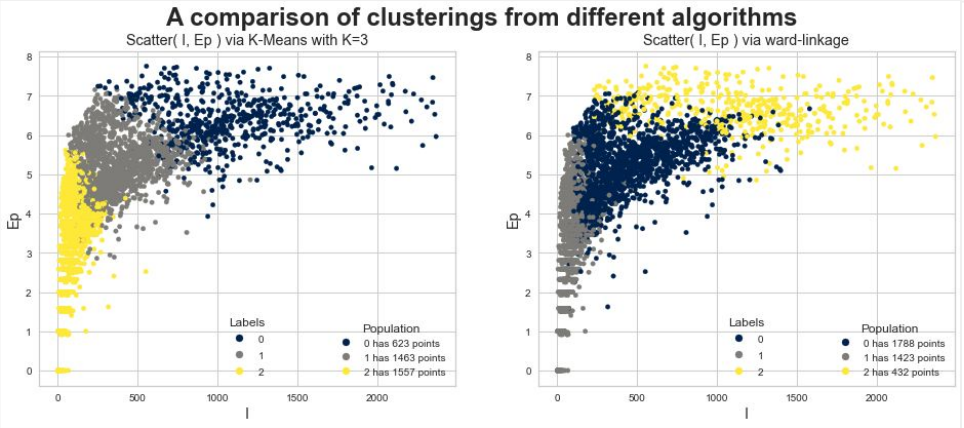
- Identification of the **optimal cut** (**longest uninterrupted segment**)
- Also, evaluation via **Internal metrics** and **Cophenetic Correlation Coefficient**



Single *Complete* *Average* *Ward*

Clustering algorithms comparison

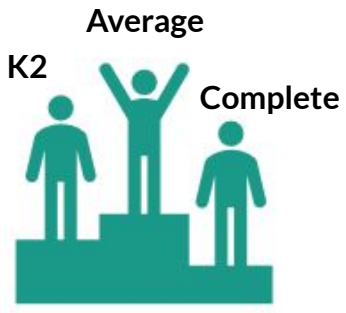




		hom	compl	v_meas
clustering_a	clustering_b			
K_2	complete	0.46	0.52	0.49
	average	0.10	0.31	0.15
K_3	ward	0.50	0.53	0.52
K_4	DBSCAN	0.23	0.42	0.30

Clustering final evaluation

	Davies_Bouldini	Silhouette	Calinski_Harabasz
K2	1.002153	0.435007	3021.802248
K3	1.160057	0.315443	2505.103841
K4	1.192858	0.321506	2172.847202
K5	1.147942	0.272066	2041.370447
K7	1.183484	0.270354	1811.916029
K8	1.127170	0.257634	1743.472511
single	0.382891	0.381495	5.938170
complete	1.078697	0.426685	2421.001945
average	0.847240	0.493721	1016.287862
ward	1.157258	0.270928	2011.862144



But also...
optional clustering algorithms !

- Optics
- Cure
- Fuzzy C-means

Task 3 - Data Classification

Naive Bayes classifiers
- Gaussian Naive Bayes
- Multinomial Naive Bayes

Support Vector Machine
- Support Vector Classification

Tree-based classifiers
- Decision tree

Neighbors-based classifiers
- K-Nearest Neighbors
- Radius-Neighbors

Machine Learning classifiers
- Feed-forward Neural Network
- Multi-layer Perceptron

Ensemble Method
- Random forest
- Voting classifier

Computing the label

Based on clustering results

- Since DBSCAN clusters were distributed following the purchasing behavior or a customer, we performed the entire classification task on its result, just to see if that label could be good.
- Problem: **the three classes were unbalanced**, since they have different support. The population of the three clusters was not equally distributed, as described before.
- We believe that this unbalanced situation had a huge impact on the models results.
- Just three models over a total of ten performed well.

Based on *Savg* indicator

It is the most representative attribute of the customers' shopping profile: it is computed on the basis of the average amount spent by each customer during a shopping session so that its relevance remains significant **to both long-standing customers and brand-new ones.**

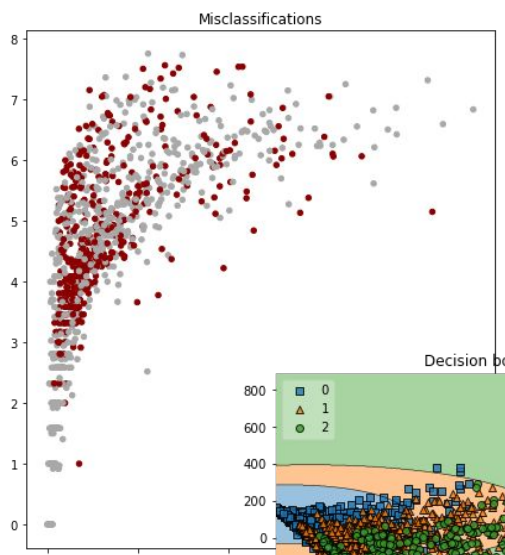
How to define the three classes:

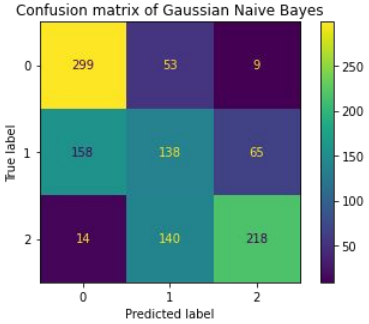
Using percentiles	• Savg < 25% → low-spend • Savg > 75% → high-spend • 25% < Savg < 75% → medium-spend
Using quantiles	• Savg < 0.33 → low-spend • Savg > 0.66 → high-spend • 0.33 < Savg < 0.66 → medium-spend

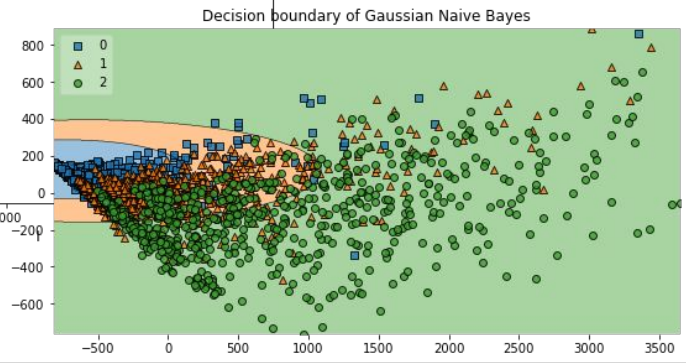
The second way provides **the most balanced distribution of customers**: this avoids incrementing the relevance of a class with respect to the others. This choice **saved us by using any kind of weight parameter** in our classifiers.

Naive Bayes classifiers

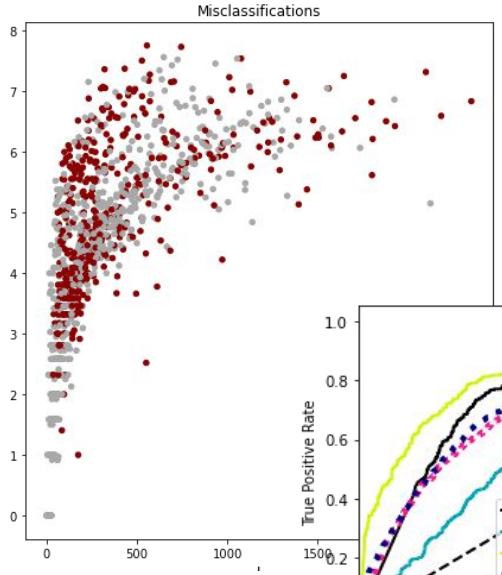
Gaussian Naive Bayes

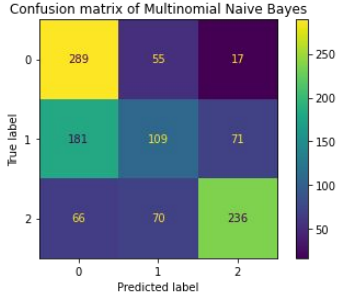


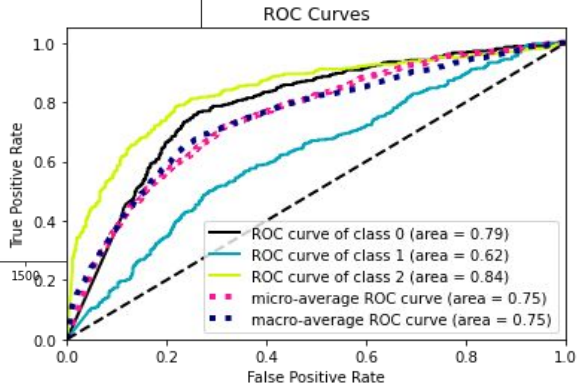




Multinomial Naive Bayes

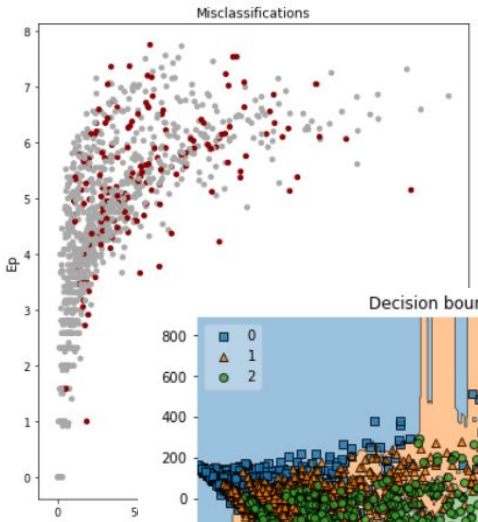


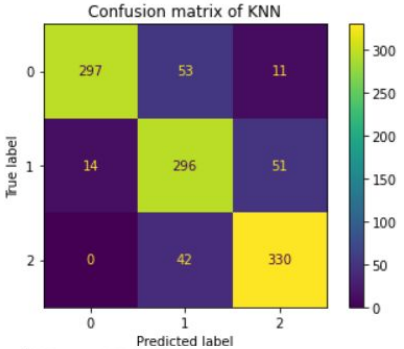


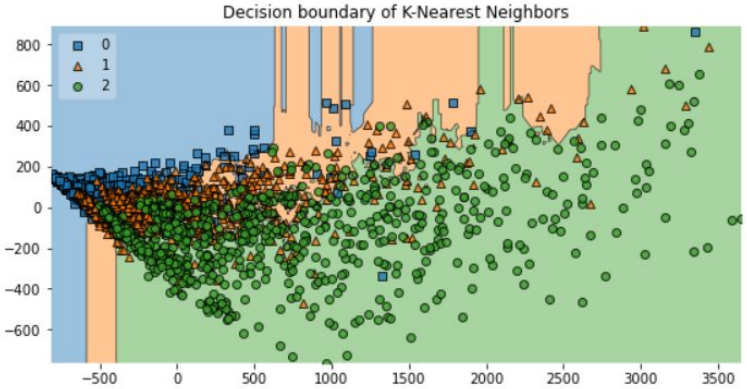


K-Nearest Neighbors and Radius Neighbors Classifier

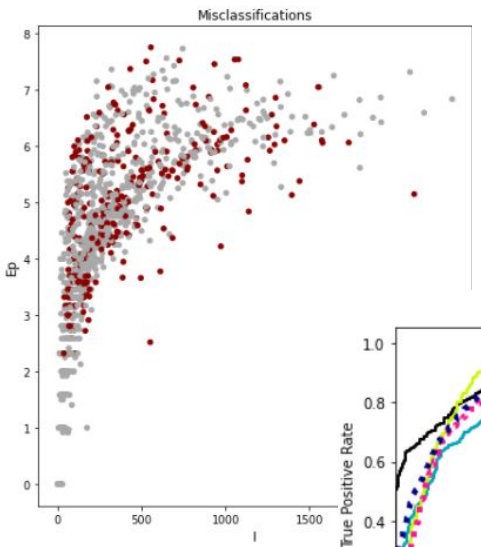
K-Nearest Neighbors

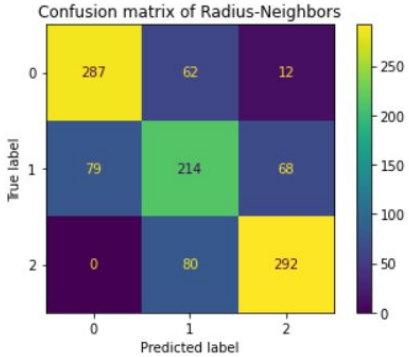


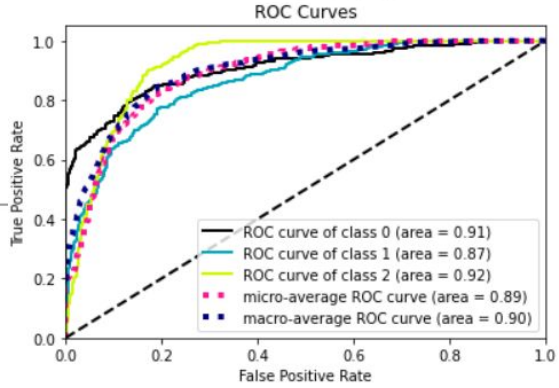




Radius Neighbors

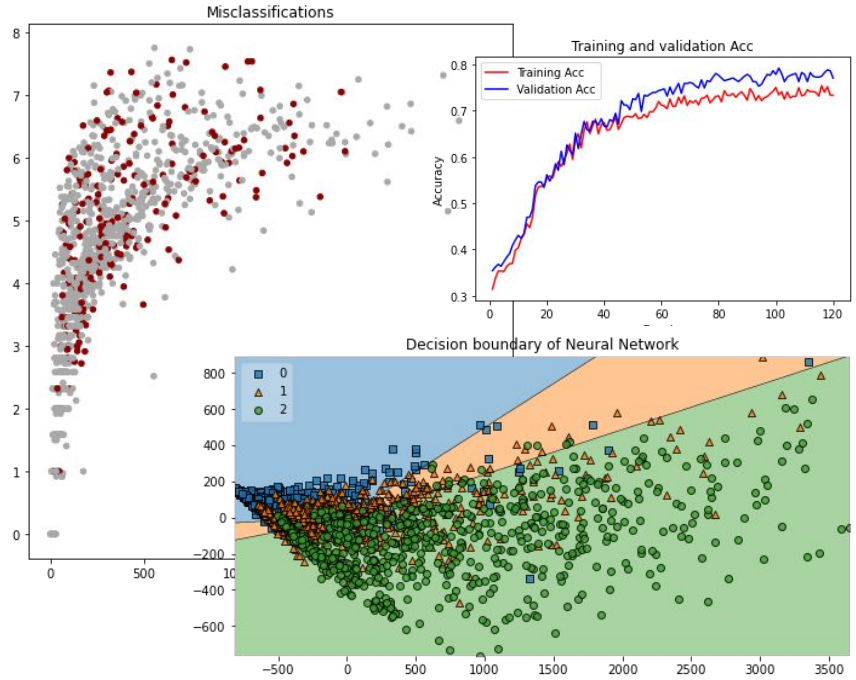




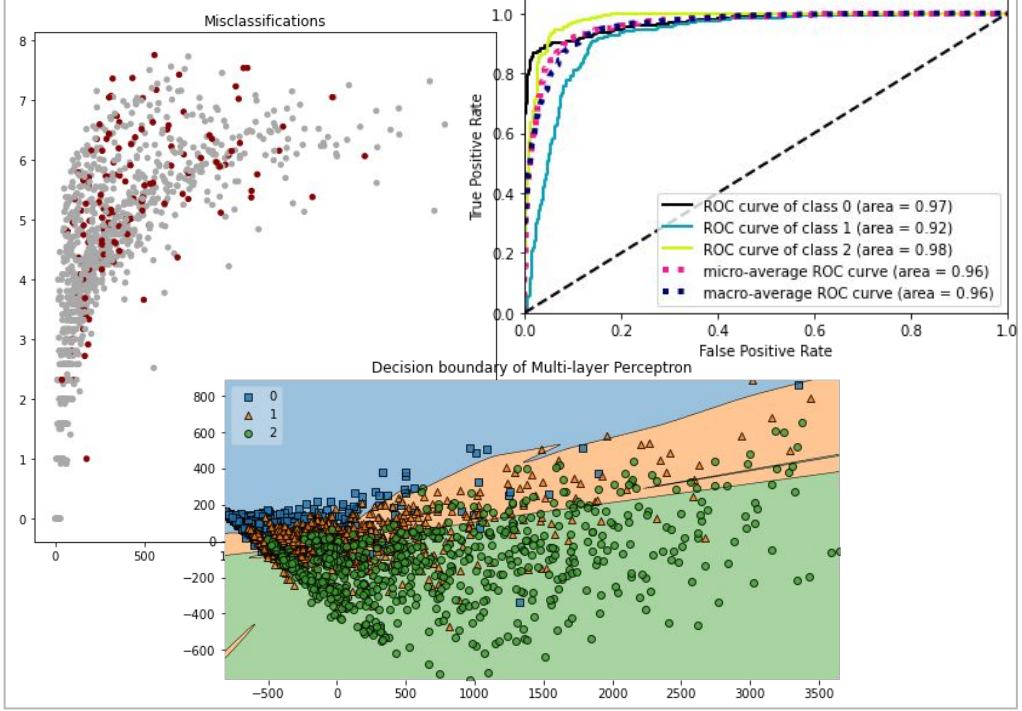


Machine Learning classifiers

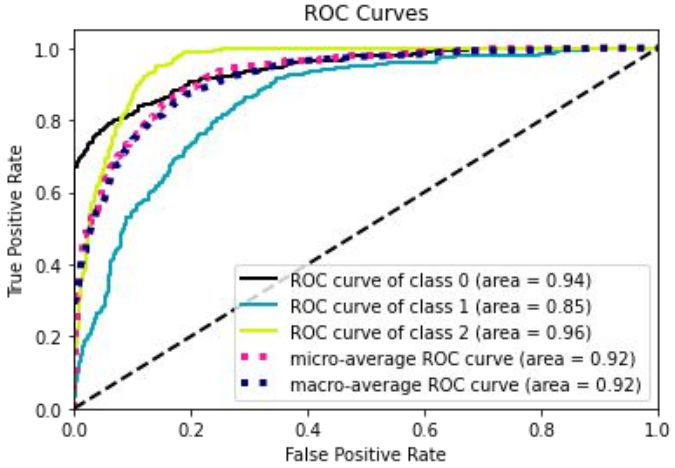
Feed-forward Neural Network

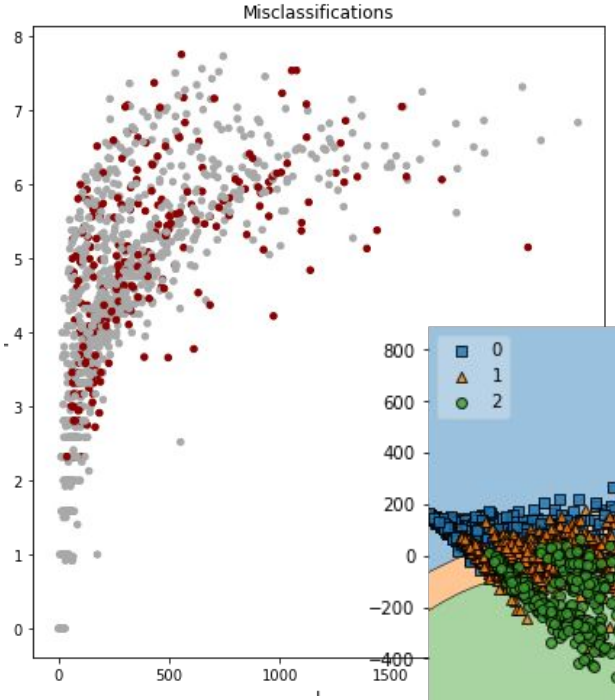


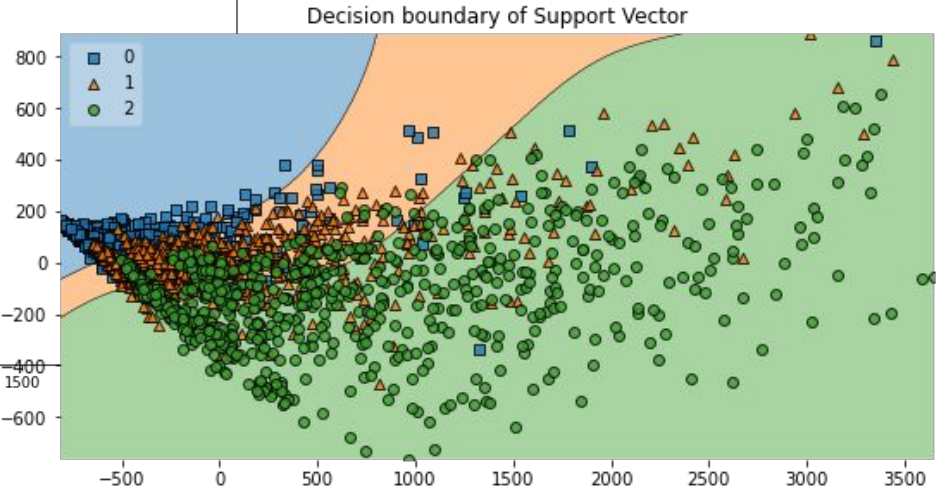
Multi-layer perceptron



SVM



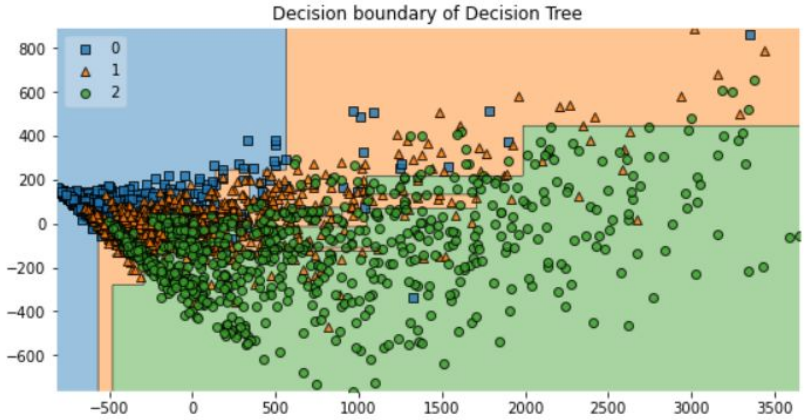


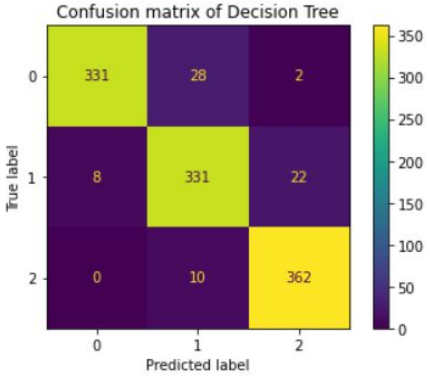


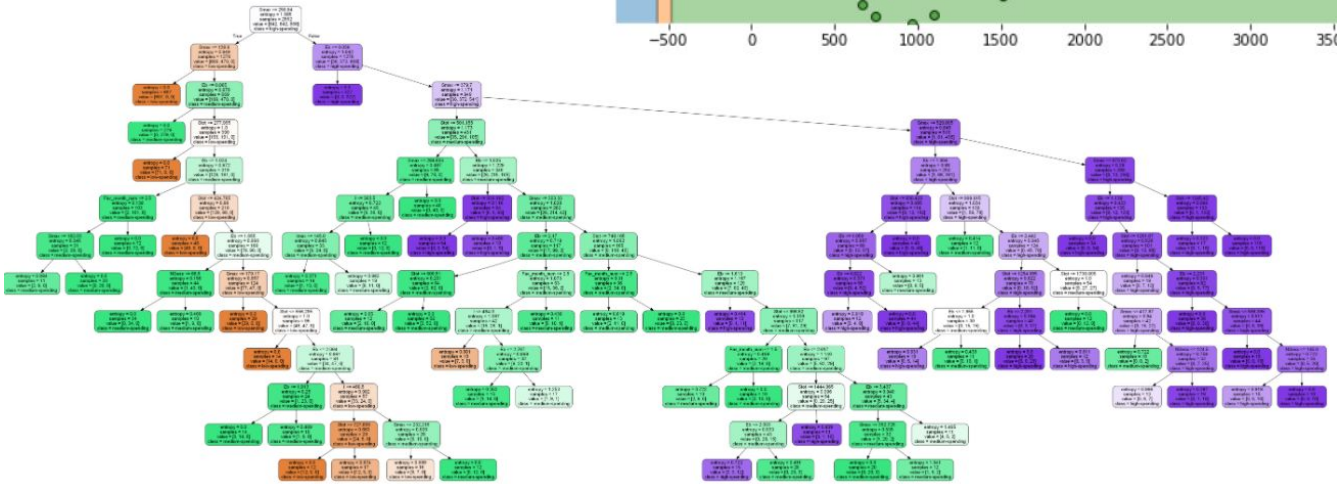
Radial Basis Function (RBF) kernel.

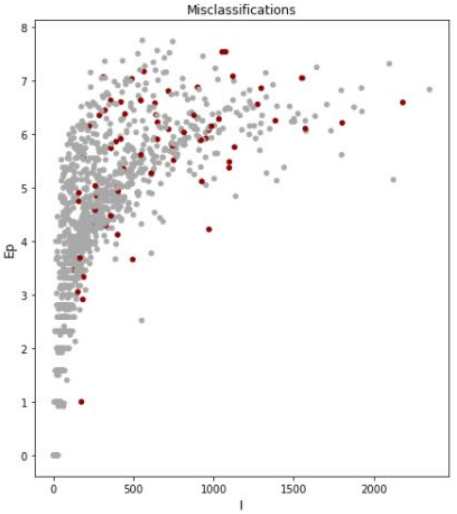
RBF Kernel is popular because of its similarity to K-Nearest Neighborhood Algorithm. It has the advantages of K-NN and overcomes the space complexity problem as RBF Kernel Support Vector Machines just needs to store the support vectors during training and not the entire dataset.

Decision Tree



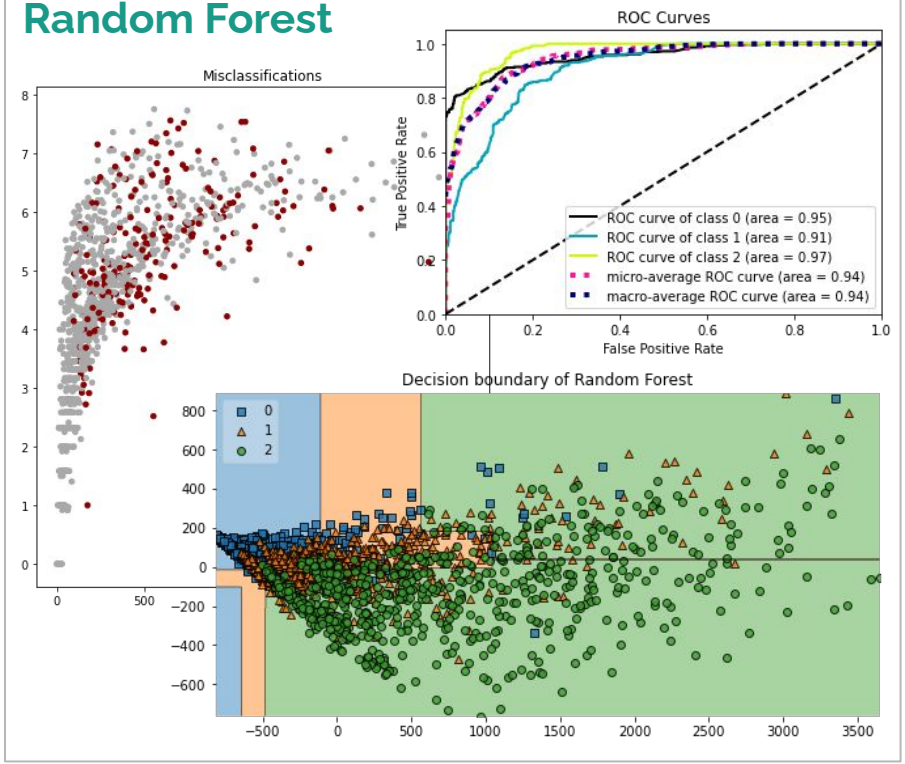




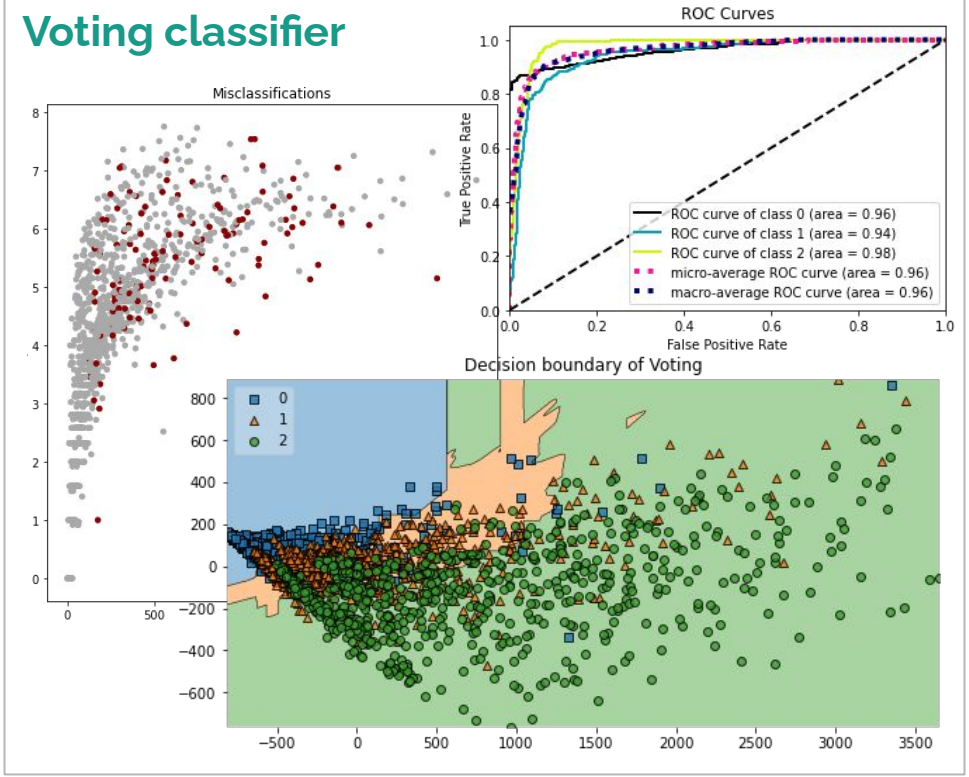


Ensemble methods

Random Forest



Voting classifier



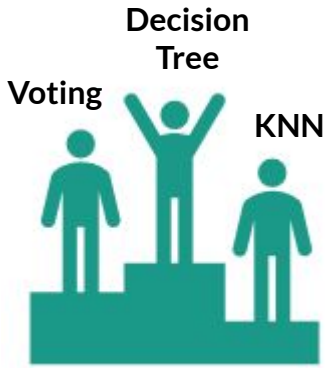
Classifiers evaluation and comparison

Reporting accuracy on test set for all the classifiers:

Model	GNB	MNB	KNN	Radius	NN	MLP	SVM	DT	RF	Voting
Acc	0.60	0.58	0.85	0.76	0.79	0.88	0.77	0.94	0.80	0.88

Not just accuracy ...

- Best **predictions**
- Lowest number of **misclassified data points**
- Best **confusion matrices** (i.e. few errors in class attribution)
- **Hyperplanes** that accurately divide data into the three classes
- Elbow of the **ROC curve** very close to (0,1), i.e. high True Positive rate



Task 4 - Sequential Pattern Mining

Sequential Pattern Mining: the process

1. **Building the sequence of baskets associated with each customer**
   - $Customer_i$ = [$Basket_1$, $Basket_2$, $Basket_3$...], with $Basket_i$ = [$Product_X$, $Product_Y$, $Product_Z$...]
2. **Excluding short sequences**
   - Customers with less than two shopping sessions
3. **Running the apriori algorithm for different percentages of minimum support**
   - Looking for the sequences of products purchased by the 5/10/15/20 % of the dataset population
4. **Substituting the ProdIDs with ProdDescr in order to improve understandability**
   - From [['23293'], ['23295']] to ['RECIPE BOX PANTRY YELLOW DESIGN'], ['SET OF 3 CAKE TINS PANTRY DESIGN ']]

Sequential Pattern Mining: analysing the results

- **Very large presence of sequences of the same or similar articles.**
  Examples:
  - [['ALARM CLOCK BAKELIKE GREEN'], ['ALARM CLOCK BAKELIKE GREEN']]
  - [['LUNCH BAG RED RETROSPOT', 'LUNCH BAG WOODLAND']]

- **Two possible explanations:**
  - Multiple purchases over time due to the **disposable nature of the products**
  - **Customers replenishing stocks** so to act as **re-sellers for external third parties**

Thank you !