

Customer Personality Analysis

Mario Avolio Rocco Gianni Rapisarda

Università Milano Bicocca - Dipartimento di Informatica Sistemistica e Comunicazione

22 gennaio 2022



Analisi del Dominio e Obiettivi

- 1 Analisi dettagliata dei clienti
- 2 Aiutare un'attività commerciale a comprendere meglio i propri compratori
- 3 Rendere più semplice la modifica e la scelta dei propri prodotti, in relazione alle esigenze richieste dagli acquirenti
- 4 Diverse personalità e comportamenti che gli acquirenti assumono durante il ruolo di potenziali clienti aziendali
 - Le aziende non possono adottare lo stesso approccio per ogni tipologia di plausibile compratore

Descrizione dei Dati

1 Informazioni Personali

- Grado educativo
- Reddito
- Numero di figli
- Età

2 Prodotti e Spese

- Spesa totale, negli ultimi due anni, di un prodotto di determinato genere.

3 Promozioni e offerte

- Offerte accettate delle diverse campagne presenti.

4 Luoghi e acquisti

- Numero di compere effettuate in un determinato luogo o in un determinato modo.

Prime Analisi: Income

	variable	n_miss	pct_miss
1	Income	24	1.07
2	ID	0	0.00
3	Year_Birth	0	0.00
4	Education	0	0.00
5	Marital_Status	0	0.00
6	Kidhome	0	0.00
7	Teenhome	0	0.00
8

Tabella: Output funzione `miss_var_summary(dataSet)`

Prime Analisi: Z_CostContact e Z_Revenue

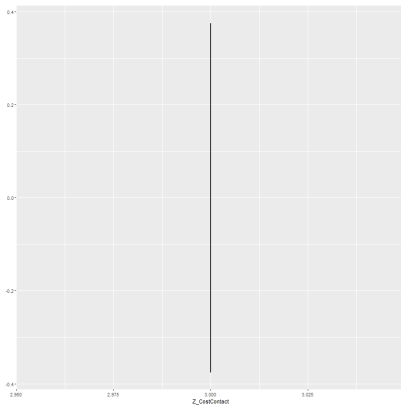


Figura: BoxPlot Z_CostContact

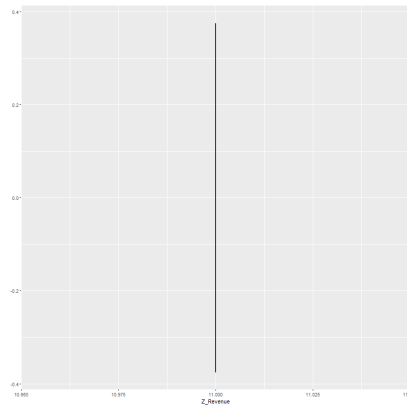


Figura: BoxPlot Z_Revenue

DataPreprocessing

- 1 Refactor del dataset
- 2 Risoluzione dei valori mancanti nella variabile *income*
- 3 Splitting del dataset in *trainingSet* e *testSet*
- 4 Feature Scaling

Refactor del Dataset

- 1 Incorporamento dei dati ridondanti
- 2 Conversione degli elementi in *factor*
- 3 Creazione di nuove variabili riassuntive
 - Age
 - Total_Spent
 - Total_Campaigns
 - Total_Childs
- 4 Rimozione delle variabili superflue
 - Z_Revenue
 - Z_CostContact
 - ID
 - Dt_customers

EDA

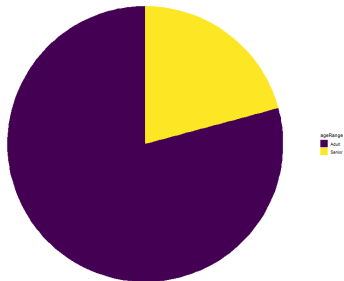


Figura: Grafico a torta di *Age*

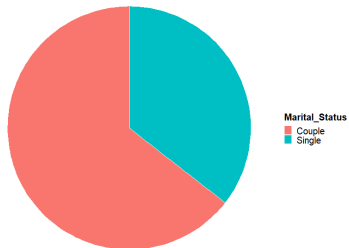


Figura: Grafico a torta di *Marital_Status*

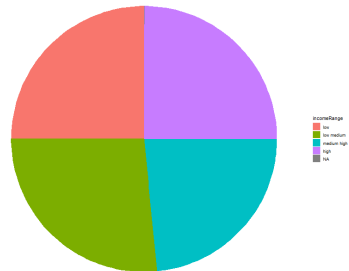


Figura: Grafico a torta di *Income*

EDA: Age, Education e Marital_Status

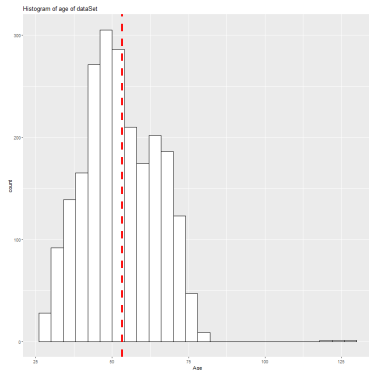


Figura: Istogramma di *Age*

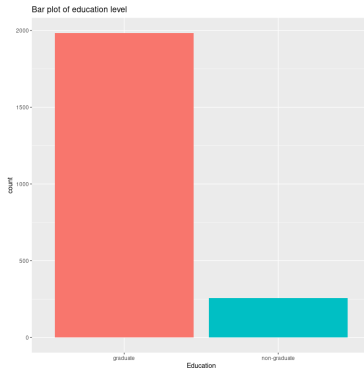


Figura: Grafico a barre di *Education*

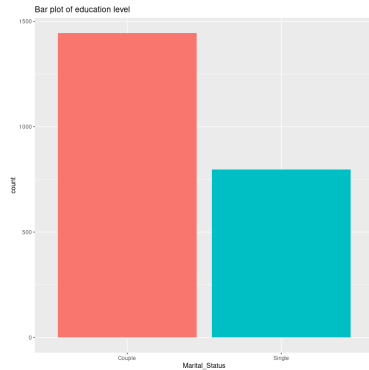


Figura: Grafico a barre di *Marital_Status*

EDA: Istogrammi delle variabili Amount

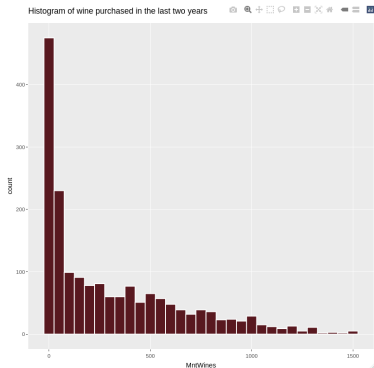


Figura: Istogramma di *Wines*

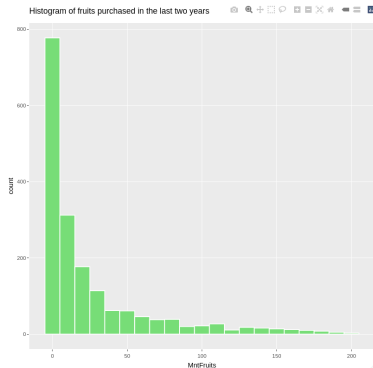


Figura: Istogramma di *Fruits*

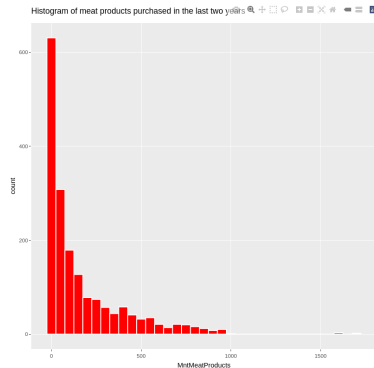


Figura: Istogramma di *Meat*

EDA: Istogrammi delle variabili Amount

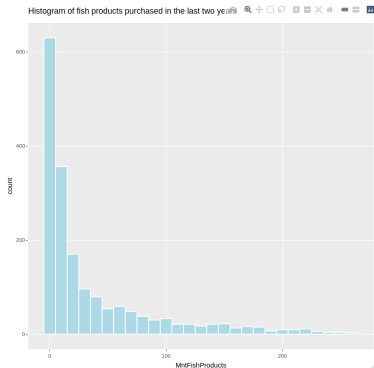


Figura: Istogramma di *Fish*

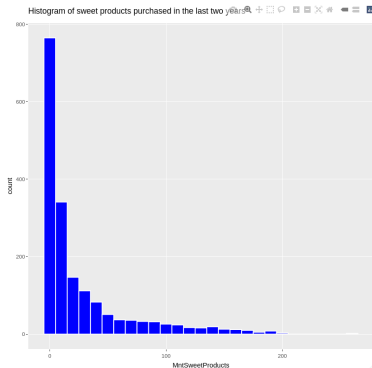


Figura: Istogramma di *Sweet*

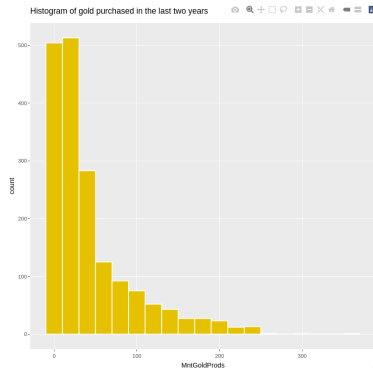


Figura: Istogramma di *Gold*

EDA: Total_Childs, MntProducts e Campaign

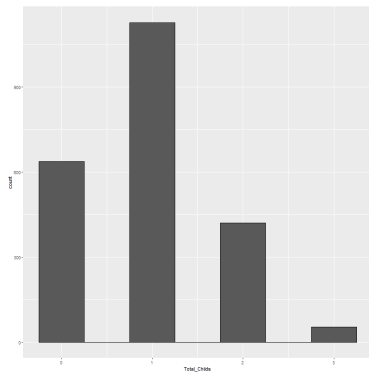


Figura: Grafico a barre di *Total_Childs*

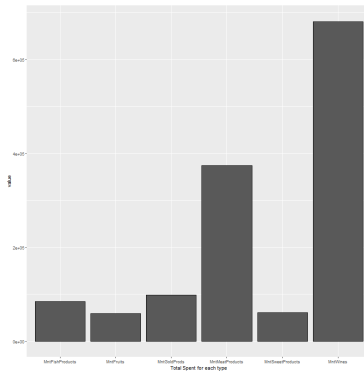


Figura: Grafico a barre del totale speso per ogni tipo di prodotto

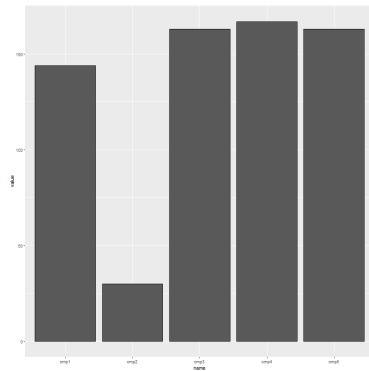


Figura: Grafico a barre del totale di istanze che hanno accettato la campagna i-esima

PCA: Varianza spiegata da ogni dimensione

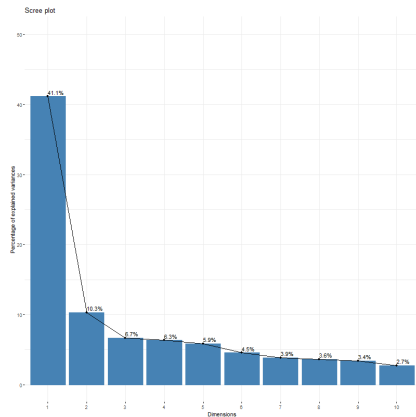
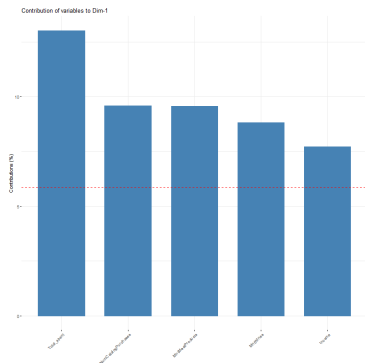


Figura: Varianza spiegata da ogni Dimensione

	eig.	v.p.	c.v.p.
Dim.1	7.00	41.15	41.15
Dim.2	1.75	10.31	51.46
Dim.3	1.14	6.69	58.15
Dim.4	1.08	6.34	64.49
Dim.5	1.00	5.86	70.34
...

Tabella: Output funzione *get_eigenvalue(pca)*

PCA: Varianza spiegata per la Dim1



- 1 Total_Spent
- 2 MntMeatProducts
- 3 NumCatalogPurchases
- 4 MntWines
- 5 Income

Figura: Varianza spiegata dalle variabili per la prima dimensione

Algoritmi scelti

- Supervisionato: **Decision Tree**
- Non Supervisionato: **K-Means**

Decision Tree

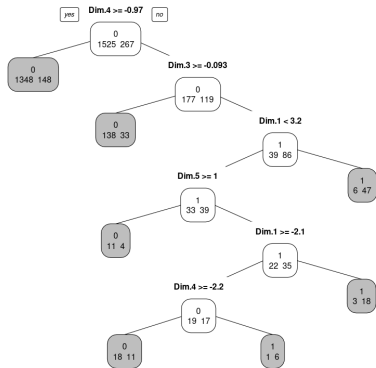


Figura: Decision Tree dataset della PCA

Positive Class: 1
Accuracy: 0.8348
Recall: 0.3478

Precision: 0.1194
F-Measure: 0.1777

Decision Tree: Riduzione Overfitting

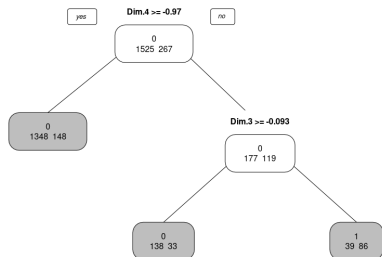


Figura: Prune Decision Tree $cp = 0.13$

Positive Class: 1

Accuracy: 0.8147

Recall: 0.1492

Precision: 0.2777

F-Measure: 0.1941

Valutazione del modello: Decision Tree

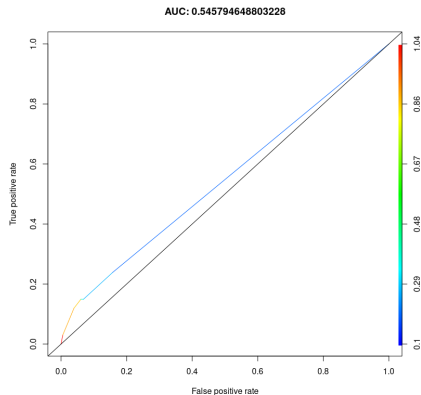


Figura: Curva ROC

	Prediction		
	—	0	1
Reference	0	1488	37
	1	200	67

Accuracy: 0.8677

Precision: 0.2509

Recall: 0.6442

F-Measure: 0.3611

K-Means: Elbow Method

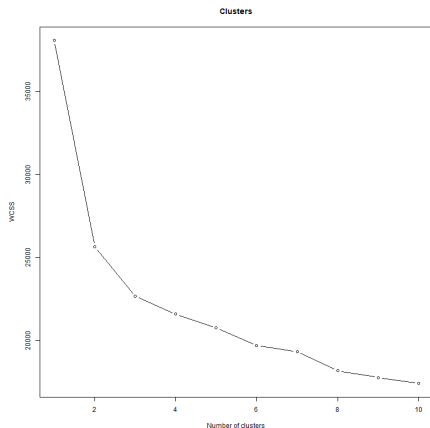


Figura: Elbow Method effettuato manualmente

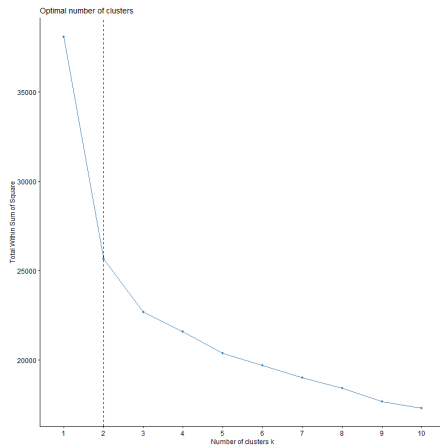
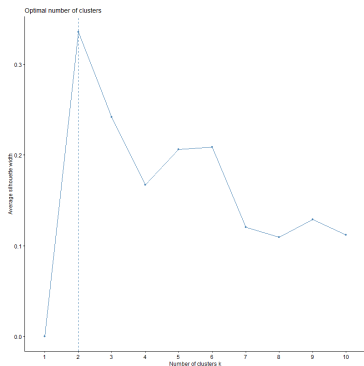


Figura: Elbow Method effettuato automaticamente dal metodo *fviz_nbclust*

K-Means: Silhouette



Sia **Elbow Method** che **Silhouette** mostrano un numero di clusters ottimo pari a due.

Figura: Silhouette effettuata automaticamente dal metodo *fviz_nbclust*

K-Means: Algoritmo



Figura: Partizionamento in clusters dei dati

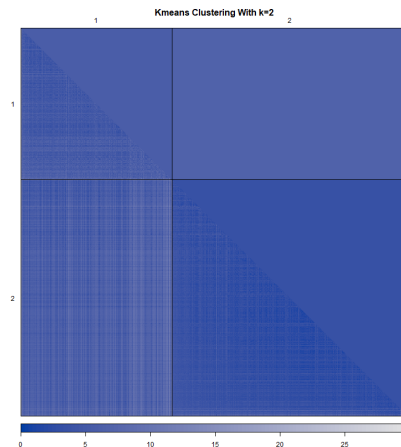


Figura: Dissimilarity matrix

K-Means: Analisi Wines

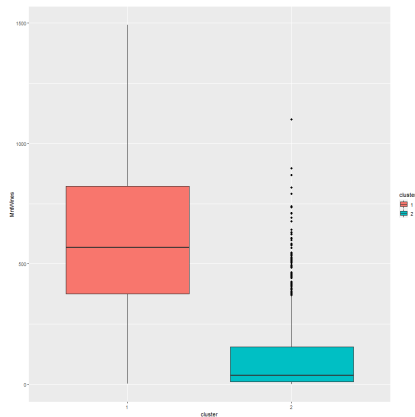


Figura: BoxPlot della variabile Wines in relazione al numero di cluster

- Spesa maggiore di vini per i *customers* all'interno del primo cluster
- La maggior parte dei clienti all'interno del secondo cluster non ha acquistato vini negli ultimi due anni

K-Means: Analisi Income

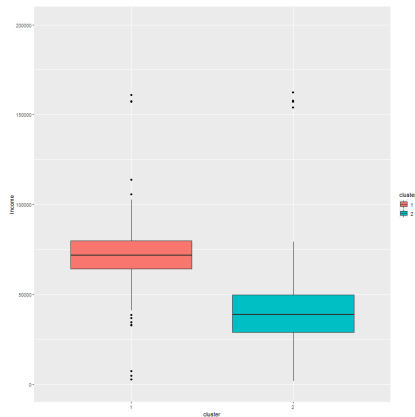


Figura: BoxPlot della variabile income in relazione al numero di cluster

- Gli elementi del primo raggruppamento godono di un reddito medio maggiore rispetto ai secondi
- La maggior parte degli elementi nel secondo cluster possiedono un reddito inferiore alla media.

K-Means: Analisi Total_spent

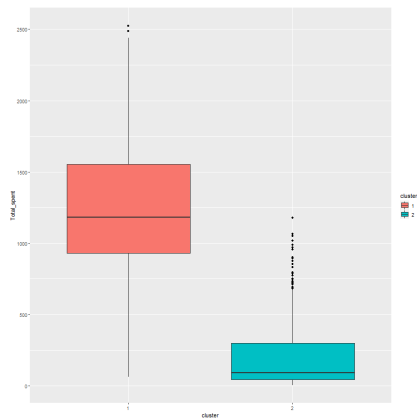


Figura: BoxPlot della variabile Total_spent in relazione al numero di cluster

- I compratori del secondo cluster generalmente spendono molto meno denaro rispetto a quelli del primo.

K-Means: Analisi NumCatalogPurchases

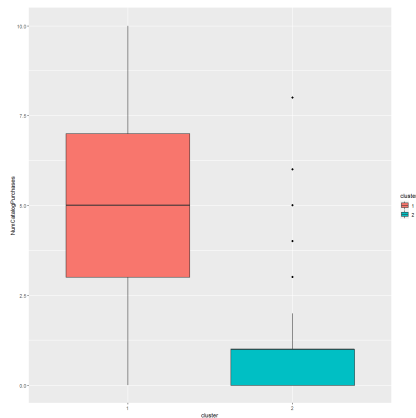


Figura: BoxPlot della variabile NumCatalogPurchases in relazione al numero di cluster

- I clienti del secondo cluster effettuano compere sul catalogo generalmente in quantità minore rispetto a quelli del primo.
- I clienti del secondo cluster acquistano mediamente 5 prodotti dal catalogo.

K-Means: Analisi MntMeatProducts

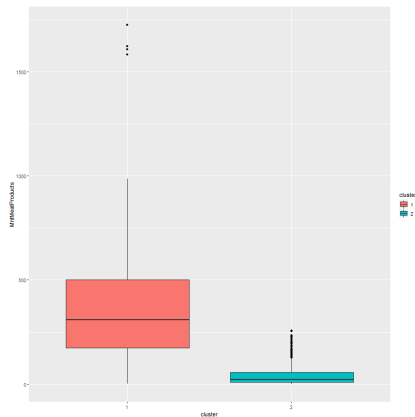


Figura: BoxPlot della variabile MntMeatProducts in relazione al numero di cluster

- Correlazione con Total_spent
- Gli acquirenti del secondo raggruppamento tendano a spendere generalmente di meno rispetto a quelli del primo

Conclusioni

1 Decision Tree

- la riduzione dell'overfitting comporta:
 - incremento di *Precision* e *F-measure*
 - decremento di *Accuracy* e *Recall*
- meno efficiente rispetto *K-Means*

2 K-Means

- Una buona suddivisione dei dati riportati può avvenire mediante l'utilizzo di due cluster.
- Il secondo cluster presenta clienti con un reddito generalmente al di sotto della media e sicuramente minore rispetto alla maggior parte dei compratori facenti parte della prima divisione.
- Riduzione delle spese totali da parte degli elementi all'interno del secondo gruppo.

Thank you for your attention!

