Relazione Caso di Studio - Ingegneria della Conoscenza a.a. 2020/2021

Valutazione di algoritmi di clustering vari applicati a un dataset costituito da dati sui clienti di una attività commerciale

Studenti:
Marco Barnaba , Mat. 699506 , m.barnaba21@studenti.uniba.it
Pierluca Lovero, Mat. 669685, p.lovero6@studenti.uniba.it

Indice dei contenuti:

Link Repository

Introduzione
L'obiettivo di questo caso di studio è l'analisi della personalità, ovvero un'analisi che aiuta l'azienda a comprendere i propri clienti rendendo più facile la modifica di prodotti e offerte in base a specifici comportamenti e alle caratteristiche riscontrate in gruppi di persone simili.
E' risultato utile procedere con diversi approcci algoritmici che hanno permesso di confrontare e valutare i diversi risultati, fornendo punti di vista diversi sullo stesso set di dati.
Allo stesso modo, è stato importante effettuare una corposa analisi esplorativa al fine di comprendere al meglio le caratteristiche dei dati trattati, avvalendosi di grafici per evidenziare e confrontare i diversi aspetti delle feature considerate.

Strumenti utilizzati
- Linguaggio di programmazione: **Python**
- Modelli di apprendimento e calcolo: **SkLearn**, **NumPy**, **Pandas**
- Visualizzazione dei dati: **Seaborn**, **Matplotlib**, **Plotly**, **Yellowbrick**

Contenuto del dataset
Feature

Persone
- ID: identificativo univoco del cliente
- Year_Birth: anno di nascita del cliente
- Education: livello di istruzione del cliente
- Marital_Status: stato civile del cliente
- Income: reddito familiare annuo del cliente
- Kidhome: numero di bambini nella famiglia del cliente
- Teenhome: numero di adolescenti nella famiglia del cliente
- Dt_Customer: data di registrazione del cliente con l'azienda
- Recency: numero di giorni dall'ultimo acquisto del cliente
- Complain: 1 se il cliente si è lamentato negli ultimi 2 anni, 0 altrimenti

Prodotti
- MntWines: importo speso in vino negli ultimi 2 anni
- MntFruits: importo speso in frutta negli ultimi 2 anni
- MntMeatProducts: importo speso in carne negli ultimi 2 anni
- MntFishProducts: importo speso in pesce negli ultimi 2 anni
- MntSweetProducts: importo speso in dolci negli ultimi 2 anni
- MntGoldProds: importo speso in oro negli ultimi 2 anni

Promozione
- NumDealsPurchases: numero di acquisti effettuati con uno sconto
- AcceptedCmp1: 1 se il cliente ha accettato l'offerta nella prima campagna, 0 altrimenti
- AcceptedCmp2: 1 se il cliente ha accettato l'offerta nella seconda campagna, 0 altrimenti
- AcceptedCmp3: 1 se il cliente ha accettato l'offerta nella terza campagna, 0 altrimenti
- AcceptedCmp4: 1 se il cliente ha accettato l'offerta nella quarta campagna, 0 altrimenti
- AcceptedCmp5: 1 se il cliente ha accettato l'offerta nella quinta campagna, 0 altrimenti
- Response: 1 se il cliente ha accettato l'offerta nell'ultima campagna, 0 altrimenti

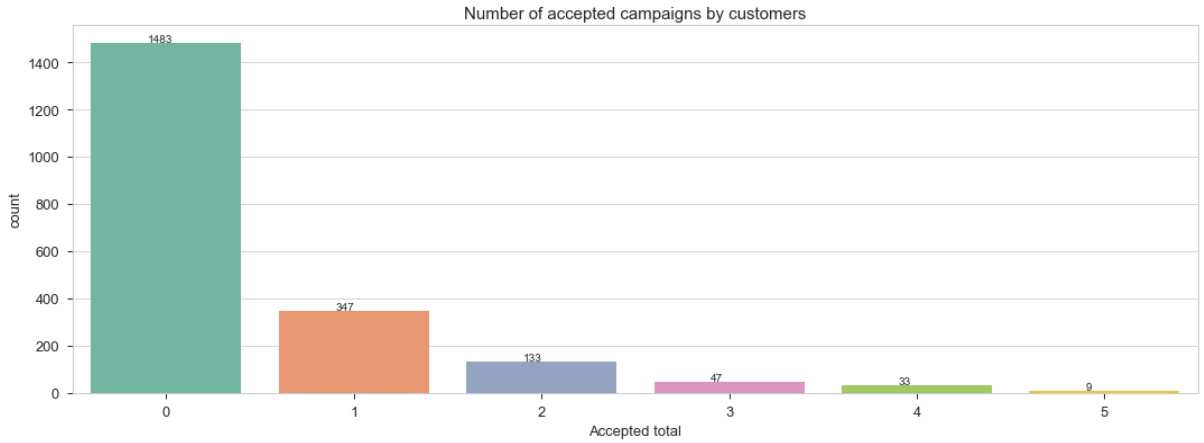
Fonte
- NumWebPurchases: numero di acquisti effettuati tramite il sito web dell'azienda
- NumCatalogPurchases: numero di acquisti effettuati usando il catalogo
- NumStorePurchases: numero di acquisti effettuati nei negozi
- NumWebVisitsMonth: numero di visite al sito web dell'azienda nell'ultimo mese

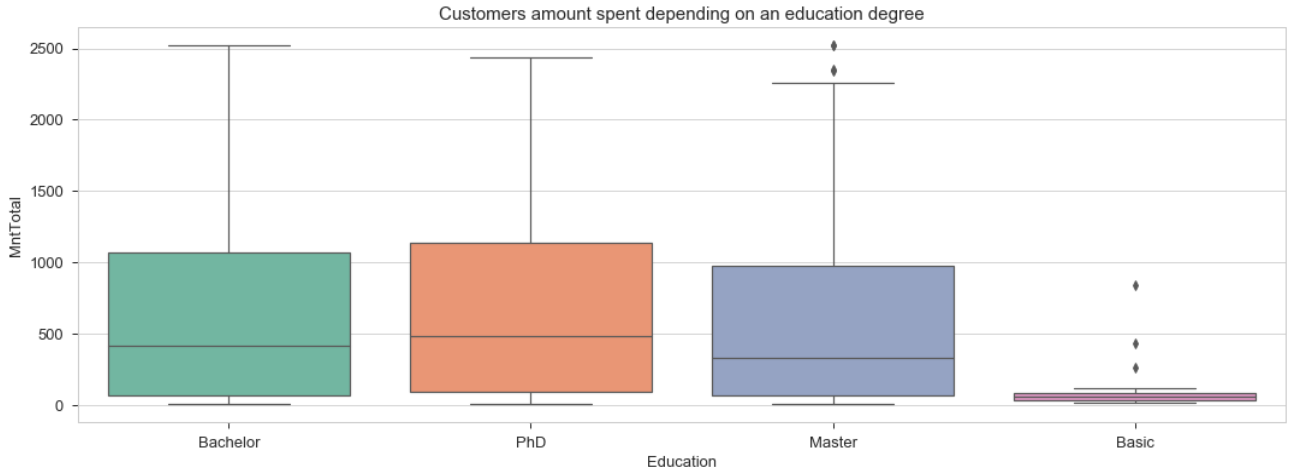
Target
E' necessario individuare i cluster in cui è possibile suddividere gli esempi.

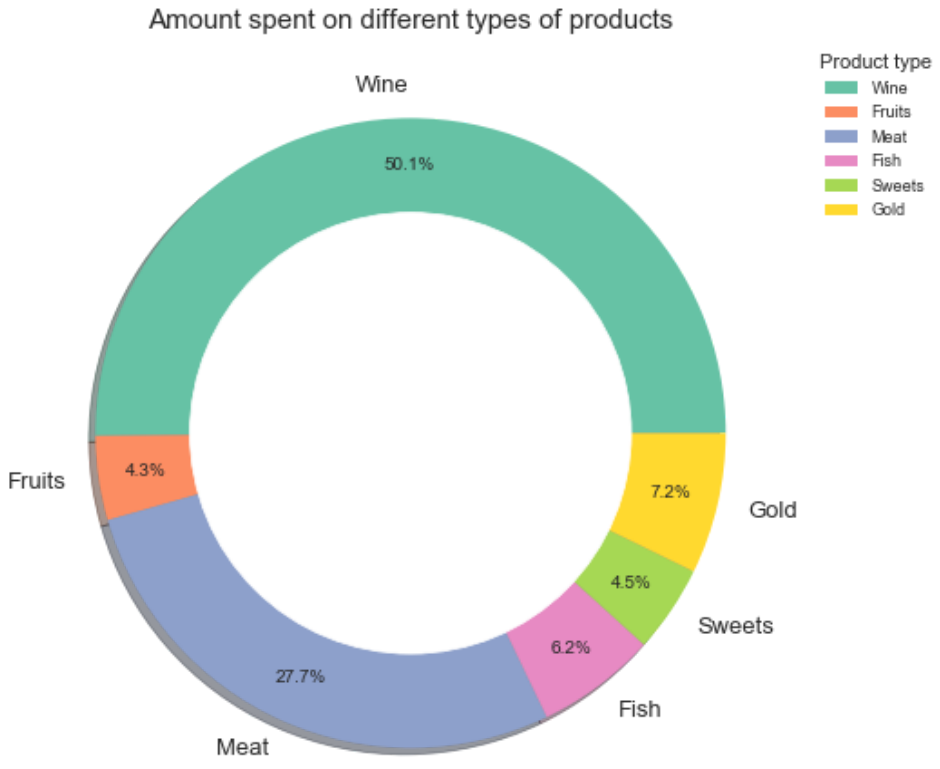
Analisi Esplorativa

Iniziando a lavorare con il dataset, ci si è concentrati innanzitutto sulla comprensione del formato dei dati e sulla validità dei valori per quanto riguarda feature categoriche e feature temporali (Date, Età, ecc.).
Successivamente si è indagato sulla presenza di outlier e valori null, in quanto essi avrebbero potuto condizionare l'efficienza degli algoritmi adoperati.

A questo punto si è posta l'attenzione sulle distribuzioni dei valori prendendo in considerazione le feature singolarmente oppure a coppie. Questo ci ha permesso di avere una prima indicazione delle relazioni tra le feature analizzate.
Di seguito alcuni esempi:


Number of accepted campaigns by customers
count
Accepted total
1483
347
133
47
33
9
0
1
2
3
4
5


Customers amount spent depending on an education degree
MntTotal
Education
Bachelor
PhD
Master
Basic

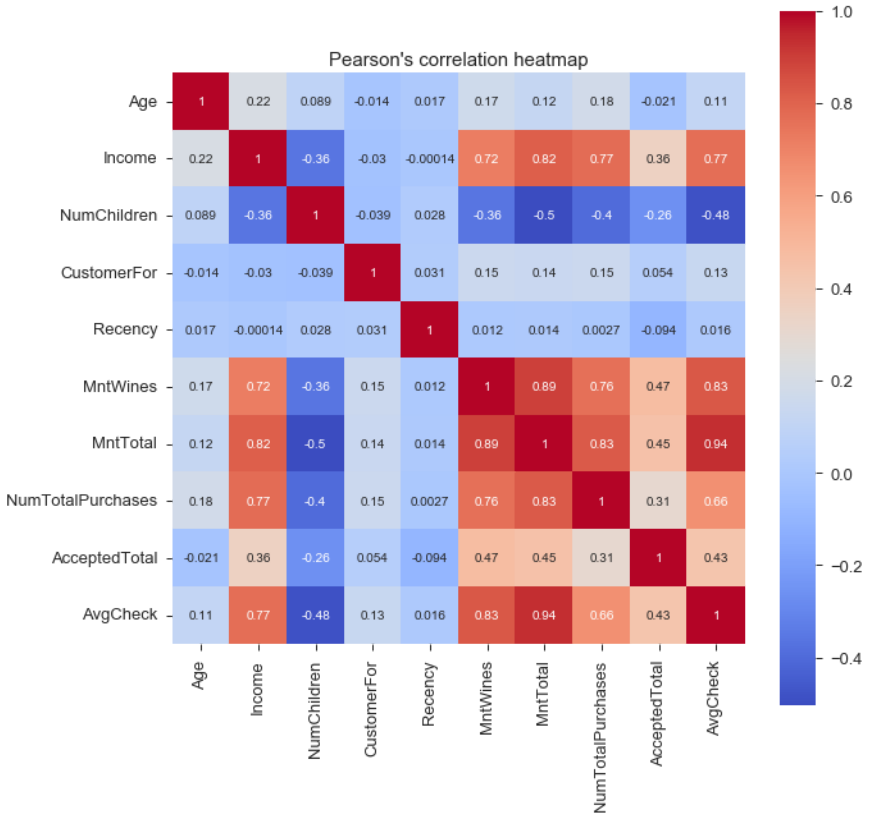

Amount spent on different types of products
Product type
Wine
Fruits
Meat
Fish
Sweets
Gold
50.1%
4.3%
7.2%
4.5%
6.2%
27.7%

Data Cleaning

Avendo individuato outlier e valori non validi, li rimuoviamo al fine di rendere più efficiente l'apprendimento. Otteniamo un dataset più pulito e coerente, anche grazie ad un limitato Feature Engineering per rendere alcuni parametri più facilmente trattabili.

Analisi Multivariata

Al fine di comprendere ancora meglio le relazioni tra le feature in esame, si è analizzata la loro correlazione.
La heatmap per l' Indice di correlazione di Pearson è molto intuitiva →



Data Pre-Processing

Le caratteristiche e il metodo di ridimensionamento sono stati scelti mediante un processo iterativo di valutazione di diverse combinazioni con il punteggio della silhouette.
Il valore della silhouette misura quanto un oggetto sia simile al proprio cluster rispetto ad altri cluster. Questa misura varia da -1 a 1, dove:

- -1 indica che i cluster sono assegnati erroneamente;
- 0 indica che i cluster sono di natura neutra;
- 1 indica che i cluster sono distinti

La formula generica per un singolo coefficiente di silhouette è la seguente:

$$(b - a)/max(a, b)$$

dove:
a: distanza media tra cluster;
b: distanza media più vicina
A questo proposito, la trasformazione Box-Cox è definita come un modo per trasformare variabili dipendenti non normali nei nostri dati in una forma normale attraverso la quale è possibile manipolare meglio i dati.

Clustering

Al fine di fornire una classificazione dei dati è stato utilizzato il metodo più conosciuto, ovvero il clustering. Questa tecnica prevede di partizionare gli esempi in classi dette cluster (da qui *clustering),* e ognuna di queste predice i valori delle feature per i suoi membri. Si ottiene un clustering migliore se al sistema di classi è associato un errore di predizione minimo.
Sono due i tipi di clustering che si conoscono:

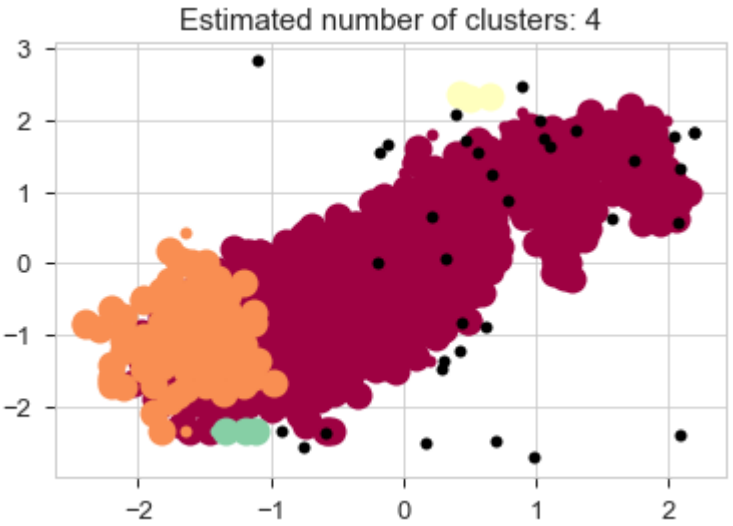
- hard clustering: ogni esempio è assegnato ad una specifica classe e in tal caso questa può permettere di predire i valori dell'esempio in questione;
- soft clustering: un esempio può essere assegnato a più classi e quindi la predizione dei valori viene ricavata attraverso una media ponderata delle predizioni di tutte le classi per quell'esempio.

DBScan

E' stato scelto di partire da questo algoritmo perché, rispetto al classico K-Means, fornisce un clustering spaziale basato su regioni a più alta densità, i cosiddetti *core*. La misura di densità è strettamente connessa alla variazione degli iperparametri:

- eps, controlla la vicinanza locale tra i campioni
- min_samples, numero minimo di campioni "vicini" appartenenti allo stesso core

E' stata effettuata una valutazione del loro valore più adatto sulla base di: numero di cluster prodotti, numero di punti considerati rumorosi, coefficiente di silhouette.
Ciò ha permesso di evincere che il numero ottimale di cluster in questo caso è 4. Tuttavia, piccole variazioni negli iperparametri hanno portato rapidamente al raggiungimento delle condizioni estreme: maggiore min_samples o minore eps indicano la necessità di una maggiore densità per formare un cluster.



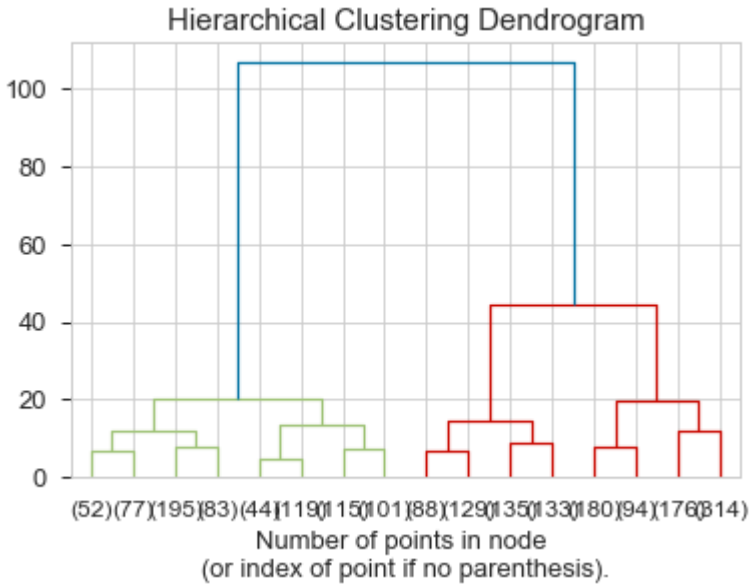
Hierarchical Clustering

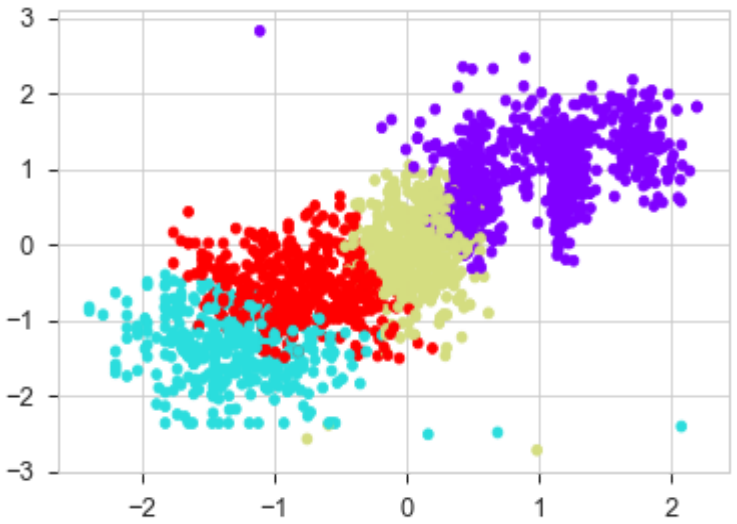
Trattasi di un metodo di analisi dei cluster che cerca di costruire una gerarchia di cluster.

Le strategie in genere sono di due tipi:

- Agglomerative:
  - Inizia con punti considerati come cluster individuali;
  - A ciascun passo, unisci le coppie di clusters più vicini;
  - Fino a quando rimane un solo cluster (o k cluster);
- Divisive:
  - Inizia con un singolo cluster che ingloba tutti i punti;
  - A ciascun passo, spezza un cluster;
  - Fino a quando ogni cluster contiene un punto (o ci sono k cluster) ;
  - E' necessario scegliere quale cluster spezzare ad ogni passo;

Nel caso in questione è stato generato un Dendrogramma che fornisce un' interpretazione più intuitiva dei risultati.
Successivamente, è stato eseguito l'algoritmo indicando dimensione crescente per il numero di cluster; anche in questo caso i risultati hanno portato a scegliere una dimensione compresa tra 4 e 8.





Nella figura, il risultato dell'algoritmo con 4 Cluster.
La misura di similarità utilizzata è quella Euclidea, mentre la distanza tra cluster (Ward Linkage) minimizza la varianza.

K-Means

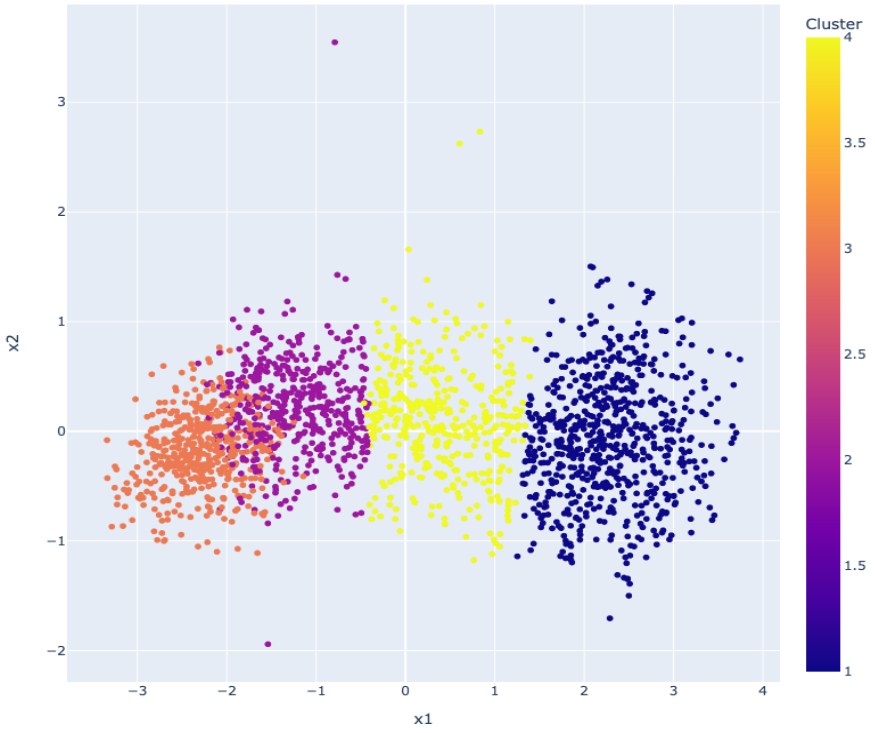
Tecnica algoritmica di apprendimento non supervisionato che trova un numero fisso di cluster in un insieme di dati in cui per ogni cluster si definisce un punto centrale chiamato *centroide*.

In linea generale, segue iterativamente i seguenti step:

1) **Inizializzazione**: si definiscono i parametri di input per eseguire l'algoritmo;
2) **Assegnazione del cluster**: ogni data points viene assegnato al cluster (o centroide) più vicino;
3) **Aggiornamento della posizione del centroide**: ricalcola il punto esatto del centroide e di conseguenza ne modifica la sua posizione.

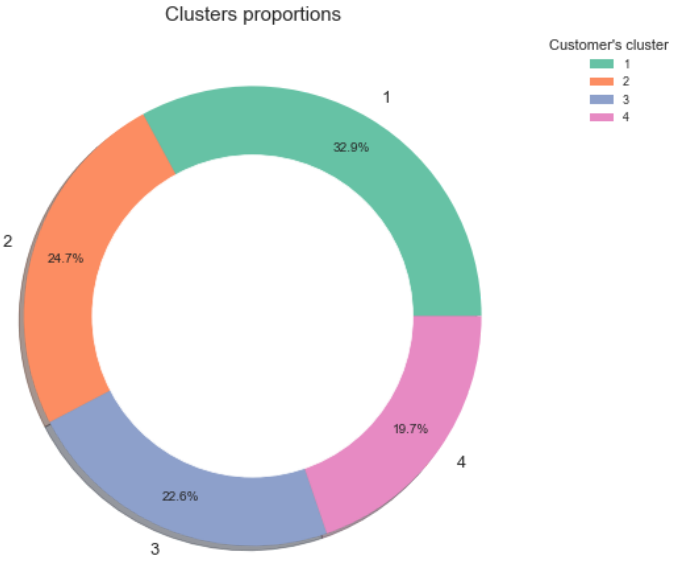
Sono stati confrontati, al fine di individuare il numero ottimale di cluster, il metodo elbow e il punteggio della silhouette. I risultati ottenuti hanno determinato la scelta di utilizzare 4 Cluster, sostenuta anche dalle precedenti esecuzioni con altri algoritmi.

A questo punto, attraverso la Principal Component Analysis, abbiamo potuto visualizzare il set di dati distribuito su un piano. Come si nota in figura, il modello confonde leggermente i Cluster 2 e 3, ma non sembra un grande problema.

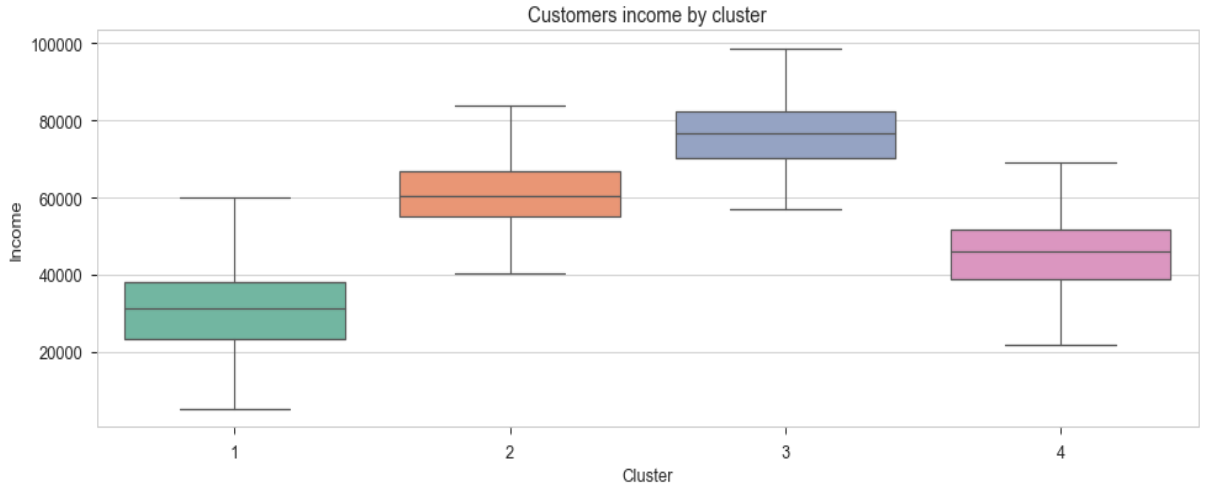


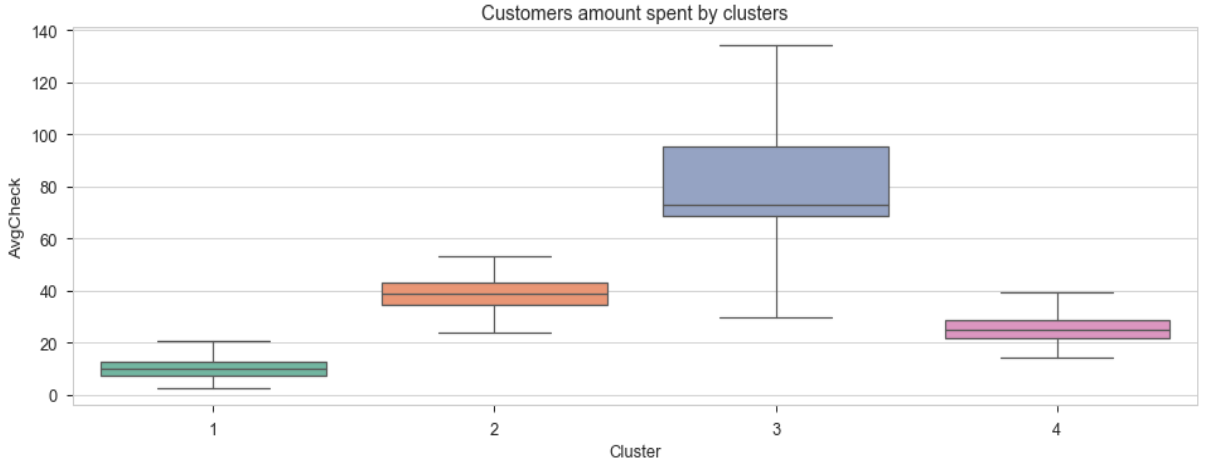
Analisi dei Cluster

A questo punto, applicato K-means al dataset, è iniziata l'analisi dei risultati: interpretare i risultati ottenuti adoperando strumenti algoritmici è importante tanto quanto scegliere il modello e i parametri appropriati per il caso in esame.
L'analisi che segue è il frutto dell'utilizzo di K-means con K = 4 e può essere visionata integralmente attraverso il Notebook allegato.
Riportiamo qui alcuni estratti:



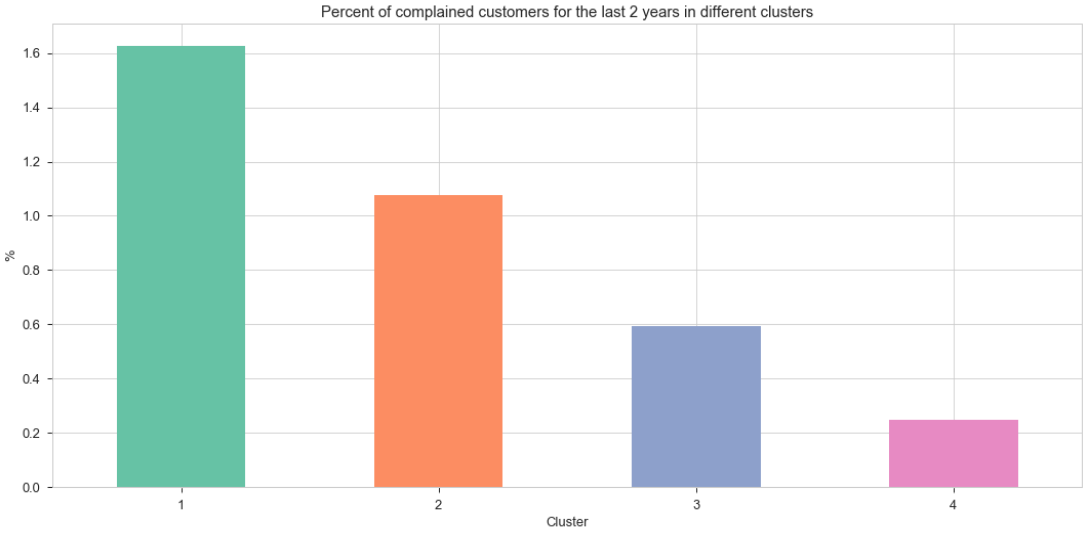
Il grafico a torta mostra come il Cluster 1 rappresenti circa ⅓ di tutti i clienti, mentre nei box-plot seguenti si possono vedere reddito e spesa media dei clienti nei diversi raggruppamenti.





Negli istogrammi seguenti è possibile valutare l'efficacia delle promozioni all'interno di ogni cluster. Come si vede, l'ultima promozione è stata efficace per tutti, mentre la numero 2 è quella che ha riscosso meno successo tra i clienti di tutti i cluster.

Infine, vediamo come la maggioranza dei reclami siano pervenuti da clienti appartenenti al primo cluster, che è anche quello che spende meno in termini di spesa media.



Risultati

Si potrebbero valutare i cluster in questo modo:

- Cluster 1 → clienti Platino
- Cluster 4 → clienti Oro
- Cluster 3 → clienti Argento
- Cluster 2 → clienti Bronzo

Clienti Platino:
- Reddito molto alto
- Average check molto alto
- Acquirenti frequenti
- Per lo più single
- Comprano per lo più da: store e catalogo, ma va bene anche il sito web
- Campagne più di successo: 1, 5, ultima
- Campagne di minor successo: 2
- Di rado visita il sito

Clienti Oro:
- Reddito alto
- Average check alto
- Acquirenti frequenti
- 80% sono genitori
- Comprano per lo più da: store e sito web
- Campagne più di successo: 4, ultima
- Campagne di minor successo: 2, 5

Clienti Argento:
- Reddito medio
- Average check medio
- Media frequenza di acquisto
- Per lo più genitori
- Comprano per lo più da: store e sito web
- Campagne più di successo: 3, ultima
- Campagne di minor successo: 1, 2, 5

Clienti Bronzo:
- Reddito basso
- Average check basso
- Scarsa frequenza di acquisto

- Per lo più genitori
- Ci sono non laureati in questo cluster
- Comprano per lo più da: store e sito web
- Visitano più di tutti il sito
- Campagne più di successo: 3, ultima
- Campagne di minor successo: 1, 2, 4, 5
- Fanno più reclami di tutti