CCT COLEGE DUBLIN

HIGHER DIPLOMA IN SCIENCE IN DATA ANALYTICS FOR BUSINESS

**LÍNIKA VICENTE FERREIRA DE ALMEIDA, ID 2023045**

**INTEGRATED CA – MACHINE LEARNING AND DATA VIZUALIZATION**

**PROFESSOR: DAVID MCQUAID, MUHAMMAD IQBAL**

Dublin, Ireland

2023

**SUMARY**

[1 INTRODUCTION 4](#_Toc152363028)

[2 DATA UNDERSTANDING 5](#_Toc152363029)

[**2.1** **Collect initial data** 5](#_Toc152363030)

[**2.2** **Exploratory Data Analyse (EDA)** 5](#_Toc152363031)

[3. DATA PREPARATION 8](#_Toc152363032)

[**3.1** **Cleaning** 9](#_Toc152363033)

[**3.2** **Transform data type** 9](#_Toc152363034)

[4. MODELLING 9](#_Toc152363035)

[**4.1 Content and Collaborative filtering** 9](#_Toc152363036)

[**4.2 Market Basket** 12](#_Toc152363037)

[**4.3 Dashboard** 13](#_Toc152363038)

FIGURE LIST

[Figure 1- Data Info](#_Toc150102059) 6

[Figure 2- Missing values 6](#_Toc150102060)

[Figure 3- Customer per Gender](#_Toc150102061) 7

[Figure 4 - Common Categories in sales 7](#_Toc150102062)

[Figure 5 - Common Purchased Size………………………………………………………………..](#_Toc150102063)8

[Figure 6 - Purchased per season](#_Toc150102064) 8

[Figure 7- Comparison RMSE ………………………………………………….……………….….](#_Toc150102065)11

[Figure 8 - Apriori and FP-Growth Algorithms………..……………………………………...…](#_Toc150102066) 12

# INTRODUCTION

The project begins with a thorough investigation of data understanding, highlighting the importance of preliminary data gathering and a thorough exploratory data analysis (EDA). The selected dataset was obtained from Kaggle and focuses on consumer shopping trends, providing important information about consumer behavior, preferences, and buying patterns.

The specific procedures for cleaning and data preparation are covered in detail in the next section. Content and Collaborative Filtering, Market Basket, and the development of an interactive dashboard specifically designed for older individuals (65+) comprise the modeling phase.

The comparison of content-based and collaborative filtering methods clarifies the subtle differences between online retail recommendation systems. Using the Apriori and FP-Growth algorithms, the market basket analysis compares support levels, frequent itemsets, association rules, and execution times.

An interactive dashboard specifically designed for older persons is the project's final product, having been developed and implemented. With attributes specific to the target audience, the dashboard summarizes the most important parts of the dataset and explains why machine learning models in an online retail company are a good fit. In order to give consumers 65 years of age and older an easy-to-use interface, the dashboard has been carefully designed and visualized.

*(195 words)*

# DATA UNDERSTANDING

For this project I selected the customer shopping trend dataset, which includes customer purchase habits. This dataset collects vital information for organizations looking to improve their client base understanding by encompassing a variety of features connected to customer purchase habits.

Age, gender, purchase amount, preferred payment methods, frequency of transactions, and feedback ratings are among the features that customers can choose from.

This dataset is useful for the project’s purpose since it has the required characteristic to apply filtering, market basket and create a dashboard.

### **Collect initial data**

The dataset was collected in the Kaggle website. This data provides information about customer purchase habits. And captures a wide range of customer attributes including age, gender, purchase history, preferred payment methods, frequency of purchases, and more.

### **Exploratory Data Analyse (EDA)**

The dataset have 18 columns and the meaning of each column in the data is as below:cust

1. Customer ID

2. Age

3. Gender - (Male/Female)

4. Item Purchased - The item purchased by the customer

5. Category - Category of the item purchased

6. Purchase Amount (USD) - The amount of the purchase in USD

7. Location

8. Size

9. Color

10. Season

11. Review Rating - Rating given by the customer for the purchased item

*(199 words)*

12. Subscription Status - Indicates if the customer has a subscription (Yes/No)

13. Shipping Type - Type of shipping chosen by the customer

14. Discount Applied - (Yes/No)

15. Promo Code Used - (Yes/No)

16. Previous Purchases

17. Payment Method - Customer's most preferred payment method

18. Frequency of Purchases

For achieve the purpose of this project I also added new columns such as Member Number, Item Quantity and Date with random numbers to complement the dataset. In the EDA was done some assumptions to a comprehension of the dataset and check missing values. Was performed the 'head' command to get an overview of the collected data and observe the first rows. Using the info() python method was some details of our features, as if there is Non-null and the type of data.

|  |  |
| --- | --- |
|  |  |
| Figure 1. Data Info | Figure 2. Missing values |

In a previous analysis is possible to notice in data.info this dataset is distributed in 3901 entries and 25 columns. And there are data with different type, being float64, int32, object and datetime64. In df.type was observed there are datetime64(1), float64(5), int32(6), object(14). In isnull().sum() shows the columns presents only one line of missing values in this dataset. According to previous analysis, it seems the dataset contains 1 line with null values in total. So this line will be dropped.

*(213 words)*

For understand the behavior of the customer and explore the dataset, was done some plots as part of the exploratory analysis. Initially in Figure 3 the customer are represented per gender, which 68% represent Male and 32% Female.

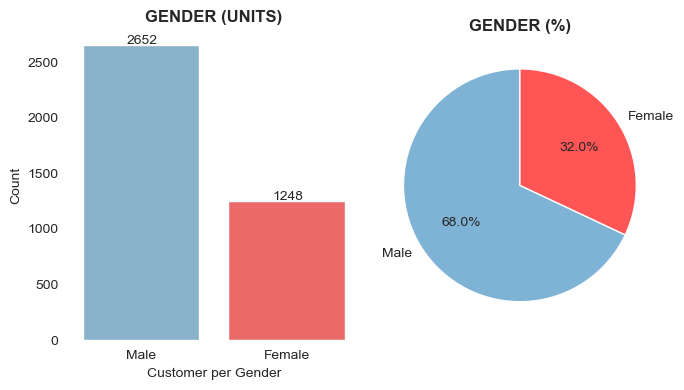


Figure 3 – Customer per Gender

The plot shows the most common category, being in first place of purchased for the customer the Category of Clothing with 44.5%, followed by Accessories, Footwear and in the last position Outerwear with 8.3%.

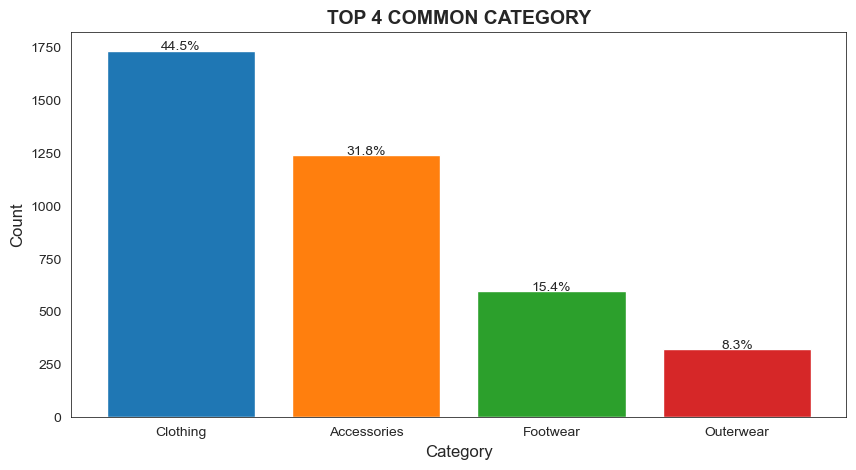


Figure 4 – Common Categories in sales

When analysis what is preference of size in sales, was observed the number one is Medium, followed by Large, Small and in the last position Extra large

*(99 words)*

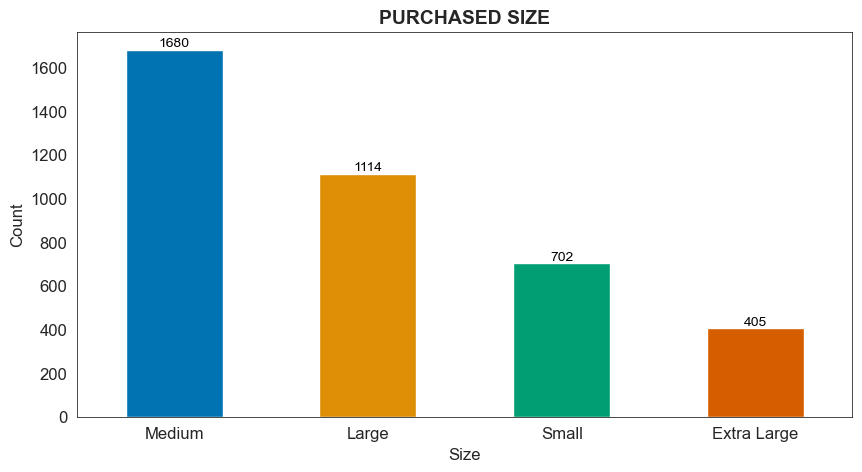


Figure 5 – Common Purchased Size

Customers usually shopping more during the spring and fall season, and during the summer do less shopping

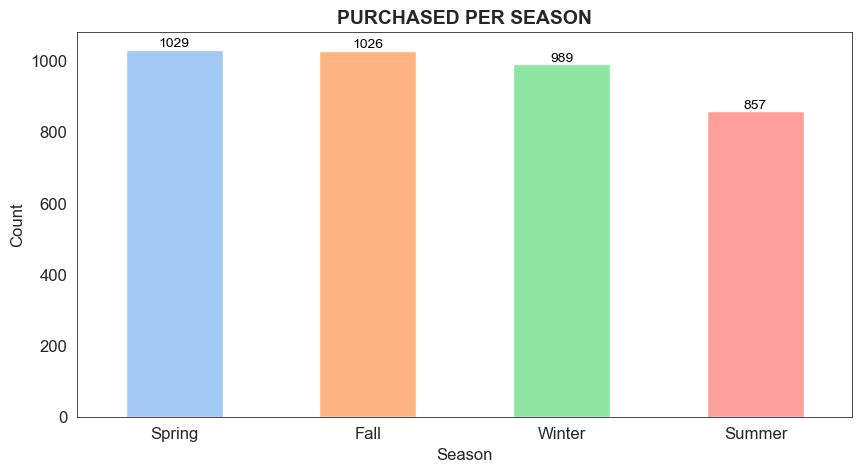


Figure 6 – Purchased per season

# 3. DATA PREPARATION

Data preparation is a crucial step in any data analysis or machine learning task. It entails cleaning and transforming raw data into a format that can be easily analyzed and utilized to train a model. The purpose of data preparation is to guarantee that the data is consistent, accurate, and devoid of errors or missing information (Pro, project 2022.

*(75 words)*

## **Cleaning**

For missing values were used a simple technique of exclusion of the lines with NaN. This approach was applied for being simple and since this dataset has only 19 values of missing values it will not impact the final results.

## **Transform data type**

In this dataset there is categorical, continuous and datetime data. Before perform the models, the categorical data will be transformed in binary data. It avoid misinterpretations, unwanted weightings and it is more compatibility with the model that is being performed. For organize the data type before perform the model fourteen categorical columns were converted in binary type by using Hot encoding.

# MODELLING

The modeling part is made up of Content and Collaborative Filtering, Market Basket and Dashboard.

## **4.1 Content and Collaborative filtering**

*Purpose of a Recommendation System for Online Retail in machine learning*

By utilizing machine learning, online retailers may increase sales, produce better shopping experiences, retain customers, improve customisation, show customers' preferences, and strategically encourage cross- and up-selling. Once the customer profile is known by using machine learning techniques, it can be utilized as a strategy to grow the firm.

Increased sales: the system suggests products that match the user's profile, which raises the likelihood of sales and, in turn, the monthly profits and growth of the online business.

Better User Experience: a pleasing, user-friendly layout encourages customers to look into more products and raises their degree of satisfaction all around.

*(240 words)*

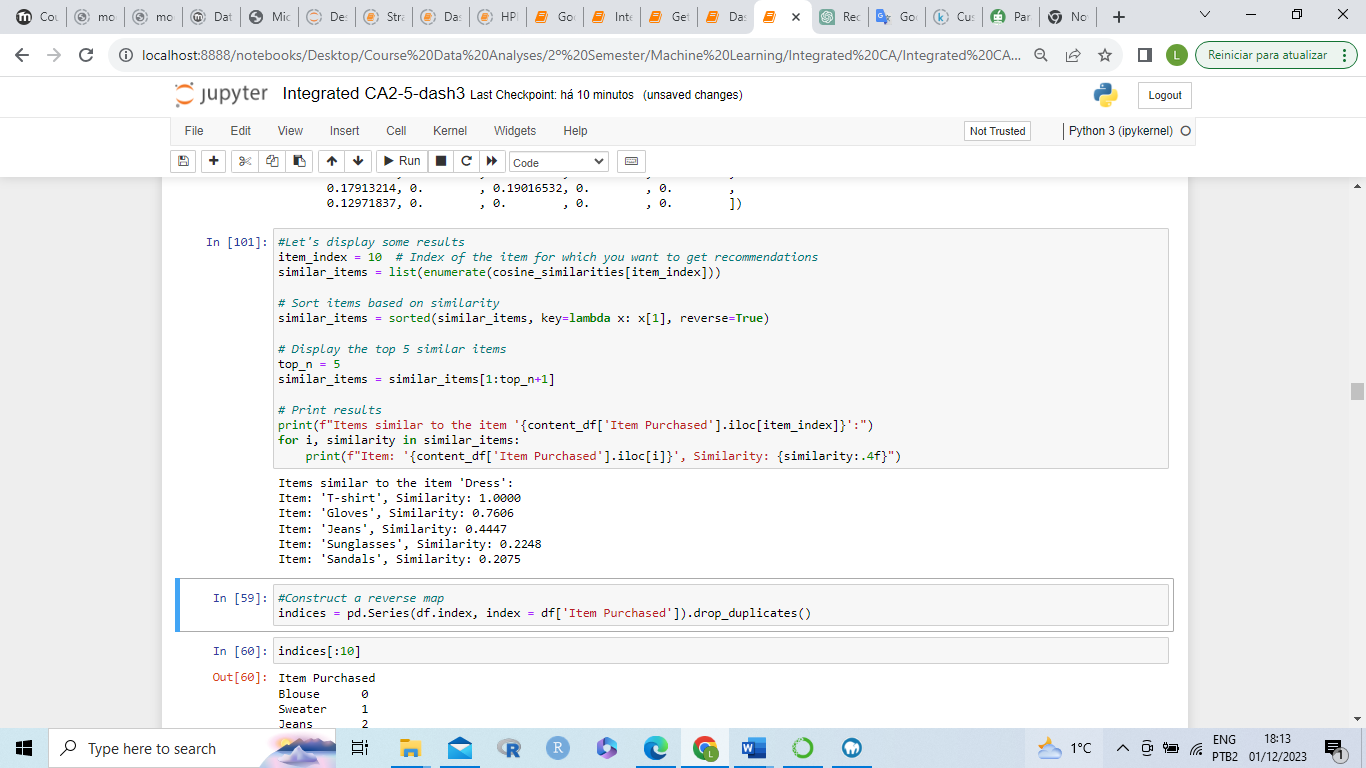
Retention customer: it keeps users engaged and entices them to return to the platform. Recommendation systems encourage repeat business and consumer retention. Furthermore, keeping existing clients generates bigger financial gains than bringing in new ones. Creating a referral program is crucial to fostering client loyalty.

User Preferences: recommendation systems are continually changing to accommodate changing user preferences and activities and catering to customer preferences is crucial.

Cross-Selling and Up-Selling: one important strategie in online retail is cross-selling and up-selling, which involves recommending complementary or superior products which maximizes average transaction value. Personalization: by making product recommendations based on tastes and habits of the custumers, a recommendation engine customizes the user's experience.

*Compare Content and Collaborative filtering and make a conceptual insight*

In the Content-Based Filtering is recommend based on the characteristics of items. It basically uses the attributes of items ('Category', 'Size', 'Color', 'Season', 'Review Rating') to find similar items. Based on that give personalized recommendations considering the individual preferences of the customers. The differentiator is it can handle new items as long as their characteristics are known. As a result of apply content for the item 10 I got the follow recommendations for the user: T-shirt, gloves, jeans and more.



In the collaborative filtering is recommend based on the behavior of similar users. It uses past interactions between users and items to identify behavioral patterns. As result give recommendations influenced by the behavior of similar users. The limitation of that is difficulty handling new items; dependent on the interaction history of similar users. After apply collaborative filtering for the item 10 I got the value 3.75 in a scale of 0-5, it means a review based on the patters behavior of other users about searched item. *(286 words)*

To implement the content-based recommendation system, textual characteristics of the items were used and the data was prepared with collaborative filtering. This consisted of the steps of data preparation, TF-IDF vectorization, cosine similarity calculation, and content-based recommendation generation. And for collaborative filtering, the Surprise library was used.

Were used cosine similarity for user-user and for item-item to check each of them demonstrate higher similarity. Then the model KNN-basic were performed an found Root Mean Squared Error (RMSE). The RMSE quantifies the degree to which the model's predictions and the actual ratings agree.

A numerical indicator of the models' accuracy is provided by the RMSE. Better performance is indicated by lower RMSE values. The better accuracy is indicated by a lower RMSE value (0.1396) when compared to the User-User model (0.7722). In comparison to User-User Collaborative Filtering, Item-Item Collaborative Filtering appears to perform better on this specific dataset, as indicated by its reduced RMSE.

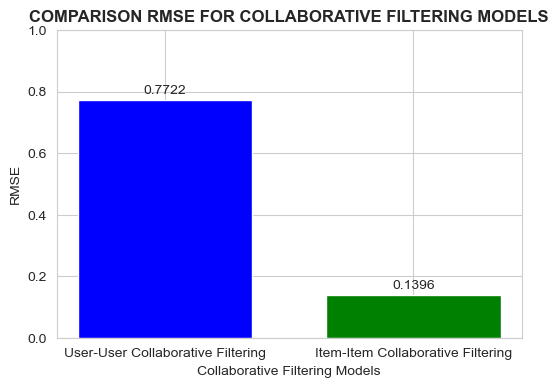


Figure 7 – Comparison RMSE

After identify the Item-Item perform better than user-ser, I considered Item-Item for the collaborative content and collaborative filtering.

## **4.2 Market Basket**

The major difference between Apriori and FP growth algorithms are both can

*(185 words)*

produce the same result, being the mainly difference is FP-growth is more efficient and scalable for large datasets. However, the choice between the two algorithms depends on factors such as dataset size, sparsity, and available memory.

Using a level-wise search approach, Apriori locates all frequent itemsets of length 1 first, then 2 and so on, until no more frequent itemsets are discoverable. However do not perform well in large datasets. Apriori might need additional RAM, particularly when the dataset is bigger and there are more candidate itemsets.

In contrast, FP-growth employs a divide and conquer strategy. In order to compress the transaction database and increase its memory efficiency, it builds an FP-tree. This FP-tree is then mined recursively to produce the frequent itemsets. In general, FP-growth outperforms Apriori. FP-growth is useful for perform big datasets because the FP-tree's compact representation eliminates the need to explicitly store several candidate itemsets, FP-growth is more memory-efficient.

After apply both algoritmical I got the results obtained from the Apriori and FP-Growth algorithms:

|  |  |
| --- | --- |
|  |  |
|  |  |

Figure 8 – Apriori and FP-Growth Algorithms

When compare both algorithms in terms of support level were observed Apriori ranges (7.77% to 47.61%) and FP-Growth ranges (7.85% to 51.10%). FP-Growth has higher support for some itemsets.

Checking top frequent itemsets I saw Apriori identifies {S}, {a} as the most

*(210 words)*

frequent, while FP-Growth identifies {s}, {e} as the most frequent. Differences in itemsets due to the algorithms' approaches.

Concerning the association rules both algorithms generate similar association, but the specific itemsets and their supports differ. Slight variations in rules but general patterns are captured. In the execution time the Apriori was faster being 0.04s while FP-Growth 0.07s.

In conclusion both algorithms provide valuable insights into frequent itemsets and association rules. Apriori is straightforward and provides clear itemsets, while FP-Growth is more efficient for large datasets. Consider the specific requirements, execution time, and interpretation ease when choosing between the two algorithms.

## **4.3 Dashboard**

This dataset is suitable for machine learning in an online retail business because it likely contains a variety of information, such as purchase amounts, item details, locations, and dates. This richness allows for a comprehensive analysis of customer behavior, preferences, and trends.

Utilizing Widget-Based Interaction allows for interactive data exploration. Such interaction can help machine learning models by enabling users to fine-tune and modify inputs for instantaneous insights and predictions. Customized Customer Data Through the comprehension of consumer preferences, trending products, and buying behaviors, machine learning models may aid in focused advertising campaigns, customized suggestions, and enhanced inventory control.

Considering this dashboard was designed for older adults (65+) the follow adjusts were made to make a friendly experience:

Font Size and Readability: the usage of a higher font size (24+) guarantees improved readability and a more user-friendly experience. It is important to take older persons who may have visual problems into mind. An interface that is simple to use and traverse is enhanced by the design of widget size and visibility, which includes sliders and selection options. The layout of the widgets is made to be visible and accessible, and the options and descriptions are clear.Minimal complexity the selection options concentrate on the most important details of the data, like the location, the item(s) bought, the purchase amount, and the date. *(322 words)*

Reducing the quantity of features and options aids in preventing cognitive overload in senior users. The distribution of purchase amounts is shown graphically through the use of a bar chart in visual representations. Even for consumers who might prefer visual information over written information, visualizations can be an efficient way to swiftly deliver information.

Real-Time Feedback this responsiveness provides immediate feedback, helping users understand the impact of their selections. Contextual instructions of clear instructions or tooltips, although not explicitly mentioned in your code, can be beneficial for guiding users through the dashboard, especially if they are not familiar with the interactive elements.

*(102 words)*