

## **HDIP. DATA ANALYTICS**

*INTEGRATED CONTINUOUS ASSESSMENT*

***DATA VISUALISATION***

***AND MACHINE LEARNING***

**Assignment Type:** Individual

**Assignment Title:** Integrated CA2\_DVT\_MLB

**Lecturer(s):** David McQuaid & Dr. Muhammad Iqba

**Issue Date:** 19th October 2023

**Submission Deadline Date:** 1st December 2023

**Students:** Bárbara Abreu Costa 2023099

SUMMARY

[1. Introduction 03](#_2vtno5fy1g9e)

[2. Project Scope and Objectives 04](#_vo3vdg4jvh2c)

[3. Data overview](#_l4s1hkaz1cmk) 05

[4](#_clnwpmii2t5)[.](#_72q4od192mai) Recommendation Systems07

6. [Market Basket Analysis](#_72q4od192mai)……………………………………………………09

7. Interactive Dashboard11

9. References12

# 1. Introduction

The present research provides the findings from Bárbara Abreu Costa's Integrated Continuous Assessment, which analysed the ever-evolving field of Machine Learning applications in the online retail industry. In order to improve consumer experiences and promote company growth inside e-commerce platforms, the project delves into the complex workings of recommendation algorithms, market basket research, and the building of customised interactive dashboards in the context of online shopping. Having that as the main goal of this project, I hope to give a thorough knowledge of these approaches, their uses, and their importance in maximising business results and user experiences through this investigation.

Firstly, this project involved the selection of a data set, carrying out exploratory data analysis, pre-processing the data and implementing machine learning algorithms that were essential for presenting the results effectively.

The "*SuperStore*" dataset, which is an extensive collection of synthetic retail data that has been meticulously generated to simulate sales, customer behaviour, and product performance inside a retail context, serves as the foundation for this project. I have chosen the "*SuperStore*" dataset because it provides insightful information on the complexities of a retail business by covering an extensive range of data.

The dataset contains information about orders, customers, products, sales, shipping details, and more.

Here are the key features included in the "*SuperStore*" dataset:

* **Order Details:** Order IDs, order dates, shipping dates, and quantities.
* **Product Information:** Product IDs, names, categories, prices, and quantities sold.
* **Customer Details:** Customer IDs, names, locations, and segments (like corporate, home office, consumer).
* **Geographical Information:** Regions, states, and cities where the store operates.
* **Sales Information:** Revenue, profits, discounts, and shipping costs.

# 2. Project Scope and Objectives

The project will involve an in-depth exploration of three core aspects:

The ***Recommendation Systems*** section delves into the purpose of recommendation systems in an online retail setting, comparing Content-Based Filtering and Collaborative Filtering. The application of machine learning models for collaborative filtering will be demonstrated, accompanied by a conceptual rationale justifying their relevance within the chosen scenario.

In sequence, the ***Market Basket Analysis*** will focus on implementing Apriori and FP Growth algorithms on the selected dataset. A detailed comparison between these methodologies will elucidate their divergences, while the machine learning results derived from both algorithms will be compared and evaluated.

And finally crux of this section an ***Interactive Dashboard*** for Older Adults (specifically aged 65+) will summarise critical aspects of the dataset, elucidating its suitability for Machine Learning models. Emphasis will be placed on the dashboard's tailored design, taking into account the unique needs and preferences of this demographic.

This report aims to thoroughly explore and study three key areas. By doing so, it seeks to show how high-tech methods in computer learning are changing online shopping. These methods are crucial in making customers more involved and boosting how well businesses do.

# 3. Data overview

As you have already seen in the *Introduction* chapter, the dataset presents a diverse array of categorical, geographical, and numerical variables that encompass 9994 rows and 21 columns. Below I have printed a dictionary of all those attributes followed by their own meanings.



3.1 Exploratory Data Analysis

An initial crucial step involved ensuring data quality through comprehensive cleaning procedures.This included looking for duplicates, missing values, and various attribute data types. Fortunately, the dataset was found to be clean, devoid of missing values or duplicates, and consisting of a mix of numerical, categorical, and datetime data types.

After the data cleaning stage, the emphasis switched to using descriptive statistics and visualisations to comprehend the nature of the dataset.Irrelevant columns like 'Row ID', 'Postal Code', and 'Ship Mode' were dropped to streamline the dataset for analysis. The distribution and features of numerical properties, such as sales, quantity, discount, and profit, were explained by descriptive statistics. Deeper insights were uncovered during the next visualisation step, which included identifying top-selling cities, states, and profitable categories in addition to showing the number of items across regions and categories.

Temporal analysis played a crucial role in understanding sales trends over time. To make this analysis easier, date-related characteristics were converted and the year and month data were extracted. After that, visualisations were created to show sales trends over time, by month, and year over year. These revealed seasonal patterns and variances in sales.

The analysis culminates in pre-processing steps where the dataset was divided into training and testing subsets, which was essential for the next step. This establishes a solid foundation for future exploratory analyses such as the application of machine learning models for predictive analysis and decision-making processes.

# 

# 5. Recommendation System

5.1 Purpose

Recommendation systems play a pivotal role in the success of online retail businesses by leveraging machine learning algorithms to deliver personalised product suggestions to users. It analyses a user's historical behaviour, preferences, and interactions with the platform to offer product suggestions tailored to their individual tastes and needs. This personalization creates a more engaging and relevant shopping experience.

5.2 Content Based Filtering

Content-based filtering recommends items to users by analysing the attributes or features of the items themselves and matching those attributes with the user's preferences. For instance, recommending movies based on their genres, actors, directors, or plot keywords.

5.3 Collaborative Filtering

Collaborative filtering recommends items to users by leveraging the preferences or behaviours of other users. It identifies patterns of interests among users and predicts what a user might like based on the preferences of users with similar tastes.

5.4 Applying Recommendation System

To begin, Content-Based Filtering was employed to generate recommendations based on item characteristics. This approach involved preprocessing the dataset by amalgamating various columns like "Category," "Sub-Category," and "Product Name" to form a comprehensive product description. This descriptive representation was transformed into numerical vectors using TF-IDF Vectorization, allowing the quantification of textual data. Subsequently, the cosine similarity metric was employed to calculate similarities between products, aiding in the identification of items with similar descriptions. The outcome was a system capable of recommending products akin to the user's preferences based on similar product descriptions.

Concurrently, Collaborative Filtering, encompassing both User-Based and Item-Based methods, was implemented to harness user behaviour for recommendation purposes. The User-Based Collaborative Filtering technique involved constructing a user-item matrix from the dataset, computing user similarities using cosine similarity, and recommending items based on similar user preferences. This method enabled the system to suggest items to users that were favoured by other users with similar purchase histories. The Item-Based Collaborative Filtering, on the other hand, computed item similarity using cosine similarity on the transposed user-item matrix. This approach suggested items similar to a chosen item based on user purchase behaviour, facilitating recommendations for customers who had purchased similar items.

5.3 Conceptual Insights

Throughout the exploration, the conceptual insights gleaned underscored the strengths and nuances of each approach. Content-Based Filtering proved advantageous in recommending items with similar attributes or characteristics, particularly beneficial when user-item interactions were limited. Meanwhile, Collaborative Filtering, whether User-Based or Item-Based, excelled in offering personalised recommendations by leveraging user behaviour data. The synergy between these methodologies was highlighted, showcasing their potential in enhancing user experience and engagement within an online retail environment.

The exploration of content-based and collaborative filtering methods showcased their distinct approaches in generating recommendations for online retail. While content-based filtering leverages product attributes, collaborative filtering delves into user-item interactions. Employing both methods in a hybrid system could provide comprehensive and accurate product recommendations tailored to user preferences, ultimately contributing to enhanced user satisfaction and increased sales.

# 

# 

# 6. Market Basket Analysis

This indispensable approach examines transactional data to reveal correlations between products that are frequently purchased together. By employing sophisticated algorithms like Apriori and FP-Growth, MBA provides valuable insights into customer behaviour, facilitating personalised marketing strategies, inventory management, and improved decision-making processes for companies across multiple industries.

The Apriori algorithm, a classic approach in MBA, was applied first. It began by transforming the dataset into a format suitable for the algorithm, creating baskets that aggregated the quantity of each product purchased in each order. It then encoded these baskets into a binary format, setting each item's presence or absence in a transaction. This step was pivotal as it allowed the algorithm to identify frequent itemsets, which are groups of items that often co-occur above a defined threshold (min\_support). The association\_rules function further derived rules based on the identified itemsets, considering parameters like confidence and lift to filter out less significant rules. These rules indicated which items were likely to be purchased together, offering insights into customer behaviour in different regions such as 'South' and 'Central'.

FP-Growth algorithm was implemented subsequently, introducing a different approach to market basket analysis. Instead of generating the itemsets explicitly as Apriori does, FP-Growth constructs a compact data structure known as an FP-tree. This tree effectively represents the transactional database, allowing for efficient mining of frequent itemsets without explicitly generating all possible itemsets, thus speeding up the process. As Apriori, FP-Growth also identifies frequent itemsets based on a specified minimum support threshold and deriving association rules. The generated rules revealed associations among items, just like Apriori, but FP-Growth's efficiency lies in its ability to handle larger datasets more effectively by compressing information into the FP-tree structure.

The resulting association rules from both algorithms shared similarities in terms of revealing item associations based on support, confidence, lift, and other metrics. However, the significant difference lies in their underlying methodologies. Apriori generates itemsets by creating a candidate set and pruning infrequent itemsets iteratively, while FP-Growth constructs an FP-tree to streamline the process by avoiding the generation of individual itemsets upfront.

In summary, while both Apriori and FP-Growth aim to uncover frequent itemsets and associated rules in market basket analysis, their fundamental approaches differ significantly. Apriori explores itemsets explicitly, while FP-Growth employs a tree-based structure for more efficient and scalable mining, especially in larger datasets.

7. Interactive Dashboard

As we delve into the realms of data interpretation and decision-making, the creation of user-friendly, insightful dashboards becomes paramount, especially when catering to specific demographics with unique needs and preferences. This chapter focuses on creating an interactive dashboard in the context of an online retail firm that is specifically designed for older adults - aged 65 and over.

7.1 Dashboard Design Considerations

Its design prioritised simplicity, better readability, guided navigation and accessibility features, ensuring a seamless experience. The considerations outlined below shaped the fundamental essence of this dashboard, ensuring it met the specific needs of the demographic in question.

***Simplicity & Uncluttered Layout:*** The design ethos centres on a clean interface that avoids overwhelming users with unnecessary complexities. By prioritising essential information, the dashboard maintains a straightforward and uncluttered layout.

***Enhanced Readability with Clear Elements:*** Larger fonts, icons, and a clear colour scheme work harmoniously to facilitate effortless comprehension. High-contrast colours improve visibility, ensuring crucial information is easily discernible.

***Intuitive Navigation & Limited Complexity:*** An uncomplicated navigation system allows straightforward access to critical information, minimising complex interactions. Prioritising easy-to-use interfaces helps people who are less experienced with technology navigate the learning curve more quickly.

***Guided Experience & Accessibility Features:*** Clear, concise explanations guide users through the dashboard, aiding in effective navigation and interpretation. Additionally, the implementation of accessibility features caters to potential visual impairments and ease-of-use concerns, ensuring inclusivity for all users.

(1860 words)

# 8. References

Bittar, I. (2023). 'Unlocking the Power of Visual Design in Dashboards', Mastering Aesthetics in Analytics: Crafting Visually Impactful Reports. Available at: <https://bootcamp.uxdesign.cc/> [Accessed 25 Nov. 2023].

Charming Data Youtube Channel - Design my Dash App Layout. Available at: <https://www.youtube.com/watch?v=1nEL0S8i2Wk> [Accessed 25 Nov. 2023].

Creating Dashboards in Python with Dash and Plotly. Available at: <https://www.youtube.com/watch?v=aS64PvDqCbU&t=29s> [Accessed 26 Nov. 2023].

Dash Bootstrap Components, Layout - Components for laying out your Dash app, including wrapping containers, and a powerful, responsive grid system. Available at: <https://dash-bootstrap-components.opensource.faculty.ai/docs/components/layout/>

[Accessed 28 Nov. 2023].

Domo, Inc. (2023). Nine Reasons Why Dashboard Design Is Essential. Available at: <https://www.domo.com/learn/article/nine-reasons-why-dashboard-design-is-essential>

[Accessed 25 Nov. 2023].

Towards data science | Hands-On Content-Based Recommender System using Python. Available at: <https://towardsdatascience.com/hands-on-content-based-recommender-system-using-python-1d643bf314e4> [Accessed 28 Nov. 2023].

NR Hosting | How to Connect HTML CSS with Python? Available at: <https://www.nrhosting.com/how-to-connect-html-css-with-python>. [Accessed 25 Nov. 2023].

Recommendation System Using Cosine Similarity | Kaggle. Available at: <https://www.kaggle.com/code/muhammadayman/recommendation-system-using-cosine-similarity> [Accessed 17 Nov. 2023].

Plotly Open Source Graphing Libraries | Built-in Continuous Color Scales in Python. Available at: <https://plotly.com/python/builtin-colorscales/#continuous-color-scales-in-dash> [Accessed 25 Nov. 2023].