Table of Content

No.	Content	Pages
1.	Executive Summary	1-2
2.	Summary of Project Context and Objective	3
3.	Methodology	4-9
4.	Result and Conclusion	10-23
5.	Conclusion	24
6.	Reflection	25-26
7.	Reference	27
8.	Appendix	-

1.0 EXECUTIVE SUMMARY

1.1 Description of the selected project

Shopping is a process in which consumers explore and evaluate products or services offered by merchants, with the intention of purchasing an assortment that meets their needs. It can take place either online or in a physical retail store. The younger generation tends to favor online shopping due to its convenience and time-saving benefits. With just a few clicks, customers can browse products, make a purchase, and have it delivered to their doorstep. However, many people still prefer the traditional shopping experience of visiting a store and physically examining the products before making a purchase.

Shopping is not limited to buying groceries or clothing, but encompasses a wide range of consumer needs. Supermarkets and retail stores offer a diverse selection of goods, including clothing, electronics, home decor, and daily essentials. In recent years, new solutions have emerged to make shopping even more convenient. For example, Grab and Foodpanda, which originally specialized in food delivery and transportation, have expanded their services to include the delivery of groceries and other essentials. Thanks to these platforms, consumers can now shop online and have their purchases delivered to them quickly and easily.

1.2 Problem to be solved

The shopping in our case is to see the preference of customers when they are shopping. From their preference, we can determine what products we should increase our sales on and what kind of new products we should produce or recommend to the customers. Other than that, we would also need to ensure that the customers use the promo codes that are given by the store to increase the in-store sales or online sales as these promo codes are the ones that attract customers to the store. This is because we are able to understand customer preferences. When we are able to understand customer's preferences, we will be able to increase the overall customer experience.

1.3 Basic Description of Data Selected

The dataset of consumer behaviour and shopping habits is from Kaggle. The table consists of 18 columns and 3900 rows.

Column	Description
Customer ID	A unique identifier for each individual customer.
Age	The age of the customer.
Gender	Gender identification of customers.
Item Purchased	The specific item selected by the customer during a transaction.
Category	The classification of which group the purchased item belongs to.
Purchase amount (USD)	The cost of purchased item(s) in USD.
Location	The geographical location where the purchase was made.
Size	Size specification of the purchased item.
Colour	Colour of the purchased item.
Season	Seasonal relevance of the purchased item.
Review Rating	Qualitative assessment provided by customers about their satisfaction with the purchased item.
Subscription Status	Indicate if the customer has opted for a subscription service.
Shipping Type	The method used to deliver purchased items.
Discount Applied	Any promotional discounts are applied to the purchase.
Promo Code Used	Notes whether promotional code is utilized during a transaction.
Previous Purchases	The number of prior purchases made by the customer.
Payment Method	The mode of payment used by the customer.
Frequency of Purchases	How often the customer engages in purchasing activities.

2.0 SUMMARY OF PROJECT CONTEXT AND OBJECTIVE

2.1 Summary of the Project Context

The primary objective of this research is to develop a comprehensive comprehension of customer behaviour by analysing a synthetic dataset called the "Customer Shopping Preferences Dataset." The dataset contains a range of client characteristics, such as age, gender, purchase history, and preferred payment methods, offering a comprehensive perspective on consumer preferences. The initial goal entails conducting a comprehensive analysis of the dataset, utilising descriptive statistics and visualisations to reveal patterns and correlations within the data.

Transitioning from exploration, the project utilises clustering techniques to divide customers into segments according to their distinct shopping inclinations. The purpose of this segmentation strategy is to identify discrete client segments, allowing organisations to customise marketing strategies for enhanced personalisation and more impactful engagement.

In addition, the research explores predictive modelling by utilising a decision tree algorithm to anticipate if clients are likely to use promotion codes throughout their purchasing activities. The capacity to forecast outcomes can have a profound effect on businesses' promotional efforts and allocation of resources.

Overall, this project is able to empower organisations with the necessary skills to make well-informed judgements, optimise marketing endeavours, and eventually improve the entire customer experience.

2.2 Objectives

The objectives of our project are as follows:

1. To conduct a detailed exploration and understanding of the Customer Shopping Preferences Dataset.
2. To analyse customer segmentation using clustering based on the similarities of their shopping preferences.
3. To predict the likelihood of a customer using a promotion code using a decision tree.

3.0 METHODOLOGY

3.1 Tools used

1) Pandas

   Pandas are used to analyze, clean, explore and manipulate data. It allows us to analyze big data and make conclusions based on statistical theories, Pandas clean messy data and make them relevant.

2) Numpy

   Numpy is a library for multi-dimensional array and matrix processing. It is capable of handling linear algebra, and Fourier transforms and allows users to manipulate matrices to easily improve machine learning performance.

3) Matplotlib

   Matplotlib is focused on creating beautiful graphs, plots, histograms and bar charts. It can plot data from SciPy, NumPy and pandas. It is also a comprehensive library for creating static, animated and interactive visualizations in Python.

4) Seaborn

   Seaborn is an open-source Python library for statistical graphics plotting. It has default styles and colour palettes to make statistical plots more attractive. The variables can be completely numerical or a category like a group, class, or division.

5) Scikit-learn

   Scikit-learn is a machine learning library built on Numpy and Scipy. Scikit-learn can support most supervised and unsupervised learning algorithms besides those that could be used for data mining, modelling, and analysis.

6) SciPy

   Scipy is a scientific computing library built on top of NumPy. The acronym for Scientific Python is SciPy. It offers additional statistical, signal processing, and optimization utility functions.

7) Kneed

Kneed is a Python package that can be used to detect the knee or elbow point by attempting an implementation of the Kneedle algorithm.

8) Lazy Predict

Lazy Predict is a Python library that offers a quick and easy prediction method. Lazy Predict is a great tool for predictive modelling.

3.2 ETL Pipeline of Preprocessing

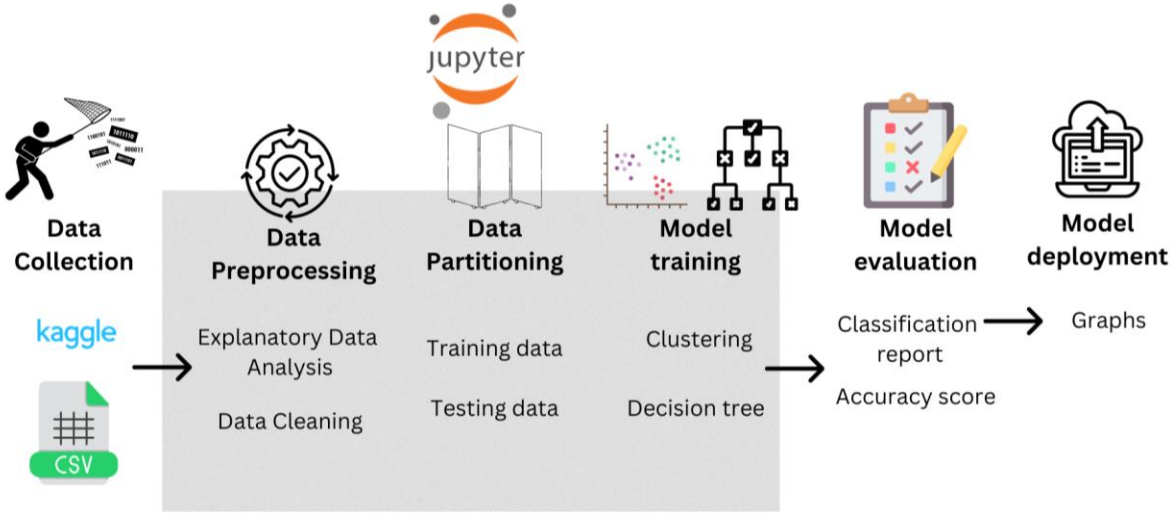


Figure 3.1 ETL Pipeline

The extract, transform and load (ETL) process is applied in this data preprocessing before moving on to modelling. The first phase is extracting the dataset "E-commerce Transaction Trends: A Comprehensive Dataset" from Kaggle and reading it in the Jupyter Notebook. In this phase, the pandas library was used to extract the dataset.

The next phase is about data transformation. This phase has several steps, including dropping unrelated columns, label encoding, one-hot encoding, and standard scalar. The "customer id" column was dropped because it didn't bring values to the final result. Label encoding is applied in both numerical and categorical columns such as '"Age", "Review Rating", "Category", "Location", and so on. Label encoding is an approach that converts the categorical columns into a numerical form using the sklearn library.

Moreover, one-hot encoding, known as dummy encoding, was also used in this phase. The technique separates category values into categorical columns and gives a binary value of either 1 or 0 to these columns. For instance, the data of the categorical column for "Item Purchased", such as blouse, sweater, jeans, and so on, will separate into another column with binary value. The standard scalar was also applied in the numerical column to standardise the feature scales. The numerical column that standardised includes "Gender", 'Subscription Status', 'Discount Applied', and 'Promo Code Used'. After data transformation, we carry out exploratory data analysis on the impact of gender on item purchase of backpack.

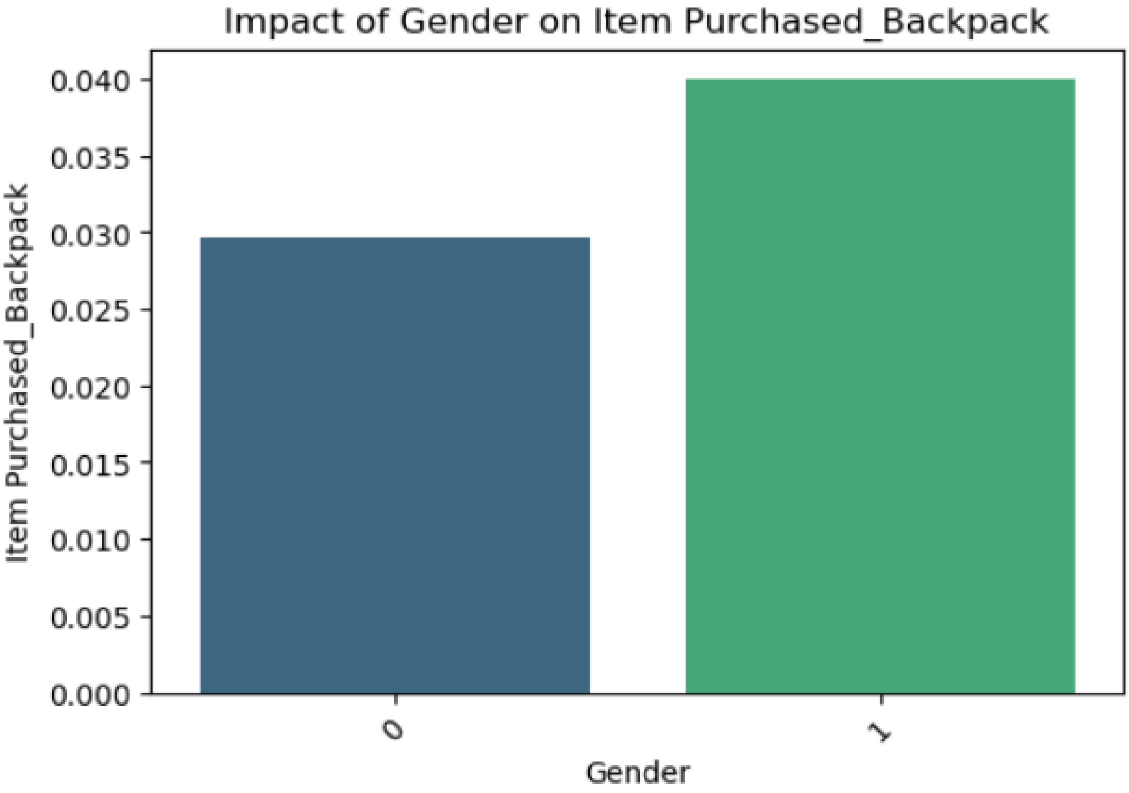


Figure 3.2 Impact of Gender on Item Purchased Backpack

From figure 3.1, 0 stands for female and 1 stands for female. We can see that the backpack is purchased more by male compared to females. It is evident that male exhibit a higher propensity to purchase the rucksack as compared to females. It is presumed that the masculine individual acquired a backpack for utilitarian purposes, such as transporting work-related materials, gym attire, or personal possessions.

3.3 Modelling

3.3.1 K-Means

K-means clustering is a widely used unsupervised machine learning approach that partitions n observations into K clusters by vector quantization, where each observation is assigned to the closest mean or centroid. The objective is to minimise the sum of squared distances between the data points and their respective cluster centroids, leading to internally homogenous and distinguishable clusters. Before using K-Means, PCA was used to diminish the feature space of the dataset data to 2 principal components. Data preprocessing is a crucial step that streamlines the dataset and can enhance clustering accuracy by eliminating noise and extraneous characteristics.

3.3.2 Hierarchical Clustering

Hierarchical clustering is a commonly used technique for categorising things, resulting in the formation of distinct groups graphically shown in a dendrogram. A dendrogram is a tree-like diagram that documents the order of mergers or divisions. The dendrogram visually depicts data points, where each leaf symbolises an individual data point, and each node represents a cluster. The vertical dimension of each node corresponds to the measure of dissimilarity between the merged clusters. In this case, the agglomerative hierarchical clustering technique was applied using the Scikit-Learn library.

3.3.3 Decision Tree

A Decision Tree is a popular non-parametric supervised learning technique for applications involving classification and regression (Saini, 2021). An example of a hierarchical model used in decision support is a decision tree, which shows choices, possible outcomes, chance occurrences, resource costs, and utility. Its internal nodes, leaf nodes, branches, and root nodes make up its hierarchical tree structure, as shown in Figure 3.1.

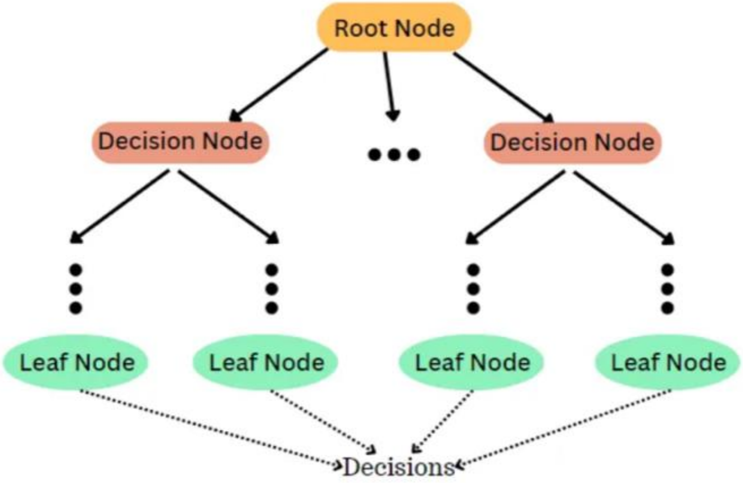


Figure 3.3 Decision Tree (Source: Nidhi, 2023 retrieved from https://medium.com/@nidhigh/decision-trees-a-powerful-tool-in-machine-learning-dd0724dad4b6)

3.4 Evaluation

3.4.1 Accuracy

According to (Gad, 2020), accuracy is a comprehensive indicator that indicates the overall performance of the model across all classes. Also, accuracy is a quantitative measure that evaluates the frequency with which a machine learning model accurately predicts the result. Accuracy may be computed by dividing the count of correct predictions by the entire count of predictions. The formula for the accuracy is shown below:

$$Accuracy = \frac{(TP + TN)}{(TP + TN + FP + FN)} \qquad (3.1)$$

3.4.2 Precision

Precision is sometimes referred to as the positive predictive value (PPV) (Sharma, 2023). The accuracy is determined by dividing the number of properly categorised positive samples by the total number of samples classified as positive, regardless of whether they were classified correctly or wrongly. The precision metric quantifies the model's ability to accurately categorise a sample as positive (Gad, 2020). Precision may be quantified on a range ranging from 0 to 1, or expressed as a percentage. Greater accuracy yields superior results. The formula for precision is shown below:

$$Precision = \frac{TP}{(TP + FP)} \qquad (3.2)$$

3.4.3 Recall

Recall, also known as sensitivity or true positive rate, refers to the ability of a model to identify positive instances correctly. The recall metric quantifies the model's capacity to identify positive samples accurately (Gad, 2020). As recall increases, the number of positive samples found also increases. The recall is determined by dividing the number of Positive samples properly categorised as Positive by the total number of Positive samples. The formula for the recall is shown below:

$$Recall = \frac{TP}{(TP + FN)} \qquad (3.3)$$

3.4.4 F1-score

The F1 score quantifies the harmonic mean of accuracy and recall (Sharma, 2023). The F1 score is often used as an assessment measure in binary and multi-class classification tasks. It combines recall and precision into a single metric, providing a more comprehensive evaluation of the model's performance. The F1 score is a statistic that incorporates both accuracy and recall, representing both symmetrically in the formula:

$$F1 - score = 2(\frac{(Precision)(Recall)}{Precision + Recall}) \qquad (3.4)$$

4.0 RESULT AND DISCUSSION

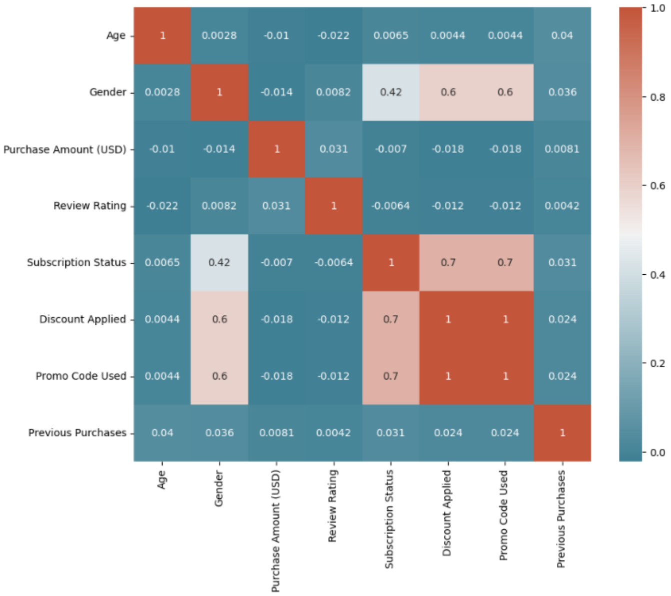


Figure 4.1 Correlation Matrix

From figure 4.1, we can see that the highest correlation is red and the lowest correlation is dark blue. For discount applied and promo code used it shows the highest correlation which is 1. For the other variables, it is either it has negative correlation or negative correlation.

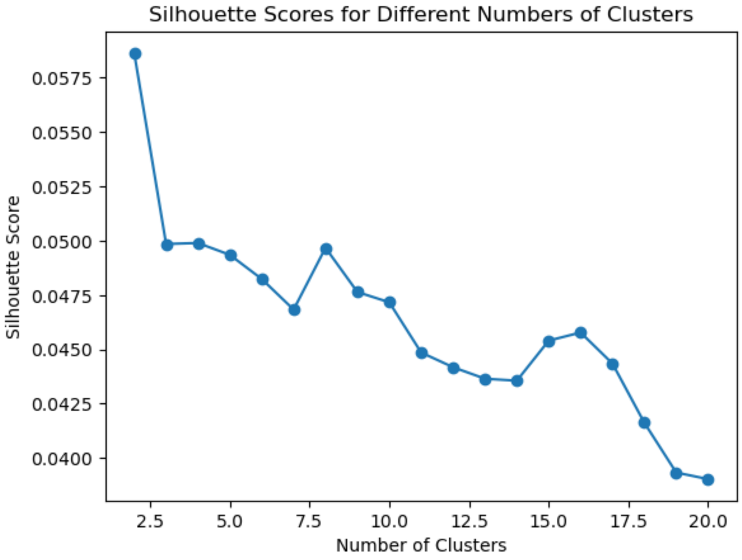
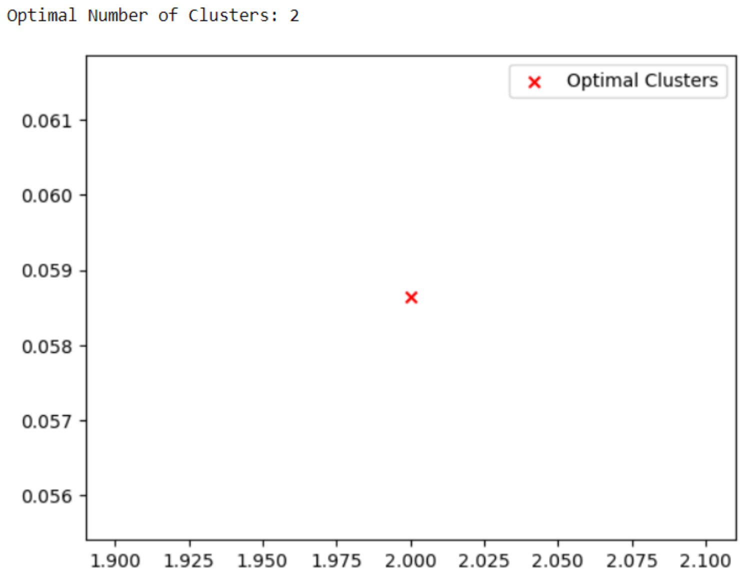


Figure 4.2 Silhouette score for different numbers of clusters

The figure shows the silhouette score where the number of clusters can be determined as 2.5.



```
Optimal Number of Clusters: 2
```

Figure 4.3 Optimal number of clusters

From the silhouette score, we assume that the number of clusters is 2.5 but with figure 4.3 we can determine that the optimal number of clusters is 2.

K-Means

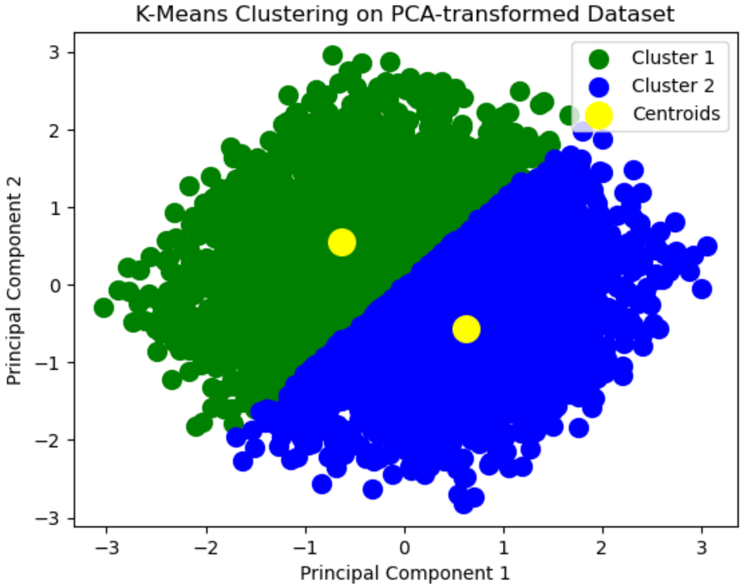


Figure 4.4 K-Means clustering on PCA transformed Dataset

Figure 4.4 shows a scatter plot with K-Means clustering after being transformed by Principal Component Analysis (PCA). Two different clusters of data points with green and blue colours can be clearly seen based on the plot. Cluster 1 is denoted as green in the upper left quadrant, while cluster 2 is represented as blue in the lower right quadrant. Each cluster contains a yellow dot representing the centroid of that cluster.

Hierarchical Clustering

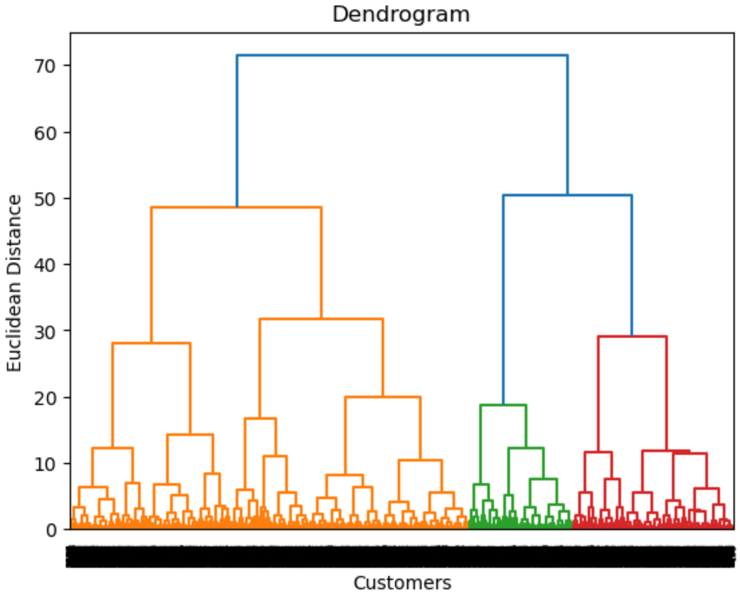


Figure 4.5 Hierarchical Clustering

To determine the number of clusters from a dendrogram in figure 4.5, we would typically look for the longest vertical lines that are not crossed by any horizontal line (extended across the entire width of the plot). There are two very prominent gaps:

- The topmost blue line suggests the data could be split into 2 clusters.
- Another significant gap can be seen a bit lower, where the dendrogram branches into three main arms.

Based on these observations, there could be 2 or 3 clusters, depending on where we choose to "cut" the dendrogram based on the specific context and domain knowledge. For the sake of simplicity, we assume there are 2 clusters.

Agglomerative Clustering

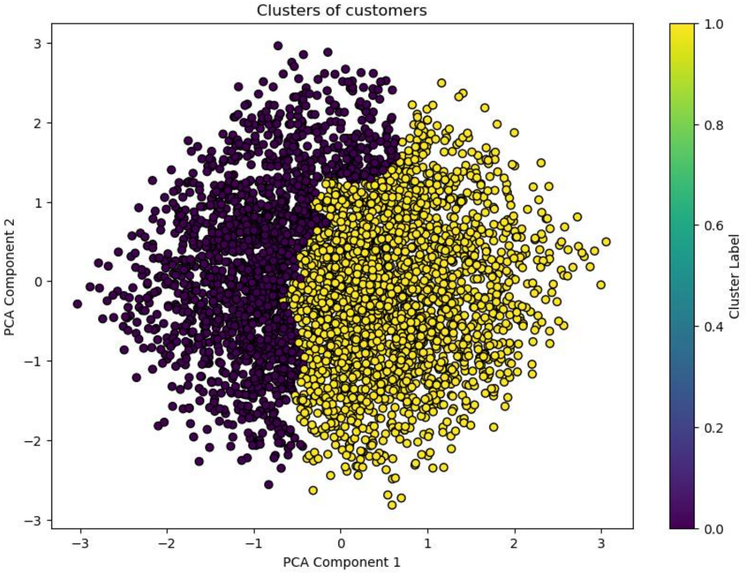


Figure 4.6 Agglomerative Clustering

Figure 4.6 shows a scatter plot with agglomerative clustering after being transformed by Principal Component Analysis (PCA). There are two different clusters of data points with purple and yellow colours. The purple cluster exhibits higher density and concentration towards lower values on both PCA components, indicating that these customers have similar behavioural patterns. Meanwhile, the yellow cluster has a more excellent dispersion towards higher values, suggesting a higher level of variability within this customer segment.

Comparing the clustering algorithms

In this comparative study of clustering algorithms applied to a dataset on shopping habits, the distinct characteristics of each method yielded varied insights. The KMeans algorithm, optimal for spherical cluster shapes and when the cluster count is pre-set, produced a silhouette score of 0.05. On the lower end, this score suggests a lack of well-defined, spherical clusters in the dataset. Furthermore, Agglomerative clustering, which offers a hierarchical perspective of data clustering and excels in situations where inter-data point relationships are pivotal, demonstrated a markedly better silhouette score of 0.3. This score suggests a much more precise and meaningful clustering structure, pointing to the potential significance of hierarchical relationships in the dataset.

The insights derived from these clustering methods regarding the dataset highlight different dimensions of customer segmentation and purchasing patterns. KMeans likely pinpointed broad customer groups, albeit with limited clarity. In contrast, Agglomerative Clustering, emphasising layered relationships, probably unveiled more intricate consumer preferences and behaviour patterns. Therefore, these methods offer a holistic view of the patterns in the shopping behaviour data, each contributing distinct perspectives that collectively facilitate a better comprehension of customer segmentation strategies.

For simplicity and a more straightforward interpretation for non-technical employees (i.e. Marketing, Sales, Business Development, etc), we choose the KMeans clustering algorithm for the e-commerce platform customer segmentation.

Interpretation of Cluster 0 and Cluster 1 for KMeans

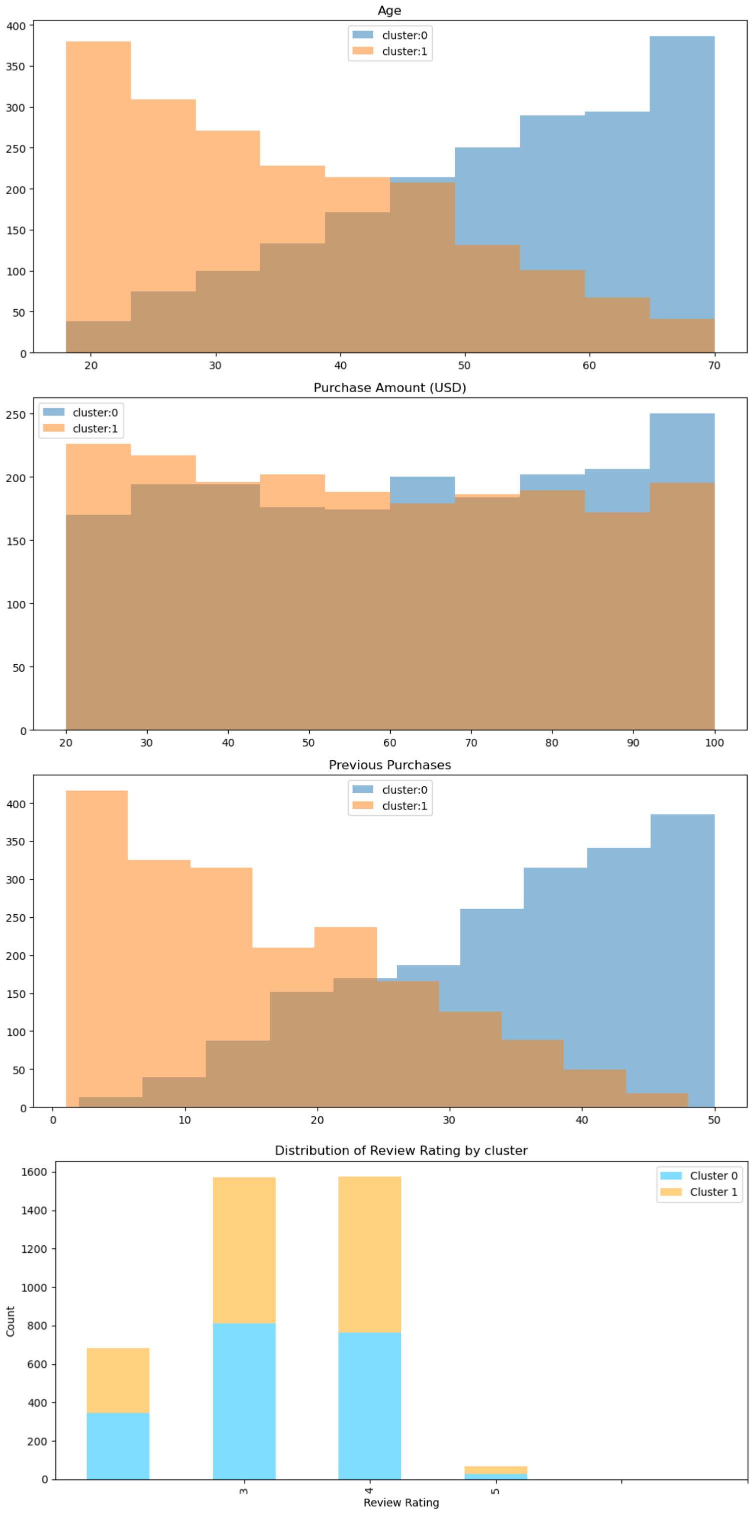


Figure 4.7 K-Means between Cluster 0 and Cluster 1 in age, purchase amount, previous purchase and review rating

Based on the comprehensive statistical analysis, we draw the following key insights based on figure 4.7:

- **Age**. Cluster 0 exhibits an inclination towards an older demographic, while Cluster 1 tends to be younger.
- **Purchase Amount (USD)**. Notably, Cluster 0 showcases a tendency for larger purchase amounts compared to Cluster 1. This aligns logically with the observation that Cluster 0 comprises an older population, often characterized by stable careers and higher income.
- **Past Purchases**. Cluster 0 demonstrates a higher frequency of past purchases in contrast to Cluster 1. This observation is consistent with the notion that the older demographic in Cluster 0 has a broader range of needs, including items for their children (toys, stationary, clothes, food, etc.) and household essentials for the family.
- **Review Rating.** Interestingly, no significant distinctions emerge between Cluster 0 and Cluster 1 concerning the Review Rating column. Both clusters assign similar star ratings, indicating that age may not correlate with increased critical evaluation or heightened expectations regarding the purchased products.
- **Binary Columns:** Analysis of binary columns reveals no substantial variations between Cluster 0 and Cluster 1.

In summary, the two identified customer segments are as follows:

- Cluster 0: Comprising older, wealthier customers with a higher frequency of past purchases.
- Cluster 1: Encompassing younger, more budget-conscious customers with fewer past purchases.

It's essential to note that while the Review Rating column and binary columns may not exhibit significant differences between clusters based on basic statistical measures, further investigation may be warranted. Additional analyses, such as advanced statistical modelling or detailed data visualization, could uncover more nuanced patterns or interactions contributing to a deeper understanding of customer behaviour and preferences.

These insights can provide valuable guidance for businesses to tailor their marketing strategies and product offerings to meet each customer segment's needs and preferences.

1. **Cluster 0: Older, Affluent Customers with Diverse Needs:**
    a. **Tailoring Products and Services:**
        i. Given the older and likely more financially stable nature of Cluster 0, businesses can develop and highlight premium products or services that align with their affluent status.
        ii. Consider offering a diverse range of products that cater to the varied needs of this demographic, including family-oriented items, luxury goods, and household essentials.
    b. **Targeted Marketing Campaigns:**
        i. Craft marketing messages that resonate with an older demographic's life experiences and values. This may include emphasizing product durability, family values, and the convenience of premium offerings.
    c. **Personalized Customer Experience:**
        i. Implement personalized marketing strategies, leveraging past purchases and preferences data to enhance the overall customer experience.
        ii. Loyalty programs or exclusive offers can be designed to reward the loyalty of this customer segment.
    d. **Channel Selection:**
        i. Choose advertising and communication channels more likely to reach and engage an older audience. This might include traditional media, such as television or print, and digital channels.

2. **Cluster 1: Younger, Budget-Conscious Customers with Potential for Future Growth:**
    a. **Affordable Product Lines:**
        i. Develop and promote affordable product lines to cater to the budget-conscious nature of Cluster 1. This could involve creating entry-level or basic versions of products to appeal to this segment.

    b. **Educational Marketing:**

i. Craft marketing campaigns that highlight the value proposition of products, emphasizing quality, functionality, and cost-effectiveness.

ii. Use educational content to inform younger customers about the benefits of products and how they align with their needs.

c. Digital and Social Media Engagement:

i. Given the likely tech-savvy nature of younger consumers, focus marketing efforts on digital and social media platforms. Leverage influencers or online communities to amplify brand visibility.

d. Customer Engagement Strategies:

i. Implement strategies to foster customer loyalty and long-term relationships. This could involve loyalty programs, interactive social media engagement, and responsive customer support.

e. Anticipation of Future Needs:

i. Recognize the potential for future growth within this segment. As younger customers advance in their careers and increase their spending capacity, our e-commerce platform should suggest more relevant products to meet their changing needs.

Classification

Predict if a customer's purchase will use a promo code

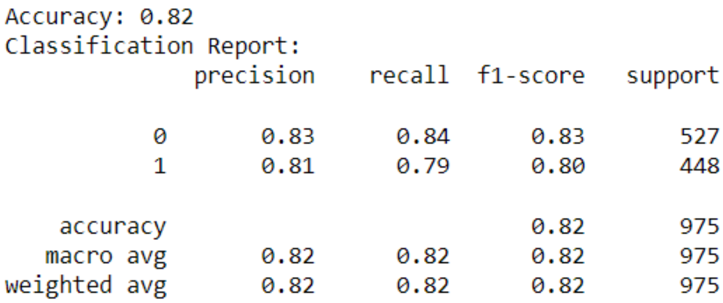
This indicates that the customer is a 'bargain hunter', a person who looks for a place or e-commerce platform to buy something at a price that is cheaper than the usual market price. For e-commerce to entice bargain hunters, they must have pricing and marketing strategies that cater to them. For example, offering first-come-first-serve promo codes for specific items at a limited time, such as the Payday Campaign (the last week of each month).

Why is 'Discount Applied' dropped from X?

Data leakage occurs when information from the target variable is inadvertently incorporated into the features during model training.

In this case, a perfect correlation coefficient of 1.0 between 'Promo Code Used' and 'Discount Applied' indicates a perfect linear relationship. This implies that if 'Discount Applied' contains information about using a promo code, keeping it as a feature in the model would lead to data leakage. Including 'Discount Applied' during training could artificially boost the model's performance, as it may unintentionally learn from the target variable.

To mitigate data leakage, we exclude the 'Discount Applied' column from the features (X). This ensures that the model learns solely from independent features, preventing any influence from information that would not be available during the prediction phase, and promotes a more accurate assessment of the model's generalisation to new, unseen data.

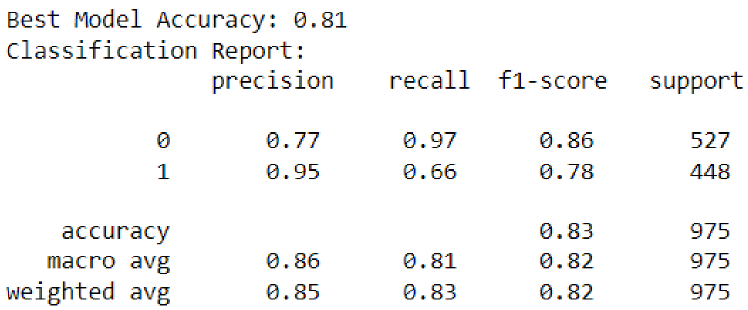


```
Accuracy: 0.82
Classification Report:
              precision    recall  f1-score   support

           0       0.83      0.84      0.83       527
           1       0.81      0.79      0.80       448

    accuracy                           0.82       975
   macro avg       0.82      0.82      0.82       975
weighted avg       0.82      0.82      0.82       975
```

Figure 4.8 Classification Report



```
Best Model Accuracy: 0.81
Classification Report:
               precision    recall  f1-score   support

           0       0.77      0.97      0.86       527
           1       0.95      0.66      0.78       448

    accuracy                           0.83       975
   macro avg       0.86      0.81      0.82       975
weighted avg       0.85      0.83      0.82       975
```

Figure 4.9 Best Decision Tree Classification Report

The model is tailored to predict whether customers will use a promo code, with two distinct classes: Class 0 (Do not use promo code) and Class 1 (Use promo code). Key performance metrics reveal an accuracy of 83%, indicating the correct classification of 83% of samples in the test dataset.

1. **Class 0 Metrics:**
    a. Precision (Positive Predictive Value): 77%
        - 77% of predictions for customers who do not use promo codes were accurate.

    b. Recall (Sensitivity or True Positive Rate): 97%
        - The model correctly identified 97% of customers who do not use promo codes out of all actual cases.

    c. F1-score: 86%
        - A balanced F1-score of 86% for Class 0.

2. **Class 1 Metrics:**
    a. Precision (Positive Predictive Value): 95%
        - 95% accuracy in predicting customers who use promo codes.

    b. Recall (Sensitivity or True Positive Rate): 66%
        - The model captured 66% of customers using promo codes out of all actual cases.

    c. F1-score: 78%
        - An F1-score of 78% for Class 1.

For predicting customers who do not use promo codes (Class 0), the model maintains high accuracy and recall, ensuring the correct identification of the majority in this category. The model emphasises high precision for predicting customers who use promo codes (Class 1), aligning with our objective of offering promo codes more generously. This results in a willingness to potentially miss some customers who would use promo codes, as indicated by the lower recall.

In conclusion, the model provides a strong foundation for predicting promo code usage. With a deliberate focus on increasing recall for Class 1, the e-commerce should offer promotion codes more generously while controlling the costs through the first-come-first-serve mechanism of limiting the promo code to be used only for the first 10,000 transactions.

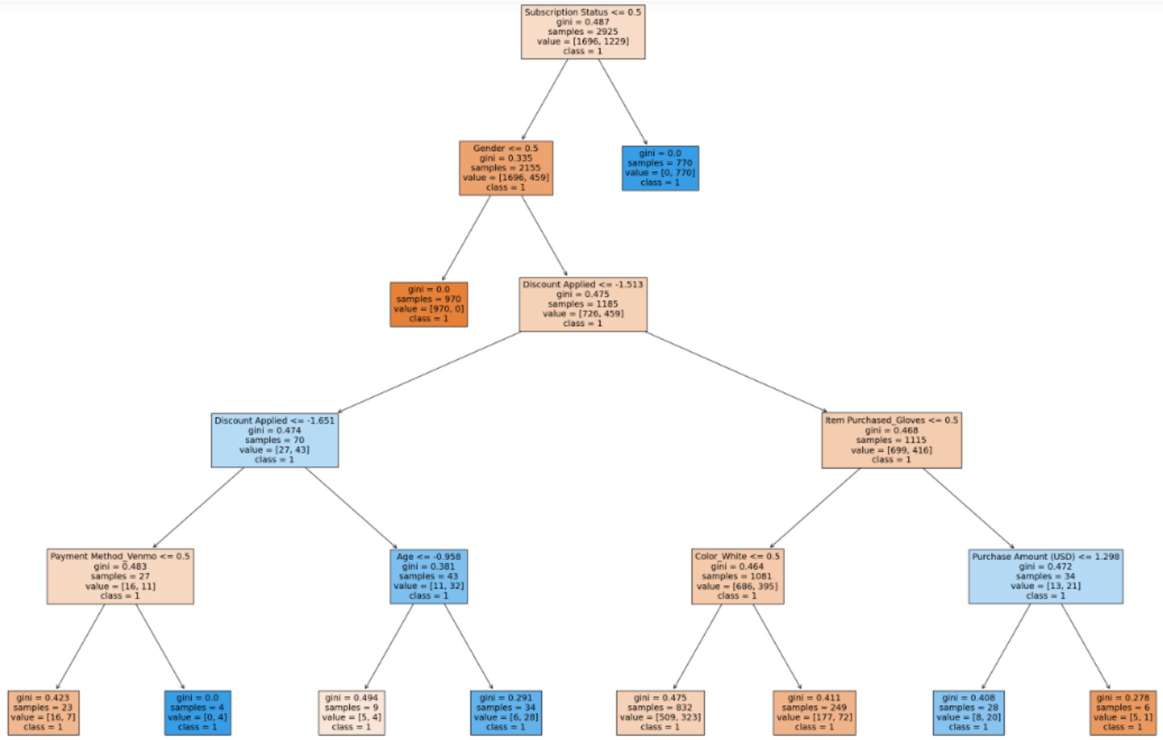


Figure 4.10 Decision Tree

If the condition (subscription <= 0.5) is true, which means subscription status is False, it will go to the left arrow and if the condition (subscription > 0.5) is false will follow the right arrow. The gini of 0.487 refers to about 48.7% of the sample would go in one direction. There are 2925 customers in the sample. Value = [1696,1229] means that 2925 customers will have 1696 customers will get a "NO" and 1229 customers will get a "GO". Next in the decision tree is deciding the gender at the left arrow of subscriptions status. The gini of 0.335 refers to about 33.5% of the sample would go in one direction. Here the sample size is 2155 customers. Value = [1696,459] means that 2155 customers will have 1696 customers will get a “NO” and 459 customers will get a “GO”. On the right arrow, the gini= 0.0 means all the samples got the same result. The sample of 770 means there are 770 customers left in this branch. The value [0,770] means that 770 customers will have 0 customers will get a “NO” and 770 customers will get a “GO”.

5.0 CONCLUSION

From this project, we have acquired valuable knowledge on forecasting client utilization of promotional codes and effectively categorizing our customer base. Our project findings indicate that there is a strategic opportunity to enhance the utilization of promotional codes on the e-commerce platform. While the current model demonstrates outstanding accuracy, the emphasis on improving customer involvement necessitates a nuanced and sophisticated approach. The acknowledgement of the potential to extend the distribution of promotions to customers who are not already availing of them aligns with the main objective of capturing a broader audience. Through our clustering analysis, we successfully categorized our clients into two distinct categories and gained insights into their preferences. Ultimately, this project provides organizations with practical knowledge obtained from data-based examinations, enabling well-informed choices, tailored marketing approaches, and an improved customer experience. By employing exploratory data analysis, clustering techniques, and predictive modelling, one can gain a comprehensive insight of client shopping habits and effectively utilize this knowledge.

6.0 REFLECTION

As a third year undergraduate student with experience working in retail as part-time (some of our group members), we have gained valuable insights into the importance of understanding customer preferences. Through our studies in data analytics, we have learned how to apply data mining techniques to analyze customer data and gain even deeper insights.

During our time working in retail, we have observed firsthand how businesses can benefit from analyzing customer data. For example, we noticed that customers were increasingly interested in eco-friendly products, and by analyzing sales data and customer feedback, the managers were able to adjust the product offerings accordingly. As a result, we saw an increase in sales and customer satisfaction.

Now, as we prepare to work as a data analyst after graduation, we are excited to apply the data mining techniques we have learned to real-world problems. In particular, we are interested in using clustering and predictive modeling to identify customer segments and tailor marketing strategies to their preferences.

The case study provided in our course on data mining applications in retail demonstrated the power of these techniques. By using KMeans clustering, we were able to identify two distinct customer segments: older, affluent customers with diverse needs and younger, budget-conscious customers with potential for future growth. This information can be used to tailor marketing strategies and product offerings to each segment, leading to increased customer satisfaction and loyalty.

Furthermore, the predictive modeling component of the study highlighted the importance of identifying "bargain hunters" and offering them targeted promotions. This strategy can be particularly effective in driving sales and attracting new customers.

Overall, we believe that the use of data mining applications in retail can lead to significant benefits, including improved customer satisfaction, increased sales, and more effective marketing strategies. As a future data analyst, we are eager to apply these techniques to real-world problems and help businesses make informed decisions based on data.

7.0 REFERENCE

Amit. (2020, May 23). *Introduction to Seaborn - Python*. GeeksforGeeks. https://www.geeksforgeeks.org/introduction-to-seaborn-python/

Fawcett, A. (2021, February 11). *Data Science in 5 Minutes: What is One Hot Encoding?* Educative: Interactive Courses for Software Developers. https://www.educative.io/blog/one-hot-encoding

Gad, A. F. (2020, October 12). *Accuracy, Precision, and Recall in Deep Learning*. Paperspace Blog. https://blog.paperspace.com/deep-learning-metrics-precision-recall-accuracy/

Matplotlib Pyplot. (n.d.). Www.w3schools.com. https://www.w3schools.com/python/matplotlib_pyplot.asp

Nidhi. (2023, June 24). *Decision Trees: A Powerful Tool in Machine Learning*. Medium. https://medium.com/@nidhigh/decision-trees-a-powerful-tool-in-machine-learning-dd0724dad4b6

Sharma, N. (2023, June 10). *Understanding and Applying F1 Score: A Deep Dive with Hands-On Coding*. Arize AI. https://arize.com/blog-course/f1-score/#:~:text=F1%20score%20is%20a%20measure

Vikasharma. (2023, January 2). *Lazy Predict Library in Python for Machine Learning*. GeeksforGeeks. https://www.geeksforgeeks.org/lazy-predict-library-in-python-for-machine-learning/

ZEE, S. (2023, October 19). *Consumer Behavior and Shopping Habits Dataset:* Www.kaggle.com. https://www.kaggle.com/datasets/zeesolver/consumer-behavior-and-shopping-habits-dataset/data