NAAN MUDHALVAN

COURSE: ARTIFICIAL INTELLIGENCE

MARKET BASKET INSIGHTS

INSTITUTE NAME: MADRAS INSTITUTE OF TECHNOLOGY, ANNA UNIVERSITY.

COLLEGE CODE: 0004

PADMAJA H 2021503034

MUGUNDH JB 2021503524

LATHA SRI SA 2021503518

PRIYADARSHNI V 2021503538

SANDHYA S 2021503552

**MARKET BASKET INSIGHTS**

**PHASE 1: OBTAINING PROJECT INSIGHTS**

**What is Market Basket Analysis?**

The retailer wants to target customers with suggestions on items that a customer is most likely to purchase. The dataset contains data of a retailer. The transaction data provides data around all the transactions that have happened over a period of time. Retailer will use result to grove in his industry and provide for customer suggestions on itemset, we be able increase customer engagement and improve customer experience and identify customer behaviour. We can solve this problem with the help of Association Rules type of unsupervised learning technique that checks for the dependency of one data item on another data item.

**Problem Statement:**

The problem is to perform market basket analysis on a provided dataset to unveil hidden patterns and associations between products. The goal is to understand customer purchasing behavior and identify potential cross-selling opportunities for a retail business. This project involves using association analysis techniques, such as Apriori algorithm, to find frequently co-occurring products and generate insights for business optimization.

**Problem Understanding:**

Market basket analysis is a crucial aspect of retail strategy. By understanding which products are frequently bought together, businesses can optimize their inventory, marketing strategies, and store layouts. The Apriori algorithm, in particular, is adept at handling large datasets and discovering these patterns efficiently.

**1 Data Understanding:**

The first step involves understanding the dataset provided. This includes grasping the structure of the data, the variables available, and the nature of the transactions. This understanding is fundamental for subsequent preprocessing and analysis.

**2. Data Preprocessing:**

Data preprocessing includes handling missing values, removing duplicates, and transforming the data into transactional format where each transaction comprises a list of purchased items.

**3. Applying the Apriori Algorithm:**

Apriori algorithm is used to perform association analysis on the pre - processed data. Appropriate thresholds are set to filter out significant associations. This step will generate frequent item sets which are groups of items frequently bought together and association rules which represent relationships between products.

**4. Interpretation of Results:**

The generated association rules are interpreted to understand the relationships between products. The products that are commonly bought together and strength of those associations are identified. This interpretation forms the basis for deriving insights into customer behavior and preferences.

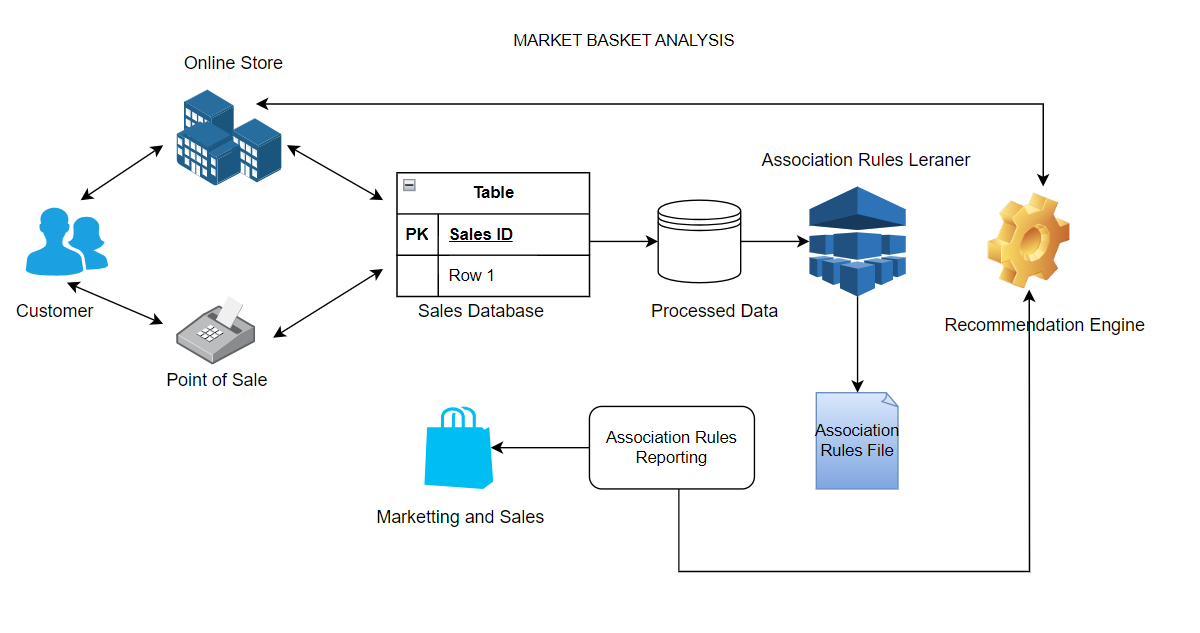
**5. Visualization:**

Visualization tools like heatmaps, network graphs, or simple bar charts can be employed to represent the discovered patterns and associations effectively.

**6. Business Recommendations:**

Based on the interpreted results and visualizations actionable recommendations can be provided to the retail business. These recommendations can include optimizing product placements in stores, creating bundled offers, or designing targeted marketing campaigns.

**MARKET BASKET ANALYSIS – OUTLINE**



**PHASE 2: EXPLORING THE INNOVATIVE METHODS TO PERFORM MARKET BASKET ANALYSIS**

Innovative techniques that can be used to improve the prediction system's accuracy and robustness are

* Ensemble Learning
* Deep Learning Architectures
* Transfer Learning
* AutoML
* Bayesian Optimization
* Reinforcement Learning
* Graph Neural Networks (GNNs)
* Self-Supervised Learning, etc.

**1) Ensemble learning:**

* Ensemble learning is a machine learning technique that enhances accuracy and resilience in forecasting by merging predictions from multiple models. It aims to mitigate errors or biases that may exist in individual models by leveraging the collective intelligence of the ensemble.
* The underlying concept behind ensemble learning is to combine the outputs of diverse models to create a more precise prediction. By considering multiple perspectives and utilizing the strengths of different models, ensemble learning improves the overall performance of the learning system. This approach not only enhances accuracy but also provides resilience against uncertainties in the data. By effectively merging predictions from multiple models, ensemble learning has proven to be a powerful tool in various domains, offering more robust and reliable forecasts.

**2) Reinforcement learning (rl):**

Reinforcement learning is a machine learning paradigm that focuses on making a sequence of decisions to maximize a cumulative reward. While RL is often associated with applications like game playing and robotics, it can also be applied to prediction systems in various domains.

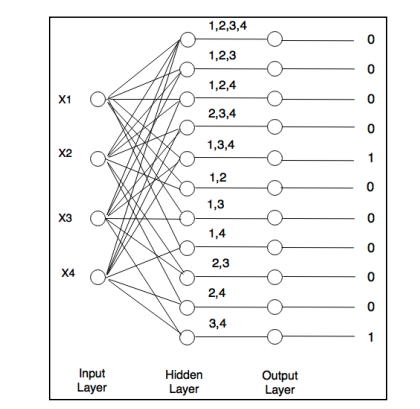
Advantages of using RL:

* Improved Model Selection
* Hyperparameter Tuning
* Exploration of Feature Engineering
* Dynamic Model Ensembles

**3) Deep learning architecture:**

**Feed Forward Neural Network**

Association rule mining aims to discover interesting relationships and patterns within large datasets particularly in transactional databases. The Apriori algorithm specifically focuses on finding frequent itemsets and generating association rules based on these itemsets. The existing Apriori algorithm can be modified by using neural network method in order to optimize the prediction results. Neural networks are a powerful tool for market basket analysis. A Feed forward Neural Network (FFNN) is a type of artificial neural network where information moves in only one direction forward, from the input nodes, through hidden layers (if any), to the output nodes. They can be used to identify patterns in customer behavior and to predict which products are likely to be purchased together.



**4) Using visualizing tools for enhanced insights presentations**

**DATA IMPORT:**

**import numpy as np # linear algebra**

**import pandas as pd**

**from matplotlib import pyplot as plt**

**df=pd.read\_excel("/kaggle/input/market-basket-analysis/Assignment-1\_Data.xlsx")**

Firstly, importing necessary python libraries such as numpy, pandas, and matplotlib. Import the datset (Assignment-1\_Data.xlsx).

**Data understanding and exploration:**

**df.info()**

This method prints information about the created data frame “df” including the index data type and columns, non-null values and memory usage. Drop any rows where item name column is null. Drop any rows where item quantity sold is 0 or less.

**df["SumPrice"]=df["Quantity"]\*df["Price"]**

Create a new column, SumPrice, that tells us total sales revenue (Quantity \* Price) of the item.

**Transformation of data using association rules:**

Market Basket Analysis using Apriori Algorithm and Association Rule Mining

* Convert the Dataset into transactional format (Each row is one bill number with every item sold in that bill in a list)
* Create a one-hot matrix of the products (Product sold = 1, Not sold = 0)
* Merge the transactional matrix and the one hot matrix
* import the mlxtend library and perform association mining and generate association rules

**Generating association rules:**

**rules = association\_rules(frequent\_itemsets, metric="lift", min\_threshold=1)**

**rules**

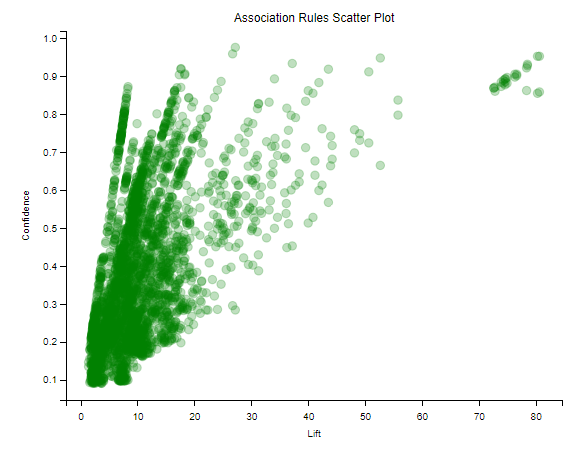
These rules can be used by retailers to make recommendations to customers, or to design marketing campaigns. This is based on “frequent\_itemsets” provided in the given dataset.

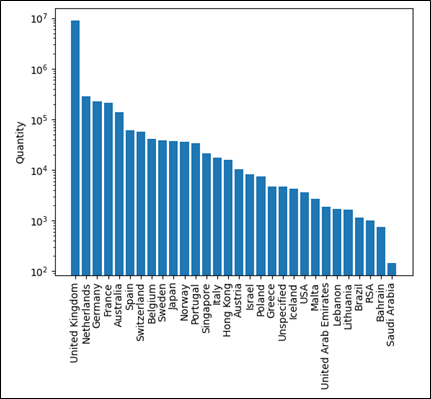
**Filtering the generated rules:**

Association rules are a powerful tool for data mining and machine learning. By filtering association rules, we can extract the most relevant and useful information from the data. Once the rules have been filtered, they can be used for a variety of tasks, such as: recommendation systems, market campaigns and fraud detection.

**Visualization of the rules:**

The scatter plot shows a number of association rules with high lift and confidence. This means that there are a number of item combinations that are frequently purchased together. Retailers can use this information to make recommendations to customers or to design marketing campaigns.





**PHASE 3: DATA PRE-PROCESSING**

**Data Source:**

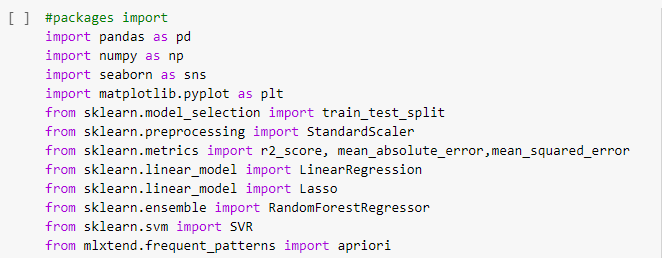
* **Google Collab:** Google Colab, a cloud-based Jupyter notebook environment, serves as our primary coding platform.
* **Python and other libraries** for association analysis and machine learning.

**Implementation Steps:**

1. **Import all necessary libraries:**

Importing all the necessary python libraries like numpy, pandas, seaborn, matplotlib, sklearn and mlxtend for performing association analysis techniques and machine learning.

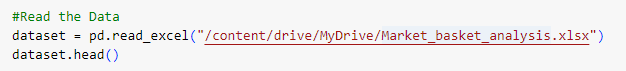
**Code:**

****

1. **Load and Explore the dataset:**

Load the excel file “Market\_basket\_analysis.xlsx” using pandas and explore the dataset to understand its structure.

**Code:**

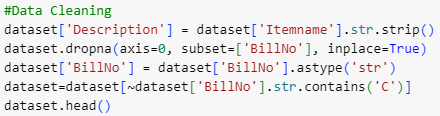


1. **Data Preparation:**

For preparing the data, Data cleaning is done. The following data cleaning steps are performed:

* **Removing duplicate data:** Removes any rows in the dataset using the dropna() function.
* **Correcting data errors:** Converts the column to a string type, which is necessary for performing string operations on the data.
* **Removing invalid data**

**Code:**

****

1. **Training Model:**

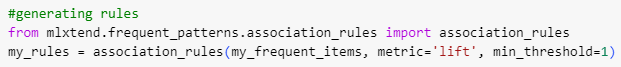
To train an association rule mining model, first identify the frequent itemsets in your data. This can be done using the Apriori algorithm. Once identified the frequent itemsets, generate association rules from them using the association\_rules() function.

Steps involved in training an association rule mining model:

* Identify frequent itemsets.
* Generate association rules.

**Code:**

****

****

1. **Marketing Recommendation:**

Now the condition based rules fileration using the the previously generated rules my\_rules is done.

**Code:**

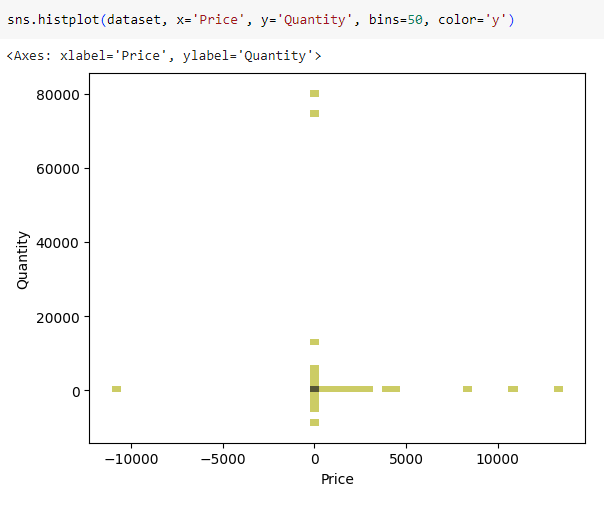
****

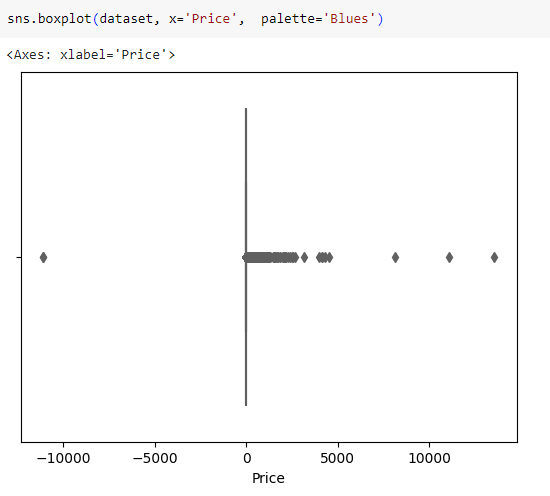
****

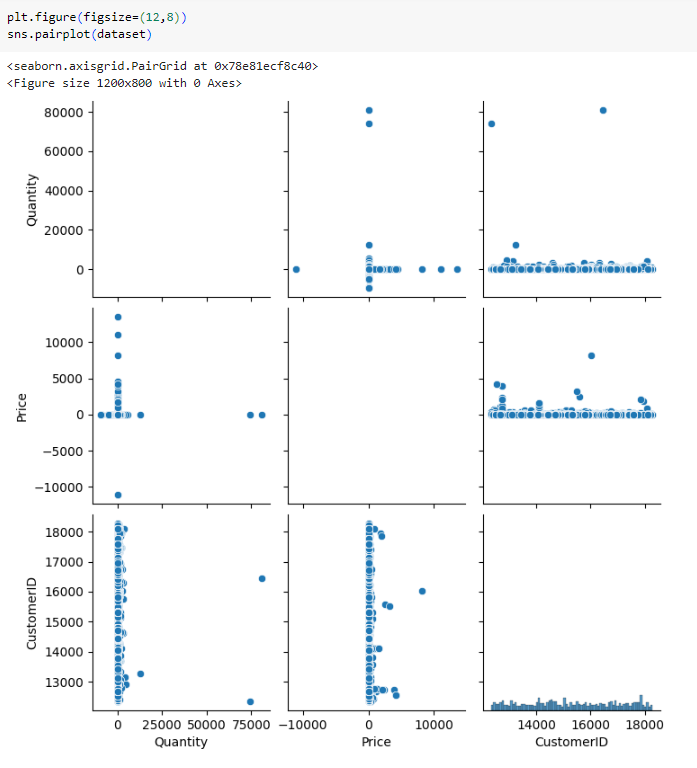
****

1. **Data Visualization:**

For visulizing the data, various plots such as 2D Histogram, boxplot, and pairplot using seaborn python library.

****

****

****

****

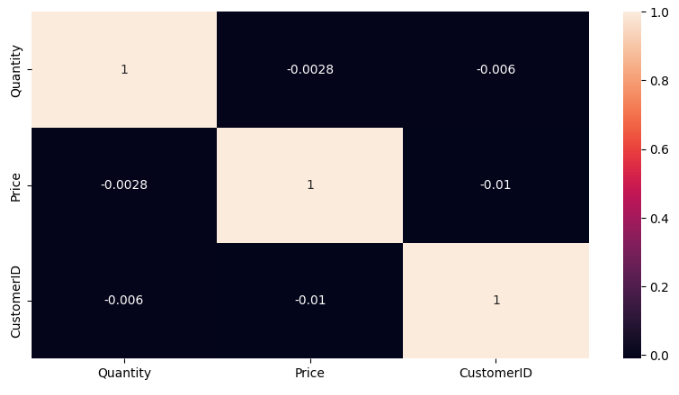
1. **Visualization Correlation:**

In this step, generating a heatmap in seaborn to visualize the correlation between variables in the dataset.

**Code:**

****

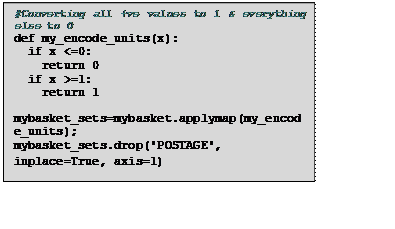
**Output:**

****

**PHASE 4: PERFORMING ASSOCIATION ANALYSIS AND GENERATING INSIGHTS**

In the first step of preparing our data, we carefully sorted through the transactions, isolating those made in 'Germany.' This step was crucial for tailoring our analysis to a specific geographic market. Next, we organized the data systematically by grouping transactions based on their unique 'BillNo' and 'Description.' For each item, we added up the quantities purchased, creating a clear and organized basket format.

The my\_encode\_units function further transforms the data into binary format, which is necessary for the Apriori algorithm. In this binary representation, items were denoted as '1' if they were present in a transaction and '0' if they were absent. This binary representation is a prerequisite for the Apriori algorithm, allowing it to discern itemset associations and pattern with precision.



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

**Apriori algorithm**

* With the preprocessed basket sets in place, the Apriori algorithm is employed. A minimum support threshold of 0.07 was set, a carefully chosen value to ensure the discovery of frequent itemsets while filtering out noise and irrelevant data.
* This step proved indispensable as it unearthed the frequent itemsets — a collection of items frequently co-occurring in transactions. These frequent itemsets serve as the foundation for the subsequent generation of meaningful association rules. These item sets represent the core patterns in customer purchases.

*#Frequent itemsets*

my\_frequent\_items = apriori(my\_basket\_sets, min\_support=0.07, use\_colnames=True)

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

**Association Rules Generation:**

Association rules are pivotal in market basket analysis, revealing valuable insights into customer behavior. After discovering frequent itemsets, these rules were meticulously generated. Using the association\_rules function, each rule was systematically examined, exploring the relationships between different items within the dataset.

Evaluation Based on 'Lift' Metric:

The evaluation of these association rules hinged significantly on the 'lift' metric. 'Lift' measures how much more likely two items are to be bought together compared to if they were bought independently. Here's a breakdown of how this evaluation occurred:

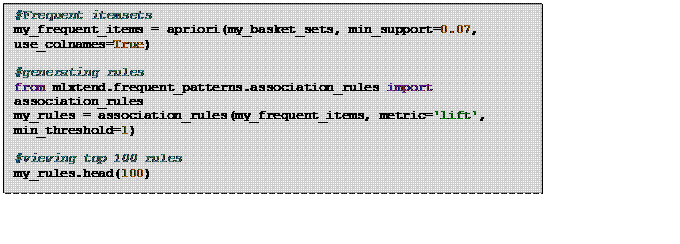
1. Understanding 'Lift':

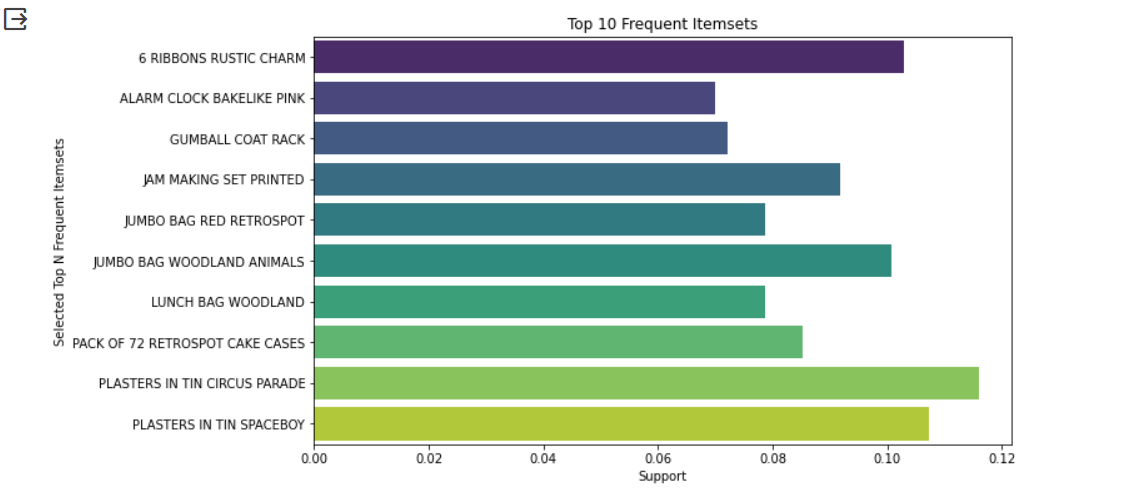
•           Lift > 1: Indicates that items in the rule are more likely to be bought together. A lift score of 1 implies items are independent of each other.

•           Lift < 1: Implies items are less likely to be bought together than if chosen randomly. It indicates a negative association between items.

2. Setting the Threshold:

A threshold of 1 was strategically chosen. This means that only rules demonstrating a significant increase in the likelihood of items being bought together were considered. By setting this threshold, the analysis focused on substantial and meaningful associations, filtering out weaker connections that might not offer actionable insights.





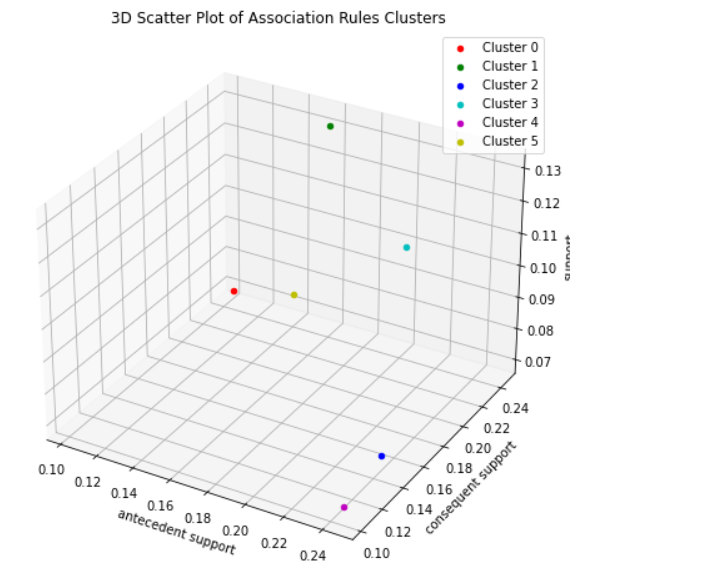
**LOGISTIC REGRESSION**

The dataset is split into features (X) and the target variable (Y) using the train\_test\_split function, allocating 75% of the data for training the model and 25% for testing. A logistic regression model is initialized and trained with the training data (X\_train and Y\_train) using the fit method. This training process involves the model learning the underlying patterns in the data. Subsequently, the trained model is used to make predictions on the testing data (X\_test), and the predicted values are stored in the predictions variable. These predictions represent the model's estimation of the target variable based on the input features.

**K-MEANS ALGORITHM**

K-Means clustering in Market Basket Analysis involves grouping similar transactions based on the items customers purchase together. Each transaction is represented as a binary vector indicating the presence or absence of specific items. K-Means algorithm identifies clusters of transactions exhibiting similar purchasing patterns. These clusters offer insights into customer behavior, enabling retailers to optimize product placements, tailor marketing strategies, and plan promotions effectively.

A 3D scatter plot is created to visualize the clusters formed by KMeans. Each point in the plot represents an association rule. The three selected features ('antecedent support', 'consequent support', 'support') are plotted on the x, y, and z axes, respectively. Different clusters are represented by different colors in the plot. This demonstrates how the association rules are grouped into clusters based on their numerical features, providing insights into the patterns and relationships within the data.



**PHASE 5: CODE COMPLETION AND PROJECT DOCUMENTATION**

**On the successful completion of the Market Basket Insight and Analysis,**

* Help businesses develop personalized product recommendations for customers. This can be done by using your insights to identify products that customers are likely to purchase together.
* Help businesses cross-sell and upsell products. This can be done by recommending related products to customers who are already purchasing a particular product.
* Help businesses optimize their marketing campaigns. This can be done by identifying customer segments that are most likely to respond to certain marketing messages.
* Help businesses improve their supply chain management. This can be done by identifying products that are likely to be in high demand.