**Market Basket Analysis on Grocery Delivery Data Set**

**A PROJECT REPORT SUBMITTED**

*in partial fulfillment for the course*

*of*

**Data Mining and Warehouse**

**MSc in Computer Science (Evening)**

**Submitted by**

Jahanzaib Khan (EP20101019)

Muhammad Ateeq (EP20101026)

Huzaifa Habib (EP20101018)

Abdur Rehman (EP20101003)



**Submitted to**

**Dr Tahseen Ahmed (Assistant Professor)**

**DEPARTMENT OF COMPUTER SCIENCE -UBIT**

**UNIVERSITY OF KARACHI**

**SESSION 2020-2021**

**Table of Contents**

|  |  |  |
| --- | --- | --- |
| **Serial No** | **Contents** | **Page Number** |
| 1 | Abstract | 1 |
| 2 | Introduction | 2 |
| 3 | Data Description | 2 |
| 4 | Methodology | 5 |
| 5 | Data Collection | 5 |
| 6 | Data Preparation | 6 |
| 7 | Data Cleaning and Manipulation | 6 |
| 8 | Sampling Strategy and Data Partitioning | 6 |
| 9 | Tools and Technologies | 6 |
| 10 | Modeling Techniques | 6 |
| 11 | Model 1 | 6 |
| 12 | Apriori Algorithm | 7 |
| 13 | FP Growth Algorithm | 7 |
| 14 | Model 2 | 9 |
| 15 | Decision Tree Algorithm | 9 |
| 16 | Random Forest Algorithm | 10 |
| 17 | Naive Bayes Algorithm | 10 |
| 18 | K Nearest Neighbor Algorithm | 11 |
| 19 | Recommendation | 13 |
| 20 | References | 14 |

# ABSTRACT

Market Basket Analysis is a technique utilized by marketers to grasp Customer behavior whereas looking in their stores. within the on-line shopping process, you will have seen a part referred to as "your suggestions" or "customers purchase this item and purchase it" within which Basket market analysis plays a crucial role. The implementation of this analysis expedited the implementation of electronic purpose of sales systems. Store house owners use written and digital records of client service generated by the purpose of the sales system. This has been with success wont to analyze a adequate quantity of information to grasp concerning customer getting behavior and pattern. In this report we are going to perceive and assist Instacart to method client performance information and concentrate on descriptive analysis of customer getting patterns, purchased things and multiple store units for simple plan and maintenance. product stock. Also, establish teams and teams of clients| of consumers of shoppers with the same shopping for and viewing approach to supply productive recommendations that concentrate on up revenue and customer expertise with classification and prediction models. Our web site contains order-based flexibility and period. Thus, order-related and time-related options were created to predict whether or not the merchandise would be redesigned or not. this can be totally used within the modeling stages and can be used for future analysis of this project. This report can permit Instacart to enhance its user expertise by suggesting a product which will be purchased for the client throughout the order method. additionally, the Report can define Instacart's selling strategy and similar vendors, together with causation customized communications to customers to prompt them to re-order, highlight the merchandise expected for that interaction.

**Introduction**

In today's world, with the appearance of the net of Things, the requirement to introduce is incredibly a lot of required. in keeping with Stastica, a applied mathematics web site, 1.8 billion folks worldwide have purchased merchandise on-line in last year, which is able to still grow within the coming back years. further analysis activities and innovations area unit current to fulfill people's wants and satisfaction. one among the most strategies wont to perceive client getting behavior is to research their group action details. Understanding is required for any style of business and its outcome can cause augmented sales. particularly in an exceedingly outlet, it will be achieved by understanding the client shopping for pattern and connected merchandise sold along. this enables for sudden purchases from customers and for understanding their common shopping for pattern and its effects on the retail market. To better perceive the market, we have a tendency to set to research Instacart activity information. Instacart is AN e-commerce web site that enables users to get groceries at a neighborhood food market on-line, and so sends a private Instacart client to select up and deliver orders created by users on identical day. These processes permit marketers to form continual purchases by users however understanding client getting patterns and behaviors will be frustrating and difficult.

With the data provided, we can find the best models for business purposes mentioned below.

**Model 1**: To predict the next possible product, the customer will make a purchase during the order process

**Model 2**: Creating a model that can predict whether a product will be reorder or not

Requirement of data for above mentioned model are

**Model 1**: Order number, Transactional details of the products, date and time of the transactions, aisles and departments where the product belongs

**Model 2**: Transactional details, Reordered details, time and day details of the products ordered, departments and aisles details

# Data Description

In this section, we will discuss more about the database provided by Instacart. A total of 6 databases provided information about customer transaction details and purchase order.

Section 1: Aisles

This dataset provides information on the aisles such as aisle ID and aisle names, through which the products were organized.

|  |  |
| --- | --- |
| **Variables** | **Description** |
| **Aisle ID** | Labels the ID of the aisles |
| **Aisle name** | Mentions the aisle name in the retail stores |

## Table 1 – Details of Aisles data set

Section 2: Department

This dataset provides information on the departments such as department names and department Id.

|  |  |
| --- | --- |
| **Variables** | **Description** |
| **Department ID** | Labels the ID of the departments |
| **Department name** | Mentions the department name in the retail stores |

**Table 2 – Details of departments data set**

Section 3: Order\_Products\_prior

This dataset gives information on the orders, products, and reordered products

Section 4: Order\_Products\_train

This dataset is same as order\_products\_prior and it is a trained dataset.

|  |  |
| --- | --- |
| **Variables** | **Description** |
| **Order ID** | Labels the ID of the order made by customer |
| **Product ID** | Labels the ID of the products purchased by customers |
| **Add to cart order** | Sequence of the order placed in the cart |
| **Reordered** | Denotes whether the products are reordered or not |

## Table 3 – Details of order\_prior\_train data set

Section 5: Orders

This dataset has information about the customer orders like order ID, order number, week day of the order, hour of the order, user ID and days since prior order.

|  |  |
| --- | --- |
| **Variables** | **Description** |
| **Order ID** | Labels the ID of the order made by customers |
| **User ID** | Labels the ID of the users who made the purchase |
| **Order number** | Denotes the order number made by the customer |
| **Order\_dow** | Denotes the day of the week, the order made by the customer |
| **Order hour of day** | Denotes the hour of the day, the order made by the customer |
| **Days since prior order** | Denotes the number of days since last order |

## Table 4 – Details of orders data set

Section 6: Products

This dataset gives information on the products such as product name, product ID, aisle and departments, which were sold to the customer.

|  |  |
| --- | --- |
| **Variables** | **Description** |
| **Product ID** | Labels the ID of the products purchased by customers |
| **Product Name** | Denotes the product name purchased by the customer |
| **Aisle ID** | Labels the ID of the aisles |
| **Departments ID** | Labels the ID of the departments |

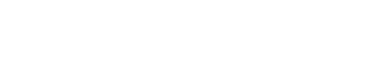
**Table 5 – Details of products data set**

# Methodology

This project focus on predicting the next products the customer tends to purchase and also whether a product is reordered or not by the customers in their next purchase.



**Planning**

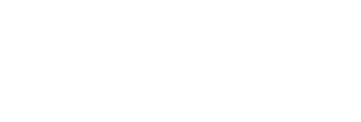


Data Validation and

Cleaning



Data Collection

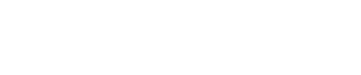


Building Predictive

Model for

reordering

and next product



Identifying the

Conclusion



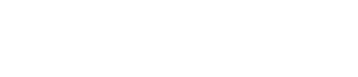
**Implementation**



**Modelling**



Data Preparation



Implementing the

Project

**Figures 1**

**–**

**Methods adopted for the purpose of analysis**

# Data Collection

Database provided by Instacart online grocery Company and taken to Kaggle for analysis. Data sets provided by Instacart contain complete information on more than 3 million wholesale orders from more than 200,000 Instacart users Both product data and customer data from Instacart cover 50,000 different products, week and time of purchase, unique production location and departments.. Understanding the details, dairy products, fruits and vegetables are widely purchased in every department and people often buy and re-order 60% of their previous orders especially on Sundays and Mondays.

# Data Preparation

In order to enhance the efficiency and computing ability we have made some change in our dataset that are mentioned in below.

# Data Cleaning and Manipulation

There are no empty or empty variable values ​​such as aisle, doors,Order\_product\_prior, order\_product\_train and product data sets. The order data set has empty values ​​in the days since the previous order variable and only 5% of the values ​​are found missing that was discarded cause no significant problem.

# Sampling Strategy and Data Partitioning

Since these data is to much to compute in our processors that’s why we have taken 1000 data of each dataset (e.g order\_products\_\_train\_chunk , Order\_chunk)

70% data has been used to train model and rest of 30% data has been used for testing.

# Tools and Technologies

These are the tools and technology that are most likely to be used in our projects

* Python Programming Language
* Google COLAB
* Scikit learn
* Apyori 1.1.2

# Modeling Techniques

**Model 1: To predict the next possible product, the customer will make a purchase during the order process**

Appriori algorithm and FP Growth algorithm are used to predict sequence of product that customer will purchase with the help of parameters Support, Confidence and Lift.

**Support** : It's the chance of deals that comprise all of the particulars in a dataset. The more the support value the more constantly the product occurs. High support values are preferred for ample quantum of unborn deals.   
**Confidence:** It's the probability that a sale that contains the particulars on the left hand side of the also contains the item on the right hand side. The more the confidence value, the lesser the liability that the product on the right hand side will be bought.   
**Lift:** It is nothing but the rate of Confidence to Anticipated Confidence. It's the probability of all of the products in a rule being together by the product of the chances of the particulars on the left and right hand side being as if there was no association between them.

**Apriori Algorithm**

The Apriori algorithm is a type of supervised literacy algorithm used for association rule mining. The algorithm quests for common objects in data sets and builds connections and associations of objects insets.Name of the algorithm is Apriori because it uses previous knowledge of frequent itemset parcels   
Steps to Apriori Algorithm   
Below are the way for the apriori algorithm

**Step-1** First of all find Support of each item set and elect the minimal support and confidence.   
**Step- 2** Elect all those particulars whose supports in the sale with advanced support value than the minimal or favored support value.   
**Step-3** Find all the rules of these subsets that have advanced confidence value than the threshold or minimal confidence.   
**Step-4** Sort the rules as the dwindling order of lift.

**Frequent Pattern Growth Algorithm**

This algorithm is an enhancement to the Apriori system. A frequent pattern is generated without the need for seeker generation. FP growth algorithm represents the database in the form of a tree called a frequent pattern tree or FP tree.   
FP Tree   
The Frequent Pattern Tree is a tree-suchlike structure made from the original accoutrements of the point. The purpose of the FP tree is to dig up the most common pattern. Each FP tree knot represents an itemset item.   
The root knot represents null while the lower bumps represent objects. The connection of the bumps with the lower bumps which are resembling to each other is maintained during the construction of the tree.   
Steps to FP Growth are

**Step-1** The first step is to overlook the database to find the circumstances of the itemsets in the database. This step is the same as the first step of Apriori. The count of 1-itemsets in the database is called support count or frequency of 1-itemset.   
**Step-2** The alternate step is to construct the FP tree. For this, produce the root of the tree. The root is represented by null.   
**Step-3** The coming step is to overlook the database again and examine the deals. Examine the first sale and find out the itemset in it. The itemset with the maximum count is taken at the top, the coming itemset with lower count and so on. It means that the branch of the tree is constructed with sale itemsets in descending order of count.   
**Step-4** The coming sale in the database is examined. The itemsets are ordered in descending order of count.However, also this sale branch would partake a common prefix to the root, If any itemset of this sale is formerly present in another branch (for illustration in the 1st sale).   
This means that the common itemset is linked to the new knot of another itemset in this sale.   
**Step-5** Also, the count of the item set is incremented as it occurs in the deals. Both the common knot and new knot count is increased by 1 as they're created and linked according to deals.

**Step-6** The coming step is to booby-trap the created FP Tree. For this, the smallest knot is examined first along with the links of the smallest bumps. The smallest knot represents the frequence pattern length 1. From this, cut the path in the FP Tree. This path or paths are called a tentative pattern base.   
Tentative pattern base is asub-database conforming of prefix paths in the FP tree being with the smallest knot **Step-7** Construct a Tentative FP Tree, which is formed by a count of itemsets in the path. The itemsets meeting the threshold support are considered in the Tentative FP Tree.   
**Step-8** Frequent Patterns are generated from the Tentative FP Tree.   
Some of the association rules which were interpreted from our methodology are mentioned below

|  |  |
| --- | --- |
|  | **Rules** |
| Rule 1 | Organic Strawberries & Organic Hass Avocado ==> Organic Baby Spinach & Bag of Organic Bananas |
| Rule 2 | Organic Hass Avocado & Organic Baby Spinach ==> Organic Strawberries & Bag of Organic Bananas |
| Rule 3 | Organic Yellow Onion & Organic Baby Spinach ==> Organic Garlic |
| Rule 4 | Organic Strawberries & Organic Garlic ==> Organic Yellow Onion |
| Rule 5 | Organic Garlic & Bag of Organic Bananas ==> Organic Yellow Onion |
| Rule 6 | Organic Yellow Onion & Organic Strawberries ==> Organic Garlic |
| Rule 7 | Organic Yellow Onion & Bag of Organic Bananas ==> Organic Garlic |
| Rule 8 | Organic Garlic & Organic Baby Spinach ==> Organic Yellow Onion |
| Rule 9 | Organic Hass Avocado & Bag of Organic Bananas ==> Organic Lemon |
| Rule 10 | Organic Lemon & Bag of Organic Bananas ==> Organic Hass Avocado |
| Rule 11 | Organic Strawberries & Organic Lemon ==> Organic Hass Avocado |
| Rule 12 | Organic Strawberries & Organic Hass Avocado ==> Organic Lemon |
| Rule 13 | Organic Strawberries & Organic Baby Spinach & Bag of Organic Bananas ==> Organic Hass Avocado |
| Rule 14 | Organic Strawberries & Organic Hass Avocado ==> Organic Raspberries |
| Rule 15 | Organic Yellow Onion & Bag of Organic Bananas ==> Organic Hass Avocado |
| Rule 16 | Organic Hass Avocado & Organic Baby Spinach ==> Organic Yellow Onion |
| Rule 17 | Organic Garlic & Bag of Organic Bananas ==> Organic Hass Avocado |
| Rule 18 | Organic Ginger Root ==> Organic Garlic |
| Rule 19 | Organic Italian Parsley Bunch ==> Organic Garlic |

## Table 6 – Association rules

**Model 2: Transactional details, Reordered details, time and day details of the products ordered, departments and aisles details**

Various modeling techniques like , decision tree classifier, Random Forest, Naive bayes formula, K Nearest Neighbor classifier were adopted to predict whether or not a product are reordered or not. the most effective model is accessed and valid with the assistance of model comparison node examination all the results of the various models used..

**The Decision Tree**

The Decision Tree may be a straightforward illustration for classifying examples. it's a Surveyed Machine reading wherever information is more and more categorised consistent with an exact parameter.

The Decision Tree contains:

**Node:** Check the value of a specific attribute.

**Edges / Branch**: Match the test result and link to the next node or leaf.

**Leaf nodes:** Terminal node predict the outcome (representing class labels or class distribution).

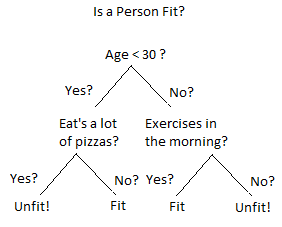


Figure 2 Decision Tree example

To understand the construct of call Tree contemplate the instance on top of. Suppose you wish to predict whether or not someone is work or not, given his or her data like age, ingestion habits, exercise, etc. choices square measure queries like 'How recent square measure you?', 'Do you exercise?', 'Do you eat heaps of pizzas'? and therefore the leaves represent effects like ‘worthiness’, or ‘inadequacy’.

**Random Forest Algorithm**

Random Forest may be a well-liked machine learning formula that's a part of a supervised learning strategy. It is supported the construct of integrated learning, that is that the method of group action multiple dividers to unravel advanced issues and improve model performance.

As the name suggests, "The Random Forest may be a subdivision that contains variety of call trees for the varied datasets set and takes mensuration to enhance the prediction accuracy of that info." rather than hoping on one call tree, the random forest takes a prediction from every tree and relies on multiple predictable votes, and predicts the ultimate outcome.

The large variety of trees within the forest results in high accuracy and majority of votes contemplate the ultimate result and it prevents the matter of over fitting.



Figure 03 Random Forest Example

**Naive Bayes algorithm**

Naive Bayes formula may be a supervised learning formula, that relies on Thomas Bayes theorem and used for resolution classification issues. I

it is a probabilistic classifier, which suggests it predicts on the premise of the likelihood of an object.

## Bayes' Theorem:

• Bayes' theorem is additionally called Bayes' Rule or Bayes' law, that is employed to work out the likelihood of a hypothesis with previous data. It depends on the contingent probability.

• The formula for Bayes' theorem is given as:

Naïve Bayes Classifier Algorithm

**Where,**

**P(A|B) is the** Probability of hypothesis A on the observed event B.

**P(B|A) is the** Probability of the evidence given that the probability of a hypothesis is true.

**P(A) is** Probability of hypothesis before observing the evidence.

**P(B) is Marginal Probability** or Probability of Evidence.

1. Convert the given dataset into frequency tables.

2. Generate probability table by finding the chances of given options.

3. Now, use Thomas Bayes theorem to calculate the posterior likelihood.

# K-Nearest Neighbor (KNN) Algorithm

K-Nearest Neighbor is one among the best Machine Learning algorithms supported supervised Learning technique. It assumes the similarity between the new case/data and obtainable cases and place the new case into the class that's most just like the obtainable classes. This formula stores all the obtainable information and classifies a replacement datum supported the similarity. this implies once new information seems then it are often simply classified into a well suite class by mistreatment K- NN formula. It are often used for Regression moreover as for Classification however largely it's used for the Classification issues. it's a non-parametric formula, which suggests it doesn't create any assumption on underlying information. it's additionally referred to as a lazy learner formula as a result of it doesn't learn from the coaching set forthwith instead it stores the

dataset and at the time of classification, it performs AN action on the dataset. At the coaching section it simply stores the informationset and once it gets new data, so it classifies that information into a class that's abundant just like the new information.

**Example:** Suppose, we've got a picture of a creature that appears just like cat and dog, however we wish to understand either it's a cat or dog. thus for this identification, we are able to use the KNN formula, because it works on a similarity live. Our KNN model can realize the similar options of the new information set to the cats and dogs pictures and supported the foremost similar options it'll place it in either cat or dog class.

**Product Models**

|  |  |
| --- | --- |
| **Mode** | **Accuracy** |
| Decision tree | **59%** |
| Random Forest | **58 %** |
| Naive Bayes algorithm | **58%** |
| K Nearest Neighbor | **56.5%** |

**Table 7 – Accuracy, misclassification for models**

As are often seen from the table on top of, accuracy was the very best for call tree model, Random forest and Naive Thomas Bayes has same accuracy and K Nearest Neighbor has lowest accuracy.

# Recommendations

Based on the legal process associated with models for predicting product restructuring, some recommendations have been made:

• It will have a positive effect on conducting promotional and marketing campaigns with the help of integration laws. Based on the following product predictions, customers can be offered additional offers by combining products together at a lower cost and customizing products according to the organization's rules.

• Based on the redesign model, personal communication can be of great benefit by reminding customers to rearrange products or can be added to the cart automatically based on customer preferences.

• We can recommend Instacart to add products directly to the customer's cart or provide a list of suggestions when purchasing to improve customer information.

• Knowing the level of re-ordered products, Instacart can use re-ordered data to analyze stocks by ensuring refinement and proper product planning to increase internal productivity.

# References

* <https://www.kaggle.com/c/instacart-market-basket-analysis/data>
* <https://searchcustomerexperience.techtarget.com/definition/market-basket-analysis>
* <https://www.sciencedirect.com/topics/computer-science/market-basket-analysis>
* <https://scikit-learn.org/stable/>
* <https://www.geeksforgeeks.org/decision-tree/>
* <https://www.analyticsvidhya.com/blog/2021/06/understanding-random-forest/>
* <https://towardsdatascience.com/naive-bayes-classifier-81d512f50a7c>
* <https://www.softwaretestinghelp.com/fp-growth-algorithm-data-mining/>
* <https://www.softwaretestinghelp.com/fp-growth-algorithm-data-mining/>
* <https://www.analyticsvidhya.com/blog/2017/09/naive-bayes-explained/>
* <https://www.analyticsvidhya.com/blog/2021/04/simple-understanding-and-implementation-of-knn-algorithm/>
* <https://builtin.com/data-science/random-forest-algorithm>
* <https://venngage.com/blog/what-is-a-decision-tree/>