



**VIT<sup>®</sup>**  
**Vellore Institute of Technology**  
(Deemed to be University under section 3 of UGC Act, 1956)

# **ONLINE SHOPPING CART ANALYSIS TO UNDERSTAND THE CUSTOMER'S ONLINE EXPENDITURE PATTERN**

*A PROJECT REPORT  
For*

**DATA MINING TECHNIQUES (ITE2006)  
in  
B.Tech – Information Technology and Engineering**

*by*

<i><b>Name</b></i>	<i><b>Registration number</b></i>
<i><b>AMAN SOMANI</b></i>	<i><b>19BIT0166</b></i>
<i><b>RITWIK SUKHIJA</b></i>	<i><b>19BIT0411</b></i>
<i><b>VISHNU PRABHAKAR</b></i>	<i><b>19BIT0043</b></i>

Under the guidance of

**Prof. Valarmathi B**

## **Abstract :**

Huge volumes of data are generated in day-to-day activities, resulting in a tremendous increase in data volume. One of the key issues for data management and mining communities is extracting information from the exploding amount of data. Furthermore, the vast majority of well-known companies acquire and keep enormous volumes of client transaction data. However, just because the organisations had a lot of data doesn't mean they had a lot of business information. The business world needs to extract useful information and insights from this massive amount of data. This leads to an examination of online shopping carts. The goal of shopping cart analysis is to figure out people's buying habits by identifying crucial connections between the things they put in their online shopping carts. It not only aids in decision-making, but it also boosts sales on various e-commerce sites such as Amazon, Flipkart, and others... The most common methods for mining frequent itemsets are Apriori and FP Growth. For both of these techniques, a predefined minimum level of support is required to identify frequent itemsets. When the minimum support is low, however, a significant number of candidate sets are formed, necessitating a lot of processing. We intend to use a strategy that has been proposed to prevent this massive calculation by limiting the dataset items with top-selling products in this project. With the help of consumer suggestions, the best-selling products will be promoted even more.

Various percentages of top selling products like 30%, 40%, 50%, 55% have been taken and for both algorithms frequent itemsets and association rules are generated. The results show that if top selling items are used, it is possible to get almost same frequent itemsets and association rules within a short time comparing with those outputs which are derived by computing all the items. From time comparison it is also found that FP Growth algorithm takes less time than the Apriori algorithm.

## **Keywords:**

Market Basket Analysis (MBA), Data Mining, Association Rule Mining (ARM), Product Recommendation system.

## **Introduction :**

Market basket analysis is a data analysis methodology based on Association data mining that merchants employ to boost sales by better understanding customer purchase patterns. It entails evaluating huge data sets, such as purchase histories, to identify product groups and products that are likely to be bought together. Customers' purchase patterns are discovered through market basket analysis, which identifies significant links between the things they place in their shopping baskets. It not only aids in the decision-making process, but it also boosts sales in many businesses. The most common methods for mining frequent itemsets are Apriori and FP Growth. Market Basket Analysis is an important part of the analytical system in the retail organisation to determine the placement of goods, designing sales promotion for different

segments of customers to improve customer satisfaction and hence the profit of the supermarkets.

## Proposed Method:

We will use the concept of Association Rule Mining in our project.

Association rule algorithms includes :

- Apriori Algorithm
- F-P Growth Algorithm (Frequent Pattern Algorithm)

So, we will be using Apriori Algorithm and F-P Growth Algorithm for the Market Basket Analysis (MBA).

## Literature Survey :

S.No.	Title of the Paper and Year	Algorithms used	Data set being used	Performance measures	Scope for future work
1.	An Implementation of the FP-growth Algorithm (2010)	FP-growth Algorithm , Apriori and Eclat algorithm	BMS-Webview-1,T10I4D100K1, census, chess, and mushroom	Accuracy - 72%	The performance of the two projection methods for projecting an FP-tree, especially, why the second is sometimes much slower than the first, needs further investigation
2.	Performance Analysis of Apriori and FP-Growth Algorithms (Association Rule Mining) (2016)	Weka workbench	Supermarket and Voter datasets	Accuracy- 87%	The performance of Apriori and FP-Growth algorithms can be analysed by taking less execution time and lesser number of scans for different instances.

3.	Study on Market Basket Analysis with Apriori Algorithm Approach (2021)	Apriori Algorithm	(Name not mentioned)	Accuracy - 78%	For future using this project a shop owner can place some items close together in future for their customers to pick more than one item whether they were previously only going to buy a single item.
4.	Market Basket Analysis with Enhanced Support Vector Machine (ESVM) Classifier for Key Security in Organisation (2019)	Association Rule Mining Algorithm Based on Probabilistic Graphical Model, FPM algorithm, ESVM algorithm	Bank Marketing Dataset	Accuracy - 86%	The works intended for the future will provide the implementation of specific features that the work was not capable of exploring by using a more reliable algorithm in the system, which in turn would facilitate the system in operating rapidly and with more efficiency. Efforts on improving the search techniques can also be useful in boosting the market and profitability.
5.	Market Basket Analysis: Identify the changing trends of market data using association rule mining (2016)	Market Basket algorithm	Extended bakery datasets	Accuracy - 70%	Authors suggested that some areas are still there which need to be focused on. Firstly, results have been influenced greatly by the manual threshold values for score, so it is needed to automate the threshold values for better recognition of

					outliers. Secondly, this approach is specifically targeted at Market Basket Data, it may perhaps be extended to other areas.
6.	Comparative Analysis of Market Basket Analysis through Data Mining Techniques (2021)	Collaborative Filtering Algorithm, ARM	Transactional Dataset	Accuracy - 74%	There are different data mining algorithms available to find out the frequent trends in consumer's buying pattern and also to give the product recommendation on the basis of past purchases made by the consumer.
7.	Analysing Online Transaction Data using Association Rule Mining: Misumi Philippines Market Basket Analysis (2019)	Apriori Algorithm	Transactional Dataset	Accuracy-70%	Producing a list of package items for consumers based on strong rules generated by association rule mining at a lesser runtime rate.
8.	Data Mining Applications for Sales Information System Using Market Basket Analysis on Stationery Company (2017)	Generalised Rule Induction Based Hash Algorithm, Apriori Algorithm	(Name not mentioned)	Accuracy- 71%	According to the questionnaire, 85% of consumers rate the appearance of good, 15% of consumers rate the appearance is very good, 100% of users assess the accuracy of the data generated very good, 25% of users rate the application, simply, 75% of users rated ease of application

					excellent, 100% of users evaluate reports produced good, 10% of users to assess the suitability of the needs of both, 90% of consumers rate the suitability to the needs of very good.
9.	MARKET BASKET ANALYSIS USING FP GROWTH AND APRIORI ALGORITHM: A CASE STUDY OF MUMBAI RETAIL STORE (2016)	FP GROWTH AND APRIORI ALGORITHM	Synthetic Transactional Dataset	Accuracy- 74%	Increasing the sample size of the dataset and accuracy of the analysis.
10.	MARKET BASKET ANALYSIS: UNDERSTANDING INDIAN CONSUMER BUYING BEHAVIOUR OF SPAIN MARKET (2016)	APRIORI ALGORITHM	Market basket dataset	Accuracy- 73%	The market basket problem can be seen as the best example of mining association rules. Discovering association rules has been a well-studied area for the past decade. Building up on previous researches by using established methods for mining association rules allowed for discovering useful information for the retailer

S.No.	Title of the Paper and Year	Algorithms used	Data set being used	Performance measures	Scope for future work
-------	-----------------------------	-----------------	---------------------	----------------------	-----------------------

11.	Consumer purchase patterns based on market basket analysis using apriori algorithms(2020)	Apriori algorithm	Transactional dataset	Accuracy-74%	It is expected that the results of consumer purchasing patterns can help minimarket managers in making decisions to get even better profits.
12.	FP-Tree Based Algorithms Analysis: FP-Growth, COFI-Tree and CT-PRO(2011)	FPGrowth, COFI-Tree, CT-PRO	Transactional dataset	Accuracy-80%	FP-Growth is the first successful tree base algorithm for mining the frequent itemsets. As for large databases its structure does not fit into main memory therefore new techniques come into pictures which are the variations of the classic FP-Tree.
13.	Online Shopping: Do Men Behave Differently than Women?(2021)	EFA, CFA and SEM tools	(Name not mentioned)	Accuracy-77%	The study proposed and validated gender specific behavioural determinants of the young Indian online shoppers. Taking this study as a base, further research investigations into gender-wise specific product category behaviour can be

					studied.
14.	Uncovering Modern Clinical Applications of Fuzi and Fuzi-Based Formulas: A Nationwide Descriptive Study With Market Basket Analysis(2021)	Apriori algorithm	Taiwan National Health Insurance dataset	Accuracy-83%	The results light up the road to the development of new Fuzi-based botanic drugs
15.	Chiller system performance management with market basket analysis(2018)	FP-growth algorithm, Association optimisation algorithm, Apriori algorithm, Self-joining algorithm	Operational Dataset	Accuracy-72%	Tailor-made optimisation strategies and the associated electricity savings can be further evaluated when developing a COP model with significant variables and predicting its maximum values under different operating conditions.
16.	Improving Efficiency of Apriori Algorithm Using Transaction Reduction (2013)	Apriori Algorithm	(Name not mentioned)	Accuracy-75%	Although this improved algorithm is optimised and efficient, but it has overhead to manage the new database after every generation of Lk. So, there should be some approach which has a very small number of scans of the database. Another solution might be the division of large databases among processors.
17.	Sales Prediction System using Machine Learning(2019)	Decision tree algorithm, XGBoost	Big Mart Companies Real-world	Accuracy-81%	In the future, we will use the output of this project as



		regressor	data set		part of the price optimization problem which we are planning to work on.
18.	Research on a Prediction Model of Online Shopping Behaviour Based on Deep Forest Algorithm(2020)	Deep forest algorithm	Customers online shopping behaviour dataset	Accuracy-79%	The combination of algorithm and business is still at the initial stage. Although machine learning algorithms are data-driven, there is no uniform standard for the construction and selection of features. Further work needs to be done on them.
19.	AN IMPROVED APRIORI ALGORITHM FOR ASSOCIATION RULES(2014)	Apriori Algorithm	Transactional database	Accuracy-78%	The time consumed to generate candidate support counts in the improved Apriori is less than the time consumed in the original Apriori; the improved Apriori reduces the time consuming by 67.38%. As this is proved and validated by the experiments, it would prove very useful in the future.
20.	A NOVELTY OF DATA MINING FOR PROMOTING EDUCATION BASED ON FP-GROWTH ALGORITHM(2018)	FP-growth algorithm	Transactional database	Accuracy-76%	From research done on some attributes not used in the resulting rule, so the selection of attributes in the dataset is very important.

S.No.	Title of the Paper and Year	Algorithms used	Data set being used	Performance measures	Scope for future work
-------	-----------------------------	-----------------	---------------------	----------------------	-----------------------

21.	Market Basket Analysis on Sales Transactions for Micro, Small and Medium Enterprises Using Apriori Algorithm to Support Business Promotion Strategy in RDA Hijab	Apriori algorithm	The sales transaction dataset	Accuracy- 73%	Further research into creating an automatic data processing system (data preprocessing) so that the process is more efficient and the data processing process is shorter. In addition, creating a priori process with python language so that the execution process is faster.
22	Personalised Market Basket Prediction with Temporal Annotated Recurring Sequences	Apriori algorithm,FP-Growth algorithm	Coop dataset	Accuracy- 72%	Furthermore, we would like to exploit TARS for developing analytical services in other domains, such as mobility data, musical listening sessions and health data. Finally, in line with, it would be interesting to study if there is an improvement in the quality of the prediction if the user-centric models are exploited for developing a collective or hybrid predictive approach
23.	Data Mining Applications for Sales Information System Using Market Basket Analysis on Stationery Company	D. Apriori Algorithm	transaction dataset	Accuracy- 85%	Further Applications can perform data mining process based on existing sales data. Then program can help decide when to make the process of bundling.
24.	A Market Basket Analysis of Beef Calf Management Practice Adoption	Apriori algorithm	Oklahoma Beef Management dataset	Accuracy-70%	we can use market basket analysis to identify which combinations of research-based recommended calf

					health management practices are bundled frequently on the ranch, which are less frequently bundled, and to gain insight as to how to help producers both recognize their importance and assist them in practice implementation to increase profitability through joint adoption of practices.
25.	MARKET BASKET ANALYSIS USING FP GROWTH AND APRIORI ALGORITHM: A CASE STUDY OF MUMBAI RETAIL STORE	GROWTH AND APRIORI ALGORITHM	Transaction dataset	Accuracy-70%	A synthetic data set has been used with 77 items each for analysis. A set of association rules are obtained by applying Apriori algorithm and FP growth.
26.	Comparing unsupervised probabilistic machine learning methods for market basket analysis	(MH-RM) algorithm	real-world point-of-sale transactions	Accuracy-81%	To infer managerial implications we determine both probability increases of other categories and expected relative basket size increases due to promoting a product category. Product categories with high expected relative increases constitute candidates for a promotion whose the objective is to increase basket size.
27.	Market basket analysis by solving the inverse Ising problem: Discovering pairwise interaction strengths	Metropolis–Hasting (MH) algorithm	transactional sample dataset	Accuracy-75%	Furthermore, the use of the couplings to obtain a network representation of the purchasing system

	among products				and the hierarchical structure that emerges from the topology of that network, is extremely interesting because it allows to understand the link that exists between the energy of a state and its revealed hierarchy
28.	Application of Market Basket Analysis for the Visualization of Transaction Data Based on Human Lifestyle and Spectroscopic Measurements	Apriori algorithm	transactional sample datase		The method has been conventionally utilized in social sciences such as marketing and has not previously been implemented for use in metabolomics or metabonomics.
29.	Mobile Agent Based Market Basket Analysis on Cloud	apriori algorithm	transactional sample datase	Accuracy-75%	In future works we consider using more automation in the application by providing information without registration process and whole transaction will happen on the mobile number provided. We can integrate the routing of GPS to provide direction and distance measurement between shop and customer.
30.	Market basket analysis of crash data from large jurisdictions and its potential as a decision support tool	Apriori algorithm	non-intersection crash data	Accuracy-75%	These investigations are necessary for developing decision support tools based on association rule mining. As with the market basket analysis in the retail sector where it is up to the data owners

					to re-shelve their items based on the results, it would be up to the agencies to act on these broad patterns discovered from the data to develop policy initiatives and/or specific solutions for reduction in injuries and fatalities on roadways.
--	--	--	--	--	---

## Gap Identified :

For both of these algorithms predefined minimum support is needed to satisfy for identifying the frequent itemsets. But when the minimum support is low, a huge number of candidate sets will be generated which requires large computation. In this project, we plan to follow an approach that has been proposed to avoid this large computation by reducing the items in the dataset with top selling products.

## Existing Systems :

1. Online Store Product Recommendation System
2. Discovering Region-Based Association Rule in the IoT Environment

## Datasets Description & Sample data

### Data Set Information:

This data is taken from : <https://www.kaggle.com/datasets>

The data involved in any sale transaction in e-commerce websites such as amazon, such as data of items purchased, time of purchase, total sales volume, item price. E-commerce companies require additional data for managers to make strategic decisions that can increase company profits, such as the most sold product information, slightly sold products, and rarely sold products. To maintain inventory, it is essential to know the pattern of consumer spending that often occurs at these websites by analyzing the data of sales transactions. The placement of the product layout is still less accurate and optimal because it is only based on management's perception by categorizing the existing products and has not been reviewed from the consumer's point of view. So that, the researcher's initiative to try to provide solutions in the placement of the product layout.

## Sample Data :

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W
1	shrimp	almonds	avocado	vegetables	green grap	whole wec	yams	cottage ch	energy dri	tomato jui	low fat yo	green tea	honey	salad	mineral w	salmon	antioxydar	frozen sm	spinach	olive oil			
2	burgers	meatballs	eggs																				
3	chutney																						
4	turkey	avocado																					
5	mineral w	milk		energy bar	whole wh	green tea																	
6	low fat yogurt																						
7	whole wh	french fries																					
8	soup	light cream	shallot																				
9	frozen veg	spaghetti	green tea																				
10	french fries																						
11	eggs	pet food																					
12	cookies																						
13	turkey	burgers	mineral w	eggs	cooking oil																		
14	spaghetti	champagn	cookies																				
15	mineral w	salmon																					
16	mineral water																						
17	shrimp	chocolate	chicken	honey	oil	cooking oi	low fat yogurt																
18	turkey	eggs																					
19	turkey	fresh tuna	tomatoes	spaghetti	mineral w	black tea	salmon	eggs	chicken	extra dark	chocolate												
20	meatballs	milk	honey	french frie	protein bar																		
21	red wine	shrimp	pasta	pepper	eggs	chocolate	shampoo																
22	rice	sparkling water																					
23	spaghetti	mineral w	ham	body spray	pancakes	green tea																	
24	burgers	grated che	shrimp	pasta	avocado	honey	white wine	toothpaste															
25	eggs																						
26	parmesan	spaghetti	soup	avocado	milk	fresh bread																	
27	ground be	spaghetti	mineral w	milk	energy bar	black tea	salmon	frozen sm	escalope														
28	sparkling water																						
29	mineral w	eggs	chicken	chocolate	french fries																		

[https://docs.google.com/spreadsheets/d/1OJsJu58wObv9py9m6\\_COXT9\\_4gWlAG44/edit?usp=sharing&oid=107248981638858011160&rtpof=true&sd=true](https://docs.google.com/spreadsheets/d/1OJsJu58wObv9py9m6_COXT9_4gWlAG44/edit?usp=sharing&oid=107248981638858011160&rtpof=true&sd=true)

## References :

<https://www.sciencedirect.com/science/article/pii/S1877050916305208>

[https://www.researchgate.net/publication/228913454\\_An\\_Implementation\\_of\\_the\\_FP-growth\\_Algorithm](https://www.researchgate.net/publication/228913454_An_Implementation_of_the_FP-growth_Algorithm)

[https://www.researchgate.net/publication/329515220\\_Performance\\_Analysis\\_of\\_Apriori\\_and\\_FP-Growth\\_Algorithms\\_Association\\_Rule\\_Mining](https://www.researchgate.net/publication/329515220_Performance_Analysis_of_Apriori_and_FP-Growth_Algorithms_Association_Rule_Mining)

<https://www.irjet.net/archives/V8/i5/IRJET-V8I5831.pdf>

<https://www.ijeat.org/wp-content/uploads/papers/v9i2/B3186129219.pdf>

<https://ieeexplore.ieee.org.egateway.vit.ac.in/document/9410737?arnumber=9410737>

<https://dl.acm.org.egateway.vit.ac.in/doi/10.1145/3377170.3377226>

<https://www.proquest.com/docview/1789980619?accountid=38207>

[https://www.researchgate.net/publication/342382244\\_Consumer\\_purchase\\_patterns\\_based\\_on\\_market\\_basket\\_analysis\\_using\\_apriori\\_algorithms](https://www.researchgate.net/publication/342382244_Consumer_purchase_patterns_based_on_market_basket_analysis_using_apriori_algorithms)

[https://www.researchgate.net/publication/251422282\\_FPTree\\_Based\\_Algorithms\\_Analysis\\_FP\\_Growth\\_COFI-Tree\\_and\\_CT-PRO](https://www.researchgate.net/publication/251422282_FPTree_Based_Algorithms_Analysis_FP_Growth_COFI-Tree_and_CT-PRO)

[https://www.researchgate.net/publication/352384736\\_Online\\_Shopping\\_Do\\_Men\\_Behave\\_Differently\\_than\\_Women](https://www.researchgate.net/publication/352384736_Online_Shopping_Do_Men_Behave_Differently_than_Women)

[https://www.researchgate.net/publication/351112666\\_Uncovering\\_Modern\\_Clinical\\_Applications\\_of\\_Fuzzy\\_and\\_Fuzzy-Based\\_Formulas\\_A\\_Nationwide\\_Descriptive\\_Study\\_With\\_Market\\_Basket\\_Analysis](https://www.researchgate.net/publication/351112666_Uncovering_Modern_Clinical_Applications_of_Fuzzy_and_Fuzzy-Based_Formulas_A_Nationwide_Descriptive_Study_With_Market_Basket_Analysis)

<http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.299.6609&rep=rep1&type=pdf>

<http://www.ijrsred.com/volume2/issue2/IJSRED-V2I2P83.pdf>

<https://sci-hub.hkvisa.net/10.1109/ICAIBD49809.2020.9137436>

<https://arxiv.org/ftp/arxiv/papers/1403/1403.3948.pdf>

<https://osf.io/preprints/inarxiv/jpsfa/>

<https://ieeexplore.ieee.org.egateway.vit.ac.in/document/9689770>

<https://ieeexplore.ieee.org.egateway.vit.ac.in/document/8262592>

<https://www.proquest.com/docview/2571494191/fulltextPDF/76BE6BE78A804ADFPQ/1?accountid=38207>

<https://www.proquest.com.egateway.vit.ac.in/docview/1789980619/fulltextPDF/DB0BE13D8FDC4517PQ/1?accountid=38207>

<https://eds.p.ebscohost.com.egateway.vit.ac.in/eds/viewarticle/render?data=dGJyMPPp44rp2%2fdV0%2bnjisfk5Ie46bNPtquyS7Ok63nn5Kx94um%2bS62orUquqLA4trCyS7imtTi%2fw6SM8Nfsi9%2fZ8oHt5Od8u6ixUbasr1Gwr6SL59q7S%2bGtskiy3OFFt9q2fKuq5E6xo997r9erfeSrsH3h2OF9s6fgWPDb4oHxnOp57N27feyc4nq72%2bKL4%2bLhPvLX5VW%2fxKR57LPjPHb6nusqKtNsKSuRbarrk%2bvqLdIs6qwPuTI8IXf6rt%2b8%2bLqjOPu8gAA&vid=2&sid=3c7404fc-9d8d-4f63-ac1a-ef52ecbce51b@redis>

<https://www.sciencedirect.com.egateway.vit.ac.in/science/article/pii/S0378437119302171#sec5>

<https://pubs.acs.org.egateway.vit.ac.in/doi/10.1021/acs.analchem.5b04182>

<https://ieeexplore.ieee.org.egateway.vit.ac.in/document/7033341>

<https://www.sciencedirect.com.egateway.vit.ac.in/science/article/pii/S0925753507001877?via%3Dihub#aep-abstract-id5>





