Privacy Preservation in Federated Market Basket Analysis using Homomorphic Encryption

Sameeka Saini
Computer Science and Engineering,
Indian Institute of Technology
Roorkee, India
sameeka_s@cs.iitr.ac.in

Durga Toshniwal
Computer Science and Engineering,
Indian Institute of Technology
Roorkee, India
durga.toshniwal@cs.iitr.ac.in

Abstract

Traditional collaborative Machine Learning model collects private datasets from multiple clients at central location for analysis, raising privacy concerns and risks of data breaches. Methods like differential privacy, secure multiparty computation(SMC) and anonymization mitigate these risks. SMC entails significant computational and communication overhead, Differential Privacy often introduces a privacy-utility trade-off, requiring noisy or perturbed data and anonymization involves high risk of re-identification attacks. The proposed work encrypts frequent mining from multiple clients in FL using Homomorphic encryption. The approach allows computations to be performed on encrypted datasets, eliminating communication overhead, privacy-utility trade-offs etc. Experiments conducted on three different transactional datasets, employing metrics like entropy, mutual information, and KL divergence, concluded that encryption maintained data integrity, indicating no significant alteration in global model post-encryption, ensuring privacy preservation.

1 Introduction

Advancements in networking, storage and processing technology have enabled creation of ultra-large databases capable of capturing and storing unprecedented amount of information from diverse users. Artificial Intelligence (AI) and Machine Learning (ML) relies heavily on this huge data to efficiently learn, generalize patterns, make accurate predictions, and perform complex tasks. With increased data volume, ML algorithms gains deeper insights into underlying structures of problems, leading to improvement in their performance and reliability. Conventional centralized ML model requires sharing private client data with central server for model training, raising significant privacy concerns (Sushama et al., 2021) due to the sensitive information (Agrawal and Srikant, 2000). Thus, centralized approach raises significant privacy concerns as sensitive information is directly exposed to server. To address this, Google in 2016 (Konečnỳ et al., 2016) introduced Federated learning (FL) that enables collaborative training of a global model among multiple nodes without sharing their raw private data. Instead, only the model parameters are shared to ensure privacy. Figure 1 illustrates the fundamental workflow of federated learning. However, despite

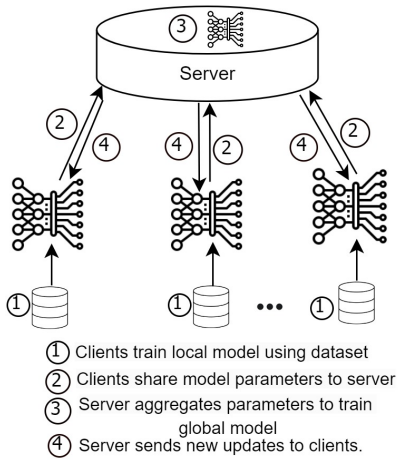

Figure 1: Federated learning.

these advancements, federated learning still poses privacy risks and challenges (Nasr et al., 2019). One significant challenge involves ensuring the security and privacy of local model parameters when they are shared with the central server for analysis. Additionally, federated learning requires frequent communication between central server and client devices, resulting in increased communication overhead, particularly when large number of clients are participating. Moreover, sharing of local model parameters by clients can attract eavesdroppers or adversaries, potentially intercepting and analyzing the data transmitted between clients and the server.

To address aforementioned challenges, various measures have been effectively employed, including anonymization, perturbation, differential privacy and blockchain based methods. In this paper,

Proceedings of the 1st International Conference on NLP & AI for Cyber Security, pages 109–118
July 29–30, 2024.

we leverage the property of Homomorphic encryption (Gentry, 2009) within the context of federated market basket analysis, which enables computations to be performed on encrypted data without decryption. The contribution of our study are:

1. We have utilized the Apriori and FP-Growth algorithms to individually find frequent items and rules by each client in federated learning. The advantage of extracting rules and items in federated learning lies in preserving the privacy of each client's data while still allowing for collective learning across multiple clients.

2. To address privacy preservation, we have applied Homomorphic Encryption to the frequent items and rules mined by each client, ensuring that privacy is maintained throughout the federated learning market basket analysis.

3. To experimentally validate our proposed work, we have assessed it on three transactional datasets. We utilized entropy, mutual information, and Kullback–Leibler (KL) divergence metrics to evaluate the integrity of the encrypted data, while also examining the execution time involved in whole process.

2 Literature review

To protect private data and ensure robust privacy in federated learning, researchers have developed several techniques including Anonymization (Li et al., 2019), Differential privacy (Abadi et al., 2016) (Wei et al., 2020), Secure multiparty computations(Mugunthan et al., 2019) and Blockchain-based methods (Zhao et al., 2020).

Association rule mining (Modi and Patil, 2016), based on Diffie-Hellman problem along with elliptic curve and digital signature was proposed to improve trustworthiness of data exchange between clients eliminating the trusted third party. However, it faces scalability issue and computational cost for large number of participating clients. Two Association rule mining (Chahar et al., 2017) was proposed for horizontally partitioned database. First scheme utilizes Elliptic curve cryptosystem that secure the site information and second scheme relies on Shamir secret sharing method that effectively addresses the vulnerability against collusion attacks. Nevertheless first scheme was susceptible to collusion attack and second was having higher computational cost. SVSM (Wang et al., 2018) address the challenge of frequent itemset mining in transactional data using local differential privacy. However, scalability issues persist.

A centralized FL framework (Molina et al., 2021) was designed for mining association rules from electronic healthcare records, ensuring global accuracy while reducing computational cost. FedFPM (Wang et al., 2022) is a local differential privacy based approach for mining frequent items efficiently with privacy. PPDMF (Wu et al., 2023) proposed for joint venture industrial collaboration for mining of high utility itemsets from multiple datasets without directly sharing the private data. The proposed method results displayed that approach is having high accuracy while preserving privacy. FedFIM (Chen et al., 2023) and FedFIM_AES uses AES encryption to rapidly mine frequent items along with adding noise in the fed_avg. FL based mining algorithm (Hong et al., 2023) considered client server method where clients possess large and diverse datasets, and the server aggregates results from each client.

While above techniques aim to protect privacy, they still have some limitations. Secure Multiparty Computation (SMC) often involves high communication overhead and intricate key sharing mechanisms. Differential privacy presents challenges in selecting an appropriate epsilon value (Lee and Clifton, 2011), impacting the accuracy of mining results. Anonymization, although effective, may not always guarantee complete privacy and Blockchain in privacy-preserving federated learning (FL) suffers from scalability issues. However, Homomorphic encryption offers a promising solution by enabling computations directly on encrypted data without decryption. This minimizes communication overhead, eliminates the need for extensive key sharing, and provides a more efficient and secure approach to privacy-preserving data mining.

3 Preliminaries

3.1 Frequent mining algorithms

Frequent item mining and Market Basket analysis identify frequent items and association rules from transactional datasets. Consider a transactional dataset, I={A, B, C,..F}, where A, B are items in the dataset and each client's data includes a subset of items from I and a pattern p, representing an item or combination of items. A pattern p is frequent if it appears in a sufficient proportion of client data, exceeding a threshold f. The support of p determines its frequency by measuring the

proportion of transactions containing p, guiding tasks like generating association rules or recommending items. Classical algorithms for frequent item mining are Apriori (Agrawal et al., 1993) and FP-Growth (Han et al., 2000). Both algorithms utilize a minimum support threshold to identify frequent itemsets, with Apriori employing an iterative candidate generation and pruning approach, while FP-Growth constructs a compact tree structure to efficiently mine frequent itemsets without explicit candidate generation.

3.2 Homomorphic encryption

Homomorphic encryption (HE) (Rivest et al., 1978) allows computations to be performed directly on encrypted data. Various researchers have used HE in various applications. In cloud environments (Fahsi et al., 2015) HE framework is used for private information retrieval to keep users safe against unauthorized access of private data. (Brakerski et al., 2014) HE yields cipher texts using specific calculations that create encrypted output but with a prerequisite for reverse computation techniques to yield plain text back. Homomorphic encryption has the property that allows operations to be performed on encrypted texts. Given E= Encryption, D= Decryption, σ= Security parameter, A= Homomorphic property, K_e= Encryption Key, ciphertexts (c_1, c_2) encrypted on messages (m_1, m_2), a new ciphertext c_3 such that $\forall m_1, m_2 \in M$ holds only when
$m_3 = m_1 + m_2$,
$c_1 = E(\sigma, (K_e, m_1))$, and
$c_2 = E(\sigma, (K_e, m_2))$ such that:
$Prob[D(A(\sigma, K_e, c_1, c_2)) \neq m_3]$ is negligible.

Different versions of Homomorphic encryption, full homomorphic encryption (FHE), partial homomorphic (PHE) and somewhat homomorphic encryption (SHE)(Fan and Vercauteren, 2012) exists. Figure 2 shows the comparison of PHE, FHE and SHE.

	FHE	PHE	SHE
Computations on encrypted data	Full	Either Add or Multiply	Limited No upto a threshold
Complexity	Complex	Simpler	Simpler
Computational Overhead	More	Less	Less
Balance Funtionality & Efficiency	Less	More	Less
Limited Funtionality	NO	YES	YES
End to End Security	YES	NO	NO

Figure 2: Comparison of FHE, PHE and SHE.

Pailler encryption (Paillier, 1999) is a type of public key based partial Homomorphic encryption that enables computations on encrypted data (either addition or multiplication) (Guo et al., 2024). It consist of four main steps:

- Key generation: From two large prime numbers p and q, generation of public key p_k and private key s_k is performed. Compute $N = pq$ and $\lambda = \text{lcm}(p-1, q-1)$, where $\text{lcm}(\cdot)$ denotes the least common multiple function. Random number g is selected so $\lambda = \gcd(L(g^\lambda \mod N^2), N) = 1$, where $\gcd(\cdot)$ signifies the greatest common divisor function and $L(x) = \frac{x-1}{N}$ with $x \in \mathbb{Z}_{N^2}$ and $x \equiv 1 \pmod N$. It generates public key as $pk = \{N, g\}$ and private key as $s_k = \lambda$.
- Encryption: Message m in $\mathbb{Z}_N$ selects a random number r in $\mathbb{Z}_{N^2}$ and computes $c = [m]_{pk} = g^m r \cdot r^N \mod N^2$.
- Decryption: For c, m and private key λ, as $m = \frac{L(c^\lambda \mod N^2)}{L(g^\lambda \mod N^2)} \mod N$.
- Addition:Two ciphertexts $[m_1]_{pk}$ and $[m_2]_{pk}$ we have $[m_1]_{pk} \cdot [m_2]_{pk} = [m_1 + m_2]_{pk}$
Because:
$[m_1]_{pk} \cdot [m_2]_{pk} = g^{m_1} r_1^N \mod N^2 \cdot g^{m_2} r_2^N \mod N^2$
$= g^{(m_1+m_2)} \cdot r_1 \cdot r_2^N \mod N^2$
$=[m_1 + m_2]_{pk}$
for multiplication:$([m_1]_{pk})^2 = [m_1 \cdot m_2]_{pk}$

3.3 Problem statement

We consider a cooperative scenario of homogeneous and horizontal partitioned dataset where p parties are semi-honest and aims to collaboratively find globally frequent itemsets without disclosing their identities. The parties uses classical mining algorithms like Apriori or FP-growth to discover frequent items and association rules. Our research approach focuses on privacy preservation in the federated learning setting, considering existing methodology limitations and leveraging Homomorphic encryption for privacy.

4 Proposed methodology

4.1 Limitation of existing work

Centralized methods for collaborative learning, while straightforward in implementation, present significant privacy concerns. In these methods, all raw data is collected and stored on a central server,

making it vulnerable to data breaches and unauthorized access (Liu et al., 2024) (Drainakis et al., 2023). The lack of data privacy can lead to the exposure of sensitive information. Additionally, this often faces scalability issues as the volume of data increases. As a solution federated learning environment is used (Rodríguez-Barroso et al., 2023) where there is no need to share the whole dataset to the server for analysis.

Differential privacy (DP) techniques add noise to the data to protect individual entries. Despite their effectiveness in preserving privacy, they have some limitations (Zhao et al., 2019). There is a utility-privacy trade-off, where higher privacy often means more noise, degrading result quality and accuracy. DP may also require extra communication rounds to ensure the noise added is effective, leading to increased overhead in FL settings. In contrast, Homomorphic encryption enables direct computations on encrypted data, eliminating the need for additional rounds of communication. This reduces communication overhead while preserving utility, accuracy, and ensuring strong privacy.

4.2 Proposed work

In federated learning settings, concerns about data privacy and security arise when data from multiple clients is aggregated at the server for model training. Particularly in Market Basket Analysis, where insights into consumer behavior are gleaned from transactional data, preserving the confidentiality of sensitive information is paramount. After mining frequent items and association rules, the data is being shared with the server for updating the global model. After the data is shared with the server, there's a potential risk of adversaries gaining access to private information or even reconstructing datasets from the shared rules and frequent items. This highlights the critical need for robust privacy-preserving techniques in federated learning settings.

Figure 3 depicts proposed methodology where clients individually train their local models using Apriori or FP-Growth algorithms to discover frequent items and associations rules. Subsequently, each client shares their results with server for global model aggregation in encrypted form. For encryption, we employ partial Homomorphic encryption supplemented by scaling and hashing techniques. The support values are in floating-point format, hence appropriately scaled before encryption

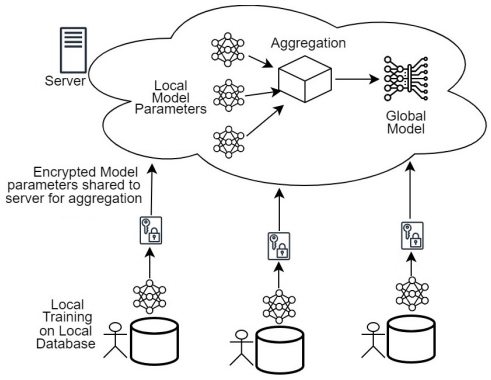

Figure 3: Proposed methodology.

using the Pailler Homomorphic encryption scheme. To secure specific item names and their support values, a secure hashing function, namely SHA-256, is utilized in conjunction with a dictionary. Upon receiving the encrypted results, the server performs aggregation (summing the values of support and confidence for respective frequent items and rules) on the encrypted data, followed by decryption, rescalling to their original values and dividing it by the number of clients. A comprehensive understanding of the systematic review methodology can be gained by referring to Algorithm 1 and Algorithm 2.

Our main concern lies in mitigating the risk of sensitive information leakage during transmission from clients to the server. In contrast to existing schemes, our proposed method circumvents high computational costs while exchanging frequent information between clients and server. Furthermore, the encryption process does not introduce any additional random noise or values to original support values. We have conducted experimental evaluations of the proposed method using metrics such as Mutual Information, Entropy and Kullback-Leibler (KL) divergence.

The product recommendation algorithm recommends products to clients after updating the global model from the server. It begins by filtering frequent items and association rules based on the items of interest and then sorts these rules using a chosen metric like support for frequent itemsets and confidence or lift for association rules. The filtered and sorted items and rules are then used to generate recommendations. The process involves determining the support, confidence, and lift of association rules, where support indicates the frequency of occurrence, confidence signifies the likelihood of purchasing one item given another, and lift measures the strength of association compared

Algorithm 1 Federated Market Basket with Homomorphic Encryption: FedMBHE

Data: D_i as transactional dataset for client i.
Input: min_supp as minimum support threshold.
Notations:
F_i:set of frequent itemsets for client i.
R_i:set of association rules for client i.
Enc(x):Paillier encryption function for plaintext x.
Dec(c):Paillier decryption function for ciphertext c.
Hash(x):SHA256 hashing function for input x.
Scale(x):scaling function converting support to integers.

procedure LOCALMODELTRAINING(D_i)
for each client i **do**
F_i = MiningAlgorithm(D_i, min_supp))
R_i = AssociationRuleGeneration(F_i, min_supp)
Enc(F_i) = Encrypt(F_i)
Enc(R_i) = Encrypt(R_i)
HashedX$_i$ = Hash(F_i)
ScaledS$_i$ = Scale($SupportValues(F_i)$)
end for
end procedure

procedure GLOBALMODELUPDATION($EncF_i$, $EncR_i$)
Enc(F) =Union($EncF_i$)
Enc(R) =Union($EncR_i$)
(F) =Decrypt(Enc_F)
(R) =Decrypt(Enc_R)
GF$_{\text{Item}}$ = ExtractFrequentItemsets(F, min_supp)
G$_{\text{Rules}}$ = ExtractAssociationRules(R, min_confi)
DivideByNumberOfClients(GF$_{\text{Item}}$, G$_{\text{Rules}}$)
ShareResultsWithClients(GF$_{\text{Item}}$, G$_{\text{Rules}}$)
end procedure

to random chance. By filtering and sorting the rules based on client interests and chosen metrics, the algorithm tailors recommendations to individual preferences, ultimately enhancing the user experience and promoting relevant product engagement.

5 Results

5.1 Experimental setup

5.1.1 Dataset & implementation

The proposed methodology uses Homomorphic encryption to provide privacy preservation in FL Market Basket Analysis. We tested the proposed work on three transactional datasets mainly Grocery[1], Telecom[2], and Retail[3] datasets available at kaggle.

Table 1 presents the sample transactional data for each dataset and Table 2 shows the characteristic of the experimental datasets. The proposed methodology was implemented in Python, considering 5 clients for our experiment. We evenly distributed the datasets among the clients horizontally. Each client in the grocery dataset comprises 1967 total

[1]Kaggle - Grocery dataset
[2]Kaggle - Telecom dataset
[3]Kaggle - Retail Transactions Dataset

Algorithm 2 Product Recommendation

Input: All available items (I), Interested items (I_{interest}), Global frequent items (GF$_{\text{Item}}$) and Global association rules (G$_{\text{Rules}}$).
Output: Set of Recommended products(R_{product}).
Notations:
S($X \rightarrow Y$): support of association rule $X \rightarrow Y$,
C($X \rightarrow Y$): confidence of association rule $X \rightarrow Y$,
L($X \rightarrow Y$): lift of association rule $X \rightarrow Y$,
F_{rules} = { Filtered rules} and S_{rules} = { Sorted rules}.
Begin
// Generate product recommendation after Filtering and sorting rules based on items of interest and chosen metric (confidence, support, or lift)
F_{rules} = FilterRules(I, I_{interest})
S_{rules} = SortRules(G$_{\text{Rules}}$, ChosenMetric)
R_{product} = GenerateRecommendations(S_{rules},GF$_{\text{Item}}$)
end

transactions, while in the telecom dataset, they possess 1500 entries, and in the retail dataset, each client is associated with 6000 transactional entries. The min_support threshold for Apriori and FP-Growth algorithm was set to 0.3% for telecom and grocery dataset, and at 0.03% for retail dataset. Association rules were evaluated using the lift metric, with thresholds of 0.01 and 0.1. From the lightphe library of Python Pailler encryption scheme and for Hashing SHA256 was used.

5.1.2 Evaluation metrics

The encrypted and original values of support were tested by metrics such as:

Entropy: It measures uncertainty or randomness in a probability or data distribution. For a discrete random variable with probability mass function $p(x)$:

$$H(X) = -\sum_x p(x) \log p(x) \tag{1}$$

For a continuous random variable with probability density function $f(x)$:

$$H(X) = -\int_{-\infty}^{\infty} f(x) \log f(x)\, dx \tag{2}$$

Mutual information: It measures the amount of information shared between two data variables.

$$I(X;Y) = \sum_{x,y} p(x,y) \log \frac{p(x)p(y)}{p(x,y)} \tag{3}$$

Kullback-Leibler (KL) divergence: It measures the difference between two data distributions.
For discrete:

$$D_{\text{KL}}(P||Q) = \sum_x P(x) \log \frac{Q(x)}{P(x)} \tag{4}$$

For continuous:

$$D_{\text{KL}}(P||Q) = \int_{-\infty}^{\infty} P(x) \log \frac{Q(x)}{P(x)}\, dx \tag{5}$$

T_{id}	Items	T_{id}	Items
{1}	a,b,a,c	{1}	a,b,f
{2}	b,c,f	{2}	d,e
{3}	a,e	{3}	a,b,c
{4}	c,f,e,b	{4}	c,a,b,e,f

Table 1: Sample transactional dataset.

Dataset	Trans	Item	Avg_Len	Density%
Grocery	9835	2201	4	20.03
Telecom	7500	1268	4	30.87
Retail	30000	116626	13	1.09

Table 2: Characteristics of experimental datasets.

5.2 Performance evaluation

5.2.1 Mining process analysis

The Apriori and FP-growth mining algorithm were applied to three transactional datasets to mine frequent items and association rules for all clients individually. Table 3 shows the number of frequent items mined at different support thresholds for all transactional datasets. Table 4 and 5 show the number of association rules mined at different lift and confidence thresholds respectively, for all transactional datasets, for a fixed min_support value. After this, each client encrypts them using Pailler Homomorphic encryption.

5.2.2 Privacy analysis

The encrypted and without encrypted support values for all three dataset for frequent items mining were tested to measure the privacy and integrity of the resulted averaged frequent items. Figure 4 shows the Entropy comparison for available datasets for the Apriori and FP-Growth with and without Encryption. Figure 5 shows Mutual Information comparison and KL divergence comparison for available datasets on the Apriori and FP-Growth algorithms.

Dataset	Client	0.1%	0.5%	1%	5%	10%
Groceries	1	108012	1330	383	32	8
	2	66423	1097	340	27	8
	3	54849	1358	413	33	10
	4	30766	912	281	28	8
	5	53297	1177	364	32	8
Telecom	1	45595	842	293	29	7
	2	17149	939	331	28	7
	3	12081	823	296	29	9
	4	12252	795	294	29	7
	5	7886	567	219	24	7
Retail	1	2344	81	81	1	0
	2	2351	81	81	1	0
	3	2333	81	81	1	0
	4	2384	81	81	1	0
	5	2387	81	81	1	0

Table 3: No of Frequent items found for all datasets using apriori and fp-growth at different support thresholds.

Dataset	Client	1%	5%	10%	50%	100%
Grocery	1	20612	20612	20612	20606	20248
	2	12816	12816	12816	12808	12564
	3	16778	16778	16778	16766	16388
	4	8932	8932	8932	8928	8622
	5	13830	13830	13830	13826	13506
Telecom	1	7904	7904	7904	7898	7588
	2	9966	9966	9966	9958	9692
	3	7104	7104	7104	7096	6798
	4	7274	7274	7274	7266	6906
	5	3898	3898	3898	3598	2812
Retail	1	15172	15172	15172	14506	11008
	2	15384	15384	15384	14712	11258
	3	15646	15646	15646	14994	11490
	4	15678	15678	15678	15004	11580
	5	15762	15762	15762	15122	11658

Table 4: No of association rules found for datasets using apriori and fp-growth with min_supp=0.3% (for retail: min_supp=0.03%) and metric=lift.

Dataset	Client	1%	5%	10%	50%	100%
Grocery	1	20612	15539	11647	2334	118
	2	12816	9430	6870	1084	35
	3	16778	12094	8956	1532	33
	4	8932	6464	4681	606	13
	5	13830	10100	7323	1034	26
Telecom	1	7904	5814	4272	566	27
	2	9966	7389	5519	880	40
	3	7104	5116	3818	494	18
	4	7274	5336	3917	540	29
	5	3898	2787	1967	162	7
Retail	1	11597	4959	4375	198	42
	2	11652	5054	4497	204	52
	3	11712	5201	4625	217	49
	4	11627	5217	4650	209	58
	5	11651	5268	4681	202	49

Table 5: No of association rules found for datasets using apriori and fp-growth with min_supp=0.3% (for retail: min_supp=0.03%) and metric=confidence.

The Entropy values of original support indicate significant diversity or variability in the frequency of items, suggesting a higher level of uncertainty or randomness. Conversely, the entropy values for the encrypted support values show a slightly lower level of uncertainty, possibly due to the regularization or compression introduced during encryption process. Mutual information value quantifies the shared information between the original and encrypted distributions, with higher values indicating a stronger relationship or dependency between the distributions. Regarding KL divergence, which measures the difference between distributions, a value close to 0 suggests a smaller difference between original and encrypted distributions, implying a higher degree of similarity. Table 6 gives the summary of evaluation metric parameters tested on all three datasets.

Dataset	Algo	O_Ent	E_Ent	MI	KL_Div
Grocery	Apriori	8.36	8.40	2.74	0.16
	FPGr	8.36	8.40	2.74	0.16
Telecom	Apriori	7.74	7.70	2.82	0.16
	FPGr	7.74	7.68	2.84	0.16
Retail	Apriori	7.53	7.52	2.01	0.37
	FPGr	7.53	7.65	1.83	0.38

Table 6: Original entropy, encrypted entropy, mutual information and kullback leibler (KL) divergence for all datasets for apriori & fp-growth algorithm.

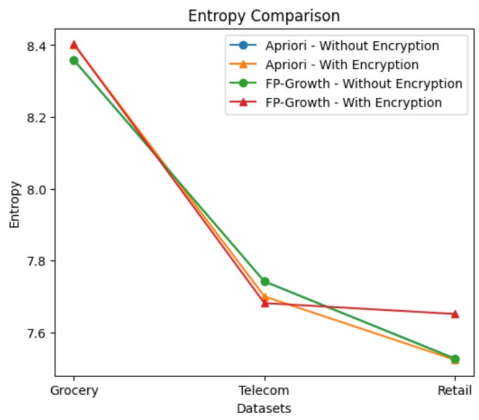

Figure 4: Entropy comparison for available datasets for apriori and fp-growth with and without encryption.

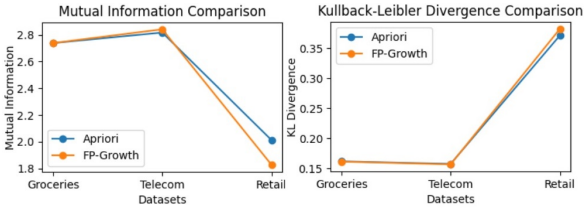

Figure 5: Mutual information & kullback-leibler divergence comparison for available datasets for apriori and fp-growth.

5.2.3 Execution time analysis

Across all datasets, execution time was measured for both regular computation and computation with Homomorphic encryption applied for privacy preservation. Figures 6 and 7 depict the execution time and encryption time associated with mining frequent items across all datasets, respectively. The time differences between encryption and non-encryption scenarios varied depending on the algorithm used. For FP-growth, the time was greater without encryption and less with encryption across all datasets. This can be attributed to the nature of algorithm, which constructs a compact data structure (FP-tree) during the initial pass over the dataset, making subsequent frequent itemset mining more efficient. When encryption is applied, the compact structure aids in reducing computational overhead associated with encryption operations, resulting in shorter execution time. Conversely, for Apriori, time with encryption was slightly greater than without encryption for all datasets.

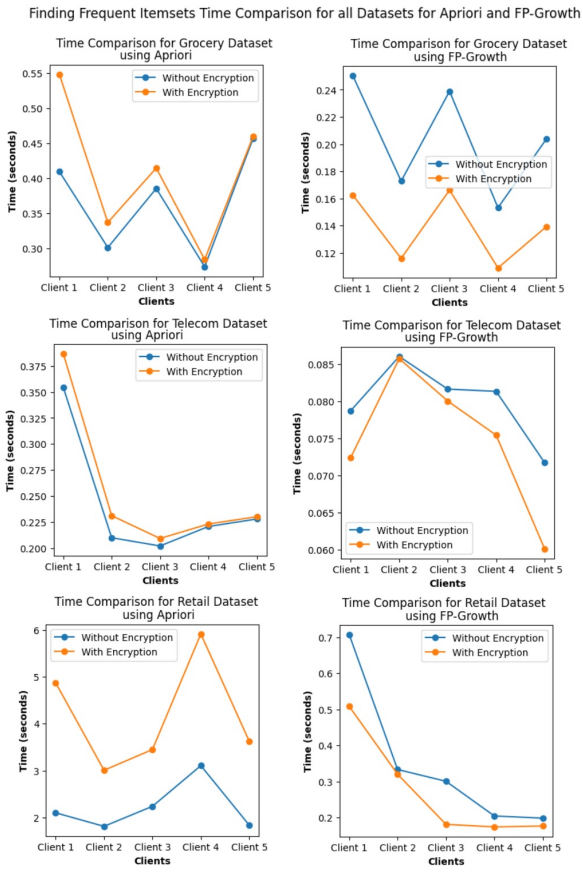

Figure 6: Execution time comparison for all datasets using apriori and fp-growth (frequent items).

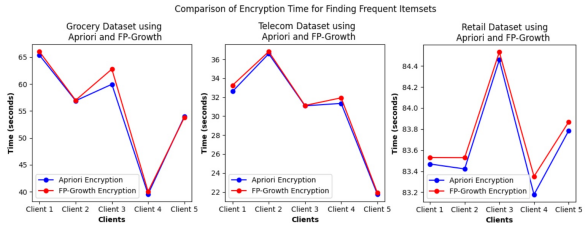

Figure 7: Encryption time comparison for all datasets using apriori and fp-growth (frequent items).

Figures 8, 9, and 10 gives execution time associated with all datasets for mining association rules. In grocery and telecom datasets, without encryption, Apriori tends to take more time compared to FP-Growth, reflecting its inherent computational complexity in generating association rules. However, with encryption applied, both algorithms exhibit almost similar time requirements. Conversely, for the retail dataset, both with and without encryption, Apriori consistently requires slightly more time compared to FP-Growth across all metrics - support, lift, and confidence. Apriori's iterative nature and the need to repeatedly scan the dataset

for candidate itemsets make it more computationally intensive, especially with larger datasets like retail. On the other hand, FP-Growth's tree-based approach allows for more efficient frequent itemset mining, resulting in shorter Execution times. These findings shows nuanced impact of encryption on different algorithms and highlight the importance of considering algorithmic characteristics when applying privacy-preserving techniques in FL.

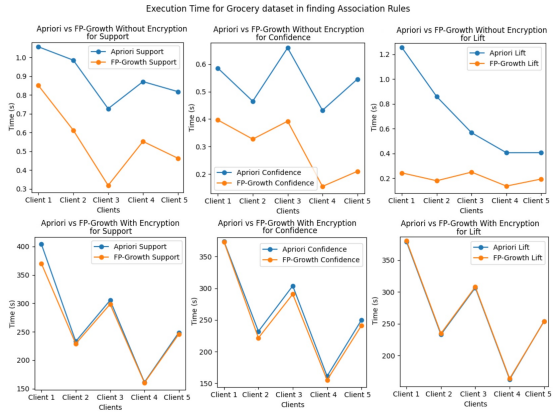

Figure 8: Execution time for grocery dataset (association rules).

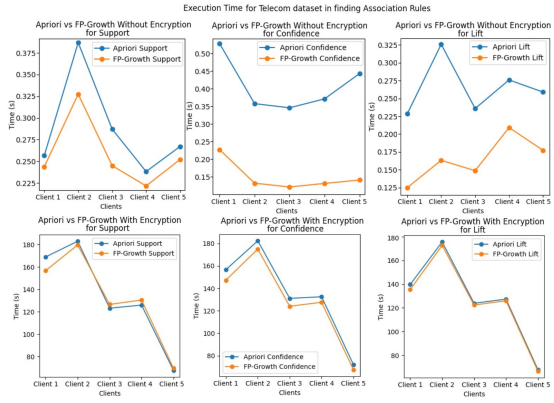

Figure 9: Execution time for telecom dataset (association rules).

For the frequent itemsets, decryption times are relatively lower compared to those for association rules. Decryption times for the grocery and retail datasets tend to be higher compared to the telecom dataset, reflecting the larger size of association rules generated with min_support and other metric such as lift and confidence. The slightly higher decryption times for Apriori compared to FP-Growth across all datasets and metrics can be attributed to the iterative nature of Apriori and the need for repeated decryption operations during candidate itemset generation. In contrast, FP-Growth's tree-based approach requires fewer decryption operations, resulting in slightly lower decryption times. Figure 11 depicts the decryption time for all datasets using Apriori and FP-Growth algorithms in decryption of frequent items and association rules.

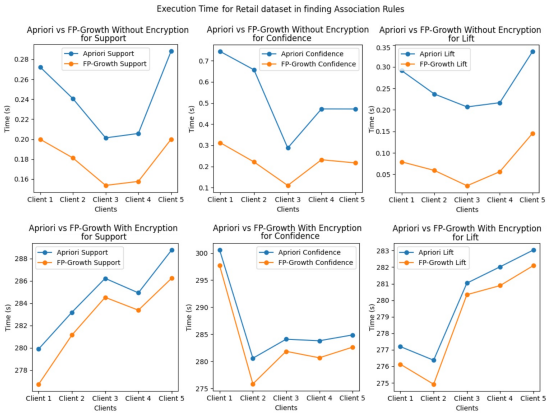

Figure 10: Execution time for retail dataset (association rules).

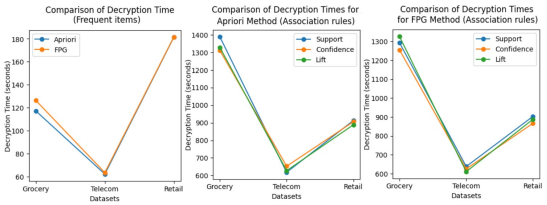

Figure 11: Comparison of decryption time between apriori and fp-growth methods across all datasets.

6 Conclusion and future work

The proposed methodology explores utilization of Federated Learning combined with Homomorphic encryption for market basket analysis, frequent itemset mining, and product recommendation. By employing the Apriori and FP-Growth algorithms on transactional datasets, frequent itemsets, association rules, and product recommendations were efficiently extracted. Homomorphic encryption was then applied to ensure the confidentiality and integrity of client results during transmission to the server for training the global model, thereby preserving privacy. Additionally, analysis using metrics such as entropy, mutual information, and KL divergence indicated that the data remained closely aligned with the original after encryption, unlike some other methods. Furthermore, the encryption and decryption time were minimal, and computational complexities were reduced, as Homomorphic encryption allows computations to be performed without decryption and does not require excessive data transmission from client to server. For future work, exploring other variants of Homomorphic encryption schemes on alternative frequent mining algorithms could be beneficial.

Limitations

In the context of large-scale federated learning, the Paillier encryption scheme, while effective for data privacy and security, presents challenges. Particularly with sizable datasets containing numerous entries, the encryption process may become computationally demanding, leading to longer encryption times. This could hinder the efficiency and scalability of federated learning systems, especially when managing numerous clients and extensive datasets. Transmitting encrypted data from multiple clients to central server for aggregation can incur high communication costs,if dataset and frequency of updates are substantial. Ensuring efficient co-ordination and synchronization among multiple clients can become crucial. Additionally, the data distribution among clients, divided equally from the same dataset, follows a non-IID pattern, further complicating the scenario.

Ethics Statement

This research strictly adheres to ethical guidelines and standards to uphold the integrity and confidentiality of data. The synthetic dataset used in this study, obtained from Kaggle. The study follows established ethical principles in data analysis and reporting, in accordance with guidelines governing research involving synthetic datasets and computational analysis. The dataset used in this study can be accessed from kaggle.

Conflict-of-interest Statement

The authors declare that they have no conflicts of interest. Additionally, we have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

I would like to express my sincere gratitude to Prof. Durga Toshniwal for her invaluable guidance and feedback throughout this research work. Her insights have significantly contributed to shaping the direction and quality of this study. I am also thankful to IIT Roorkee for providing access to research papers and other hardware and software requirements, which were essential for conducting the experiments and analyses presented in this paper. Lastly, I extend my appreciation to my family, friends, and colleagues for their unwavering encouragement and understanding throughout this endeavor.

References

Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. 2016. *Deep learning with differential privacy*.

Rakesh Agrawal, Tomasz Imieliński, and Arun Swami. 1993. *Mining association rules between sets of items in large databases*.

Rakesh Agrawal and Ramakrishnan Srikant. 2000. *Privacy-preserving data mining*.

Zvika Brakerski, Craig Gentry, and Vinod Vaikuntanathan. 2014. *(Leveled) fully homomorphic encryption without bootstrapping*. 3. ACM New York, NY, USA.

Harendra Chahar, BN Keshavamurthy, and Chirag Modi. 2017. *Privacy-preserving distributed mining of association rules using Elliptic-curve cryptosystem and Shamir's secret sharing scheme*. 12. Springer.

Yao Chen, Wensheng Gan, Yongdong Wu, and S Yu Philip. 2023. *Privacy-preserving federated mining of frequent itemsets*, volume 625. Elsevier.

Georgios Drainakis, Panagiotis Pantazopoulos, Konstantinos V Katsaros, Vasilis Sourlas, Angelos Amditis, and Dimitra I Kaklamani. 2023. *From centralized to Federated Learning: Exploring performance and end-to-end resource consumption*, volume 225. Elsevier.

Mahmoud Fahsi, Sidi Mohamed Benslimane, and Amine Rahmani. 2015. *A framework for homomorphic, private information retrieval protocols in the cloud*. 5. Modern Education and Computer Science Press.

Junfeng Fan and Frederik Vercauteren. 2012. *Somewhat practical fully homomorphic encryption*.

Craig Gentry. 2009. *A fully homomorphic encryption scheme*. Stanford university.

Yuqi Guo, Lin Li, Zhongxiang Zheng, Hanrui Yun, Ruoyan Zhang, Xiaolin Chang, and Zhixuan Gao. 2024. *Efficient and Privacy-Preserving Federated Learning based on Full Homomorphic Encryption*.

Jiawei Han, Jian Pei, and Yiwen Yin. 2000. *Mining frequent patterns without candidate generation*. 2. ACM New York, NY, USA.

Tzung-Pei Hong, Ya-Ping Hsu, Chun-Hao Chen, and Jimmy Ming-Tai Wu. 2023. *A Federated Mining Framework for Complete Frequent Itemsets*.

Jakub Konečnỳ, H Brendan McMahan, Felix X Yu, Peter Richtárik, Ananda Theertha Suresh, and Dave Bacon. 2016. *Federated learning: Strategies for improving communication efficiency*.

Jaewoo Lee and Chris Clifton. 2011. *How much is enough? choosing ε for differential privacy*.

Wenqi Li, Fausto Milletarì, Daguang Xu, Nicola Rieke, Jonny Hancox, Wentao Zhu, Maximilian Baust, Yan Cheng, Sébastien Ourselin, M Jorge Cardoso, et al. 2019. *Privacy-preserving federated brain tumour segmentation*.

Bingyan Liu, Nuoyan Lv, Yuanchun Guo, and Yawen Li. 2024. *Recent advances on federated learning: A systematic survey*. Elsevier.

Chirag N Modi and Ashwini R Patil. 2016. *Privacy preserving association rule mining in horizontally partitioned databases without involving trusted third party (TTP)*.

Carlos Molina, Belen Prados-Suarez, and Beatriz Martinez-Sanchez. 2021. *Federated Mining of Interesting Association Rules Over EHRs*. IOS Press.

Vaikkunth Mugunthan, Antigoni Polychroniadou, David Byrd, and Tucker Hybinette Balch. 2019. *Smpai: Secure multi-party computation for federated learning*, volume 21.

Milad Nasr, Reza Shokri, and Amir Houmansadr. 2019. *Comprehensive privacy analysis of deep learning: Passive and active white-box inference attacks against centralized and federated learning*.

Pascal Paillier. 1999. *Public-key cryptosystems based on composite degree residuosity classes*.

Ronald L Rivest, Len Adleman, Michael L Dertouzos, et al. 1978. *On data banks and privacy homomorphisms*, volume 4. Citeseer.

Nuria Rodríguez-Barroso, Daniel Jiménez-López, M Victoria Luzón, Francisco Herrera, and Eugenio Martínez-Cámara. 2023. *Survey on federated learning threats: Concepts, taxonomy on attacks and defences, experimental study and challenges*, volume 90. Elsevier.

C Sushama, M Sunil Kumar, and P Neelima. 2021. *WITHDRAWN: Privacy and security issues in the future: A social media*. Elsevier.

Tianhao Wang, Ninghui Li, and Somesh Jha. 2018. *Locally differentially private frequent itemset mining*.

Zibo Wang, Yifei Zhu, Dan Wang, and Zhu Han. 2022. *FedFPM: A unified federated analytics framework for collaborative frequent pattern mining*.

Kang Wei, Jun Li, Ming Ding, Chuan Ma, Howard H Yang, Farhad Farokhi, Shi Jin, Tony QS Quek, and H Vincent Poor. 2020. *Federated learning with differential privacy: Algorithms and performance analysis*, volume 15. IEEE.

Jimmy Ming-Tai Wu, Qian Teng, Shamsul Huda, Yeh-Cheng Chen, and Chien-Ming Chen. 2023. *A privacy frequent itemsets mining framework for collaboration in IoT using federated learning*, volume 19. ACM New York, NY.

Jingwen Zhao, Yunfang Chen, and Wei Zhang. 2019. *Differential privacy preservation in deep learning: Challenges, opportunities and solutions*, volume 7. IEEE.

Yang Zhao, Jun Zhao, Linshan Jiang, Rui Tan, Dusit Niyato, Zengxiang Li, Lingjuan Lyu, and Yingbo Liu. 2020. *Privacy-preserving blockchain-based federated learning for IoT devices*, volume 8. IEEE.