Designing a Novel Fuzzy Association Rule Mining Algorithm for Federated Environments

Carlos Fernandez-Basso[1,2,3][0000−0002−8809−8676], M. Dolores Ruiz[1,3][0000−0003−1077−3173], and Maria J. Martin-Bautista[1,3][0000−0002−6973−477X]

[1] Research Centre for Information and Communications Technologies (CITIC-UGR), University of Granada, Granada 18014, Spain
[2] Causal Cognition Lab, Division of Psychology and Language Sciences, University College London, London, United Kingdom
[3] Department of Computer Science and Artificial Intelligence, University of Granada, Granada 18014, Spain
[4] * Corresponding author cjferba@decsai.ugr.es

Abstract. This paper introduces a novel fuzzy association rule mining algorithm explicitly developed for federated environments. With exponential growth in datasets and increasing data privacy concerns, solutions such as federated learning have become at the forefront of secure and efficient data analysis. However, efficiently finding meaningful and relevant patterns in data across decentralized databases remains challenging. To address this, we propose integrating fuzzy logic with association rule mining in a federated setting. The ability of fuzzy logic to handle uncertainty and nuance in data combined with the distributed data mining process of federated systems, creates an efficient, secure, and powerful tool for pattern discovery. Our proposed algorithm respects data privacy and effectively manages communication overhead, an innate challenge in federated systems. Experimental results demonstrate the efficacy of the proposed algorithm. The system has significant implications for the healthcare sector, where data volume and privacy concerns are paramount.

Keywords: Association rules · federated learning · data mining · privacy preserving

1 Introduction

The burgeoning field of data mining has become an integral part of extracting meaningful patterns and relationships from vast datasets. Among the various data mining techniques, association rule mining has proven valuable for uncovering transactional data correlations. In recent times, advancements in distributed computing have given rise to federated learning or mining environments, where machine learning and data mining tasks are performed across a network of decentralized nodes, each holding local data samples. This paradigm shift brings

about novel opportunities and challenges, bearing implications for privacy, efficiency, and scalability.

Federated learning, by design, preserves the privacy of local datasets by training shared models without the need to transfer data between nodes. This feature is highly beneficial when data privacy is paramount, such as in healthcare or financial services. Despite the advantages, the decentralized nature of federated learning poses significant analytical challenges, particularly when applying traditional data mining algorithms like association rule mining. Conventional methods often require centralizing datasets or iterative communication that may compromise privacy or incur heavy communication costs.

Furthermore, in real-world scenarios, the crisp classification of data points often fails to capture the nuances and uncertainties inherent in the data. Fuzzy set theory has been employed to address this limitation by allowing objects to belong to multiple classes with varying degrees of membership, providing a more natural representation of data relationships and patterns.

This study introduces a novel algorithm that combines federated mining with fuzzy set theory to perform association rule mining. The proposed fuzzy association rule mining algorithm accommodates the uncertainties found in real-world data while adhering to the principles of federated environments. It aims to efficiently discover meaningful and interesting patterns across decentralized databases while mitigating the challenges associated with privacy and communication overheads.

Our contribution is twofold: firstly, we present a novel approach for integrating fuzzy logic within the federated mining framework to enhance the interpretability and relevance of association rules. Secondly, we provide a comprehensive algorithm that addresses the constraints of federated environments, such as limited communication bandwidth and data privacy concerns. The proposed algorithm lays the foundation for a new generation of data mining tools optimized for the distinctive requirements of federated systems.

The following sections discuss the relevant literature on association rule mining, federated learning, and fuzzy logic. We then outline the methodology underpinning our novel algorithm, followed by experimental results demonstrating its efficacy. Finally, we conclude with discussions on the implications and potential applications of our findings, as well as future research directions.

2 Preliminary concepts and related work

Unsupervised methods are used when no information is given about data, that is, data have no labels. These methods allow for discovering hidden patterns, data grouping, and other insights. We will now cover federated implementations of some of the most popular existing unsupervised learning/data mining algorithms.

2.1 Federated Mining

In recent years, advanced Artificial Intelligence (AI) methodologies, such as deep learning requiring vast data, have seen significant growth. They are employed extensively to grapple with complex challenges across diverse sectors like healthcare, finance, and transportation. The ability of these techniques to effectively generate accurate predictions or classifications heavily relies on the volume of available data.

The more data we collect, the more worried people are about privacy, who owns the data, and how it is shared. Storing all this personal information raises concerns about how it is being used. To address this, regulations such as the European Union's General Data Protection Regulation (GDPR) have been enforced, safeguarding user's data from unauthorized sharing.

Federated Learning (FL), a solution initially introduced by McMahan et al.[9] and expansively articulated by Yang et al.[11], provides a decentralized framework that facilitates multi-party collaboration on machine learning or data mining tasks without necessitating the sharing of private raw data. This mechanism empowers organizations and individuals to undertake collaborative projects while maintaining the privacy and security of their data. Federated systems hinge on two key ideas: how data are divided and how many devices are involved.

Data can be split horizontally or vertically. Horizontal means that each device has the same features (like age, income) but different data points (specific people). Vertical means that devices share some data points (same people) but have different features (one might have income, another location). Most current applications use horizontal partitioning.

The number of devices involved can also be categorized as cross-device or cross-silo. Cross-device involves many devices (think millions of phones) with limited data and processing power each. Cross-silo involves fewer, more powerful devices (think hospitals or banks) with larger datasets.

2.2 Association Rules

Association rules were formally defined for the first time by Agrawal et al. [1]. The problem consists in discovering implications of the form $X \rightarrow Y$ where X, Y are subsets of items from $I = \{i_1, i_2, ..., i_m\}$ fulfilling that $X \cap Y = \emptyset$ in a database formed by a set of n transactions $D = \{t_1, t_2, ..., t_n\}$ each of them containing subsets of items from I. X is usually referred as the antecedent and Y as the consequent of the rule.

The problem of discovering association rules is divided into two sub-tasks:

- Finding all the sets above the minimum support threshold, where support is defined as the percentage of transactions in the set. Itemsets exceeding the imposed threshold for the support, often called *minSupp* are known as frequent itemsets.
- Then, rules are discovered as those exceeding the minimum threshold for confidence or another assessment measurement generally given by the user.

However, real-world data can come in different formats: numbers, categories, or even imprecise descriptions. For numbers (like height), we can group them into ranges (e.g., 1.70-1.90 meters). But how we define these ranges can affect results. Fuzzy sets offer a better solution. We can use labels like "tall" instead of a strict range. This makes the data more user-friendly and captures the inherent fuzziness of some concepts (e.g., not everyone agrees exactly where "tall" begins). For truly imprecise data, even fuzzy sets might not work. That is where fuzzy transactions and fuzzy association rules come in [2, 5]. These tools are designed specifically to handle data that's too vague for traditional methods.

2.3 Association Rules in a Federated Environment

The works found for federated association rules are developed using privacy preserving implementations. As far as we know, the first work was published by Kantarcioglu et al., who proposed a secure algorithm based on commutative encryption and Secure Multiparty Computation (SMC) [8]. However, their algorithm had excessive information leaks and was computationally optimizable. Tassa et al. [10] corrected the privacy leaks of it and offered a more optimal algorithm. A different proposal is that of Chahar et al. which suffers from lengthy processing times because it relies on homomorphic encryption [3]. In the following, we briefly overview the Tassa's approach, the one that we are going to be based in order to mine fuzzy association rules.

2.4 Tassa's approach

Tassa's algorithm was proposed in [10]. Its main features are the use of a *secure multi-party protocol* for computing the union of private subsets held by different participants, and a protocol which tests the inclusion of an element held by one participant in a subset held by another. Their approach is based on the Fast Distributed Mining (FDM) algorithm [4], a method for finding frequent itemsets in a distributed dataset without revealing individual data. Its underlying principle is that any globally frequent itemset must also be frequent locally in at least one participating site. We summarize the main steps:

- **Initialization:** Participants have already found all globally frequent itemsets with size $k-1$ (denoted as F_s^{k-1}, i.e. the set of all $k-1$-itemsets that are s-frequent). Now, the goal is to find globally frequent itemsets with size k, i.e. (F_s^k).
- **Candidate Sets Generation:** Each participant finds frequent itemsets of size $k-1$ in their local data and are also known to be globally frequent (achieved through a separate step not mentioned here). Then, they use the Apriori algorithm to generate candidate itemsets of size k (denoted as $B_s^{k,m}$).
- **Local Pruning:** Each participant removes any candidate itemset from $B_s^{k,m}$ that are not frequent in their local data. This refined set is denoted as $C_s^{k,m}$.
- **Unifying Candidate Itemsets:** All participants share their local frequent itemsets ($C_s^{k,m}$) to create a combined set of candidate itemsets (C_s^k).

- **Computing Local Supports:** Each participant calculates the support of all candidate itemsets in C_s^k within their local data.
- **Broadcast Mining Results:** Participants share their local support calculations. This allows everyone to determine the global support for each candidate itemset.
- **Identifying Frequent Itemsets:** The final step involves identifying all itemsets in C_s^k that have a global support greater than or equal to *minsupp*. These itemsets form the set of globally frequent itemsets with size k, named F_s^k.

But the FDM algorithm raises some privacy concerns. Broadcasting locally frequent itemsets and their support sizes from individual databases reveals information about the underlying data, potentially compromising confidentiality. To address these shortcomings, Tassa et al. proposed secure multi-party protocols. These protocols enable participants to compute the union (or intersection) of their private datasets without revealing the individual data points. Additionally, they introduced a separate protocol that allows participants to securely check if a specific item exists within another participant's private subset.

2.5 Fuzzy Association Rules

For introducing fuzzy association rules, we first define what we consider a fuzzy transaction [2, 5].

Definition 1 *Let I be a set of items. A fuzzy transaction, τ, is a non-empty fuzzy subset of I in which the membership degree of an item $i \in I$ in τ is represented by a number in the range [0, 1] and denoted by $\tau(i)$.*

By this definition a crisp transaction is a special case of fuzzy transaction. We denote by $\tilde{D}$ a fuzzy transactional database. For an itemset, $A \subseteq I$, the degree of membership in a fuzzy transaction τ is calculated as the minimum of the membership degree of all its items $\tau(A) = \min_{i \in A} \tau(i)$.

Then, a fuzzy association rule $X \to Y$ is satisfied in $\tilde{D}$ if and only if $\tau(X) \leq \tau(Y)$ for all $\tau \in \tilde{D}$, i.e. the degree of satisfiability of Y in $\tilde{D}$ is greater than or equal to the degree of satisfiability of X for all fuzzy transactions t in $\tilde{D}$. Using this model the support and confidence measures are defined using a semantic approach based on the evaluation of quantified sentences as proposed in [2, 5]. Using the GD-method [5] and the quantifier $Q_M(x) = x$ the support of a fuzzy rule $X \to Y$ results:

$$FSupp(X \to Y) = \sum_{\alpha_i \in \Lambda(X \cap Y)} (\alpha_i - \alpha_{i-1}) \frac{|(X \cap Y)_{\alpha_i}|}{|\tilde{D}|} \tag{1}$$

where $\Lambda(X \cap Y) = \{\alpha_1, \alpha_2, \ldots, \alpha_p\}$ is an ordered set of α-cuts with $\alpha_i > \alpha_{i+1}$ and $\alpha_{p+1} = 0$.

Analogously the confidence is computed as follows:

$$FConf(X \to Y) = \sum_{\alpha_i \in \Lambda(X \cap Y)} (\alpha_i - \alpha_{i-1}) \frac{|(X \cap Y)_{\alpha_i}|}{|X_{\alpha_i}|} \tag{2}$$

3 Fuzzy association Rules in a federated environment

Our proposal aims to extend the functioning of Tassa's algorithm when each of the participants contain fuzzy transactional data.

The outline of the proposal is the following:

- **Definition of Λ:** a set of predefined α-cuts for the unit interval is defined.
- **Mining Frequent Itemsets per α-cut:**
  For each α-cut level:
  - We apply Tassa's algorithm to identify frequent itemsets within the data.
  - The output is a set of frequent itemsets with their associated support in that level.
- **Computing Final Support and Confidence:**
  - The final support of a fuzzy rule is determined using the individual supports from each α-cut and formula (1).
  - Following a similar level-wise approach, confidence for each frequent fuzzy rule is computed using formula (2).

4 Experimentation

In our experimental setup, we focused on evaluating the performance of the implemented algorithm for the case of fuzzy transactions in a federated environment. For that, we analyze its performance under various parameters. For this purpose, we use an adaptable configuration technique which provides the agility needed to assess the algorithm under diverse conditions and scenarios.

4.1 Dataset

For our experimentation, we opted for the comprehensive 'CDC Diabetes Health Indicators' dataset from the UCI Machine Learning repository. This dataset contains healthcare statistics about people and their associated diagnosis about diabetes. It comprises 253680 instances and an expanded 34 fully binary features, offering a rich source for extracting significant information about diabetes and associated health factors. Although our analysis here focuses on this dataset, it serves as a foundational evaluation of our proposed algorithm. We plan to investigate its performance across a broader range of datasets in future work.

A process of fuzzification of the dataset has been carried out using the library, and the process published in [7].

Next section explains the architecture employed in order to simulate a federated environment.

4.2 Simulated Environment

We have selected Python for implementing the simulate environment. This is primarily due to its fast development speed, familiarity with potent libraries such as Pandas and Numpy, and easy integration with a variety of tools including Docker.

For the system architecture we employed Docker and Docker Compose [6] to replicate the federated data mining environment and its collaborative dynamics accurately (see Figure 1). Docker containerization technology encapsulates each participant node as an independent, autonomous unit, thus mirroring real-world entities in a federated setup. Docker Compose coordinates these containers, recreating the interaction between distributed nodes in a controlled setting.

Indispensable to the architecture are our participating nodes that simulate independent collaborators in the federated data mining process. In addition, the system includes a 'Database controller' designed to manage data distribution through the participants in the experimental setting, which allow users to replicate conditions mimicking real-world scenarios.

In this way, the implementation and system architecture aim at creating a flexible and scalable environment for testing and refining our federated data mining process and facilitating the transition from an experimental to a real-world setting.

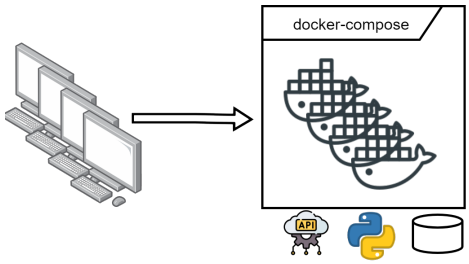

Fig. 1. Representation of a federated setting with the described technologies.

4.3 Results

In this section, we delve into the significance of the experimental parameters, exploring how variations in the number of participants, transactions, features, and data distribution can help us understand the algorithms' scalability, efficiency, and effectiveness. To perform the experiments, a subset of 5000 instances of the main dataset have been selected, with no data splitting and a default min_sup value of 0.5, with the number of participants being set to 10.

Number of participants. This parameter determines the size and complexity of our simulated federated environment. Varying the number of participants allows us to understand the scalability of our federated association rule mining algorithms. Due to hardware constraints, the number of participants we can simulate does not enter into the cross-device category (10^{10}), but enough insight can be obtained even from this small variance. For this experiment, the number of participants has been iteratively set to values in the range [3, 20], with all other parameters remaining equal.

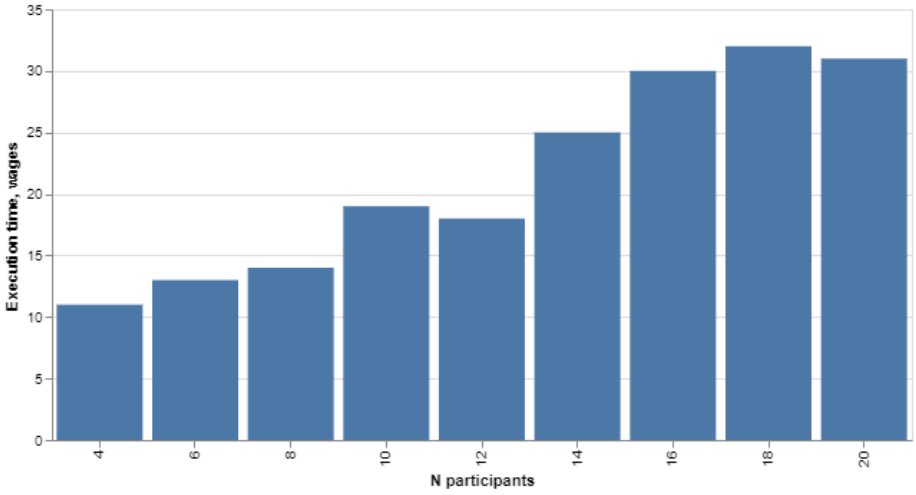

Fig. 2. Total execution time vs a number of participants for the algorithm.

However, it is interesting to observe that in Tassa's algorithm, the sent message size growth is exponential, as can be seen in Figure 3; this is due to the utilization of Shamir's secret sharing technique, which requires participants to send N shares of their information to other participants.

After this experiment, it is clear that Tassa's algorithm performs well. However, further testing would be needed to understand how a large number of participants would affect the algorithm's exponential communications.

Number of transactions. The number of transactions in our dataset is crucial as it impacts the data participants need to share and process. This parameter influences the communication overhead and computational requirements of the federated learning process, making it essential to explore how the algorithms perform under varying transaction loads. To perform this experiment, the number of transactions has been iteratively set to values in the range [1000, 50000] with all other parameters remaining equal (Participants = 10, minsupp = 0.5 and n_items = 36).

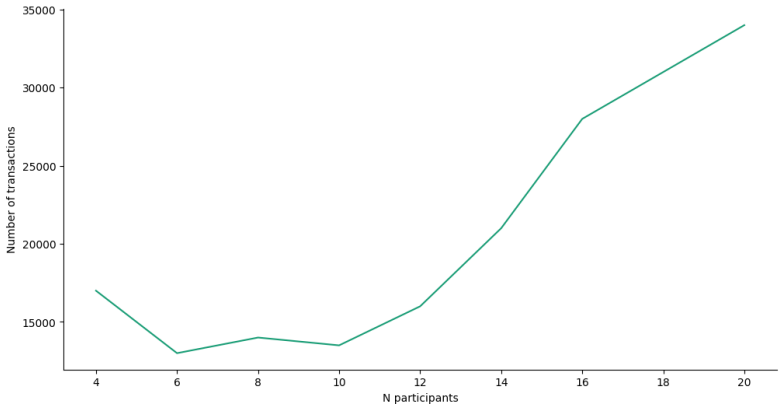

Fig. 3. Effect of the parameter N_participants on the size of the shared messages.

The shared data size can be examined to understand the effect of data dimensionality on the implemented algorithm. In Figure 4, we can deduce that the growth pattern is linear, meaning that the algorithm reacts stably to a growth in the number of data instances.

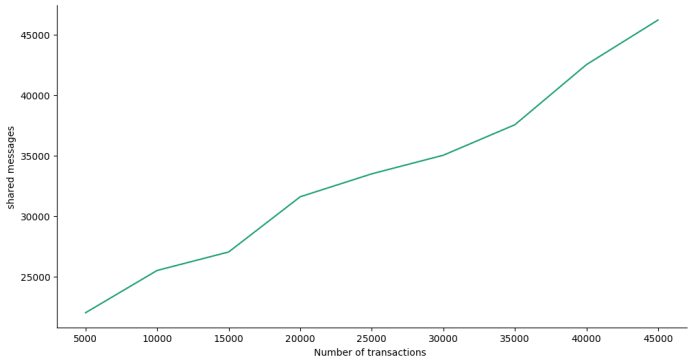

Fig. 4. Effect of the number of transactions on the size of the sent messages in Tassa's algorithm.

After an analysis of other metrics, no significant difference has been found in the performance of the algorithms when exposed to the same dimensionality changes.

Number of features. The number of features in our dataset is significant because it affects the complexity of the mining task. Increasing the number of features can challenge the algorithms in terms of computational efficiency and

the discovery of meaningful association rules. Exploring this parameter helps assess algorithm adaptability to high-dimensional data. For this experiment, the number of selected features has been iteratively set to values in the range [5, 36] with all other parameters remaining equal (Participants = 10, minsupp = 0.5 and n_transactions = 5000).

The first metric to consider is the total execution time. We can see a linear time in Figure 5.

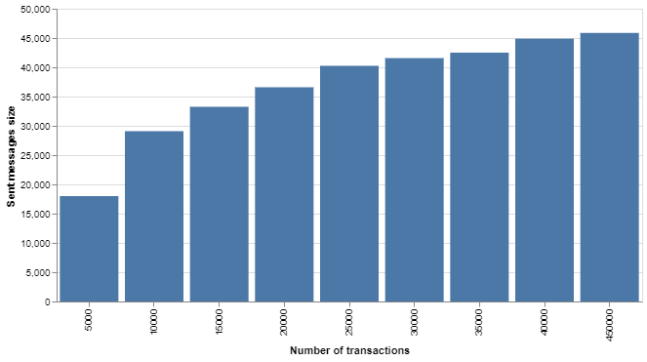

Fig. 5. Effect of the number of features on the execution times of Tassa's algorithm.

After this experimentation, we can conclude that a higher data dimensionality affects the algorithm performance.

This experiment has been very useful for analyzing the performance of an algorithm when varying one of the most important parameters in frequent itemset and association rule mining, the minimum support threshold. It has been observed that, while most metrics behave similarly, mining costs are drastically diverse and could pose a real challenge in a production setting.

4.4 Discussion

In examining the experimental results, we have closely investigated the performance and behavior of our approach in different scenarios. The experiments have offered insights into various aspects, including mining efficiency, and overall execution. We will now cover the main takeaways from these experiments, in hopes of highlighting the current limitations of the implemented algorithms.

Some of the key aspects inferred from the experimentation are:

– The communication cost of Tassa's algorithm has exponential growth, and while performing correctly in a non-massive setting, further testing needs to be implemented to study how a massive number of participants could negatively impact the algorithm's performance.

– The importance of the base frequent itemset mining algorithm cannot be understated. The experimentation has allowed for the understanding that Tassa's use of the Apriori algorithm presents a massive challenge in certain settings, and thus an alternative needs to be considered.

5 Conclusions

This work addresses the challenge of mining fuzzy association rules in a federated setting while preserving data privacy. Our proposed algorithm enables secure pattern discovery from participants' fuzzy transactional data.

The experimental results exhibited the versatility and scalability of our algorithm in variable settings. It was tested in various scenarios with differing numbers of transactions, features, and participants. Despite these variances, the algorithm maintained correct functionality and demonstrated operational adaptability and scalability. Whether dealing with higher volumes of transactions, a greater number of features, or a larger set of participants, the algorithm consistently delivered accurate and efficient performance. These results underscore the capability of the algorithm to scale seamlessly and cater to diverse operational requirements, demonstrating its strength when applied to high-level and complex data environments.

While this work demonstrates the effectiveness of fuzzy association rule mining in federated environments, there are exciting avenues for future exploration. One direction involves exploring methods for incorporating background knowledge or domain expertise into the fuzzy logic framework that could further enhance the interpretability and relevance of the mined rules. Furthermore, investigating the integration of advanced privacy-preserving techniques beyond those employed in this work could offer even stronger guarantees for data confidentiality in sensitive domains like healthcare. Finally, applying the proposed algorithm to other real-world datasets can prove the system's scalability.

Acknowledgements

We would like to acknowledge support for this work from FederaMed project: Grant PID2021-123960OB-I00 funded by MCIN/AEI/10.13039/501100011033 and by ERDF/EU. And from DesinfoScan project: Grant TED2021-129402B-C21 funded by MCIN/AEI/10.13039/501100011033 and by the European Union NextGenerationEU/PRTR.

In addition, the Ministry of Universities has partially supported this research through the EU-funded Margarita Salas programme NextGenerationEU and the pre-competitive project of the Plan Propio of the "University of Granada".

Finally, the research reported in this paper is also funded by the European Union (BAG-INTEL project, grant agreement no. 101121309).

References

1. Agrawal, R., Imielinski, T., Swami, A.: Mining associations between sets of items in large databases. In: ACM-SIGMOD International Conference on Data. pp. 207–216 (1993)
2. Berzal, F., Delgado, M., Sánchez, D., Vila, M.: Measuring accuracy and interest of association rules: A new framework. Intelligent Data Analysis **6**(3), 221–235 (2002)
3. Chahar, H., Keshavamurthy, B., Modi, C.: Privacy-preserving distributed mining of association rules using elliptic-curve cryptosystem and shamir's secret sharing scheme. Sādhanā **42**(12), 1997–2007 (2017)
4. Cheung, D., Han, J., Ng, V., Fu, A., Fu, Y.: A fast distributed algorithm for mining association rules. In: Fourth International Conference on Parallel and Distributed Information Systems. pp. 31–42 (1996). https://doi.org/10.1109/PDIS.1996.568665
5. Delgado, M., Marín, N., Sánchez, D., Vila, M.: Fuzzy association rules: General model and applications. IEEE Transactions on Fuzzy Systems **11**(2), 214–225 (2003)
6. Docker: Docker compose documentation, https://docs.docker.com/compose/compose-file/compose-file-v3/
7. Fernandez-Basso, C., Gutiérrez-Batista, K., Morcillo-Jiménez, R., Vila, M.A., Martin-Bautista, M.J.: A fuzzy-based medical system for pattern mining in a distributed environment: Application to diagnostic and co-morbidity. Applied Soft Computing **122**, 108870 (2022)
8. Kantarcioglu, M., Clifton, C.: Privacy-preserving distributed mining of association rules on horizontally partitioned data. IEEE transactions on knowledge and data engineering **16**(9), 1026–1037 (2004)
9. McMahan, B., Moore, E., Ramage, D., Hampson, S., y Arcas, B.A.: Communication-efficient learning of deep networks from decentralized data. In: Artificial intelligence and statistics. pp. 1273–1282. PMLR (2017)
10. Tassa, T.: Secure mining of association rules in horizontally distributed databases. IEEE Transactions on Knowledge and Data Engineering **26**(4), 970–983 (2013)
11. Yang, Q., Liu, Y., Chen, T., Tong, Y.: Federated machine learning: Concept and applications. ACM Transactions on Intelligent Systems and Technology (TIST) **10**(2), 1–19 (2019)