

Association Rules Machine Learning complete intersection Calabi-Yau 5-Folds and 6-Folds

Kaniba Mady Keita^{a,b*}

a Centre de Calcul de Modélisation et de Simulation: CCMS

*Department of Physics, Faculty of Sciences and Techniques, University of Sciences, Techniques
and Technologies of Bamako, FST-USTTB, BP:E3206, Mali.*

b Centre de Recherche en Physique Quantique et de ses Applications: CRPQA, Bamako, Mali.

October 29, 2025

Abstract

Association rule machine learning is applied to the dataset of complete intersection Calabi–Yau 5-folds and 6-folds in order to uncover hidden patterns among their Hodge numbers. These Hodge numbers—six for the 5-folds and nine for the 6-folds—serve as the items in our analysis. For the 5-folds, we discover 60 significant association rules. For example, within the dataset, if $h^{1,3} = 0$ and $h^{2,2} = 5$, then $h^{1,1} = 3$ with 99.43% confidence. Similarly, if $h^{2,1} = 0$, $h^{1,3} = 0$, and $h^{2,2} = 5$, then $h^{1,1} = 3$ with 99.42% confidence. For the 6-folds, we identify 160 association rules across a dataset of 1,482,022 examples. A particularly striking observation is that $h^{1,2} = h^{1,3} = h^{1,4} = h^{2,3} = 0$ for all entries in this dataset. These types of association rules are especially valuable because the Hodge numbers of complete intersection Calabi–Yau 5-folds have only been computed for approximately 53% of the dataset, while those of 6-folds remain largely undetermined. The discovered patterns provide predictive insights that can guide future computations and theoretical developments.

*E-mail: madyfalaye@gmail.com, kanibamady.keita@usttb.edu.ml

1 Introduction

Topological invariants of complete intersection Calabi–Yau three-folds (CICY3s) have been computed, and the corresponding dataset consists of 7890 CICY3s [1]. Similarly, there are 921,497 known complete intersection Calabi–Yau four-folds [2]. In the case of complete intersection Calabi–Yau five-folds (CICY5s), 27,068 spaces were obtained by the authors of [3], but the cohomological data has been computed for only 12,433 of them (approximately 53.7%).

The space of Calabi–Yau six-folds is even more intriguing. Edward Hirst and Tancredi Schettini Gherardini identified approximate Hodge numbers for a dataset containing 1,482,022 Calabi–Yau six-fold candidates with a total weight sum less than 200 [4]. These Hodge numbers were shown to match exactly with the values computed by V. N. Dumashev [5]. Additionally, these six-dimensional weighted projective spaces with transverse polynomials were confirmed to be consistent by Maximilian Kreuzer and Harald Skarke [6].

In recent years, machine learning has been successfully applied to the study of complete intersection Calabi–Yau manifolds [7–23]. These approaches are believed to simplify the computation of topological invariants of such manifolds. Most of the machine learning applications in this context have focused on classification, regression, and clustering algorithms. While these techniques are quite effective for fully determined datasets, this is unfortunately not the case for Calabi–Yau 5-folds and 6-folds.

For incomplete datasets, one particularly useful technique is association rule machine learning (also known as association rule data mining) [24], which can be employed to propose logical rules for estimating the missing data entries.

In this work, we apply association rule machine learning using the Apriori algorithm [25] and some of its improved versions (see, for instance, [26–28]). Our implementation uses the `arulesViz` package [29] in RStudio. The main evaluation metrics for an association rule are the *support*, *confidence*, and *lift* of the rule. These metrics will be defined in Section 2.

The remainder of this paper is organized as follows. In Section 2, we introduce the basic theoretical concepts of association rules and describe how they are implemented using the `arulesViz` package. Section 3 discusses the challenges involved in computing the cohomological datasets for Calabi–Yau 5-folds and 6-folds and motivates the application of association rule techniques. Section 4 presents the results of applying the Apriori algorithm to these datasets. Finally, Section 5 offers concluding remarks and outlines directions for future work.

2 Basic Theoretical Concepts of Association Rules

The basic concepts we present here can be found in [30, 31].

Let us denote our dataset by $\mathcal{D}$, where each row corresponds to a transaction. The first column of $\mathcal{D}$ represents the identity of each transaction, while the remaining columns correspond to items.

In our context, each transaction represents a specific Calabi–Yau manifold, and the identity refers to its sequential label within the dataset. For instance, the CICY3 dataset

contains 7,890 such identities, whereas the CICY6 dataset contains 1,482,022.

The items in the dataset are the individual Hodge numbers $h^{i,j}$ associated with each manifold. By definition, an *itemset* X is a collection of one or more such items. If X contains k items, it is referred to as a k -*itemset*.

A **association rule** is a logical implication between two itemsets, typically expressed as:

$$X \Rightarrow Y,$$

which means that if the itemset X appears in a transaction, then the itemset Y is likely to appear as well. Here, X is called the **left-hand side (LHS)** and Y is called the **right-hand side (RHS)** of the rule.

When $X = \emptyset$, the association rule

$$\emptyset \Rightarrow Y$$

means that the itemset Y is likely to appear in all transactions, regardless of the presence of any other items. In other words, the rule holds unconditionally— Y is always true.

The **support** of the association rule $X \Rightarrow Y$, denoted $\text{support}(X \Rightarrow Y)$, is the fraction of transactions in the dataset $\mathcal{D}$ that contain both itemsets X and Y ; that is, transactions containing $X \cup Y$.

The *confidence* of the rule, denoted by $\text{conf}(X \Rightarrow Y)$, is, by definition, the conditional probability $p(Y \mid X)$. In other words, confidence is the proportion of transactions in $\mathcal{D}$ that contain X and also contain Y . A few statistical parameters for association rules are given in Table 1 below:

Measure	Definition	Formula
Support(X)	The empirical probability, $P(X)$, of the itemset X	$\frac{N(X)}{\text{Total number of transactions}}$
Support($X \Rightarrow Y$)	Proportion of transactions that contain both X and Y	$\text{supp}(X \cup Y)$
Confidence($X \Rightarrow Y$)	The conditional probability $P(Y \mid X)$	$\frac{\text{supp}(X \cup Y)}{\text{supp}(X)}$
Lift($X \Rightarrow Y$)	This measure evaluates the strength of the rule $X \Rightarrow Y$	$\frac{\text{conf}(X \Rightarrow Y)}{\text{supp}(Y)}$
Conviction($X \Rightarrow Y$)	This measure evaluates the violation of the rule $X \Rightarrow Y$	$\frac{1 - \text{supp}(Y)}{1 - \text{conf}(X \Rightarrow Y)}$
Leverage($X \Rightarrow Y$)	It assesses the added value of the rule $X \Rightarrow Y$	$\text{supp}(X \cup Y) - \text{supp}(X) \cdot \text{supp}(Y)$

Table 1: Definitions and formulas of common statistical measures for association rules.

where $N(X)$ = Number of transactions containing X . Hence,

$$P(X) = \frac{\text{Number of transactions containing } X}{\text{Total number of transactions}}.$$

To apply association rules, we use the **Apriori algorithm**, a widely used technique in association rule mining. The Apriori method follows an iterative approach, starting with the identification of *frequent itemsets*.

1. First, a **minimum support threshold** is defined.
2. All itemsets with **support** greater than or equal to this threshold are identified. These are known as *frequent itemsets* or *large itemsets*.
3. The algorithm then iteratively generates larger itemsets from smaller frequent itemsets, using the Apriori property which states that *all subsets of a frequent itemset must also be frequent*.
4. Once all frequent itemsets are identified, the next step is to generate **association rules** from them.
5. For each rule of the form $X \Rightarrow Y$, where X and Y are itemsets and $X \cap Y = \emptyset$, the **confidence** is calculated as:

$$\text{Confidence}(X \Rightarrow Y) = \frac{\text{Support}(X \cup Y)}{\text{Support}(X)}$$

6. Only rules with confidence greater than or equal to the **minimum confidence threshold** are considered as **strong association rules**.

These **strong association rules** are then considered as **the association rules of the data** under consideration.

3 Motivation

The computation of topological invariants of Complete Intersection Calabi-Yau (CICY) manifolds is an active and important area of research. While the invariants of CICY threefolds were computed in [1], yielding 7890 distinct configuration matrices and a list of 921,497 configuration matrices for CICY four-folds were worked out in ref. [2], much less is known about their higher-dimensional analogues, such as Calabi-Yau 5-folds and 6-folds. These higher-dimensional CICYs are particularly interesting in M-theory and F-theory models, where compactification spaces of dimension five or more naturally arise. The size and complexity of configuration matrices grow rapidly, making computations expensive using traditional algebraic geometry algorithms.

The topological invariants of CICYs are not merely mathematical curiosities—they have direct physical implications. For instance, the Hodge numbers determine the number of moduli fields in the effective theory, while intersection numbers influence Yukawa couplings, anomaly cancellation conditions, and gauge symmetry breaking. Moreover, the so-called mirror symmetry provides dual pairs of Calabi-Yau manifolds whose Hodge diamonds exhibit symmetric structures. Understanding these symmetries in higher-dimensional settings remains an important open problem, both mathematically and physically.

Given the computational challenges of traditional methods, alternative such as data-driven approaches offer a powerful and scalable direction. In particular, *association rule mining*—a machine learning technique can be adapted to uncover frequently co-occurring patterns among topological invariants. Instead of computing each invariant individually,

this method searches for statistically significant rules of the form “if X , then Y ,” where X and Y represent combinations of properties (e.g., Hodge number values or configuration features). Applying these techniques to higher-dimensional CICYs could provide new insights into their topological invariants, guiding further theoretical developments where conventional tools are no longer feasible.

4 Results of Association Rules for CICY5 and CICY6 Datasets

To apply association rule mining to the datasets of complete intersection Calabi–Yau five-folds (**CICY5**) and six-folds (**CICY6**), we first transform the given Hodge numbers into a set of discrete items. The correspondence between the Hodge numbers and item labels is given in Table 2.

CICY5		CICY6	
Item	Value	Item	Value
item ₁	$h^{1,1}$	item ₁	$h^{1,1}$
item ₂	$h^{2,1}$	item ₂	$h^{1,2}$
item ₃	$h^{1,3}$	item ₃	$h^{1,3}$
item ₄	$h^{1,4}$	item ₄	$h^{1,4}$
item ₅	$h^{2,2}$	item ₅	$h^{1,5}$
item ₆	$h^{2,3}$	item ₆	$h^{2,2}$
—	—	item ₇	$h^{2,3}$
—	—	item ₈	$h^{2,4}$
—	—	item ₉	$h^{3,3}$

Table 2: Item encoding of Hodge numbers for CICY5 and CICY6 datasets.

The Apriori algorithm is applied to the dataset of complete intersection Calabi–Yau 6-folds (CICY6). The minimum support is set to 10%, and the minimum confidence is set to 80%. These threshold values are chosen based on the relative size of the dataset under consideration, ensuring a balance between rule significance and result comprehensiveness. The algorithm is implemented in **RStudio** using the **arulesViz** package and executed on the CCMS computing cluster.

The first ten results of the association rule mining applied to the CICY5 dataset are presented in Table 3, and there are 60 such rules in total. The remaining rules are provided in reference [32].

Table 3: Top 10 Association Rules for CICY5

rule	LHS	RHS	Support	Confidence	Coverage	Lift	Count
1	{}	{item3=0}	0.9570	0.9570	1.0000	1.0000	11899
2	{}	{item2=0}	0.9770	0.9770	1.0000	1.0000	12147
3	{item5=4}	{item1=3}	0.1055	0.9661	0.1092	2.4962	1312
4	{item5=4}	{item3=0}	0.1075	0.9838	0.1092	1.0280	1336
5	{item5=4}	{item2=0}	0.1091	0.9993	0.1092	1.0228	1357
6	{item5=7}	{item1=4}	0.1130	0.9256	0.1221	2.1287	1405
7	{item5=7}	{item3=0}	0.1194	0.9776	0.1221	1.0215	1484
8	{item5=7}	{item2=0}	0.1219	0.9987	0.1221	1.0222	1516
9	{item5=5}	{item1=3}	0.1953	0.9870	0.1979	2.5501	2428
10	{item5=5}	{item3=0}	0.1931	0.9760	0.1979	1.0198	2401

In the same manner, the first ten results of the association rule mining applied to the CICY6 dataset are presented in Table 4. Our investigation yielded 160 association rules, the complete details of which are also provided in reference [32].

Table 4: Top 10 Association Rules for CICY6

rule	LHS	RHS	Support	Confidence	Coverage	Lift	Count
1	{}	{item2=0}	1.0000000	1.0000	1.0000000	1.0000	1482022
2	{}	{item3=0}	1.0000000	1.0000	1.0000000	1.0000	1482022
3	{}	{item4=0}	1.0000000	1.0000	1.0000000	1.0000	1482022
4	{}	{item7=0}	1.0000000	1.0000	1.0000000	1.0000	1482022
5	{item5=5}	{item2=0}	0.1319083	1.0000	0.1319083	1.0000	195491
6	{item5=5}	{item3=0}	0.1319083	1.0000	0.1319083	1.0000	195491
7	{item5=5}	{item4=0}	0.1319083	1.0000	0.1319083	1.0000	195491
8	{item5=5}	{item7=0}	0.1319083	1.0000	0.1319083	1.0000	195491
9	{item5=4}	{item2=0}	0.1691729	1.0000	0.1691729	1.0000	250718
10	{item5=4}	{item3=0}	0.1691729	1.0000	0.1691729	1.0000	250718

These results provide strong evidence of the usefulness of applying association rule mining to Calabi–Yau datasets, as they uncover statistically significant and interpretable relationships between different Hodge numbers.

5 Conclusions and Future Outlooks

In this paper, association rule machine learning is applied to datasets of complete intersection Calabi–Yau (CICY) 5-folds and 6-folds. For the 5-folds, we discover 60 significant association rules, while for the 6-folds, we identify 160 association rules across a dataset containing 1,482,022 examples. The detailed rules are provided in Ref. [32]. Our investigation reveals rich and non-trivial correlations among the Hodge numbers of these higher-dimensional Calabi–Yau manifolds. In particular, the structural patterns (rules) may guide future computations. Furthermore, this work demonstrates the potential of data-driven and interpretable machine learning methods in uncovering hidden relationships within topological invariants of CICYs datasets.

6 Acknowledgment

The author is indebted to Bobby Samir Acharya for pointing out the importance of CICY 6-folds as the next non-trivial even-dimensional class of complete intersection Calabi–Yau manifolds.

References

- [1] P.S. Green, T.H. Hübsch, and C.A. Lütken, **All the Hodge numbers for all Calabi-Yau complete intersections**, Class. Quantum Grav. **6** (1989) 105–124.
<https://www-thphys.physics.ox.ac.uk/projects/CalabiYau/cicylist/>
- [2] J. Gray, A.S. Haupt, and A. Lukas, **All Complete Intersection Calabi-Yau Four-Folds**, JHEP **07** (2013) 070.
<https://www-thphys.physics.ox.ac.uk/projects/CalabiYau/Cicy4folds/index.html>
- [3] R. Alawadhi, D. Angella, A. Leonardo, and T. Schettini Gherardini, **Constructing and Machine Learning Calabi-Yau Five-folds**, Fortschr. Phys. **2023**, 2300262, arXiv:2310.15966 [hep-th].
- [4] E. Hirst and T. Schettini Gherardini, **Calabi-Yau Four/Five/Six-folds as $\mathbb{P}_w^n$ Hypersurfaces: Machine Learning, Approximation, and Generation**, Phys. Rev. D **109**, 106006 (2024).
doi:10.1103/PhysRevD.109.106006
- [5] V. Dumachev, **Complete intersection Calabi-Yau six-folds**, Appl. Math. Sci. **9** (2015) 7121.
<http://dx.doi.org/10.12988/ams.2015.510620>
- [6] M. Kreuzer and H. Skarke, **Numerical data of Calabi-Yau manifolds of various dimensions and the program PALP**,
<http://hep.itp.tuwien.ac.at/~kreuzer/CY/CYff.html>
- [7] K.M. Keita, **Machine learning complete intersection Calabi-Yau 3-folds**, Phys. Rev. D **110** (2024) 12, 126002, arXiv:2404.11710 [hep-th].
- [8] K.M. Keita and Y.H. Dicko, **Machine Learning Calabi-Yau Three-Folds, Four-Folds, and Five-Folds**, arXiv:2503.00139 [hep-th].
- [9] Y.-H. He, **The Calabi-Yau Landscape: from Geometry, to Physics, to Machine-Learning**, arXiv:1812.02893 [hep-th].
- [10] Y.-H. He, **Deep-Learning the Landscape**, arXiv:1706.02714 [hep-th].
- [11] Y.-H. He, Z.-G. Yao, and S.-T. Yau, **Distinguishing Calabi-Yau Topology using Machine Learning**, arXiv:2408.05076 [math.AG].
- [12] W. Cui, X. Gao, and J. Wang, **Machine Learning on Generalized Complete Intersection Calabi-Yau Manifolds**, arXiv:2209.10157 [hep-th].
- [13] H. Erbin and R. Finotello, **Deep learning complete intersection Calabi-Yau manifolds**, arXiv:2311.11847 [hep-th].
- [14] A. Ashmore, Y.-H. He, and B. Ovrut, **Machine learning Calabi-Yau metrics**, Fortschr. Phys. **68** (2020) 9, 2000068.
- [15] D. Aggarwal et al., **Machine-learning Sasakian and G2 topology on contact Calabi-Yau 7-manifolds**, arXiv:2310.03064 [math.DG].
- [16] D.S. Berman, Y.-H. He, and E. Hirst, **Machine learning Calabi-Yau hypersurfaces**, Phys. Rev. D **105** (2022) 066002, arXiv:2112.06350 [hep-th].

- [17] R. Deen, Y.-H. He, S.-J. Lee, and A. Lukas, **Machine learning string standard models**, Phys. Rev. D **105**, 046001 (2022).
- [18] Y.-H. He and A. Lukas, **Machine learning Calabi-Yau four-folds**, Phys. Lett. B **815** (2021) 136139.
- [19] F. Ruehle, **Data science applications to string theory**, Phys. Rep. **839** (2020) 1.
- [20] K. Bull, Y.-H. He, V. Jejjala, and C. Mishra, **Machine learning CICY Three-folds**, Phys. Lett. B **785** (2018) 65–72, arXiv:1806.03121 [hep-th].
- [21] K. Bull, Y.-H. He, V. Jejjala, and C. Mishra, **Getting CICY High**, Phys. Lett. B **795** (2019) 700–706, arXiv:1903.03113 [hep-th].
- [22] H. Erbin and R. Finotello, **Inception neural network for complete intersection Calabi-Yau 3-folds**, Mach. Learn.: Sci. Technol. **2** (2021), 02LT03, arXiv:2007.13379 [hep-th].
- [23] H. Erbin, R. Finotello, R. Schneider, and M. Tamaazousti, **Deep multi-task mining Calabi-Yau four-folds**, Mach. Learn.: Sci. Technol. **3** (2022) 015006.
- [24] C. Zhang and S. Zhang (Eds.), **Association Rule Mining: Models and Algorithms**, Springer, 2002.
<https://link.springer.com/book/10.1007/3-540-46027-6>
- [25] R. Agrawal and R. Srikant, **Fast Algorithms for Mining Association Rules**, Proc. 20th Int. Conf. Very Large Data Bases (VLDB), Santiago de Chile, 1994, pp. 487–499.
<https://www.vldb.org/conf/1994/P487.PDF>
- [26] J. He, L. Lang, B. Li, and W. Xiong, **Unveiling Valuable Insights: Leveraging Apriori Algorithm for Association Rule Mining and Analysis**, 2023 4th Int. Conf. on Computer, Big Data and Artificial Intelligence (ICCBD+AI), Guiyang, China, pp. 612–616.
doi:10.1109/ICCBD-AI62252.2023.00112
- [27] M. Al-Maolegi and B. Arkok, **An Improved Apriori Algorithm for Association Rules**, arXiv:1403.3948.
<https://ui.adsabs.harvard.edu/abs/2014arXiv1403.3948A/abstract>
- [28] F.H. Al-Zawaidah, Y.H. Jbara, and A.L. Marwan, **An Improved Algorithm for Mining Association Rules in Large Databases**, World of Computer Science and Information Technology Journal (WCSIT), **1**(7), 311–316 (2011).
- [29] M. Hahsler, **arulesViz: Interactive Visualization of Association Rules with R**, R Journal **9**(2) (2017), 163–175.
doi:10.32614/RJ-2017-047, <https://journal.r-project.org/archive/2017/RJ-2017-047/RJ-2017-047.pdf>
- [30] J. Han, M. Kamber, and J. Pei, **Data Mining: Concepts and Techniques**, 4th ed., Morgan Kaufmann, 2022.
- [31] P.J. Azevedo and A.M. Jorge, **Comparing Rule Measures for Predictive Association Rules**, in: J.N. Kok, J. Koronacki, R.L. de Mántaras, S. Matwin, D.

Mladeníć, and A. Skowron (eds), *Machine Learning: ECML 2007*, Lecture Notes in Computer Science, vol. 4701, Springer, Berlin, Heidelberg, 2007.
https://doi.org/10.1007/978-3-540-74958-5_47

- [32] Kaniba Mady Keita , KEITA, Kaniba Mady (2025), **Association Rules Machine Learning complete intersection Calabi-Yau 5-Folds and 6-Folds**, Mendeley Data, V1, doi:10.17632/jjrrhr9vw.1