Project Final Report

Analysis of association between student participation and academic performance in online education

Student:

Maitriben SHAH: 200510593

Behnush Kiani: 200491333

University of Regina

2024



# **Abstract**

Online education or e-learning has become a prominent part of students’ lives. The viewpoint that students can only focus on studies in traditional classroom has changed, especially during the period like Covid-19 pandemic. Unlike the traditional classroom, online education has few to no interaction among students, also it is hard for the teacher to judge the performance of a particular student. To address this issue, we have focused on finding relationship between student participation and their academic performance. This project is based on the research paper “Relationship between Student Engagement and Performance in e-Learning Environment Using Association Rules”. They have used Apriori which is a Frequent Pattern Mining algorithm typically used for association in market basket analysis which uses boolean values for features to find relationship between student engagement and their academic performance. The aim of our project is to find association with quantitative features, where there are weights in the features instead of boolean values for boolean association rule. After a preliminary study of the available algorithms and data review, it became apparent that for this problem we have to use High Utility Itemset Mining algorithms. We have applied association between itemsets using Two-Phase High Utility Itemset Mining algorithm. Experimental analysis shows that, students with more participation during the course, often have better academic performance.

Table of Contents

[**Abstract** 1](#_Toc162821925)

[List of Figures 3](#_Toc162821926)

[List of Tables 3](#_Toc162821927)

[**Introduction** 4](#_Toc162821928)

[**Background** 4](#_Toc162821929)

[A. Association Rules 4](#_Toc162821930)

[B. Two-phase HUIM Algorithm: 6](#_Toc162821931)

[**Example of Two-phase application in market basket analysis:** 8](#_Toc162821932)

[**Related Works** 9](#_Toc162821933)

[**Aim** 10](#_Toc162821934)

[**Scope** 10](#_Toc162821935)

[**Data Science lifecycle process** 10](#_Toc162821936)

[**Modules** 10](#_Toc162821937)

[Data Discovery 11](#_Toc162821938)

[Data Preparation 13](#_Toc162821939)

[Model Planning 16](#_Toc162821940)

[Model Building 17](#_Toc162821941)

[Implementation 17](#_Toc162821942)

[Analysis 20](#_Toc162821943)

[Problems faced and solved 20](#_Toc162821944)

[Problems we were not able to solve 20](#_Toc162821945)

[Lessons Learned from the Project 20](#_Toc162821946)

[Output of the project 21](#_Toc162821947)

[**Conclusion** 21](#_Toc162821948)

[A summary of project and findings 21](#_Toc162821949)

[Recommendations 22](#_Toc162821950)

[References 23](#_Toc162821951)

# List of Figures

Fig. 1 Itemsets lattice related to the example in Table 1. ɛ’ = 120. Itemsets in circles (solid and dashed) are the high transaction-weighted utilization itemsets in transaction-weighted utilization mining model. Gray-shaded boxes denote the search space. Itemsets in solid circles are high utility itemsets. Numbers in each box are transaction-weighted utilization / number of occurrences. 8

Fig. 2 Highest Utility Itemsets with MinUtil - 20000 9

Fig. 3 Relationship Between visited resources and Class 13

Fig. 4 Relationship between announcement view and class 14

Fig. 5 Relationship between Raised hands and class 14

Fig. 6 Relationship between Discussion and Class 15

Fig. 7 Correlation Matrix: All features included 15

Fig. 8 Correlation matrix: Quantitative features 16

Fig. 9 High Itemset Utility with Minutil-0 18

Fig. 10 HUIM with MinUtil-8k 18

Fig. 11 HUIM with MinUtil-20K 19

Fig. 13Testing for best possible Minimum threshold 19

Fig. 14 High Utility itemset in student database 21

# List of Tables

Table 1 Transaction Database 7

Table 2 Transaction Database 8

Table 3 Utility table 8

Table 4 Transaction Utility Table 9

Table 6 Dataset Description 12

Table 7 Utility Table 17

# **Introduction**

With the increase in the use of smart devices, e-learning has become a new medium of education for many people. E-learning observed a sudden shift during Covid-19 pandemic [1]. Before pandemic in 2019, only 8% of first-years (21 and older) and 20% of seniors (25 and older) took all spring courses online. In 2021, however, fully 65% of first-year students and 66% of seniors took courses that were mostly remote [1]. With sudden shift toward online education, many MOOCs platform observed vast number of enrolments in online courses, for example, enrolment in courses offered by Coursera increased by more 1.5 times from what was in 2019(45 million) [2]. However, the increase in enrolment does not show much increase in the certificates awarded or completion of courses [3]. According to survey of 2015, the completion rates can increase up to 40% but they remain largely at 15% [4]. A study conducted in 2018, projected that less than 15% complete the online course they are enrolled in [5]. This is one of the problems faced in e-learning, that how to keep student engaged and interested. As in online education the environment for each student is different, for example, some may have a quiet learning environment and some may have to study with outdoor noises. Therefore, it is important to keep them engaged and motivated to make them feel included though their pace might be slower or faster than others [6].

In this project we study the comprehensive relationship between students’ engagement and their academic performance. There are many literatures that focus on this relationship but only few of them used comprehensive set of engagement metrics. This project studies frequency related engagement metrics as well as overall engagement level of student using Two-Phase High Utility Itemset Mining Algorithm.

## **Background**

1. Association Rules:

Association rule learning is a type of rule-based machine learning algorithms that aims to discover interesting relations between items in large databases [7]. The idea is to produce rules that can predict the occurrence of an item based on

occurrences of other items [8]. Agrawal et al. provide a more formal definition as follows [9]:

Let I = {i1, i2, ..., in} be a set of n attributes/items and

T = {t1, t2, ..., tm} be a set of m transactions forming a database. Each transaction it includes a subset of the items available in I. A rule can be defined as X ⇒ Y where

X, Y ⊆ I, i.e. X and Y (known as itemsets) are a subset of the items available. In other words, a rule can be thought of as a predictable transaction within the database. X and Y are commonly referred to as the antecedent and the consequent of the rule respectively.

Association rules can be beneficial in an e-learning environment as they can detect correlations between different features within the dataset. In particular, they can be used to correlate student behaviours with their performance to determine what is positively or negatively impacting their learning experience.

In order to evaluate the importance and interestingness of an association rule, several measures have been proposed. In what follows, three measures are presented.

* **Support:** Support of an itemset is an indication of how frequent an itemset appears in the transactions’ database. It can be thought of as the probability of occurrence of the considered itemset by counting the number of transactions

in which the itemset appears relative to the total number of transactions. More formally, the support of an itemset X with respect to a set of transactions T is calculated as [8]:

supp(X) = |t ∈ T; X ⊆ t|

|T|

* **Confidence:** Confidence is a measure of how frequent the rule is within the transactions’ database. In layman terms, it is the portion of transactions that contain both itemsets X and Y forming the rule relative to the transactions that contain X in general. Hence, the confidence of rule X ⇒ Y can be defined as [8]:

Confidence (X ⇒ Y) = supp (X ∪ Y)

supp(X)

* **Lift:** The lift of a rule is a measure of how interesting the rule is. The lift determines the probability of the rule occurring relative to the probability of the antecedent and consequent being independent. Therefore, the lift of an association rule is defined as [21]:

Lift (X ⇒ Y) = supp (X ∪ Y)

supp(X) × supp (Y)

A lift of 1 implies that the two itemsets comprising the rule are independent and hence the rule associating them together is not truly a rule. However, if the lift is > 1, it can be concluded that the two itemsets are dependent on each other. This makes the corresponding rule possibly useful in predicting future occurrences of the consequent if the antecedent occurs.

Association rules are often used for categorical (non-numeric) datasets and have been considered for a variety of applications. Several association rules algorithms have been developed. In what follows, a brief description of well-known algorithm that is the basis of the research paper is given.

#### Apriori Algorithm:

Apriori algorithm is a breadth-first search-based algorithm that depends on the frequency of the itemsets to identify a set of association rules [**10**]. The algorithm identifies the itemsets that appear in at least C transactions within the database, where C is the minimum threshold chosen by the user. It adopts a” bottom up” approach where a frequent subset is extended by one item in each iteration. This means that the algorithm starts with itemsets of length 1 (i.e. only one item within the itemset) and determines the itemset that have a frequency higher than the considered threshold C. This is repeated until no newer frequent itemsets are found. The length is then incremented by 1 and the same process is adopted again. This continues until there is no more possible extensions of the itemsets. The popularity of the Apriori algorithm stems from the fact that it can be easily implemented and parallelized as well as makes use of the large itemset property [**10**]. However, the algorithm does suffer from some drawbacks. One of the main drawbacks is that it requires several database scans in order to produce the rules [**10**]. This is mainly due to the itemset extension property of the algorithm. Another drawback is the fact that it can be slow due to its dependence on the size of the database, number of items within it, and the choice of minimum

support [8].

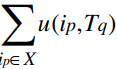
Traditional Association rules mining (ARM) model treat all the items in the data base equally by only considering if an item is present in a transaction or not [**11]**. Frequent itemsets identified by ARM may only contribute a small portion of the overall profit, whereas non-frequent itemsets may contribute a large portion of the profit. In reality, a retail business may be interested in identifying its most valuable customers (customers who contribute a major fraction of the profits to the company). These are the customers, who may buy full priced items, high margin items, or gourmet items, which may be absent from a large number of transactions because most customers do not buy these items. In a traditional frequency-oriented ARM, these transactions representing highly profitable customers may be left out [**11]**.

The goal of utility mining is to identify high utility itemsets that drive a large portion of the total utility. Traditional ARM problem is a special case of utility mining, where the utility of each item is always 1 and the sales quantity is either 0 or 1 [**11]**.

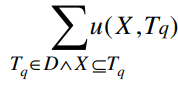
1. Two-phase HUIM Algorithm:

Two-phase HUIM algorithm is one of many High Utility Itemset Mining algorithms and in quantitative customer transactions, HUIM reveal the high profit yielding itemsets when bought together by customers. The more formal definition is as follows:

Let I = {i1, i2, …, im} is a set of items.

* D = {T1, T2, …, Tn} be a transaction database where each transaction Ti ∈ D is a
  + subset of I.
* o(ip, Tq), local transaction utility value, represents the quantity of item ip in transaction Tq. For example, o(A, T8) = 3, in Table 1(a).
* s(ip), external utility, is the value associated with item ip in the Utility Table. This
  + value reflects the importance of an item, which is independent of transactions.
  + For example, in Table 1(b), the external utility of item A, s(A), is 3.
* u(ip, Tq), utility, the quantitative measure of utility for item ip in transaction Tq, is
  + defined as o(ip,Tq) × s(ip). For example, u(A, T8) = 3 × 3 = 9, in Table 1
* u(X, Tq), utility of an itemset X in transaction Tq, is defined as

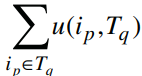
where X = {i1, i2, …, ik} is a k-itemset, X ⊆ Tq and 1≤ k ≤ m.

* u(X), utility of an itemset X, is defined as

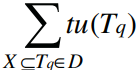
Utility mining is to find all the itemsets whose utility values are beyond a user specified threshold. An itemset X is a high utility itemset if u(X) ≥ ε, where X ⊆ I and ε is the minimum utility threshold, otherwise, it is a low utility itemset. For example, in Table 1, u({A, D, E}) = u({A, D, E}, T4) + u({A, D, E}, T8) = 14 + 32 = 46. If ε = 120, {A, D, E} is a low utility itemset [**11]**.

#### Phase I:

**Definition 1. (Transaction Utility)**: The transaction utility of transaction Tq, denoted

as tu(Tq), is the sum of the utilities of all the items in Tq: tu(Tq) = Table 1

(c) gives the transaction utility for each transaction in Table 1.

**Definition 2. (Transaction-weighted Utilization):** The transaction-weighted utilization of an itemset X, denoted as twu(X), is the sum of the transaction utilities of all the transactions containing X:

For the example in Table 1, twu(AD) = tu(T4) + tu(T8) = 14 + 57 = 71.

**Definition 3. (High Transaction-weighted Utilization Itemset)**: For a given itemset

X, X is a high transaction-weighted utilization itemset if twu(X) ≥ ε’, where ε’ is the

user specified threshold.

#### **Phase II:**

In Phase II, one database scan is performed to filter the high utility itemsets from high transaction-weighted utilization itemsets identified in Phase I. The number of high

transaction-weighted utilization itemsets is small when ε’ is high. Hence, the time saved in Phase I may compensate for the cost incurred by the extra scan in Phase II. In Figure

1, the high utility itemsets ({B}, {B, D}, {B, E} and {B, D, E}) are covered by the high transaction-weighted utilization itemsets. One database scan is performed in Phase II

to prune 5 of the 9 itemsets since they are not high utility itemsets.

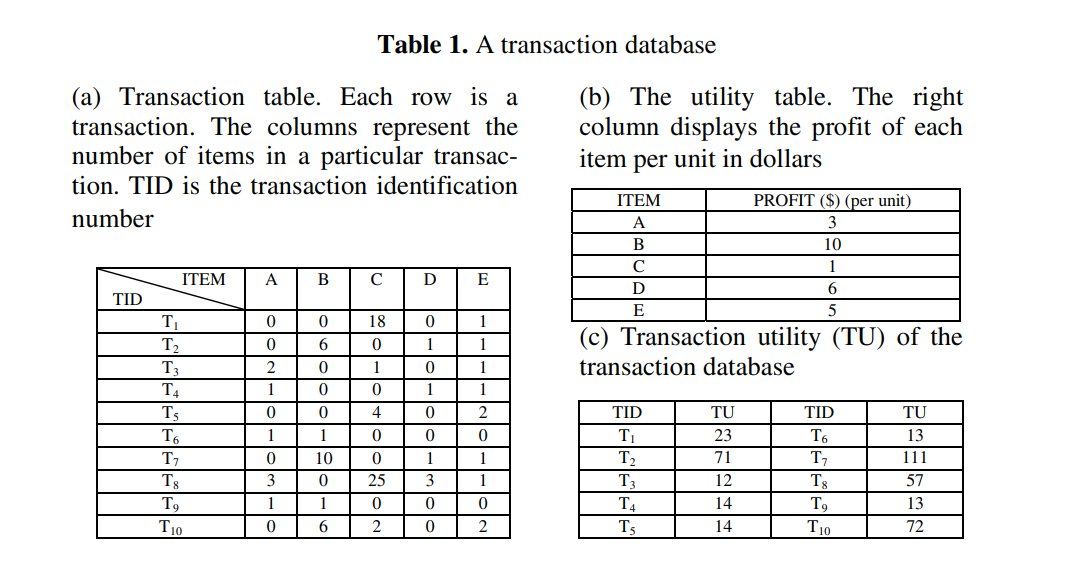


Table 1 Transaction Database

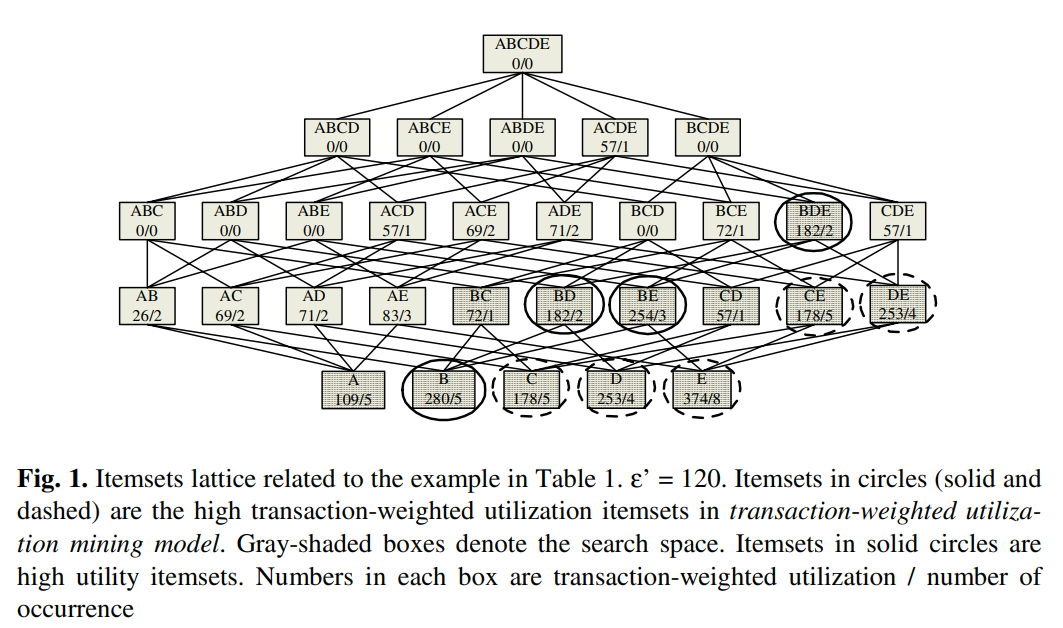


Fig. 1 Itemsets lattice related to the example in Table 1. ɛ’ = 120. Itemsets in circles (solid and dashed) are the high transaction-weighted utilization itemsets in transaction-weighted utilization mining model. Gray-shaded boxes denote the search space. Itemsets in solid circles are high utility itemsets. Numbers in each box are transaction-weighted utilization / number of occurrences.

## **Example of Two-phase application in market basket analysis:**

Let us understand in laymen’s terms:

Consider the transaction table as below:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Tid Items | Apple | Orange | Plum | Watermelon |
| T1 | 0 | 8 | 0 | 1 |
| T2 | 8 | 0 | 6 | 2 |
| T3 | 5 | 2 | 1 | 2 |
| T4 | 10 | 2 | 0 | 0 |
| T5 | 0 | 0 | 0 | 2 |
| T6 | 0 | 0 | 2 | 1 |

Table 2 Transaction Database

The utility table will be the price of the product:

|  |  |
| --- | --- |
| Product | Price (in $/unit) |
| Apple | 1 |
| Orange | 2 |
| Plum | 3 |
| Watermelon | 10 |

Table 3 Utility table

Transaction Utility (TU) of the transaction database:

|  |  |
| --- | --- |
| TID | TU |
| T1 | 26 |
| T2 | 46 |
| T3 | 32 |
| T4 | 14 |
| T5 | 20 |
| T6 | 16 |

Table 4 Transaction Utility Table

With minimum utility threshold of 20 we get the following output

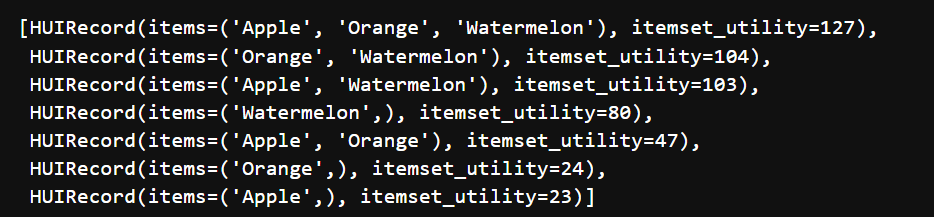


Fig. 2 Highest Utility Itemsets with MinUtil - 20000

This clearly shows the itemsets that generates more profit to the store. Here, the Itemset utility of watermelon as a single item is highest though it is bought in less quantity than other fruits. Therefore, it shows that quantity of itemset is as important as association with another item.

## **Related Works**

Several works in the literature have explored the notion of student engagement levels and methodologies on how to define and determine these levels and their impact on the academic performance of the student.

Wei Wang studies the online learning frequency, learning engagement and mental health of college students during Covid-19. He conducted online survey with online tool and through exploratory factor analysis, it was observed that the online learning engagement of college students during the epidemic period is composed of three factors, namely "emotional engagement", "cognitive engagement" and "behavioural engagement".

When it comes to mobile devices for online learning, 82.5 percent of students chose laptops, followed by mobile phones with 69 percent, tablets with 11.9 percent and desktop computers with 7.5 percent. Through independent sample T test, it was known the mean value of positive affect of college students in urban areas was significantly lower than that of college students in rural areas. In terms of learning engagement, the mean emotional engagement of college students in urban areas was significantly lower than that of college students in rural areas. The correlation analysis showed that there was a significant positive correlation between positive affect and emotional engagement, cognitive engagement and behavioural engagement, the correlation coefficients are respectively.

Also, Chih-Hsien Hsia, Bryan Chiang, Liang-Ying Ke, Zih-Yan Ciou, Chin-Feng Lai studies student engagement using facial expression in online course. This study collects bigdata images through computer vision, collects data on the students’ emotional and behavioural dimensions under multidimensionality, and analyses these data through deep learning (DL) with the student engagement model. Student engagement is considered to be synonymous with educational quality and is positively correlated with a student’s perseverance, satisfaction, learning efficiency, and degree completion. Student engagement is a reflection of a student’s internal psychological state, which includes behaviour, cognition and emotion. Student engagement is related to the student’s psychological participation in activities, and the quality and frequency of participation in the process can further predict learning outcomes.

In the prediction of student participation, emotional and behavioural dimension data obtained from different students is used to find out the weight relations between these two engagement dimensions and engagement value through the regression model of DL and evaluate participants’ engagement status in online learning.

It is used in the detection of macro expression and micro expression of human face in biometric image recognition, as well as the participation detection and evaluation of students’ learning by recording class notes, to give teachers an objective standard to measure students’ learning outcomes.

They used CNN(Efficient-B0) and LSTM on two regions on student’s face to determine the concentration of the student.

However, most of the works that studied relationship between student engagement and their academic performance were based on traditional or classroom learning. Moreover, the few works that investigate this relationship in online learning environment did not use frequency related metrics or Quantitative metrics.

## **Aim**

Our aim in this project is to find High Utility transactions from student e-learning dataset, which can be able to improve the academic performance of the student.

**Scope**

Analysis of student engagement relationship with their academic performance in the e-learning environment.

# **Data Science lifecycle process**

## **Modules**

* Data Discovery
* Dataset
* Data Preparation
* Model Planning
* Model Building
* Implementation
* Analysis
* Problems faced and solved
* Problems we were not able to solve
* Lessons learned from the project
* Results and Findings

### Data Discovery

Data discovery was the most crucial step as many of the dataset does not or have very few quantitative features which can be used to generate the relationship. Finally, we were able to select a suitable dataset which is described as follows:

**Dataset**: Students' Academic Performance Dataset

Above dataset was obtained from: <https://www.kaggle.com/datasets/aljarah/xAPI-Edu-Data>

This is an educational data set which is collected from learning management system (LMS) called Kalboard 360. Kalboard 360 is a multi-agent LMS, which has been designed to facilitate learning through the use of leading-edge technology. Such system provides users with a synchronous access to educational resources from any device with Internet connection.

The data is collected using a learner activity tracker tool, which called experience API (xAPI). The xAPI is a component of the training and learning architecture (TLA) that enables to monitor learning progress and learner’s actions like reading an article or watching a training video. The experience API helps the learning activity providers to determine the learner, activity and objects that describe a learning experience.

The dataset consists of 480 student records and 16 features. The features are classified into three major categories: (1) Demographic features such as gender and nationality. (2) Academic background features such as educational stage, grade Level and section. (3) Behavioral features such as raised hand on class, opening resources, answering survey by parents, and school satisfaction.

The dataset consists of 305 males and 175 females. The students come from different origins.

The dataset is collected through two educational semesters: 245 student records are collected during the first semester and 235 student records are collected during the second semester.

The data set includes also the school attendance feature such as the students are classified into two categories based on their absence days: 191 students exceed 7 absence days and 289 students their absence days under 7.

This dataset includes also a new category of features; this feature is parent parturition in the educational process. Parent participation feature have two sub features: Parent Answering Survey and Parent School Satisfaction. There are 270 of the parents answered survey and 210 are not, 292 of the parents are satisfied from the school and 188 are not.

The dataset consists of 480 distinct instances. There are 17 columns in the dataset which are described below:

|  |  |  |  |
| --- | --- | --- | --- |
| Feature | Description | Type | Range of Values |
| **Gender** | student's gender | Nominal | 'Male' or 'Female’ |
| **Nationality** | student's nationality | Nominal | Kuwait, Lebanon, Egypt, Saudi Arabia, USA, Jordan, Venezuela, Iran, Tunis, Morocco, Syria, Palestine, Iraq, Lybia |
| **Place of birth** | student's Place of birth | Nominal | Kuwait, Lebanon, Egypt, Saudi Arabia, USA, Jordan, Venezuela, Iran, Tunis, Morocco, Syria, Palestine, Iraq, Lybia |
| **Educational Stages** | educational level student belongs | Nominal | lower level, MiddleSchool, High School |
| **Grade Levels** | grade student belongs | Nominal | G-01 to G-12 |
| **Section ID** | classroom student belongs | Nominal | A, B, C |
| **Topic** | course topic | Nominal | English, Spanish, French, Arabic, IT, Math, Chemistry, Biology, Science, History, Quran, Geology |
| **Semester** | school year semester | Nominal | First or second |
| **Parent responsible for student** | Parents more engaged with student | Nominal | Mom or father |
| **Raised hand** | how many times the student raises his/her hand on classroom | Numeric | 0-100 |
| **Visited resources** | how many times the student visits a course content | Numeric | 0-100 |
| **Viewing announcements** | how many times the student checks the new announcements | Numeric | 0-100 |
| **Discussion groups** | how many times the student participate on discussion groups | Numeric | 0-100 |
| **Parent Answering Survey** | parent answered the surveys which are provided from school or not | Nominal | Yes or no |
| **Parent School Satisfaction** | the Degree of parent satisfaction from school | Nominal | Yes or no |
| **Student Absence Days** | the number of absence days for each student | Nominal | above-7 or under-7 |
| **Class** | final grade of student | Nominal | Lower-level (0 to 69), middle-level(70 to 89), high-level (90-100) |

Table 5 Dataset Description

### Data Preparation

Cleaning data for missing values. Nominal columns such as Gender, Nationality, place of birth cannot be changed to quantitative features as it will turn the columns into ranks.

Converting each instance or transaction into Lists of Tuples with features and their respective weights, as required by the algorithm.

Only the columns, **Raised hand, Visited resources**, **Viewing announcements**, **Discussion groups** and **class** are taken into consideration as all other columns are of less importance or nominal.

The Values of class column is converted into ranks. Lower-lever is equal to 1, Midde-level is equal to 2 and High-level is equal to 3.

Creating a utility table based on user’s point of view and the importance of the features. The values of the features range between 0 – 100 which is quite high when generating high Utility transactions. Therefore, the values for each feature in the utility table range from 0 – 1. The values are decided according to the importance of the variable in the dataset.

Performed Exploratory Data Analysis to find relationship between features. Following are some important insights from data:

* Relationship between Visited resources and Class of Students

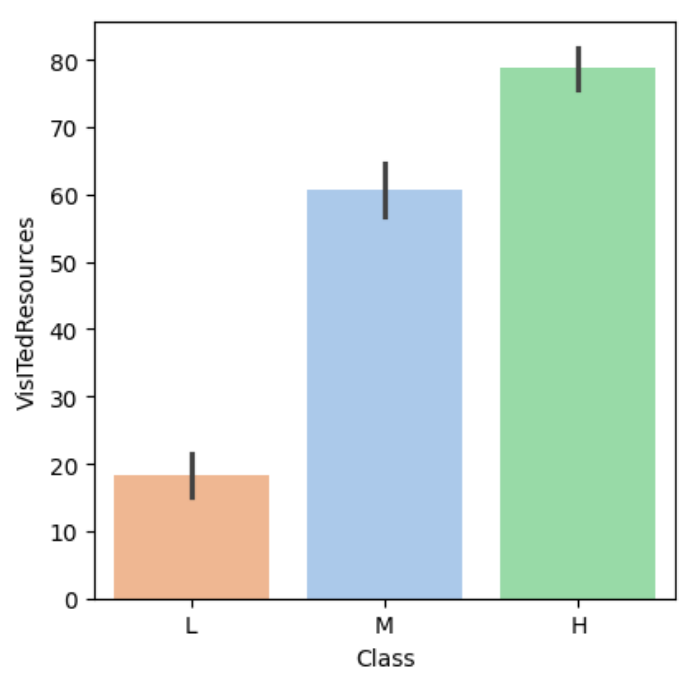


Fig. 3 Relationship Between visited resources and Class

* Relationship between Announcement view and Class of the students

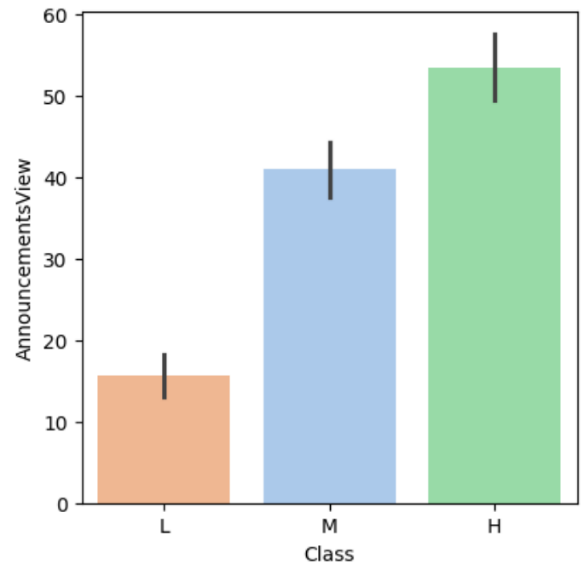


Fig. 4 Relationship between announcement view and class

* Relationship between Raised hands and class

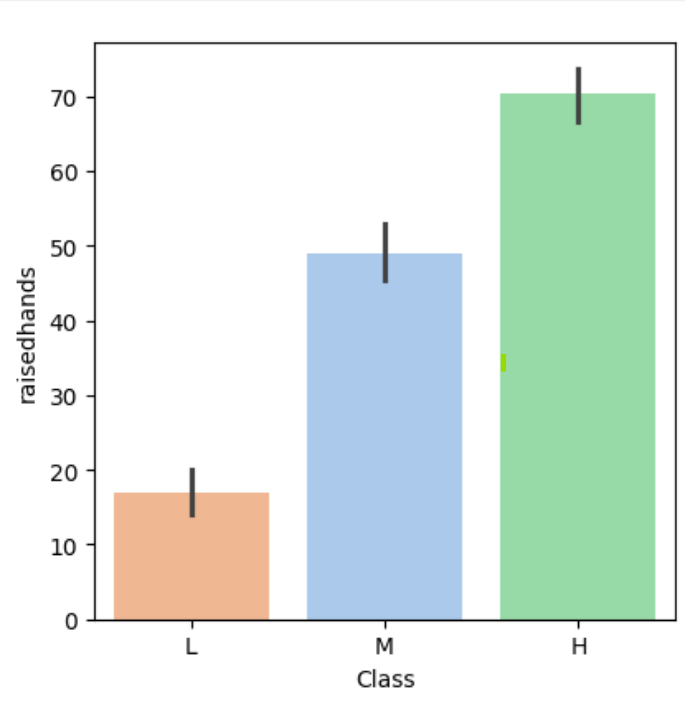


Fig. 5 Relationship between Raised hands and class

* Relationship between Discussion and Class of the students

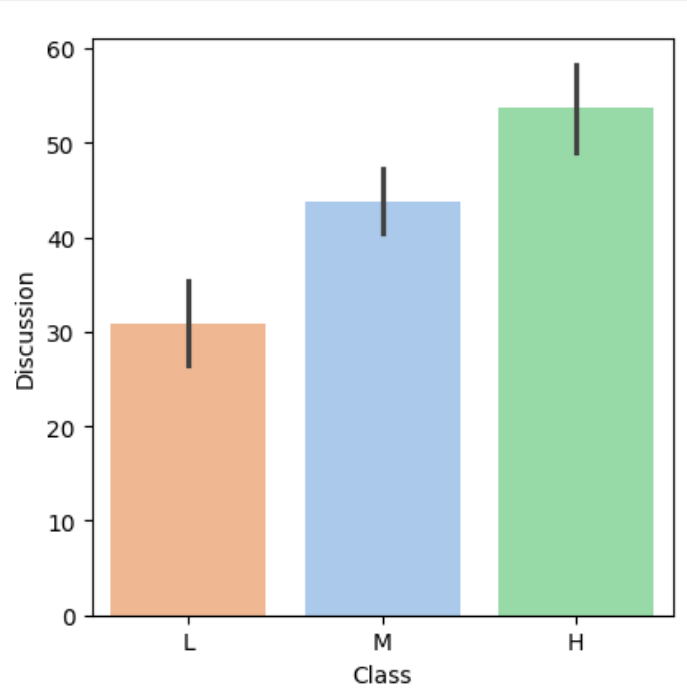
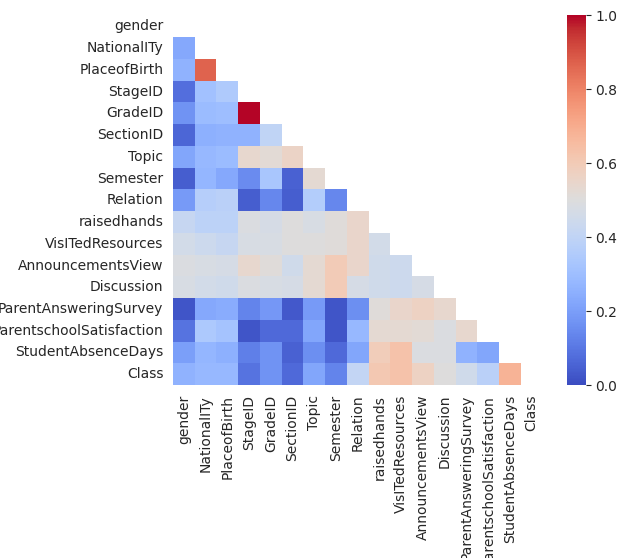


Fig. 6 Relationship between Discussion and Class

* Correlation between all the Features:

Fig. 7 Correlation Matrix: All features included



* Correlation between Quantitative features

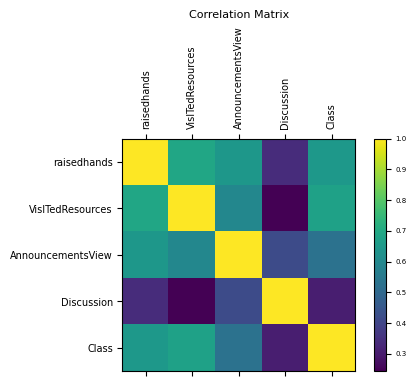


Fig. 8 Correlation matrix: Quantitative features

### Model Planning

We tried to use multiple HUIM algorithms. However, Two-phase was the fastest so we implemented that.

#### **Two-Phase:**

Let’s understand the algorithm in laymen’s term. Suppose you are a teacher who takes online classes for different grades. During your online classes, only few students have questions and they raised hands. However, many students who does not participate during lectures, are more actively participating on the discussion forum.

So, is the one participating on discussion forum is less engaged in education than the one asking question during online lectures? No. Therefore, using this algorithm we derived the relationship between student participation through any medium and their academic performance.

This algorithm works in two phases:

In phase 1, All the transactions (List of tuples with features and their weights) are multiplied with the utility values in the utility table. Transactions having value greater or equal to the minimum utility threshold provided by the user, are considered for the next phase. All other transactions are pruned.

In phase 2, one database scan is performed to filter the high utility itemsets from high transaction-weighted utilization itemsets identified in Phase I.

### Model Building

* Two-Phase HUIM bases on Utility table and minimum Utility threshold set by the user.
* In Market basket analysis the Utility table consist of the price or the amount of the money the product is worth of.
* In our case, the utility values must be something meaningful just like the price. Therefore, we decide that the utility values will be according to the importance of the features with respect to class of the student.
* We derived Correlation matrix to observe the associativity between quantitative variables and final grade of the student and assigned the values in utility table accordingly.
* For minimum threshold we tried and tested multiple values till we got certain desired itemsets of association.

### Implementation

Columns taken into consideration:

**Raised Hands**: Number of times student raised hand

**Visited Resources**: Number of times students visited resources

**Announcement View**: number of times student view announcements

**Discussion**: number of times students discussed on forums

**Class**: Final grades or result obtained by students

Utility table of the features:

|  |  |
| --- | --- |
| Features | Utility |
| **Raised Hands** | 0.5 |
| **Visited Resources** | 0.6 |
| **Announcement View** | 0.4 |
| **Discussion** | 0.3 |
| **Class** | 0.1 |

Table 6 Utility Table

Two-phase HUIM is applied on the transactions and Utility table is used to obtain Transaction Weighted utility Twu.

High Utility Itemsets are obtained by applying minimum threshold to all the itemset in each transaction having Twu.

* Trying multiple values for minimum threshold to select best possible threshold:
* With minimum threshold of 0 we get:

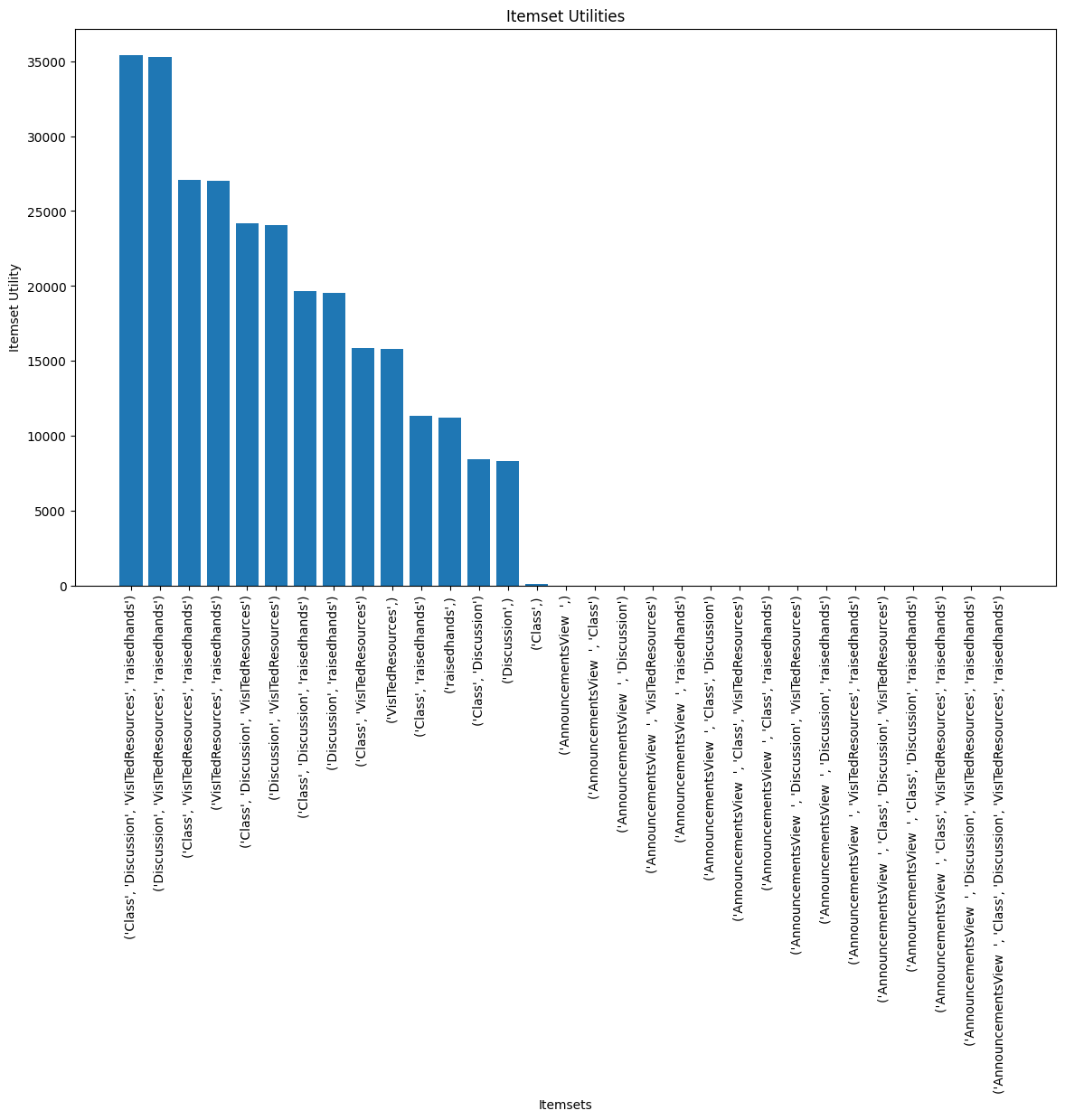


Fig. 9 High Itemset Utility with Minutil-0

* With minimum threshold of 8000 we get:

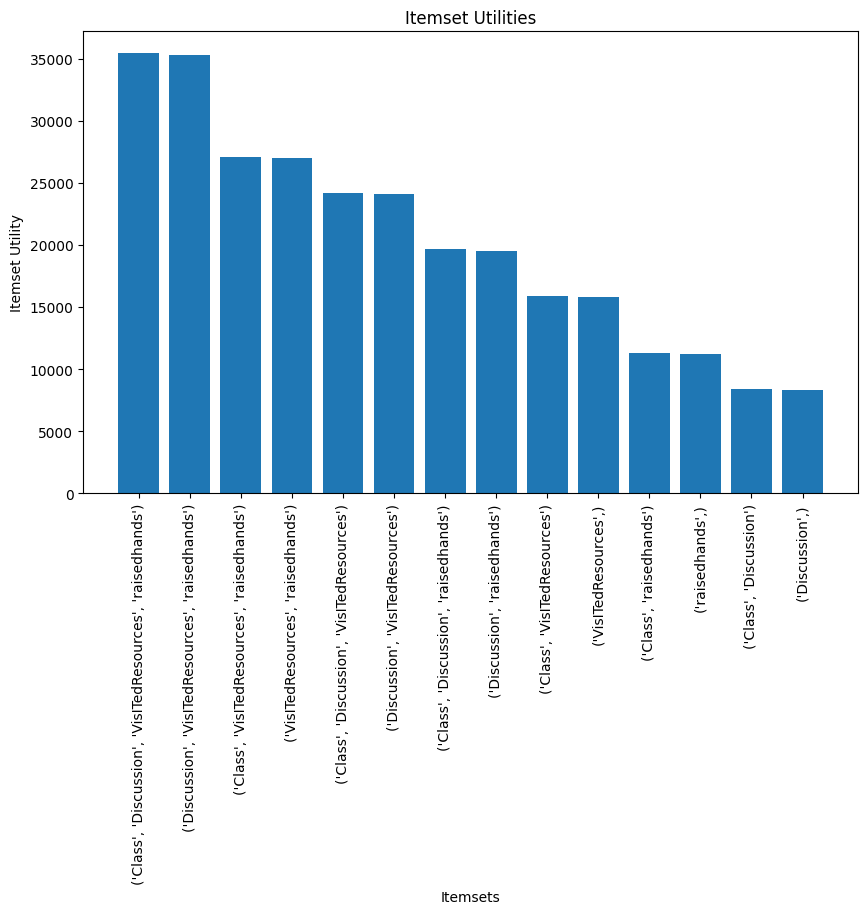


Fig. 10 HUIM with MinUtil-8k

* With minimum threshold of 20000 we get:

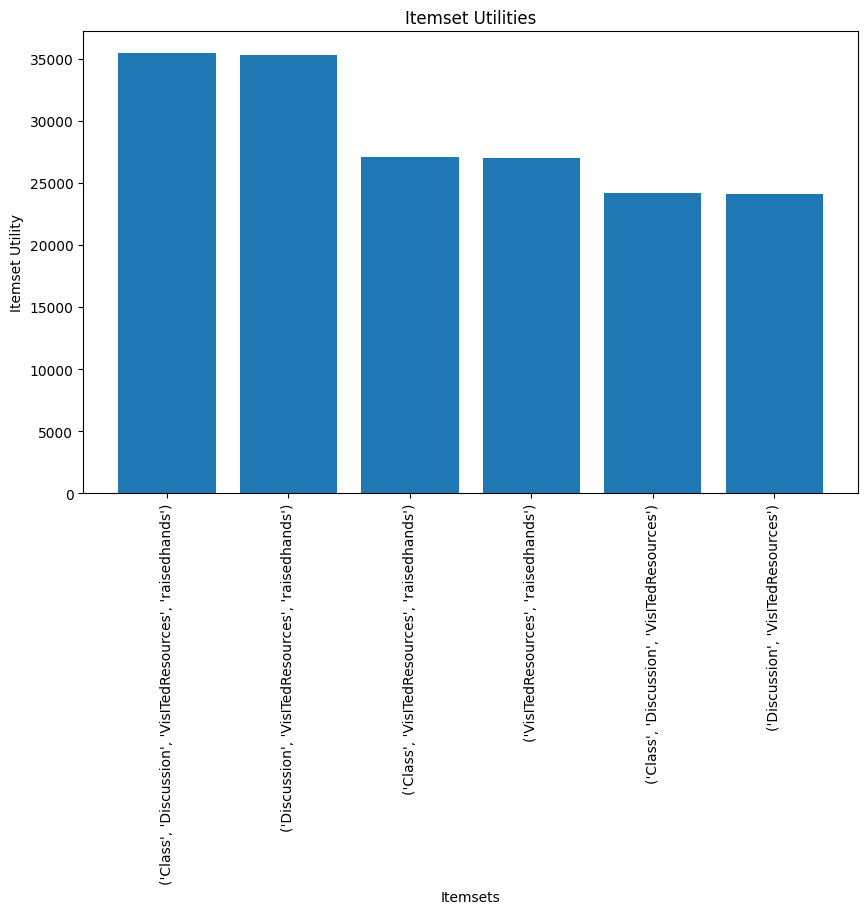


Fig. 11 HUIM with MinUtil-20K

* Testing for multiple values, we derived that MinUtil = 20000 gives the best possible itemset

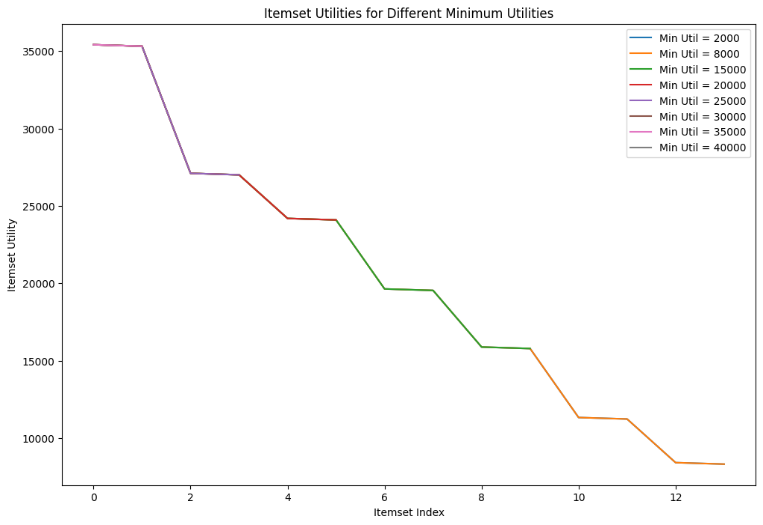


Fig. 12Testing for best possible Minimum threshold

### Analysis

Association between features and student academic performance shows positive relationship. Activities like visited resources have more contribution towards academic performance of a student.

### Problems faced and solved

* Dataset finding was a problem at the start of the project as most of the student dataset are either for offline or traditional classroom and very few of them are of online education. Moreover, if the dataset is of the online education, it does not have quantitative features that can be used to assess student performance. Finally found dataset that has few but quantitative features.
* Another problem faced was to select utility values in such a way that does not compromise the actual importance of the features. This was solved by finding the co-relation between features and class of the student.

### Problems we were not able to solve

* Certain features such as Semester, Student Absence Days etc. were important but could not be converted to quantitative features as converting them will assign ranks to the classes in that feature.

### Lessons Learned from the Project

* Frequent Pattern Itemset Mining algorithm often ignore items that have more benefit to the results.
* Student having more participation in activities often tent to get good academic results that other student who don not participate.
* Two-phase algorithm is faster and more reliable, but it has constraints, such as it can only be applied to quantitative features and only to dataset which is in transactional pattern, that is, no item is repeated in same transaction.
* Moreover, the weight of feature plays important role in HUIM. Therefore, sudden shift in weight can be impactful to the outcome. Therefore, the dataset needs to be balanced. For example, if half of the instances have high values in weights of features and half of them have very less values, this can affect the output and it can be biased.
* The scale of the quantitative features should be same to provide equal importance to all features.

### Output of the project

* With minimum threshold of 20000, following itemset were the Highest Utility itemsets according to the Two-Phase HUIM:

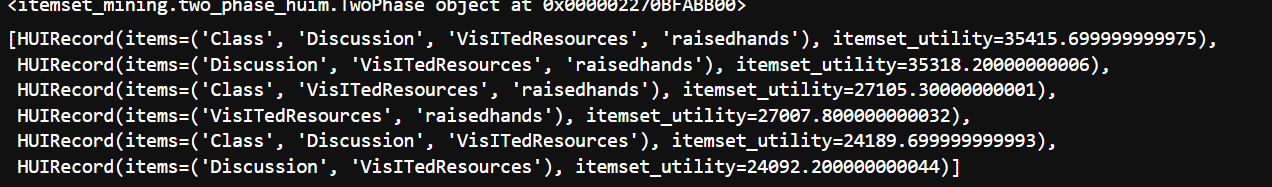


Fig. 13 High Utility itemset in student database

# **Conclusion**

## A summary of project and findings

Positive relationship between student engagement and academic performance in online learning. Thus, it can be concluded that with more activities, student get more engaged in learning.

Findings:

* Student engage in learning more when considered included. Therefore, not only academic activities but also cultural and sport activities can be helpful.
* Activities participation can predict which student need more help than other student. For example, less participation in verbal activities says that the student is shy or hesitant to speak. So, more such activities can be included where it does not require to participate verbally, if there are a greater number of such students.

Contribution:

The contribution of this work is summarised as follows:

* Using a comprehensive set of frequency-related metrics.
* Using Quantitative features for association instead of only qualitative features.
* Studying the impact of considered engagement metrics and academic performance.

# Recommendations

* Quantitative features can give more idea about the performance of the student than qualitative features,
* Other HUIM algorithms were very slow compare to two-Phase algorithm.
* Further research can be done by collecting more quantitative features and a large database.
* Minimum utility threshold plays important role in defining the HUI.
* Scaling the features is a very important steps if the features are unscaled.
* Balancing the dataset is recommended analysis before application of algorithm.

# References

[[1]Pandemic & Student Engagement: Annual Results 2021: Annual Results: Research: Evidence-Based Improvement in Higher Education: Indiana University](https://nsse.indiana.edu/research/annual-results/2021/story1.html)

[2][Coursera’s 2020: Year in Review — Class Central](https://www.classcentral.com/report/coursera-2020-year-review/)

[3] [Online Course Completion Rates (And Other 2021 Learning Stats) (elitecontentmarketer.com)](https://elitecontentmarketer.com/online-education-stats/)

[4]<http://www.katyjordan.com/MOOCproject.html>

[5][Benefits and Costs of MOOC-Based Alternative Credentials (wixstatic.com)](https://docs.wixstatic.com/ugd/cc7beb_5803e625ebee463ebc6f4796027366f1.pdf)

[6] [e-Learning: Challenges and Solutions (valuebound.com)](https://www.valuebound.com/resources/blog/e-learning-challenges-and-solutions)

[By the Numbers: MOOCs During the Pandemic — Class Central](https://www.classcentral.com/report/mooc-stats-pandemic/)

[7]G. Piatetsky-Shapiro, “Discovery, analysis and presentation of strong rules,” in Knowledge Discovery in Databases, G. Piatetsky-Shapiro and W. J. Frawley, Eds. AAAI Press, 1991, pp. 229–248.

[8]L. Antonie, “INTRODUCTION TO DATA MINING ASSOCIATION RULES,” University of Guelph

[9]R. Agrawal and R. Srikant, “Fast algorithms for mining association rules,” in Proc. of 20th Intl. Conf. on VLDB, 1994, pp. 487–499.

[10]Y. Liu, W.-k. Liao, and A. Choudhary, “A two-phase algorithm for fast discovery of high utility itemsets,” in Advances in Knowledge Discovery and Data Mining (T. B. Ho, D. Cheung, and H. Liu, eds.), (Berlin, Heidelberg), pp. 689–695, Springer Berlin Heidelberg, 2005.

[12] [Relationship between Student Engagement and Performance in e-Learning Environment Using Association Rules](https://ieeexplore.ieee.org/document/8451005)

[13][A Two-Phase Algorithm for Fast Discovery of High Utility Itemsets](https://link.springer.com/chapter/10.1007/11430919_79)