Identifying At-Risk Course Combinations for Freshman Students Using Market Basket Analysis

Fikrewold Bitew and Khoi To

About the Authors

Fikrewold Bitew, PhD, and Khoi To, PhD, have a combined 25 years of experience in the field of institutional research. They currently work in the Office of Institutional Research and Analysis at the University of Texas at San Antonio. They are interested in applying data mining, artificial intelligence and machine learning, and statistical modeling to support decision-making at various levels.

Abstract

As institutions increasingly focus on improving student retention and graduation rates, understanding factors that influence student success has become crucial. Course failures in students' first terms have been shown to have strong associations with student retention. While extensive research has analyzed failure rates of individual courses, this study advances the field by using market basket analysis to identify courses that, when taken during the same semester, lead to high probabilities of failure. These insights are not available when courses are analyzed individually.

Using data from 7,466 first-time, full-time freshman students across five cohorts (Fall 2018–Fall 2022) at The University of Texas at San Antonio, we applied the Apriori algorithm to analyze 15,698 course enrollment records with failing grades (F, D–, D, D+). Our findings identified seven high-risk course combinations, with mathematics courses (MAT 1053, MAT 1073) appearing in four combinations and a writing course (WRC 1013) appearing in five combinations. Chi-square analyses revealed that students taking both courses in these combinations had significantly lower first-term retention rates (69.3%–81.0% vs. 83.0%–86.5%) and first-year retention rates (34.6%–54.0% vs. 54.6%–59.6%) compared to students taking only one course from the same combination.

The AIR Professional File, Fall 2025

Article 181

https://doi.org/10.34315/apf1812025

These findings provide actionable insights for academic advisors and curriculum designers to implement targeted intervention strategies, to enhance course scheduling guidance, and to develop support systems for high-risk course combinations. The study demonstrates the value of data-driven approaches in higher education and establishes a methodological framework that other institutions can replicate to improve student success outcomes.

Keywords: freshman students, at-risk course combinations, market basket analysis (MBA)

INTRODUCTION

Student success is a critical measure of institutional effectiveness in higher education, with retention and graduation rates serving as key indicators. However, despite substantial resources dedicated to supporting students, many of them still face challenges in achieving academic success. Understanding the factors that influence student success or failure is, therefore, crucial for developing strategies to improve academic outcomes. A key factor influencing student success is course failures in students' first semester, which research has identified as a strong predictor of academic performance (Slim et al., 2016). From that standpoint, being able to identify two or more courses that would pose a high risk of failure when taken together can help institutions design better support systems and interventions, enhance academic advising, and ultimately improve retention and graduation rates.

Much work has been done in analyzing failure rates of single courses. This study will take it one step farther by using market basket analysis (MBA) to identify courses that, if taken together, would lead to the high probability that a student will fail one or both courses. This data-mining technique, traditionally used in retail to identify associations between products that are often purchased together, offers a unique approach to examining these factors in an educational setting. By applying MBA to academic records, we can uncover patterns and associations among courses taken by students who encounter academic difficulties. The findings will provide actionable recommendations for academic advisors and curriculum designers, and will inform targeted intervention strategies, ultimately contributing to improved student performance.

LITERATURE REVIEW

The concept of "student success" encompasses multiple dimensions depending on stakeholder perspectives (Ezarik, 2023). It could include retention and persistence rates, degree completion, post-graduation outcomes such as employment or graduate school enrollment, or the acquisition of specific skills such as communication and critical thinking. Within this analysis, we focus on retention as our primary metric of student success; we define "retention" as the rate of return among students from term to term or from year to year (Soika, 2021).

Research has shown that certain courses, often referred to as "high-risk courses" or "bottleneck courses," have higher rates of failure and can significantly impact student retention and eventual graduation. A study by the Colorado State University Office of Institutional Research, Planning and Effectiveness (2012) found a strong negative association between unsuccessful course attempts and retention for first-time, full-time freshman students.

Similarly, a study by Michaels and Milner (2021) revealed significant gaps in retention and graduation among students receiving different numbers of D/F grades in the first term. On average, students saw an 8.5% retention differential with one D/F grade in the first term and a 35% retention differential with more than one D/F grade, as compared to students with no D/F grade in the first term. Although the data were not causal, early grades were clearly correlated with student success.

A recent study conducted by the University of Wyoming's Office of Institutional Analysis (Zong & Koller, 2023) also pointed out that students taking fewer high-risk courses are more likely to retain after the first year. The study identified high-risk courses as those with a pass rate of less than 80% over 5 years. These courses were predominantly in science, technology, engineering, and mathematics (STEM) fields, with quantitative reasoning and physical sciences being particularly challenging courses for students.

Recent advances in business intelligence have enabled researchers to apply data mining techniques to analyze student course-taking patterns. MBA has emerged as a pioneering approach to identify course combinations associated with high failure rates. This data mining technique, originally developed for retail to identify product associations, has been effectively adapted for educational settings (Papadogiannis et al., 2024; Romero & Ventura, 2020).

This approach provides insights that are not observable if individual courses were examined in isolation. For instance, students who enroll in multiple high-risk courses in their first semester are more likely to struggle academically and therefore are more vulnerable to attrition. Additionally, the analysis can reveal hidden patterns, such as the impact of taking certain elective courses alongside core courses (McKinsey & Company, 2023).

Safour et al. (2024) applied MBA to identify associations between courses in a university setting, finding that certain course combinations were more likely to result in student dropouts. Similarly, González et al. (2008) used MBA to explore the relationship between course sequences and student performance, revealing that some course sequences were strongly associated with lower grades.

Bautista (2005) applied MBA to analyze first-year engineering student course combinations, identifying the top 20 combinations with high failure rates. The study highlighted significant financial implications of course failures, since unsuccessful attempts result in extended enrollment periods, creating additional financial burdens for students and their families, and creating pressure on government funding sources in public institutions.

RESEARCH QUESTIONS AND SIGNIFICANCE OF THE STUDY

Given the demonstrated impact of course failures on student retention, this study analyzes data from first-time, full-time freshman students at The University of Texas at San Antonio (UTSA), using MBA to explore first-semester course-taking patterns. Two primary research questions guide this investigation:

Research Question 1: Among the courses taken by first-time, full-time freshman students at UTSA in their first semester, which course combinations, when taken together, lead to high probabilities of failure?

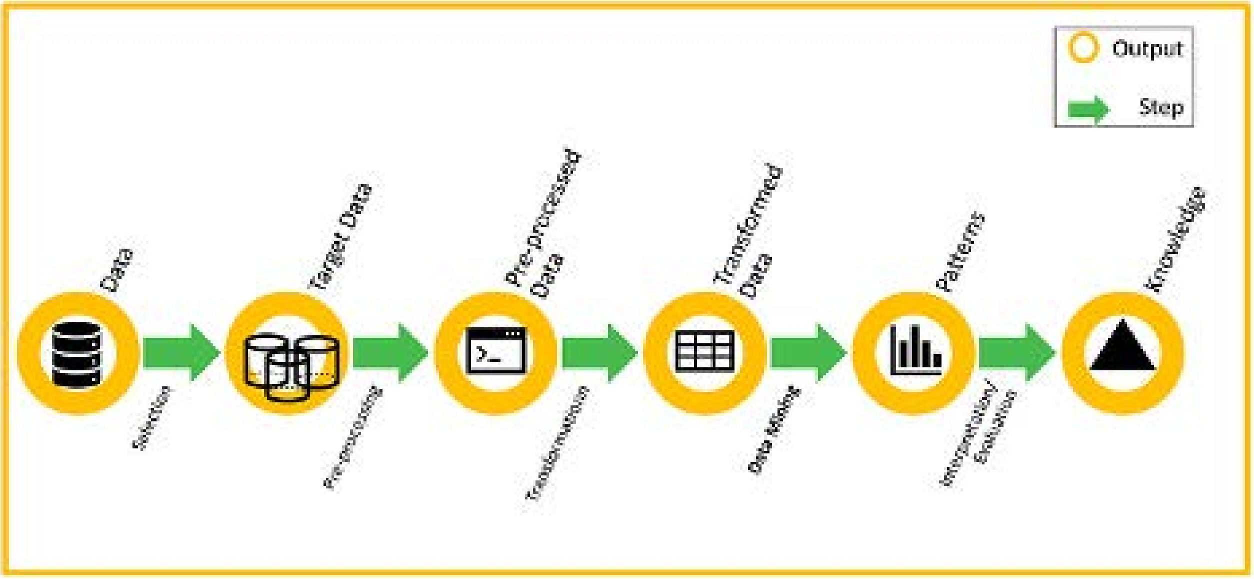
Research Question 2: Are there significant relationships between first-term retention and first-year retention when comparing two groups of students—those taking only one course from a high-risk combination versus those taking both courses from the same combination?

Findings from these research questions contribute to understanding how certain course combinations lead to high failure rates and adversely affect student retention. This knowledge enables university administrators, academic advisors, and curriculum planners to develop effective strategies for teaching methodology, curriculum design, and targeted student support systems.

DATA AND METHODS

This study applies the Knowledge Discovery in Databases (KDD) approach that was developed by Usama Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth (see Figure 1). This framework is recognized as being one of the most reliable research methodologies for academic purposes, since it was designed to identify hidden patterns, unseen trends, and correlations in databases to inform future decision-making (Bautista, 2005).

Figure 1. Knowledge Discovery in Databases (KDD) Approach in Data Mining



The Knowledge Discovery in Databases (KDD) approach was systematically applied to analyze student course data at UTSA, with each step detailed below in this section, as well as in Results and Discussion.

Data Selection

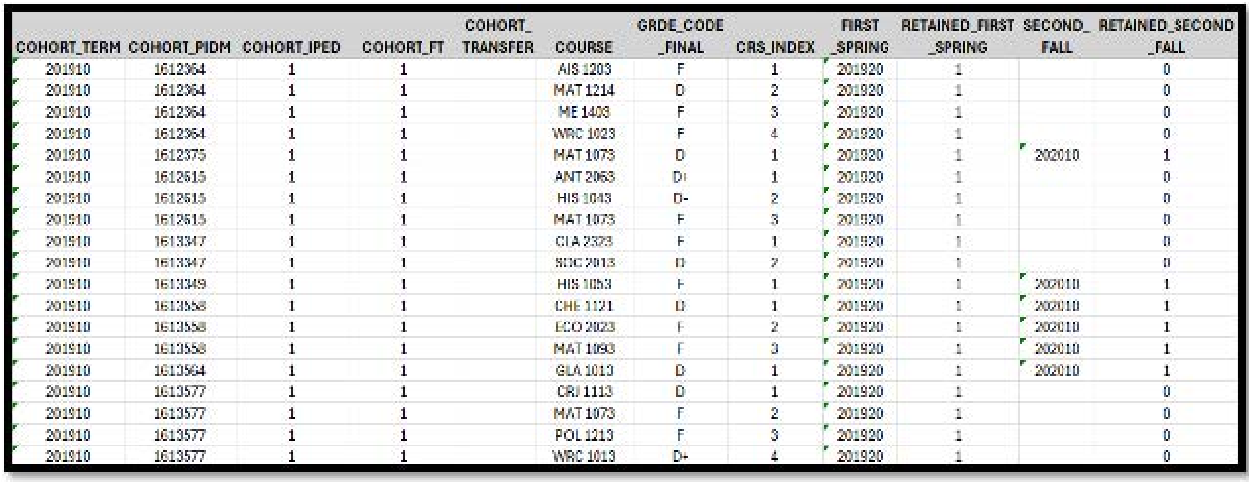
The study uses anonymized data from UTSA, covering course enrollments and final grades of five first-time, full-time freshman cohorts (Fall 2018–Fall 2022). Courses with final grades indicating failure (F, D–, D, D+) were retained for the analysis.

We use Oracle SQL to retrieve data from the university's student database (DMARTPROD), extracting student-course records needed for the analysis. The dataset has a total of 15,698 records and was structured in a transactional layout, with each row representing a student-course enrollment ("a transaction") in each given term. This layout is also known as the "long format."

The dataset was exported as an Excel file and imported into Python for processing and analysis. The following describes each variable in the dataset, with a representative sample presented in Figure 2.

- COHORT_TERM: Identifies the student cohort at UTSA, representing the student's first enrollment term (201910, 202010, 202110, 202210, and 202310, corresponding to Fall 2018, Fall 2019, Fall 2020, Fall 2021, and Fall 2022, respectively).
- COHORT_PIDM: Unique student identifiers (anonymized to protect privacy).
- COHORT_IPED: Binary indicator of cohort status (1 = Roadrunner, 0 = CAP). "Roadrunner" refers to fully enrolled UTSA students (represented by the university mascot). "CAP" refers to students in the University of Texas at Austin Coordinated Admission Program (CAP) who begin their studies at UTSA. All students in this dataset are Roadrunners.
- COHORT_FT: Binary indicator of full-time enrollment status (1 = full time, 0 = part time). All students in this dataset are full time.
- COHORT_TRANSFER: Indicates student type (1 = new transfer, null = first timer). All students in this dataset are first timers.
- COURSE: Lists courses taken during the first semester that resulted in final grades of F, D–, D, or D+. Each row represents one course per student, so students with multiple failing grades appear in multiple rows.
- GRDE_CODE_FINAL: Final grade received (F, D–, D, or D+) for the corresponding course.
- CRS_INDEX: Sequential counter of failing courses per student in the first semester (reference only, not used in analysis). For example, a student with four failing courses would have index values 1–4 across their respective rows.
- FIRST_SPRING: Spring term enrollment indicator showing whether the student returned for the subsequent Spring semester.
- RETAINED_FIRST_SPRING: Binary indicator derived from FIRST_SPRING (1 = retained for Spring semester). Used in chi-square tests to examine relationships between course combinations and first-term retention.
- SECOND_FALL: Fall term enrollment indicator showing whether the student returned for the second Fall semester (first-year retention).
- RETAINED_SECOND_FALL: Binary indicator derived from SECOND_FALL (1 = retained for second Fall semester). Used in chi-square tests to examine relationships between course combinations and first-year retention.

Figure 2. Sample of Dataset Queried Out from The University of Texas at San Antonio's Student Database (DMARTPROD)



COHORT_TERM	COHORT_PIDM	COHORT_IPED	COHORT_FT	COHORT_ TRANSFER	COURSE	GRDE_CODE _FINAL	CRS_INDEX	FIRST _SPRING	RETAINED_FIRST _SPRING	SECOND_ FALL	RETAINED_SECOND _FALL
201910	1612364	1	1		AIS 1203	F	1	201920	1		0
201910	1612364	1	1		MAT 1214	D	2	201920	1		0
201910	1612364	1	1		ME 1403	F	3	201920	1		0
201910	1612364	1	1		WRC 1023	F	4	201920	1		0
201910	1612375	1	1		MAT 1073	D	1	201920	1	202010	1
201910	1612615	1	1		ANT 2063	D+	1	201920	1		0
201910	1612615	1	1		HIS 1043	D-	2	201920	1		0
201910	1612615	1	1		MAT 1073	F	3	201920	1		0
201910	1613347	1	1		CLA 2323	F	1	201920	1		0
201910	1613347	1	1		SOC 2013	D	2	201920	1		0
201910	1613349	1	1		HIS 1053	F	1	201920	1	202010	1
201910	1613558	1	1		CHE 1121	D	1	201920	1	202010	1
201910	1613558	1	1		ECO 2023	F	2	201920	1	202010	1
201910	1613558	1	1		MAT 1093	F	3	201920	1	202010	1
201910	1613564	1	1		GLA 1013	D	1	201920	1	202010	1
201910	1613577	1	1		CRJ 1113	D	1	201920	1		0
201910	1613577	1	1		MAT 1073	F	2	201920	1		0
201910	1613577	1	1		POL 1213	F	3	201920	1		0
201910	1613577	1	1		WRC 1013	D+	4	201920	1		0

Data Pre-processing

From this point forward, we utilize Python and Jupyter Notebook, taking advantage of their robust data manipulation, analysis, and visualization capabilities in data mining.

Relevant packages are imported into Python, and the Excel data file is loaded and set up for pre-processing.

```
# Load relevant packages into Python
import pandas as pd
import numpy as np
from mlxtend.frequent_patterns import apriori, association_rules
from mlxtend.preprocessing import TransactionEncoder
import matplotlib.pyplot as plt
import seaborn as sns
```

```
# Load the data from Excel file
df = pd.read_excel('student_course_data.xlsx')
print(f"Data shape: {df.shape}")
print(f"Columns: {list(df.columns)}")
```

Data quality assessment revealed no missing values or duplicate records. All variable data types were verified as appropriate for the intended analyses.

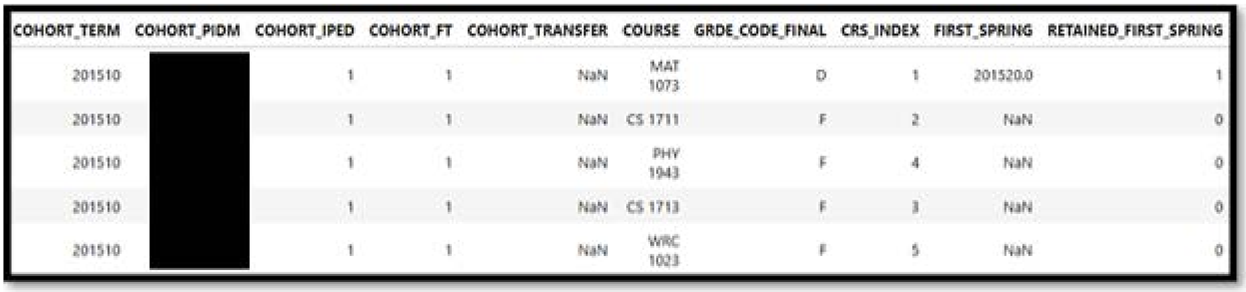
```
# Check for missing values
print("Missing values per column")
print(df.isnull().sum())

# Check for duplicate records
print(f"\nDuplicate records: {df.duplicated().sum()}")

# Display basic statistics
print("\nDataset overview:")
print(df.info())

# Display first few rows
print("\nFirst 5 rows of the dataset")
df.head()
```

Following is a screenshot of the first few rows of the dataset loaded in Python.



COHORT_TERM	COHORT_PIDM	COHORT_IPED	COHORT_FT	COHORT_TRANSFER	COURSE	GRDE_CODE_FINAL	CRS_INDEX	FIRST_SPRING	RETAINED_FIRST_SPRING
201510	[redacted]	1	1	NaN	MAT 1073	D	1	201520.0	1
201510	[redacted]	1	1	NaN	CS 1711	F	2	NaN	0
201510	[redacted]	1	1	NaN	PHY 1943	F	4	NaN	0
201510	[redacted]	1	1	NaN	CS 1713	F	3	NaN	0
201510	[redacted]	1	1	NaN	WRC 1023	F	5	NaN	0

Table 1 presents a summary of the data. Several key patterns emerge from these data as described below.

1| Distribution of grades: F grades are consistently the most common low grade across all semesters, comprising 9,623 instances (61.3% of all low grades). D grades are the second-most common (4,348 or 27.7%), while D+ (815 or 5.2%) and D– (912 or 5.8%) occur less frequently.

2| Pandemic impact: The data suggest possible effects of the COVID-19 pandemic, with Fall 2020 showing a notable increase in F grades (2,134) while having fewer D-range grades, potentially indicating greater academic challenges during this period.

3| Recent developments: The most recent data (Fall 2022) show some improvement, with fewer F grades compared to the previous year while still maintaining relatively high counts of D– grades.

Table 1. Descriptive Summary of the Dataset in Figure 2

Cohort	Final Grade				Total
	F	D–	D	D+	
Fall 2018	1,657	107	1,126	99	2,989
Fall 2019	1,494	66	1,019	87	2,666
Fall 2020	2,134	102	845	73	3,154
Fall 2021	2,415	310	679	295	3,699
Fall 2022	1,923	327	679	261	3,190
Total	**9,623**	**912**	**4,348**	**815**	**15,698**

Data Transformation

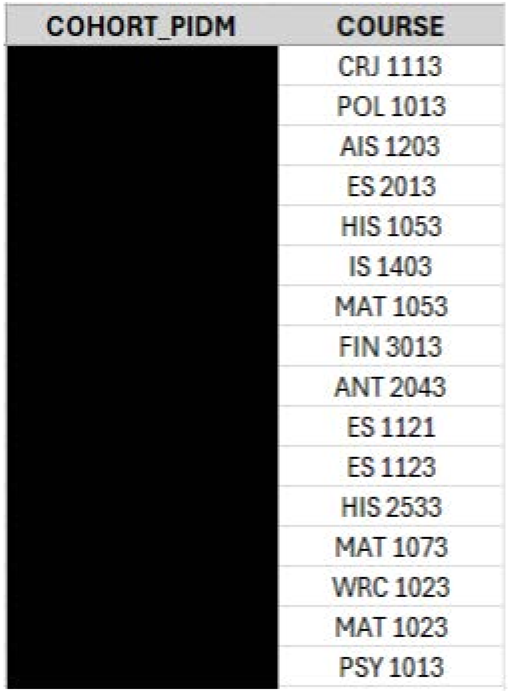
Before we perform MBA, we must convert the dataset into a format compatible with the Apriori algorithm. Specifically, the data require transformation from their current transactional layout (long format) into a tabular structure (wide format) comprising Boolean values. A “True” value indicates student enrollment in a course, while a “False” value indicates non-enrollment.

The first step in data transformation involves removing variables that are not relevant to MBA. Several columns were excluded, including cohort identifiers, full-time status, retention indicators, and grade information; the remainder focused exclusively on student-course relationships. The analysis retains only student identifiers (COHORT_PIDM) and courses taken (COURSE).

```
# Select relevant columns for transformation
columns_to_keep = ['COHORT_PIDM', 'COURSE']
df_transformed = df[columns_to_keep].copy()

# Display the structure of the filtered data
print("Filtered dataset structure")
print(df_transformed.head())
print(f"\nShape after filtering: {df_transformed.shape}")
print(f"Unique students: {df_transformed['COHORT_PIDM'].nunique()}")
print(f"Unique courses: {df_transformed['COURSE'].nunique()}")
```

The following screenshot displays a sample of the filtered dataset.



COHORT_PIDM	COURSE
[redacted]	CRJ 1113
[redacted]	POL 1013
[redacted]	AIS 1203
[redacted]	ES 2013
[redacted]	HIS 1053
[redacted]	IS 1403
[redacted]	MAT 1053
[redacted]	FIN 3013
[redacted]	ANT 2043
[redacted]	ES 1121
[redacted]	ES 1123
[redacted]	HIS 2533
[redacted]	MAT 1073
[redacted]	WRC 1023
[redacted]	MAT 1023
[redacted]	PSY 1013

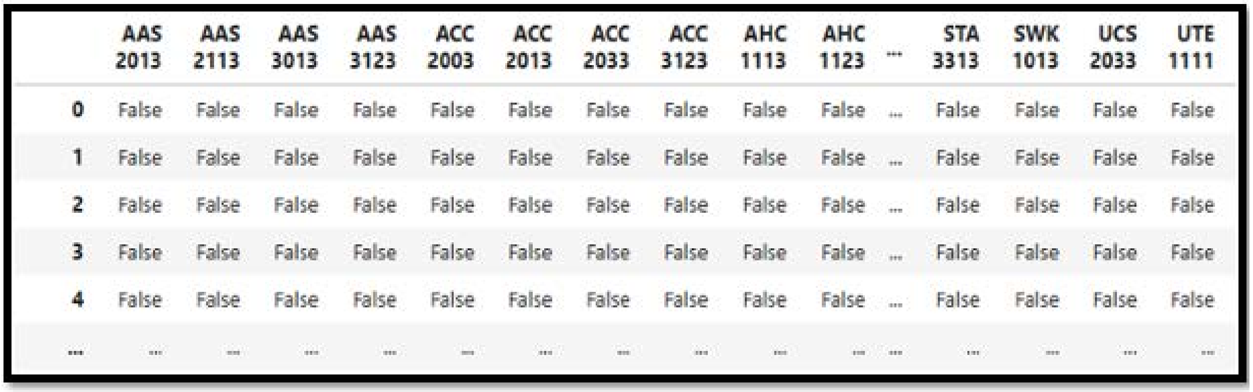
The next step involves transforming the data into a tabular format suitable for the Apriori algorithm, where each row represents a student and each column represents a course. The values in this transformed dataset are either “True” or “False.” A “True” value indicates that a student takes a specific course, and a “False” value means the student does not take that course.

```
# Create a pivot table to transform from long to wide format
basket = df_transformed.groupby(['COHORT_PIDM', 'COURSE']) .size().unstack().
fillna(0)

# Convert to boolean values (True/False)
def encode_units(x):
    if x <= 0:
        return False
    if x >= 1:
        return True
basket_sets = basket.applymap(encode_units)
print("Transformed dataset structure:")
print(f"Shape: {basket_sets.shape}")
print("Sample of transformed data:")
print(basket_sets.head())
```

After transformation, the dataset contains 7,466 rows, each representing a unique student, and 373 columns, each representing a distinct course. Since full-time students typically enroll in four courses per semester, each student row contains four “True” values among the 373 available columns, with the remaining columns showing “False” values. In the MBA analogy, students represent “baskets” and courses represent “items,” hence the term “market basket analysis.” The transformed dataset is now ready for analysis (Figure 3).

Figure 3. Sample of Transformed Dataset Ready for Market Basket Analysis



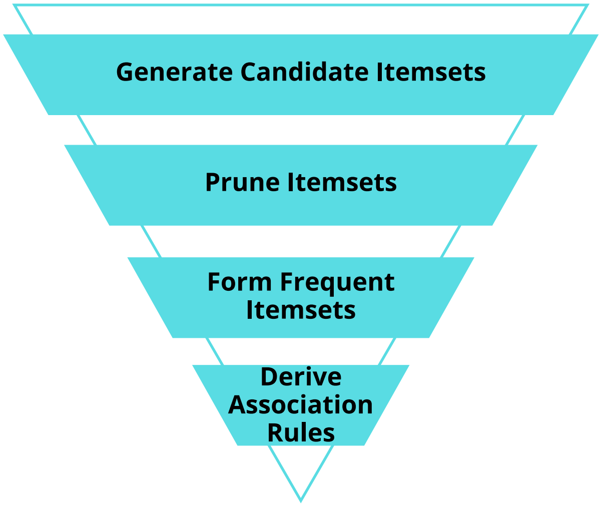
	AAS 2013	AAS 2113	AAS 3013	AAS 3123	ACC 2003	ACC 2013	ACC 2033	ACC 3123	AHC 1113	AHC 1123	...	STA 3313	SWK 1013	UCS 2033	UTE 1111
0	False	False	False	False	False	False	False	False	False	False	...	False	False	False	False
1	False	False	False	False	False	False	False	False	False	False	...	False	False	False	False
2	False	False	False	False	False	False	False	False	False	False	...	False	False	False	False
3	False	False	False	False	False	False	False	False	False	False	...	False	False	False	False
4	False	False	False	False	False	False	False	False	False	False	...	False	False	False	False
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...

Data Mining

There are several algorithms in MBA: they include Apriori, artificial immune system (AIS), sort and merge scan (SETM), and Frequent Pattern–Growth (FP–Growth). In this study, we use Apriori, a well-established and commonly used algorithm in MBA, to identify frequent itemset and association rules within large datasets. The Apriori algorithm was selected for its simplicity and ease of implementation. The algorithm's straightforward approach to generating candidate itemset and applying support and confidence thresholds facilitates both implementation and interpretation. However, Apriori has computational limitations, and requires multiple dataset scans that can demand significant memory and processing resources, particularly with large datasets containing numerous frequent patterns.

Figure 4 illustrates the steps in Apriori algorithm process. The following describes each step from initial data processing through identification of significant item combinations:

Figure 4. Steps in Apriori Algorithm Process



Source: Mwiti 2025.

- **Step 1: Generate candidate itemset:** The algorithm identifies individual items and counts their occurrences to determine frequent items. The fundamental principle states that, if an itemset is frequent, all its subsets must also be frequent. This assumption reduces the number of itemsets requiring evaluation, and improves algorithmic efficiency.
- **Step 2: Prune itemsets:** Itemsets occurring below the minimum support threshold are eliminated from further consideration.
- **Step 3: Form frequent itemsets:** The algorithm creates larger itemsets by combining frequent smaller itemsets, iterating until no additional frequent itemsets can be identified.
- **Step 4: Derive association rules:** The algorithm extracts meaningful association rules based on support, confidence, and lift values to identify significant relationships among items.

In this analysis, itemsets are defined as combinations of courses taken by students; our goal is to identify rules that can predict course failure based on these combinations.

Metrics Used to Evaluate Association Rules Among Items

The Apriori algorithm uses four key metrics—support, confidence, lift, and Zhang's metric—to identify meaningful association rules among items. These metrics are essential for selecting significant association rules, particularly when analysis generates numerous potential relationships (Alangari & Alturki, 2020). Each metric is detailed next (Derouiche, 2024):

SUPPORT

Definition: Support measures the proportion of course combinations in the dataset that contain a particular course set.

Interpretation: High support suggests that the course set occurs frequently in the dataset. The higher the support, the more frequently the itemset occurs. Rules with a high support are preferred since they are likely to be applicable to a large number of future course combinations. For instance, if the course set {Math, Physics} appears in 20 out of 100 course combinations, the support would be 0.20 or 20%.

Formula:

$$\textit{Support } (A \rightarrow B) = \frac{\textit{Number of transactions containing both A and B}}{\textit{Total number of transactions}}$$

Confidence

Definition: Confidence measures the likelihood that Course B is taken when Course A is taken.

Interpretation: High confidence indicates a strong association between Courses A and B. The higher the confidence, the greater the likelihood that Course B (the right-hand side) will be taken when Course A (the left-hand side) is taken. For example, if 15 out of 20 course combinations that contain mathematics also contain physics, the confidence is 0.75 or 75%.

Formula:

$$\textit{Confidence } (A \rightarrow B) = \frac{\textit{Support}(A \cup B)}{\textit{Support}(A)}$$

LIFT

Definition: Lift measures how much more likely that Course B is taken when Course A is taken, compared to the likelihood of taking Course B independently.

Interpretation: A lift value greater than 1 indicates a positive association between Courses A and B, suggesting they are more likely to be taken together than independently. For instance, a lift of 1.5 means that enrolling in mathematics increases the likelihood of enrolling in physics by 50%.

Formula:

$$Lift(A \rightarrow B) = \frac{Confidence(A \rightarrow B)}{Support(B)}$$

ZHANG'S METRIC

Zhang's metric offers a valuable alternative measure for evaluating association rules in MBA. Unlike lift, which can sometimes overemphasize rare item relationships, Zhang's metric provides a more-balanced evaluation by considering both the presence and the absence of items. This metric offers a normalized measure (–1 to 1), where positive values indicate positive correlation, zero indicates independence, and negative values indicate negative correlation.

Zhang's metric is calculated using the following formula:

$$Zhang's\ Metric\ (A \rightarrow B) = \frac{Confidence(A \rightarrow B)-Support(B)}{1-Support(B)}$$

RESULTS AND DISCUSSION

Identifying Association Rules for Course Failure Patterns

With the dataset properly formatted, the Apriori algorithm was applied to identify frequent course combinations and their association rules. The minimum support threshold was set at 0.01 (1%) to capture meaningful patterns while avoiding overly restrictive filtering. This threshold requires course combinations to appear in at least 1% of total student records, equivalent to 75 students (7,466 × 1%). The minimum confidence was set at 0.2 (20%), indicating that, when a student fails one course in a combination, there is at least a 20% probability that the student will fail the associated course.

These moderate threshold values were selected considering the large course catalog (373 total courses) and the extensive number of possible combinations. Higher thresholds would eliminate potentially meaningful patterns due to the natural diversity in student course selections. Effective association rule mining requires careful threshold selection based on domain knowledge, data characteristics, and iterative refinement to identify actionable insights.

The analysis specifications targeted two-course combinations (length = 2) with minimum support of 0.01 (support > = 0.01) and minimum confidence of 0.2 (metric = “confidence,” min_threshold = 0.2).

```
# Filter frequent itemsets by length & support
apriori_model_colnames = apriori_model_colnames[
    (apriori_model_colnames['length'] == 2) &
    (apriori_model_colnames['support'] >= 0.01)
]

# Generate association rules
rules = mlxtend.frequent_patterns.association_rules(
    apriori_model_colnames,
    metric="confidence",
    min_threshold=0.2,
    support_only=False
)
```

The model identifies the top seven course combinations that are associated with failure (F, D–, D, D+ grades), if taken together. Table 2 shows these course combinations and their performance metrics of support, confidence, lift, and Zhang’s. The course combinations and the interplay among their performance metrics are discussed below.

Table 2. Course Combinations Leading to Failure for First-Time, Full-Time Freshman Students

Antecedents	Consequents	Support	Confidence	Lift	Zhang’s
BIO 1404	MAT 1073	0.023	0.460	3.080	0.711
AIS 1203	WRC 1013	0.020	0.315	1.695	0.437
CHE 1073	MAT 1073	0.030	0.285	1.908	0.532
HIS 1053	WRC 1013	0.017	0.277	1.488	0.350
HIS 1043	WRC 1013	0.015	0.276	1.481	0.343
MAT 1053	WRC 1013	0.014	0.238	1.282	0.234
MAT 1073	WRC 1013	0.035	0.235	1.261	0.243

For reference, the following are the full course titles for the identified combinations:

- AIS 1203–Academic Introduction and Strategies
- BIO 1404–Biosciences I
- CHE 1073–Basic Chemistry
- MAT 1053–Mathematics for Business
- MAT 1073–Algebra for Scientists and Engineers
- HIS 1043–United States History: Pre–Columbus to Civil War Era
- HIS 1053–United States History: Civil War Era to Present
- WRC 1013–Freshman Composition I

The support metric indicates the frequency of co-occurrence, with approximate values ranging from 0.014 to 0.035 across the top seven rules. While these support values may appear low, they are significant in educational context, where failing multiple courses when taken together is relatively uncommon. The confidence metric, whose approximate values range from 0.235 to 0.460, represents the conditional probability of failing the consequent course, given failure in the antecedent course.

The lift metric provides a complementary perspective on association strength. Lift values exceeding 1 (ranging from 1.261 to 3.080) confirm positive association. Zhang's metric also offers useful insight into the performance of the association rules. For the top seven rules identified, the values of Zhang's metric range from 0.711 to 0.243, indicating positive association within each course combination.

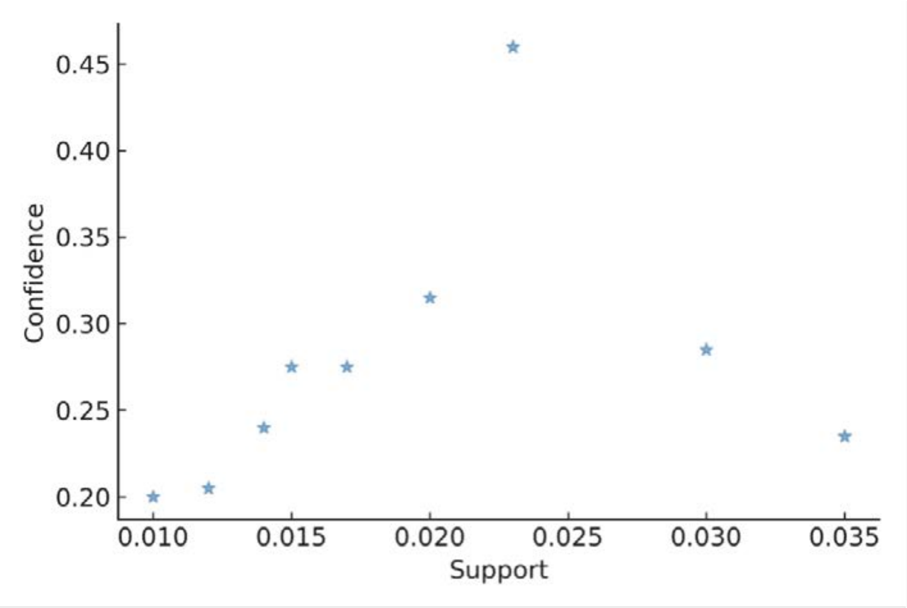
- Rule 1 (BIO 1404 and MAT 1073) is obtained under a high degree of confidence and lift. It signifies that those students who struggled in BIO 1404 also struggled in MAT 1073.
- Rule 2 (AIS 1203 and WRC 1013) also has high degree of confidence and lift. It illustrates that students who struggled in AIS 1203 also had difficulty in WRC 1013.
- Rule 3 (CHE 1073 and MAT 1073) is selected because it has a high degree of support. It shows that these two courses were taken together frequently; this rule is applicable to many future course combinations.
- Rules 4 and 5 (HIS 1043/HIS 1053 and WRC 1013) are based on a high degree of confidence and support. Those rules indicate association of student performance in HIS 1043/HIS 1053 and student performance in WRC 1013.
- Rules 6 and 7 (MAT 1053/MAT 1073 and WRC 1013) are obtained based on a high degree of confidence and support. Those rules show that students who struggled in MAT 1053/MAT 1073 also struggled in WRC 1013.

The analysis reveals critical patterns: mathematics courses (MAT 1053/MAT 1073) appear in four of the seven high-risk combinations, while the writing course (WRC 1013) appears in five combinations. This frequency suggests that targeted intervention strategies should prioritize mathematics and writing support, since these subjects frequently co-occur with failures in other disciplines. These cross-disciplinary failure patterns would remain undetected through individual course analysis.

A Closer Look at the Performance Metrics of the Association Rules for Course Failure Patterns

The scatter plot in Figure 5 illustrates a fundamental tradeoff that exists in association rule mining. Confidence and support are two primary metrics used to evaluate the strength and relevance of association rules; their relationship reveals important insights about the underlying data patterns.

Figure 5. Plot of Confidence and Support Levels of Association Rules



Looking specifically at Figure 5, we can observe several key characteristics.

- **Distribution pattern:** The points are scattered across different support and confidence values, with no rules appearing in the upper-right quadrant (high support and high confidence). Instead, we see a cluster of points with approximate support values ranging from 0.010 to 0.035 and approximate confidence values ranging from 0.200 and 0.460.
- **Highest confidence rule:** The rule with the highest confidence (approximately 0.460) has a moderate support value of about 0.023. This suggests that, while this rule is highly reliable when it applies, it affects only a moderate number of students.
- **Highest support rules:** The rules with the highest support values (around 0.030 to 0.035) have relatively lower confidence values (around 0.235), indicating these combinations occur more frequently but are less reliable as predictors.
- **Moderate confidence–support balance:** Several rules fall in the middle range, with confidence values between 0.240 and 0.320 and support values between 0.015 and 0.020, representing potentially valuable insights that balance frequency and reliability.

The absence of rules in the upper-right quadrant of the plot reflects a natural constraint in most real-world datasets, particularly in educational contexts. This absence of rules occurs for several interconnected reasons.

- **Statistical dilution effect:** Courses with high support (frequently failed courses) naturally appear across many different student records and combinations. This widespread distribution means they co-occur with many different courses, and not just with specific ones. Consequently, the conditional probability (confidence) of one high-support course leading to another specific course tends to decrease.
- **Specificity versus generality tradeoff:** Association rules that capture highly specific relationships (high confidence) typically involve more-distinctive or more-specialized course combinations that fewer students take, resulting in lower support. Conversely, rules with high support often capture patterns that are more general, and that apply to many students but are less predictive of specific outcomes.
- **Course selection diversity:** Students follow diverse academic paths based on their majors, interests, and requirements. This diversity naturally limits the number of students taking identical course combinations, creating a ceiling effect on support values for highly predictive rules.

In educational data mining specifically, the above analysis of the performance metrics suggests that the most actionable insights often come from rules with moderate support and confidence, representing course combinations that both affect several students and demonstrate reliable predictive power. For academic advisors and curriculum designers, these middle-ground rules often identify the most promising targets for intervention, since they balance relevance (affecting enough students to matter) with predictive accuracy (reliably identifying problematic course combinations). As explained at the beginning of Results and Discussion, the minimum support and minimum confidence are set with moderate values since there are many individual courses (373 courses in total) that students can choose from, and a significant number of possible combinations that can be made of these courses. Additionally, choosing the appropriate support and confidence values requires domain knowledge, understanding of the data at hand, experimentation, and iterative refinement to find values that yield actionable insights.

In the next phase of the analysis, we analyze the relationships between the course combinations identified above with student retention rates. Specifically, we will examine whether there are statistically significant relationships with first-term retention and first-year retention between two groups of students: those taking only one course from a course combination compared with those taking both courses from the same combination.

Chi-square tests of independence are conducted for the seven course combinations discussed previously, and all show significant relationships between two groups of students and retention rates. The results of combinations BIO 1404–MAT 1073, CHE 1073–MAT 1073, and HIS 1053–WRC 1013 are presented next.

BIO 1404–MAT 1073

First-Term Retention

From Table 3 we see that 1,023 students take either BIO 1404 or MAT 1073, and 174 students take both courses. The first-term retention rates of the two groups are 83.4% and 81.0%, respectively.

Table 3. BIO 1404–MAT 1073, First-Term Retention

BIO 1404–MAT 1073	Retained after first term		Total	% Retained
	No	Yes		
Taking either course	170	853	1,023	83.4%
Taking both courses	33	141	174	81.0%

A chi-square test is performed to examine the relationship between the two groups of students in Table 3 and first-term retention rates. There is significant relationship between the two variables, with X^2 (1, N = 1,197) = 5.0, p < .05. Students taking both courses are less likely to retain after the first term than are students taking either course.

First-Year Retention

First-year retention rates of the two groups of students are 55.7% and 54.0%, respectively (Table 4).

Table 4. BIO 1404–MAT 1073, First-Year Retention

BIO 1404–MAT 1073	Retained after first year		Total	% Retained
	No	Yes		
Taking either course	453	570	1,023	55.7%
Taking both courses	80	94	174	54.0%

A chi-square test is performed to examine the relationship between the two groups of students in Table 4 and first-year retention rates. There is significant relationship between the two variables, with X^2 (1, N = 1,197) = 156.5, p < .001. Students taking both courses are less likely to retain after the first year than are students taking either course.

CHE 1073–MAT 1073

First-Year Retention

Table 5 shows that there are 1,452 students taking either CHE 1073 or MAT 1073, and that 222 students are taking both courses. The retention rates of the two groups are 86.5% and 71.6%, respectively.

Table 5. CHE 1073–MAT 1073, First-Term Retention

CHE 1073–MAT 1073	Retained after first term		Total	% Retained
	No	Yes		
Taking either course	196	1,256	1,452	86.5%
Taking both courses	63	159	222	71.6%

A chi-square test is performed to examine the relationship between the two groups of students in Table 5 and first-term retention rates. There is significant relationship between the two variables, with X^2 (1, ***N*** = 1,674) = 32.6, *p* < .001. Students taking both courses are less likely to retain after the first term than are students taking either course.

First-Year Retention

First-year retention rates of the two groups of students are 59.6% and 39.6%, respectively (Table 6).

Table 6. CHE 1073–MAT 1073, First-Year Retention

CHE 1073–MAT 1073	Retained after first year		Total	% Retained
	No	Yes		
Taking either course	586	866	1,452	59.6%
Taking both courses	134	88	222	39.6%

A chi-square test is performed to examine the relationship between the two groups of students in Table 6 and first-year retention rates. There is significant relationship between the two variables, with X^2 (1, *N* = 1,674) = 31.4, *p* < .001. Students taking both courses are less likely to retain after the first year than are students taking either course.

HIS 1053–WRC 1013

First-Year Retention

There are 1,597 students taking either HIS 1053 or WRC 1013, and 127 students taking both courses. The retention rates of the two groups are 83.0% and 69.3%, respectively (Table 7).

Table 7. HIS 1053–WRC 1013, First-Term Retention

HIS 1053–WRC 1013	Retained after first term		Total	% Retained
	No	Yes		
Taking either course	272	1,325	1,597	83.0%
Taking both courses	39	88	127	69.3%

A chi-square test is performed to examine the relationship between the two groups of students in Table 7 and first-term retention rates. There is significant relationship between the two variables, with X^2 (1, *N* = 1,724) = 18.5, *p* < .001. Students taking both courses are less likely to retain after the first term than are students taking either course.

First-Year Retention

First-year retention rates of the two groups of students are 54.6% and 34.6%, respectively (Table 8).

Table 8. HIS 1053–WRC 1013, First-Year Retention

HIS 1053–WRC 1013	Retained after first year		Total	% Retained
	No	Yes		
Taking either course	725	872	1,597	54.6%
Taking both courses	83	44	127	34.6%

A chi-square test is performed to examine the relationship between the two groups of students in Table 8 and first-year retention rates. There is significant relationship between the two variables, with X^2 (1, *N* = 1, 724) = 262.9, *p* < .001. Students taking both courses are less likely to retain after the first year than are students taking either course.

The bivariate analyses above show that the seven course combinations have statistically significant relationship with retention rates. Students taking both courses in a course combination are less likely to retain after the first term and after the first year than are students taking only one course from the same course combination.

SUMMARY AND IMPLICATIONS

The findings from this study provide valuable insights into the course combinations that are negatively correlated with retention of first-time, full-time freshman students. We identified seven pairs of courses that, when taken together, significantly increase the likelihood of student attrition. This information is crucial for university administrators, academic advisors, and curriculum planners aiming to improve student retention rates. Such insights would not have been available if courses were analyzed individually rather than in combination.

The identification of the high-risk course combinations above allows for targeted interventions. Academic advisors can use this information to guide first-time, full-time freshman students in selecting their courses more strategically. For example, advisors might recommend against taking CHE 1073 and MAT 1073 in the same semester, or they might suggest additional support resources for students enrolled in these courses.

These findings also have implications for curriculum design and institutional policy. Universities could consider restructuring the timing and prerequisites of high-risk courses to reduce the likelihood of students encountering these problematic combinations. Additionally, supplemental instruction or tutoring programs could be developed specifically for the identified high-risk course pairs.

The analysis of high-risk course combinations also offers practical financial implications to consider. Failing a course would mean an additional semester or year enrolled, hence increasing financial responsibilities for students, their families, and the government. Being able to identify high-risk course combinations would help university administrators design strategies to support students taking these courses and help relieve the financial burden of the parties involved.

The application of MBA has proven to be a valuable tool in identifying course combinations that are associated with high rates of failure. Understanding these associations provides colleges and universities with insights to develop effective strategies that support student success. From that perspective, this study highlights the importance of data-driven approaches in higher education and sets the stage for further research in this area.

LIMITATIONS OF THE STUDY AND FUTURE RESEARCH

Limitations of the Study

The study acknowledges limitations on data generalizability and temporal changes. On the one hand, the findings in this analysis are based on data from a single institution (UTSA) and might not be readily generalizable to other contexts. On the other hand, the data are from Fall 2018–Fall 2022 cohorts. Therefore, changes in curriculum and academic policies after the study period could affect subsequent analyses.

An additional limitation involves the exclusion of demographic variables (race/ethnicity, gender, age groups) from the current analysis due to the study’s broad scope. Future research could apply MBA within specific demographic subgroups to identify population-specific patterns, as demonstrated by Çiçekli and Kabasakal (2021).

MBA provides valuable pattern identification but has inherent limitations. Association rules identify correlations rather than causal relationships, requiring careful interpretation when developing interventions. Additionally, the algorithm's multiple database scans demand substantial computational resources, particularly with large datasets containing numerous frequent patterns.

Future Research

Future research should aim to replicate this study across multiple institutions to validate the results. Additionally, future studies could explore the underlying reasons why these specific course combinations lead to higher failure rates, such as course content difficulty, teaching methods, or student preparedness. Another study that can be considered to follow up from this one is to explore how these course combinations could be associated with student attrition. This can be done by treating student attrition as the response variable and course combinations among the predictors.

REFERENCES

Alangari, N., & Alturki, R. (2020). Association rule mining in higher education: A case study of computer science students. In R. Mehmood, S. See, I. Katib, & I. Chlamtac (Eds.), *Smart infrastructure and applications* (pp. 311–328). EAI/Springer Innovations in Communication and Computing. https://doi.org/10.1007/978-3-030-13705-2_13

Bautista, R. M. (2005). Clustering failed courses of engineering students using association rule mining. *Journal of Theoretical and Applied Information Technology, 96*(4), 875–886. https://www.jatit.org/volumes/Vol96No4/3Vol96No4.pdf

Çiçekli, U. G., & Kabasakal, I. (2021). Market basket analysis of basket data with demographics: A case study in e-retailing. *Alphanumeric Journal, 9*(1), 1–12. https://doi.org/10.17093/alphanumeric.752505

Colorado State University Office of Institutional Research, Planning and Effectiveness. (2012). *Unsuccessful course completion and student success.* https://irpe-reports.colostate.edu/pdf/ResearchBriefs/Unsuccessful_Courses_Student_Success.pdf

Derouiche, M. (2024). *Association rule mining for MBA.* https://www.kaggle.com/code/mohammedderouiche/association-rule-mining-for-mba

Ezarik, M. (2023). *Defining, measuring and unifying approaches to student success.* https://www.insidehighered.com/news/student-success/life-after-college/2023/11/22/how-college-campus-professionals-speak-about

González, J., Romero, C., & Ventura, S. (2008). Data mining in course management systems: Moodle case study and tutorial. *Computers & Education, 51*(1), 368–384. https://doi.org/10.1016/j.compedu.2007.05.016

McKinsey & Company. (2023). *Strengthening higher education outcomes through partnerships, alliances, and mergers.* https://www.mckinsey.com/industries/education/our-insights/higher-ed-is-consolidating-transforming-the-sector

Michaels, K., & Milner, J. (2021). *Powered by publics learning memo: The big ten academic alliance cluster exploring foundational course DFW rates, equity gaps, and progress to degree.* https://www.aplu.org/wp-content/uploads/powered-by-publics-learning-memo-the-big-ten-academic-alliance-cluster.pdf

Mwiti, D. (2025). *Apriori algorithm explained: A step-by-step guide with Python implementation.* https://www.datacamp.com/tutorial/apriori-algorithm

Papadogiannis, I., Wallace, M., & Karountzou, G. (2024). Educational data mining: A foundational overview. *Encyclopedia* 2024, *4*(4), 1644–1664. https://doi.org/10.3390/encyclopedia4040108

Romero, C., & Ventura, S. (2020). Educational data mining and learning analytics: An updated survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 10*(3), e1355. https://doi.org/10.1002/widm.1355

Safour, H., Essgaer, M., Alshareef, A. (2024). Unraveling academic failure: Examining the influence of course correlations on student performance through association rules algorithms. *2024 International Conference on Computer and Applications (ICCA)*, Cairo, Egypt. https://doi.org/10.1109/ICCA62237.2024.10927837

Slim, A., Heileman, G., Al-Doroubi, W., & Abdallah, C. (2016). The impact of course enrollment sequences on student success. *2016 IEEE 30th International Conference on Advanced Information Networking and Applications (AINA)*, Crans-Montana, Switzerland. https://doi.org/10.1109/AINA.2016.140

Soika, B. (2021). *What is student success? New insight into a complex question.* https://rossier.usc.edu/news-insights/news/what-student-success-new-insight-complex-question

Zong, C., & Koller, S. (2023). *The relationship between high-risk courses and Fall-to-Fall retention of first-time full-time students at UW* [University of Wyoming]. https://www.uwyo.edu/oia/_files/research-reports/high-risk-course-research-report-oct-2023.pdf