

TC3283 DATA MINING

PROJECT 1

ASSOCIATION RULES MINING

NAME	NO MATRIC
MUHAMMAD AJRUL AMIN BIN MOHD ZAIDI	A194789
MUHAMMAD IZZUL ISLAM BIN FASAL	A200363
VILAASINI A/P KUMAR	A195632

1.0 INTRODUCTION

This project uses Association Rule Mining (ARM) to explore patterns and extract meaningful insights from the Malaysia University Enrollment 2022-2023 dataset, sourced from Kaggle (*Malaysia Public University Enrolment 2022-2023*, 2024). The dataset includes a wide range of information about student enrollment across Malaysian universities such as university name, academic year, gender, degree type (e.g., Diploma, Bachelor's, Master's, PhD), institution category (public or private), and enrollment figures. This dataset acts as a valuable tool for understanding enrollment behaviors and identifying common trends within the Malaysian higher education system.

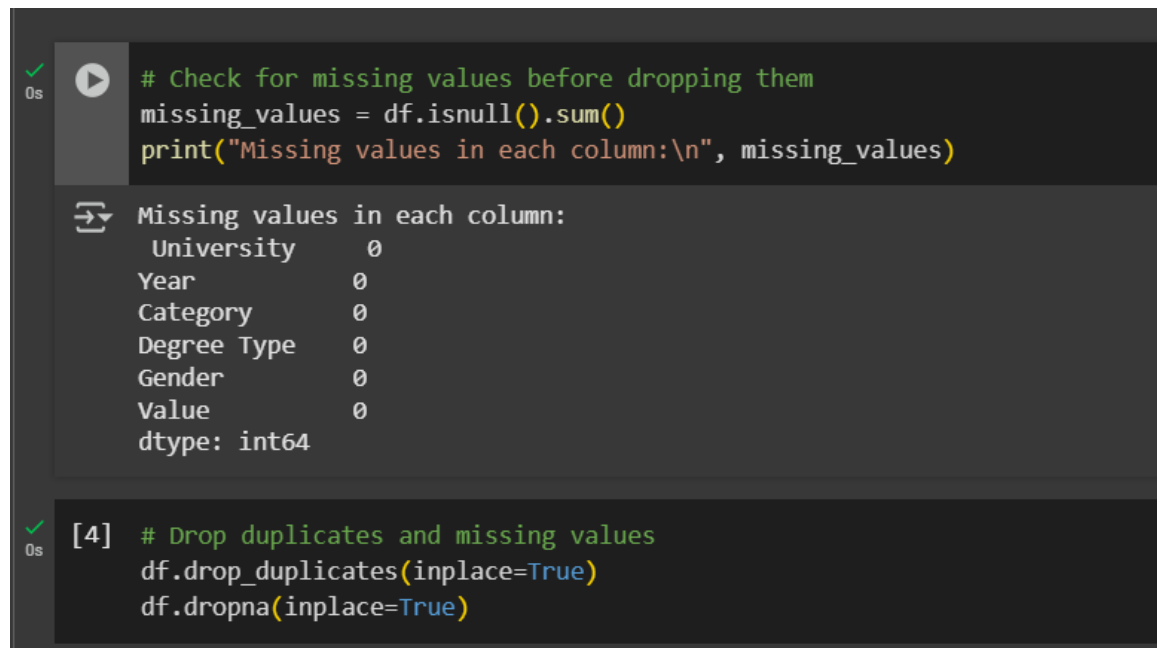
Association Rule Mining (ARM) is a widely used data mining technique that discovers hidden patterns, correlations, or associations among categorical variables in large datasets. It is particularly effective in uncovering frequent itemset such as combinations of attributes that occur together and generating rules that express relationships between these items. ARM relies on three core evaluation metrics: (i) support, which measures how often an itemset appears in the dataset; (ii) confidence, which indicates how often the rule holds true; and (iii) lift, which shows how much more likely the consequent is to occur when the antecedent is present compared to random chance. These metrics help determine the strength and relevance of each discovered rule (Herath, 2024).

In this project, ARM is applied to educational data to uncover co-occurrence patterns such as gender preferences in degree programs, trends in public vs. private university enrollments, and shifts in program popularity across academic years. Such informations are valuable for policymakers, university administrators, and educational planners, as they support data-driven decision-making in areas like academic program design, enrollment forecasting, and targeted outreach strategies. By applying ARM to this dataset, the project aims to show actionable associations that may not be immediately apparent through traditional analysis. This not only deepens our understanding of student behavior in Malaysian universities but also represents the strategy and strength of ARM in extracting knowledge from large-scale categorical data.

2.0 METHODOLOGY

I. Data Collection & Preprocessing

The process began with data collection, utilizing a publicly available dataset titled Malaysia University Enrollment 2022–2023 which was downloaded in CSV format from Kaggle. This dataset consist of categorical attributes such as university name, academic year, gender, degree type, and institution category, followed by a numeric column labeled "Value" representing the number of students for each unique record combination. Before conducting any analysis, it was essential to ensure data quality. Therefore, data cleaning steps were undertaken, which included the removal of missing values and duplicate entries to avoid bias and redundancy in the association rules. Figure 1 shows the code snippet to check and drop duplicate records from the dataset.



```
# Check for missing values before dropping them
missing_values = df.isnull().sum()
print("Missing values in each column:\n", missing_values)
```

Missing values in each column:

University	0
Year	0
Category	0
Degree Type	0
Gender	0
Value	0
dtype:	int64

```
[4] # Drop duplicates and missing values
df.drop_duplicates(inplace=True)
df.dropna(inplace=True)
```

Figure 1: Code Snippet for Identifying and Dropping Duplicate and Missing Values from the Dataset

Next, the dataset was examined to identify categorical features appropriate for transaction-based analysis. The unique value counts for attributes such as university, category, degree type, and gender were reviewed to assess their suitability. Figure 2 shows the code snippet to check the categorical data.

```
✓ 0s # Check for categorical data (potentially suitable for transactions)
print(df.select_dtypes(include='object').nunique())

↕ University      20
    Category       3
    Degree Type    8
    Gender         3
    dtype: int64
```

Figure 2: Code Snippet to Overview the Categorical Features and Their Uniqueness in the Dataset

Following data cleaning, the next step involved converting the dataset into a format suitable for Association Rule Mining (ARM). Specifically, each row was transformed into a transaction representing a unique student enrollment profile, combining multiple categorical attributes such as university, gender, and degree type. These transactions were then expanded based on the count in the “Value” column to experiment student-level data. This process is showed in Figure 3, which contains the code that performs the transformation and expansion. The expanded dataset resulted in over 3.7 million transactions, met the minimum requirement of 1,500 records necessary for meaningful ARM analysis. Figure 3 shows the code snippet to create a list of transaction based on “Value”.

```
✓ 0s [7] # Choose features to include in each "transaction"
df['Transaction'] = df[['University', 'Year', 'Category', 'Degree Type', 'Gender']].astype(str).agg(', '.join, axis=1)

# Create list of transactions based on 'Value'
expanded_transactions = []
for _, row in df.iterrows():
    items = row['Transaction'].split(', ')
    count = int(row['Value'])
    expanded_transactions.extend([items] * count)

print(f"\nTotal transactions generated: {len(expanded_transactions)}")

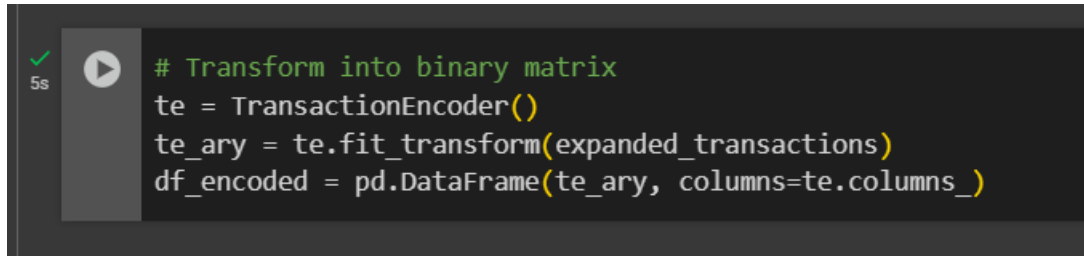
↕ Total transactions generated: 3728320

✓ 0s [8] # Check for minimum 1500 transactions
if len(expanded_transactions) > 1500:
    print("Dataset meets the minimum transaction requirement (1500).")

↕ Dataset meets the minimum transaction requirement (1500).
```

Figure 3: Transaction Transformation and Expansion Based on Student Count

Furthermore, to prepare the data for the FP-Growth algorithm, one-hot encoding was applied using the TransactionEncoder from the mlxtend library. This encoding converted the list of transactions into a binary matrix, where each column represented a unique item (e.g., “Public”, “Bachelor”, “Female”) and each row represented a transaction with binary values indicating the presence (1) or absence (0) of that item. Figure 4 shows the code snippet for one-hot encoding step and transformation into a structured dataframe.

The image shows a code editor window with a dark background. On the left, there is a green checkmark and a play button icon, with '5s' below them. The code is written in a light green font and consists of four lines: a comment, an initialization of TransactionEncoder, a fit_transform call, and a DataFrame creation.

```
# Transform into binary matrix
te = TransactionEncoder()
te_ary = te.fit_transform(expanded_transactions)
df_encoded = pd.DataFrame(te_ary, columns=te.columns_)
```

Figure 4: One-Hot Encoding of Transactions Using TransactionEncoder

II. Algorithm Selection

In this project, the FP-Growth algorithm was chosen, a widely adopted technique in Association Rule Mining (ARM), to uncover hidden patterns and associations within the dataset. Unlike the Apriori algorithm, FP-Growth does not generate candidate itemsets explicitly, making it significantly more efficient when working with large datasets (GeeksforGeeks, 2025). It constructs a compact data structure called the FP-tree and extracts frequent itemsets directly from it (Zeng et al., 2015). This makes it an ideal choice for our dataset, as we were interested in identifying associations between different categorical attributes such as university, gender, and degree type.

To build the algorithm for the analysis, a minimum support threshold of 0.05 was set, meaning that an itemset must appear in at least 5% of the transactions to be considered frequent. This ensured that only the most common associations were focused on, filtering out less relevant patterns. Additionally, a minimum confidence threshold of 0.5 was defined, which implied that only the rules with at least a 50% probability of being valid would be generated. The FP-Growth algorithm was implemented using the mlxtend library, which provided an efficient and straightforward implementation for the needs. Figure 5

shows the code snippet to generate the top frequent itemsets using the FP-Growth algorithm and the resulting association rules.

```
✓ [29] # Generate association rules  
Ds    rules = association_rules(frequent_itemsets, metric="confidence", min_threshold=0.5)  
  
      # Sort rules by lift  
      rules_sorted = rules.sort_values(by="lift", ascending=False)
```

Figure 5: Frequent Itemset Mining and Association Rule Generation

III. Association Rule Generation

Once frequent itemsets were identified through the FP-Growth algorithm, association rules were generated based on these itemsets. The rules were created by pairing antecedents (items on the left-hand side of the rule) with consequents (items on the right-hand side of the rule). The association rules function from the mlxtend library was used to generate the rules, with each rule being evaluated using three key metrics such as support, confidence, and lift.

Support was used to measure the frequency of an itemset in relation to the top 10 rules. Confidence provided the probability that the consequent would appear if the antecedent was present. The metric of lift was used to measure the strength of the association by comparing the observed support to the expected support if the two itemsets were independent. Lift values greater than 1 indicated a positive relationship, while values less than 1 suggested weaker or negative associations. Rules were filtered to retain only those with significant relationships, based on these evaluation metrics.

Figure 6 shows the top 10 association rules generated from the frequent itemsets using FP-Growth, sorted by default order. Each rule is presented with its antecedents, consequents, and key evaluation metrics: support, confidence, and lift.

```
[ ] # Show top 10 rules
print(rules[['antecedents', 'consequents', 'support', 'confidence', 'lift']].head(10))
```

Figure 6: Code Snippet for Displaying Top 10 Association Rules with Key Evaluation Metrics

IV. Evaluation and Interpretation

After the generation of association rules, their strength was primarily evaluated using lift values, as lift provides an effective measure of the significance of the relationships. Lift values greater than 1 indicated strong positive associations, suggesting that the two items co-occurred more often than expected by chance. A lift value below 1 indicated weaker or less frequent co-occurrence, suggesting a lack of meaningful association.

The generated rules were categorized into three groups based on their lift values. Rules with a lift greater than 1.5 were classified as strong associations, indicating a statistically significant relationship. Rules with a lift between 1.0 and 1.5 were considered moderate associations, indicating a reasonable connection, though with some uncertainty. Rules with a lift below 1.0 were classified as weak or irrelevant associations, suggesting that the relationship between the antecedent and consequent was either weak or not meaningful.

Figure 7 shows the classification of generated association rules based on their lift values. The strength of each rule is interpreted as strong, moderate, or weak.

```

# Collect filtered rule data into a list
rule_data = []

for index, row in top_rules_eval.iterrows():
    lift = row['lift']

    if lift <= 1.00:
        continue

    rule = f"{'', '.join(list(row['antecedents']))} → {'', '.join(list(row['consequents']))}"

    if lift > 1.5:
        strength = "Strong Positive Association"
    elif lift > 1.2:
        strength = "Moderate Positive Association"
    elif 1.05 < lift <= 1.2:
        strength = "Weak Positive Association"
    else:
        strength = "Irrelevant Association"

    rule_data.append([rule, round(lift, 2), strength])

```

Figure 7: Code Snippet for Interpreting Association Rule Strength based on Lift

To help in the interpretation of the results, visual tools, such as bar plots and scatter plots, were employed to display the distribution of support, confidence, and lift values. These visualizations helped to identify the most significant rules. Additionally, network graphs were created to represent the relationships between items for the interpretation of the associations found in the dataset. Through these evaluation and interpretation steps, valuable insights into student enrollment patterns and behaviors were uncovered, which could be used to inform decision-making and policy development in the field of higher education.

3.0 RESULTS

I. Frequent Item Sets and Association Rules

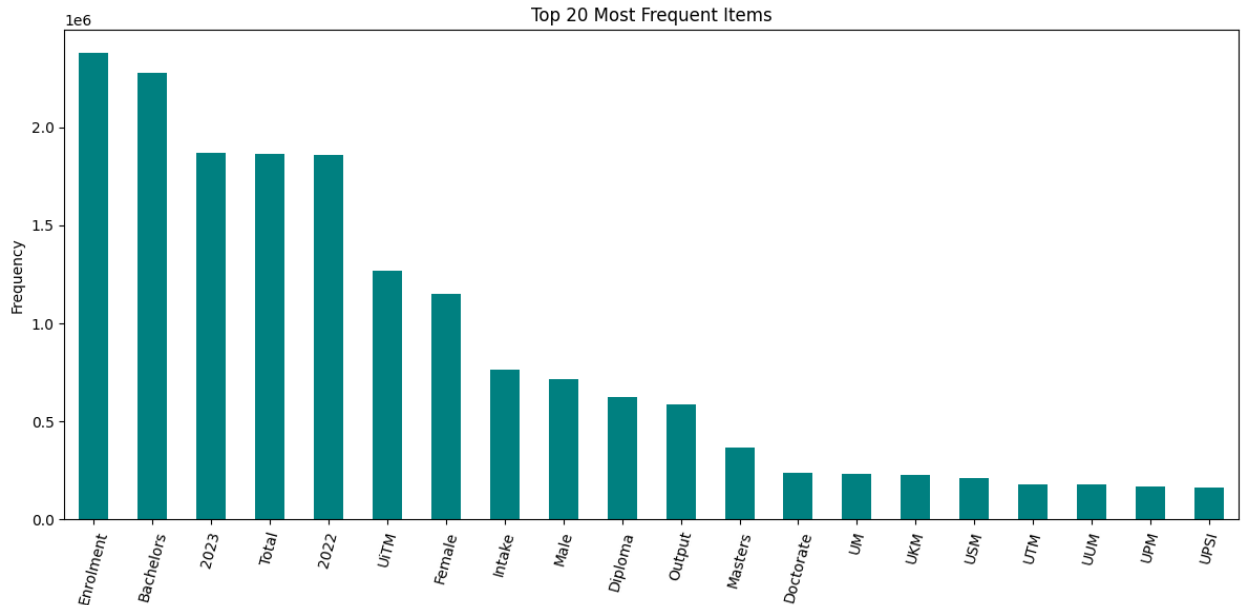


Figure 8: Top 20 Most Frequent Items

Figure 8 shows a bar chart titled "Top 20 Most Frequent Items" illustrates the most commonly occurring terms within the dataset, which appears to be related to higher education statistics in Malaysia. The horizontal axis displays the top 20 items, while the vertical axis represents their corresponding frequency counts. Among the most frequent terms are "Enrolment", "Bachelors", "2023", "Total", and "2022", each appearing over a million times. These terms suggest that the dataset focuses heavily on student enrollment numbers, academic levels, and yearly statistics. Additionally, gender-related terms such as "Female" and "Male" are also prominent, indicating the inclusion of demographic data. Various education levels like "Diploma", "Masters", and "Doctorate" are frequently mentioned, reflecting the academic qualifications being analyzed. Furthermore, the presence of university names such as UiTM, UM, UKM, USM, UPM, and others indicates that the dataset covers multiple higher education institutions across Malaysia. Overall, this

chart provides a clear overview of the most prevalent terms in the dataset, highlighting its focus on educational enrollment and institutional data.

Frequent Itemsets Found:		
	support	itemsets
0	0.501618	(2023)
1	0.204888	(Intake)
2	0.191434	(Male)
3	0.062399	(UM)
4	0.498382	(2022)

Figure 9: Frequent Item sets Found

	antecedents	consequents	support	confidence	lift
0	(2023)	(Enrolment)	0.318160	0.634267	0.994704
1	(Bachelors)	(2023)	0.306172	0.500932	0.998633
2	(2023)	(Bachelors)	0.306172	0.610369	0.998633
3	(Bachelors, 2023)	(Enrolment)	0.201648	0.658609	1.032879
4	(Enrolment, 2023)	(Bachelors)	0.201648	0.633794	1.036958
5	(Intake)	(2023)	0.103547	0.505386	1.007511
6	(Intake)	(Total)	0.102444	0.500000	1.000000
7	(Intake)	(Bachelors)	0.109493	0.534406	0.874347
8	(Bachelors, Intake)	(2023)	0.055336	0.505382	1.007504
9	(Intake, 2023)	(Bachelors)	0.055336	0.534402	0.874341

Figure 10: Top 10 Rules

The association rules generated from the FP-Growth algorithm provide insightful relationships between frequently occurring item sets in the dataset. As shown in Figure 9, the most frequently occurring item sets include the years 2023 (50.16%) and 2022 (49.84%), followed by terms like Intake, Male, and UM. These item sets reflect the common attributes shared among the dataset’s transactions.

Based on these frequent item sets, the association rules were derived and sorted by lift value to prioritize the strongest correlations. As seen in Figure 10, a notable rule is the association between (2023) and (Enrolment) with a confidence of 63.43% and a lift of 0.99, suggesting that when 2023 appears in a transaction, Enrolment is also likely to appear, albeit

with nearly independent correlation (lift ≈ 1). Similarly, the rule (Bachelors, 2023) \rightarrow Enrolment exhibits a higher lift of 1.03, indicating a slightly stronger positive association, meaning students enrolled in 2023 who are also in Bachelors programs are more likely to have Enrolment tagged in the dataset. Moreover, the rule (Intake, 2023) \rightarrow Bachelors with a confidence of 53.44% and a lift of 0.87 shows a moderate association, though the lift value below 1 implies a negative correlation or weaker dependency.

Overall, the rules with lift values above 1 indicate a meaningful association, which can be used for strategic decision-making such as tailoring academic program offerings or enrollment campaigns based on student intake patterns. These rules align with the principles of market basket analysis, where the FP-Growth algorithm effectively discovers associations without candidate generation, making it efficient for large datasets (Han et al., 2000).

II. Evaluation Metrics

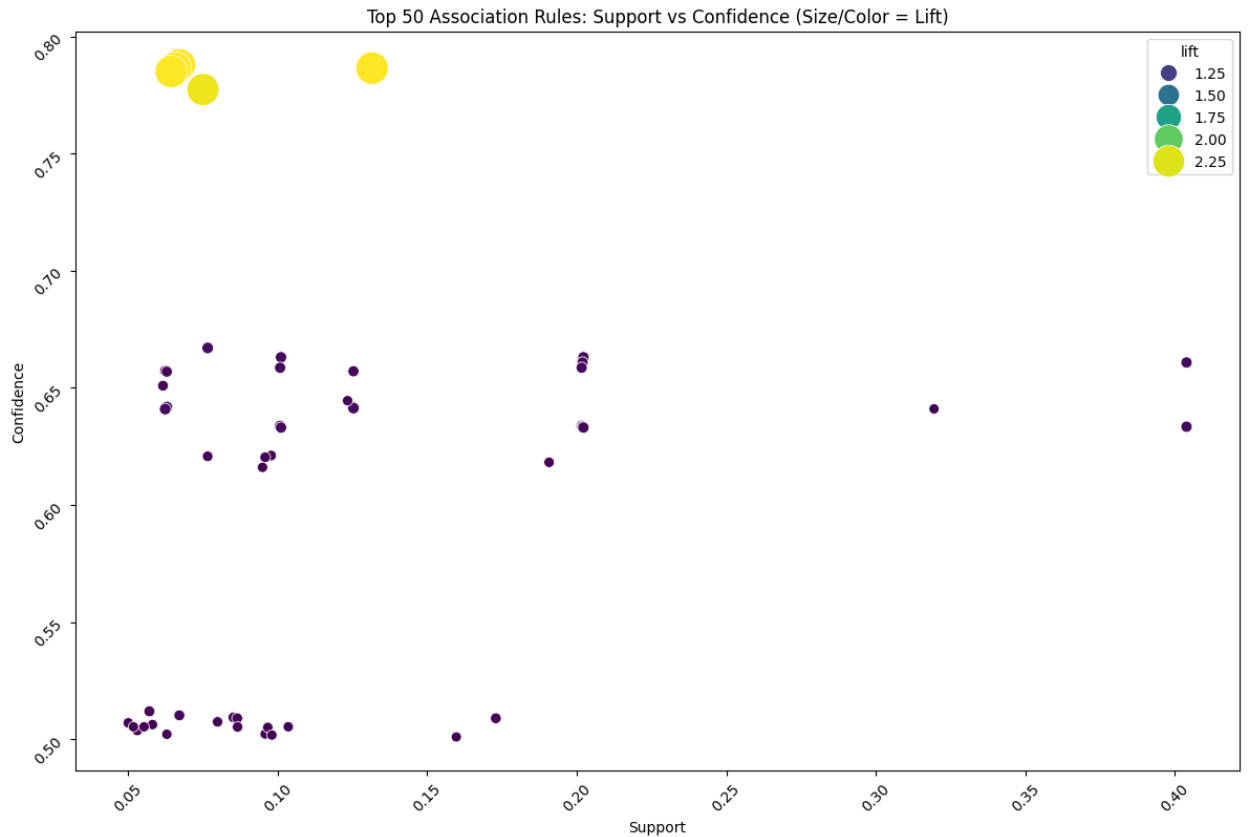


Figure 11: Scatter Plot for Top 50 Association Rules: Support VS Confidence

The scatter plot presented in Figure 11 illustrates the top 50 association rules, visualized based on their support and confidence values, with both color and size representing the lift of each rule. This plot helps to identify the strongest and most relevant rules derived from the dataset.

On the x-axis, we see the support of the rules, indicating how frequently the itemsets occur in the dataset. On the y-axis, the confidence of the rules is shown, representing the likelihood that the consequent occurs when the antecedent is present. Most of the rules cluster in the lower support range (between 0.05 and 0.15), with a few extending up to 0.40, indicating that only a few rules are highly frequent.

Notably, several points appear near the top-left corner, showing high confidence (above 0.75) but low support, which typically suggests strong but less common associations. These points are colored bright yellow and have larger sizes, indicating a lift greater than 2.0. This means these rules are not only confident but also significantly more likely to occur than by random chance, highlighting their potential importance.

The clustering of smaller, darker points near the middle of the graph indicates rules with moderate confidence and support but relatively low lift values (close to 1.0), suggesting these associations are close to being statistically independent and may not be particularly actionable.

Overall, Figure 11 supports the earlier findings by visualizing that while some rules (e.g., those involving "Bachelors", "2023", and "Enrolment") are both frequent and confident, only a few achieve high lift, underscoring their strength and potential for decision-making. This type of visualization is essential for prioritizing association rules in practical applications such as academic planning, enrollment targeting, or curriculum design.

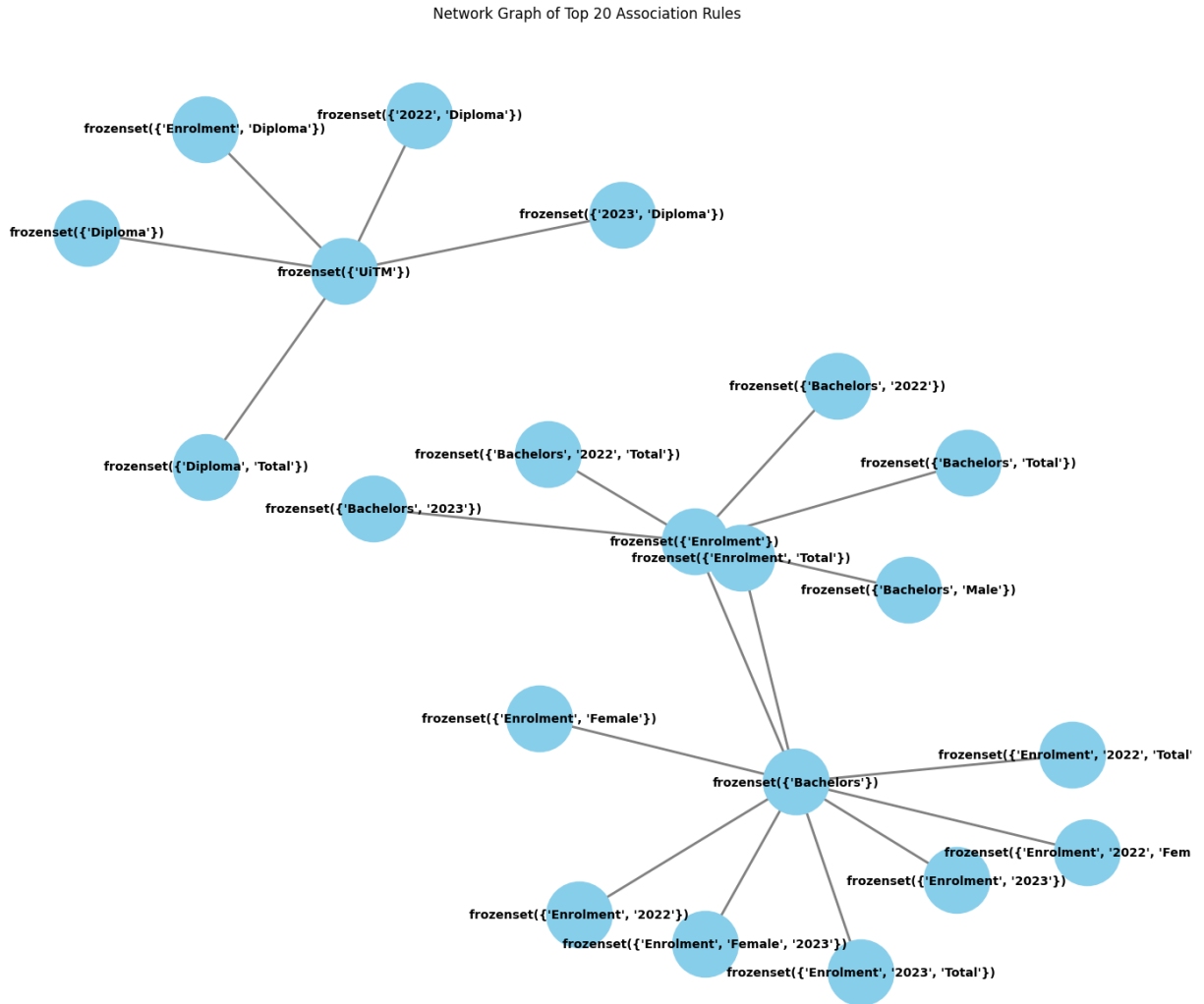


Figure 12: Network Graph of Top 20 Association Rules

Based on the network graph presented in Figure 12, the results clearly illustrate the strong interrelationships between academic-related attributes such as education level, enrollment year, gender, and institution. This network visualization helps to interpret how frequently certain combinations appear together in the dataset and which elements act as central connectors in association rules.

One of the most prominent observations from the graph is the central role of the "Bachelors" itemset. It connects with multiple other item sets such as '2022', '2023', 'Total', 'Male', and 'Enrolment'. This indicates that students enrolled in bachelor's programs have frequent

associations with those specific years, total counts, and gender, suggesting that the bachelor's category is a key driver in understanding broader enrollment patterns.

Another noticeable hub in the graph is "UiTM", which is strongly connected to the item sets involving 'Diploma', 'Total', and various years like '2022' and '2023'. This shows that the institution UiTM is frequently associated with diploma programs, hinting at a trend where UiTM sees a significant intake of diploma students across these academic years. "Enrolment" also appears as a highly connected node, linking with attributes like 'Female', '2022', '2023', and 'Total'. This suggests that the enrollment attribute plays a key role in forming significant association rules, potentially being influenced by or influencing various demographic and academic factors.

The clustered structure of the graph indicates that:

- Diploma programs tend to form their own sub-network, especially around UiTM.
- Bachelors programs dominate another cluster with broader connections across gender and year.
- Enrolment-related rules span across both degree types and gender, showing more diverse associations.

Overall, the network graph reveals that enrollment behaviors are strongly structured around academic level, year, and institution. These associations, represented as connections in the graph, reflect meaningful trends in the dataset—such as common combinations of degree level and year, or gender-based participation—that can be used for strategic academic planning or targeted interventions.

	antecedents	consequents	support	confidence	lift
122	(2022, Diploma)	(UiTM)	0.067135	0.787845	2.318587
124	(Diploma, Total)	(UiTM)	0.065784	0.786460	2.314510
117	(Diploma)	(UiTM)	0.131568	0.786460	2.314510
120	(2023, Diploma)	(UiTM)	0.064433	0.785022	2.310277
126	(Enrolment, Diploma)	(UiTM)	0.075053	0.777330	2.287640
88	(Enrolment, 2022, Female)	(Bachelors)	0.062951	0.641832	1.050110
93	(Enrolment, Female)	(Bachelors)	0.125353	0.641400	1.049403
81	(Enrolment, Female, 2023)	(Bachelors)	0.062402	0.640965	1.048691
25	(Bachelors, Male)	(Enrolment)	0.076599	0.667047	1.046112
41	(Bachelors, 2022)	(Enrolment)	0.202255	0.663060	1.039859

Figure 13: Top Association Rules

Figure 13 presents the top 10 association rules derived from the dataset, highlighting strong patterns in student enrolment attributes. The most prominent rules indicate a strong association between diploma-level students and UiTM. Specifically, rules such as (2022, Diploma) → UiTM and (Diploma, Total) → UiTM demonstrate high confidence values (approximately 78%) and lift values exceeding 2.31. These values suggest that diploma students, particularly in 2022 and 2023, are significantly more likely to be enrolled in UiTM than would be expected by chance. The individual rule (Diploma) → UiTM, with a support of 13.15%, confirms that a large proportion of diploma students across the dataset are enrolled at UiTM, emphasizing UiTM's dominant role in diploma-level education.

In contrast, rules involving female enrolment in bachelor's programs—such as (Enrolment, Female, 2023) → Bachelors—show moderate confidence levels around 64% and lower lift values (approximately 1.05). These indicate a consistent, but not particularly strong, relationship between female enrolment and bachelor's program selection. Similarly, rules like (Bachelors, Male) → Enrolment and (Bachelors, 2022) → Enrolment reflect common trends

with confidence values near 66%, but again, the lift values slightly above 1 suggest these are typical patterns rather than exceptional associations.

Overall, the rules in Figure 13 reinforce the insights shown in Figure 11 and Figure 12, where UiTM and diploma programs emerged as central and highly associated elements within the data. The combination of high lift, support, and confidence for these rules indicates that these patterns are both statistically significant and practically meaningful in understanding enrolment trends.

4.0 ANALYSIS AND INTERPRETATION

Rule	Lift	Strength
2022, Diploma → UiTM	2.32	Strong Positive Association
Total, Diploma → UiTM	2.31	Strong Positive Association
Diploma → UiTM	2.31	Strong Positive Association
2023, Diploma → UiTM	2.31	Strong Positive Association
Enrolment, Diploma → UiTM	2.29	Strong Positive Association
Enrolment, 2022, Female → Bachelors	1.05	Weak Positive Association
Enrolment, Female → Bachelors	1.05	Irrelevant Association
Enrolment, 2023, Female → Bachelors	1.05	Irrelevant Association
Bachelors, Male → Enrolment	1.05	Irrelevant Association
2022, Bachelors → Enrolment	1.04	Irrelevant Association
2022, Bachelors, Total → Enrolment	1.04	Irrelevant Association
Enrolment, 2023, Total → Bachelors	1.04	Irrelevant Association
Enrolment, 2023 → Bachelors	1.04	Irrelevant Association
Bachelors → Enrolment	1.04	Irrelevant Association
Enrolment → Bachelors	1.04	Irrelevant Association
Bachelors, Total → Enrolment	1.04	Irrelevant Association
Enrolment, Total → Bachelors	1.04	Irrelevant Association
Enrolment, 2022 → Bachelors	1.04	Irrelevant Association
Enrolment, 2022, Total → Bachelors	1.04	Irrelevant Association
2023, Bachelors → Enrolment	1.03	Irrelevant Association
2023, Bachelors, Total → Enrolment	1.03	Irrelevant Association
2023, Bachelors, Female → Enrolment	1.03	Irrelevant Association
Female, Bachelors → Enrolment	1.03	Irrelevant Association
2022, Bachelors, Female → Enrolment	1.03	Irrelevant Association
Female, UiTM → 2022	1.03	Irrelevant Association
UiTM, Diploma → 2022	1.02	Irrelevant Association
Diploma → 2022	1.02	Irrelevant Association
Total, UiTM → 2022	1.02	Irrelevant Association
UiTM → 2022	1.02	Irrelevant Association
2022, Male → Enrolment	1.02	Irrelevant Association
Output → Bachelors	1.02	Irrelevant Association
Enrolment, Male → Bachelors	1.02	Irrelevant Association
2022, Female → Bachelors	1.02	Irrelevant Association
Bachelors, UiTM → 2022	1.01	Irrelevant Association
Output → 2023	1.01	Irrelevant Association
Female → Bachelors	1.01	Irrelevant Association
Enrolment, Bachelors, UiTM → 2022	1.01	Irrelevant Association
Masters → 2023	1.01	Irrelevant Association
Male → Enrolment	1.01	Irrelevant Association
Bachelors, Male → 2023	1.01	Irrelevant Association
2023, Female → Bachelors	1.01	Irrelevant Association
Female, Bachelors → 2022	1.01	Irrelevant Association
Enrolment, Female, Bachelors → 2022	1.01	Irrelevant Association
Intake, Total → 2023	1.01	Irrelevant Association
Intake → 2023	1.01	Irrelevant Association
Intake, Bachelors → 2023	1.01	Irrelevant Association
Male → 2023	1.01	Irrelevant Association
Enrolment, Female → 2022	1.01	Irrelevant Association
2022 → Enrolment	1.01	Irrelevant Association
Enrolment, Total → 2022	1.01	Irrelevant Association
2022, Total → Enrolment	1.01	Irrelevant Association
Enrolment → 2022	1.01	Irrelevant Association
Enrolment, Bachelors → 2022	1.0	Irrelevant Association
Enrolment, Bachelors, Total → 2022	1.0	Irrelevant Association
Female → 2022	1.0	Irrelevant Association
Enrolment, Total, UiTM → 2022	1.0	Irrelevant Association
Enrolment, UiTM → 2022	1.0	Irrelevant Association
2022 → Bachelors	1.0	Irrelevant Association
2022, Total → Bachelors	1.0	Irrelevant Association
2023, Male → Enrolment	1.0	Irrelevant Association

Figure 14: Positive Association rules

Figure 14 shows that most rules generated from the dataset are weak or irrelevant. Some however, do show interesting results. There are 4 strong associations with Lift values above 1.5. those rules being:

Table 1: Strong Positive Association Rules

Association Rules	Lift Value
2022, Diploma > UiTM	2.32
Diploma, Total > UiTM	2.31
Diploma > UiTM	2.31
2023, Diploma > UiTM	2.31
Enrolment, Diploma > UiTM	2.29

The association rules with positive correlations suggest that students who are interested in enrolling in a Diploma course are highly likely to choose UiTM as their university of choice, especially when compared to other prominent universities such as UKM, UM, and USM. This trend can be attributed to UiTM's extensive network of campuses spread across various states in Malaysia, making it a convenient option for students nationwide. Moreover, UiTM offers one of the most diverse and comprehensive ranges of Diploma programs among public universities in the country. This wide selection of diploma courses allows potential students to choose from a variety of fields, thus enhancing its appeal and accessibility.

In contrast, the other rules with positive associations either indicate weak correlations or are deemed irrelevant to the analysis ($\text{Lift} < 1.05$). These rules do not provide significant insights that would be applicable or valuable in a real-world context, thereby limiting their practical use in decision-making processes.

Rule	Lift	Strength
2023, Male → Bachelors	0.98	Irrelevant Association
Male → Bachelors	0.98	Irrelevant Association
2022, Male → Bachelors	0.98	Irrelevant Association
2023, Bachelors, UiTM → Enrolment	0.97	Irrelevant Association
Bachelors, UiTM → Enrolment	0.96	Irrelevant Association
Bachelors, Total, UiTM → Enrolment	0.96	Irrelevant Association
2022, Bachelors, UiTM → Enrolment	0.96	Irrelevant Association
2023, UiTM → Enrolment	0.94	Weak Negative Association
2023, Total, UiTM → Enrolment	0.94	Weak Negative Association
Masters → Enrolment	0.94	Weak Negative Association
UiTM → Enrolment	0.93	Weak Negative Association
Total, UiTM → Enrolment	0.93	Weak Negative Association
Female, UiTM → Enrolment	0.93	Weak Negative Association
2022, Total, UiTM → Enrolment	0.91	Weak Negative Association
2022, UiTM → Enrolment	0.91	Weak Negative Association
Diploma → Enrolment	0.91	Weak Negative Association
UiTM, Diploma → Enrolment	0.89	Weak Negative Association
Intake, 2022 → Bachelors	0.87	Weak Negative Association
Intake → Bachelors	0.87	Weak Negative Association
Intake, Total → Bachelors	0.87	Weak Negative Association
Intake, 2023 → Bachelors	0.87	Weak Negative Association
Enrolment, 2022, UiTM → Bachelors	0.86	Weak Negative Association
Enrolment, UiTM → Bachelors	0.86	Weak Negative Association
Enrolment, Total, UiTM → Bachelors	0.86	Weak Negative Association
Enrolment, 2023, UiTM → Bachelors	0.85	Weak Negative Association
Female, UiTM → Bachelors	0.84	Weak Negative Association
2023, UiTM → Bachelors	0.83	Weak Negative Association
UiTM → Bachelors	0.83	Weak Negative Association
Total, UiTM → Bachelors	0.83	Weak Negative Association
2022, UiTM → Bachelors	0.82	Weak Negative Association

Figure 15: Negative Association rules

Figure 15 illustrates the association rules with negative or weak associations (i.e., Lift < 1). These rules generally indicate that the occurrence of one item is less likely to be associated with the other. Upon inspection, none of these rules appear to be of strong analytical interest, as most fall within the “irrelevant” or “weak negative” range. However, the rules with the lowest lift values suggest that students are slightly less inclined to choose UiTM as their university of choice for their bachelor's degree.

This observation may be attributed to the fact that other top universities in Malaysia—such as UKM, USM, UM, and UTM—offer a wide variety of Bachelor's degree programs, making the competition for undergraduate enrollment much stiffer. In contrast, UiTM appears to dominate in Diploma-level offerings, where it provides significantly more options than most other institutions. As a result, UiTM is more strongly associated with Diploma programs than with Bachelor's degree programs.

Overall, while these negative associations are not particularly strong or conclusive, they provide subtle insights into the preference shifts students may exhibit when progressing from Diploma to bachelor's level education.

5.0 CONCLUSION

This project successfully showed the application of Association Rule Mining (ARM) to extract meaningful insights from the Malaysia Public University Enrollment 2022–2023 dataset. Using the FP-Growth algorithm, the study efficiently identified frequent itemsets and generated association rules that reflect real patterns in student enrollment behavior across Malaysian higher education institutions.

Methodological Strengths:

The use of FP-Growth over more traditional methods like Apriori proved to be advantageous due to its computational efficiency and ability to handle a large dataset of over 3.7 million transaction records. The project followed a structured methodology, starting with data cleaning and transformation, followed by one-hot encoding and then frequent pattern mining. The final step involved rule evaluation using metrics such as support, confidence, and lift, which provided a reliable way to gauge the significance and strength of each rule.

Key Findings:

The results showed strong positive associations between Diploma programs and UiTM. For example, rules such as (2022, Diploma) \rightarrow UiTM and (Diploma, Total) \rightarrow UiTM exhibited high confidence ($\sim 78\%$) and lift values (> 2.3), clearly indicating that UiTM is the most common choice for students pursuing diploma-level education. These insights are not merely statistical but reflect the institutional structure of Malaysian higher education—where UiTM's wide geographical coverage and broad diploma program offerings make it a popular and accessible choice.

On the other hand, Bachelor's programs displayed more moderate associations, often with lift values close to 1.0. For example, (Bachelors, 2023) \rightarrow Enrolment showed only a slight positive association, indicating that such combinations are common but not uniquely strong.

This suggests that students pursuing Bachelor's degrees are more evenly distributed among different institutions like UM, USM, UKM, and others, possibly due to broader academic choices and specialized offerings.

Interpretations and Implications:

The visualizations such as bar charts, scatter plots, and network graphs provided intuitive interpretations of the relationships among gender, academic level, institution, and enrollment year. Notably, the network graph highlighted how UiTM and the "Diploma" category serve as central nodes, further reinforcing their importance in enrollment trends.

From a practical standpoint, the discovered rules can inform various stakeholders:

University administrators can better understand student preferences to tailor academic offerings.

Policymakers can allocate resources or create policies to balance enrollment across institutions and programs.

Educational planners can develop targeted outreach strategies for underrepresented groups or program types.

Final Reflection:

In conclusion, this project showcases the power of ARM—especially the FP-Growth algorithm—in uncovering non-obvious and actionable insights from large-scale categorical data. It bridges the gap between raw data and strategic decision-making in the higher education sector. The clear and interpretable rules generated can be used to support data-driven planning, helping to align institutional offerings with student demands and national education goals.

6.0 REFERENCES

- GeeksforGeeks. (2025, April 5). *Frequent Pattern Growth Algorithm*. GeeksforGeeks. <https://www.geeksforgeeks.org/frequent-pattern-growth-algorithm/>
- Han, J., Pei, J., & Yin, Y. (2000). *Mining frequent patterns without candidate generation: A frequent-pattern tree approach*. Proceedings of the 2000 ACM SIGMOD international conference on Management of data.
- Herath, S. (2024, January 22). Fundamentals of Associate Rule mining - Data Science and Machine Learning - Medium. *Medium*. <https://medium.com/image-processing-with-python/fundamentals-of-associate-rule-mining-468801ec0a29>
- Jodha, R. (2023, June 12). FP Growth Algorithm in Data Mining - Scaler Topics. *Scaler Topics*. <https://www.scaler.com/topics/data-mining-tutorial/fp-growth-in-data-mining/>
- Malaysia Public University enrolment 2022-2023*. (2024, October 15). Kaggle. <https://www.kaggle.com/datasets/andrewong74/malaysia-public-university-enrolment-2022-2023>
- Patil, M., & Patil, T. (2022). Apriori Algorithm against Fp Growth Algorithm: A Comparative Study of Data Mining Algorithms. *SSRN Electronic Journal*. <https://doi.org/10.2139/ssrn.4113695>
- Wang, K., Tang, L., Han, J., & Liu, J. (2002). Top down FP-Growth for Association rule mining. In *Lecture notes in computer science* (pp. 334–340). https://doi.org/10.1007/3-540-47887-6_34
- Zeng, Y., Yin, S., Liu, J., & Zhang, M. (2015). Research of improved FP-Growth Algorithm in association Rules Mining. *Scientific Programming*, 2015, 1–6. <https://doi.org/10.1155/2015/910281>