TDS3301 DATA MINING: TRIMESTER 2130

Name	**MUHAMMAD WAIEE BIN ZAINOL**
Student ID	**1191103225**
Lecture/Tutorial	**TC1L/TT1L**
Email	**1191103225@student.mmu.edu.my**

TABLE OF CONTENTS

1. Introduction

In the context of higher education, student performance evaluation is a critical aspect of assessing the effectiveness of educational programs and identifying areas for improvement. Traditional methods of evaluation often focus on grades, but there are numerous factors that contribute to a student's academic success. These factors can include study habits, attendance, socio-economic background, extracurricular activities, and more. Association Rule Mining can help uncover hidden patterns or relationships among these factors, showing which combinations of variables are more likely to result in positive or negative outcomes in students' performance.

The motivation behind this work is to help identify the key factors that significantly impact students' academic performance. For example, it may reveal that students who participate in certain extracurricular activities tend to perform better academically. Other than that, resource optimization is also an important aspect in this work. Institutions can optimise resource allocation based on identified associations. For example, if certain study habits are strongly associated with success, resources can be directed towards promoting those habits among students.

To have a clear understanding in Association Rule Mining application, we first reviewed 3 related previous research papers. “Oversampling technique in student performance classification from engineering course” paper focuses on combining oversampling minority class data with various kinds of classifier model for student performance classification from engineering courses. Some oversampling techniques they used might be similar with this work such as SMOTE, the difference is they only focus on sampling while in this work, we focus on finding relationships and patterns, not balancing data. In terms of dataset, this paper used student data records from the Faculty of Engineering of Rajamagala University of Technology, Thailand, which grouped into 6,882 records by student ID attribute. Similar to this work, the dataset has been grouped by student ID. However, the difference we can see is the number of attributes for this paper is only 15 while for this association rule mining work, we have 31 attributes.

Other than that, “Student performance classification and prediction in a fully online environment using Decision tree” objective is to study the regulating factors of education via digital platforms during Covid-19 Pandemic, by using Decision tree Classifier. In terms of dataset, this paper developed four separate datasets from “X-University” and Microsoft Team software., while in this work we will be using only one dataset. The similarity between these datasets contains both numeric and nominal data. For the method, the paper used Decision tree algorithm to find patterns, and not Apriori algorithm like this work.

For the last paper “Comparison of Data Mining Classification Algorithms for Student Performance”, it focuses on finding which classification model has the best performance

related to student performance data. This study used seven methods such as K- nearest neighbour, random forest etc. This study used a “Student Performance” dataset. The dataset is related to secondary education performance while “Higher Education Students Performance Evaluation" focuses on higher education. The similarity between this dataset is most of both dataset consisted of student grades, social, and same attributes. However, the “Student Performance” dataset contains 394 students while “Higher Education Students Performance Evaluation" only has 145 students information.

All these previous works give us clear understanding on the similarities and differences between datasets and the method, so that we could deliver this work better with the references.

2. Formulating Exploratory Questions

Exploratory Questions

Question 1: Is there any correlation between the **mode of transportation to the university** and **the likelihood of attending classes regularly**?

- By analysing association rules related to transportation and attendance, we aim to uncover insights into whether specific modes of transportation are linked to a higher or lower likelihood of attending classes regularly.

Question 2: Does a student's **parental status** give impact to their C**umulative Grade Point Average (CGPA)?**

- The goal is to identify whether certain transportation choices exhibit a correlation with students' academic performance. This analysis can provide insights into potential factors influencing CGPA.

Question 3: Are there any patterns between the **total salary of students** and **the number of hours they spend studying per week**?

- By exploring association rules related to total salary and weekly study hours, we aim to identify patterns that reveal how financial considerations may impact students' study habits and time commitment to academic activities.

Question 4: Are there any relationships between **students having a partner or not** and the **number of hours they dedicate to studying each week**?

- To uncover whether there is a correlation between relationship status and academic commitment. This analysis can shed light on how personal relationships may influence students' study habits and time management.

Potential Subjective Interestingness

Question 1: Is there any correlation between the **mode of transportation to the university** and **the likelihood of attending classes regularly**?

- The potential subjective interestingness in this analysis lies in revealing insights into how students' choice of transportation might influence their commitment to attending classes. Identifying patterns in transportation habits and attendance can inform university administrators and policymakers about potential interventions to improve attendance rates and student engagement.

Question 2: Does a student's **parental status** give impact to their C**umulative Grade Point Average (CGPA)?**

- Understanding the potential correlations between transportation choices and Cumulative Grade Point Average (CGPA) can provide valuable insights into the influence on academic performance. This analysis may be interesting to educators, researchers, and policymakers seeking to implement targeted support systems for students with specific transportation and family-related circumstances.

Question 3: Are there any patterns between the **total salary of students** and **the number of hours they spend studying per week**?

- The subjective interest in exploring the relationship between total salary and weekly study hours is that we could uncover the potential impact of financial considerations on students' study habits. This information can be of interest to university support services, financial aid offices, and researchers aiming to address challenges faced by students managing work and academic commitments.

Question 4: Are there any relationships between **students having a partner or not** and the **number of hours they dedicate to studying each week**?

- Investigating the association between having a life partner and weekly study hours can be subjectively interesting as it provides insights into the interplay between personal relationships and academic dedication. This analysis may be relevant to university counsellors, relationship educators, and students themselves, contributing to a better understanding of the balance between personal life and academic success.

3. Data Exploration

Dataset Description

The dataset called Higher Education Students Performance Evaluation, collected from students at the Faculty of Engineering and Faculty of Educational Sciences in 2019, is designed for predicting students' end-of-term performances using machine learning techniques. With 145 instances and 31 features, the dataset is multivariate and characterised by integer-type features, indicating a numerical nature. The associated task is classification, emphasising the prediction of categorical outcomes. The dataset is structured into three sections: personal questions (1-10), family questions (11-16), and education habits (remaining questions). The absence of missing values ensures data has been cleaned and there are no null values.

Data Exploration

I. Statistics Summary

	1	2	3	4	5	6	\
count	145.000000	145.000000	145.000000	145.000000	145.000000	145.000000	
mean	1.620690	1.600000	1.944828	3.572414	1.662069	1.600000	
std	0.613154	0.491596	0.537216	0.805750	0.474644	0.491596	
min	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	
25%	1.000000	1.000000	2.000000	3.000000	1.000000	1.000000	
50%	2.000000	2.000000	2.000000	3.000000	2.000000	2.000000	
75%	2.000000	2.000000	2.000000	4.000000	2.000000	2.000000	
max	3.000000	2.000000	3.000000	5.000000	2.000000	2.000000	

	7	8	9	10	...	23	\
count	145.000000	145.000000	145.000000	145.000000	...	145.000000	
mean	1.579310	1.627586	1.620690	1.731034	...	1.337931	
std	0.495381	1.020245	1.061112	0.783999	...	0.614870	
min	1.000000	1.000000	1.000000	1.000000	...	1.000000	
25%	1.000000	1.000000	1.000000	1.000000	...	1.000000	
50%	2.000000	1.000000	1.000000	2.000000	...	1.000000	
75%	2.000000	2.000000	2.000000	2.000000	...	2.000000	
max	2.000000	5.000000	4.000000	4.000000	...	3.000000	

	24	25	26	27	28	29	\
count	145.000000	145.000000	145.000000	145.000000	145.000000	145.000000	
mean	1.165517	2.544828	2.055172	2.393103	1.806897	3.124138	
std	0.408483	0.564940	0.674736	0.604343	0.810492	1.301083	
min	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	
...							
75%	3.000000	7.000000	5.000000				
max	4.000000	9.000000	7.000000				

To start exploring the dataset, we calculate the statistics summary for the dataset using the describe() function. This summary provides key statistical measures, including the count, mean, standard deviation, minimum, 25th percentile, median (50th percentile), 75th percentile, and maximum values for each numerical feature in the dataset. By examining these statistics, we gained insights into the central tendencies, spread, and distribution of the data.

Based on the statistics summary above, there is not much information and insights we can gain because the values inside the dataset are nominal and categorical values (0,1,2 etc.), even though it is in numeric form. Thus, the statistics could not give much insight for us.

II. Data Types

```
STUDENT ID    object
1              int64
2              int64
3              int64
4              int64
5              int64
6              int64
7              int64
8              int64
9              int64
10             int64
11             int64
12             int64
13             int64
14             int64
15             int64
16             int64
17             int64
18             int64
19             int64
20             int64
21             int64
22             int64
23             int64
24             int64
...
29             int64
30             int64
COURSE ID      int64
GRADE          int64
dtype: object
```

Based on the data type checking, it was observed that the dataset comprises two primary data types: 'Object' and 'Int64'. The 'Object' data type is associated with the 'Student ID' column, indicating that it likely contains alphanumeric identifiers. The majority of the remaining columns, consisting of various features, are of 'Int64' data type, representing numerical values.

III. Class Distribution

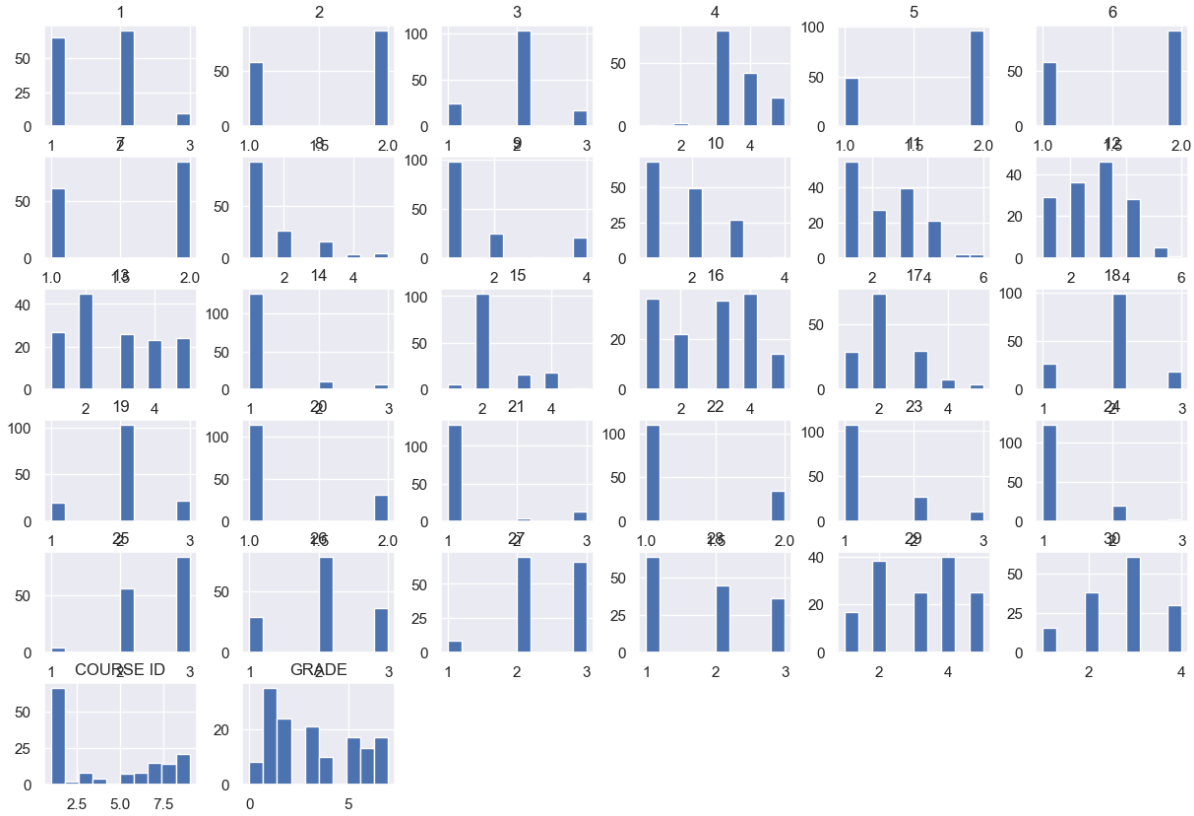

Figure 1: Histogram of all variables

To check the class distribution, we used histogram because it provides a visual representation of the distribution of values within each variable of the dataset. This is just for a quick overview of their frequency distribution. By examining the histograms, we gain insights into the shape of each feature's distribution, potential outliers, and the spread of values. For instance, column '16' (Father Occupation) histogram is skewed to the right, showing that many students have a father who has already retired, which is labelled as 1 in value.

Generally, based on the **Figure 1** above, we can observe that each category has various distributions and has different shapes. But we can't interpret any information yet because histogram is for numerical value, we are just looking into their distribution.Thus, we need to dive deeper into the dataset to gain more insights.

IV. Heatmap

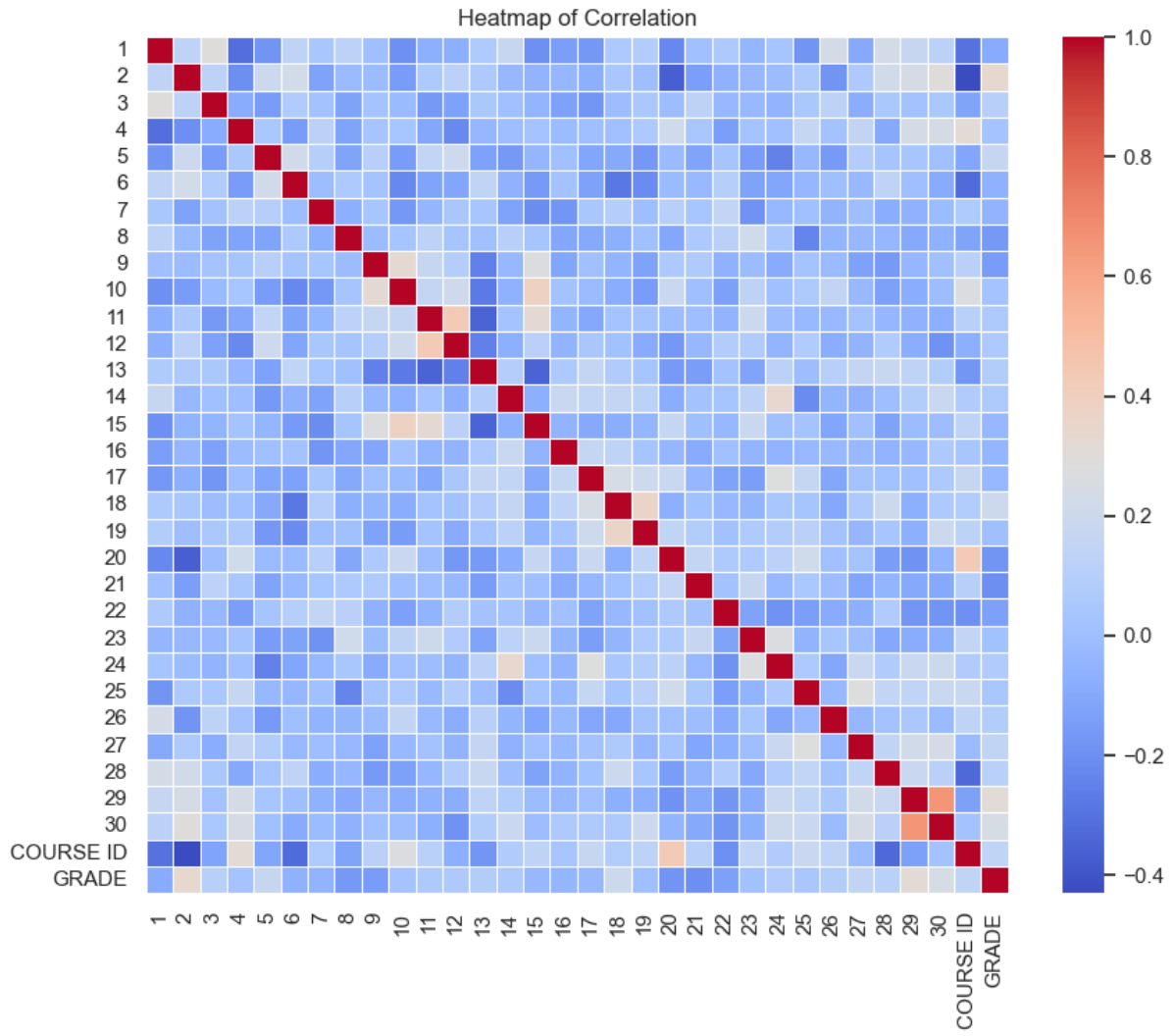

Figure 2: Heatmap of all variables

Figure 2 above represents all variables in the dataset. The analysis was conducted to explore the correlation between numerical features in the dataset. The figure provides a visual representation of the correlation matrix, with each cell colour indicating the strength and direction of the correlation between corresponding feature pairs. The colour scale ranges from blue (indicating a negative correlation) through white (no correlation) to red (indicating a positive correlation).

Based on **Figure 2**, we can observe that column '29'(CGPA) and '30'(Expected CGPA) have the highest correlation. Other than that, we can see that column '2'(Sex) and '20'(Attendance to the seminars/conferences related to the department) have low correlation between each other. Since this is a categorical variable represented as a numerical value, the information is not enough.

V. Variable Exploration using Histogram & Countplot

I. Count Plot of Grade

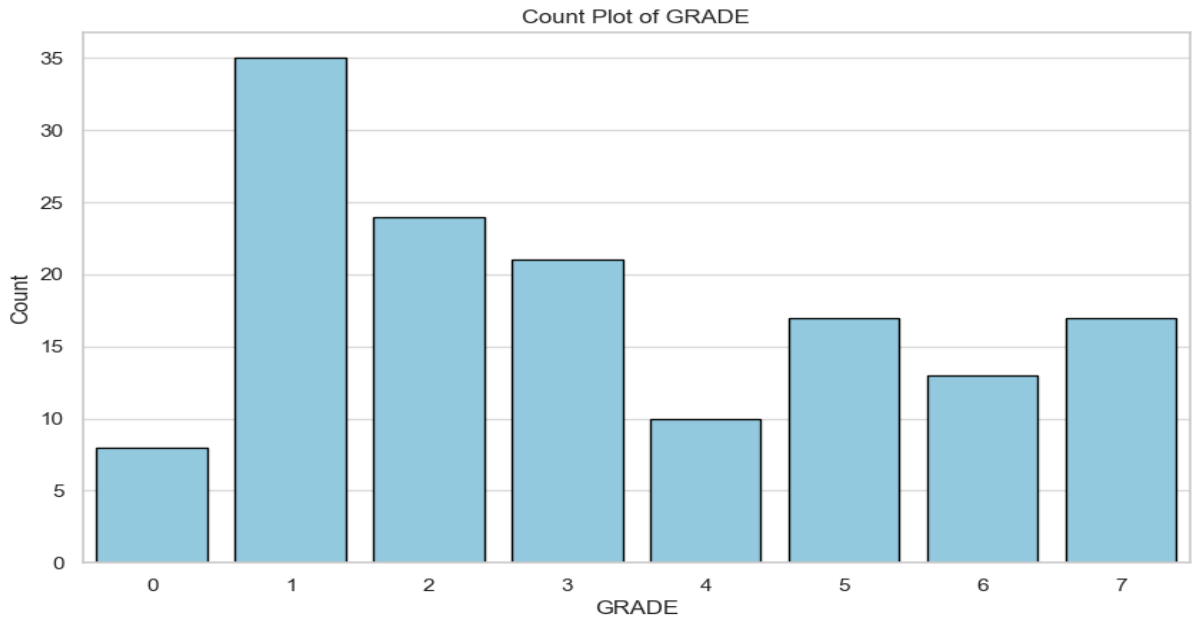

Figure 3: Count Plot of GRADE

Figure 3 above shows the count plot of Grade. Based on the count plot, we can observe that the highest number of students are getting grade 'DD' while the lowest number of students are getting 'Fail'. The second highest number is the students who get the grade 'DC'.

II. Count Plot of CGPA

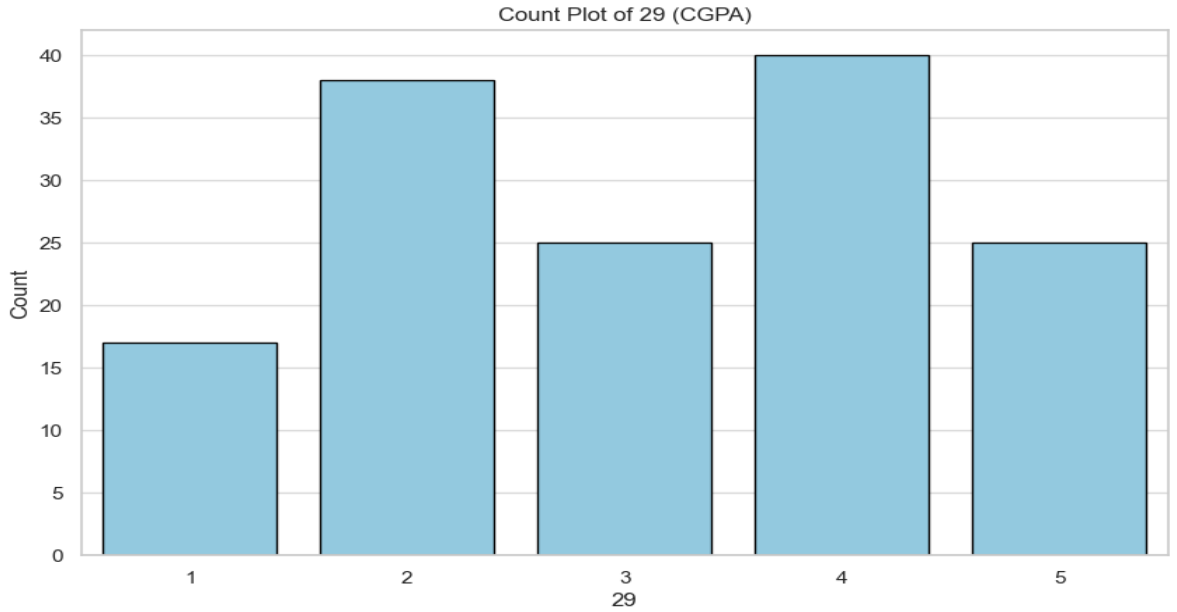

Figure 4: Count Plot of CGPA

Figure 4 above shows the histogram of column '29'(CGPA). Based on the count plot, we can observe that the highest number of students are getting 3.00-3.49 CGPA score while the lowest number of students are getting less than 2.00 CGPA Score. The second highest is the students who get 2.00-2.49 CGPA score.

III. Count Plot of Transportation to the university and Attendance to Classes

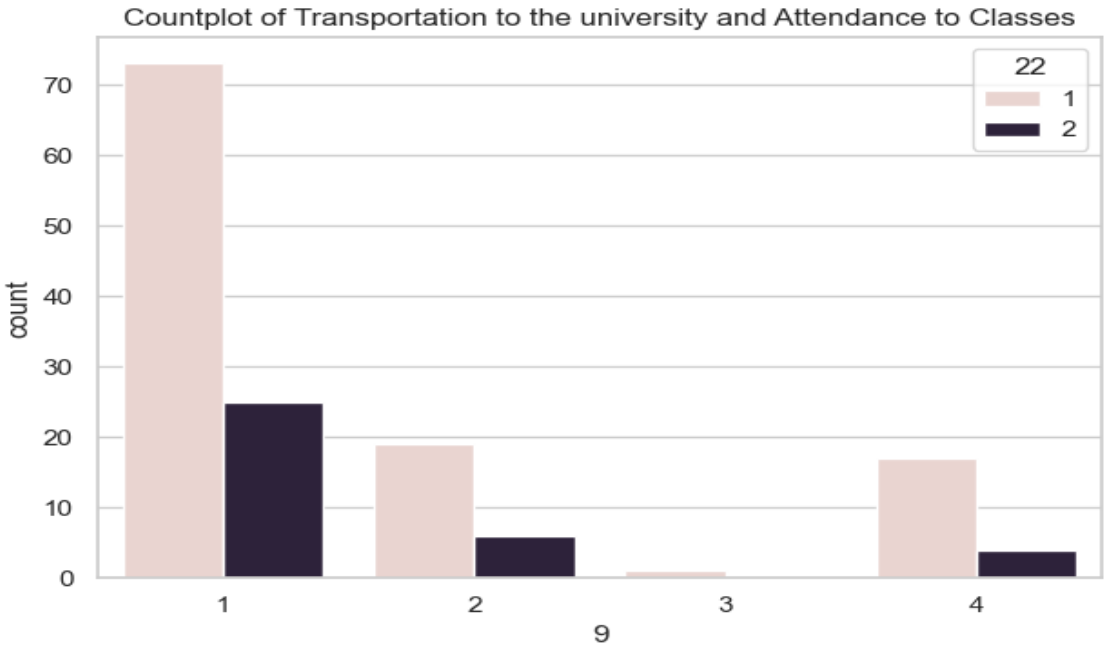

Figure 5: Count Plot of Transportation to university and Attendance to classes

Figure 5 above represents the column '9'(Transportation to the university) and '22'(Attendance to Classes). Based on the plot, we can observe that most of the students use 'Bus' as transportation to the university and always attend classes. Other than that, the lowest amount of transportation used is bicycles, even though all bicycle users always attend classes.

IV. Parental Status and Cumulative Grade Point Average (CGPA)

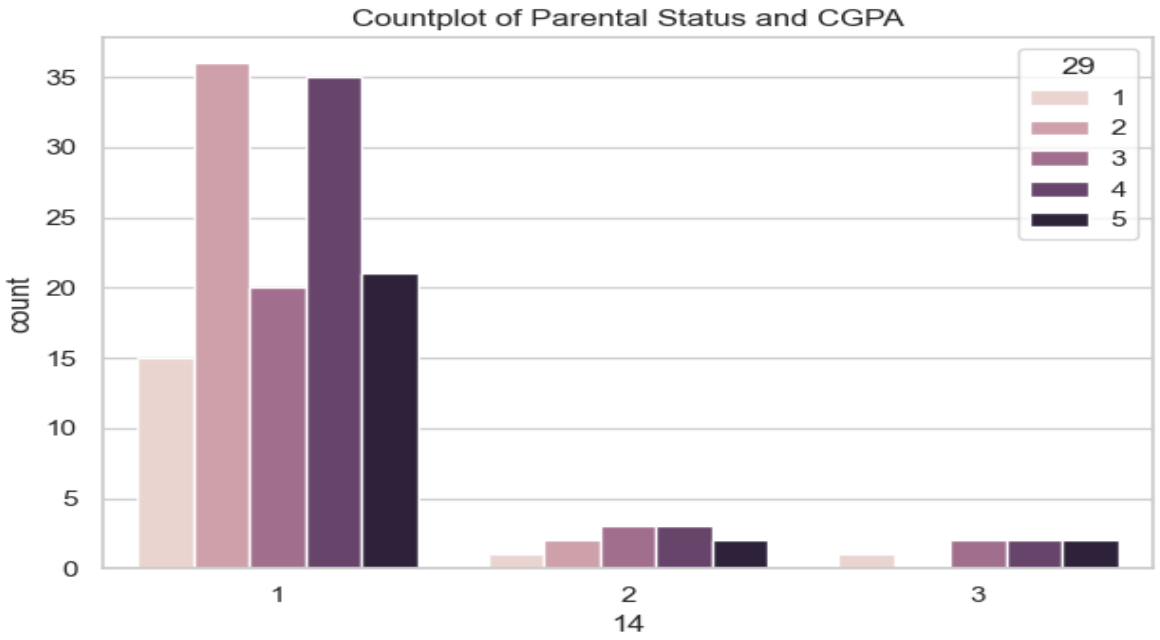

Figure 6: Count Plot of Parental Status and CGPA

Figure 6 above represents the column '14'(Parental Status) and '29'(CGPA). Based on the plot, we can observe that the highest number of students have married parents and get 2.00-2.49 CGPA, followed by students who have married parents and get 3.00-3.49 CGPA. Moreover, the lowest number of students have divorced and died parents, while getting less than 2.00 CGPA.

V. Count Plot of Total Salary and Weekly Study Hours

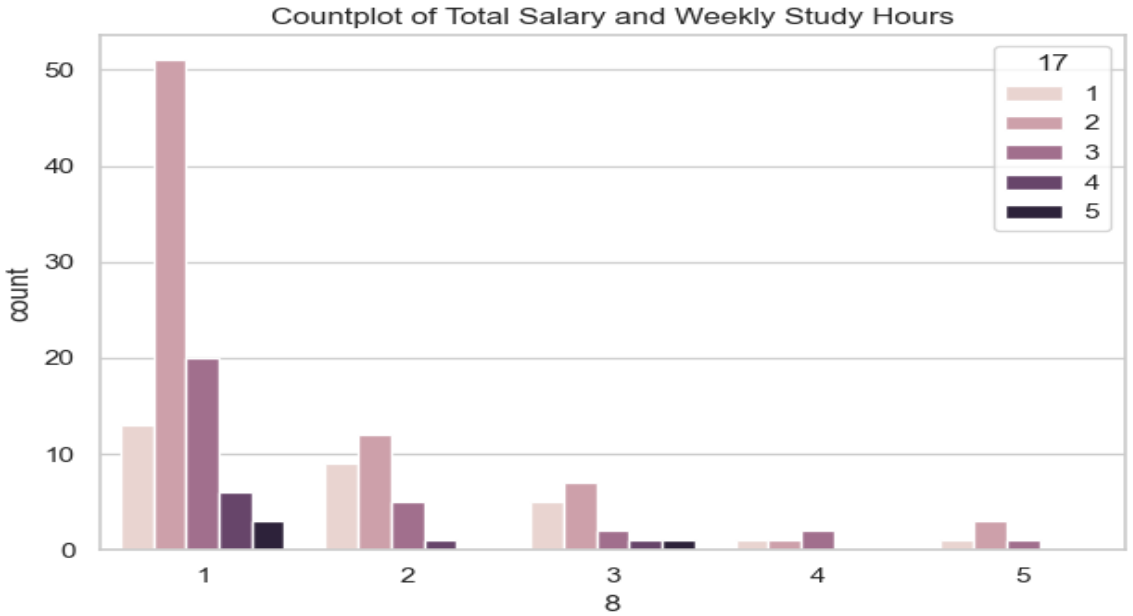

Figure 7: Count Plot of Total Salary and Weekly Study Hours

Figure 7 above represents the column '8'(Total Salary) and '17'(Weekly Study Hours). Based on the plot, we can observe that the majority of students have total salary USD 135-200 and have less than 5 hours of weekly study hours, while the least amount is the students who have a total salary above 410 and have None weekly study hours.

VI. Count Plot of Having a Partner and Weekly Study Hours

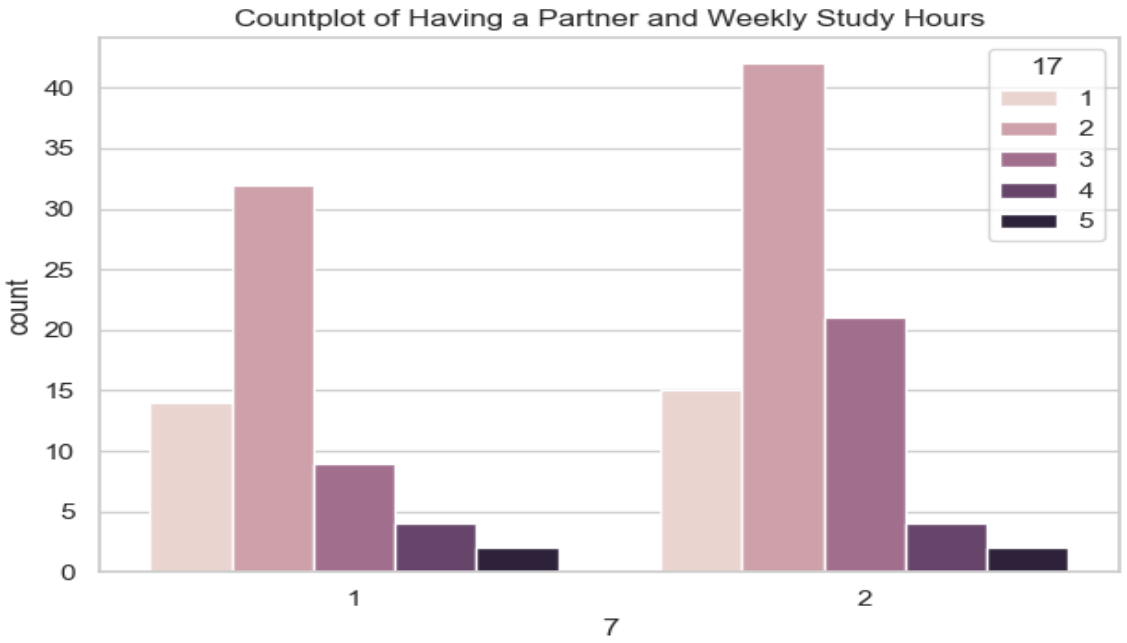

Figure 8: Count Plot of Having a Partner and Weekly Study Hours

Figure 8 above represents the column '7'(Having a Partner) and '17'(Weekly Study Hours). Based on the plot, we can observe that the majority of students are not having a partner and have less than 5 hours of weekly study hours. On the other hand, the least amount is the students who are either having a partner or not, and have more than 20 hours of weekly study hours.

4. Data Preprocessing

Data Cleaning

I. Null Values

As mentioned in the data description above, this dataset has been cleaned and there are no null values. To confirm this claim, we did check whether there is null value or not. Based on the checking, we can confirm that there are no null values in this dataset.

Transformation

I. Column Name

We changed the column name by mapping the column numbers to its value. The main purpose of this process is to make the dataset more readable, and clear to understand. Changes take place as below:

Before transformation:

	STUDENT ID	1	2	3	4	5	6	7	8	9	...	23
0	STUDENT1	2	2	3	3	1	2	2	1	1	...	1
1	STUDENT2	2	2	3	3	1	2	2	1	1	...	1
2	STUDENT3	2	2	2	3	2	2	2	2	4	...	1
3	STUDENT4	1	1	1	3	1	2	1	2	1	...	1
4	STUDENT5	2	2	1	3	2	2	1	3	1	...	2

After transformation:

	STUDENT ID	Student_Age	Sex	Grad_High_Sch_Type	Scholar_Type	Additional_work	Reg_Or_Sport	Partner
0	STUDENT1	2	2	3	3	1	2	2
1	STUDENT2	2	2	3	3	1	2	2
2	STUDENT3	2	2	2	3	2	2	2
3	STUDENT4	1	1	1	3	1	2	1
4	STUDENT5	2	2	1	3	2	2	1
...	...	...	...	...	...	...	...	...

II. Values

After column names, we transform all values and map it to the right meaning instead of categorising them by the number 0,1,2,3,4,5 etc. This process helps us to understand the value better and gives us insights on the next process whether we need more transformation or not.

Before transformation:

	STUDENT ID	Student_Age	Sex	Grad_High_Sch_Type	Scholar_Type	Additional_work	Reg_Or_Sport	Partner
0	STUDENT1	2	2	3	3	1	2	2
1	STUDENT2	2	2	3	3	1	2	2
2	STUDENT3	2	2	2	3	2	2	2
3	STUDENT4	1	1	1	3	1	2	1
4	STUDENT5	2	2	1	3	2	2	1
...	...	...	...	...	...	...	...	...

After transformation:

	STUDENT ID	Student_Age	Sex	Grad_High_Sch_Type	Scholar_Type	Additional_work	Reg_Or_Sport	Partner
0	STUDENT1	22-25	male	other	50%	Yes	No	No
1	STUDENT2	22-25	male	other	50%	Yes	No	No
2	STUDENT3	22-25	male	state	50%	No	No	No
3	STUDENT4	18-21	female	private	50%	Yes	No	Yes
4	STUDENT5	22-25	male	private	50%	No	No	Yes

Discretization

Data discretization is a process in data preprocessing where continuous variables are transformed into discrete categories or bins. This is often done to simplify the data, handle noise or outliers, and make it more suitable for certain types of analyses or machine learning algorithms. For this dataset, data discretization is applied to three continuous variables: 'CGPA,' 'Total_Salary,' and 'Weekly_Study'.

Firstly, we bin it into two categories, which are 0 and 1. For 'CGPA', if it is less than 2.99 then it will become 0, else it will be 1. After that, 'Total Salary' is set to 0 if it is less than USD 340, else is 1. For the last variable, 'Weekly_Study' is set to 0 if it is less than 5 hours or None, other than that would be labelled as 1.

	CGPA	Total_Salary	Weekly_Study
0	0	0	1
1	0	0	0
2	0	0	0
3	0	0	1
4	0	0	0
...	...	...	...
140	0	0	1
141	1	0	1
142	1	0	0
143	1	1	0
144	1	0	1

Once categorised them into 0 and 1, we set the value 0 as ‘Low’ and 1 as ‘High’ to indicate that there are two big categories.

	CGPA	Total_Salary	Weekly_Study
0	Low	Low	High
1	Low	Low	Low
2	Low	Low	Low
3	Low	Low	High
4	Low	Low	Low
...	...	...	...
140	Low	Low	High
141	High	Low	High
142	High	Low	Low
143	High	High	Low
144	High	Low	High

The discretization process was employed to transform variables into binary categories. This will simplify the dataset, making it more readable to certain machine learning algorithms that perform optimally with categorical input.

5. Association Rule Mining

Application

I. One-Hot Encoding

	Transport_1	Transport_2	Transport_3	Transport_4	Attend_Class_always	Attend_Class_sometimes
0	True	False	False	False	True	False
1	True	False	False	False	True	False
2	False	False	False	True	True	False
3	True	False	False	False	True	False
4	True	False	False	False	True	False
...	...	...	...	...	...	...
140	True	False	False	False	True	False
141	False	False	False	True	False	True
142	True	False	False	False	True	False
143	False	True	False	False	True	False
144	True	False	False	False	True	False

One-Hot Encoding technique has been applied in order to represent categorical variables as binary vectors. The process involves creating binary columns for each category in the original variable. By doing this, we can proceed to apply algorithms. Example shown in figure above, One-Hot Encoding applied to column ‘Transport’ and ‘Attend_Class’.

II. Apriori Algorithm

The Apriori algorithm is a classic association rule mining algorithm used to discover frequent itemsets within a dataset. Developed by Agrawal and Srikant, the algorithm is fundamental in extracting patterns and relationships in transactional data. The key concept of the Apriori algorithm is its use of the "Apriori property," which states that if an itemset is frequent, then all of its subsets must also be frequent. In this application, we set the minimum support of 0.3 to find the frequent patterns.

Interestingness Measures

Interesting measures ”Lift” are used in this study. This measure is a common and valid approach in association rule mining. It is a measure that helps assess the strength and significance of association rules by comparing the observed support of a rule to what would be expected if the items in the rule were independent. By these measures, we can decide

whether our rules are interesting or not. For this measure, we will filter the result to having a lift score above 1.0.

Patterns

	support	itemsets
0	0.675862	(Transport_1)
1	0.758621	(Attend_Class_always)
2	0.503448	(Transport_1, Attend_Class_always)

	support	itemsets
0	0.875862	(Parental_Status_married)
1	0.448276	(CGPA_High)
2	0.551724	(CGPA_Low)
3	0.386207	(Parental_Status_married, CGPA_High)
4	0.489655	(CGPA_Low, Parental_Status_married)

	support	itemsets
0	0.937931	(Total_Salary_Low)
1	0.710345	(Weekly_Study_Low)
2	0.668966	(Total_Salary_Low, Weekly_Study_Low)

	support	itemsets
0	0.579310	(Partner_No)
1	0.420690	(Partner_Yes)
2	0.710345	(Weekly_Study_Low)
3	0.393103	(Partner_No, Weekly_Study_Low)
4	0.317241	(Weekly_Study_Low, Partner_Yes)

After the Apriori algorithm has been applied, these are the result of frequent patterns. All of these patterns have been filtered by having minimum support more than 0.3. Based on the patterns above, we can conclude that the highest support values between two variables are ("Total_Salary_Low, Weekly_Study_Low") while the lowest is ("Weekly_Study_Low, Partner_Yes").

6. Result Discussion

I. Is there any correlation between the mode of transportation to the university and the likelihood of attending classes regularly?

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction	zhangs_metric
0	(Transport_1)	(Attend_Class_always)	0.675862	0.758621	0.503448	0.744898	0.981911	-0.009275	0.946207	-0.053778
1	(Attend_Class_always)	(Transport_1)	0.758621	0.675862	0.503448	0.663636	0.981911	-0.009275	0.963653	-0.070909

Based on the result above, we can observe that (Attend_Class_always) → (Transport_1) has the highest lift measure which is 0.981911. This result might have happened due to lack of discretization. Even though we are looking for a lift above 1.0, the score is quite close to 1.0. This result can be considered as positively correlated. Looking back to our formulated exploratory question, we can conclude that there is a correlation between type of transportation to the university with attendance to classes, which students that have buses as transportation, will most likely to attend classes.

II. Does a student's parental status give impact to their Cumulative Grade Point Average (CGPA)?

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction	zhangs_metric
0	(CGPA_Low)	(Parental_Status_married)	0.551724	0.875862	0.489655	0.887500	1.013287	0.006421	1.103448	0.029252
1	(Parental_Status_married)	(CGPA_Low)	0.875862	0.551724	0.489655	0.559055	1.013287	0.006421	1.016626	0.105634

Based on the result above, we can observe that (Parental_Status_married) → (CGPA_Low) has the highest lift measure which is 1.013287. This result can be considered as positively correlated since the lift score is above 1.0. Based on the result, we can observe that students who have married parents, will most likely get a low CGPA score. However, if we refer back to our formulated exploratory questions, married couples do not seem to give much impact to student's CGPA, because most of them still get low CGPA scores. We are expecting a pattern where parents' issues will affect their academic performance. Thus, we can conclude that this parental status most likely will not give much impact to a student's CGPA, and future exploration must be done to find another potential factor.

III. Are there any patterns between the total salary of students and the number of hours they spend studying per week?

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction	zhangs_metric
0	(Total_Salary_Low)	(Weekly_Study_Low)	0.937931	0.710345	0.668966	0.713235	1.004069	0.002711	1.010080	0.065292
1	(Weekly_Study_Low)	(Total_Salary_Low)	0.710345	0.937931	0.668966	0.941748	1.004069	0.002711	1.065517	0.013991

Based on the result above, we can observe that (Total_Salary_Low) → (Weekly_Study_Low) has the highest lift measure which is 1.004069. This result can be considered as positively correlated since the lift score is above 1.0. Based on the result, we can observe that students who have a low total salary that is below USD340, will have low weekly study hours, which is less than 5 hours in this context. If we refer back to our formulated exploratory questions, we can observe that the pattern is students who have low salary still having low weekly study hours. We are expecting an interesting pattern where higher salary students would have lower weekly study hours since they are working. Thus, we can conclude that low total salary probably will affect student's weekly study hours, and there might be another factor we can observe in future on why their weekly study hours are still low even though the salary is low.

IV. Are there any relationships between students having a partner or not and the number of hours they dedicate to studying each week?

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction	zhangs_metric
0	(Weekly_Study_Low)	(Partner_Yes)	0.710345	0.420690	0.317241	0.446602	1.061595	0.018407	1.046824	0.200311
1	(Partner_Yes)	(Weekly_Study_Low)	0.420690	0.710345	0.317241	0.754098	1.061595	0.018407	1.177931	0.100155

Based on the result above, we can observe that (Partner_Yes) → (Weekly_Study_Low) has the highest lift measure which is 1.061595. This result can be considered as positively correlated since the lift score is above 1.0. Based on the result, we can observe that students

who are having a partner will most likely have low weekly study hours. If we refer back to our formulated exploratory questions, we can observe that the students who have a partner will have low weekly study hours. Thus, we can conclude that this pattern is interesting and obvious, that having a partner will most likely decrease students' weekly study hours.

7. Conclusion

In conclusion, we can conclude that there is an interesting relationship and patterns we can find in this dataset based on the formulated exploratory questions that will be useful for future research. From the first exploratory question, we found that students who take buses as their transportation to university will most likely attend classes. Next, we can observe that even though students have married parents, their CGPA score is still low since we are expecting family issues would affect their academic performance, thus there should be other crucial factors that affect CGPA score. Other than that, we know students who have a low total salary, will still have low weekly study hours. This indicates that there is another factor out there on why these students have low weekly study hours. For the last question, we can conclude that having a partner will most likely have low weekly study hours. This result shows that having a partner might affect the students' study hour.

From this result, there are potential use cases that we can conclude and do future research. Since students who take buses as their transportation to their university will most likely attend classes, this information is important to all academic institutions, and the government. Higher authorities should improve public transportation efficiency for students, in order to encourage students to attend classes and make sure they are not skipping classes. Moreover, since students who have a partner will most likely have low weekly study hours, this information needs to be addressed properly. Universities could run a campaign of encouraging students to have a study group with their partners and friends, so that this will encourage them to enjoy studying and to not ignore their academic performance.

For future studies, we can try to use a different approach of association rule mining to evaluate whether this is the best approach or not. This is to ensure that we are having the best result. Other than that, we can still use the same approach but we could test it on another potential variable for different formulated exploratory questions. This will overcome our limitations where the result is not really satisfying and , and also proving that there is another factor in influencing some important variables.