Journal of Mathematics and Computing Science, 2021,
Volume 7, No 2, 81- 88

Analyzing Course Affecting the Results of Inferential Statistics Students in UiTM Seremban using Association Rule Mining

Jaida Najihah Jamidin[1]*, Siti Sarah Januri [2], Dayang Masrafisha Awang Raffinur [3], Nur Asyikin Mohd[4] and Nursyaza Nisa Sazali [5]

[1,2]Fakulti Sains Komputer dan Matematik, Universiti Teknologi MARA Cawangan Negeri Sembilan, Kampus Seremban, Seremban, Negeri Sembilan, Malaysia

*jaida5698@uitm.edu.my

Abstract: Inferential Statistics was recognized as a high failure rate statistics course taken by the students of Bachelor of Science (Hons.) Statistics. Inferential Statistics is the most challenging course compared to the other courses since a majority of the students failed this course. This study aimed to analyze the courses affecting the result of Inferential Statistics. An Association Rule Mining was used to 126 graduated students from matriculation and foundation. Results showed that 239 rules had been obtained during the process of association rule mining. After removing the redundancy, only 106 most appropriate rules are associated with the objectives. This study focused on providing an effective mechanism for lecturers to improve students' performance in the high failure rate of statistics courses. The university administration needs to improve the teaching and learning method to produce good academic performance.

Keywords: Inferential Statistics Course, Association Rule Mining

1 Introduction

Higher education has experienced new development for the past few decades. The infusion and integration of the new information technologies in teaching and learning significantly impacted the educational environment [1]. Many information systems have been developed and successfully implemented to support the educational process. These systems technically save almost every data regarding students from the enrollment into courses set by the university until graduation, as was discussed by [2]. Moreover, the system also stores the demographic profile of all students, such as name, age, semester, academic background, results from every semester, and others. If these data are analyzed and visualized correctly, it can provide valuable information and knowledge that can be used to enhance students learning skills and teaching process.

In university, training a high-level talent through teaching and education is the most crucial task, and due to that, teaching work has always been at the centre of the school and university. To build a higher level of education, the university administrator has to improve the quality of education and students' performance. So, it is important to analyze the data of students' performance to investigate the trend of students' performances. Some factors can affect the students' performances, and one of the factors is the subjects or the courses assigned to each program. In each program, there can be subjects that have higher failure rates than other subjects.

In UiTM, Inferential Statistics is a common and compulsory subject or course for statistics students. Previously, the Inferential Statistics course was taken by the third-year student, and when some students failed the course, they had to take the supplementary examination to qualify them for internship. If the student does not pass the supplementary examination, the student will have to take the course next semester.

eISSN 0128-0767 © Unit Penerbitan UiTM Kelantan

This is a crucial issue because it affects the student to graduate on time and affects their CGPA. Hence, this research aimed to investigate the relationship between courses in Semester 1 until three that affect the result of the Inferential Statistics course. In addition, this study also wants to determine the interesting rules form by the courses taken in Semester 1 until three and the result of the Inferential Statistics course using association rule mining.

Association rule mining is a popular and well-researched data mining method to discover interesting relations between variables in large databases [3]. Research about association rule mining has grown extensively in recent decades. Many areas have used this method for problems such as fatal pedestrian crashes [4]–[5], road accidents [6]–[8], discovering symptom patterns of COVID-19 patients [9], identification of cancer-related genes [10], and construction [11]. Association rules are if or then statements that help to uncover the relationships between unrelated data in the database, relational database, or other information. Association rules are used to find the relationships between the frequently used objects. Association rules are very effective for a large set of unsupervised data by providing key insights that can be used for data-driven decision-making [4]. Besides that, it had higher predicting accuracy than statistical methods, as there is no need for pre-assumption for them about the data distribution and the relationship between the dependent and independent [8],[6]. Applications of association rules are basket data analysis, classification, cross-marketing, clustering, catalogue design, and loss-leader analysis [12]. In this approach, all the possible rules are listed first, and those rules that do not satisfy the given condition are pruned [9].

The use of different data mining techniques can be viewed as potential groundwork for a systemic change, and it can have a significant positive impact if it is seen and serve as an instrument that can help higher education institutions find solutions for their most specific issues [13]. Educational data mining (EDM) is defined as the area of scientific inquiry centred around the development of methods for making discoveries within the unique kinds of data that come from educational settings, and this method is used to improve in understanding students and the setting they learn [14]-[15]. It includes the area of guiding academic programs design and assessment [16]-[17], evaluation process in the e-learning [18], and educational technology classroom research [19]. According to the previous study, association rule mining was applied to determine whether there was any significant association existed between students' performances in different subjects. If the relationship existed, this information could be used as a guide for the students to improve their overall performance. Based on the results, the rules discovered through association rule mining were used to predict the outcome of the forthcoming examination. So, according to (Chandrakar and Saini, 2013), the prediction could be used to identify the subject that the student needed to focus more on at the beginning of the semester.

Based on (Damaševičius 2009), association rule mining was applied in analyzing the Informatics course examination results, ranking course topics following their importance for final course marks, and proposing which specific course topic should be improved to achieve higher student learning effectiveness and progress. Moreover, association rule mining is also used to identify the weak courses in the current semester based on the previous semester's results to improve the student results. Strong association rules results will identify the weak courses in the current semester using the previous results. Therefore, as discussed by (Tadiparthi et al. 2011) stated from the generated strong rules, it identified that when a student ever failed the specific courses from the previous semester, the student will fail the targeted course in the current semester. Hence, this method was applied to achieve the objectives of the study.

2 Methodology

This research used secondary data, where the data was collected from academic welfare (HEA). The research takes 126 students from matriculation as a sample from batch 2015 till 2019. The data used were the students' results from previous courses taken to identify the courses that affected the result of the Inferential Statistics course.

The method used in this study was Association Rule Mining. Association Rule Mining is sometimes referred to as "Market Basket Analysis", as it was the first application area of association mining. Association rule mining is an "if-then" statement that helps to show the probability of associations and/or correlation relationships among large data sets of data items. An association rule has two parts which are antecedent (if) and a consequent (then). An antecedent is an item found within the data, while a consequent is an item found in combination with the antecedent. Association rules are created by searching data for frequent if-then patterns and using the criteria support in confidence in equation (1) to identify the most important relationships. The measures of the effectiveness of the rule are as follows:

Confidence in equation (1) includes A's transaction and contains the conditional probability of B, which indicates the number of times the if-then statements is found true.

$$Confidence = \frac{Number\ of\ transactions\ with\ both\ A\ and\ B}{Total\ number\ of\ transactions\ with\ A} = \frac{P(A \cap B)}{P(A)} \quad (1)$$

The association rule mining can be viewed as a two-step process in which was the first process will find all frequent item-sets which was each of these item-sets will occur at least as frequently as a predetermined minimum support count. The second process will be generated strong association rules from the frequent item-set which was the rules must satisfy minimum support and confidence. These rules are called strong rules.

The dependent variable for this study was the result of the Inferential Statistics subject by Bachelor of Science (Hons.) Statistics students during Semester 4. On the other hand, the independent variable for the study was the factor influencing the passing rates of the Inferential Statistics subject. The data set needs to be cleaned during the cleaning process by removing the inconsistencies. In the cleaning process, three subjects were removed from the data set which is Islamic Thought and Civilization (CTU) in Malay, National Kesateria (HBU), and the third language subject which are Introductory Mandarin (TMC) and Introductory Arabic (TAC). In addition, the students who were dismissed from UiTM also were removed from the data set. Furthermore, students who fail a particular subject and repeat it to pass were grouped in a new code. The new code was created based on the duration taken for the student to pass the subject they failed, followed by the passing grade obtained to facilitate the process of Association Rule Mining (ARM).

R studio software version 3.5.2 and 1.1.463 were used to run this method. The results showed that 239 rules can be generated, including the redundant rules. The process of removing redundancy was indispensable to avoid redundant rules. After removing redundant rules, only 106 rules total from three categories are obtained and used to be made as to the result of the following process. Rules obtained included independent variables on the left-hand side followed by the dependent variable on the right-hand side and followed by the value of confidence.

3 Result and Finding

This study was conducted to identify the causes of the failure rate of the Inferential Statistics (STA560) course is increased. The relationship between the courses in Semester 1 to 3 with the result of the Inferential Statistics course was recorded. It sought to determine whether the result of the Inferential Statistics course was influenced by the results of the other courses that the students in the past semester took. The result was categorized into three categories: Excellent, Moderate, and Failed. Before a redundancy process was performed, the total rules were 162 rules, 32 rules, and 45 rules, respectively, for each category. However, after the redundancy process was performed, only 65 rules, seven rules, and 34 rules were relevant to be applied in this study. The independent variables that covered all the courses taken in Semester 1 until Semester 3 were set at the left-hand side and the dependent variable which was the Inferential Statistics course was set as the right-hand side.

The confidence values explained the accuracy and strength of each existing rule which demonstrated the importance of the relationship. The rules were classified into three categories Excellent, Moderate, and Failed through the finding. However, the interest rules were Moderate and Failed category which indicated the increase in student performance on achieving an outstanding result in Inferential Statistics. Besides, this study also aimed to reduce students' failure in Inferential Statistics. Table 1, Table 2, and Table 3 show the confidence value for each category.

For category A, the Excellent category, Table 1 indicates the confidence value is 1.0, which concluded that this value contributed to the dependent variable Inferential Statistics to get an A was significantly dependent on the independent variables. There were 162 rules under this category before removing the data redundancy and only 65 rules left after removing the data redundancy. It shows the first ten rules in the Excellent Category and shows that IF MAT441=A+ AND STA400=A AND STA470=A THEN STA560="EXCELLENT" with confidence value 100%. Those courses with 100% confidence value significantly impact the students to get an excellent result in Inferential Statistics.

Table 1: Excellent Category

Number	**Courses**			**Confidence**
	Semester 1	**Semester 2**	**Semester 3**	

1	MAT441=A+
	STA400=A
		STA470=A	-	1.000

2	MAT441=A+	QMT437=A
		STA470=A
			-	1.000

3	MAT441=A+	STA470=A	MAT523=A	1.000

4	MAT441=A+	STA450=A
		STA470=A
			-	1.000

5	MAT441=A
	STA400=A
		QMT437=A	-	1.000

6	STA400=A
	STA420=A
		-	STA500=A-	1.000

7	STA400=A	-	ITS472=B
			STA500=A-
				1.000

8	MAT441=A
	STA400=A
		-	ITS472=B	1.000

9	STA400=A	MAT423=A
		STA470=A-
			-	1.000

10	MAT441=A
	STA400=A
		MAT423=A	-	1.000

Secondly, Table 2 for the Moderate category consists of grades B+, B, and B- mostly shows the confidence value is also 1.0, which means the Inferential Statistics result will be 100% influenced by the results of independent variables. It can be concluded that this value contributed to the dependent variable (Inferential Statistics) to get a B+, B, and B- was significantly depending on the independent variables. There were 32 rules under this category before removing the data redundancy, and after removing the data redundancy, there were left with seven rules as in Table 2. The result shows that IF STA420=B- AND MAT423=C AND QMT437=C THEN STA560="MODERATE" with confidence value 100%. Those courses with 100% confidence value significantly impact the students to get a moderate result in Inferential Statistics.

Table 2: Moderate Category

Number	**Courses**			**Confidence**
	Semester 1	**Semester 2**	**Semester 3**	

1	STA420=B-	MAT423=C
		QMT437=C
			-	1.000

2	STA400=B	QMT437=A-
		STA470=B-
			-	1.000

3	CSC415=C+
	MAT441=B
	STA420=B
		-	-	1.000

4	-	ITS432=B-
		STA450=A-
			MAT491=C+	1.000

5		MAT422=B+
		STA450=A-
			MAT491=C+	1.000

6		ITS432=B-
		MAT422=B+
		STA450=A-
			MAT491=C+	1.000

7	MAT441=B+	ITS432=A
		MAT523=A
			STA500=A	1.000

Lastly, for the Failed category, which was graded by C-. This category mostly shows the confidence value is 1.0, which concluded that this value contributed to the dependent variable (Inferential Statistics) to get a C- was significantly dependent on the independent variables as shown in Table 3. There were 45 rules under this category before removing the data redundancy and 34 rules after removing the data redundancy. The interesting rules for this category are shown in Table 3. The table shows that IF STA450=B- AND STA470=C+ AND STA500=C THEN STA560="FAILED" with confidence value 100%. Those courses with 100% confidence value significantly impact the students to fail in Inferential Statistics.

For the last part of the analysis, the summary of the crucial course for Semester 1, 2, and 3 was obtained. Table 4 was obtained from Table 1 and 2, which were the courses that students need to excel or pass to get at least moderate results in Inferential Statistics. The results show that fourteen courses had given a high contribution towards Inferential Statistics results.

Table 3: Failed Category

Number	Courses			Confidence
	Semester 1	Semester 2	Semester 3	

1	-	STA450=B-
		STA470=C+
			STA500=C	1.000

2	-	STA450=B-	STA500=C	1.000

3	-	MAT422=A-
		STA450=B-
			STA500=C	1.000

4	-	STA450=B-	ITS472=C+
			STA500=C
				1.000

5	-	MAT422=A-	STA470=C+
			STA500=C
				1.000

6	-	MAT422=A-	MAT491=C
			STA500=C
				1.000

7	-	MAT422=A-	ITS472=C+
			STA500=C
				1.000

8	-	STA450=B-
		STA470=C+
			MAT491=C	1.000

9	-	MAT422=A-
		STA450=B-
		STA470=C+
			-	1.000

10	-	STA450=B-
		STA470=C+
			ITS472=C+	1.000

Table 3: Failed Category *(Cont…)*

Table 4: All Courses Influenced to Inferential Statistics Result

Semester	Course
1	1. CSC415 2. MAT441 3. STA400 4. STA420
2	1. ITS432 2. MAT422 3. MAT423 4. QMT437 5. STA450 6. STA470
3	1. ITS472 2. MAT491 3. MAT523 4. STA500

4 Conclusions

The goals of this study had been successfully achieved with 239 rules had been mined by using the Association Rules Mining technique. After removing redundancy, only 106 most appropriate and effective rules were associated with the objectives of this research. The interesting rules show 72 that out of 106 rules were in 100% confidence level, which significantly affects the Inferential Statistics result. This study provides significant patterns of the courses affecting the results of Inferential Statistics amongst first-timers. These patterns can then be used in predicting the future results of the students who will be taking the Inferential Statistics course in semester 4. The important courses that the students should focus more on have been identified. The subjects that occurred most frequently in the rules for all three categories were Calculus 2, Probability Statistics and, Statistical Method. These frequent items show that they are the most significant subjects that influent the performance of students in the Inferential Statistics course. For students predicted to fail this subject, the lecturers should take intervention as early as in semester 1. The main contribution of this study to association rule mining is describing unique item sets and providing an efficient mining process. This study is applicable only for UiTM Seremban and suggested extending this study to all branches in UiTM.

5 References

[1] V. S. Chomal and J. R. Saini, "A study and analysis of paradigm shifts in education triggered by technology," International Journal of Research in Economics & Social Sciences 3(1): 14–28, 2013.

[2] O. Chandrakar and J. R. Saini, "Predicting Examination Results using Association Rule Mining," International Journal of Computer Applications 116(1): 7–10, 2015.

[3] J. Arora, N. Bhalla and S. Rao, "A Review on Association Rule Mining Algorithms," International Journal of Innovative Research in Computer and Communication Engineering 1(5): 1246–1251, 2013.

[4] S. Das, R. Tamakloe, H. Zubaidi, I. Obaid and A. Alnedawi, "Fatal pedestrian crashes at intersections: Trend mining using association rules," Accident Analysis and Prevention 160(November 2020): 106306, 2021.

[5] S. K. Sivasankaran, P. Natarajan and V. Balasubramanian, "Identifying Patterns of Pedestrian Crashes in Urban Metropolitan Roads in India using Association Rule Mining," Transportation Research Procedia 48(2019): 3496–3507, 2020.

[6] F. Jiang, K. K. R. Yuen and E. W. M. Lee, "Analysis of motorcycle accidents using association rule mining-based framework with parameter optimization and GIS technology," Journal of Safety Research 75: 292–309, 2020.

[7] C. Arteaga, A. Paz and J.W. Park, "Injury severity on traffic crashes: A text mining with an interpretable machine-learning approach," Safety Science 132(May): 104988, 2020.

[8] S. Momeni Kho, P. Pahlavani and B. Bigdeli, "Classification and association rule mining of road collisions for analyzing the fatal severity, a case study," Journal of Transport and Health 23(September): 101278, 2021.

[9] M. Tandan, Y. Acharya, S. Pokharel, and M. Timilsina, "Discovering symptom patterns of COVID-19 patients using association rule mining," Computers in Biology and Medicine 131(December 2020): 104249, 2021.

[10] C. Gakii and R. Rimiru, "Identification of cancer related genes using feature selection and association rule mining," Informatics in Medicine Unlocked 24: 100595, 2021.

[11] H. Abdirad and P. Mathur, "Artificial intelligence for BIM content management and delivery: Case study of association rule mining for construction detailing," Advanced Engineering Informatics 50(August): 101414, 2021.

[12] T. A. Kumbhare and S. V. Chobe, "An Overview of Association Rule Mining Algorithms," International Journal of Computer Science and Information Technologies 5(1): 927–930, 2014.

[13] H. Aldowah, H. Al-Samarraie and W.M. Fauzy, "Educational data mining and learning analytics

for 21st century higher education: A review and synthesis," Telematics and Informatics 37(January): 13–49, 2019.

[14] R. S. J. Baker, "Data Mining for Education," In International Encyclopedia of Education. 3rd editio pp. 1–13. Oxford, UK: Elsevier, 2013.

[15] G. Czibula, A. Mihai and L. M. Crivei, "S PRAR: A novel relational association rule mining classification model applied for academic performance prediction," Procedia Computer Science 159: 20–29, 2019.

[16] A. A. Yahya and A. Osman, "Using Data Mining Techniques to Guide Academic Programs Design and Assessment," Procedia Computer Science 163: 472–481, 2019.

[17] F. Martínez-Abad, A. Gamazo and M. J. Rodríguez-Conde, "Educational Data Mining: Identification of factors associated with school effectiveness in PISA assessment," Studies in Educational Evaluation 66(December 2019), 2020.

[18] M. W. Rodrigues, S. Isotani and L. E. Zárate, "Educational Data Mining: A review of evaluation process in the e-learning," Telematics and Informatics 35(6): 1701–1717, 2018.

[19] C. Angeli, S.K. Howard, J. Ma, J. Yang and P.A. Kirschner, "Data mining in educational technology classroom research: Can it make a contribution?," Computers and Education 113: 226–242, 2017.

[20] R. Damaševičius, "Analysis of academic results for informatics course improvement using association rule mining," Information Systems Development: Towards a Service Provision Society 357–363, 2009.

[21] A. Tadiparthi, R. S. Prasad and S. N. T. Rao, "Identifying Weak Subjects using Association Rule Mining," International Journal of Scientifics & Engineering Research 2(11): 1–3. 2011.