Enhanced Knowledge in Sciences and Technology Vol. 3 No. 1 (2023) 237-246

EKST

Homepage: http://publisher.uthm.edu.my/periodicals/index.php/ekst
e-ISSN : 2773-6385

Analysis on Students' Academic Performance in Relation to The Results of Pre-University Examination

Chong Qi[1], Sabariah Saharan[2*]

[1] Jomrun Sdn. Bhd.,
Block C2, Level 3, No.8, CBD Perdana 3, Jalan Perdana,
63000 Cyberjaya, Selangor, MALAYSIA

[2] Department Mathematics and Statistics,
Faculty of Applied Sciences and Technology,
Universiti Tun Hussein Onn Malaysia (Pagoh Campus),
84600 Pagoh, Muar, Johor, MALAYSIA.

*Corresponding Author Designation

DOI: https://doi.org/10.30880/ekst.2023.03.01.027
Received 14 January 2023; Accepted 11 April 2023; Available online 3 August 2023

Abstract: There is a great deal of uncertainty regarding the factors that influence their final year grade, which includes their entry qualification. This paper investigates the impact of entry qualification and pre-university CGPA on student performance at the university level. Entry qualifications are critical for educational institutions or educational providers to ensure the quality of the graduates. The goal of this study is to analyse and compare performance of Bachelor of Science (Industrial Statistics) with Honours (BWQ) students. Total of 52 students were selected form the Faculty of Applied Sciences and Technology (FAST), Universiti Tun Hussein Onn Malaysia (UTHM). The students with entry qualification of Malaysian Higher School Certificate (STPM) and Malaysian Matriculation Programme were analysed and compared. Paired t test and Z test were carried out to analyze the impact of pre-university's CGPA and each semester's GPA as well as impact of entry qualification towards their final year grade. Classification and Regression Tree (CART), K-Nearest Neighbors and Naïve Bayes were used to develop and predict the students' performance. The findings show that there is no relation between the result obtained from previous semester towards the next semester. Meanwhile, students from STPM outperform Matriculation in terms of their GPA per semester, pre-university CGPA as well as their final CGPA. The K-Nearest Neighbors and Naïve Bayes models have been documented as the most efficient data mining techniques in predicting student performance with the highest percentage of accuracy of 100%.

Keywords: Student Performance, Paired t test, Z test, Data Mining, Classification and Regression Tree, K-Nearest Neighbors, Naïve Bayes

**Corresponding author: sabaria@uthm.edu.my*

1. Introduction

In Malaysia, there are 20 public universities, also known as Universiti Awam [1]. After completing the Malaysian Higher School Certificate or Sijil Tinggi Persekolahan Malaysia (STPM) and Malaysian Matriculation Programme, most students will continue their education at one of these universities. Universiti Tun Hussein Onn Malaysia (UTHM) is one of the UA that offers a Bachelor of Science (Industrial Statistics) with Honours programme for students. Students from STPM and Malaysian Matriculation Programme or Diploma who have a Cumulative Grade Point Average (CGPA) of A in Mathematics can apply to this programme.

There are 584,576 students enrolled in public universities in Malaysia in 2020, an increase of 8.13% (43938 students) from 2015 [2]. A country's poverty rate is highly associated with its educational level [3]. Thus, it is critical for Malaysian students to enroll in higher education for Malaysia to achieve Shared Prosperity Vision 2030, which is to make Malaysia a nation that achieves sustainable growth while ensuring fair and equitable distribution across income groups, ethnicities, regions, and supply chains.

The number of graduates in Malaysia in 2020 was 5.36 million, a 4.4% increase over the previous year [4]. Greater university academic performance will lead to more employment opportunities, as most human resource managers will hire employees based on their prestige, academic performance, and university rank after reviewing stacks of applications [5]. The unemployment rate in a country will fall as employment opportunities increase as one's educational level rises [6].

Students need to choose their desired courses based on their pre-university results to access public university. STPM and Matriculation have different study periods and recognition of foreign countries. Furthermore, STPM allows students to retake semesters 1 and 2 to improve their overall result, whereas Matriculation does not. There has been no research on students' university performance based on their types of pre-university examinations. Since students' pre-university performance is important for enrolment to university. As a result, more research into this topic is required in order to have a clearer figure for the university performance of students' academic performance based on their types of pre-university examination.

Academic achievement is influenced by a variety of factors, including socioeconomic position, student temperament and motivation, peer and family support [7]. Three objectives were discussed, including analysing the performance of BWQ students within their semesters' GPA with their pre-university and university CGPA by using paired t test, comparing the performance of BWQ students between their entry qualification such as STPM and Matriculation by using Z test as well as analysing and predicting the final grade by using CART, K- Nearest Neighbors and Naïve Bayes.

2. Materials and Methods

This section briefly explaining about the methods used in this study. The data collection, data pre-processing and the analysis used were explained in the following parts.

2.1 Data Collection

Data were collected from Universiti Tun Hussein Onn Malaysia (UTHM)'s Academic Department, Faculty of Applied Science and Technology (FAST). The data depicts the academic performance of students enrolled in the September 2018 intake batch. This is the most recent batch of BWQ graduates from the FAST Department with 54 students.

Table 1: Example of the Degree Classification variable

Variable	Example of Each Variable
First Class (1st)	CGPA greater than or equal to 3.70
Upper Second Class (U 2nd)	CGPA greater than or equal to 3.00 but lesser than 3.70
Lower Second Class (L 2nd)	CGPA greater than or equal to 2.30 but lesser than 3.00
Third Class (3rd)	CGPA greater than or equal to 2.00 but lesser than 2.30

Table 1 shows the example of the Degree Classification variable. First Class, Upper Second Class, Lower Second Class and Third Class are indicated by 1st, U 2nd, L 2nd and 3rd respectively.

Table 2: Sample of Data

Gender	Entry Qualification	Pre-U CGPA	Sem 1 GPA	Sem 2 GPA	Sem 3 GPA	Sem 4 GPA	Sem 5 GPA	Sem 6 GPA	Sem 7 GPA	Sem 7 CGPA	Degree Classification
F	M	2.38	2.52	2.48	2.35	2.8	2.19	3.23	3.33	2.82	U 2nd
M	STPM	2.81	2.98	2.61	3.27	3.13	2.65	2.92	3.64	3.03	L 2nd

Table 2 shows the example of the data. For gender, Female and Male are indicated by F and M respectively while higher certificate of Malaysian Education and Malaysian Matriculation Programme are represented by STPM and M respectively for Entry Qualification.

2.2 Data Pre-processing

Data cleansing is required because the data obtained from the FAST Academic Department contains missing data for the 42nd student in Year 2 Semester 2 and 21st student in Year 3 Semester 1. Because these students postponed his or her studies for one semester, the data will be deleted. Therefore, the total sample of the study is 52 students. Gender, entry qualification, Grade Point Average (GPA) of each semester, Cumulative Grade Point Average (CGPA) of final year and their degree classification are the variables used in this study.

Since the sampling distribution of the sample mean becomes more symmetric as sample size increases, the sample mean will be roughly equal to the population mean. When the sample size was greater than 30, the central limit theorem could be applied to the situations [8]. Because the sample size is greater than 30, the central limit theorem is used to determine the type of data.

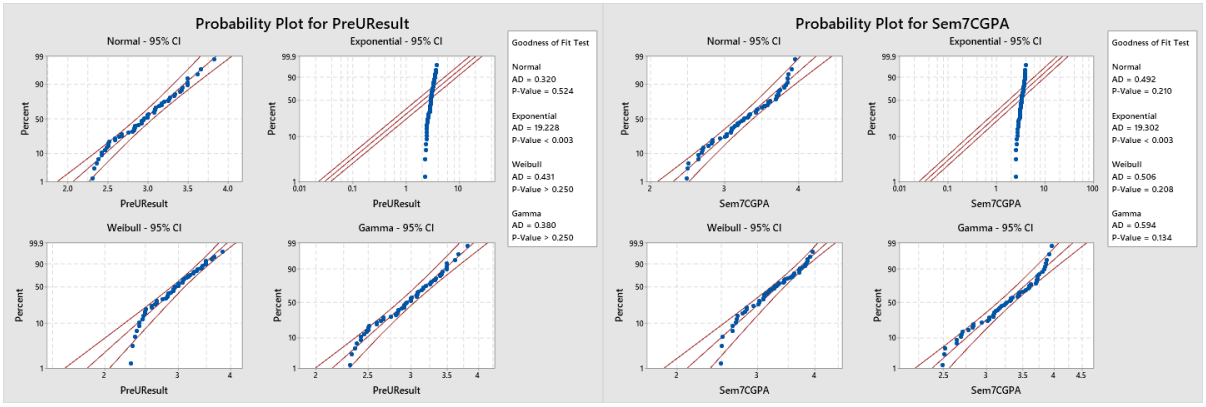

Figure 1: Probability plot for pre-university and university result in terms of CGPA

Figure 1 shows the four probability plots which includes Normal, Exponential, Weibull and Gamma distribution. Among the four plots, Normal distribution yields the smallest Anderson-Darling value with 0.320 and 0.492 for pre-university and university CGPA respectively, indicating that the pre-university and university result in terms of CGPA dataset fits best in Normal distribution.

2.3 Data Analysis

Statistical data analysis was performed using software such as Minitab, Microsoft Excel, Phyton and R Studio. This software is used to carry out the hypothesis testing of paired t-test and Z test as well as CART, K-Nearest Neighbors and Naïve Bayes.

2.4 Paired t test

A paired t-test is used to determine whether the assumption that excellent pre-university results will lead to excellent university results [9] is valid for BWQ intake September 2018 students. Equation for test statistics and critical values of paired t test are shown as the following:

$$t = \frac{\Sigma d}{\sqrt{\frac{n(\Sigma d^2)-(\Sigma d^2)}{n-1}}} \sim t_{\alpha,n-1} \qquad Eq.\ 1$$

Based on the Equation 1, d is the differences between previous result and the result afterwards while n is the sample size. Furthermore, α is the significant level and it is equal to 0.05. The null hypothesis states that two variables are independent and have no relationship, whereas the alternative hypothesis states that previous result will affects to the results in the future.

2.5 Z test

The Z test is used to determine whether the entry qualification affects university academic performance for BWQ intake September 2018 students. Equation for test statistics and critical values of Z test are shown as the following:

$$Z = \frac{\bar{x}_1-\bar{x}_2-\triangle}{\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}} \sim Z_\alpha \qquad Eq.\ 2$$

Based on the Equation 2, $\bar{x}$ is the estimated sample mean, $\triangle$ is the hypothesized difference between two samples, s is the estimated sample standard deviation while n is the sample size. Furthermore, α is the significant level and it is equal to 0.05. The null hypothesis states that two variables are independent and have no relationship, whereas the alternative hypothesis states that STPM students perform better than Matriculation Students.

2.6 Predictive Models

First, the data was divided into two categories: 80% of the training dataset and 20% of the testing dataset before the CART, K-Nearest Neighbors and Naïve Bayes models were carried out. For K-Nearest Neighbors to be carried out, the data for gender and entry qualification were transformed from categorical to binary data for the purpose of making them available to be analysed.

2.6.1 Classification and Regression Tree (CART)

A tree-structured classifier is a decision tree that predicts a class variable from one or more predictor variables. The CART is used to predict BWQ students' final year grade performance. CART uses the Gini index of node impurity to split the data in this study, and the formula is shown below [10]. Equation of Gini Index is shown as the following:

$$Gini(t) = 1 - \sum_{j=0}^{1}\left(\frac{n(j|t)}{n(t)}\right)^2 \qquad Eq.\ 3$$

Based on the equation above, t is the number of nodes, j is the class of target variable where $j = 0$ means STPM, $j = 1$ means Matriculation, $n(j|t)$ is the number of records of node t belongs to class j while $n(t)$ is the total record number in node t.

2.6.2 K-Nearest Neighbours

The K-Nearest-Neighbors (KNN) method is a straightforward but effective classification technique. When using KNN to classify, data normalization is required. Since the variables in this study are a combination of numerical and categorical, Euclidean Distance can be used to calculate the distance between the data points [11]. Equation for normalization and Euclidean Distance are shown as the following:

$$X_s = \frac{X - Min}{Max - Min} \qquad Eq.\ 4$$

$$d_{(x,y)} = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \qquad Eq.\ 5$$

Based on Equation 4, X_s is the standardized distance, X is the distance Min is the minimum distance in the training set while Max is the maximum distance in the training set. For Equation 5, d is the distance between two points, x is the position of x coordinate, y is the position of y coordinate while n is the number of points.

2.6.3 Naive Bayes

The Naïve Bayes classifier greatly simplifies learning by assuming that features are independent of class and to calculate the likelihood of one event occurring given another, conditional probabilities were required [12]. Equation for conditional probabilities is shown as the following:

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)} \qquad Eq.\ 6$$

Based on the equation above, $P(A|B)$ is the probability of A happens given that B happens, $P(B|A)$ is the probability of B happens given that A happens, $P(A)$ is the probability of A happens, $P(B)$ is the probability of B happens while $(b_1, b_2, b_3, \cdots, b_n)$ are the features of B.

2.7 Confusion Matrix

Since classification accuracy is the most important concern in classification problems required. To calculate the Overall Accuracy, divide the total number of correctly classified elements by the total number of elements in the confusion matrix. Equation for accuracy is shown as the following:

$$Accuracy = \frac{Tp + Tn}{Tp + Tn + Fp + Fn} \qquad Eq.\ 7$$

Based on the equation above, Tp is known as the number of cases belongs to a class and actually belong to it while Fp is known as number of cases belong to a class but reality does not. Oppositely, Tn is explained on the number of cases does not belong to a class and actually does not belong to it but Fn explained in it does not belong to a class but reality it does.

3. Results and Discussion

This section presents the descriptive analysis and results obtained by using paired t test, Z test, CART, K-Nearest Neighbors and Naïve Bayes.

3.1 Descriptive Analysis

The descriptive analysis was used in this study to look into the overall picture of the data.

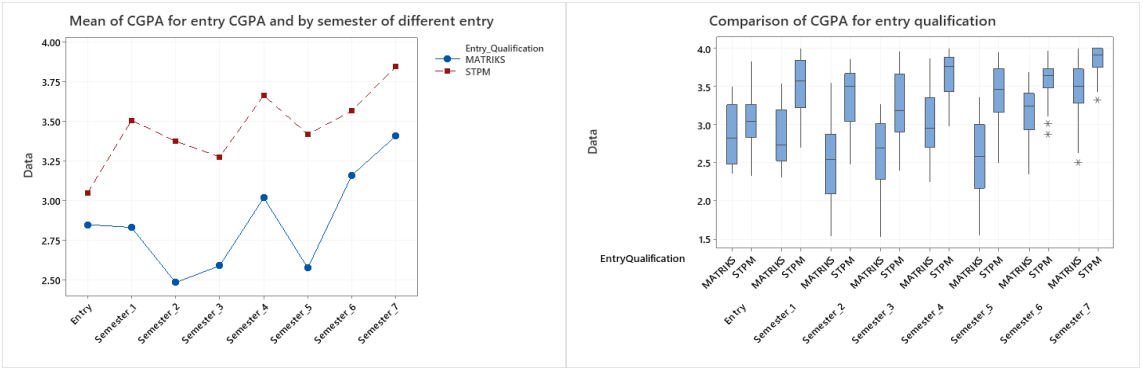

Figure 2: Line Chart and Boxplot of CGPA for entry qualification

Figure 2 represents the visualization of the performance based on their entry qualification and boxplot comparison of STPM and Matriculation students' performance for each semester. It shows that the results for STPM are better than the Matriculation students throughout the whole semesters. It also gives some information that students have decreased their performance on Semester 2 and Semester 5 but they catch up again on their final semester. Also, it demonstrates that most STPM students have a CGPA and GPA greater than 3.00. For Matriculation students, the median CGPA and GPAS ranging from 2.50 to 3.00. However, it has increased in the last two semesters, with Matriculation students' CGPA and GPAs ranging from 3.00 to 3.50.

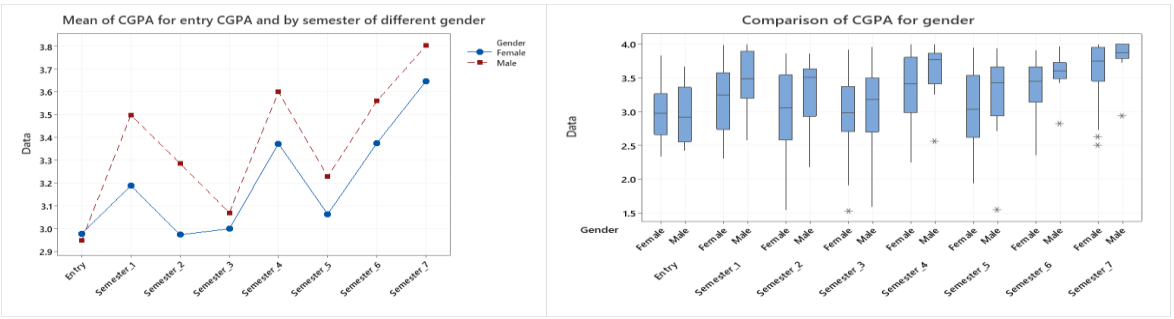

Figure 3: Line Chart and Boxplot of CGPA for gender

Figure 3 represents a gender performance comparison for each semester and boxplot comparison of male and female students' performance for each semester. This visualization, however, cannot be a fair comparison because there are three times as many female students as male students. As a result, male performance is superior to female performance, the results will be skewed. Also, it demonstrates that most male students have a CGPA and GPA greater than 3.50. For the result obtained shows that male outperform female students and it is probably due to the huge different in sample size which cannot support for the comparison between male and female students.

3.2 Analysis on students' performance in terms of the effects of previous result

The paired t test was also used to examine how the performance of BWQ students' final year grades is affected by their pre-university and how the performance of BWQ students' next semester is affected by the previous semester result.

Table 3: Paired t test for BWQ students

Paired t test							
Pre	Mean	SD	Post	Mean	SD	*p*-value	Decision
Pre-U CGPA	2.9613	0.3866	University CGPA	3.2811	0.4141	0.000	Do not reject H_0
Sem 1 GPA	3.2165	0.6472	Sem 2 GPA	2.9887	0.6472	0.000	Do not reject H_0
Sem 2 GPA	2.9887	0.6472	Sem 3 GPA	2.9656	0.6062	0.657	Reject H_0
Sem 3 GPA	2.9656	0.6062	Sem 4 GPA	3.3807	0.5013	0.000	Do not reject H_0
Sem 4 GPA	3.3807	0.5013	Sem 5 GPA	3.0630	0.6004	0.000	Do not reject H_0
Sem 5 GPA	3.0630	0.6004	Sem 6 GPA	3.4024	0.3639	0.000	Do not reject H_0
Sem 6 GPA	3.4024	0.3639	Sem 7 GPA	3.6609	0.3663	0.000	Do not reject H_0

Based on the Table 3, the assumption of better previous result will also bring better result later is not applicable for BWQ students as the result only shows that good semester 2 result will also bring good semester 3 result, while the others are not affected by the previous result.

3.3 Analysis on students' performance in terms of their entry qualification

The Z test is then used to examine the relationship between BWQ students' academic performance and entry qualification.

Table 4: Z test for Entry Qualification

Entry Qualification				
	Mean	SD	*p*-value	Decision
Pre-U CGPA	2.9613	0.3866	0.0453	Reject H_0
Sem 1 GPA	3.2165	0.6472	0.0000	Reject H_0
Sem 2 GPA	2.9887	0.6472	0.0000	Reject H_0
Sem 3 GPA	2.9656	0.6062	0.0000	Reject H_0
Sem 4 GPA	3.3807	0.5013	0.0000	Reject H_0
Sem 5 GPA	3.0630	0.6004	0.0000	Reject H_0
Sem 6 GPA	3.4024	0.3639	0.0000	Reject H_0
Sem 7 GPA	3.6609	0.3663	0.0060	Reject H_0
University CGPA	3.2811	0.4141	0.0000	Reject H_0

Based on the result obtained as shown in Table 4, there is a difference between student's performance in term of entry qualification. The result indicates that STPM outperform Matriculation students.

3.4 Predictive Models

3.4.1 Classification and Regression Tree

From CART result obtained, the most contributed variable is result of GPA in the first semester. It shows that if the students want to graduate in First Class, he/she need to get higher GPA since semester 1 and maintain the results throughout the whole semester.

From the CART, it is predicted that if the students get more than 3.00 GPA in semester 1, they will probably graduate with Lower Second Class. However, if the students got less than 3.00 GPA in semester 1 and got less than 4.00 GPA in semester 7, the student can graduate with Upper Second Class. Otherwise, the student will graduate with Lower Second Class.

Table 5: Confusion Matrix for Classification and Regression Tree

Test_actual	*Test_pred*			Row total
	First Class	Lower Second Class	Upper Second Class	
First Class	4	0	0	4
Lower Second Class	0	4	1	5
Upper Second Class	0	0	2	2
Column Total	4	4	3	11

However, this finding had some issues where the real situation is supposed if the student got more than 3.00 GPA in semester 1, then the student will be graduated with Upper Second Class. This result is not 100% accurate because of this misclassification issue and it can be proven by findings in Table 5 which shows 1 misclassified where the student supposed to be graduated with Upper Second Class, but it predicts that this student will be graduated with Lower Second Class. This might happen due to the small data set (52 samples) as CART mostly efficient when the data is large (more than 100 samples). In this study, the accuracy of CART is only 90.91% as there is overall 0.09% of misclassification rate in predicting the students' degree classification for the testing dataset.

3.4.2 K-Nearest Neighbors

Table 6: Confusion Matrix for K-Nearest Neighbor

Test_actual	*Test_pred*			Row total
	First Class	Lower Second Class	Upper Second Class	
First Class	2	0	0	2
Lower Second Class	0	2	0	2
Upper Second Class	0	0	7	7
Column Total	2	2	7	11

The confusion matrix as shown in Table 6, predicts 7 students got Upper Second Class, and 2 students are predicted to graduate with Lower Second Class and First Class. There is no misclassification issue from K-Nearest Neighbors result. As a result, the overall misclassification rate in predicting the students' degree classification for the testing dataset is 0%.

3.4.3 Naïve Bayes

The Naïve Bayes Classifier was also tested to compare its performance in predicting student degree classification. This method computed the probability for a class based on various feature values such as Gender, Entry Qualification, CGPA of Pre-university, GPA of Semester 1, GPA of Semester 2, GPA of Semester 3, GPA of Semester 4, GPA of Semester 5, GPA of Semester 6, GPA of Semester 7, and final CGPA.

Table 7: Confusion Matrix for Naïve Bayes

Test$Degree_Classification	*Test_pred*			Row total
	First Class	Lower Second Class	Upper Second Class	
First Class	2	0	0	2
Lower Second Class	0	2	0	2
Upper Second Class	0	0	7	7
Column Total	2	2	7	11

From Table 7, it shows that there is no misclassified of the final year grade. All the prediction is same as the real data. This model achieved 100% accuracy, indicating that it is a good sensitivity model for predicting the probability of students' academic performance with no students misclassified. Thus, for the testing dataset, the misclassification rate in predicting the probability of students' academic performance in terms of Degree Classification is 0%.

3.5 Discussion

The three predictive data mining models are analyzed and compared in this study based on their accuracy value. The accuracy of the three different predictive data mining models is shown in Table 8.

Table 8: Accuracy for the models

Predictive Model	Accuracy (%)
Classification and Regression Tree	90.91
K-Nearest Neighbor	100
Naïve Bayes	100

A classifier's accuracy was calculated by Equation 9. Based on Table 8, CART shows the minimum efficient result with the lowest percentage of accuracy of 90.91% while K- Nearest Neighbors and Naïve Bayes share the highest percentage of accuracy of 100 %. Small sample size of the data set brings the result of K- Nearest Neighbors and Naïve Bayes outperform CART.

4. Conclusion

According to the results of the paired t test, only the result in semester 2 will affects the result in semester 3, supporting the assumption that excellent previous results will also bring excellent results later, while the others do not have any effect. Meanwhile, the results of Z test shows that the entry qualification does matter for students' performance, as the Z test results show that STPM students outperform Matriculation students.

Three of the models mentioned were used to achieve the last objective, which is to analyse and predict the final grade by using CART, K-Nearest Neighbors, and Nave Bayes. The model built with CART was found to be the least efficient data mining technique in predicting students' performance in terms of degree classification, with the lowest percentage of accuracy of 90.91%, while K-Nearest Neighbors and Nave Bayes had the highest percentages of accuracy of 100%. Data mining has the potential to have a greater impact on students' performance-related investigations; thus, the models developed would be extremely useful in the academic field for students to achieve better results. Some suggestions should be included to encourage the continuation of the current research. K-Nearest Neighbors and Nave Bayes outperform classification and regression trees due to the small sample size of the data set.

For future research, a larger data set should be used to analyse students' performance within their semester GPA as well as pre-university and university CGPA. Other variables that can be included for future research are students’ family background, parents’ occupation, learning skills, parental background, peer influence, teachers' quality, learning infrastructure and lastly there should not have any gap between gender in terms of sample size.

Acknowledgement

The authors would also like to express their gratitude to the Faculty of Applied Sciences and Technology at Universiti Tun Hussein Onn Malaysia for their assistance.

References

[1] Senarai UA Universiti Awam Di Malaysia - Info UPU, 2022. [Online]. Available: https://www.infoupu.com/senarai-ua-universiti-awam-di-malaysia/. [Accessed Jan 8, 2023]

[2] Department of Statistics Malaysia Official Portal. [Online]. Available: https://www.dosm.gov.my/v1/index.php?r=column/cthree&menu_id=Z1hCMUVLQWVOL2tScVlhSmo5eEd3QT09

[3] Wanka, F. A., "The Impact of Educational Attainment on Household Poverty in South Africa: A Case Study Of Limpopo Province", May 2014. [Online]. Available: https://core.ac.uk/download/pdf/58915334.pdf.

[4] Mahidin, D. S. D. M. U., Department of Statistics Malaysia Official Portal. Department of Statistics, Malaysia (DOSM), 2021. [Online]. Available: https://www.dosm.gov.my/v1/index.php?r=column/cthemeByCat&cat=476&bul_id=U1ltVWpwNXRNRUR2NlhRSHZmenRMUT09&menu_id=Tm8zcnRjdVRNWWlpWjRlbmtlaDk1UT09

[5] Vasyl Taras, Grishma Shah, Marjaana Gunkel, and E. T, " Graduates of Elite Universities Get Paid More. Do They Perform Better? Harvard Business Review", 2020. [Online]. Available: https://hbr.org/2020/09/graduates-of-elite-universities-get-paid-more-do-they-perform-better

[6] Ahmed Manaa and Muhammad Abrar ul Haq., "View of The Effects of SMEs, Population and Education level on Unemployment in Kingdom of Bahrain", 2021.[Online]. Available: https://iksp.org/journals/index.php/ijbe/article/view/45/84

[7] Masud, S. etal., "Academic Performance in Adolescent Students: The Role of Parenting Styles and Socio-Demographic Factors – A Cross Sectional Study from Peshawar", Frontiers in Psychology, 2019. [Online]. Available: https://doi.org/10.3389/FPSYG.2019.02497/BIBTEX

[8] Islam, M. R., "Sample size and its role in Central Limit Theorem (CLT)", Computational and Applied Mathematics Journal, 2018. [Online]. Available: http://www.aascit.org/journal/camj

[9] Alyahyan, E., and Düştegör, D., "Predicting academic success in higher education: literature review and best practices" International Journal of Educational Technology in Higher Education, 2020. [Online]. Available: https://doi.org/10.1186/S41239-020-0177-7/TABLES/15

[10] Loh, W. Y., "Tree-structured classifiers", Wiley Interdisciplinary Reviews: Computational Statistics, 2010. [Online]. Available: https://doi.org/10.1002/WICS.86

[11] Guo, G., Wang, H., Bell, D., Bi, Y., and Greer, K., "KNN model-based approach in classification", Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2003. [Online]. Available: https://doi.org/10.1007/978-3-540-39964-3_62/COVER

[12] Zhang, Z., "Naïve Bayes classification in R". Annals of Translational Medicine, 2016. [Online]. Available: https://doi.org/10.21037/ATM.2016.03.38