Predictive Analytics on Academic Performance in Higher Education Institution during COVID-19 using Regression Model

Muhammad Farezuan Zulfikri[1], Shazlyn Milleana Shaharudin[1]*,
Noor Azrin Abdul Rajak[1], and Muhammad Safwan Ibrahim[2]

[1]Department of Mathematics,
Faculty of Science and Mathematics,
Universiti Pendidikan Sultan Idris,
35900 Tanjong Malim, Perak, Malaysia

[2]Faculty of Science and Technology,
Universiti Sains Islam Malaysia,
71800 Bandar Baru Nilai, N.Sembilan, Malaysia

Abstract- The coronavirus disease 2019 (COVID-19) outbreak in December 2019 had affected the way of living for people around the world including students in educational institutions. These students had to prepare for the continuity of their study mentally and physically by adapting to the online teaching and learning approach, which can significantly impact their academic performance. Hence, this study examines the students' academic performance on the online teaching and learning approach, thus predicting their academic performance for the upcoming semester. This study enables various actions to be taken in improving and maintaining the student's performance in their study activities. The study was conducted on undergraduate students from the Faculty of Science and Mathematics, Universiti Pendidikan Sultan Idris. This paper presents the prediction of the student's academic performance with a linear regression model. Evidently, the result shows that the student's academic performance continually improves while adapting to the online teaching and learning approach. It also shows that there were few respondents affected while adapting to this new norm approach. Hence, the future development of the regression model can be improved by having a more comprehensive range of Malaysian universities' data.

Keywords- Student's Academic Performance, Prediction, Grade Point Average, Linear Regression Model, COVID-19

I. Introduction

The novel coronavirus, which was first reported by the Wuhan Municipal Health Commission of the Republic of China in December 2019, had been pronounced by the World Health Organisation (WHO) as a pandemic. The virus is a super contagious disease and named COVID-19 that could trigger respiratory problems to the affected person. Consequently, the virus had been spreading and infecting all countries around the globe. Malaysia's first detected case was on January 24, 2020 [1]. Since then, the cumulative cases of COVID-19 in Malaysia had reached 50,030 cases as of November 2020 [2]. Malaysian Control Order (MCO) has been implemented since March 18, 2020, by the government to prevent and stop the chain of spreading this contagious virus [3]. The MCO is mainly an inclusive ban of mass gatherings and movements such as social, sports, religious, and cultural events. Educational institutions including schools and universities were also affected by the order and forced to be closed [4].

The implementation of the MCO has restricted educational institutions from conducting physical teaching and learning activities. Therefore, universities have to adopt the new norm namely online learning to allow educational activities to continue. The Ministry of Higher Education (MOHE) Malaysia had proposed and suggested that every higher educational institution conducts its educational activities using an online platform on May 27, 2020 [5]. Responding to the government's call, Universiti Pendidikan Sultan Idris (UPSI) took the initiative and implemented fully online teaching and learning activities. Online-based learning is a digital teaching and learning via technology such as Zoom, Google Meet, MOOC, YouTube, and Webex [6, 7]. The materials of

teaching and learning activities are also varied in such a way as visual graphics, words, animations, video or audio to enable students' accessibility to information.

Although some educators use blended learning as their teaching and learning pedagogical approach, fully online learning activities are scarce in Malaysia. However, during these unprecedented times and the occupancy of MCO, all educational institutions have the urgency to execute and find the alternatives by implementing a fully online teaching and learning approach. Online learning has affected the students in many ways such as decreased self-motivation, unorganized study time, unconducive learning environment, perplexity among family members, and various technical problems that lead to online overloaded [8]. Hence, due to online learning obstacles, students have been struggling to maintain their academic performance, which can affect their future.

There are various types of predictive models that can be applied to determine factors affecting a student's academic performance. The use of these models is to predict the chosen factors by observing their patterns, discovering the trends and later, the outcome will be drawn. Some of the well-known predictive models are decision tree, singular spectrum analysis, support vector machine, and linear regression [9, 10, 11]. These models have a different style of an algorithm to produce an outcome. The predictive model implemented in this study was the Linear Regression Model (LR). The regression model is a statistical analytic model that can help to predict a trend of data. The LR model can be used in various applications such as trend analysis, business planning, marketing, prediction of finance, time series prediction, biomedical and drug response modelling and environmental modelling. Linear regression assumes that the relationship between x and y as a straight line (linear). This model has the simplest form of a predictive model using the formula of a straight line ($y = mx + c$). The independent variable x is called a predictor while the dependent variable y is a response. The application of this model towards predicting academic performance had been widely used since the model is easily understandable without needing a mastery of mathematical skills [12, 13].

This paper will focus on predicting student's Grading Point Average (GPA) using the Linear Regression Model. By observing the results, one can help to improve their teaching and learning approach such as improvising learning methods, using suitable learning materials and making better preparation for teaching and learning activities.

II. DATA

For this study, the data employed were the Grade Point Average (GPA) of the undergraduate students, who are currently in semester 3 to semester 7, from the Faculty of Science and Mathematics, Universiti Pendidikan Sultan Idris. The data collected through a survey using an online platform, particularly, in Google Forms. From the data, 82.5% are female, and 17.5% are male,

where 66% of them are Mathematics students, 7% are Biology students, 15% are Chemistry students, 4% are Physics students, and 8% are Science students.
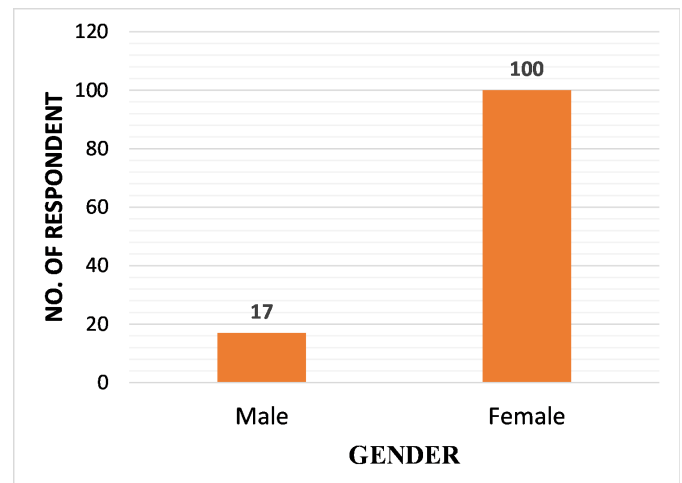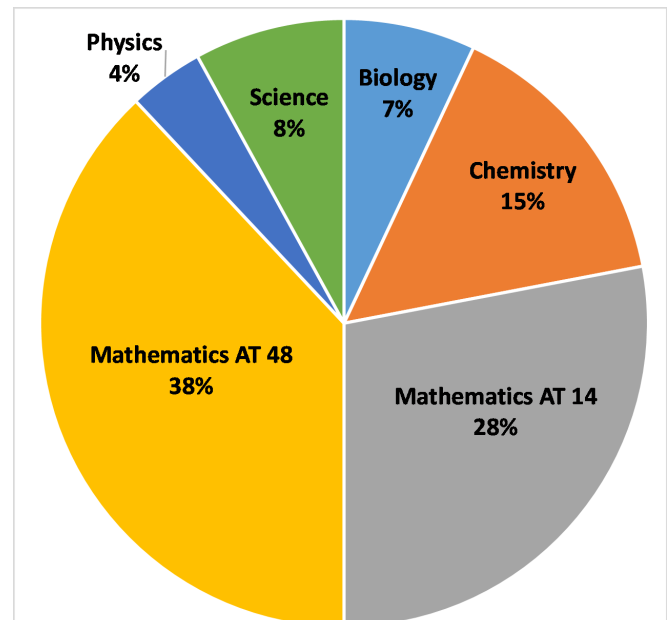

Fig. 1: Student's Gender

Fig. 2: Student's Program

III. METHODOLOGY

A. Regression

Regression is one of the analytical models that helps in fitting an equation into a dataset. The simplest form of regression is in the form of a straight line using the formula $y = ax + b$. The values of a and b need to be determined to get the equation line, to give the results of the prediction of y that depends on variable x. The regression task started with a known value of the target. A regression model is not commonly used for academic purposes; therefore, not many researchers use this model to predict scholastic achievements [14]. However, the linear regression model will be applied in this research to

project undergraduate students' academic performance after online teaching and learning activities.

B. Methodology of Regression Model

To construct a linear regression model, the values of the parameters need to be determined for the use of the prediction model. Symbols are used to represent the relationship of the equation. The process of regression is to estimate the value of a continuous target (y) as a function (F) which depend on the variable of the predictors $((x^1, x^2, x^3, \ldots, x^n)$, set of parameters $(\beta^1, \beta^2, \ldots, \beta^n)$ and error (e) as in Equation (1) [15].

$$Y = F(x, \beta) + e \tag{1}$$

The independent variables (x) refer to the predictors and the dependent variable (y) refers to the target. The error is known as residual, obtained by the discrepancy between the expected and predicted value of the predictor. The parameter of regression is known as regression coefficients. The regression model training process was involved in obtaining the parameter values that will minimize the measure of the error, for example, the sum of squared error.

C. Linear Regression

The use of the Linear Regression Model is observing the relationship between the independent variables (predictor) and dependent variable (target) [12]. The model can be represented using a straight-line equation as follows:

$$y = \beta^2 x + \beta^1 + e \tag{2}$$

The slope of the line (β^2), and y-intercept (β^1) are the two parameters in linear regression which (β^2) is the angle between a data point and the regression line while (β^1) is the point of x crosses y-axis $(x = 0)$.

D. Coefficient of Determination

R^2 is denoted as the coefficient that acts as a vital output of a regression analysis. The form of this coefficient is in a proportion variance in the dependent variable that is predicted from the stated variable. The detail of this coefficient is given as follows:

- The range of coefficient is between 0 to 1.

- If $R^2 = 0$, the dependent variable cannot be predicted from independent variable or else the prediction is depends on error, e.

- If R^2 between 0 to 1, the dependent variable is predictable. For example, R^2 of 0.10 indicates 10% of variance in Y is predictable from X.

E. Prediction Model

The prediction model needs to be constructed to predict the academic performance of the students. In this study, academic performance can be observed from a student's GPA, which will be the target of the prediction (independent variable, y). Age was chosen to be the independent variable x. The prediction model is based on Equation (2).

IV. RESULTS AND DISCUSSION

Regression analysis on the data of the undergraduate students was performed using XLSTAT. XLSTAT can perform various data analyses, presentation of functions, statistical analysis and graphical data depiction. The summary statistics shown in Table 1 describes the target (dependent variable, y), which is the GPA, and predictors (independent variable, x), which is the AGE.

Table 1: Summary Statistics

Variable	Observations	Obs. with missing data	Obs. without missing data
GPA	227	0	227
AGE	227	0	227

Minimum	Maximum	Mean	Std. deviation
2.050	5.450	3.796	0.263
20.000	26.000	21.863	1.497

From Table 2, R^2 is 0.068, which can be translated as 6.8% of the variance is predictable using the variable age. The Root Mean Square Error (RMSE) signifies the expected error when using the variable age is 0.254 years, equivalent to 3 months. Given the p-value of the F statistic generated in Table 3, with a significance level of 5%, the information from the explanatory variables is substantially better than that of a basic mean.

Table 2: Goodness of fit statistics (GPA)

Statistics	Training set	Validation set
Observations	227	1
Sum of weights	227	1
DF	225	-1
R^2	0.068	
Adjusted R^2	0.064	
MSE	0.065	
RMSE	0.254	
MAPE	4.597	0.0000
DW	1.724	
Cp	2.000	
AIC	-619.531	
SBC	-612.681	
PC	0.949	

Table 3: Analysis of variance (GPA)

Source	DF	Sum of squares	Mean squares	F	$Pr > F$
Model	1	1.06	1.06	16.42	<0.0001
Error	225	14.56	0.07		
Corrected Total	226	15.62			

Table 4: Model Parameters (GPA)

Source	Value	Standard error	t
Intercept	4.797	0.248	19.373
AGE	-0.046	0.011	-4.052

$Pr > |t|$	Lower bound (95%)	Upper bound (95%)
<0.0001	4.309	5.285
<0.0001	-0.068	-0.024

From Table 4, the value of parameters of the slope of the line $\beta^2 = 4.797$, and y-intercept $\beta^1 = -0.046$. The equation of the prediction model of the GPA is $y = 4.797x - 0.046$. The model was then used to construct a prediction of the students' GPA collected from the surveys. The results are shown in Figure 3.
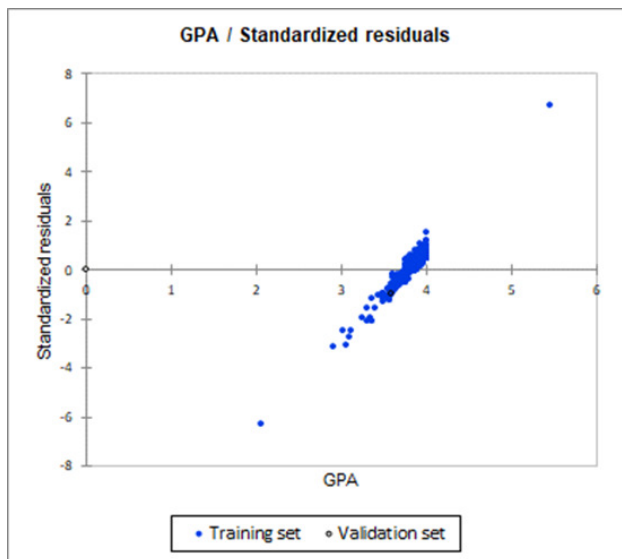

Fig. 3: Scatter plot shown of the actual GPA and standardized residuals

The scatter plot shown in Figure 3 explains the trend of student's GPA from their past results. This trend demonstrates that the GPA is higher throughout the year.

From the scatter plot in Figure 4, the undergraduate students' GPA had increased drastically since the scatter plot cumulated at the right side of the graph. There is a slight trend that the undergraduate student's GPA will increase over time since the results show that most students either maintain or get a higher GPA for the upcoming semester. There is only a slight error for the prediction model; hence the prediction model is still reliable to predict the student's GPA.
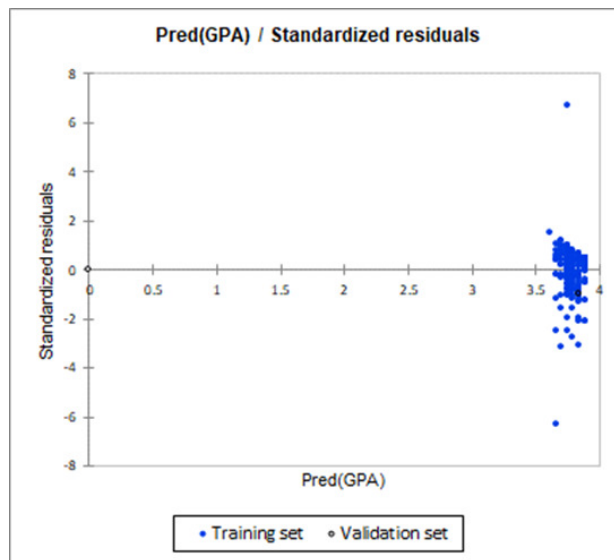

Fig. 4: Scatter plot shown of the prediction GPA and standardized residuals
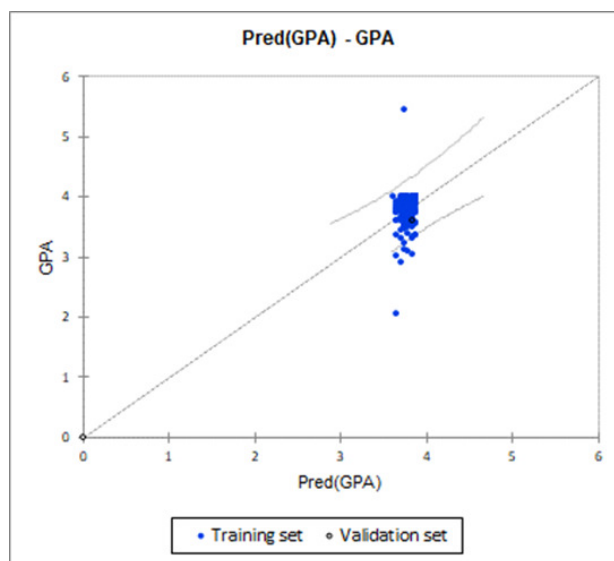

Fig. 5: Scatter plot shown of the error of actual GPA-prediction GPA

Based on the results, by using the linear regression model, the result is suitable for predicting the student's academic performance since it shows significant improvement of student's GPA through online teaching and learning activities during the MCO. From the data collected, the model could be constructed using the linear regression method by determining the variables for the equation. By referring to Figure 3, the linear re-

gression model for the prediction using predictor AGE showed that students' academic performance would increase through their upcoming semester. The graph is inclined to positive infinity and shows a significant increase in the student's academic performance. Figure 4 shows that the graph is skewed to the right and most students are predicted to have higher GPAs. It indicates that the students can obtain higher results in their upcoming semester although they have to learn through an online platform.

Certain reasons would cause an increase in students' GPA result even though they have to go through online teaching and learning approach during the MCO period. One of the reasons is that students feel safe and at ease during class because they have more time to prepare for themselves before starting their class. Most lecturers will have a synchronous class which means students need to attend the class using online platforms such as Google Meet, MOOC, Zoom and Microsoft Teams; or asynchronous classes that require them to watch videos prepared by their lecturers. Due to this, students have more courage to ask questions without worry or self-conscious about what their peers will think. They will be more motivated to gain more knowledge since they do not have to worry about their feelings of afraid to be judged by others. Besides, students could expand their knowledge on using the technology to help them learn and complete their tasks. Most of the lecturers did give tasks or assignments to the students to brush up on their knowledge. However, since the MCO period, students cannot move around and meet their group mates, thus students need to have other alternatives to complete the task given. Online group platforms such as Google Slides, Google Docs, and Powtoon can help students cooperate to complete their tasks. They will be more competent in completing the task and perform the submission on time, which leads to getting a good grade.

On the other hand, some students cannot keep up with online teaching and learning activities, resulting in a drop in their GPA. They could have a lack of motivation to go through the online teaching and learning processes. The lack of access to the internet can be one of the reasons students have a hard time attending their online class, completing the tasks given, which results in a course fail. For this reason, lecturers need to keep revising and improving their teaching and learning approach and prepare suitable material to help students engage with the course. Students also need to be more focused and ready to learn during online classes so that they will not be left behind.

Online teaching and learning activities will not affect a student's study badly; instead, it helps students improve their academic performance. Implementing an online teaching and learning approach will be a great way to ensure the continuity of teaching and learning activities of the undergraduate students during the MCO period.

V. Conclusion

The purpose of applying a statistical method is to improve students' academic performance by predicting their academic results to provide the necessary aid. Using the traditional way of the predicating method will only take a longer time since there are many complexities of the learning environment and a variety of learning tools that can limit the study. In this study, a linear regression model was implemented to project students' academic performance based on the current semester of their GPA. The student's current GPA (online learning) was compared with the previous GPA result (face-to-face learning) from the past semester. It can be concluded that online learning will not affect a student's academic performance since majorities of them gain benefits from it which increases the student's academic performance. Conclusively, this linear regression model can be applied as a generic tool to predict a student's academic performance.

Acknowledgment

References

[1] World Health Organization: *Novel Coronavirus (2019-nCoV) Advice for the Public*, https://www.who.int/emergencies/diseases/novel-coronavirus-2019/advice-for-public (2020).

[2] Ministry of Health Malaysia: *Situasi Semasa Pandemik Covid-19 Di Malaysia*, http://covid-19.moh.gov.my/.

[3] M.S. Samsudin, A. Azid, N. L. A. Rani, K. M. K. K. Yusof, S. M. Shaharudin, K. Yunus, M. A. Razik, M. H. Sidik, N. M. Rozar: *Evidence of Recovery from the Restriction Movement Order by Mann-Kendall during the COVID-19 Pandemic in Malaysia*, Journal of Sustainability Science and Management, **16** (2020), 55-69.

[4] M. F. M. Fuad, S. M. Shaharudin, S. Ismail, N. A. M. Samsudin, M. F. Zulfikri: *Comparison of singular spectrum analysis forecasting algorithms for student's academic performance during COVID-19 outbreak*, International Journal of Advanced Technology and Engineering Exploration, **8**(74) (2021), 178 – 189.

[5] A. Shahzad, R. Hassan, A. Y. Aremu, A. Hussain, R. N. Lodhi: *Effects of COVID-19 in E-learning on higher education institution students: the group comparison between male and female*, Quality & Quantity, (2020).

[6] L. Mishra, T. Gupta, A. Shree: *Online teaching-learning in higher education during lockdown period of COVID-19 pandemic*, International Journal of Educational Research Open, **1** (2020), 100012.

[7] S. Maity, T. N. Sahu, N. Sen: *Panoramic view of digital education in COVID-19: A new explored avenue*, Review of Education, (2020).

[8] A. P. Aguilera-Hermida: *College students' use and acceptance of emergency online learning due to COVID-19*, International Journal of Educational Research Open, **1** (2020), 100011.

[9] S. M. Shaharudin, S. Ismail, M. L. Tan, N. S. Mohamed, N. A. F. Sulaiman: *Predictive Modelling of Covid-19 Cases in Malaysia based on Recurrent Forecasting-Singular Spectrum Analysis Approach*, International Journal of Advanced Trends in Computer Science and Engineering, **9**(1.4) (2020), 175 − 183.

[10] N. A. F. Sulaiman, S. M. Shaharudin, N. H. Zainuddin, S. A. M. Najib: *Improving Support Vector Machine Rainfall Classification Accuracy based on Kernel Parameters Optimization for Statistical Downscaling Approach*, International Journal of Advanced Trends in Computer Science and Engineering, **9**(1.4) (2020), 652 − 657.

[11] M. N. R. Ayán, M. T. C. García: *Prediction of University Students' Academic Achievement by Linear and Logistic Models*, The Spanish Journal of Psychology, **11**(1) (2008), 275 − 288.

[12] R. R. Rajalaxmi, P. Natesan, N. Krishnamoorthy, S. Ponni: *Regression Model for Predicting Engineering Students Academic Performance*, International Journal of Recent Technology and Engineering (IJRTE), **7**(6S3) (2019), 71 − 75.

[13] P. M. Arsad, N. Buniyamin, J. A. Manan: *Prediction of engineering students' academic performance using Artificial Neural Network and Linear Regression: A comparison*, Proceeding of the 2013 IEEE 5th Conference on Engineering Education (ICEED), (2013), 43 − 48.

[14] S. J. H. Yang, O. H. T. Lu, A. Y. Q. Huang, J. C. H. Huang, H. Ogata, A. J. Q. Lin: *Predicting Students' Academic Performance Using Multiple Linear Regression and Principal Component Analysis*, Journal of Information Processing, **26** (2018), 170 − 176.

[15] P. K. Srimani, M. M. Patil: *Regression Model for Edu-data in Technical Education System: A Linear Approach*, ICT and Critical Infrastructure: Proceedings of the 48th Annual Convention of Computer Society of India- Vol II, (2014), 785 − 793.