Modeling Student's Academic Performance During Covid-19 Based on Classification in Support Vector Machine

Nor Ain Maisarah Samsudin[1], Shazlyn Milleana Shaharudin[2], Nurul Ainina Filza Sulaiman[3], Muhammad Fakhrullah Mohd Fuad[4], Muhammad Fareezuan Zulfikri[5], Nurul Hila Zainuddin[6]

[1,2,3,4,5,6]Department of Mathematics, Faculty of Science and Mathematics, Universiti Pendidikan Sultan Idris, Malaysia.

Abstract: This study proposed a statistical investigate the pattern of students' academic performance before and after online learning due to the Movement Control Order (MCO) during pandemic outbreak and a modelling students' academic performance based on classification in Support Vector Machine (SVM). Data sample were taken from undergraduate students of Faculty of Science and Mathematics, Universiti Pendidikan Sultan Idris (UPSI). Student's Grade Point Average (GPA) were obtained to developed model of academic performances during Covid-19 outbreak. The prediction model was used to predict the academic performances of university students when online classes was conducted. The algorithm of Support Vector Machine (SVM) was used to develop a model of students' academic performance in university. For the Support Vector Machine (SVM) algorithm, there are two important parameters which are C (misclassification tolerance parameter) and epsilon need to identify before proceed the further analysis. The parameters was applied to four different types of kernel which is linear kernel, radial basis function kernel, polynomial kernel and sigmoid kernel and the result was found that the best accuracy achieved by SVM are 73.68% by using linear kernel and the worst accuracy obtained from a sigmoid kernel which is 67.99% with parameter of misclassification tolerance C is 128 and epsilon is 0.6.

Keywords: Students' Academic Performance, Prediction, Grade Point Average, Support Vector Machine;

Introduction

The global higher education landscape has dramatically changed for the past few months due to the spread of the coronavirus (COVID-19) [1]. In December 2019, a pneumonia of unknown etiology was reported in Wuhan, China. On 31 December 2019, the outbreak was trace to a novel strain of coronavirus [2]. Malaysia is also not immune from the COVID-19 epidemic. The positive rate of Covid-19 Malaysian patients initially in a controlled rate declined where the number of patients increased with the arrival of foreigners to Malaysia [3]. On 16 March 2020, the Prime Minister, Tan Sri Muhyiddin Yassin declared the Movement Control Order (MCO) in Malaysia [4]. The effect of this MCO has indirectly caused schools and institutions of higher learning institutions not to be allowed to run during the MCO period [5]. On 27 May 2020, Ministry of Higher Education (MOHE) informed that all teaching and learning activities for students must be implemented online until 31 December 2020 [6]. Universiti Pendidikan Sultan Idris (UPSI) will be implementing online learning and teaching (PdP) in line with Higher Education Ministry (KPT)'s directive [7].

Face-to-face learning is a process of learning and teaching directly and indirectly between teachers or lecturers and students. Face-to-face learning is also known as conventional learning [8]. Online learning can be defined as that learning occurs partially or entirely through internet access [9]. Online learning is not stranger among university students. However, conducting the concept of digital lectures as a whole due to the pandemic outbreak is seen to have a huge impact on university students, especially on academic achievement [10]. But in [11], Smith and Stephens found that online students tended to have higher pre-course GPA as a positive, since online learning required discipline and self-motivation. Many studies have been conducted to study online learning[12], but it is very difficult to find studies related to online learning during pandemics. In [13], online learning has effect to the students learning outcomes where student interaction during online learning has a significant effect on student academic results. Recent studies show that interactive activities are one of the factors that affect student results [14]. In addition, Oye, et al., and Keshavarz said that e-learning gives a positive impact on students' academic achievement because of the reduces costs, saving time and increase accessibility of education [15].Shahiri et al. [16] stated that there are several techniques used to evaluate the academic performance of students. Data mining is one of the most familiar techniques use to examine the academic performance of students. Support Vector Machine (SVM) algorithm is used in this research which is a technique of supervised learning. Through SVM algorithm the prediction is performed and data is analysed using classification. [17] proposed SVM which is very useful technique for the data classification. [18] investigates students' academic performance prediction using support vector machine. This study examines the association between preadmission academic profile of students and final performance of academics. The support vector machine outperforms other machine learning algorithms.

Besides, [19] proposed classification of students based on quality of life and academic performing using SVM and [20] use the SVM as machine learning too due to its advantage to improve the accuracy of classification procedure especially in data mining. Moreover, the result from [21] shows that SVM is one the model that capable of predicting with scoring high accuracy not less than 92%.

This study mainly focused on finding the pattern of students' academic performance before during online learning due to the COVID-19 pandemic outbreak by referring on their Grade Point Average (GPA). Moreover, this study also to analyse the prediction of academic performance of Universiti Pendidikan Sultan Idris (UPSI) undergraduates' students after they completely attend whole one semester by studying online based on classification in Support Vector Machine (SVM).

Methodology

In order to find the pattern of students' academic performance before and after online learning and the proposed modelling students' academic performance was shown in Figure 1.
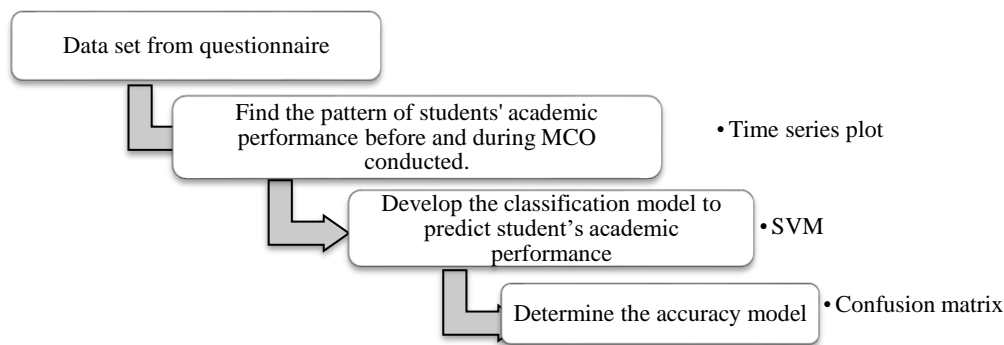

Figure 1 Flowchart of modelling students' academic performance

Data

The dataset used in this study were from undergraduate student from Universiti Pendidikan Sultan Idris (UPSI). The data was collected by using questionnaire and was distributed through online platform. The data collected based on the questionnaire are GPA's students before and during online learning, ages of students and current Cumulative Grade Point Average (CGPA). The questionnaire was responded by undergraduate students from semester 3 to semester 7 from Faculty of Science and Mathematics, UPSI. The data collected shows that, 82.5% are female and 17.5% are male. The data also consist of 72.5% from department Mathematics, 5.7% from department Biology, 12.2% from department Chemistry, 3.1% from department Physics and 6.6% from department Science and the total respondents are 225 students.

Time Series Plot

In order to find the pattern of students' academic performance before and after online learning, a graph has been constructed by using excel. The graph uses GPA's students before and during online learning as a data.

Support Vector Machine (SVM)

The Support Vector Machine (SVM) is a powerful machine learning tools which was proposed by [22] and become more attracted of machine learning researchers and community. The algorithm of SVM has been proven effectively to be used in regression and classification methods. Based on previous studies from [23], the study reported that the SVM is generally able to result the best accuracy of classification compare than other methods. Also, SVM can performs linear and nonlinear classification with high efficiently. However, the challenging of using SVM in classification or regression is to find the best of penalty term parameter and kernel parameters. It is because of SVM is very sensitive to the parameter used.

Consider that dataset from PCs were divided into two sets which are training data and testing data. The training data with two classes $[(x_1, y_1), (x_2, y_2), \ldots \ldots, (x_i, y_i)]$ and the input vector is x_i, the output is y_i. The output was labelled by $y_i \epsilon \{+1, -1\}$. The classifier for the problem of binary classification is

$$f(x) = sign[w^T \cdot \phi(x) + b] \tag{1}$$

where the input vector (x) was mapped with a feature space by non-linear function $\phi(x)$. Then, the w and b are the classifier parameter. By solving the optimization problems are equivalent with determine the SVM classifier from theory,

Nor Ain Maisarah Samsudin, Shazlyn Milleana Shaharudin, Nurul Ainina Filza Sulaiman, Muhammad Fakhrullah Mohd Fuad, Muhammad Fareezuan Zulfikri, Nurul Hila Zainuddin

$$\text{min} \qquad \frac{1}{2}w^T \cdot w + C \sum_{i=1}^{l} \xi_i$$
$$\text{subject to } y_i[w^T \cdot \phi(x_i) + b] \geq 1 - \xi_i \qquad (2)$$
$$\xi_i \geq 0, \quad i = 1, \ldots, l$$

where ξ_i is a non-negative slack variables that influence objective function when data misclassified. Then, C is a penalty parameter with positive value. The optimization problems will be solved by using Lagrange multiple, α_i where $0 \leq \alpha_i \leq C$. Hence, the classifier will be eqn. (3) by a series of mathematical derivation.

$$f(x) = sign\left(\sum_{ij=1}^{l} \alpha_i y_i \phi(x_i)^T \cdot \phi(x_j) + b\right) \qquad (3)$$

The kernel function, $K(x_i, x_j) = \phi(x_i)^T \cdot \phi(x_j)$ was introduced to calculated the inner products. There are a quite number of kernel might be used in SVM classification but the standard kernels used are linear, polynomial, radial basis function and sigmoid. A most popular and capable kernel is the radial basis function with parameter γ.

$$K(x, x') = \begin{cases} x^T.x' & linear \\ (x^T.x' + 1)^d & polynomial \\ \exp(-\gamma||x - x'||^2) & RBF \\ \tanh(\gamma x.x' + C) & sigmoid \end{cases} \qquad (4)$$

In the final classifier, only nonzero Lagrange multiple will be take part as indicated in eq. (3). The data which having nonzero corresponding Lagrange multiple will be named as support vector. Then, the classifier will be written as,

$$f(x) = sign\left(\sum_{k=1}^{m} \alpha_k y_k K(x_k, x) + b\right) \qquad (5)$$

where x_k is the support vector and m is the number of support vector. In the support vector classifier, there have two parameters to calibrate which are C(misclassification tolerance parameter) and $epsilon$ [24]. By determine those parameter, the Lagrange multiple and parameter b in eq. (5) can be find by the SVM algorithm.

Result and discussion

Figure 2 shows the result of GPA students before and during pandemic outbreak using two different types of classes. The blue line represents the student's GPA which conducted face-to-face classes and the orange line represent the student's GPA which conducted via online classes. Meanwhile, x-axis represent the students and y-axis represent the GPA. From the Figure 2, the pattern clearly shows that most of the students got excellent results when conducted using online classes. This is because, the orange line are mostly on top of blue line where it indicates that majority of the students score higher during online classes compared to the face-to-face classes. Thus, the prediction model was used to estimate the student's GPA performance for next coming semester if online classes are continued.
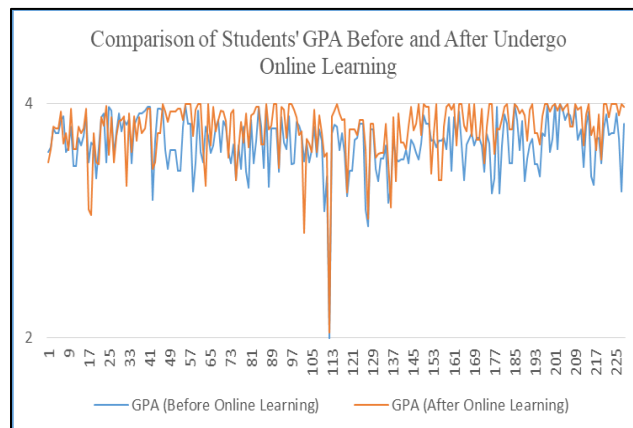

Figure 2 GPA of academic performance students before and during COVID-19 outbreak.

In this study, SVM classification model is used to predict students' academic performance. Therefore, it is most important to identify the best parameter in the SVM classification. The selection of the best parameters will improve the accuracy of classification. And, the selection process has called by turning parameter in SVM. Table 1 shows the turning parameter process and the best parameter was chosen by depending the smallest misclassification error. The selected values of parameter C are 4, 8, 16, 32, 64 and 128. Then, the selected values of $epsilon$ are 0, 0.2, 0.4, 0.6, 0.8 and 1 for each values of C respectively. Hence, based on the table 1 shown that the smallest misclassification error is 0.195398.

Table 1 The Result of Turning Parameter Process

	Epsilon	C	Misclassification error
1	0.0	4	0.263636
3	0.2	4	0.225100
5	0.4	4	0.202496
7	0.6	4	0.195400
9	0.8	4	0.203949
11	1.0	4	0.228148
12	0.0	8	0.263637
14	0.2	8	0.225242
16	0.4	8	0.202496
18	0.6	8	0.195400
20	0.8	8	0.203949
22	1.0	8	0.228148
23	0.0	16	0.263636
25	0.2	16	0.225300
27	0.4	16	0.202496
29	0.6	16	0.195400
31	0.8	16	0.203949
33	1.0	16	0.228148
34	0.0	32	0.263637
36	0.2	32	0.225275
38	0.4	32	0.202496
40	0.6	32	0.195400
42	0.8	32	0.203949
44	1.0	32	0.228148
45	0.0	64	0.263637
47	0.2	64	0.225285
49	0.4	64	0.202496
51	0.6	64	0.195400
53	0.8	64	0.203949
55	1.0	64	0.228148
56	0.0	128	0.263636
58	0.2	128	0.225368
60	0.4	128	0.202496
62	0.6	128	0.195398
64	0.8	128	0.203949
66	1.0	128	0.228148

However, based on the table 1 the patterns of misclassification error are slightly different. Hence, the result by plotting graph to measure the performance of turning parameter will be helping to identify the best pair of parameter. The figure 3 shown the graph between cost against epsilon and resulting the misclassification error.

The misclassification error was measured by using the reference scale of colour at right side in figure 3. Then, the figure 3 clearly shown that when epsilon became 0.6 the colour become darker and when it's approaching higher value of cost the colour became more darkest than other values. Hence, the result of turning parameter has summarized that the best pair of parameter are 128 as the C value and 0.6 as epsilon value selected. By using the selected pair of parameter, the SVM classifier will be resulting the best accuracy of classification model. In way to get the best result by using SVM, the selected of kernels also is the main point to highlight in this study. The types of kernel were used in this study are radial basis function, sigmoid, polynomial and linear. Each kernel function has a particular parameter that must be optimized to obtain the best result performance [25].
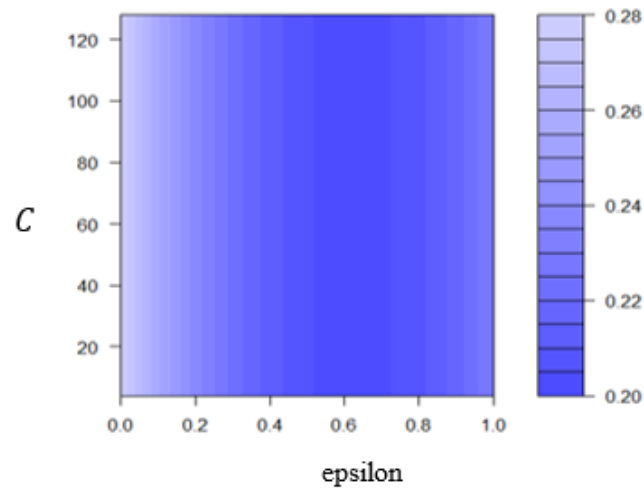

Figure 3 Performance of turning parameter

Table 2 Result of Support Vector Machine by Classification

Type of kernel	Parameter		Result	
	C	$epsilon$	No. of support vectors	Accuracy model (%)
Linear	128	0.6	59	73.68
Radial Basis Function	128	0.6	58	73.69
Polynomial	128	0.6	60	73.69
Sigmoid	128	0.6	56	67.99

As shown in Table 2, SVM was resulting the number support vector and misclassification error. The number of support vector are representing the data which approaching or far away from hyperplane during classification. The suitable value of number of support vector is medium value where it is representing the classification is overfitting or underfitting. Based on Table II, the highest number of support vector is 60 and the lowest number of support vector is 56. So, the medium value will be 58 and 59 which represent for radial basis function and linear kernel respectively. Then, the accuracy of model to be classification model can be determining by subtracts misclassification error with 1. The values of misclassification model were obtained by using matrix confusion between prediction and current class. The best accuracy of model achieved when type of kernel be a linear kernel with 73.68% and the worst accuracy get from a sigmoid kernel which is 67.99%.

Conclusion
In this study, GPA's students between face-to-face learning (before MCO) and GPA's students during online learning (during MCO) was studied to find the pattern of the students' academic performance. From the result, students outperform very well when online classes were conducted compared to the face-to-face classes. Most of the students achieve high GPA during online learning compared to the last semester's GPA which is before online learning. Meanwhile, SVM was applied by stressing on the classification method to predict student's academic performance based on their current CGPA and their ages. From the study, by using SVM, the best accuracy of model achieved when type of kernel be a linear kernel with 73.68%. The most accuracy of SVM was achieved by applying linear kernel with pair of misclassification tolerance, C, 128 and epsilon 0.6. The best number of support vector for this study is 59 which not be overfitting or underfitting. Therefore, this SVM classification model are successfully predicted students' academic performance by using misclassification tolerance, C, and epsilon as a parameters. For future work, the accuracy of predicting students' academic performance in this study can be compared by using other models in machine learning such as Artificial Neural

Network (ANN) and Relevance Vector Machine (RVM). Predicting students' academic performance can be very useful in many contexts especially in management such as identifying excellent students for scholarship programs, admissions, and to help to quickly identify students who are unlikely to graduate.

Acknowledgement

References

1. World Health Organization. Novel Coronavirus (2019-nCoV) Advice for the Public. Available athttps://www.who.int/emergencies/diseases/novel-coronavirus-2019/advice-for-public, accessed 2020.
2. World Health Organization. Pneumonia of Unknown Cause. Available at: https://www.who.int/csr/don/05-january-2020-pneumonia-of-unkown-cause-china/en/, accessed 2020.
3. World Health Organization. (2020, June 30). Timeline of WHO's response to COVID-19. Available at: https://www.who.int/malaysia/emergencies/coronavirus-disease-(covid-19)-in-malaysia, accessed June 2020.
4. Tang and Ashley. (2020). Malaysia announces movement control order after spike in Covid-19 cases (updated). The Star. Archived from https://www.thestar.com.my/news/nation/2020/03/16/malaysia-announces-restricted-movement-measure-after-spike-in-covid-19-cases, accessed 2020.
5. NS Times. Covid-19: Movement Control Order imposed with only essential sectors operating. New Straits Times. 16 March 2020. Archived from https://www.nst.com.my/news/nation/2020/03/575177/covid-19-movement-control-order-imposed-only-essential-sectors-operating, accessed 2020.
6. Y Palansamy. Higher Education Ministry: All university lectures to be online-only until end 2020, with a few exceptions. Retrieved from https://www.malaymail.com/news/malaysia/2020/05/27/higher-education-ministry-all-university-lectures-to-be-online-only-until-e/1869975, accessed 2020.
7. Bernama. UPSI to implement online teaching in line ministry's instruction. The Sun Daily. Retrieved fromhttps://www.thesundaily.my/local/upsi-to-implement-online-teaching-in-line-ministry-s-instruction-DB2175401, accessed 2020.
8. MDIBBaba and GJ Pendek. Keberkesanan Pengajaran Dan Pembelajaran Dan Kaitannya Terhadap Prestasi Akademik Pelajar Uthm.2009.
9. U.S. Department of Education, Office of Planning, Evaluation, and Policy Development. Evaluation of Evidence-Based Practices in Online Learning: A Meta-Analysis and Review of Online Learning Studies, Washington, D.C. Available at: https://www2.ed.gov/rschstat/eval/tech/evidence-based-practices/finalreport.pdf, accessed 2010.
10. Ahn and Nguyen. The Impact of Online Learning Activities on Student Learning Outcome in Blended Learning Course. *Journal of Information & Knowledge Management.*2017, DOI: 16. 10.1142/S021964921750040X.
11. D Smith and B Stephens. Marketing Education: Online vs Traditional. Proceedings of the American Society of Business and Behavioral Sciences, 17, 810-814.Gonzalez, D., & Louis, R. St. (2018). Online Learning. In J. I. Liontas (Ed.), The TESOL Encyclopedia of English Language Teaching (1st ed.). Retrieved from https://doi.org/10.1002/9781118784235.eelt0423, accessed 2010.
12. C Gonzalez. Conceptions of, and approaches to, teaching online: a study of lecturers teaching postgraduate distance courses. Higher Education, 57(3), 299–314. Available at: https://doi.org/10.1007/s10734-008-9145-1, accessed 2009.
13. EKayode, -O and TL Teng.The Impact of Transactional Distance Dialogic Interactions On Student Learning Outcomes in Online and Blended Environments. *Comput Educ*, 2014; **78**, 414–427.
14. I Mushtaq and SN Khan. Factors Affecting Students' Academic Performance. *Global journal of management and business research*. 2012; **12**, 9.
15. NA Oye et al. The impact of e-learning on students' performance in tertiary institutions. *International Journal of Computer Networks and Wireless Communications*, 2012; **2**, 2,121-30.
16. AM Shahiri, W Husain, NA Rashid. A review on predicting students' performance using data mining techniques. *The 3rd Information System International Conference, Elsevier*. 2015; **72**, 414–422.
17. CW Hsu, CC Chang, and CJ Lin. A practical guide to support vector classification. 2003.
18. K Nahar, MA Ottom, F Alshibli and M Abu Shquier. Air Quality Index Using Machine Learning – A Jordan Case Study. *Compusoft: An International Journal of Advanced Computer Technology*. 2020; **9**, 9, 3831-3840.

19. SA Oloruntoba and JL Akinode. Student academic performance prediction using support vector machine. *International Journal of Engineering Sciences and Research Technology*. 2017; **6**, 12, 588–597.

20. Z Raihana and AM Farah Nabilah. Classification of students based on quality of life and academic performance by using support vector machine. *Journal of Academia UiTM Negeri Sembilan*. 2018; **6**, 1, 45–52.

21. NAF Sulaiman, SM Shaharudin, NH Zainuddin, SAM Najib. Improving Support Vector Machine Rainfall Classification Accuracy based on Kernel Parameters. 2020.

22. O Yamini and S Ramakrishna. A Study on Advantages of Data Mining Classification Techniques, *International Journal of Engineering Research & Technology (IJERT)*. 2015; **4**, 9,969-972.

23. SM Shaharudin, S Ismail, SMCM Nor and N Ahmad. An Efficient Method to Improve the Clustering Performance using Hybrid Robust Principal Component Analysis-Spectral Biclustering in Rainfall.2013.

24. RG Brereton and GR Lloyd. Support Vector Machines for Classification and Regression, The Royal Society of Chemistry. *Analyst*. 2010; **135**, 230–267.

25. Y Zhang and L Wu. Classification of fruits using computer vision and a multiclass support vector machine.*Sensors (Switzerland)*. 2012; **12**, 9, 12489-12505.