ISSN 2409-2665
Journal of Logistics, Informatics and Service Science
Vol. 9 (2022) No.3, pp.64-77
DOI:10.33168/LISS.2022.0306

Predicting Student Performance from Video-Based Learning System: A Case Study

Chin-Wei Teoh[1], Sin-Ban Ho[1+], Khairi Shazwan Dollmat[1], Chuie-Hong Tan[2]
[1] Faculty of Computing and Informatics, Multimedia University, 63100 Cyberjaya, Malaysia
[2] Faculty of Management, Multimedia University, 63100 Cyberjaya, Malaysia

[+]*sbho@mmu.edu.my (corresponding author)*

Abstract. The major impact of the COVID-19 pandemic on the shift of education norms from physical classroom learning to MOOCs (Massive Open Online Courses) could accelerate the big data era growth for the e-learning platform. This circumstance has provided an opportunity for a teacher to use MOOC data to help students learn and perform better. Moreover, this research study goal is to propose a combination of machine learning algorithms and the feature selection benefit with the SMOTE (Synthetic Minority Oversampling Technique) algorithm for balancing the output features number to predict student performance in a video-based learning platform. As a result, the proposed machine learning classifier, Naïve Bayes algorithm with the combination of chi-square test and SMOTE has shown the highest accuracy in prediction of more than 90%. Results by the proposed classifier with feature selection and SMOTE have outperformed the traditional machine learning classifiers.

Keywords: machine learning, educational data mining (EDM), smote, student performance

1. Introduction

The most typical approach of passing on knowledge to students was formerly regarded to be having instructors physically present to educate (Hoyos & Velasquez, 2020). However, the educational technology domain advancement has resulted in a new educational standard, with many students using online e-learning platform like Udemy, Coursera, and Edpuzzle to master innovative skills (Al Kurdi et al.,2020).

MarketDigits research analysis projected the MOOCs market to grow at 32.8 percent in Compound Annual Growth Rate (CAGR) between 2021 and 2026 (MarketDigits, 2021). To improve learners' performance, the emerging MOOC data provides a good research avenue to educational data mining (EDM). Many researchers have used artificial intelligence knowledge to construct a prediction model on MOOC data to predict student success in this scenario (Mubarak et al., 2021). Educators might benefit from analyzing a trend in student learning and gathering information to assist them make better judgments when constructing instructional approaches.

Video was a primary source of MOOC course creation in the MOOC system. With at least one video form learning material being uploaded to each learning course (Teoh et al., 2022). According to Edgar Dale, the creator of the Cone of Learning Theory, learner would like to retain 90% on doing, 70% on saying and writing, 30% on seeing, 20% on hearing, and 10% on reading (Masters, 2020). Furthermore, the hearing and seeing parts within videos, allow learners to retain at least 50% of the information delivered. Moreover, the ability of combining graphic, animation, text, and audio together, video has become essential e-learning material to every student on today.

According to a YouTube Marketing report, during the time of the covid-19 pandemic in January and February 2020, just over 300 educational videos with online teaching or distance learning in the title were uploaded to YouTube, and that number increased by over 23,000 in March 2020 (FutureSource, 2020). Therefore, the increased video-based learning popularity has stimulated the educational video learning analytics growth.

Nevertheless, due to the limited interaction between students and educators in online learning, educators find it challenging to determine students who are at risk of poor performance and intervene as soon as possible. As a result, instructors require an autonomous predictive system that can analyze MOOC data and provide predictions on individual student performance once they have completed the system. The goal of this study is to implement a prediction model based on the number of videos watching rate to predict student performance in video-based learning.

2. Related Works

Research in education has shown that spent to watch all the allocated videos on a single day causes students to have a hefty study stress. Furthermore, most participants hanged on till the final day to watch the assigned video. They tended to skip insignificant video topics to learn faster (Nielsen, 2020). An alternative previous finding shown that in video-based learning, video watching rate increased on the day prior to assignments and tests (Ahn & Bir, 2018). Simultaneously, a rise in video viewing before tests and assignments has prompted the measurement of the total amount of time spent by students watching videos.

Furthermore, survey research of 357 papers in student performance identified factors that had the greatest impact on learner performance, including student engagement, demographics, and psychomotor skills (Hellas et al., 2018). As a result, the purpose of this study was to collect demographic information about participants in terms of their previous academic success, gender, and age with their learning style that can have an impact on student success in MOOCs.

Data from MOOCs have used widely in machine learning field for predicting student performance. According to Xu et al. research, the ensemble learning technique has been applied to predict student academic performance (Xu et al.,2017). Furthermore, the heterogeneous ensemble learning algorithm has been proven that outperformed the main machine learning method in predicting student performance (Pujianto et al., 2020).

The author used several machine learning algorithms to classify student performance in a study related to video learning analytics, including naive bayes (NB), support vector machine (SVM), and logistic regression, with the random forest model achieving approximately 88 percent accuracy (Hassan et al., 2020). Another study used the AdaBoost classifier, an ensemble learning algorithm that successfully enhanced the prediction model's accuracy to around 80% in predicting student performance (Kumar et al., 2020).

The study of whether it can increase the predictive model performance has focused on feature selection. The use of a mix of chi-square test and the ensemble learning models as a feature selection strategy to improve the predictive model's accuracy has proven crucial (Fitni & Ramli, 2020). Ebrahimi-Khusfi et al. found that adding the feature selection approach to a predictive model improves its performance (Khusfi et al., 2020). Furthermore, it is possible combine methods such as SMOTE and ensemble learning algorithms for addressing imbalance data in advancing the prediction model (Huang, 2021). For high imbalance datasets, an integration between feature selection and SMOTE has been applied to improve the prediction model's performance (Huang, 2021). The SMOTE method was applied to student performance datasets, resulting in a 10% increase in NB and random forest accuracy when compared to performance without the SMOTE method (Rattan, 2021).

As a result, most of the past relate works indicates that the imbalance data technique and feature selection are important variables to boost predictive model performance. Furthermore, prior comparable research has determined that gathering student's previous academic achievement and learning style was more important than focusing solely on learners' involvement with video viewing. Therefore, this research would focus on gathering information on the students' learning styles and previous academic grades.

3. Methodology

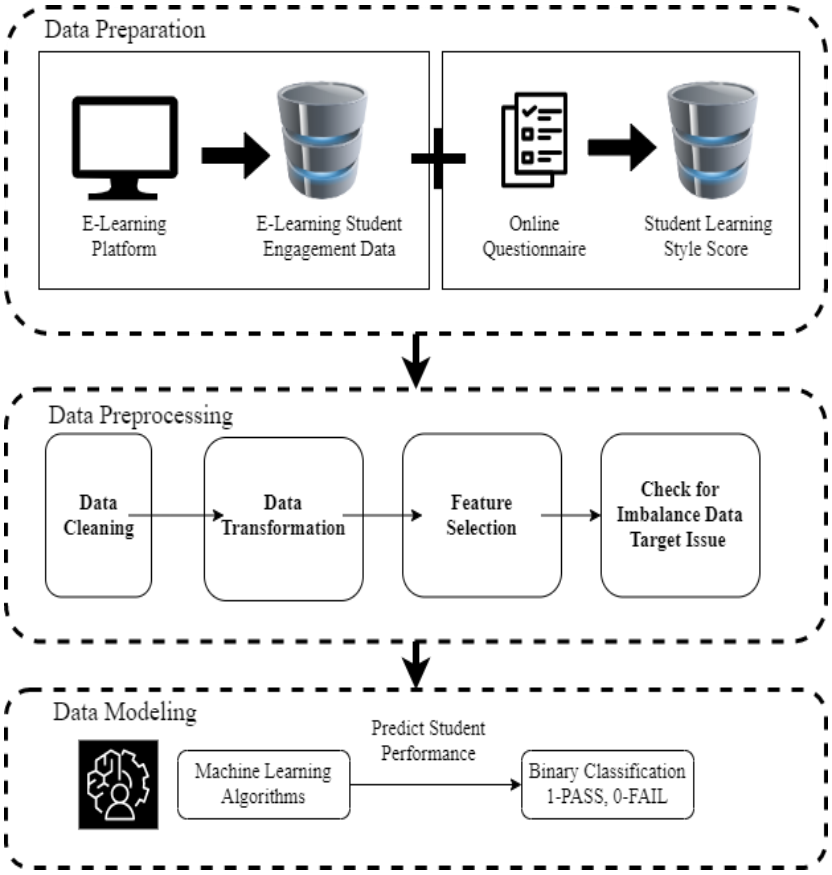

Fig. 1: Overall research framework.

There are 110 university students in total were allocated to watch a software design video content in an e-learning platform when the COVID-19 pandemic was in full swing at Malaysia. In this case, 22 features have been collected from video watching rate and learning style score were compiled into a comprehensive dataset**.** Figure 1 shows that two types of input features were collected: video engagement features and the learning style of the student. We captured the number of views for

each video portion in the e-learning or MOOCs system. Meanwhile, we also captured each student duration in watching videos. On the other side, an online questionnaire was used to determine learning type. The Felder Silverman (FS) learning style model was used to determine student learning styles in this scenario.

Next, the original datasets first went through a data cleaning process and then followed by data transformation. The Chi-Square test will be used to identify significant features against the data output, and then the SMOTE technique will be used to rectify the imbalance data issue in the original datasets.

The datasets were split into two groups throughout the data modelling phase where 30% for the test set and 70% for the training set. There were 10 different types of machine learning algorithms created. Finally, accuracy was used to assess the performance of all machine learning approaches. The model's final prediction was a binary classification, with 1 indicating a student's PASS outcome and 0 indicating a student's FAIL result.

3.1. Dataset description

There are 22 data attributes in all from the source datasets. Two of the 22 features are nominal data types, whereas the other 20 are integer data types. Non-numeric values, such as a label, are available in nominal data types. All 21 input features, except for Grade, which is the output, could be divided into three groups: learner's video watch engagement, learner's learning style attributes, and learner's demographic background as shown in Table 1.

Table 1: Data categories descriptions

Data Categories	Data Features
Student demographic background	Gender, CGPA_class
Student video watch engagement	Time_1, Time_2, Time_3, Time_4, Time_5, Time_6, Time_7, Time_8, Time_9, Time_10, Time_Spent, Grade_Q1, Grade_Q2, Grade_Q3, Rating
Student learning style	A/R Score, S/I Score, Vi/Vb Score, S/G Score

3.2. Experimental design for prediction framework

Table 2: Experiment design for data modelling.

Group	Feature Selection	SMOTE
C1	No	No
C2	Yes	No
C3	Yes	Yes

The experiment has been designed into three experimental groups as illustrated in the Table 2 to compare the performance of the prediction models. In each group of experiment, 10 different type of machine learning algorithms have been evaluated.

C1 consider as baseline models where all prediction models have been designed with datasets from data cleaning and transformation but without went through feature selection and SMOTE. C2 uses data that went through feature selection only but not SMOTE whereas C3 uses data that went through feature selection and SMOTE.

3.3. Machine learning algorithm

We applied a total of 10 different machine learning algorithms on the experiment test as listed in the Table 3. Each machine learning algorithms have tested under different case conditions.

Table 3: Machine learning algorithms tested.

No	Machine Learning Algorithms
1	Linear Support Vector Machine (LSVM)
2	Radial Support Vector Machine (RSVM)
3	Bagging
4	AdaBoost
5	Logistic Regression (LR)
6	Random Forest (RF)
7	Decision Tree (DT)
8	K-Nearest Neighbors (KNN)
9	Naïve Bayes (NB)
10	Gradient Boosting (GB)

3.4. Model evaluation

In the context of model evaluation, the performance of each machine learning models was evaluated based on accuracy. In addition, confusion metrics were applied for the analysis of prediction outcome where each column of the matrix represents the different predicted classes whereas each row of the matrix represents the actual classes as illustrated in Table 4.

Table 4: Confusion matrix for classification.

	FAIL (Predicted -0)	PASS (Predicted -1)
FAIL (Actual - 0)	a	c
PASS (Actual -1)	b	d

The upper row label in the confusion matrix has the following meaning where label a is the correct negative prediction or true negative (TN), classified as grade FAIL by the classifier, label c is incorrect positive prediction or false positive (FP), classified as grade PASS by the classifier. While the bottom label b is incorrect

negative prediction or false negative (FN), classified as grade FAIL by the classifier, label d is correct positive prediction or true positive (TP), classified as grade PASS by the classifier. Moreover, the accuracy formula can be referred to Equation 1.

$$Accuracy = \frac{(a+d)}{(a+b+c+d)} \quad (1)$$

4. Results and Discussions

4.1. Result of data pre-processing

Firstly, we checked the collected dataset for null values during data preprocessing. From the 22 features of the original datasets, there was no null value found among the 110 entries. Table 5 shows the Gender and CGPA Class nominal attributes were then transformed into numerical data types.

Table 5: Result of data types converted.

Data Features	Categorical to Numerical
Gender	Female = 0, Male =1
CGPA Class	2.00 - 2.66 = 1, 2.67 - 3.32 = 2, 3.33 - 3.66 = 3, 3.67 - 4.00 = 4

4.2. Summary of video engagement data

The number of people who watched each video interval was measured in this study, as shown in Table 6. The experimental materials have ten video time intervals. In this case, each video time interval has 31 minutes. Different video contents are included in each video time interval. Based on each video time interval, the maximum, minimum, and average number of views are measured.

Table 6 shows that across all video time intervals, the video interval between 2 min 4 sec and 2 min 35 sec had the most views where the students watched multiple time to gain comprehensive understanding on the knowledge of the singleton class constructor. Each video time has been ignored, implying that the pupil has skipped it. Every video time interval has at least one number of view on average.

Table 6: Number of view on each video interval.

No	Video Time Interval	Video Delivered Content	Maximum Watched	Minimum Watched	Average Number of Watched
1	0:00 -0:31	Introduction of video theme	6	0	1.3
2	0:31 - 01:02	Introduction of Singleton	7	0	1.3
3	01:02 - 01:33	Introduction of Singleton	5	0	1.2
4	01:33 - 02:04	How to Implement Singleton	4	0	1.2
5	02:04 - 02:35	Explain of singleton class constructor	9	0	1.6
6	02:35 - 03 : 06	Explain of singleton class variable	4	0	1.1
7	03:06 : 03:37	Implementation of GetInstance class	5	0	1.2
8	03:37 - 04:08	Concepts Lazy Creation	5	0	1.2
9	04:08 - 04:39	explain of Lazy Creation implementation	5	0	1.1
10	04:39 - 05:11	Example of Singleton concepts in application and summary of video	6	0	1.3

According to the results in Figure 2, most learners took 6 minutes in average watching the entire video. The original video duration provided in this study consists of five minutes and 24 seconds. At the same time, the maximum time of video watched was 10 minutes, implying the learner may have repeated certain video interval portions.

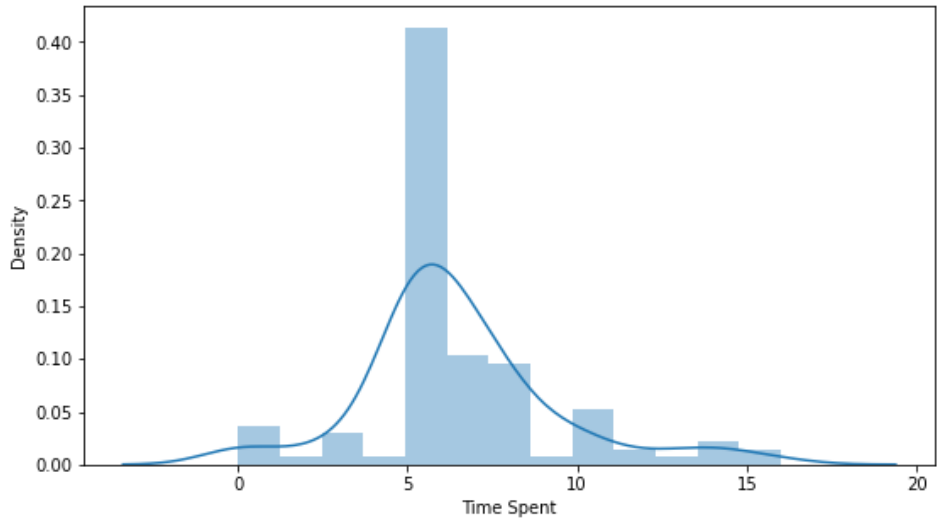

Fig. 2: Relationship between time spent variable and the video watched.

Figure 3 compares the video time spent variable to the student grade, with 0 denoting a FAIL and 1 denoting a PASS. Hence, learners who watched the video multiple times or took more than five minutes were more likely to receive a PASS mark at the end than learners who spent less than 5 minutes watching the video. These findings have indicated that if learners spend more time watching the video, they will have gathered enough information and knowledge to prepare for the quiz.

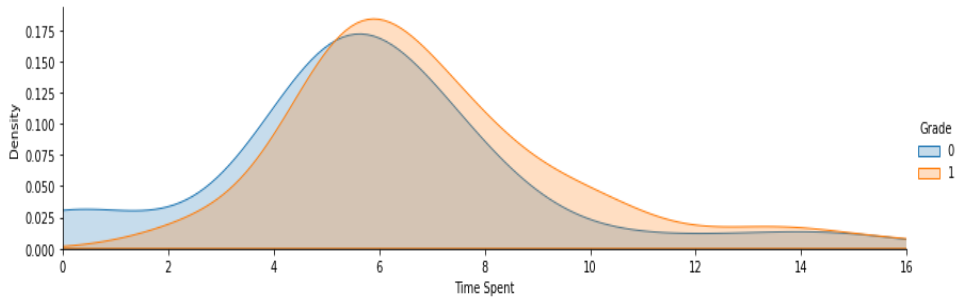

Fig. 3: Relationship between time spent variable and student grade.

4.3. Outcome of feature selection and SMOTE

The Chi Square values computed from the Chi-Square method on each input feature are shown in Figure 4. The Chi-Square value determines the independence between the input features and the output feature, Grade, in this case. Consequently, a total of ten input characteristics were chosen because they had a higher Chi-Square value and were significantly reliant on the output feature, as shown in Table 7.

Table 7: Top 10 input features with high chi-square value.

Input Features	Chi-Square Value
Time_Spent	5.60
Time_9	2.19
Time_8	1.71
S/G Score	1.63
Time_3	1.34
Time_7	1.31
Time_6	1.22
Time_5	1.13
Time_4	1.10
Grade_Q2	0.63

The initial datasets comprise a total of 65 students who receive a PASS result and 45 students who receive a FAIL grade. As a result, using SMOTE as an oversampling strategy, a balanced data target of 65 students was created for both classes.

4.4. Accuracy results of machine learning models

Table 8 illustrates that the NB algorithm attained the highest accuracy of approximately 92.3% under case condition 3 (C3) which involved Chi-Square feature selection and SMOTE method. However, not all machine learning algorithms have followed the similar improvement as NB classifier.

Table 8:. Machine learning models accuracy based on case conditions.

Algorithms	C1	C2	C3
LSVM	63.6	63.6	38.5
RSVM	63.6	63.6	30.8
Bagging	54.5	72.7	53.8
Adaboost	54.5	63.6	46.2
LR	54.5	63.6	38.5
RF	36.4	63.6	61.5
DT	36.4	63.6	61.5
KNN	45.5	63.6	61.5
NB	45.5	45.5	92.3
GB	27.3	54.5	53.8

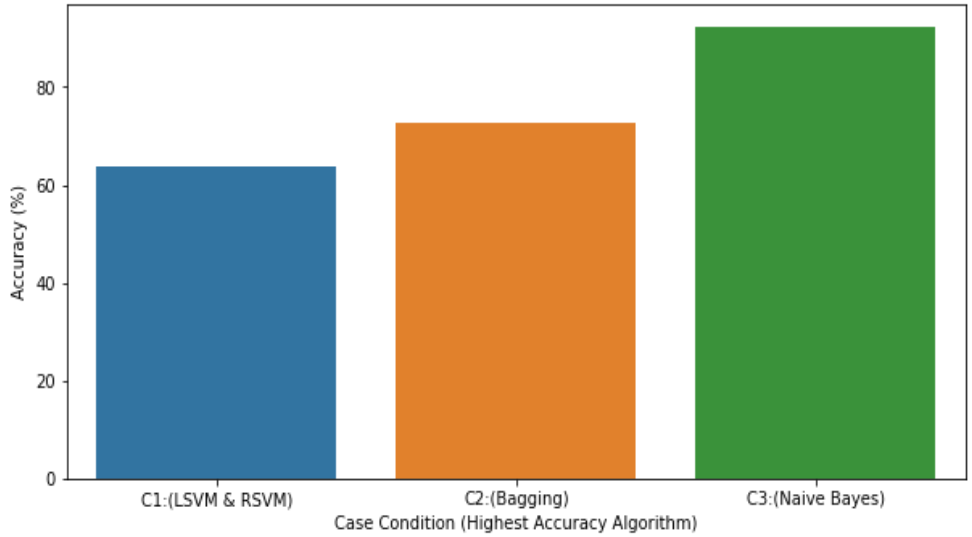

Fig. 4: Comparison of accuracy on each case condition.

By comparison on the highest accuracy rate on each case condition as illustrated in the Figure 4, case condition 3 has achieved the highest accuracy by 92.3%. In this case, LSVM and RSVM has achieved the similar highest accuracy of 64.6 % under the condition without Chi-Square test and SMOTE. While in the condition of only involved Chi-Square test, bagging classifier has achieved the highest accuracy of 72.7 %.

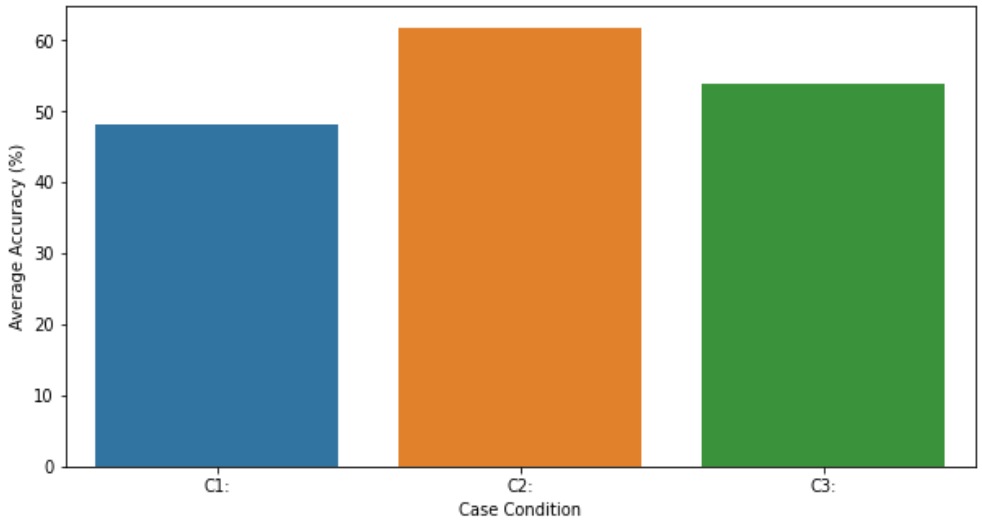

Fig. 5: Comparison of mean accuracy on each case condition.

By comparison on the average accuracy on each case condition as illustrated in the Figure 5, case condition 2 and 3 has achieved more higher average accuracy by 61.8% and 53.8% respectively compared to case condition 1 with only 48.2%. These finding has indicated that effect of Chi-Square test, SMOTE or both combinations are the key factors to enhance the accuracy of the classifier on prediction of student performance.

5. Conclusion

The goal of this research study was to implement and compare ten various types of machine learning algorithms for predicting student performance in video-based learning. The proposed methods integrate the benefits of the feature selection and the Chi-Square test as well as the SMOTE technique for strengthening the performance of predictive models. This study contributes to the following key findings: Firstly, the Naïve Bayes algorithm has the best prediction performance at about 92.7% accuracy after passing the Chi-Square test, followed by SMOTE. Secondly, the average accuracy prediction performance for the entire 10 machine learning algorithms under the proposed methodology outperforms all standard machine learning classifiers that without went through any feature selection, as well as SMOTE for dealing with data imbalance issues. Both finding has shown a consistent outcome with previous research works, indicating that feature selection and the technique of handling the imbalance data problems were critical variables in improving model effectiveness in predicting student success. Deep learning approaches should be investigated further and compared to present performance in future research.

Acknowledgments

The authors are grateful for this research project was funded by the Fundamental Research Grant Scheme (FRGS/1/2019/SS06/MMU/02/4).

References

Ahn, B. & Bir, D. D. (2018). student interactions with online videos in a large hybrid mechanics of materials course. *Advances in Engineering Education, 6*(3).

Al Kurdi, B., Alshurideh, M., & Salloum, S. A. (2020). Investigating a theoretical framework for e-learning technology acceptance. *International Journal of Electrical and Computer Engineering (IJECE), 10*(6), 6484-6496.

Ashraf, M., Zaman, M., & Ahmed, M. (2020). An intelligent prediction system for educational data mining based on ensemble and filtering approaches. *Procedia Computer Science, 167*, 1471-1483.

Ebrahimi-Khusfi, Z., Nafarzadegan, A. R., & Dargahian, F. (2021). Predicting the number of dusty days around the desert wetlands in southeastern Iran using feature selection and machine learning techniques. *Ecological Indicators*, *125*, 107499.

Fitni, Q. R. S., & Ramli, K. (2020, July). Implementation of ensemble learning and feature selection for performance improvements in anomaly-based intrusion detection systems. In *2020 IEEE International Conference on Industry 4.0, Artificial Intelligence, and Communications Technology (IAICT)* (pp. 118-124). IEEE.

FutureSource. (2020). social media and microlearning combine to expand access to education content. [Online]. Available: https://www.futuresource-consulting.com/insights/social-media-and-microlearning-combine-to-expand-access-to-education-content/?locale=en.

Hasan, R., Palaniappan, S., Mahmood, S., Abbas, A., Sarker, K. U., & Sattar, M. U. (2020). Predicting student performance in higher educational institutions using video learning analytics and data mining techniques. *Applied Sciences*, *10*(11), 3894.

Hellas, A., Ihantola, P., Petersen, A., Ajanovski, V. V., Gutica, M., Hynninen, T., & Liao, S. N. (2018, July). Predicting academic performance: A systematic literature review. In *Proceedings companion of the 23rd annual ACM conference on innovation and technology in computer science education*, 175-199.

Hoyos, A. A. C., & Velásquez, J. D. (2020). Teaching analytics: current challenges and future development. *IEEE Revista Iberoamericana de Tecnologias del Aprendizaje*, *15*(1), 1-9.

Huang, M. W., Chiu, C. H., Tsai, C. F., & Lin, W. C. (2021). On combining feature selection and over-sampling techniques for breast cancer prediction. *Applied Sciences*, *11*(14), 6574.

Kumar, M., Mehta, G., Nayar, N., & Sharma, A. (2021). EMT: Ensemble meta-based tree model for predicting student performance in academics. In *IOP Conference Series: Materials Science and Engineering*, 1022(1), 012062. IOP Publishing.

MarketDigits. (2021). Massive Open Online Courses (MOOCs) Market Growing at a CAGR of 32.8% during 2021-2026. [Online]. Available: https://www.globenewswire.com/en/news-release/2021/01/06/2154487/0/en/Massive-Open-Online-Courses-MOOCs-Market-Growing-at-a-CAGR-of-32-8-during-2021-2026-Growing-Popularity-Booming-Segments-Emerging-Trends-and-Investors-Seeking-Growth-Top-Players-Cou.html.

Masters, K. (2020). Edgar Dale's pyramid of learning in medical education: Further expansion of the myth. *Medical Education*, *54*(1), 22-32.

Mubarak, A. A., Cao, H., & Ahmed, S. A. (2021). Predictive learning analytics using deep learning model in MOOCs' courses videos. *Education and Information Technologies*, *26*(1), 371-392.

Nielsen, K. L. (2020). Students' video viewing habits during a flipped classroom course in engineering mathematics.

Pujianto, U., Prasetyo, W. A., & Taufani, A. R. (2020, December). Students' academic performance prediction with k-nearest neighbor and C4. 5 on SMOTE-balanced data. In *2020 3rd International Seminar on Research of Information Technology and Intelligent Systems (ISRITI)*, IEEE, 348-353.

Rattan, V., Mittal, R., Singh, J., & Malik, V. (2021, March). Analysing the application of SMOTE on machine learning classifiers. In *2021 International Conference on Emerging Smart Computing and Informatics (ESCI)*, IEEE, 692-695.

Teoh, C. W., Ho, S. B., Dollmat, K. S., & Chai, I. (2022, February). An evolutionary algorithm-based optimization ensemble learning model for predicting academic performance. In *2022 11th International Conference on Software and Computer Applications*, 102-107.

Xu, J., Han, Y., Marcu, D., & Van Der Schaar, M. (2017, February). Progressive prediction of student performance in college programs. In *Thirty-First AAAI Conference on Artificial Intelligence*