Predicting Engineering Students' Grade on Introductory Physics Using Machine Learning

Purwoko Haryadi Santoso[1, 2, *], Syamsul Bahri [3], Wahyudi [4], Johan Syahbrudin [1]

[1] *Educational Research and Evaluation, Postgraduate Program, Universitas Negeri Yogyakarta, Indonesia*
[2] *Physics Education, Faculty of Teacher Training and Education, Universitas Sulawesi Barat, Indonesia*
[3] *Physics Education, Faculty of Teacher Training and Education, Universitas Musamus, Indonesia*
[4] *SMK N 1 Budong-Budong, Central Mamuju*
[*] *Corresponding author. Email: purwokoharyadi.2021@student.uny.ac.id*

ABSTRACT

Introductory physics is a compulsory course for the first-year engineering college for providing students the underlying concepts of the future course throughout their study. A sudden shift of distance learning during the disruption of COVID-19 in the middle of 2021 has generated an extensive collection of educational data that can potentially be mined for educational purposes. Educational data mining (EDM), a branch of machine learning research, has offered some tools to perform this task. In this study, a logistic regression classifier was employed to early identify students' performance in introductory physics courses for engineering majors. Data were collected at a public university (N=180) through a learning management system engaged throughout a semester. This study successfully trained the model with an 80% identification rate to predict the low-performing student in the course. The finding is necessary for the educator to do the review and give feedback to their class for providing some help, particularly for the low-performing student. It is suggested for the further development of the model to make prediction more accurate with another model ensemble that has been proven decisive in the recent study of machine learning.

Keywords: *Course, Machine learning, Physics, Predition*

1. INTRODUCTION

Physics is a fundamental science to explain physical phenomena as an introductory physics course that is necessary for complementing the principal knowledge of the engineering students [1]. Ref. [2] described that introductory physics course as an approach to "scope, generate, evaluate, and realize ideas." Students should understand multiple concepts to have experience in solving multidimensional real-world problems. Introductory physics should be an excellent course to address the underlying idea of their engineering problems [3], [4]. The first-year engineering students who are still in the process of adoption and have no experience in the college atmospheres need it to have mastery of fundamental knowledge for their engineering world. At a broad destination, future challenges of global development as the industrial revolution 4.0 and artificial intelligence push the current generation to integrate their physical and digital knowledge into their lives. Fortuitously, the emergency transition of COVID-19 has increasingly transformed our direct interaction through distance learning. Students and instructors have been accustomed to the remote courses through various learning management systems (LMS). Therefore, introductory physics courses through LMS are vital for administering classes during current circumstances.

The delivery of remote courses has established the opportunity for massive educational data. The enormous data will have an essential feature to portrait the physics learning process. Educational data mining (EDM) and learning analytics (LA) is a purview of data science using the method of machine learning (ML) from

artificial intelligence (AI) studies. As a general thought, educators have to regularly evaluate the learning process and monitor their student's progress for a better learning experience. EDM is one approach to perform this essential task by analyzing the students' data generated from the LMS. In this work, authors implement the ML algorithm to model students' learning on introductory physics course at an Indonesian public university, particularly in predicting their performance at the end of the classes. This work is imperative to provide an early monitoring system so instructors have an initial description of their low-performing students that require additional intervention to help their learning.

EDM also offers a further advantage for the emerging technology that provides university benefits by enabling scholars to gather students' interactions. After the meeting in an undisturbing way, thus one can observe how engineering students engage during the learning process [7]. Through this approach, one can study, model, and predict the students' performance as we are doing in this study. EDM studies should recommend best practices for the educators to effectively take the educational generated data. This data is helpful to maintain their students' engagement in their class even if they may have limited effort to approach their students one by one among their other activities outside the class. Data mining (not specifically EDM) is considered a tool for other fields' technical issues by employing the artificial intelligence system, introductory physics courses for the engineering students. This tool is beneficial to support the creative and innovative learning process that has to be developed even more. Since there are no individual students that desire to be failed in their college course therefore their learning experience should be taken into account by the educators. It pushes a challenging mission for the instructors to assist their students' success using the ML algorithm through the EDM [8].

In this study, the employed student's dataset was gathered at an Indonesian public college generated through the LMS channel developed by the university's IT center during an introductory physics course. Students were taught over 16 weeks in the semester encompassing the physical concept of mechanics, heat, and sound through the LMS. The interaction between students and the LMS was recorded into comma-separated values (CSV) formats. We implement one of the machine learning algorithms to model students' learning through the widely used open-source programming language, python. The model will predict students' achievement on the four-point scale in the early semester. With this work, the authors propose one research question to be investigated in the model: how well students' performance can be predicted through the machine learning algorithm employed in this study?

2. METHOD

2.1. Course Details

Students' dataset for this study was collected through the introductory physics course (INF 092219) at Universitas Sulawesi Barat (Unsulbar). This course was administered through the second semester of the 2020/2021 academic year for the first-year engineering students at the Department of Informatics Engineering. Due to the pandemic outbreaks still present within us, students and instructors cannot manage their lectures through online platforms such as learning management system (LMS). During the second semester of the 2020/2021 academic year, 180 students were participating in this study in which they were divided into four classes managed with the same team of lecturers.

In this study, the researchers set the observation at the three points in the early semester for the study period. Each meeting takes three 50-min lectures and one session a week. Students were initially assigned a 15-min quiz about specific topics discussed in the current sessions to prepare them for the topic discussed in a weekly meeting.

2.2. Data

Students' activity was collected from their interaction between students in the classroom and the LMS system's learning process. The dataset was comprised of pre-class quizzes (Quiz1, Quiz2, Quiz3), students' submission of homework (PR1, PR2, PR3)., UN(physics), UTBK(physics), Calculus, and IPK (Sem1). The raw data should be pre-processed through several time-consuming processes based on the typical way of the EDM research [11]. After the data has been overviewed within exploratory data analysis, it will be divided into two subsets of training and testing data.

2.3. Machine Learning Model

In this study, researchers employed a freely accessible python library, scikit-learn [12], to build our machine learning model. Two features for each student, namely in-class and institutional variables was applied. A logistic regression classifier was implemented to predict the final grade of students' performance.

3. RESULTS & DISCUSSION

This study aimed to build a machine learning model to predict engineering students' performance in the introductory physics course. To evaluate the model, we should identify the essential features to predicting students'

performance at the period of study both with the institutional variable and using the in-class variable.

3.1. Logistic Regression Model Using Institutional Variables

Table 1 describes the collected institutional dataset employed to model the optimal logistic regression. Unstandardized variables were decided to report the odds-ratio and 95% CI of each logistic regression model. For instance, an increase of UTBK at one point will make the 10.80 times odds. Since standardized continuous variables were required to normalize the odds ratio, we should normalize the variables on a continuous scale. We calculate the z-score to create normalized data in the range [-3, 3] by subtracting the data with the class mean, and the result will be divided by the standard deviation. After the data has been normalized, the increase of one standard deviation would make the odds 3.35 times that is 235 percent of the previous UTBK odds ratio. Intercept is reported to describe the base odds when the independent variables are zero. Due to the majority of unstandardized intercepts being zero, zero UN or UTBK will lead the students to a small probability of passing the course. Instead, the standardized intercept reflects the students' odds with the zero categorical variables with the dichotomous format and mean values of whole continuous variables at A or B grade on the physics course.

Table 1. Optimal model using institutional data

Feature	Odds ratio	95% CI	Norm odds ratio	Z score	p
Intercept	0.00	[0.00, 0.00]	4.82	–8.53	<0.001
UN	1.14	[1.05, 1.02]	1.75	3.84	<0.001
UTBK	10.80	[6.56, 19.46]	3.35	8.58	<0.001
Calculus	1.03	[0.14, 0.55]	1.63	2.51	0.008
IPK	0.23	[0.12, 0.51]	0.35	–3.82	<0.001

There were four features required to build the optimal logistic regression model using institutional variables, namely UN, UTBK, Calculus, and UN. The possibility of students passing the introductory physics course with grades A or B may be increased with higher UN, UTBK, and Calculus. Calculus is mathematical knowledge that has been provided for the first engineering students in the first semester. UTBK is a nationally standard assessment comprised of six subjects in which physics is set as one of the aspects. Students' grades on UTBK highly influence their success in the introductory physics course. Students with a higher score of UTBK (physics) have a bigger chance to pass the course (A or B). Moreover, IPK became the most critical institutional variable because increasing IPK by about one point will change the odds increase 235 percent (A or B). UN grades are found as less important than other variables within the dataset. A higher UN score on physics was unable to ensure that students will have increased odds to obtain A or B on the course. The result was paramount, and it brought insights to the physics education community. Even high school physics has been examined in the UN, and it has a lack of contribution to guarantee students' performance on the introductory physics course. Implicitly, high school physics should be treated as preparation before college physics due to high achievement on the university was irrelevant with their basic understanding of the physics knowledge.

3.2. Logistic Regression Model Using In-Class Variables

Each week, in-class variables data were collected to predict students' grades using a logistic regression model. The modeling results' description using in-class features is presented in Table 2. Quizzes and homework are chosen as a predictor due to the accessibility of the date within the course. Those data are common in most college courses everywhere. Table 2 summarizes our optimal logistic regression model using the in-class variables.

Table 2. Optimal model using in-class data

Feature	Odds ratio	95% CI	Norm odds ratio	Z score	p-value
Week 1					
Intercept	0.00	[0.00, 0.00]	0.82	–6.43	<0.001
Quiz1	1.07	[1.03, 1.06]	1.91	4.51	<0.001
PR1	1.02	[1.01, 1.04]	2.65	6.30	<0.001
Week 2					
Intercept	0.00	[0.00, 0.00]	1.72	–8.02	<0.001
Quiz2	1.10	[1.02, 1.08]	2.28	5.14	<0.001
PR 2	1.02	[1.03, 1.11]	4.71	7.52	<0.001
Week 3					
Intercept	0.00	[0.00, 0.00]	0.43	–9.14	<0.001
Quiz 3	1.07	[1.05, 1.06]	2.64	5.69	<0.001
PR 3	1.04	[1.06, 1.09]	6.55	8.66	<0.001

Quizzes and homework (PR) grades were employed for the optimal logistic model measured at three weeks in the early semester. During the semester's journey, the importance of normalized homework variables (PR) gradually increased, which can be illustrated by the growing shift at the odds ratio of 2.65 from the first week through the larger odds ratio of 6.55 at week 3. The most significant normalized odds ratio of quiz grades was obtained in week 3 in the same manner on the homework features. The quiz grades' importance increased, contributing to predicting students' grades in the course, representing weekly students' attention to their attendance. Therefore, two features demonstrated the same manner for their predictive importance to the logistic regression model using the weekly variables.

It might be such unimpressed findings if the predictive importance of homework and quiz grades were significant to the model. As educators, one might often employ the method of quizzes and homework in their formative assessment. This proposed model is in line with the aim of this educational routine. Students will finally succeed in the course if they enjoy the learning process as they do the quizzes, submit the assignments, and ensure their full attendance in class.

Table 3. Optimal model using the ensemble of institutional and in-class data

Feature	Odds ratio	95% CI	Norm odds ratio	Z score	p
Week 1					
Intercept	0.00	[0.00, 0.00]	1.26	–9.43	<0.001
Quiz1	1.03	[1.01, 1.04]	1.83	3.57	<0.001
PR1	1.05	[1.03, 1.06]	1.92	3.81	<0.001
UN	1.04	[1.04, 1.07]	1.64	3.55	<0.001
IPK	11.85	[6.63, 21.1]	3.51	8.42	<0.001
Calculus	0.28	[0.18, 0.44]	0.21	–3.96	<0.001
UTBK	0.22	[0.11, 0.59]	0.30	–4.02	<0.001
Week 2					
Intercept	0.00	[0.00, 0.00]	1.32	–9.43	<0.001
Quiz2	1.04	[1.03, 1.08]	2.45	4.38	<0.001
PR2	1.04	[1.01, 1.06]	2.76	4.61	<0.001
UN	1.05	[1.05, 1.09]	1.56	3.57	<0.001
IPK	8.81	[4.85, 15.0]	3.01	6.93	<0.001
Calculus	0.43	[0.33, 0.90]	0.45	–2.52	0.008
UTBK	0.42	[0.26, 0.85]	0.47	–2.91	<0.009
Week 3					
Intercept	0.00	[0.00, 0.00]	1.12	–9.79	<0.001
Quiz3	1.02	[1.02, 1.07]	2.71	4.58	<0.001
PR3	1.04	[1.01, 1.09]	3.80	5.75	<0.001
UN	1.08	[1.04, 1.05]	1.42	2.38	<0.001
IPK	7.52	[4.12, 12.2]	2.77	6.51	<0.001
Calculus	0.38	[0.17, 0.43]	0.24	–3.29	<0.001
UTBK	0.34	[0.23, 0.81]	0.38	–2.55	0.002

3.3. Logistic Regression Model Using Institutional and In-Class Variables

This study finally tried to combine the institutional and in-class features with building the logistic model. Table 3 represents the summary of the logistic modeling results using the combination of a dataset. During the weeks' progress, there are not the most significant variables of the in-class features based on the normalized odds ratio of both quizzes and homework grades. Those values were slightly equal, implying their equal contribution to the predictive model. IPK was the most influential variable for the institutional features from the students and had the highest normalized odds ratio than the in-class variables. UN, UTBK, and Calculus's odds ratio were explicitly less important compared with the IPK grades.

The ensemble method combining the institutional and in-class variables implies that in-class and institutional data cannot accurately capture the prediction towards the students' grades. These results recommended that both features contribute equally to evaluating the students' performance in the introductory physics course.

3.4. Accuracy of the Model

The accuracy of the classifier algorithm is more improved over the course progress, and students have done their assignments. Table 4 below summarizes the evaluation of our model based on accuracy, kappa (κ), and AUC using the in-class features and the combination of in-class and institutional features at the three points early in the semester. Institutional and in-class features are employed at the institutional model, which created an optimal model.

Firstly, the in-class model outperformed the institutional model in the third week in the early semester based on the kappa and AUC. Kappa and AUC are more recommended for the evaluation metrics due to

their ability to overcome the factor of random guessing rather than accuracy. During the first week over the last observation, the ensemble of institutional and in-class models outperformed the in-class only models each week. The models were indistinguishable statistically based on DeLong's test and calculating the AUC of different ROC curves at the third week of the semester. This result suggested that instructors might not need to administer the final examination at the end of the semester because they have accurately predicted the students' outcomes. Logistic classifiers recommend that it was necessary to consider the previous institutional features of the prediction model to know the low-performing students in the early semester.

Table 4. Evaluation of logistic regression model performance

Model	Features	Evaluation metrics			
		Accuracy	κ	AUC	R^2
Baseline	None	0.64	0.00	0.50	
Institutional	All	0.72	0.36	0.76 [0.74, 0.81]	0.34
	Optimal	0.71	0.31	0.78 [0.71, 0.80]	0.32
Week 1	In-Class	0.70	0.25	0.72 [0.64, 0.75]	0.18
	In-Class and Institutional	0.74	0.38	0.80 [0.79, 0.82]	0.35
Week 2	In-Class	0.75	0.48	0.80 [0.76, 0.87]	0.31
	In-Class and Institutional	0.78	0.53	0.88 [0.85, 0.90]	0.46
Week 3	In-Class	0.80	0.56	0.89 [0.86, 0.95]	0.54
	In-Class and Institutional	0.78	0.54	0.84 [0.83, 0.91]	0.48

3.5. How Well Students' Performance can be Predicted Through Machine Learning Employed in the Study ?

Physics education should address the importance of formative assessment to facilitate feedback for the students. Lecturers should be aware of the formative evaluation and respond to students with constructive feedback about their learning progressions. It is time-consuming [19] if they want to reach all the students, even in a department, for individual treatment. It is challenging to scale to all students, especially in large classes, and it will be predicted if the distance learning may remain for the general physics course in the future.

The findings obtained in this study may initiate educational data to be mined for educational purposes. An extensive collection of datasets assist the teachers in monitoring their student's progress during the semester by providing real-time feedback through a learning management system. A logistic classifier that is employed in the modeling process enables educators to offer personal feedback that will help and construct their students within the learning process to explain the physical concepts of engineering problems.

Improvements for future studies are recommended to overcome the study's limitations. The first weakness is that only a logistic classifier is trained in this study. Hence, future work is suggested to invite another model to predict students' performance in introductory physics course. For example, recurrent neural networks (RNN) would be assumed as a better prediction model based on reward and punishment theory. The second limitation is the educational feature engaged in the model. Although institutional and in-class variables are employed, they have not created the best predictions. There is a bias within the false-positive rate among the model based on confusion matrices analysis. Therefore, our findings' generalizability may still be an arguable position due to the localized educational environment participated in the study. Consequently, large-scale participants should be more invited to construct the models.

4. CONCLUSION

Logistic regression was employed in this study to predict students' grades on introductory physics courses for engineering students. The machine learning model should be improved before it might be implemented in the classroom. It produced good predictions at an 80% identification rate with the feature of in-class variable and institutional variable, and it recommended a proof of concept. Observation at the five points in the early semester should be considered an arguable conclusion even with its high accuracy. While it leads to a significant reduction in the number of students with potentially lower achievement, the model does not decrease enough to ensure the model's viability for providing the individual intervention as instructors that would assist the students.

AUTHORS' CONTRIBUTIONS

PHS, BK and H design the study and write the first draft. SB, W, and JS provided assistance in the data c.

ACKNOWLEDGMENTS

We would like to express our highest gratitude for the Ministry of Education, Culture, Research, and

Technology cooperating with the Indonesia Endowment Fund for Education that have provided doctoral fundings for the first author through the Indonesian Educational Scholarship 2021.

REFERENCES

[1] G. Pahl, W. Beitz, J. Feldhusen, and K. H. Grote, *Engineering design: A systematic approach*. 2007. doi: 10.1007/978-1-84628-319-2.

[2] A. Van den Beemt *et al.*, "Interdisciplinary engineering education: A review of vision, teaching, and support," *Journal of Engineering Education*, vol. 109, no. 3. 2020. doi: 10.1002/jee.20347.

[3] M. E. Jordan and R. R. McDaniel, "Managing Uncertainty During Collaborative Problem Solving in Elementary School Teams: The Role of Peer Influence in Robotics Engineering Activity," *Journal of the Learning Sciences*, vol. 23, no. 4, 2014, doi: 10.1080/10508406.2014.896254.

[4] K. Y. Lin, Y. T. Wu, Y. T. Hsu, and P. J. Williams, "Effects of infusing the engineering design process into STEM project-based learning to develop preservice technology teachers' engineering design thinking," *International Journal of STEM Education*, vol. 8, no. 1, 2021, doi: 10.1186/s40594-020-00258-9.

[5] A. M. M. S. Ullah and K. H. Harib, "Tutorials for integrating CAD/CAM in engineering curricula," *Education Sciences*, vol. 8, no. 3, 2018, doi: 10.3390/educsci8030151.

[6] D. Um, *Solid modeling and applications: Rapid prototyping, CAD and CAE theory: Second Edition*. 2018. doi: 10.1007/978-3-319-74594-7.

[7] C. Vieira, M. Hathaway Goldstein, Ş. Purzer, and A. J. Magana, "Using Learning Analytics to Characterize Student Experimentation Strategies in the Context of Engineering Design," *Journal of Learning Analytics*, vol. 3, no. 3, 2016, doi: 10.18608/jla.2016.33.14.

[8] H. S. Lee, G. H. Gweon, T. Lord, N. Paessel, A. Pallant, and S. Pryputniewicz, "Machine Learning-Enabled Automated Feedback: Supporting Students' Revision of Scientific Arguments Based on Data Drawn from Simulation," *Journal of Science Education and Technology*, vol. 30, no. 2, 2021, doi: 10.1007/s10956-020-09889-7.

[9] Y. Dwiyono, W. G. Mulawarman, P. O. Pramono, N. A. Salim, and M. Ikhsan, "Implementation of national examination based on Computer Based Test at Vocational School 1 North Sangatta," *Cypriot Journal of Educational Sciences*, vol. 16, no. 1, 2021, doi: 10.18844/cjes.v16i1.5510.

[10] LTMPT, "Informasi UTBK-SBMPTN 2020," *Lembaga Tes Masuk Perguruan Tinggi*, vol. 1, 2020.

[11] G. Mahajan and B. Saini, "Educational Data Mining: A state-of-the-art survey on tools and techniques used in EDM," *International Journal of Computer Applications & Information Technology*, vol. 12, no. 1, 2020.

[12] F. Pedregosa *et al.*, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, 2011.

[13] F. Ratzinger *et al.*, "Machine learning for fast identification of bacteraemia in SIRS patients treated on standard care wards: a cohort study," *Scientific Reports*, vol. 8, no. 1, 2018, doi: 10.1038/s41598-018-30236-9.

[14] D. G. Altman, *Practical Statistics for Medical Research*. 1990. doi: 10.1201/9780429258589.

[15] P. Black and D. Wiliam, "Assessment and Classroom Learning, Assessment in Education: Principles, Policy & Practice," *Assessment in Education*, vol. 5, no. 1, 1998.

[16] W. E. Maddox, T. D. Thiede, S. H. Cobb, S. R. Hickman, and J. Crofton, "Design Considerations In Engineering Physics: Integrating Design Across The Engineering Physics Curriculum," 2020. doi: 10.18260/1-2--8266.

[17] L. K. Berland and K. C. Busch, "Negotiating STEM epistemic commitments for engineering design challenges," 2012. doi: 10.18260/1-2--21726.

[18] S. R. Bartholomew and G. J. Strimel, "Factors influencing student success on open-ended design problems," *International Journal of Technology and Design Education*, vol. 28, no. 3, 2018, doi: 10.1007/s10798-017-9415-2.

[19] S. R. Bartholomew, G. J. Strimel, and E. Yoshikawa, "Using adaptive comparative judgment for student formative feedback and learning during a middle school design project," *International Journal of Technology and Design Education*, vol. 29, no. 2, 2019, doi: 10.1007/s10798-018-9442-7.

[20] F. Ye, Q. Huang, S. Wu, and Y. Chen, "Talking Avatar: An intelligent mobile application based on third-party cloud services," *International Journal of Technology and Human Interaction*, vol. 15, no. 3, 2019, doi: 10.4018/IJTHI.2019070101.