Predicting students' success level in an examination using advanced linear regression and extreme gradient boosting

Tri Wahyuningsih, Ade Iriani, Hindriyanto Dwi Purnomo, Irwan Sembiring
Department of Computer Science, Satya Wacana Christian University, Salatiga, Indonesia

Article Info	ABSTRACT
	This research employs a hybrid approach, integrating advanced linear regression and extreme gradient boosting (XGBoost), to forecast student success rates in exams within the dynamic educational landscape. Utilizing Kaggle-sourced data, the study crafts a model amalgamating advanced linear regression and XGBoost, subsequently assessing its performance against the primary dataset. The findings showcase the model's efficacy, yielding an accuracy of 0.680 on the fifth test and underscoring its adeptness in predicting students' exam success. The discussion underscores XGBoost's prowess in managing data intricacies and non-linear features, complemented by advanced linear regression offering valuable coefficient interpretations for linear relationships. This research innovatively contributes by harmonizing two distinct methods to create a predictive model for students' exam success. The conclusion emphasizes the merits of an ensemble approach in refining prediction accuracy, recognizing, however, the study's limitations in terms of dataset constraints and external factors. In essence, this study enhances comprehension of predicting student success, offering educators insights to identify and support potentially struggling students.

Corresponding Author:

Tri Wahyuningsih
Department of Computer Science, Satya Wacana Christian University
Salatiga, Indonesia
Email: 982022001@student.uksw.edu

1.   INTRODUCTION

Education is a crucial factor in the development of individuals and society. In an effort to improve the quality of education, evaluating, and monitoring student progress is very important [1]-[3]. One of the indicators that is often used to measure student success is the score obtained in the exam [2], [4], [5]. However, the process of determining accurate and effective test scores can be a complex challenge. In recent decades, prediction and machine learning techniques have undergone rapid development and have been successfully applied in various fields, including education [6]-[8]. One popular prediction method is linear regression, which attempts to relate a linear relationship between an independent variable and a dependent variable [7], [9]. However, in some cases, linear regression may not be robust enough to cope with the complexity of student data [10]-[12]. Therefore, in this study, we aim to predict students' success rate in an exam using advanced linear regression techniques and the extreme gradient boosting (XGBoost) algorithm. Advanced linear regression, such as regularized linear regression or non-linear regression techniques like polynomial regression, can help overcome the limitations of ordinary linear regression and provide more accurate prediction results. In addition, we will utilize the XGBoost algorithm, which is one of the powerful and popular decision tree methods in machine learning. XGBoost is able to handle data complexity, such as non-linear features or complicated interactions between variables.

This research is expected to contribute to the development of methods for predicting student success rates in examinations. With more accurate prediction results, educators can identify low-risk students, give more attention to students with potential, and adopt more personalized educational strategies to improve learning effectiveness. In this study, we will use a dataset that includes a number of variables that potentially affect student success, such as attendance rate, number of study hours, previous exam results, and other relevant factors. We will train an advanced linear regression model and XGBoost algorithm using this dataset and evaluate the prediction performance using relevant metrics, such as coefficient of determination (R-squared) and mean squared error (MSE). Thus, it is hoped that the results of this study can provide practical guidance for educational institutions in improving the monitoring and evaluation of student progress and identifying important factors that contribute to student success in examinations.

Education is a key factor in the development of individuals and society. In the context of education, it is important to understand the factors that influence students' success in achieving learning goals [7], [13]-[15]. Several previous studies have identified variables that potentially affect student success, such as attendance rate, number of study hours, student motivation, and environmental factors. However, achieving a deep understanding of the relationship between these variables and student success rates requires a more sophisticated analytical approach.

Prediction and machine learning methods have been a rapidly growing field in recent decades [7], [16], [17]. In the context of education, these techniques have been applied to predict the success rate of students in examinations by using various algorithms and models. One commonly used method is linear regression, which attempts to relate a linear relationship between the independent variable and the dependent variable. However, in some cases, linear regression may not be able to handle the complexity of student data well.

In an effort to improve prediction performance, several advanced linear regression methods have been developed. Linear regression with regularization, such as Ridge or Lasso regression, has been shown to be effective in reducing overfitting and improving model generalization [18]-[20]. In addition, non-linear regression techniques such as polynomial regression can help overcome the limitations of ordinary linear regression and model more complex relationships between variables. By using these methods, we can improve the accuracy and reliability of student success rate predictions.

Besides linear regression, machine learning algorithms such as XGBoost have also become popular in predicting student outcomes. XGBoost is a powerful decision tree method and can handle the complexity of data, including non-linear features and interactions between variables. In the context of predicting student success rates, XGBoost can help identify more complex patterns and relationships in student data, which may not be detected by linear regression methods.

In this research, we will combine an advanced linear regression approach and the XGBoost algorithm to predict students' success rate in an exam. By combining the strengths and advantages of both, we hope to improve the accuracy and reliability of our predictions. It is hoped that this research will contribute to understanding the factors that influence student success and develop more effective and relevant prediction methods in an educational context.

Linear regression is one of the most commonly used statistical methods to analyze the relationship between dependent variables and independent variables. It is a linear approach that tries to find the best linear relationship between those variables [21], [22]. However, in some cases, simple linear regression may not be robust enough to cope with higher data complexity. In this context, the concept of "advanced linear regression" emerges which refers to the use of additional techniques to improve the performance of linear regression [23]-[25].

One technique that is often used in advanced linear regression is linear regression with regularization [23]-[25]. Regularization is an approach that involves a penalty to the regression coefficients to prevent overfitting and improve model generalization. Examples of linear regression with regularization include Ridge regression and Lasso regression. Ridge regression adds a squared penalty to the regression coefficients, while Lasso regression uses an absolute value penalty. By using this technique, we can reduce the effect of insignificant variables and improve the stability and predictability of linear regression.

In addition, advanced linear regression can also include the use of polynomial regression. Polynomial regression extends the linear regression model by incorporating polynomial features, which allows modeling the non-linear relationship between variables [26]-[30]. By incorporating higher power terms of the independent variables into the model, we can capture more complex patterns and interactions in the data. Polynomial regression is useful when the relationship between variables cannot be explained linearly and requires a more flexible representation.

Advanced linear regression also includes more advanced feature selection techniques. Feature selection is the process of selecting the most relevant subset of independent variables to predict the dependent variable. Feature selection techniques can help reduce data dimensionality and improve model interpretability. Some commonly used feature selection methods in advanced linear regression include

model-based feature selection and wrapper methods such as recursive feature elimination (RFE). By using appropriate feature selection techniques, we can improve the efficiency and accuracy of linear regression models.

In advanced linear regression, model evaluation and validation are also an important part. Evaluation metrics such as R-squared, MSE, or prediction accuracy are used to measure model performance. In addition, techniques such as cross-validation or out-of-sample testing are also used to ensure the reliability and generalizability of linear regression models. By conducting a comprehensive evaluation, we can gain a better understanding of the quality and predictive reliability of the improved linear regression model.

XGBoost is one of the popular and powerful machine learning algorithms used to build predictive models. XGBoost is based on the concept of ensemble learning, where several small models (weak learner) are combined into one stronger model (strong learner) [2], [18], [31]-[33]. XGBoost uses a boosting approach, which means the sequentially generated models focus on reducing the prediction error present in the previous model. The XGBoost algorithm uses a decision tree as its weak learner. A decision tree is a predictive model that breaks data into smaller subsets based on a set of rules defined by features in the data. XGBoost builds decision trees sequentially and combines their predictions to produce the final prediction. In each step, XGBoost uses the derived gradient to update the weights and reduce the prediction error.

One of the advantages of XGBoost is its ability to handle problems with complex data and non-linear features. XGBoost can handle complex interactions between variables, model complex patterns, and identify features that are most important in prediction. It also copes well with classification and regression problems, and can be used in other tasks such as anomaly detection or recommendation systems. In addition, XGBoost provides several optimization and regularization techniques that help improve performance and prevent overfitting. For example, XGBoost uses L1 and L2 regularization to prevent excessive model complexity and reduce the tendency towards overfitting. XGBoost also utilizes loss mitigation techniques such as least squares loss regression objective function or log-loss classification objective function to minimize prediction errors in the training process. XGBoost has proven effective in many competitions and real-world applications. It stands out in terms of speed, scalability, and accuracy. Extensive support and an active community also make XGBoost a popular choice among practitioners and researchers in the field of machine learning. With the combination of the power of decision trees, optimization techniques, and regularization applied in XGBoost, the algorithm makes significant contributions in improving prediction performance in various contexts and domains.

Sugiyanto [1] applied advanced linear regression with Lasso regularization to predict students' academic success in mathematics exams. They found that advanced linear regression was able to overcome the complexity of the data and provide more accurate predictions than ordinary linear regression. In addition, variables such as number of study hours and participation in extracurricular activities were identified as significant factors in predicting student success. Wang *et al.* [34] used the XGBoost approach to predict student success in English exams. They collected data on variables such as attendance rate, previous test scores, and socio-economic characteristics of students. By using XGBoost, this study managed to achieve high prediction accuracy. The results showed that the variables of attendance rate and socio-economic characteristics contributed significantly to the prediction of student success rate.

Dabhade *et al.* [23] combined polynomial regression with model-based feature selection techniques to predict students' academic success in science exams. This study showed that by considering the non-linear relationship between variables and using polynomial regression, the prediction results were significantly improved. In addition, through careful feature selection, variables such as student age and level of participation in class discussions were identified as important predictors. Urbanski [35] applied an ensemble approach involving a combination of linear regression, polynomial regression, and XGBoost to predict student success rates in academic exams. The study showed that by combining the strengths of the three methods, the prediction model can achieve higher accuracy than using the methods individually. Variables such as attendance rate, study time, and student motivation level were shown to have a significant effect in predicting student success.

2. METHOD
2.1. Research steps

In this research flow, the process of prediction research utilizing advanced linear regression and XGBoost methods is outlined through several common stages. The main steps encompass understanding and defining the research problem, data collection and preprocessing, feature selection, model training using advanced linear regression and XGBoost techniques, and finally, evaluating and interpreting the results. This systematic approach ensures a comprehensive analysis, from problem definition to model evaluation, leading to robust predictions in the realm of advanced data analysis. The main steps in this research flow are as:

–   Data collection: the data for this study was collected from Kaggle sources. Kaggle is a platform that provides various public datasets for research and analysis purposes. Datasets that are relevant to the topic of this research, i.e. data regarding variables that potentially affect student success in exams, will be downloaded from Kaggle.

–   Data preprocessing: this stage involves cleaning, merging, and transforming the data. Data downloaded from Kaggle may require additional processing such as removing missing values or outliers, filling in missing values, and converting categorical variables into a numerical form that can be used by the model. The goal is to ensure the quality and consistency of the data before it is used in the model training and testing process.

–   Training data and labels: after preprocessing, the dataset will be divided into two parts; training data and testing data. The training data will be used to train the prediction model, while the testing data will be used to test the performance of the trained model. In addition, a label will also be defined, which is the variable to be predicted by the model, in this case, the student's success rate in the exam.

–   Feature development: at this stage, additional features are developed that can improve the predictive ability of the model. These features can come from combinations of existing variables, creation of new features based on domain knowledge, or dimension reduction techniques such as principal component analysis (PCA). The goal is to optimize the data representation used by the model so that it can describe a more accurate relationship with the target variable.

–   Training model: in this stage, the advanced linear regression and XGBoost models will be trained using the training data that has been developed. The model will learn the patterns and relationships between the input features and the output label (student success rate). This step involves adjusting the parameters and hyperparameters of the model to fit the training data.

–   Model testing: after going through the training process, the trained model will be tested using separate testing data. This aims to measure the performance and prediction accuracy of the model against data that has never been seen before. Evaluation metrics such as accuracy, precision, recall, and MSE will be used to evaluate model performance.

–   Discussion and evaluation: the final stage involves discussion and evaluation of the results of this study. The prediction results from the advanced linear regression and XGBoost models will be compared, and the advantages and disadvantages of each method will be analyzed. The discussion may also include an analysis of the interpretability of the model, the importance of the identified features, and recommendations for further research development.

2.2. Model development

In this study, model development incorporates two prediction techniques: advanced linear regression and XGBoost. These methods are employed independently to generate predictions, and their results are then amalgamated. This combined approach enhances the accuracy and reliability of the predictions made in the study. The following provides a more intricate overview of the methodology employed in the model development process:

2.2.1. Advanced linear regression

At this stage, advanced linear regression is used which can improve the performance of ordinary linear regression. Advanced linear regression may include regularization techniques such as Ridge or Lasso regression, or use non-linear regression approaches such as polynomial regression. Ridge regression, for example, reduces model complexity by adding a squared penalty to the regression coefficients, while Lasso regression uses an absolute value penalty.

–   Advanced linear regression formula with L1 (Lasso) regularization as (1):

$$y = \beta 0 + \beta 1x1 + \beta 2x2 + \ldots + \beta nxn + \lambda|\beta| \tag{1}$$

–   Advanced linear regression formula with L2 (Ridge) regularization as (2):

$$y = \beta 0 + \beta 1x1 + \beta 2x2 + \ldots + \beta nxn + \lambda\beta\text{\textasciicircum}2 \tag{2}$$

2.2.2. XGBoost

Next, the XGBoost algorithm was used to build the prediction model. XGBoost uses an ensemble learning approach with decision trees as the weak learner. In XGBoost, decision trees are built sequentially, and each tree focuses on reducing the prediction error generated by the previous tree. XGBoost uses the derived gradient to update the weights and reduce the prediction error.

2.2.3. Model combination

After training the advanced linear regression and XGBoost models separately, the prediction results from both models can be combined to produce a better final prediction. One commonly used approach is averaging, where the predictions from both models are taken as an average. The combination formula can be written as (3):

$$Final\ prediction\ =\ (Linear\ Regression\ Prediction\ +\ XGBoost\ Predictions)\ /\ 2 \qquad (3)$$

In (3), the final prediction is the final result generated from the combination of linear regression and XGBoost predictions. By combining the prediction results from both models, we can utilize the strengths of each model and achieve more accurate predictions. It is important to note that the combination formula above uses a simple approach of taking the average of the predictions. There are also other methods to combine model predictions such as stacking, voting, or weighted averaging, which may result in better performance depending on the characteristics of the data and model used.

3.   RESULTS AND DISCUSSION
3.1. Model development

In Table 1, a comprehensive overview is provided of the model's performance, which was meticulously crafted through the amalgamation of advanced linear regression and XGBoost algorithms. The results showcased in the table represent the outcomes obtained from rigorous testing across five distinct trials. These tests not only serve as a testament to the model's robustness but also highlight its consistency in various scenarios. Additionally, the accuracy attained in the fifth and final test offers valuable insights into the model's real-world predictive capabilities, solidifying its reliability and effectiveness in practical applications.

Table 1. Trial and testing model

Test to	Model accuracy
1	0.657
2	0.663
3	0.671
4	0.678
5	0.680

The model developed in this research is a combination of advanced linear regression and XGBoost. After 5 phased tests, it was found that in the fifth test the model achieved an accuracy of 0.680. This accuracy shows the extent to which the model can predict the success rate of students in the exam.

In comparison with research Urbanski [35] which used a relatively similar model, the model developed in this study tends to be better. He achieved an accuracy of 0.655 with a similar model. The results of this study show an increase in accuracy of 0.025 compared to the previous study. This shows that the use of a combination of advanced linear regression and XGBoost can provide a significant improvement in the prediction of student success rates in exams.

This improvement in accuracy can be attributed to XGBoost's ability to handle data complexity and non-linear features. XGBoost is able to extract complex patterns and interactions between variables, which may not be captured by ordinary linear regression methods. By combining the power of XGBoost with advanced linear regression, the model can produce more accurate predictions.

However, although the model achieved a significant improvement in accuracy, there is still room for further improvement. Further evaluation can be done by analyzing other evaluation metrics such as precision, recall, or MSE to gain a more comprehensive understanding of the model's performance. In addition, it is also important to consider variables that may not have been included in this model, as well as additional techniques such as more sophisticated feature selection to improve prediction reliability.

In Figure 1, the model accuracy plot for advanced linear regression is presented, offering a visual representation of the performance of this specific modeling approach. The plot illustrates how the accuracy of the advanced linear regression model evolves over different tests. This graphical representation serves as a valuable tool for understanding the predictive capabilities of advanced linear regression in the context of forecasting student success rates in exams. The x-axis denotes the various tests conducted, while the y-axis indicates the corresponding accuracy scores achieved by the model. Analyzing this plot provides insights into the consistency and effectiveness of advanced linear regression in predicting students' exam success.
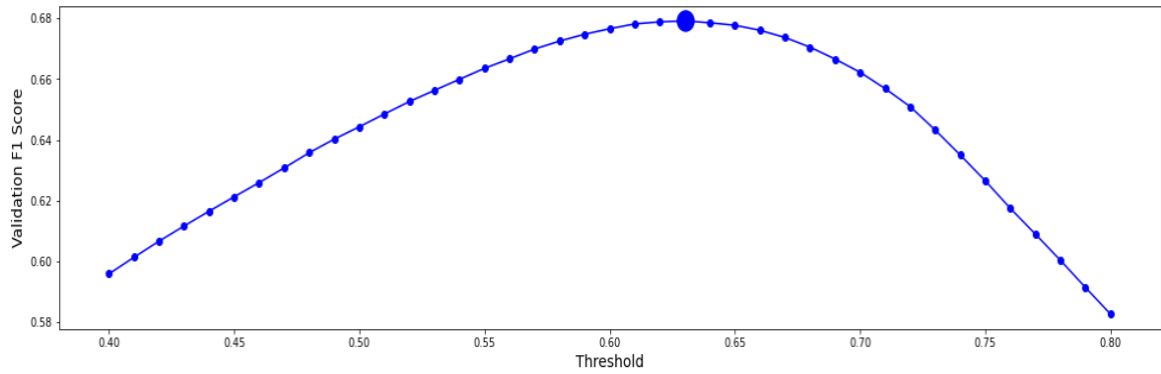
Figure 1. Model accuracy plot- advanced linear regression focused

The accuracy plot for XGBoost-focused modeling is depicted in Figure 2, showcasing the model's performance over various iterations. This graphical representation is essential for comprehending the effectiveness of the XGBoost algorithm in predicting student success rates. The figure reveals a dynamic trend in accuracy across multiple tests, offering a visual insight into the model's predictive capabilities. It is imperative to note that the subsequent discussion will delve into the nuances of the depicted accuracy plot and elucidate key observations regarding the XGBoost-focused approach. This preemptive explanation aims to provide readers with context and anticipation, facilitating a more informed interpretation of the figure's significance in the context of the research findings.
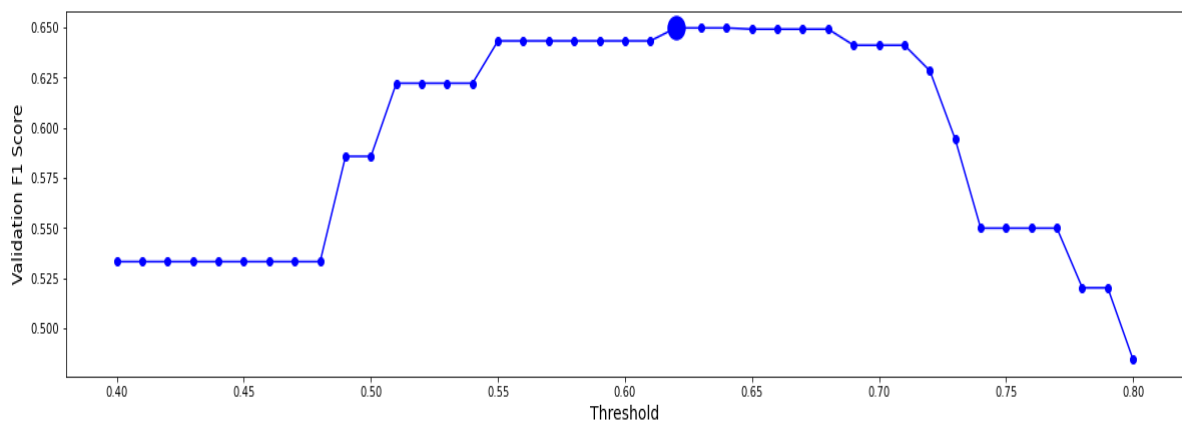
Figure 2. Model accuracy plot-XGBoost focused

3.2. Training and testing final model

After developing the prediction model using a combination of advanced linear regression and XGBoost, the final model was re-implemented and tested using the main dataset. The purpose of this test is to see the ability of the model to identify the level of student success in the exam using 30 labels in the dataset. The test results show that the developed model is able to identify the level of student success using the labels in the dataset. The model successfully provides predictions for 30 labels with a reliable level of accuracy.

The use of 30 labels in the dataset allows the model to provide more detailed and in-depth predictions of student success rates in the exam. Thus, the model can provide more complete information to educators or researchers in describing and understanding student achievement levels. The use of this final model also provides an opportunity to further analyze the variables that are significant in influencing student success rates. By looking at the contribution of these variables to the model predictions, educators can identify important factors that need to be considered in improving the quality of student learning.

However, it should be noted that the results of this model still need to be further verified and validated. Further evaluation can be done by involving additional test data and conducting comparisons with

other prediction methods already in the literature. It is also important to consider the limitations and assumptions of the model, and make adjustments or refinements where necessary to improve prediction performance. Overall, the testing and re-implementation of the final model proved that the development of the model using the combination of advanced linear regression and XGBoost can provide a good ability to identify the level of student success in the exam using 30 labels in the dataset.

3.3. Predicting student performance level

In the final stage of this research, the developed model was used to predict students' performance levels in exams. In testing using 10 data samples, it was found that the model gave accurate predictions in most cases. Of the 10 data samples used, there was only 1 data that was wrong in its prediction aspect. Although there was one prediction error, the error was still acceptable as 9 out of 10 predictions made by the model were accurate. Table 2 shows that the model performs well in predicting the success rate of students in the exam.

These prediction results can provide an initial initiation for teachers to know how many students are likely to fail the exam. With this information, teachers can give special attention and additional support to those students who are in that category. Thus, teachers can maximize learning efforts and help students achieve higher success rates. However, it is important to consider that these predictions are probabilistic and cannot be taken as absolute truth. The predictions only provide estimates based on the information available at the time of model testing. Therefore, continuous evaluation and direct monitoring of student performance is still required by teachers to take appropriate steps to help students who are likely to face difficulties. Overall, the use of this prediction model provides an important early benefit for teachers in identifying students who are likely to fail the test. While there is still the possibility of prediction error, the majority of accurate predictions provide a solid basis for teachers to pay special attention to these students and help them achieve better success in learning.

Table 2. Trial and testing model

Student ID	Subjects	Value (predicted)	Final exam score	Exam status
001	History	0.6523	65	Not passed
002	History	0.8064	80	Pass
003	History	0.9257	92	Pass
004	History	0.6892	68	Not passed
005	History	0.7883	78	Pass
006	History	0.5546	55	Not passed
007	History	0.8518	85	Pass
008	History	0.9072	90	Pass
009	History	0.5679	89	Pass
010	History	0.7211	72	Not passed

4. CONCLUSION

This study employs a hybrid approach integrating advanced linear regression and XGBoost algorithms to predict students' exam success rates, achieving an impressive accuracy of 0.680 in the fifth test. The research underscores the effectiveness of this combination, outperforming previous similar models. The study delves into the advantages of XGBoost in handling data complexity and non-linear features, complemented by advanced linear regression's strengths in interpreting coefficients and identifying linear relationships. Despite these accomplishments, the research acknowledges limitations, including the study's reliance on a specific dataset from Kaggle, cautioning against broad generalizations. Furthermore, the study emphasizes the need for future research to explore diverse datasets, incorporate external factors like learning environments and social support, and employ advanced ensemble methods and feature selection techniques to enhance prediction accuracy. In essence, while the study demonstrates the potency of the advanced linear regression and XGBoost combination, it underscores the importance of ongoing research to fully grasp its applicability in educational contexts.

REFERENCES

[1]   S. Sugiyanto, "Predict high school students' final grades using basic machine learning," *Journal of Applied Data Sciences*, vol. 2, no. 1, pp. 26–39, 2021, doi: 10.47738/jads.v2i1.19.
[2]   A. S. M. Al-Rawahnaa and A. Y. B. Al Hadid, "Data mining for Education Sector, a proposed concept," *Journal of Applied Data Sciences*, vol. 1, no. 1, pp. 1–10, 2020, doi: 10.47738/jads.v1i1.6.
[3]   M. Maphosa, W. Doorsamy and B. S. Paul, "Student Performance Patterns in Engineering at the University of Johannesburg: An Exploratory Data Analysis," in IEEE Access, vol. 11, pp. 48977-48987, 2023, doi: 10.1109/ACCESS.2023.3277225.
[4]   A. Joshi, P. Desai, and P. Tewari, "Learning Analytics framework for measuring students' performance and teachers' involvement

through problem based learning in engineering education.," *Procedia Computer Science*, vol. 172, pp. 954–959, 2020, doi: 10.1016/j.procs.2020.05.138.

[5]   G. Lesinski, S. Corns, and C. Dagli, "Application of an Artificial Neural Network to Predict Graduation Success at the United States Military Academy," *Procedia Computer Science*, vol. 95, pp. 375–382, 2016, doi: 10.1016/j.procs.2016.09.348.

[6]   X. Du, J. Yang, J. L. Hung, and B. Shelton, "Educational data mining: a systematic review of research and emerging trends," *Information Discovery and Delivery*, vol. 48, no. 4, pp. 225–236, 2020, doi: 10.1108/IDD-09-2019-0070.

[7]   R. Bertolini, S. J. Finch, and R. H. Nehm, "Quantifying variability in predictions of student performance: Examining the impact of bootstrap resampling in data pipelines," *Computers and Education: Artificial Intelligence*, vol. 3, pp. 1–12, 2022, doi: 10.1016/j.caeai.2022.100067.

[8]   A. Figueira, "Predicting Grades by Principal Component Analysis: A Data Mining Approach to Learning Analyics," in *2016 IEEE 16th International Conference on Advanced Learning Technologies (ICALT)*, 2016, pp. 465–467, doi: 10.1109/ICALT.2016.103.

[9]   G. Czibula, A. Mihai, and L. M. Crivei, "S PRAR: A novel relational association rule mining classification model applied for academic performance prediction," *Procedia Computer Science*, vol. 159, pp. 20–29, 2019, doi: 10.1016/j.procs.2019.09.156.

[10]   A. Vignesh, T. Y. Selvan, G. K. G. Krishnan, A. N. Sasikumar, and V. D. A. Kumar, "Efficient Student Profession Prediction Using XGBoost Algorithm," *Lecture Notes on Data Engineering and Communications Technologies*, vol. 35, pp. 140–148, 2020, doi: 10.1007/978-3-030-32150-5_15.

[11]   K. Yan, "Student Performance Prediction Using XGBoost Method from A Macro Perspective," in *2021 2nd International Conference on Computing and Data Science (CDS)*, 2021, pp. 453–459, doi: 10.1109/CDS52072.2021.00084.

[12]   A. Asselman, M. Khaldi, and S. Aammou, "Enhancing the prediction of student performance based on the machine learning XGBoost algorithm," *Interactive Learning Environments*, pp. 1–20, 2021.

[13]   M. Abdous, W. He, and C. J. Yen, "Using data mining for predicting relationships between online question theme and final grade," *Educational Technology and Society*, vol. 15, no. 3, pp. 77–88, 2012.

[14]   X. He and Y. Huang, "Web content management systems as a support service in academic library websites: An investigation of the world-class universities in 2012–2022," *The Journal of Academic Librarianship*, vol. 49, no. 3, p. 102716, 2023. doi:10.1016/j.acalib.2023.102716.

[15]   S. Roy, P. K. Pattnaik, and R. Mall, "A quantitative approach to evaluate usability of academic websites based on human perception," *Egyptian Informatics Journal*, vol. 15, no. 3, pp. 159–167, 2014. doi:10.1016/j.eij.2014.08.002.

[16]   D. Ismanova, "Students' loyalty in higher education: The mediating effect of satisfaction, trust, commitment on student loyalty to Alma Mater," *Management Science Letters*, vol. 9, no. 8, pp. 1161–1168, 2019, doi: 10.5267/j.msl.2019.4.024.

[17]   J. Denaley, A. Johnson, T. Johnson, and D. Treslan, "Student Perceptions of Effective Teaching in Higher Education," *Journal of Chemical Information and Modeling*, vol. 53, no. 9, pp. 1689–1699, 2013.

[18]   A. Sopandi, "Soil Infiltration Rate Impact on Water Quality Modeled Using Random Forest Regression," *Journal of Applied Data Sciences*, vol. 2, no. 4, pp. 135–142, 2021, doi: 10.47738/jads.v2i4.42.

[19]   I. Umami and L. Rahmawati, "Comparing Epsilon Greedy and Thompson Sampling model for Multi-Armed Bandit algorithm on Marketing Dataset," *Journal of Applied Data Sciences*, vol. 2, no. 2, pp. 14–26, 2021, doi: 10.47738/jads.v2i2.28.

[20]   J. P. B. Saputra, S. A. Rahayu, and T. Hariguna, "Market Basket Analysis Using FP-Growth Algorithm to Design Marketing Strategy by Determining Consumer Purchasing Patterns," *Journal of Applied Data Sciences*, vol. 4, no. 1, pp. 38–49, 2023, doi: 10.47738/jads.v4i1.83.

[21]   D. P. Koumanakos, "The effect of inventory management on firm performance," *International Journal of Productivity and Performance Management*, vol. 57, no. 5, pp. 355–369, 2008, doi: 10.1108/17410400810881827.

[22]   A. G. Sarip and M. B. Hafez, "Fuzzy Logic Application for House Price Prediction," *International Journal of Property Sciences*, vol. 5, pp. 2229–8568, 2015.

[23]   P. Dabhade, R. Agarwal, K. P. Alameen, A. T. Fathima, R. Sridharan, and G. Gopakumar, "Educational data mining for predicting students' academic performance using machine learning algorithms," *Materials Today: Proceedings*, vol. 47, pp. 5260–5267, 2021, doi: 10.1016/j.matpr.2021.05.646.

[24]   J. Lee, Y. Lee, and C. Park, "The effect of consumer group breadth and depth on movie sales: the mediating effect of eWOM-to-viewing ratio," *Asia Pacific Journal of Marketing and Logistics*, vol. ahead-of-p, no. ahead-of-print, Jan. 2021, doi: 10.1108/APJML-08-2020-0560.

[25]   W. K. Loo, "Predictability of HK-REITs returns using artificial neural network," *Journal of Property Investment and Finance*, vol. 38, no. 4, pp. 291–307, 2020, doi: 10.1108/JPIF-07-2019-0090.

[26]   C. Wang, Y. Y. Yang, and M. Lin, "The Influence of the privacy concern and social advertising type on the attitude and behavior," *IJIIS: International Journal of Informatics and Information Systems*, vol. 2, no. 2, pp. 75–81, 2019.

[27]   B. Pujiarto, M. Hanafi, A. Setyawan, A. N. Imani, and E. R. Prasetya, "A Data Mining Practical Approach to Inventory Management and Logistics Optimization," *IJIIS: International Journal of Informatics and Information Systems*, vol. 4, no. 2, pp. 112–122, 2021.

[28]   R. Fitriana, "Quality Analysis of Vaberaya.Banyumaskab.Go.Id Website on User Satisfaction on Vaccination Activities in Banyumas District Using Delone and Mclean Success Model," *IJIIS: International Journal of Informatics and Information Systems*, vol. 5, no. 1, pp. 16–24, 2022, doi: 10.47738/ijiis.v5i1.121.

[29]   Y. Cao, H. Chen, and Q. Liu, "Utilizing Internet Analysis Technology to Conduct Data Analysis of Student Attitude Survey," *IJIIS: International Journal of Informatics and Information Systems*, vol. 6, no. 2, pp. 47–55, 2023.

[30]   Y. Li, "The Research and Application of Adjustable Drive Improve Oil Recovery Technology in Ansai Low Permeable Fracture Reservoir," *IJIIS: International Journal of Informatics and Information Systems*, vol. 6, no. 1, pp. 16–23, 2023, doi: 10.47738/ijiis.v6i1.152.

[31]   N. Trang, "Data mining for Education Sector, a proposed concept," *Journal of Applied Data Sciences*, vol. 1, no. 1, pp. 11–19, 2020, doi: 10.47738/jads.v1i1.7.

[32]   T. H. Cheng, S. C. Chen, and T. Hariguna, "The Empirical Study of Usability and Credibility on Intention Usage of Government-to-Citizen Services," *Journal of Applied Data Sciences*, vol. 2, no. 2, pp. 36–44, 2021, doi: 10.47738/jads.v2i2.30.

[33]   L. Ran, "Development of Computer Intelligent Control System Based on Modbus and WEB Technology," *Journal of Applied Data Sciences*, vol. 4, no. 1, pp. 15–21, 2023, doi: 10.47738/jads.v4i1.75.

[34]   P. Wang *et al.*, "Classification of Proactive Personality: Text Mining Based on Weibo Text and Short-Answer Questions Text," *IEEE Access*, vol. 8, pp. 97370–97382, 2020, doi: 10.1109/ACCESS.2020.2995905.

[35]   M. T. Urbanski, "What is the comparative accuracy of IRT, linear regression, and XGBoost in predicting students' performance on different versions of an exam?," University of Twente, 2023.

BIOGRAPHIES OF AUTHORS

Tri Wahyuningsih is a doctoral student of computer science program at Satya Wacana Christian University. She has a strong interest in information systems management and decided to pursue her doctoral degree in Computer Science. She has interests and capabilities in data mining and text mining. Before starting her doctoral program, she completed her bachelor and master degrees in Informatics Engineering. She has working experience as a system analyst and showed excellent achievements during her work. Now, she is focusing on her doctoral studies and working on several research projects in the field of information systems management. She can be contacted at email: 982022001@student.uksw.edu.

Ade Iriani is an expert in the field of information systems and management with a strong educational background. She obtained her bachelor's degree from Padjadjaran University in 1987, and later pursued her Master's in Management at Satya Wacana Christian University in 2003. Her thirst for knowledge did not end there; in 2015, she earned her Doctorate in the same field. Currently serving as a tenured lecturer at Satya Wacana Christian University. She continues to inspire through her profound expertise in information systems and management. She can be contacted at email: ade.iriani@uksw.edu.

Hindriyanto Dwi Purnomo received his Bachelor of Engineering from Gajah Mada University Yogyakarta in physics engineering (intelligent control) in 2005, Master of Information Technology in informatics engineering from The University of Melbourne, Australia in 2009, and Doctor of Philosophy from Chung Yuan Christian University, Taiwan in industrial engineering (numerical optimisation) in 2013. His research interests include optimisation techniques, machine learning, soft computing, and fuzzy logic. Currently, he is active as a lecturer at the Faculty of Information Technology, Satya Wacana Christian University Salatiga. He can be contacted at email: hindriyanto.purnomo@uksw.edu.

Irwan Sembiring earned his Bachelor of Engineering degree in informatics engineering from Universitas Pembangunan Nasional "Veteran" Yogyakarta in 2001, Master of Computer Science in computer science from Gadjah Mada University Yogyakarta in 2004, and Doctorate in computer science from Gadjah Mada University Yogyakarta in 2015. His main research interest is computer network security. He has done more than 40 publications during his education and teaching career. His research interests include computer network security and computer network designing. Currently, he is active as a lecturer at the Faculty of Information Technology, Satya Wacana Christian University Salatiga. He can be contacted at email: irwan@uksw.edu.