Data Mining Approach for Predicting Student Performance

Janssen Yang
Computer Science Department
Bina Nusantara University
Jakarta, Indonesia
janssen.yang@binus.ac.id

Rudy Kurniawan Efendy
Computer Science Department
Bina Nusantara University
Jakarta, Indonesia
rudy.efendy@binus.ac.id

Budi Oetomo
Computer Science Department
Bina Nusantara University
Jakarta, Indonesia
budi.oetomo@binus.ac.id

Muhammad Fajar
Computer Science Department
Bina Nusantara University
Jakarta, Indonesia
muhammad.fajar@binus.ac.id

Puti Andam Suri
Computer Science Department
Bina Nusantara University
Jakarta, Indonesia
puti.suri@binus.ac.id

***Abstract*—This study focuses on the application of data mining technique to predict student graduation based on their academic performance. The study used three different algorithms, namely Deep Learning, Linear Regression, and Neural Network. The study's purpose is to improve the accuracy of the development of more accurate models for predicting outcomes or scores and to have a better understanding of data using a data mining approach. The dataset for this study was collected from the grades of computer science students for 3 semesters in 2018. The results of this study indicate that Deep Learning achieved the highest overall accuracy, which is 94.38%, outperforming both Neural Network with an accuracy of 93.48% and Linear Regression with an accuracy of 91.25%. The study provides valuable insights into the application of data mining techniques for predicting student graduation, aiding decision-makers in making informed decisions based on accurate predictions and derived insights.**

Keywords—component, data mining; linear regression; machine learning; clustering; neural network; support vector machine; deep learning; rentention rates

I. Introduction

According to Murtopo [1], Graduation is one factor that can influence the accreditation of a university. One of the assessments of the university is the university's efficiency in providing an ideal education to the students. In this regard, if students can graduate on time, it can help the accreditation assessment of the university. In addition to graduating on time, students will not have to pay more expenses to repeat semesters and can start looking for jobs and working faster after graduation.

Yet, students' graduation cannot be predicted from the beginning due to several factors that can affect students' graduation during their study period. Grades are one of the factors that can determine students' graduation when taking a course. With high scores, students can graduate on time and success in their studies. So, it is important to predict student graduation with high accuracy and calculation to assist in making decisions regarding student graduation.

Many students fail to graduate on time due to repeating semesters or repeating failed courses. Repeating occurs due to difficulties in the courses taken and an inability to keep up with them, resulting in unsatisfactory exam scores. Additionally, several non-academic factors can also affect students' academic performance, such as family problems or financial issues that cause them not to take their courses seriously, affecting their grades. Students must repeat courses that were declared failed because the results did not meet the passing grade requirements. This research aims to investigate a model that can predict student graduation based on their academic performance data throughout their study period. The model used will be able to predict whether a student will pass or fail based on their grades. From the prediction results, it is expected that students will work harder and improve their learning performance to reduce the risk of failing and repeating the semester.

Jixia Tu et al. (2019) used Support Vector Machines (SVM) approach such as random forest (RF), support vector machine (SVM), sine cosine algorithm (SCA), and chaotic local search to increase an adaptive guide vector gadget framework, known as RF-CSCA-SVM, for predicting university students' entrepreneurial aim in advance. As shown, the test accuracy can be achieved at 83.50%, and the sensitivity, specificity, and MCC of the proposed RFCSCA-SVM on the test set are shown at 91.25%, 72.22%, and 0.6602, respectively. The consequences of the confusion matrix display that maximum entrepreneurship college students may be appropriately expected as entrepreneurship. The reputation mistakes charge particularly happens whilst the scholars who are hired are misclassified as entrepreneurs [3].

The contribution given in this study is to improve the accuracy of the development of more accurate models for predicting outcomes or scores and to have a better understanding of data using a data mining approach. By analyzing this study, decision-makers will make better-informed decisions based on accurate predictions and insights derived from data.

This study will conduct a data mining approach to predict students' graduation using three different algorithms as Deep Learning, Linear Regression, and Neural Network models. The achieved results from this study are to compare the results of three models. Based on this research, a proposed solution is to create a model that can predict student graduation using their academic performance data, the prediction results can be used as a consideration or motivation to improve their learning performance during the study period.

This research contains five sections. Section 1 explains the research problem, the aims that want to achieve, previous research about students' academic performance prediction, the contribution given of the study, and the methodology.

Section 2 describes the recent related works about predicting student academic performance. Section 3 explains the data collection process and the algorithms used in the study. Section 4 shows the results and discussions, and section 5 concludes the study and future works.

II. Related Works

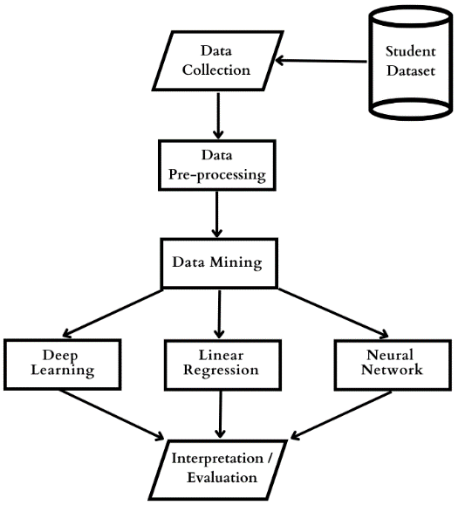
In recent years, several studies have been conducted to predict students' academic performance using machine-learning techniques. These studies have provided insights into the development of machine learning models for predicting academic performance.

Orynbassar et al. (2022) explored the use of ROC curve analysis for predicting students' passing grades in a course based on prerequisite grades in a private university in Kazakhstan. The study found that the ROC method outperformed the currently used practice of setting Prerequisite grade necessities in figuring out viable screw-ups in Calculus 2 [2]. Jixia Tu et al. (2019) used Support Vector Machines (SVM) approach such as random forest (RF), support vector machine (SVM), sine cosine algorithm (SCA), and chaotic local search to increase an adaptive guide vector gadget framework, known as RF-CSCA-SVM, for predicting university students' entrepreneurial aim in advance. As shown, the test accuracy can be achieved at 83.50%, and the sensitivity, specificity, and MCC of the proposed RFCSCA-SVM on the test set are shown at 91.25%, 72.22%, and 0.6602, respectively. The consequences of the confusion matrix display that maximum entrepreneurship college students may be appropriately expected as entrepreneurship. The reputation mistakes charge particularly happens whilst the scholars who are hired are misclassified as entrepreneurs [3]. Czibula et al. (2019) proposed a novel Relational Association Rule (RAR) mining classification model, S PRAR, for academic performance prediction using grades of students at CS undergraduate courses offered in the first, second, and third semesters at Babes¸-Bolyai University from Romania. The obtained experimental results highlighted that their classifier is better than, or comparable to, the supervised classifiers already applied in the EDM literature for students' performance prediction [4]. Muhammad Imran et al. (2019) used AI Supervised Learning techniques such as the C4.5/J48 Decision Tree algorithm, the Non-Nested Generalized Exemplars (NNge) algorithm, and Multilayer Perceptron (MLP) algorithm as models to predict student academic performance. C4.5/J48 algorithm achieved high accuracy among others at 95.78% [5].

Yacoob et al. (2020) used Decision tree, logistic regression, random forest, K-nearest neighbor, and neural network algorithm to predict student dropout of Computer Science undergraduate students after three years of enrolment in Universiti Teknologi MARA. The results revealed that the Logistic regression classifier performed best (with the highest overall classification accuracy) followed by k-NN, Random Forest, Neural Network, and Decision Tree. accuracy above 80% which means the error rate was low and predictions were reliable. The detection on the sensitivities of Logistic Regression, k-NN, Random Forest, Neural Network, and Decision Tree were 89.9%, 87.8%, 85.2%,82.4%, and 80.9% respectively [6]. Sriudaya Damuluri et al. (2020) used Support Vector Machine (SVM) to predict students' final grades in the course. The result shows that the model successfully identifies students who are at risk of failing the class with the accuracy obtained at 70,21%. Diego Buenaño-Fernández et al. (2019) used the Machine Learning and Decision Trees algorithm to predict students' performance. In this study, the result obtained for predicting the final grades with accuracy was 96.5% [8]. Lau et al. (2019) used Artificial Neural Network (ANN) approach with both conventional statistical analysis and neural network modeling/prediction of students' performance. the ANN has achieved a good prediction accuracy of 84.8% and a good AUC value of 0.86 [9]. Altabrawee et al. (2019) used a Machine Learning approach to predict students' performance using four machine learning techniques the Artificial Neural Network (ANN), Naïve Bayes, Decision Tree, and Logistic Regression. The results of four models are compared with the ANN model obtained a ROC index that equals 0.807 with an accuracy of 77.04 is higher than the others [10].

III. Methodology

The study aims to predict student graduation based on their academic performance using data mining techniques in RapidMiner and compare the performance of three different algorithms or models such as Deep Learning, Linear Regression, and Neural Network. The students' grades dataset was used for the analysis.



3.1 Data Collection

The dataset for this study was collected from the grades of computer science students for 3 semesters in 2018. The dataset has several features such as student ID, student streaming/minor, enrichment 1, enrichment 2, course ID, course name, score, STRM, IPK, thesis topic, thesis title, and semester.

3.2 Data Preprocessing

The data will be extracted and analyzed to identify the relevant features that impact student graduation. The key features that will be analyzed are the student's IPK and the student's score. In addition to these features, other factors that will be considered in the data analysis include the student ID, course name, and semester. These factors will be used to identify any trends or patterns in the data that may impact a student's grades.

First, courses that are more essential and have more impact on student graduation were chosen from the dataset. This step involved carefully analyzing the correlation between each course and the graduation outcome. Next, attributes that did not contribute to the prediction of student graduation were removed. This step involved identifying sections that had minimal influence on the graduation outcome. By eliminating these irrelevant sections, the dataset was further refined, focusing only on the most influential factors. Then, missing values were erased to refine the dataset. Certain attributes' names and data types have been modified and aligned to match the topic. Finally, short-term semester grades were removed from the dataset. By excluding short-term grades, the model focused solely on original grades, which provided a more accurate representation of a student's capabilities and likelihood of graduating.

3.3 Data Mining

This study uses Data Mining tool's applied techniques to predict student's graduation based on three different algorithms. The three different algorithms that were applied to the dataset are Deep Learning, Linear Regression, and Neural Network. The models were trained using a Data Mining tool, which offers a comprehensive set of data mining and machine learning algorithms. In this study, three different algorithms are using 10 folds to avoid the model can cause overfitting and keep the model for generalizing well for new data.

3.3.1 Deep Learning

Deep learning is a machine learning approach based on multi-layer feed-forward artificial neural networks. It utilizes advanced features like adaptive learning rate, momentum training, dropout, and regularization (L1 or L2) to model and solve complex problems. Deep learning is typically trained using stochastic gradient descent with backpropagation. It can be executed using various frameworks or libraries, such as H2O 3.30.0.1. Deep learning drives many AI applications and services that automate tasks without human intervention. It is used in everyday products like digital assistants and voice-enabled remotes, as well as emerging technologies like self-driving cars.

3.3.2 Linear Regression

Linear Regression is a statistical technique used to predict numerical values by analyzing the relationship between a dependent variable and one or more independent variables. Linear Regression uses the Akaike criterion for model selection. Linear Regression fits a linear equation to observed data and is employed when making predictions about continuous outcomes, such as estimating the salary based on work experience or student graduation based on student's grades.

3.3.3 Neural Network

Neural Network is an algorithm that is based on the structure and functionality of biological neural network. It comprises interconnected artificial neurons and processes information using a connectionist approach to computation. This Neural Network algorithm that will be used is based on feed-forward neural network trained by a back propagation algorithm (multi-layer perceptron). Neural networks are adaptive systems that can modify their structure based on internal or external information during the learning phase. These networks are commonly used to model complex relationships between inputs and outputs or discover patterns in data. Neural Network is used for a wide range of tasks and applications across various fields, such as prediction and forecasting, pattern recognition, regression analysis, etc.

The performance of the models was evaluated by different metrics, Absolute Error (AE), Root Mean Squared Error (RMSE), and Coefficient of Determination (R-Squared). The AE and RMSE were used to measure the accuracy of the predictions, while R-Squared was used to measure the goodness of fit. The study's results were presented in tables, showing the performance of the three models across different metrics.

3.4 Interpretation/Evaluation

The Confusion matrix of Deep Learning, Linear Regression, and Neural Network is presented comparatively to show which of the three algorithms that is applied gives the best results. The study revealed that Deep Learning surpassed that of both Linear Regression and Neural Network. Although the performance of Linear Regression and Neural Network was commendable, they did not match the accuracy achieved by Deep Learning. The study provided valuable insights into the application of data mining techniques for predicting student graduation.

IV. Result and Discussion

The dataset that has been preprocessed will be used to be applied to three different algorithms, namely Deep Learning, Linear Regression, and Neural Network. In terms of prediction, the performances of three algorithms are presented as confusion matrices in Table 1. for the Deep Learning algorithm, Table 2. for the Linear Regression algorithm, and Table 3. for the Neural Network model. The dataset and the model of the algorithm are presented in github [11].

In the Deep Learning model, the optimal configuration for the model includes the following settings. The activation function used is the rectifier. The training process is performed over 25 epochs, representing the number of times the model iterates through the entire dataset. Various parameters are employed to enhance the learning process, where Epsilon is set to 1.0E-8, determining the minimum threshold for the update in the weights during optimization and Rho, with a value of 0.99, represents the coefficient used in the adaptive learning rate. Regularization techniques are also incorporated, where L1 regularization with a coefficient of 1.0E-5 encourages sparsity in the model's weights, while L2 regularization is not utilized and is set to 0.0. The maximum weight limit is restricted to 10.0, preventing excessively large weights that could negatively impact the model's performance. To handle missing values in the dataset, a mean imputation approach is employed, where missing values are replaced by the mean of the available values. Finally, the maximum runtime is set to 0 seconds, indicating no specific time limit for the training process. These settings contribute to the accuracy and stability of the Deep Learning model, enabling it to effectively learn complex patterns and make accurate predictions.

For the Linear Regression model, the following settings are employed. The feature selection technique used is M5 Prime, which helps identify the most relevant features for the regression analysis. The minimum tolerance is set to 0.05, indicating the acceptable level of error in the model's predictions. Additionally, a ridge parameter of 1.0E-8 is utilized to apply ridge regression, which helps mitigate the issue of multiple regression in the dataset. These settings contribute to the accuracy and stability of the Linear Regression model's predictions.

In the Neural Network model, the following settings are applied. The training cycle is set to 200, representing the number of iterations performed during the training phase. A learning rate of 0.01 is chosen, determining the step size at which the model's weights are updated during backpropagation. A momentum value of 0.9 is utilized to introduce an element of inertia in weight updates, helping to accelerate convergence and overcome local minima. Furthermore, an error epsilon of 1.0E-4 is set as the threshold for considering the network's convergence. These settings collectively contribute to the effective training and performance of the Neural Network model in capturing complex patterns and making accurate predictions.

The dataset is applied to three different algorithms to perform predictive analysis, and the results demonstrate their respective performance in terms of accuracy and other evaluation metrics, Deep Learning achieved the highest overall accuracy of 93.39%, followed by Neural Network with an accuracy of 92.92%, and Linear Regression with an accuracy of 91.25%. Furthermore, additional evaluation metrics were considered to assess the performance of each algorithm. When measuring the average Absolute Error (AE), Deep Learning achieved an AE of 8.8%, while Linear Regression and Neural Network exhibited AEs of 8.7% and 9% respectively. The Root Mean Squared Error (RMSE) provides another measure of prediction accuracy, where Deep Learning demonstrated an RMSE of 21.7%, Linear Regression exhibited an RMSE of 29.4%, and Neural Network achieved an RMSE of 21.6%. Additionally, the R-squared statistic, which represents the proportion of variance explained by the model, showed that Deep Learning obtained an R-squared of 65.3%, while Linear Regression and Neural Network achieved R-squared values of 45.9% and 61% respectively are presented in Table 4. Comparatively.

Although the Neural Network achieved the lowest average Absolute Error (AE) among the three algorithms, Deep Learning emerged as the top-performing approach in terms of overall results. Deep Learning demonstrated the highest accuracy, the lowest Root Mean Squared Error (RMSE), and the highest R-squared value. These outcomes signify the superior performance of Deep Learning, making it the most effective algorithm for the given dataset.

Table 1. Confusion Matrix for Deep Learning algorithm

	Graduate on time	Not graduate on time
Graduate on time	895	37
Not graduate on time	26	126

Table 2. Confusion Matrix for Linear Regression algorithm

	Graduate on time	Not graduate on time
Graduate on time	905	82
Not graduate on time	16	117

Table 3. Confusion Matrix for Artificial Neural Network algorithm

	Graduate on time	Not graduate on time
Graduate on time	891	43
Not graduate on time	30	156

Table 4. Comparison of models' performance

	Deep Learning	Linear Regression	Neural Network
Accuracy	94.38%	91.25%	93.48%
AE	8.8%	8.7%	9%
RMSE	21.7%	29.4%	21.6%
R-squared	65.3%	45.9%	61%

V. Conclusion and Future Works

Predicting student graduation is essential for ensuring adequate quality in the educational system. By accurately predicting student graduation, educational institutions can implement targeted interventions and support systems to enhance student success and retention rates. The study aims to predict student graduation based on their academic performance using data mining techniques. The dataset used for analysis was collected from the grades of computer science students for 3 semesters in 2018 and includes features such as student ID, course name, score, IPK and semester. The data preprocessing stage involves extracting and analyzing relevant features that impact student graduation, such as IPK and scores. Irrelevant attributes are removed, missing values are erased, and short-term semester grades are excluded from the dataset to focus solely on original grades.

In data mining phase, three different algorithms, namely Deep Learning, Linear Regression, and Neural Network are applied to the dataset. The results show that in terms of accuracy, Deep Learning achieved the highest overall accuracy, which is 94.38%, outperforming both Neural Network with an accuracy of 92.92% and Linear Regression with an accuracy of 91.25%. The assessment of additional evaluation metrics revealed that Deep Learning achieved an average Absolute Error (AE) of 8.8%, while Linear Regression and Neural Network exhibited AEs of 8.7% and 9%, respectively. The Root Mean Squared Error (RMSE) analysis indicated that Deep Learning achieved an RMSE of 21.7%, whereas Linear Regression exhibited a higher RMSE of 29.4%, and Neural Network achieved an RMSE of 21.6%. Moreover, the R-squared statistic, demonstrated that Deep Learning attained an R-squared value of 65.3%, whereas Linear Regression and Neural Network obtained R-squared values of 45.9% and 61%, respectively. While Neural Network has the lowest AE, Deep Learning still has the best overall result with the highest accuracy, lowest RMSE and the highest R-Squared, making it the most effective algorithm for the given dataset. For future studies, it would be beneficial to incorporate non-academic data to improve the accuracy and comprehensiveness of the prediction models rather than only using academic data. Expanding the analysis to include non-academic data can provide valuable insights into additional factors that may influence student graduation such as economic background, extracurricular activities, part-time work experience, and participation in community service or leadership roles.

References

[1] A. A. Murtopo, "Prediksi Kelulusan Tepat Waktu Mahasiswa STMIK YMI Tegal Menggunakan Algoritma Naïve Bayes", Comp. Scie. Resea. and. Its. Dev. J., vol. 7, no. 3, pp. 145-154, October 2015.

[2] A. Orynbassar, Y. Sapazhanov, S. Kadyrov, I. Lyublinskaya, "Application of ROC Curve Analysis for Predicting Students' Passing Grade in a Course Based on Prerequisite Grades", Mathematics, vol. 10, no. 12, pp. 2084, 2022, https://doi.org/10.3390/math10122084.

[3] J. Tu, A. Lin, H. Chen, Y. Li, "Predict the Entrepreneurial Intention of Fresh Graduate Students Based on an Adaptive Support Vector Machine Framework", Math. Prob. in. Engi., vol. 2019, pp. 1-16, 2019, https://doi.org/10.1155/2019/2039872.

[4] G. Czibula, A. Mihai, L. M. Crivei, "S PRAR: A novel relational association rule mining classification model applied for academic performance prediction", Pro. Comp. Scie., pp. 20-29, 2019, https://doi.org/10.1016/j.procs.2019.09.156.

[5] M. Imran, S. Latif, D. Mehmood, and M. S. Shah, "Student Academic Performance Prediction using Supervised Learning Techniques", Int. J. Emerg. Technol. Learn., vol. 14, no. 14, pp. 92–104, Jul. 2019, https://doi.org/10.3991/ijet.v14i14.10310.

[6] W. F. W. Yacoob et al, "Predicting Student Drop-Out in Higher Institution Using Data Mining Techniques", J. Phys.: Conf. Ser., 1496, 012005, 2020, doi: 10.1088/1742-6596/1496/1/012005.

[7] S. Damuluri, K. Islam, P. Ahmadi, N. S. Qureshi, "Analyzing Navigational Data and Predicting Student Grades Using Support Vector Machine", Emerg. Scie. J., vol. 4, no. 4, pp. 243-252, August 2020, doi: https://doi.org/10.28991/esj-2020-01227.

[8] D. B. Fernandez, D. Gill, S. L. Mora, "Computer Engineering Students: A Case Study", sustainability, vol. 11, no. 10, pp. 2833, 2019, https://doi.org/10.3390/su11102833.

[9] E. T. Lau, L. Sun, and Q. Yang, "Modelling, prediction and classifcation of student academic performance using artifcial neural networks", SN Appl. Sci., vol. 1, no. 982, p. 10, 2019. https://doi.org/10.1007/s42452-019-0884-7.

[10] H. Altabrawee, O. A. J. ali, S. q. Ajmi, "Predicting Students' Performance Using Machine Learning Techniques", J. of. Uni. of. Babylon., vol. 27, no. 1, pp. 194-205, 2019.

[11] https://github.com/RudyKurniawanEfendy/Data-Mining-Approach-for-Predicting-Student-Performance