The Impact of Data Splitting on Graph-Based Dropout Prediction Using Subgraph Matching and Graph Edit Distance

Meilia Nur Indah Susanti
Computer Science Department, Universitas Bina Nusantara, Jakarta, Indonesia | Faculty of Energy Telematics, Institut Teknologi PLN, West Jakarta, Indonesia
meilia.susanti@binus.ac.id (corresponding author)

Yaya Heryadi
Computer Science Department, Universitas Bina Nusantara, Jakarta, Indonesia
yayaheryadi@binus.edu

Yusep Rosmansyah
School of Electrical Engineering and Informatics, Institut Teknologi Bandung, Bandung, Indonesia
yusep@itb.ac.id

Widodo Budiharto
School of Computer Science, Universitas Bina Nusantara, Jakarta, Indonesia
wbudiharto@binus.edu

Received: 24 December 2025 | Revised: 20 January 2026 and 27 January 2026 | Accepted: 29 January 2026

 DOI: https://doi.org/10.48084/etasr.17152

ABSTRACT

Student dropout remains a persistent issue in higher education, affecting institutional effectiveness and student success rates. This paper proposes a graph-based predictive model that employs subgraph matching and Graph Edit Distance (GED) to identify students at high risk of dropout. By modeling students and courses as an undirected bipartite graph, the system detects structural similarities between student profiles. The proposed model was evaluated using a dataset of 282 students from a private university in Indonesia under three data-splitting scenarios: 70/30, 79/21, and 89/11. Evaluation metrics include precision, recall, F1-score, and accuracy. The model achieved its best performance at the 89/11 split, with an accuracy of 91%, precision of 1.00, recall of 0.88, and an F1-score of 0.94. Results suggest that increasing the proportion of training data enhances generalization and prediction accuracy. GED demonstrated effectiveness in capturing subtle structural distinctions among student–course relationships, enabling early dropout risk identification. The primary contribution of this study is the development of a graph-analytic framework for dropout prediction, offering an alternative to traditional models such as logistic regression and decision trees that lack relational awareness. Future work will incorporate behavioral and socio-economic attributes to further improve prediction outcomes.

Keywords-data split; dropout prediction; graph-based modeling; Graph Edit Distance (GED); subgraph matching

I. INTRODUCTION

Student dropout remains one of the most significant challenges in higher education worldwide, including in Indonesia. Despite various initiatives aimed at reducing dropout rates, higher education institutions, particularly private universities, still face high dropout rates [1]. This phenomenon is influenced by various factors, such as students' economic conditions, academic preparedness, motivation to learn, as well as the quality of academic and non-academic services provided by higher education institutions [2]. The high dropout rate not only impacts students individually but also harms institutions in terms of reputation, financial sustainability, and academic performance. Therefore, a comprehensive, data-driven strategy is needed to identify dropout risk factors early and design effective interventions.

Establishing a precise conceptual definition of "college dropout" is a complex task that goes beyond the theoretical realm and is embodied in policies, measures, and studies

developed by universities and countries around the world. It is also a difficult phenomenon to measure because it requires knowledge of what we want to measure, precise institutional data, and the necessary time frame. Thus, arriving at a complete definition of "college dropout" is a complex process because of the many variables that influence this decision; therefore, universities introduce a variety of curricular and institutional policies to reduce dropout [3]. This alarming statistic highlights the need for effective intervention strategies to identify and address the factors contributing to dropout risk early in a student's academic journey [4].

Many educational institutions rely on traditional methods to monitor and predict student dropout based on academic performance, socio-economic factors, and institutional data to address this issue. However, these manual methods are becoming increasingly inefficient due to the rising student-to-staff ratio [5, 6]. The advent of machine learning technologies has offered a promising avenue for analyzing large volumes of educational data, allowing for more accurate predictions and timely interventions [7-9]. Yet, the integration of graph analytic methods for student dropout prediction has received limited attention despite their potential to model complex relationships between students and their courses [10, 11].

Graph analytics offers a flexible and robust framework for capturing the intricate relationships between students and courses, allowing for a more comprehensive analysis of dropout risk factors [12]. Graph-based models can more effectively explore and predict dropout risks than traditional machine learning approaches by representing students and courses as vertices in a graph and using their relationships as edges. Subgraph matching, a key technique in graph analytics, enables the identification of patterns and similarities within the relationships between student features and course data, providing valuable insights into the underlying causes of dropout [13].

This study proposes using subgraph matching combined with Graph Edit Distance (GED) to model student dropout risk and improve the accuracy of graduation predictions. By applying this approach to student–course data, we aim to uncover hidden patterns and relationships contributing to student success or failure. This research demonstrates the effectiveness of graph-based models in predicting dropout risks and enhancing the precision of educational interventions. To operationalize this graph-based analysis, it is crucial to employ a robust similarity metric that can quantify structural correspondences between student graphs.

GED serves as an important metric to enable subgraph matching operations. Specifically, GED measures the similarity between two graph structures by calculating the minimum number of editing operations—such as node insertions, deletions, and replacements—required to transform one graph into the other. By integrating GED into the subgraph matching process [14], the model can detect partial structural alignments between students' academic paths and established patterns of dropout or graduation, enabling more accurate identification of at-risk individuals. This hybrid approach provides a more flexible and robust matching framework, capable of identifying at-risk students even when an exact structural match is lacking, thereby enhancing the model's ability to generalize across diverse student profiles.

Based on this background, the main objective of this study is to develop a graph-based student dropout prediction model by integrating subgraph matching and GED to capture complex student–course relationships more effectively. The key contribution of this research lies in demonstrating how the proposed hybrid approach improves dropout risk identification accuracy and provides actionable insights to support early, data-driven intervention strategies in higher education institutions.

II. MATERIALS AND METHODS

The object of study in this research includes students who dropped out of college, especially private colleges in Indonesia. Table I shows the distribution of student dropout rates in private colleges in Indonesia from 2020 to 2023.

TABLE I. THE SPREAD OF DROPOUT RATES IN INDONESIA

Year	Dropout percentage (%)	Number of students
2020	79.50	478,826
2021	76.58	367,908
2022	66.88	250,891
2023	60.56	213,465

Based on Table I, the dropout rate of private universities in Indonesia shows that in 2020, it was 79.50% or 478,826 students; in 2021 it was 76.58% or 367,908 students; in 2022, it was 66.88% or 250,891 students, and in 2023 it was 60.56% with a total of 213,465 students. Although dropout rates are decreasing, they are still high, with more than 200,000 students dropping out of school by 2023. The dataset used in this study consists of data from 282 students at private universities in West Jakarta, Indonesia. The features include 77 courses and the corresponding course grade weights for each student, which were used to predict graduation status.

The dataset includes data on students from private universities in Indonesia from 2020 to 2023, representing the COVID-19 pandemic phase and the early post-pandemic transition period, with coverage of the annual dropout rate distribution. These data were used as the primary data (training data) in the development and analysis of a predictive model to capture patterns of dropout behavior that occurred during the crisis and early recovery period. The study specifically focuses on analyzing the dropout phenomenon in the post-COVID-19 period as a critical transition phase for private universities in Indonesia, given the structurally impactful effects of the pandemic on students' economic conditions, learning patterns, and institutional stability.

A. Research Design

This section provides a detailed overview of the proposed methodology. The main objective is to predict students who are detected to drop out early and examine how existing datasets affect the ability to predict students who are at risk of dropout. This method begins with Research Clarification, which involves identifying evidence that supports the assumptions

used to formulate the research objectives. The next stage is Descriptive Study I, which analyzes empirical data to gain an understanding of the research and determine the factors that influence the success of the study. The next step is Predictive Study, in which there are two main processes, namely Data Preparation and Modeling. The Data Preparation process includes collecting student data in the form of grade weights, graduation status, and other academic information. The last stage is Descriptive Study II, which consists of two processes: Evaluation and Deployment. The Evaluation stage evaluates the accuracy of the prediction model using metrics such as GED. The Deployment process includes implementing the prediction model into a practical system. The prediction results are visualized through reports and graphs to support higher education decision-making.

Figure 1 outlines the stages of the proposed research. This study adopts the Design Research Methodology (DRM) [14, 15] to develop a student graduation prediction model based on artificial intelligence and machine learning.

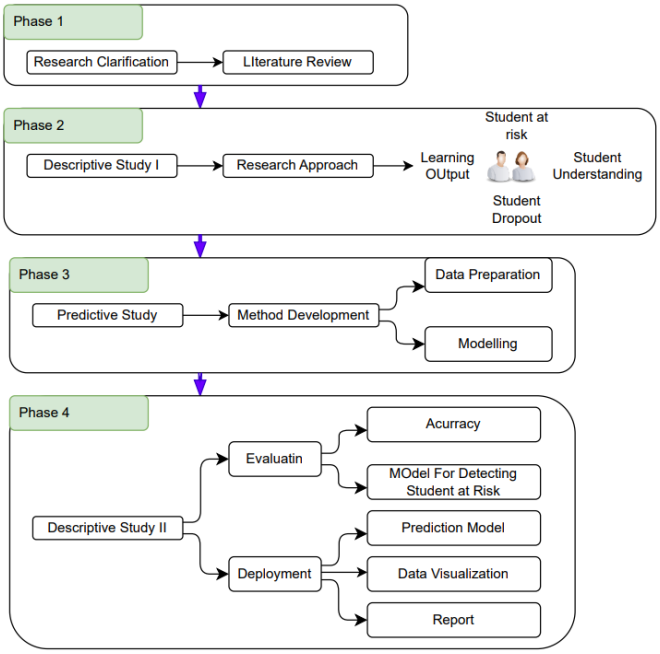

Fig. 1. Research design of the proposed student dropout prediction methodology.

In the Research Clarification phase, a literature review is conducted to identify factors influencing graduation, such as academic risk, material comprehension, and dropout patterns, while highlighting the Graph Neural Network (GNN) with the subgraph matching technique as an innovative approach.

The Descriptive Study I phase aims to map student data characteristics, identify relationships between academic performance and graduation rates, and establish a strong research framework. In the Predictive Study phase, the data undergo selection, cleaning, and construction before being applied to the GNN model to analyze graduation patterns based on graph structures. Subgraph matching enables the model to identify students at risk of not graduating by exploring relationships between academic entities.

The Descriptive Study II phase evaluates the model's performance using the GED metric. During this stage, the concept of data splitting is applied, where a single dataset is divided into two subsets: a training set for model training and a testing set for model validation. Several studies recommend using an 80/20 (training/testing) split for applications such as used car price prediction [16] and heart disease prediction [17], whereas others have employed a 90/10 split for predicting material properties [18]. Increasing the proportion of training data from 80% to 90% can lead to noticeable improvements in model performance during testing. This demonstrates that the size of the training dataset can significantly impact the overall performance of the model.

The primary objective of this study is to evaluate model performance across different data split ratios using educational datasets, specifically student data, and to identify the most effective split ratio for achieving optimal predictive accuracy. This framework provides the foundation for integrating the predictive model into a practical decision-support system within the educational context.

Accordingly, this study not only assesses the influence of varying data split ratios on model performance but also promotes the implementation of the model in a predictive system equipped with data visualizations. This approach is expected to enhance decision-making processes in higher education institutions, support improved student graduation rates, and contribute meaningfully to dropout prevention strategies through more targeted and effective interventions.

B. Subgraph Matching

Subgraph matching is a generalization of the graph isomorphism problem in graph theory. Using only structural information, the graph-matching algorithm identifies the mapping between a set of vertices of two similar graphs. Isomorphism describes the connectivity between two graphs, where one graph (called a subgraph) can be found within the other, which is referred to as the target graph [19, 20]. Two graphs are considered isomorphic if $G_1 = (V_1, E_1)$ and $G_2 = (V_2, E_2)$ have the same structure, although their vertices may be labeled differently [21]. Based on the presence of labels on nodes and edges, graphs can be categorized as labeled or unlabeled [22, 23]. In the context of simple graphs, two graphs are considered isomorphic if there exists a one-to-one correspondence between their nodes and edges, preserving both structural and topological relationships [22].

C. Graph Edit Distance

GED is a graph-matching method designed to identify the optimal set of transformations required to convert graph g_1 into a graph g_2 through edit operations such as insertion, deletion, and substitution of vertices and their corresponding edges in graph g_1 [24, 25]. Let $g_1 = (V_1, E_1, \mu_1, \zeta_1)$ and $g_2 = (V_2, E_2, \mu_2, \zeta_2)$ [26]. The GED between g_1 and g_2 is defined as:

$$\mathrm{GED}(g_1, g_2) = \min_{e_1, e_2, \ldots, e_k \in \gamma(g_1, g_2)} \sum_{i=1}^{k} c(e_i) \qquad (1)$$

where $c(e_i)$ represents the cost function that measures the extent of modification for an edit operation e_i and $\gamma(g_1, g_2)$ denotes the set of edit paths that transform g_1 into g_2. In general, edit operations include substitution, deletion, and insertion of both vertices and edges. The substitution of two vertices u and v is represented as the operation $(u \rightarrow v)$, the deletion of a vertex u is denoted as $(u \rightarrow \varepsilon)$, and the insertion of a vertex v is represented as $(\varepsilon \rightarrow v)$.

D. Confusion Matrix

A confusion matrix is a tabular representation used to evaluate the performance of a machine learning model. It presents the number of correct and incorrect predictions made by the model on a test dataset. The confusion matrix consists of four primary components:

- True Positives (TP): the number of correctly predicted positive instances.
- True Negatives (TN): the number of correctly predicted negative instances.
- False Positives (FP): incorrect predictions where the model classifies an instance as positive when it is negative (false alarm).
- False Negatives (FN): incorrect predictions where the model classifies an instance as negative when it is positive (missed detection) [24].

The values in a confusion matrix are used to calculate key performance metrics that assess the model's classification accuracy [25]. These metrics include accuracy, precision, recall, and F1-score, which are crucial for evaluating machine learning models in different application contexts. The formulas for these metrics are as follows:

$$\text{Accuracy} = \frac{\text{TP+TN}}{\text{TP+TN+FP+FN}} \quad (2)$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}} \quad (3)$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP+FN}} \quad (4)$$

$$\text{F1} - \text{score} = 2 * \frac{\text{Precision} * \text{Recall}}{\text{Precision+Recall}} \quad (5)$$

III. RESULTS AND DISCUSSION

In this study, a graph was constructed to illustrate the relationships between 282 students and 77 courses. All data were represented as a graph in which students and courses are depicted as nodes, and each connection between a student and a course is represented as an edge. Figure 2 presents an example of such a graph, showing the relationships between 23 student nodes and 11 course nodes. In this graph, pink nodes represent students who successfully graduated, whereas blue nodes indicate students who dropped out. The remaining nodes represent courses (yellow nodes). Each course is positioned at the center of the relationships within the graph, based on the assumption that all students enrolled in 11 courses.

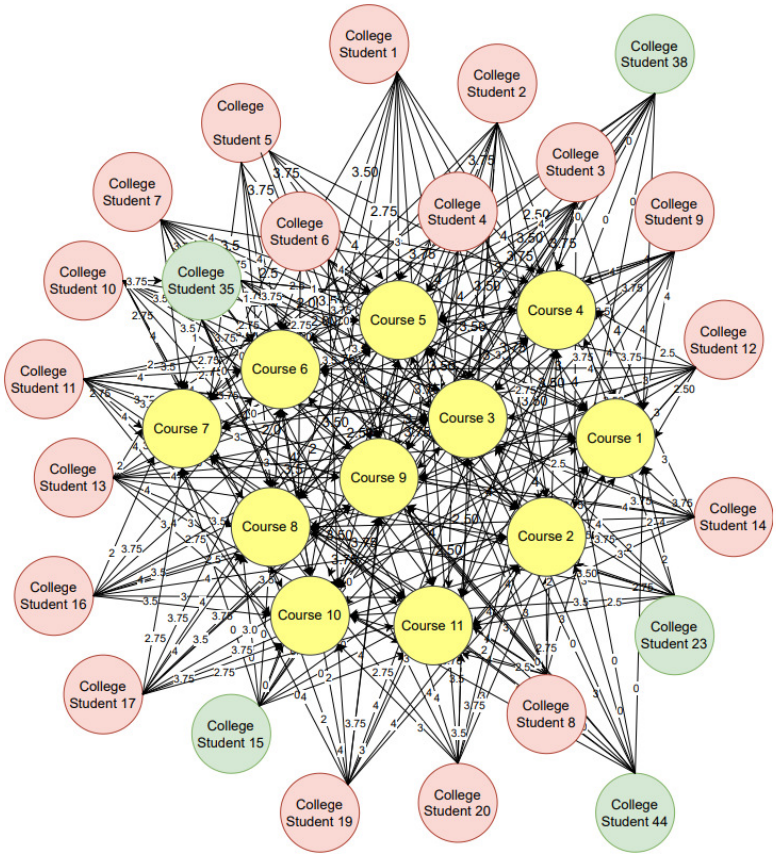

Fig. 2. Undirected bipartite graph representing students and courses.

In the graph, students who graduated (pink nodes) exhibited higher and more stable grade patterns across various courses than students who dropped out (blue nodes). Students who dropped out, as indicated by the nodes College Student 15, College Student 23, College Student 35, College Student 38, and College Student 44, tended to have low grades in most courses. In contrast, students who completed their studies demonstrated better grade patterns in the same courses. This relationship provides valuable insights for identifying academic risk factors that may contribute to dropout.

The connectivity between student nodes and course nodes, represented by edges, reflects student performance in each course through weighted values. Table II illustrates the edge weights for selected students and courses in the graph.

TABLE II. EXAMPLE OF STUDENT–COURSE EDGE WEIGHTS

Student	Nodes	Edge weights
College Student 1	Course 1, Course 5	2.75, 3.50
College Student 2	Course 4, Course 1	3.25, 2.75
College Student 3	Course 1, Course 6	3.50, 3.75
College Student 4	Course 3, Course 7	2.90, 3.30
College Student 5	Course 5, Course 8	2.70, 3.60
College Student 6	Course 6, Course 9	3.40, 2.85
College Student 7	Course 10, Course 11	3.10, 2.90
College Student 8	Course 2, Course 9	3.20, 3.00
College Student 9	Course 7, Course 3	3.40, 2.80
College Student 10	Course 4, Course 8	3.10, 2.70
College Student 11	Course 5, Course 11	3.00, 3.25
College Student 12	Course 1, Course 2	3.50, 2.60
College Student 13	Course 6, Course 7	2.75, 3.20
College Student 14	Course 3, Course 9	3.40, 2.80
College Student 15	Course 8, Course 10	3.00, 2.90
College Student 16	Course 11, Course 2	2.60, 3.30
College Student 17	Course 4, Course 10	3.10, 2.70
College Student 19	Course 5, Course 9	3.40, 2.80
College Student 20	Course 7, Course 3	3.00, 3.50
College Student 23	Course 2, Course 6	3.25, 3.75
College Student 35	Course 5, Course 10	2.90, 3.20
College Student 36	Course 6, Course 8	3.40, 2.80
College Student 38	Course 5, Course 10	2.75, 3.10
College Student 44	Course 3, Course 6	3.20, 2.90

Table II provides an example of the analysis results, showing the relationships between students and courses and the weights of the connections between nodes. This structure allows the identification of significant connectivity patterns, such as groups of students with similar academic behaviors or students with weak connectivity to courses.

In this study, subgraph matching was applied to identify connectivity patterns by comparing individual student subgraphs with subgraphs of students previously known to be at academic risk. The GED metric was used to enable early detection of students vulnerable to dropping out.

A. Results

The data analysis process involved a total of 282 student records by conducting nine experiments using different data splits. Each experiment was repeated three times, and the results of the data-split tests are presented below for selected split ratios.

1) Experimental Results Using 70/30 Train–Test Split

The first experiment employed a 70/30 data split, and the resulting classification accuracy was evaluated. In this setting, 197 records were used for training, whereas 85 records were allocated for testing. Table III presents a comparison between the actual outcomes and the predicted results obtained from this evaluation.

TABLE III. SAMPLE TESTING WITH 70/30 DATA SPLIT

Index	Test graph	Training graph	GED	Actual	Predicted
1	College Student 197	College Student 141	24.25	Graduated	Graduated
2	College Student 198	College Student 031	25.0	Graduated	Graduated
3	College Student 199	College Student 120	15.0	Graduated	Graduated
4	College Student 200	College Student 130	19.25	Graduated	Graduated
5	College Student 201	College Student 128	23.5	Graduated	Dropout
6	College Student 202	College Student 036	10.25	Graduated	Graduated
7	College Student 203	College Student 015	0.0	Dropout	Dropout
8	College Student 204	College Student 073	20.0	Dropout	Graduated

GED is used to measure the level of structural difference between two graphs by calculating the minimum cost required to transform one graph into another. In this context, each graph represents a student profile built from relevant academic attributes and relationships. The transformation is carried out through a series of edit operations, such as substitution, addition, or deletion of nodes and edges, each of which has an associated cost. The GED value is obtained by summing the costs of all edit operations that result in the minimum distance between the two graphs. The GED value is calculated using (1) based on the data in Table III (index 1), as follows:

$$\begin{aligned}\mathrm{GED}(g_{197}, g_{141}) &= c(e_1) + c(e_2) + c(e_3) \\ &= 10.00 + 8.25 + 6.00 \\ &= 24.25\end{aligned}$$

The values of $c(e_1)$, $c(e_2)$, and $c(e_3)$ are obtained from the cost evaluation process for each graph edit operation required to transform the College Student 197 graph into the College Student 141 graph. The value of $c(e_1) = 10.00$ represents the cost of substituting node attributes, which is calculated from the differences in key academic attributes, such as the variation in average grades or course completion status, in accordance with the cost function defined in the model. The value of $c(e_2) = 8.25$ results from the operation of removing a single edge, reflecting the loss of certain academic relationships, such as connections to courses or academic activities that are no longer the same between the two graphs. The classification report obtained from this test is shown in Table IV.

TABLE IV. CLASSIFICATION REPORT FOR 70/30 TRAIN–TEST SPLIT

Class	Precision	Recall	F1-score	Support
Dropout	0.56	0.93	0.70	15
Graduated	0.98	0.85	0.91	17
Accuracy	–	–	0.86	32
Macro avg	0.77	0.89	0.80	32
Weighted avg	0.91	0.86	0.87	32

Table IV shows that the model achieves an overall accuracy of 86%, with excellent performance on the Graduated class (precision = 0.98, recall = 0.85, F1-score = 0.91). However, although the recall for the Dropout class is high (0.93), its precision is low (0.56), resulting in an F1-score of only 0.70. The unequal distribution between the majority and minority classes causes the model to tend to be biased toward the Graduated class, as reflected by the weighted average F1-score of 0.87 and the macro average F1-score of 0.80.

2) Experimental Results Using 89/11 Train–Test Split

This experiment employed a data split ratio of 89/11. In this setting, 251 records were used for training, whereas 31 records were allocated for testing. Table V shows a comparison between the actual outcomes and the predicted results obtained from this evaluation.

TABLE V. SAMPLE TESTING WITH 89/11 DATA SPLIT

Index	Test graph	Training graph	GED	Actual	Predicted
1	College Student 251	College Student 089	19.25	Graduated	Graduated
2	College Student 253	College Student 049	19.25	Graduated	Graduated
3	College Student 255	College Student 029	17.25	Graduated	Graduated
4	College Student 259	College Student 128	23.25	Graduated	Dropout
5	College Student 261	College Student 247	18.0	Graduated	Graduated
6	College Student 267	College Student 151	23.75	Graduated	Dropout
7	College Student 275	College Student 208	15.25	Graduated	Graduated
8	College Student 278	College Student 232	7.25	Dropout	Dropout

From these results, a classification report was obtained, as shown in Table VI.

TABLE VI. CLASSIFICATION REPORT FOR 89/11 TRAIN–TEST SPLIT

Class	Precision	Recall	F1-score	Support
Dropout	0.70	1.00	0.82	7
Graduated	1.00	0.88	0.94	25
Accuracy	–	–	0.91	32
Macro avg	0.85	0.94	0.88	32
Weighted avg	0.93	0.91	0.91	32

The classification report in Table VI indicates that the model achieved an overall accuracy of 91% in classifying student status. For the Dropout class, the model attained a precision of 0.70, a perfect recall of 1.00, and an F1-score of 0.82, indicating that all actual dropout cases were correctly identified, although the precision remained relatively low due to the presence of false positive predictions. In contrast, the model exhibited optimal performance for the Graduated class with a precision of 1.00, recall of 0.88, and F1-score of 0.94. Overall, the macro average F1-score of 0.88 and weighted average F1-score of 0.91 reflect robust classification performance.

The evaluation results for the different data split ratios show that increasing the proportion of training data has a positive effect on both the accuracy and balance of model performance. At a ratio of 70/30, the model obtained an accuracy of 86% with a dropout class F1-score of 0.70. Although the 79/21 split data experiment is not shown here, its performance metrics indicated a notable improvement, with a dropout F1-score of 0.79 and overall accuracy around 90%. The 89/11 split achieved the highest accuracy of 91% with a dropout F1-score of 0.82. In addition, the macro average F1-score increased gradually from 0.80 to 0.88 as the proportion of training data increased. These findings demonstrate that a larger amount of training data has a positive contribution to improving the model's ability to classify student status, especially in increasing precision and F1-score for the dropout minority class.

B. Discussion

The exploration of different data split ratios in this study, aimed at predicting students at risk of dropout, reveals several key insights. The experimental results show that model accuracy varies with the data split ratio, ranging from 86% to 91%. Four factors affect the optimum accuracy value based on the data split ratio, as follows:

1. Proportion of training and testing data: The higher the proportion of training data, the more information the model can learn so that it can generalize to new data. This is reflected in the increase in accuracy from 86% at a 70/30 split to the highest accuracy of 91% at 89/11.
2. Amount of training data: Higher data split ratios, such as 89/11 or 90/10, allow the model to capture more complex patterns, increasing prediction accuracy. However, this improvement is not always linear. At a ratio of 90/10, the accuracy decreases slightly to 90%, indicating that adding training data beyond a certain point does not always result in significant improvements.
3. Test data variability: Smaller test sets, such as 10% or 11% of the data, reduce variability in the evaluation and may positively bias accuracy due to limited testing samples. On the other hand, larger test sets, such as 30%, provide a more realistic assessment of the model's generalization ability, even if the accuracy is slightly lower.
4. Data complexity and graph structure: Model performance is also affected by the complexity and variety of graph structures represented in the data. The more complex the relationships between entities, the harder it is for the model to generalize patterns with limited training data.

As shown in the comparative analysis of nine experiments using different data split ratios to predict students at risk of dropping out (Table VII), the 89/11 split yields the highest model accuracy of 91%.

TABLE VII. MODEL PERFORMANCE FOR DIFFERENT TRAIN–TEST SPLITS

No	Data split	Training samples	Testing samples	Precision	Recall	F1-score	Accuracy
1	70/30	197	85	0.98	0.85	0.91	0.86
2	75/25	212	71	1.00	0.87	0.93	0.89
3	79/21	223	59	1.00	0.88	0.93	0.90
4	80/20	226	56	1.00	0.87	0.93	0.89
5	83/17	234	48	1.00	0.85	0.92	0.88
6	85/15	240	42	1.00	0.86	0.92	0.88
7	88/12	248	34	1.00	0.86	0.92	0.89
8	89/11	251	31	1.00	0.88	0.94	0.91
9	90/10	254	28	1.00	0.86	0.93	0.90

Table VII presents the evaluation of model performance across the nine data split scenarios, ranging from 70/30 to 90/10. Overall, the model shows an increasing trend in performance as the proportion of training data grows. For the 70/30 split, the model achieved a precision of 0.98, recall of 0.85, F1-score of 0.91, and accuracy of 0.86. Performance improves consistently with larger training sets, reaching a precision of 1.00 in all scenarios except the 70/30 split.

The highest accuracy of 91% occurs at the 89/11 split, with 251 training records and 31 testing records. In this scenario, the performance metrics are calculated as follows:

- $\text{Precision} = \frac{22}{22+0} = 1$
- $\text{Recall} = \frac{22}{22+3} = 0.88$
- $\text{F1} - \text{Score} = 2 * \frac{1*0.88}{1+0.88} = 0.94$
- $\text{Accuracy} = \frac{22+7}{22+7+3+0} = 0.91$

The values above are calculated based on the model's predictions for a specific class. The precision is calculated by dividing the number of true positives (22) by the total number of positive predictions (22 true positives + 0 false positives), resulting in a precision of 1.00. This indicates that all positive predictions made by the model are correct, with no false positives. Furthermore, recall is obtained by dividing the number of true positives (22) by the total number of actual positives (22 true positives + 3 false negatives), yielding 0.88. This shows that the model correctly identifies 88% of all actual positives. The F1-score, calculated as the harmonic mean of precision and recall, is 0.94, indicating a good balance between prediction accuracy and coverage of actual positive cases. Finally, accuracy is 0.91 (91%), calculated as the ratio of correctly classified instances (true positives + true negatives = 29) to the total test data (32), demonstrating that the model correctly classifies the majority of the test samples.

Figure 3 shows the confusion matrix of the classification model for predicting student status, where a value of 1 represents a graduated student and 0 represents a dropout. The vertical axis (Actual) shows the true student status, whereas the horizontal axis (Predicted) shows the model's predictions.

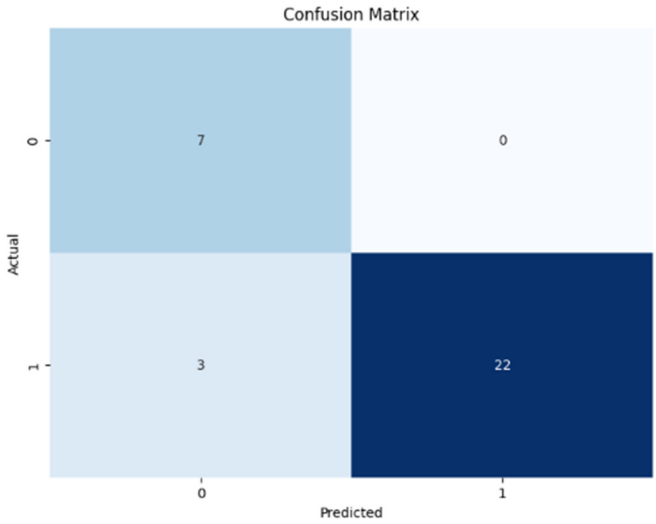

Fig. 3. Confusion matrix for 89/11 train–test split.

From the matrix, the model correctly identified 7 students as dropouts, representing True Negatives (TN: 0 → 0). There were no False Positives (FP: 0 → 1), indicating that no dropout students were incorrectly predicted as graduates. Three False Negatives (FN: 1 → 0) occurred, where students who graduated were mistakenly predicted as dropouts. The model achieved 22 True Positives (TP: 1 → 1), meaning 22 graduated students were correctly predicted.

This distribution indicates that the model is highly effective in detecting dropout students, with no misclassification of dropouts as graduates (no false positives). Overall, these results align with the evaluation metrics: precision = 1.00, recall = 0.88, F1-score = 0.94, and overall accuracy = 0.91. The high consistency between the confusion matrix and these metrics demonstrates the model's robustness and its balanced capability to classify both target classes. The results suggest that a sufficiently large training dataset, combined with an appropriate testing split, enhances prediction stability and generalization to unseen data.

The analysis results have clear practical implications for education, particularly in the efforts to detect the risk of student dropout early. Graph-based modeling can be used by educational institutions to identify academic patterns and learning relationships that indicate potential declines in student performance. By understanding the similarities among students, administrators can design more targeted interventions, such as academic counseling, workload adjustments, or additional learning support. The relevance of this research is further strengthened because the issue of dropout is a real problem that directly impacts graduate quality and institutional efficiency. Therefore, the research results are not only valuable methodologically but also provide an applicable basis for educational institutions to make data-driven decisions.

Finally, the results highlight the importance of the data splitting strategy. Differences in the proportion of training and testing data significantly affect model performance, influencing both subgraph matching and GED-based methods in terms of prediction stability and generalization. When the training data are more dominant, the model can form richer graph structure representations, allowing the patterns of student academic relationships to be identified more accurately. Conversely, in more balanced data splits or with limited training data, the complexity of the graph structure tends to decrease, affecting the accuracy of subgraph matching and increasing GED values. These findings emphasize that prediction quality is not only determined by the algorithm used but is also greatly influenced by the data splitting strategy before modeling. Furthermore, this study contributes methodologically by showing that data splitting is a critical yet often overlooked factor in graph mining. It also advances dropout prediction research by integrating and comparing two structural approaches that are sensitive to data composition. Overall, these findings provide a foundation for developing and evaluating more reliable graph-based dropout prediction models, particularly for imbalanced and dynamic educational datasets.

IV. CONCLUSION

This study proposed a graph-based analytical approach to predict student dropout risk by utilizing Graph Edit Distance (GED) and subgraph matching techniques. By representing students and courses as an undirected bipartite graph and analyzing structural similarities between graduated and dropout students, the model successfully identified at-risk students with high accuracy. Through a series of experiments using different data split ratios, the best performance was achieved with an 89/11 train–test split, yielding an overall accuracy of 91%, a precision of 1.00, a recall of 0.88, and an F1-score of 0.94.

The findings highlight that increasing the proportion of training data significantly improves the model's ability to generalize, enhancing both precision and recall, especially for the minority dropout class. Additionally, GED proved effective in capturing the degree of structural similarity between student graphs, enabling early detection of dropout patterns. The proposed method demonstrates the potential of graph analytics as a predictive tool within Educational Data Mining (EDM) to support student retention strategies.

Future work should consider integrating broader academic, socio-economic, and behavioral attributes to further refine the predictive power and real-world applicability of the model within educational decision-support systems. Moreover, this study can be extended by including verified dropout data for 2024 and 2025 once official records become available. These data can be used to externally validate the model trained on 2020–2023 data and to gradually update the modeling scheme (for example, using 2020–2024 data to predict 2025). Such an approach is expected to enhance the accuracy and practical relevance of the model for strategic dropout prevention in private higher education institutions.

ACKNOWLEDGMENT

The authors would like to express sincere gratitude for the support received during this research, which was made possible by the Institut Teknologi PLN Ph.D. Scholarship (Agreement No. 0004A.PJK/3/A0/2021) and Bina Nusantara University.

REFERENCES

[1] K. Oqaidi, S. Aouhassi, and K. Mansouri, "Towards a Students' Dropout Prediction Model in Higher Education Institutions Using Machine Learning Algorithms," *International Journal of Emerging Technologies in Learning*, vol. 17, no. 18, pp. 103–117, Sept. 2022, https://doi.org/10.3991/ijet.v17i18.25567.

[2] B. Alsubhi *et al.*, "Effective Feature Prediction Models for Student Performance," *Engineering, Technology & Applied Science Research*, vol. 13, no. 5, pp. 11937–11944, Oct. 2023, https://doi.org/10.48084/etasr.6345.

[3] E. J. Lizarte Simón and J. Gijón Puerta, "Prediction of early dropout in higher education using the SCPQ," *Cogent Psychology*, vol. 9, no. 1, Dec. 2022, Art. no. 2123588, https://doi.org/10.1080/23311908.2022.2123588.

[4] D. González-González, M. Arias-Corona, A. Cárdenas-Cruz, and A. Vicente-Bújez, "The impact of academic dropout at the University of Granada and proposals for prevention," *Frontiers in Education*, vol. 8, Feb. 2023, Art. no. 1110491, https://doi.org/10.3389/feduc.2023.1110491.

[5] W. Villegas-Ch, J. Govea, and S. Revelo-Tapia, "Improving Student Retention in Institutions of Higher Education through Machine Learning: A Sustainable Approach," *Sustainability*, vol. 15, no. 19, Oct. 2023, Art. no. 14512, https://doi.org/10.3390/su151914512.

[6] K. M. Sujon *et al.*, "The Effects of Imbalanced Datasets on Machine Learning Algorithms in Predicting Student Performance," *JOIV: International Journal on Informatics Visualization*, vol. 8, no. 3–2, pp. 1599–1605, Nov. 2024, https://doi.org/10.62527/joiv.8.3-2.2449.

[7] N. Mduma, K. Kalegele, and D. Machuve, "Machine learning approach for reducing students dropout rates," *International Journal of Advanced Computer Research*, vol. 9, no. 42, pp. 156–169, May 2019, https://doi.org/10.19101/IJACR.2018.839045.

[8] C. C. Gray and D. Perkins, "Utilizing early engagement and machine learning to predict student outcomes," *Computers & Education*, vol. 131, pp. 22–32, Apr. 2019, https://doi.org/10.1016/j.compedu.2018.12.006.

[9] L. Kemper, G. Vorhoff, and B. U. Wigger, "Predicting student dropout: A machine learning approach," *European Journal of Higher Education*, vol. 10, no. 1, pp. 28–47, Jan. 2020, https://doi.org/10.1080/21568235.2020.1718520.

[10] M. A. Hassan, A. H. Muse, and S. Nadarajah, "Predicting Student Dropout Rates Using Supervised Machine Learning: Insights from the 2022 National Education Accessibility Survey in Somaliland," *Applied Sciences*, vol. 14, no. 17, Aug. 2024, Art. no. 7593, https://doi.org/10.3390/app14177593.

[11] Y. Zhang, Y. Yun, H. Dai, J. Cui, and X. Shang, "Graphs Regularized Robust Matrix Factorization and Its Application on Student Grade Prediction," *Applied Sciences*, vol. 10, no. 5, Mar. 2020, Art. no. 1755, https://doi.org/10.3390/app10051755.

[12] Q. Hu and H. Rangwala, "Academic Performance Estimation with Attention-based Graph Convolutional Networks," in *Proceedings of The 12th International Conference on Educational Data Mining*, Montréal, Canada, 2019, pp. 69–78, https://doi.org/10.48550/arXiv.2001.00632.

[13] M. Anwar, A. E. Hassanien, V. Snášel, and S. H. Basha, "Subgraph Query Matching in Multi-Graphs Based on Node Embedding," *Mathematics*, vol. 10, no. 24, Dec. 2022, Art. no. 4830, https://doi.org/10.3390/math10244830.

[14] C. Piao, T. Xu, X. Sun, Y. Rong, K. Zhao, and H. Cheng, "Computing Graph Edit Distance via Neural Graph Matching," *Proceedings of the VLDB Endowment*, vol. 16, no. 8, pp. 1817–1829, Apr. 2023, https://doi.org/10.14778/3594512.3594514.

[15] L. T. M. Blessing and A. Chakrabarti, *DRM, a Design Research Methodology*. London, UK: Springer, 2009, https://doi.org/10.1007/978-1-84882-587-1.

[16] A. Haque *et al.*, "Implication of Different Data Split Ratio on the Performance of Model in Price Prediction of Used Vehicles Using Regression Analysis," *Data and Metadata*, vol. 3, pp. 425–425, Jan. 2024, https://doi.org/10.56294/dm2024425.

[17] K. Ayyavvu, B. I. Panneer, A. Sreenivasan, and A. K. A. Muthukrishnan, "Heart disease prediction using machine learning," *AIP Conference Proceedings*, vol. 2857, no. 1, Aug. 2023, Art. no. 020065, https://doi.org/10.1063/5.0165188.

[18] D. Jha *et al.*, "Enhancing materials property prediction by leveraging computational and experimental data using deep transfer learning," *Nature Communications*, vol. 10, no. 1, Nov. 2019, Art. no. 5316, https://doi.org/10.1038/s41467-019-13297-w.

[19] X. Gao, B. Xiao, D. Tao, and X. Li, "A survey of graph edit distance," *Pattern Analysis and Applications*, vol. 13, no. 1, pp. 113–129, Feb. 2010, https://doi.org/10.1007/s10044-008-0141-y.

[20] C. Solnon, "AllDifferent-based filtering for subgraph isomorphism," *Artificial Intelligence*, vol. 174, no. 12, pp. 850–864, Aug. 2010, https://doi.org/10.1016/j.artint.2010.05.002.

[21] Y. Liu, "ORB Feature Based Neighbor Graph Construction Method for Graph Regularized Non-Negative Matrix Factorization," *ICIC Express Letters Part B: Applications*, vol. 7, no. 10, pp. 2197–2203, Oct. 2016, https://doi.org/10.24507/icicelb.07.10.2197.

[22] C. Schröer, F. Kruse, and J. M. Gómez, "A Systematic Literature Review on Applying CRISP-DM Process Model," *Procedia Computer Science*, vol. 181, pp. 526–534, Jan. 2021, https://doi.org/10.1016/j.procs.2021.01.199.

[23] D. B. Blumenthal, N. Boria, J. Gamper, S. Bougleux, and L. Brun, "Comparing heuristics for graph edit distance computation," *The VLDB Journal*, vol. 29, no. 1, pp. 419–458, July 2019, https://doi.org/10.1007/s00778-019-00544-1.

[24] A. Vanacore, M. S. Pellegrino, and A. Ciardiello, "Fair evaluation of classifier predictive performance based on binary confusion matrix," *Computational Statistics*, vol. 39, no. 1, pp. 363–383, Feb. 2024, https://doi.org/10.1007/s00180-022-01301-9.

[25] L. Lavazza and S. Morasca, "Common Problems With the Usage of F-Measure and Accuracy Metrics in Medical Research," *IEEE Access*, vol. 11, pp. 51515–51526, 2023, https://doi.org/10.1109/ACCESS.2023.3278996.

[26] A. Sanfeliu and K.-S. Fu, "A distance measure between attributed relational graphs for pattern recognition," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. SMC-13, no. 3, pp. 353–362, May 1983, https://doi.org/10.1109/TSMC.1983.6313167.