Review of Computer Engineering Research
2025 Vol. 12, No. 4, pp. 218-227
ISSN(e): 2410-9142
ISSN(p): 2412-4281
DOI: 10.18488/76.v12i4.4559

A data-driven model for student retention in a Philippine higher education institution

Bell S. Campanilla[1+]
Jonathan O. Etcuban[2]
Patrick D. Cerna[3]

[1,2]*Cebu Technological University, Cebu City, Philippines.*
[1]*Email: bell.campanilla@ctu.edu.ph*
[2]*Email: joetcuban@gmail.com*
[3]*State University of Northern Negros, Sagay City, Philippines.*
[3]*Email: ptcerna@sunn.edu.ph.com*

(+ Corresponding author)

Article History
Received: 23 June 2025
Revised: 18 September 2025
Accepted: 21 October 2025
Published: 2 December 2025

Keywords
Decision tree
Degree completion prediction
Educational data mining
Knowledge discovery in databases
Student dropout analysis.

ABSTRACT

The impact of higher education inevitably enhances economic, social, and human capital development. In the Philippines, many students require assistance to complete their college degrees, resulting in high dropout rates. The study aimed to determine the attributes that contributed to students' degree completion, predict them using the Decision Tree (DT) algorithm, and assist university policymakers with information to help them create effective early intervention plans for at-risk students. The Knowledge Discovery in Databases (KDD) process extracts accurate information from 3,417 student records, which consist of academic and socio-economic attributes, after undergoing selection, preprocessing, transformation, data mining, and interpretation/evaluation. MS Excel and RapidMiner were used to explore data, build predictive models, and generate insights. The results show that the DT model achieves an accuracy of 82.50%. To finish a degree, students must have a grade point average (GPA) of 2.55 or lower and be no older than 19. Parents' educational background and living outside the city also affect the decision outcomes. Living outside the city and parents' academic backgrounds also influence GPA outcomes. The researchers recommend reviewing, analyzing, and, as necessary, revising previous policies to enhance university programs that address this long-standing issue.

Contribution/Originality: This article contributes to the development of a predictive model utilizing a Decision Tree algorithm to forecast graduates from colleges in the Philippines. This is the first analysis to combine academic, socioeconomic, and demographic aspects for machine learning applications. The originality lies in its localized approach, providing data-driven insights to support universities in improving retention strategies and student success interventions.

1. INTRODUCTION

With the advent of information technology (IT), the digital revolution has had a profound impact on academic organizations more than ever. Students these days have almost unrivaled access to knowledge due to the availability of internet-based databases and libraries. Individual talents and shortcomings are accommodated via adaptive learning platforms, and abstract ideas are brought to life through interactive tools and virtual simulations. Thanks to online platforms, students from all around the world can collaborate globally. Additionally, IT improves decision-making through data analysis and expedites administrative procedures.

Higher education institutions (HEIs) have historically been recognized in the Philippines for supporting the country's human resource needs by developing the personal, social, and economic capacities of their students.

However, there is a new development: many young people are failing to complete college, resulting in a significant increase in college dropouts. There has been variability in Philippine HEIs' dropout rates between and among the country's geographic regions. Between 2020 and 2021, the attrition rate was 15.90%. However, for the 2021–2022 period, it increased to 37.79%, and further to 40.98% in 2022–2023 [1].

The Commission on Higher Education (CHED) Chairperson stated that dropout rates have remained high since 2016, averaging 34%, particularly in private higher education institutions (HEIs) [2]. The COVID-19 pandemic accelerated this issue, especially between 2021 and 2023. Even before the pandemic, multiple factors contributed to college dropout rates. Several factors influenced dropout rates prior to the pandemic. These include cultural and economic factors, such as the availability of school-based infrastructure and resources, and geographical access to these.

On the other hand, students leave school for a variety of reasons, including financial difficulties, inadequate academic support, or issues within their family or community.

And it is not just individuals whose lives are impacted by the high dropout rate. This inability to utilize the brain's full potential influences the division of labor and resources within the country.

Dropouts, who are likely to earn far less than graduates, can become derelicts dependent on the state welfare system, which carries all societal implications. Due to the financial impact on HEIs, a decrease in tuition payments harms instructors' employment, infrastructure maintenance, and the provision of academic programs.

In one of the most prominent universities in the Philippines, in terms of student population, approximately 750 students (15%) of the total enrollment dropped out during the 2018–2019 academic year. This study constructed an empirical model that characterizes graduation trends for the 2021–2024 cycle. The student records from the 2018–2021 academic year were collected for this research, and the privacy and security of the data were protected by anonymizing personal identities a simple yet powerful technique for generating classifiers based on conditions. The DT algorithm is effective even with small datasets, which makes it suitable for this research. The model aimed to develop strategies to enhance retention and graduation rates by first identifying the causes of student dropouts.

2. OBJECTIVES OF THE STUDY

This study aimed to determine the accuracy of the DT model in predicting college students' degree completion in private HEIs in the Philippines. Specifically, it sought answers to the following questions: 1) Identify critical attributes contributing to student degree completion through data mining; 2) Forecast students' degree completion utilizing the DT algorithm; and 3) Provide information to university administrators to inform them of the development of early intervention strategies for students identified as at-risk.

3. SCOPE OF THE STUDY

The following was the scope of the study:

- The study was conducted among one of the largest student populations in Cebu, Philippines.
- The dataset covers the academic year 2021-2024.
- All graduates of the university's baccalaureate programs.

4. SIGNIFICANCE OF THE STUDY

The study has several positive implications for the following people:

a. *Students.* The study findings may identify students at an early stage who are at risk of dropping out or are more likely to do so. This implies that struggling students can receive the required assistance before it is too late.

b. *School Administrators.* The study is important to them because it provides a method for identifying which students are most likely to drop out of college prematurely. If identified early, school administrators can intervene and provide these students with the support they need to succeed.
c. *Faculty.* The results will equip faculty members to recognize and assist at-risk students. They can collaborate with other stakeholders, such as parents, counselors, and administrators, to support students.
d. *Researchers.* The study is essential to the proponents because it will provide valuable information about current patterns in student dropouts and how to address the issue in the future by leveraging the available technological resources.
e. *Future researchers.* Following the output model, they can take concrete steps to improve the current situation by introducing specific modifications to the algorithm or utilizing it to achieve breakthroughs. This will provide them with specific knowledge and skills in advance. Additionally, this research will serve as a blueprint and guide for those intending to upgrade it.

5. RELATED STUDIES

5.1. Attributes Affect Students' Dropout

The studies in the latter group reflect a diversity of models of student performance. Hoque et al. [3] had access to a comprehensive set of details regarding demographic, academic, and behavioral variables, including extracurricular activities, attendance, and prior-year GPA, to better understand the predictors. Conversely, Yaacob et al. [4] narrowed the scope of prediction by essentially considering academic features, including the final CGPA and the course grade. Tamayo et al. [5] also highlighted the importance of pre-admission factors, such as entrance examination scores and school type, suggests that academic preparedness is a key consideration. This research further contributes to this line of inquiry by integrating these three lines of research, utilizing a multidimensional dataset with demographic, behavioral, and academic variables. Unlike existing literature, which is predominantly focused on isolated academic or pre-admission information, the innovative approach of using aggregations to enhance the qualities of both parent and precursor records provides comprehensive insight into student success in various educational scenarios.

The studies reviewed here offer several approaches to predicting student performance, each with its advantages and disadvantages. One study utilized a wide range of demographic, academic, and behavioral variables to gain a deeper understanding of the predictors of performance. However, this multi-variable model is also more complex for data handling and analysis. In contrast, a second study focused on a narrow scope, primarily examining GP and course grades. While this aids in generating the modeling exercise, it also reduces the predictive ability if significant behavioral and socio-demographic determinants are left out. A second study looked at preadmission factors, which represent baseline academic readiness. However, it disregards in-program factors, which may be key predictors of longer-term outcomes. Building on this literature, the current study also benefits from being able to include a wide range of measures, including demographic, behavioral, and school-related features. This holistic perspective is unprecedented in the literature and was selected to achieve the most accurate prediction and improved comprehensibility of student performance in various learning settings.

5.2. Utilizing the DT Algorithm

The reviewed papers imply that predictive modeling using DTs can predict student performance. Cardona et al. [6] underlined the capability of DT to predict direct predictors of dropout/completion. Similarly, Al-Kmali et al. [7] also showed that the prediction of DT over the semester in GPA was performed exceptionally well, and Mengash [8] emphasized that the interpretability of nuanced decision rules (simplified by the DT model) was clarified. Alsayed, et al. [9] observe that DT can handle numerical data without complicated preprocesses, which is another advantage, especially for small datasets. Using these works as a foundation, the current paper uses DT to integrate several academic, demographic, and behavioral characteristics, but also advances this literature by applying the approach to the larger, multidimensional set of data with greater predictive strength.

Introduction of additional longitudinal tracking and other psychosocial factors into the study may increase the stability and predictability of the DT model. Variable selection methods can help determine the most informative predictors, thereby simplifying the model without compromising model accuracy. Providing transparency for models and user-friendly visualization tools can offer teachers and school heads practical insights. For HEIs, such trends enable the earlier identification and prevention of dropouts, thereby increasing the possibility of intervening and supporting students in a personalized manner. This model of proactive outreach can contribute to decreased attrition and increased persistence, streamlining resources to individual circumstances more effectively and justly, resulting in higher levels of academic success and institutional effectiveness.

6. METHODOLOGY

The researchers requested permission from the university administrators to conduct the study. After the transmittal letter was approved, the researchers then prepared the necessary and precautionary measures to gather the data. They strictly adhered to the ethical considerations and the Philippines' Data Privacy Act of 2012 during the conduct of the research.

In this study, KDD was utilized, referring to the non-trivial process of discovering valid, novel, potentially useful, and ultimately understandable patterns in data. It was coined in 1989 to describe the process of discovering knowledge in data and to emphasize the "high-level" use of specific data mining techniques [10]. This process involves extracting both useful data and previously unknown, potentially valuable information from a large dataset. It is an iterative process and requires multiple iterations to extract accurate knowledge from the data [11].

In the first step, the data sources are defined, where data and results are analyzed using a subset of the original features, providing the most information for the output [11]. Reducing the dimensionality of the data set by moving irrelevant and outlier data to facilitate the data mining step [12].

Afterward, the dataset is filtered and assembled into a tabular form that is easy to comprehend and digest [13]. It presents data cleaning, the dismissal of irrelevant data, and methods for imputing missing values, all while improving the dataset [11]. The researchers used Microsoft Excel (MS Excel) to conduct data hygiene, ensuring that formats were strictly adhered to.

During the transformation step, the initial data processing and the derivation of new variables from the existing ones were strictly followed using RapidMiner. It aimed to understand the variables, their importance, interactions, and relationships among the identified variables. This step involves reducing the data by applying feature selection techniques to decrease the dataset's dimensionality.

The *data mining* step represents the core of the KDD, comprising statistical and machine-learning models that can uncover structures, correlations, patterns, rules, and anomalies in the data, resulting in the best model to apply for prediction.

The final step was the interpretation and evaluation step, which involves passing the data for interpretation and documentation. The curated output has been cleaned, transformed, and filtered based on relevant characteristics, and then packaged into visual displays for human consumption to assess the curated output [14].

6.1. Selection Step

In this study, the prediction of degree completion is based on the dataset of graduates from the 2021-2024 batch at the university. The dataset used in this study was retrieved from the university's data center. The database contains 3,417 entities, nine semesters, and 17 socio-demographic attributes. For academic attributes, namely, students' junior high school (JHS) GPA, four semesters of senior high school (SHS) GPA, and four semesters of college GPA. Socio-demographic attributes include the student's gender, age, living outside the city, living in a boarding house, scholarship status, number of units enrolled, living with family, type of junior high school (private or public), mother's and father's educational attainment, and parents' monthly income. These attributes were retrieved from the university data center warehouse.

Table 1 depicts the attributes, descriptions, and types in this study.

Table 1. Attributes of the study.

Attribute	Description	Type
JHS_AVE	Junior High School	Numeric
171_SHS	1^{ST} Semester SHS	Numeric
172_SHS	2^{nd} Semester SHS	Numeric
181_SHS	1^{ST} Semester SHS	Numeric
182_SHS	2^{nd} Semester SHS	Numeric
191_GPA	1^{st} Year– 1^{st} Sem GPA	Numeric
192_GPA	1^{st} Year – 2^{nd} Sem GPA	Numeric
201_GPA	2^{nd} Year– 1^{st} Sem GPA	Numeric
202_GPA	2^{nd} Year – 2^{nd} Sem GPA	Numeric
SHS_GPA_AVERAGE	SHS GPA Average	Numeric
COLL_GPA_AVERAGE	College GPA Average	Numeric
191_UNTS	1^{st} Sem Total Units -1^{st} Year	Numeric
192_UNTS	2^{nd} Sem Total Units – 1^{st} Year	Numeric
201_UNTS	1^{st} Sem Total Units – 2^{nd} Year	Numeric
202_UNTS	2^{nd} Sem Total Units – 2^{nd} Year	Numeric
UNITS_AVERAGE	College Average Units	Numeric
Sex	Student Gender	Text
Age	Student Age	Numeric
IS_PROVINCE	Living outside city	Numeric
IS_BOARDING	Renting Boarding House	Numeric
JHS_TYPE	Junior High School Type	Text
IS_SCHOLAR	Student Scholarship	Numeric
IS_LIVING_WITH_FAMILY	Living with family	Numeric
FATHER EDUC BACKGROUND	Father's educational background	Numeric
MOTHER EDUC BACKGROUND	Mother's educational background	Numeric
PARENT_MONTLY_INCOME	Parents Monthly Income	Numerical
IS_GRADUATE (target attribute)	Student Graduate	Text

6.2. Step 2: Preprocessing

Data processing is one of the primary steps in data analysis. The dataset for this study was available in a CSV file. Data purification, processing of missing information, identification of duplicate records, and transformation of categorical variables into numerical ones were computerized using Microsoft Office Excel software. In this phase, we had to hygiene-check the dataset to eliminate unnecessary attributes that could lead to flawed results.

6.3. Step 3: Transformation

The gain ratio was applied to rank the attributes. From 17 attributes, the number was reduced to 11 after conducting the gain ratio analysis. The *Select Attributes operator* (Figure 1) was utilized, resulting in the reduction of irrelevant attributes by manually selecting specific attributes. The operator was used to narrow down the dataset to the most essential characteristics.

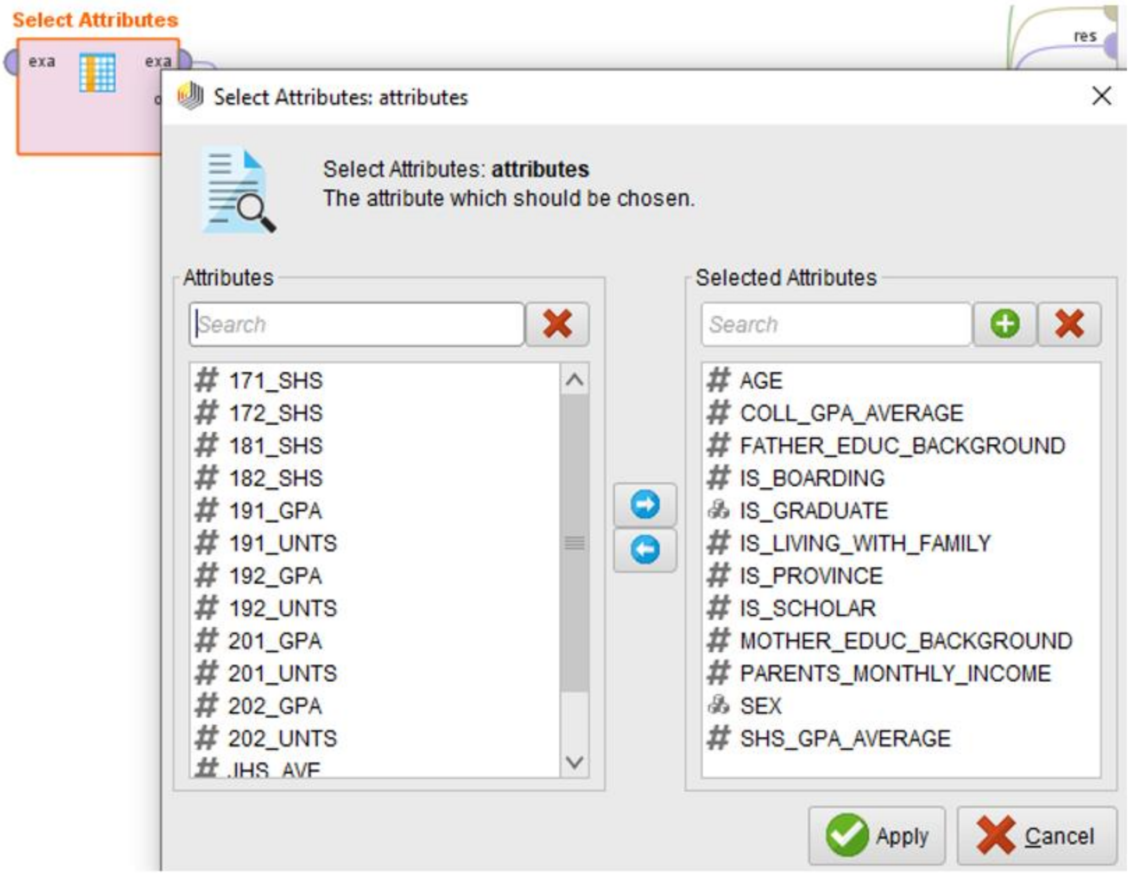

Figure 1. Preprocessing using Select Attributes.

The *Set Role* operator selected the target variable or attribute to predict. In this study, the target attribute was IS_Graduate, as shown in Figure 2.

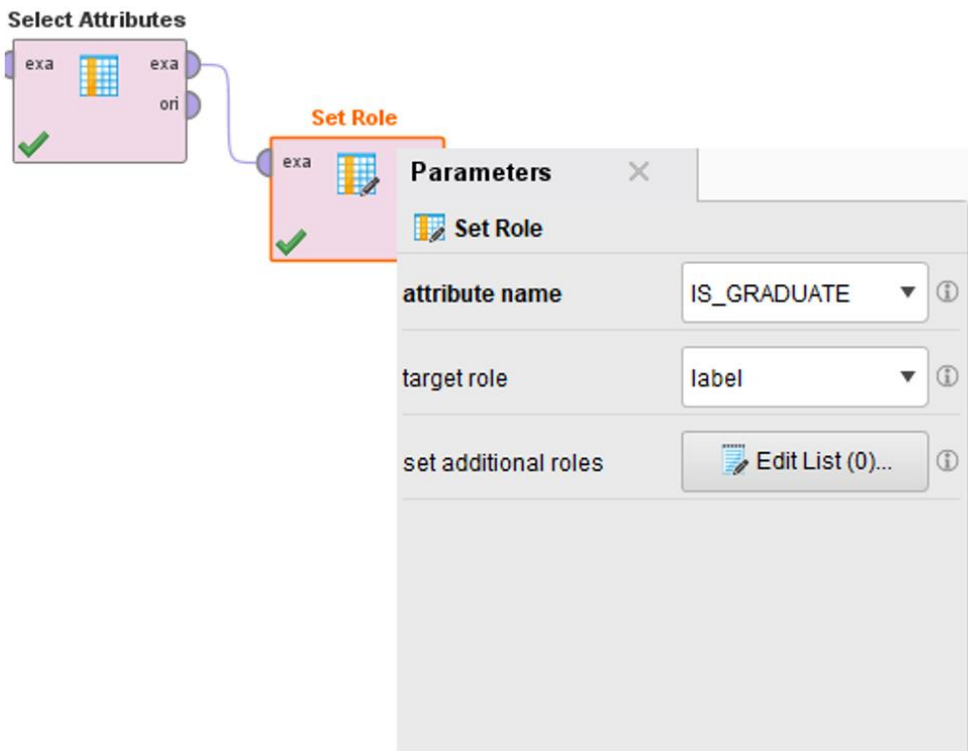

Figure 2. Selecting the target attribute using the set role.

Cross-validation was employed in this study to ensure the model's effectiveness and to prevent overfitting or underfitting (Figure 3). The dataset was divided into training and testing sets. The number of folds, *k*, was set to 10, and stratified sampling was employed, a type of sampling that leads to a more accurate and reliable model evaluation.

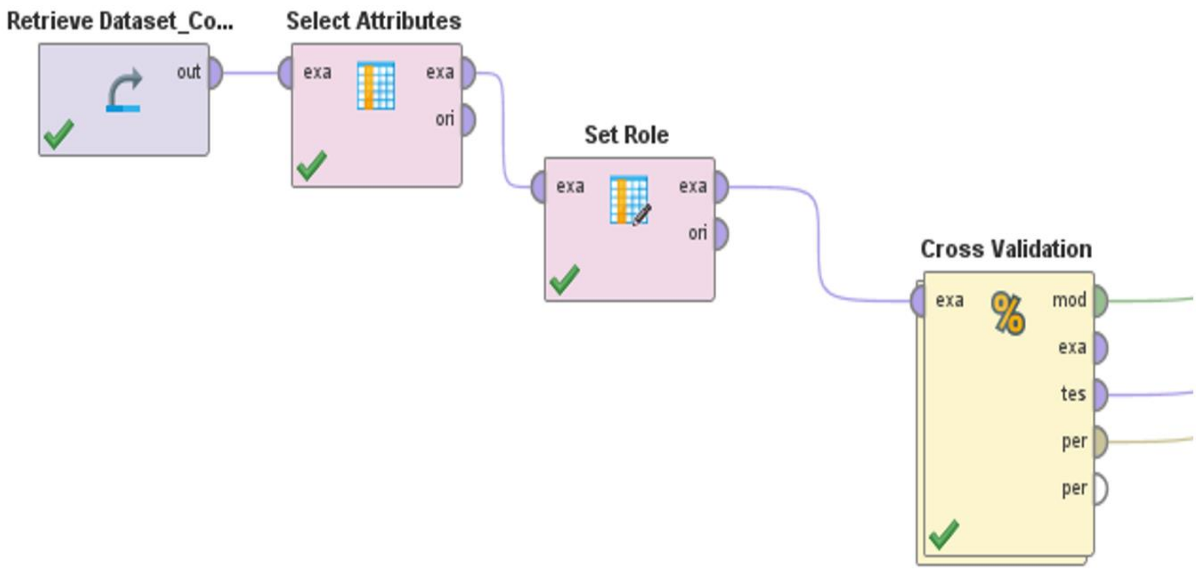

Figure 3. Cross-validation for overfitting and underfitting.

6.4. Step 4: Data Mining

The DT algorithm (Figure 4) created a model that predicted the target value, *IS_GRADUATE, for the classification process.* The parameters were the following: maximal depth = 10, confidence = 0.1, minimal gain = 0.01, minimal leaf size = 2, minimal size for split = 4, and several prepruning alternatives = 3. This study set the maximal depth to 10, limiting the tree's size to avoid overfitting. The confidence level was set to 0.1 to simplify the model without compromising prediction accuracy. The minimum leaf size was set to 2 to ensure that splits occurred only when there was a significant increase in prediction and appropriate data were available for each leaf. In contrast, the minimal size for the split was set to 4 to avoid overly specific splits on the datasets.

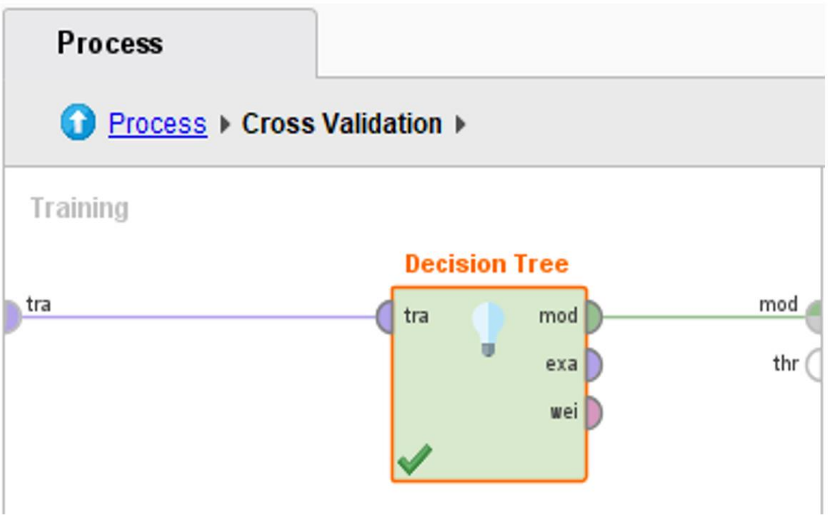

Figure 4. Decision Tree algorithm for data mining.

6.5. Step 5. Interpretation/Evaluation

In Figure 5, the *Apply Model* operator was used to predict the label value for each entry in the test set. A new column, titled *prediction (IS_GRADUATE)*, was added to the results, representing the predicted outcomes. After generating this prediction column, the model's performance was evaluated using the *Performance* operator (Figure 6) to assess its accuracy and effectiveness.

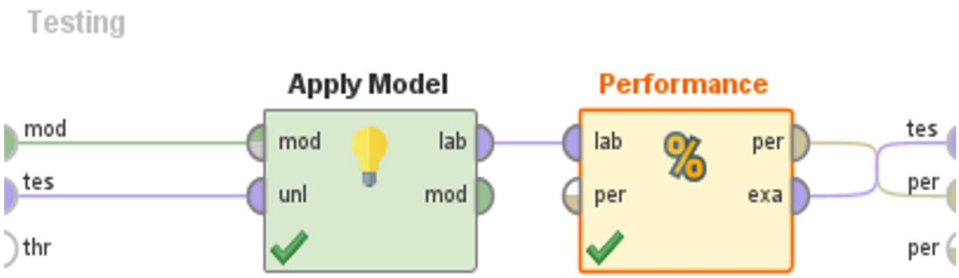

Figure 5. Apply model and performance – evaluation of models.

7. RESULTS AND DISCUSSIONS

The primary attribute in this study, which was the target, was *IS_Graduate*. A classification model was created to predict the likelihood of degree completion. *Not Completed* indicates that students are unlikely to graduate, while *Completed* indicates that students are likely to graduate.

Only 11 of the 26 attributes in the dataset were adjusted: two academic attributes, namely, SHS GPA and college GPA; nine socio-demographic attributes, namely, sex, age, living outside the city, living in a boarding house, living with family, scholar status, father's and mother's educational background, and parents' income.

The adjusted attributes and utilization of the DT algorithm in predicting degree completion present an accuracy rate of 82.50% (Figure 6). This means that the model correctly predicted instances where the actual outcome was either *"Completed"* or *"Not Completed."*

Moreover, this implies that the 3,417 records resulted in 1,207 students who finished their college degree programs and were correctly predicted by the model, and about 511 students who did not complete their degree program but were incorrectly predicted as graduates.

PerformanceVector

```
PerformanceVector:
accuracy: 82.50% +/- 1.74% (micro average: 82.50%)
ConfusionMatrix:
True:   Completed       Not Completed
Completed:      1207    511
Not Completed:  87      1612
precision: 94.95% +/- 2.17% (micro average: 94.88%) (positive class: Not Completed)
ConfusionMatrix:
True:   Completed       Not Completed
Completed:      1207    511
Not Completed:  87      1612
recall: 75.93% +/- 2.87% (micro average: 75.93%) (positive class: Not Completed)
ConfusionMatrix:
True:   Completed       Not Completed
Completed:      1207    511
Not Completed:  87      1612
f_measure: 84.33% +/- 1.74% (micro average: 84.35%) (positive class: Not Completed)
ConfusionMatrix:
True:   Completed       Not Completed
Completed:      1207    511
Not Completed:  87      1612
```

Figure 6. Performance accuracy, precision, recall, and F1.

In contrast, in the *True Negative* (TN) category, Figure 6 shows that 1,612 students were correctly predicted to be unable to graduate from their college degree programs; however, 87 students who did complete the degree programs were incorrectly predicted as non-completers.

The model's precision correctly predicted 94.95%, which means that there were only 87 students who were incorrectly predicted as at-risk out of 1,612. The model was effective at identifying students who are at risk of not completing their degrees.

It generated a recall score of 75.93%, indicating that the model accurately identified students who did not complete their college degrees. Specifically, it correctly predicted 1,612 students as being at risk of not graduating, but incorrectly predicted that 511 students who failed to complete their degrees would graduate. Furthermore, there were only 87 students who were wrongly labeled as at risk but actually graduated.

The results present the factors that influence whether students complete their studies. Students with a 2.55 college GPA are often classified as "Not Completed," especially for students over 19.5 years old. Living outside the city and parents' educational background also affect college GPA outcomes. Students with higher SHS GPAs are more likely to complete a degree. Additionally, students who are scholars and live with their families are more likely to complete a degree, but this depends on other factors, such as college GPA and age. Based on the micro average squared error *(squared_error)*, which is 0.0128, the result indicates that the model's predictions are close to the actual values, with only minor average errors.

8. CONCLUSIONS

The study concluded that the DT model accurately predicted college students' degree completion in a private university in Cebu City, Philippines. This indicates that the model's performance has been satisfactorily predicted at 82%. The predicted effects were mirrored in the observations. The students' propensity to complete college is primarily driven by a combination of their demographic profile, family background, and initial academic success. Based on the findings, it was concluded that the DT algorithm is an effective tool for predicting degree completion. DT is straightforward to understand and can handle both numerical and categorical data.

The researchers recommend that university policymakers proactively strengthen and address the perennial problem of student dropouts by establishing in-house dormitories for students who live outside the city. Additionally, they recommend developing academic programs, such as course tutorials, peer collaboration, and adaptive bridging courses for students who struggle to adjust to college life. Empowering faculty is crucial for reducing student dropouts, as teachers are often the first line of support, providing extra guidance and assistance to students at risk.

Funding: This study received no specific financial support.
Institutional Review Board Statement: The study involved minimal risk and adhered to ethical guidelines for social science fieldwork. Formal approval from an Institutional Review Board was not required under the policies of Cebu Technological University, Philippines. Informed verbal consent was obtained from all participants, and all data were anonymized to ensure participant confidentiality.
Transparency: The authors state that the manuscript is honest, truthful, and transparent, that no key aspects of the investigation have been omitted, and that any differences from the study as planned have been clarified. This study followed all writing ethics.
Competing Interests: The authors declare that they have no competing interests.
Authors' Contributions: All authors contributed equally to the conception and design of the study. All authors have read and agreed to the published version of the manuscript.

REFERENCES

[1] Z. Sarao, *Dropout rate in universities, colleges at 35.15% in SY 2023–2024. Inquirer.net.* Makati City, Philippines: Philippine Daily Inquirer, 2024.

[2] R. G. Cruz, *CHED reports higher enrollment, but graduation, drop out rates unchanged.* Quezon City, Philippines: ABS-CBN News, 2024.

[3] M. I. Hoque, A. K. Azad, M. A. H. Tuhin, and Z. U. Salehin, "University students result analysis and prediction system by decision tree algorithm," *Advances in Science, Technology and Engineering Systems Journal*, vol. 5, no. 3, pp. 115-122, 2020. https://doi.org/10.25046/aj050315

[4] W. F. W. Yaacob, N. M. Sobri, S. A. M. Nasir, N. D. Norshahidi, and W. Z. W. Husin, "Predicting student drop-out in higher institution using data mining techniques," presented at the Journal of Physics: Conference Series (Vol. 1496, No. 1, p. 012005). IOP Publishing. 2020.

[5] J. D. Tamayo, N. V. Francisco, M. E. P. Malonzo, and A. P. Bugay, "Predicting students' degree completion using decision trees," *International Journal of Computer*, vol. 28, no. 1, pp. 75–89, 2018.

[6] T. A. Cardona, E. A. Cudney, and J. Snyder, "Predicting degree completion through data mining," presented at the 2019 ASEE Annual Conference & Exposition, 2019.

[7] M. Al-Kmali, H. Mugahed, W. Boulila, M. Al-Sarem, and A. Abuhamdah, "A machine-learning based approach to support academic decision-making at higher educational institutions," presented at the 2020 International Symposium on Networks, Computers and Communications (ISNCC), Montreal, QC, Canada. https://doi.org/10.1109/ISNCC49221.2020.9297177, 2020, pp. 1-5.

[8] H. A. Mengash, "Using data mining techniques to predict student performance to support decision making in university admission systems," *IEEE Access*, vol. 8, pp. 55462-55470, 2020. https://doi.org/10.1109/ACCESS.2020.2981905

[9] A. O. Alsayed *et al.*, "Selection of the right undergraduate major by students using supervised learning techniques," *Applied Sciences*, vol. 11, no. 22, p. 10639, 2021. https://doi.org/10.3390/app112210639

[10] A. Azevedo, *Data mining and knowledge discovery in databases. In Advanced methodologies and technologies in network architecture, mobile computing, and data analytics.* M. Khosrow-Pour, Ed. Hershey, PA, USA: IGI Global. https://doi.org/10.4018/978-1-5225-7598-6.ch037, 2019, pp. 502–514.

[11] H. M. Safhi, B. Frikh, and B. Ouhbi, "Assessing reliability of big data knowledge discovery process," *Procedia Computer Science*, vol. 148, pp. 30–36, 2019. https://doi.org/10.1016/j.procs.2019.01.005

[12] D. Oreški and T. Novosel, "Comparison of feature selection techniques in knowledge discovery process," *TEM Journal*, vol. 3, no. 4, pp. 285-290, 2014.

[13] C. A. Palacios, J. A. Reyes-Suárez, L. A. Bearzotti, V. Leiva, and C. Marchant, "Knowledge discovery for higher education student retention based on data mining: Machine learning algorithms and case study in Chile," *Entropy*, vol. 23, no. 4, p. 485, 2021. https://doi.org/10.3390/e23040485

[14] N. Hotz, "KDD and data mining. Data Science Process Alliance," Retrieved: https://www.datascience-pm.com/kdd-and-data-mining/, 2024.