Nanotechnology Perceptions
ISSN 1660-6795
www.nano-ntp.com

Enhancing BSIT Track Selection Using Predictive Modeling: A Multi-Class Classification Approach

Rosemarie M. Bautista

College of Information and Communications Technology, Bulacan State University, City of Malolos, Bulacan, Philippines
Email: rosemarie.bautista@bulsu.edu.ph

In today's evolving educational landscape, aligning students' skills with industry demands is essential. This study employs predictive modeling to improve the Bachelor of Science in Information Technology (BSIT) track selection at Bulacan State University. The BSIT program includes foundational courses in the first two years, followed by specialized tracks — Business Analytics (BA), Service Management (SM), and Web and Mobile Application Development (WMAD) — in the final two years. The study aimed to determine the characteristics of students who completed the requisite course requirements per track and graduated within the stipulated timeframe. It analyzed 481 student records admitted to the BSIT program during the 2019-2020 academic year, divided into a training set and a test set. Using the Cross-Industry Standard Process for Data Mining (CRISP-DM) and data-driven methods, this research recommends suitable specialization tracks - BA, SM, or WMAD - based on students' demographic and academic profiles. By analyzing historical data from timely graduates, significant attributes influencing track selection were identified. The study addresses class imbalance using the Synthetic Minority Over-sampling Technique (SMOTE) and evaluates model performance through accuracy, kappa statistics, and confusion matrix. The developed model is incorporated into a prototype prediction system, enhancing decision-making for personalized educational pathways and supporting timely graduations.
Keywords: predictive modeling, track selection, multi-class classification, data mining.

1. Introduction

In today's rapidly advancing needs in the industry sector, higher education institutions face the challenge, specifically in the IT industry, of preparing students for a dynamic job market, where graduates often need more practical skills required by employers despite having a theoretical foundation. This mismatch affects employability and productivity, creating challenges in meeting the demand for highly skilled IT professionals in the Philippines. The gap is caused by outdated curricula and the rapid evolution of technology (DICT, 2022; CHED, 2020; PSIA, 2020). Industry stakeholders emphasize that aligning the BSIT curriculum with modern industry needs requires closer collaboration between academic institutions and IT companies, ensuring the development of relevant skills and competencies (PSIA, 2020; World Bank, 2021). To address this, many universities have revisited and restructured their curricula to provide initial broad-based learning, followed by specialized tracks designed to prepare students for specific roles in the workforce. However, ensuring students select the most appropriate track based on their skills, interests, and academic performance remains a persistent challenge.

The advent of big data and machine learning that revolutionized various sectors, including education, presents a potential solution to this challenge. In the educational context, these technologies can significantly enhance decision-making processes by identifying student characteristics patterns to predict performance and guide students toward the right track. Utilizing a machine learning algorithm provides valuable insights for academic planning and enhances the track selection process to support timely graduation (Salman et al., 2019; Hariri et al., 2020). By leveraging data-driven insights, educational institutions can improve student outcomes and operational efficiencies (Kumar & Chadha, 2020; Li et al., 2021).

Bulacan State University, true to its mission to produce highly competent professionals, has adopted a curriculum that offers broad foundational knowledge in the first two years of its Bachelor of Science in Information Technology (BSIT) program, followed by specialized tracks in Business Analytics (BA), Service Management (SM), and Web and Mobile Application Development (WMAD). This program structure aims to equip students with the technical and analytical skills necessary to prosper in the modern IT landscape. However, the challenge lies in ensuring students are guided toward the track that best aligns with their abilities. The current track selection process is based on the crafted criteria of the BSIT curriculum committee, which revolves around students' choices and grades in relevant foundational courses, which may only sometimes account for the set of factors influencing academic success.

This study seeks to enhance the BSIT track selection process by employing a predictive modeling approach. Within this framework, the author leverages historical data from students who graduated on time to identify key demographic and academic attributes that influence successful track completion. Predictive analytics has been shown to improve academic planning and student outcomes, particularly when used to tailor educational experiences to individual needs (Tiwari et al., 2019; Ali et al., 2021). By integrating these insights into a web-based prediction system, this study aims to create a practical tool that enhances decision-making for both students and academic advisors, ultimately supporting timely graduation and improved educational outcomes. The insights gained from this study

will not only enhance the academic experience for BSIT students at Bulacan State University but also provide a valuable model for other institutions seeking to optimize their specialization programs through data-driven decision-making.

The main objective of this study is to provide track selection recommendations for incoming third-year BSIT students using a predictive model. Specifically, the study aims to:

1. Identify the most distinctive attributes that significantly contribute to predicting and recommending appropriate BSIT tracks.
2. Address class imbalances in the dataset using data augmentation techniques.
3. Determine the machine learning algorithm best suited to the task.
4. Evaluate and validate the model's performance using appropriate evaluation metrics.

2. Methodologies

This research applies the Cross-Industry Standard Process for Data Mining (CRISP-DM), a widely adopted methodology for structuring data mining projects, as shown in Figure 1. CRISP-DM allows integration at each step, reducing the risk of not meeting project objectives and providing a structured guide for achieving project goals (Wirth & Hipp, 2000).

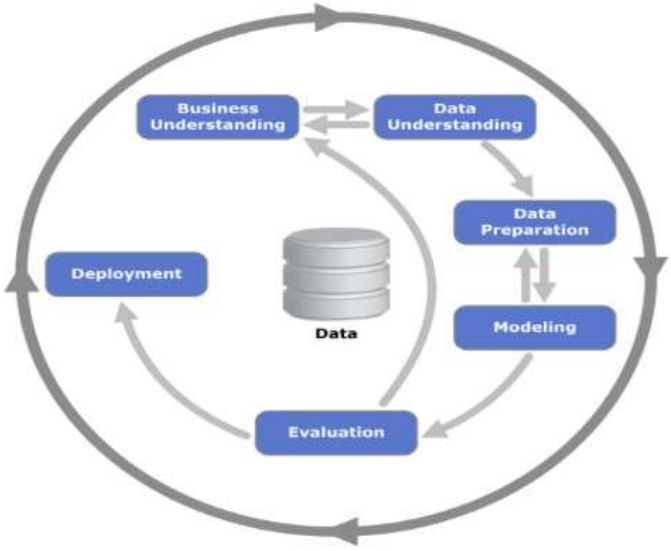

Figure 1. The Cross-Industry Standard Process for Data Mining

The research follows the CRISP-DM methodology, which consists of the following phases:

Business Understanding

In this initial phase, the researcher assessed the college's current situation to define the purpose of the study and convert this knowledge into a data mining problem. Specifically, the study aims to develop a predictive model to recommend the most suitable BSIT track for

each student, enhancing the track selection process and improving student outcomes. This involved analyzing the BSIT curriculum, student demographics, and academic data to determine the factors influencing successful track selection.

Data Understanding

In the Data Understanding phase, the researcher familiarized herself with the student's demographic data and academic performance, mainly their grades in the first and second years. These data were gathered from the university's Registrar's Office and surveys. The initial dataset comprised combined academic and demographic profiles, which were expected to contribute to the predictive model. During this phase, the researcher analyzed the structure and types of variables, reviewed the initial statistics, and assessed data quality by checking for missing values, inconsistencies, and other impurities (Han et al., 2011).

Data Preparation

During the Data Preparation phase, the researcher merged data from the registrar's office and surveys, ensuring the proper linking of student IDs to demographic and academic records. Microsoft Excel and OpenRefine were used to clean and organize the dataset, addressing inconsistencies and missing values. Data transformation processes were carried out to enhance data quality, including smoothing and discretization. Attribute selection using WEKA was performed to identify the most significant attributes for building the predictive model. Additionally, the SMOTE (Synthetic Minority Over-sampling Technique) method was applied to address class imbalance (Chawla et al., 2002). This step ensured the data was ready for further processing and model development.

Modeling

In this phase, the researcher selected machine learning algorithms based on their suitability for the classification problem. Algorithms such as J48, Random Forest, and Random Tree were chosen due to their proven effectiveness in handling classification tasks and their capacity to produce interpretable models, particularly in educational data mining (Zhao et al., 2020). The algorithms were trained and tested on the prepared dataset to predict the most suitable BSIT track for students.

Results Evaluation

The researcher evaluated the performance of the models using several key metrics, including accuracy, kappa statistics, mean absolute error, true positive and false positive rates, precision, recall, and execution time. These metrics were selected to comprehensively assess the model's performance. For instance, accuracy provides a primary measure of overall correctness, while kappa statistics help account for random chance in the predictions (Sun et al., 2019). The evaluation phase was crucial in validating the machine learning model and ensuring its effectiveness and reliability in predicting student outcomes.

Deployment

Although the CRISP-DM process includes a deployment phase where the findings are put into a practical application, this study focused on the results evaluation phase and did not extend to deployment. However, the prototype developed in this study could be used to predict the most suitable BSIT specialization track for students. The web-based prediction

system prototype was built using Extreme Programming (XP) methodology, ensuring the system could be easily adapted and iterated based on future needs (Beck, 1999).

3. Results and Discussion

Data Pre-processing and Attribute Selection

Data pre-processing is a crucial, yet often time-consuming, task that all data enthusiasts and researchers must complete before engaging in machine learning algorithms. Failing to execute this step accurately can significantly affect the model's performance and results. By cleansing the data effectively, one can enhance machine learning models' accuracy and reliability (Zhu et al., 2019; Kotsiantis et al., 2022).

The first data processing task involved integrating academic records from the registrar's office with student demographic data, using students' ID numbers as a reference, resulting in a comprehensive dataset organized for further analysis. After integration, various tasks, such as resolving inconsistencies, discretization, and encoding, were performed to simplify the data and ensure its reliability for subsequent analytics (Wickham & Grolemund, 2021). The researcher employed OpenRefine software to facilitate these steps.

A significant component of data pre-processing was discretizing household income into categories, as presented in Table 1. The discretization step facilitated a more straightforward interpretation of students' socioeconomic status (Padillo, 2023).

Table 1 Discretization of Social Status

Social Status	Household Income Range	Label
Low Income Class	Less than 18,200 a month	1
Lower Middle Income Class	18,201 – 36,400 a month	2
Middle Income Class	36,401 – 63,700 a month	3
Upper Middle Income Class	63,701 – 109,200 a month	4
High-Income Class	109,201 and above a month	5

Table 1 shows how the continuous variable, household income, is binned into five discrete meaningful categories: low income, lower-middle income, middle income, upper-middle income, and high income, in alignment with social status classifications. Afterward, label encoding is applied, where each category (social class) is assigned a numerical value (1, 2, 3, 4, 5).

After thoroughly cleansing the dataset, the researcher obtained the most significant attributes using the CFS Subset Evaluator with the Best First and Greedy Stepwise (Forward) search methods. Figure 2 shows the significant attributes identified.

```
Search Method:
        Greedy Stepwise (forwards).
        Start set: no attributes
        Merit of best subset found:    0.11

Attribute Subset Evaluator (supervised, Class (nominal): 43 Specialization):
        CFS Subset Evaluator
        Including locally predictive attributes

Selected attributes: 3,5,7,11,22,24,25,32,33,36,38,39,41 : 13
                     Sex
                     FamilyStructure
                     HouseholdIncome
                     Strand
                     STS101
                     ARP101
                     PAL101
                     IT103
                     IT104
                     IT107
                     IT202
                     IT203
                     IT205
```

Figure 2 Attributes Selection: CfsSubsetEval with GreedyStepwise

The variables in Figure 2, "sex," "family structure," etc., represent a mix of demographic and academic factors identified as essential for predicting student track or specialization. These attributes were chosen because they demonstrated the highest predictive power for the target class "Track" while maintaining low redundancy.

Addressing Dataset Imbalance

A significant challenge in the dataset was class imbalance across different tracks in the BSIT program, as shown in Figure 3.

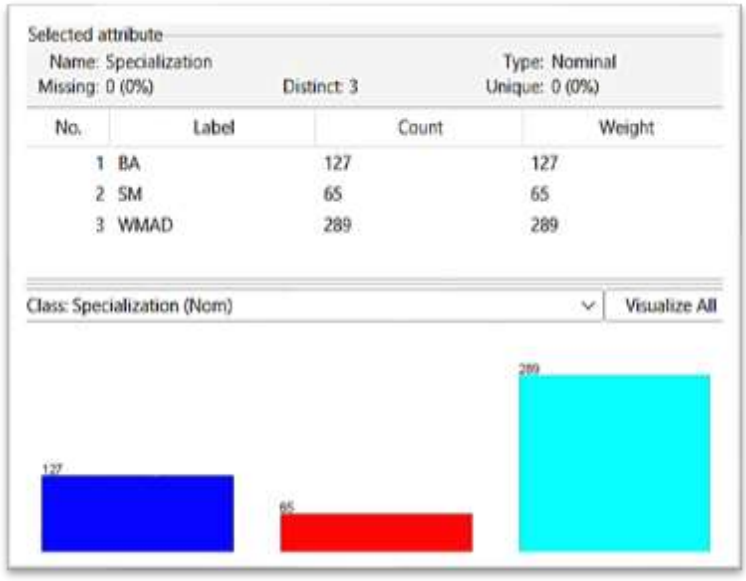

Figure 3 Original Distribution of Instances per Class

The WMAD specialization was over-represented, with 289 records, while BA and SM had only 127 and 65 records, respectively. This imbalance poses a risk of developing biased models that favor the majority class (WMAD) and underperform in recognizing the minority classes (BA and SM), which is a common issue in both binary and multi-class classification (He & Ma, 2019; Haixiang et al., 2020).

To address this, the researcher applied the Synthetic Minority Over-sampling Technique (SMOTE), a widely used method for handling imbalanced data. SMOTE generates synthetic samples for the minority classes, helping to balance the dataset and reduce bias toward the majority class. Although computationally expensive, SMOTE enhances the model's robustness by creating a more diverse and balanced dataset (Sole, 2023).

The dataset's SMOTE percentages for all three classes (WMAD, BA, and SM) are computed by determining how much to oversample each minority class, tracks BA and SM, to match the majority class size of 289 of the WMAD track. The SMOTE percentage is calculated using the formula:

$$\text{SMOTE Percentage} = \frac{(\text{target sample} - \text{current sample})}{\text{current sample}} \times 100$$

The SMOTE percentages for the BA and SM tracks were calculated to oversample these classes to match the majority class (WMAD), as shown in Figure 4.

Figure 4 Distribution of Instances per Class After Treatment

The BA track was oversampled by 127.56%, and the SM track by 344.62%. After applying SMOTE, each class in the dataset had approximately 289 samples, achieving a balanced dataset ready for further analysis.

Selecting the Best Machine Learning Algorithm

Machine learning was employed to build predictive models based on the balanced dataset, which consisted of 289 instances per class (WMAD, BA, and SM). The researcher divided the dataset into training (70%) and test sets (30%), with 606 instances in the training set and

261 in the test set.

The researcher tested three machine learning algorithms, J48, Random Forest, and Random Tree, to profile students likely to graduate on time based on the identified attributes. Each classifier's performance was evaluated based on the number of correctly classified instances, measured by percentage accuracy. The J48 algorithm achieved a 92.08% accuracy on the training set and 90.10% on the test set. On the other hand, Random Forest delivered 91.09% accuracy on the training set. While, Random Tree had 92.57% accuracy on the training set.

While both J48 and Random Tree yielded similar accuracy, the tree size generated by Random Tree was significantly larger (277 nodes) compared to J48 (146 nodes, pruned to 29). The J48 model's tree was pruned to 29 nodes, further simplifying the model without compromising its accuracy, which remained high at 85.5%, as shown in Figure 5.

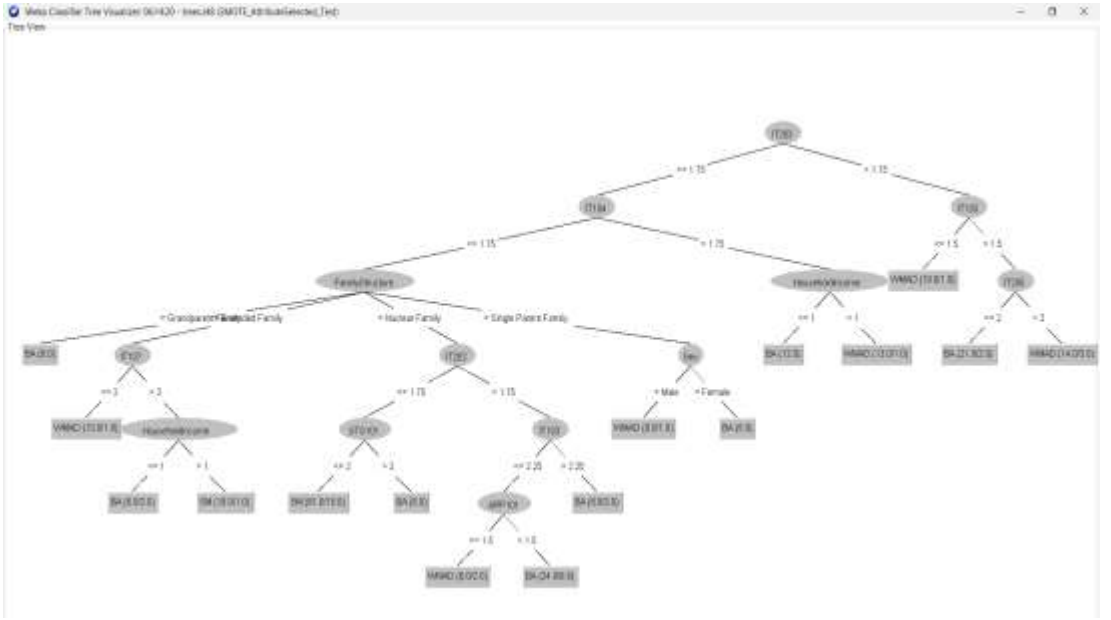

Figure 5 Resultant Tree Generated by the Pruned J48

This simplicity and its strong performance led the researcher to select J48 as the preferred model. The decision tree in Figure 5 delineates the relationships between significant attributes such as family structure, household income, sex, and academic performance in courses like IT203 and IT104, offering a transparent and interpretable model for predicting student specialization.

The decision tree model classifies data points based on multiple attributes, beginning with IT203, which splits the data into two branches based on whether the value is less than or equal to 1.75. Subsequent nodes, such as IT104, FamilyStructure, HouseholdIncome, and Sex, further divide the branches. These divisions lead to leaf nodes representing class labels like BA, WMAD, and SM, along with counts of data points in each class.

Model's Performance Validation

Validation testing is crucial for an initial assessment of the model's performance. Without it, a model's accuracy can be jeopardized due to bias, overfitting, or underfitting, making it less effective with new data. In this study, the researcher built the model on the train set and validated the model's robustness using the test set. This test set provides a realistic measure of the model's performance on unseen data, further ensuring its generalizability.

In addition, the detailed analysis of the accuracy by class and the confusion matrix, as shown in Figure 6, including the parameters such as true positive rates, false positive rates, precision, recall, F-measure, and ROC areas, is also essential. These metrics comprehensively present the model's strengths and weaknesses, confirming its robustness and accuracy across different classes.

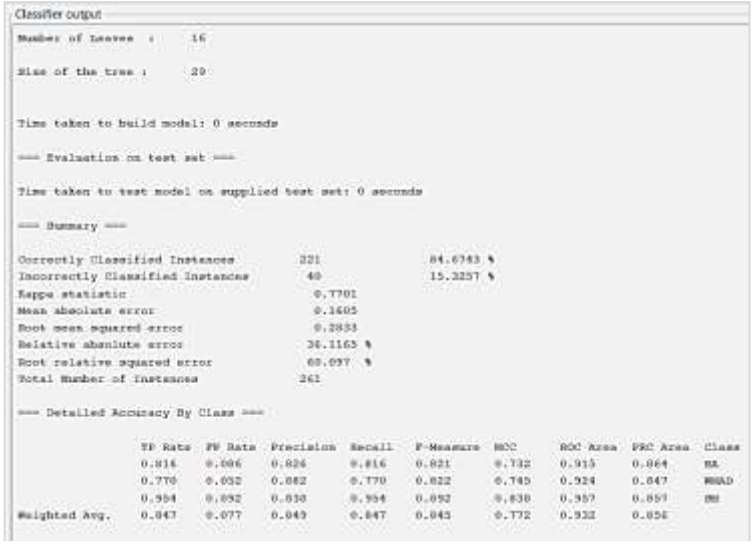

Figure 6 Evaluation Results of the Pruned J48 Classifier

Figure 6 displays the evaluation of the predictive model using the J48 pruned tree algorithm, revealing a well-performing and efficient model. The model comprises 16 leaves and 29 nodes, indicating a relatively simple decision tree structure. The model's overall accuracy is reflected in the 84.67% of correctly classified instances, while 15.33% were incorrectly classified. The Kappa statistic of 0.7701 suggests a substantial agreement between the predicted and actual classifications, surpassing chance. This means that a Kappa value of 0.7701 signifies that the model's predictions are consistently accurate and align closely with the correct classifications rather than occurring due to random chance. This level of agreement demonstrates that the model is reliable and effective in distinguishing between different classes. In addition, the mean absolute error (MAE) is 0.1605, and the root mean squared error (RMSE) is 0.2833, indicating that the model's predictions are pretty close to the actual values. The relative absolute and root relative squared errors are 36.12% and 60.10%, respectively, highlighting the model's efficiency in minimizing errors.

Figure 6 also shows the model's performance in providing prediction by class. The class-

wise performance metrics indicate high accuracy and robustness across different classes. For the BA class, the true positive (TP) rate is 0.816, and the false positive (FP) rate is 0.086, with a precision of 0.826 and a recall of 0.816. The F-measure stands at 0.821, with an MCC of 0.732; the ROC and PRC areas are 0.915 and 0.864, respectively. The WMAD class shows a TP rate of 0.770 and an FP rate of 0.052, with a precision of 0.882 and a recall of 0.770. The F-measure for WMAD is 0.822, with an MCC of 0.745; the ROC and PRC areas are 0.924 and 0.847. The SM class exhibits a high TP rate of 0.954 and an FP rate of 0.092, with a precision of 0.838 and a recall of 0.954. The F-measure is 0.892, with an MCC of 0.838, and the ROC and PRC areas are 0.957 and 0.857. The ROC and PRC areas for all classes are above 0.80, suggesting excellent model performance regarding true positive rates versus false positive rates and precision-recall.

The confusion matrix, as shown in Figure 7, further elaborates on the model's performance.

```
=== Confusion Matrix ===

  a  b  c   <-- classified as
 71  9  7 |  a = BA
 11 67  9 |  b = WMAD
  4  0 83 |  c = SM
```

Figure 7 The Model's Confusion Matrix

The confusion matrix, as shown in Figure 7, further elaborates on the model's performance. For the BA class, 71 instances were correctly classified, with nine misclassified as WMAD and seven as SM. In the WMAD class, 67 instances were correctly classified, with 11 misclassified as BA and nine as SM. The SM class had 83 correctly classified instances, with four misclassified as BA.

The developed model is integrated into a prototype prediction system, as shown in Figures 8 and 9, facilitating informed decision-making for personalized educational pathways and promoting timely graduations.

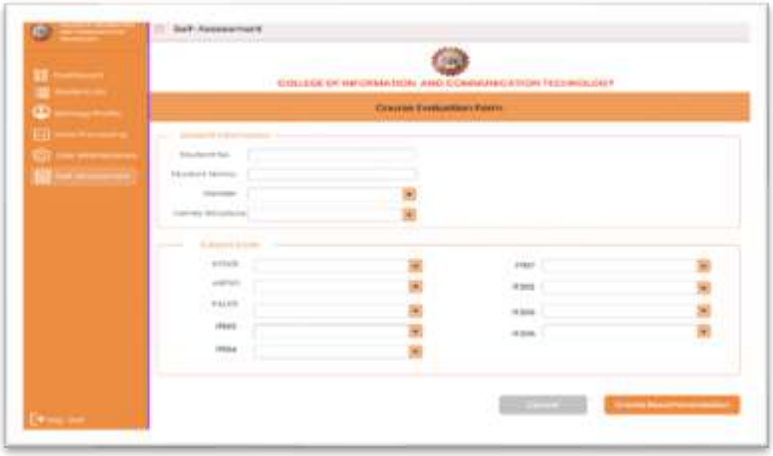

Figure 8 Self-Assessment Module

Figure 8 is part of a prediction system that enables users to perform individualized self-assessments. Users can input personal details, including student number, name, gender, and family structure, alongside subject-specific codes, allowing the system to evaluate their progress and provide tailored recommendations. The "Create Recommendation" button initiates a personalized analysis and recommendation for students.

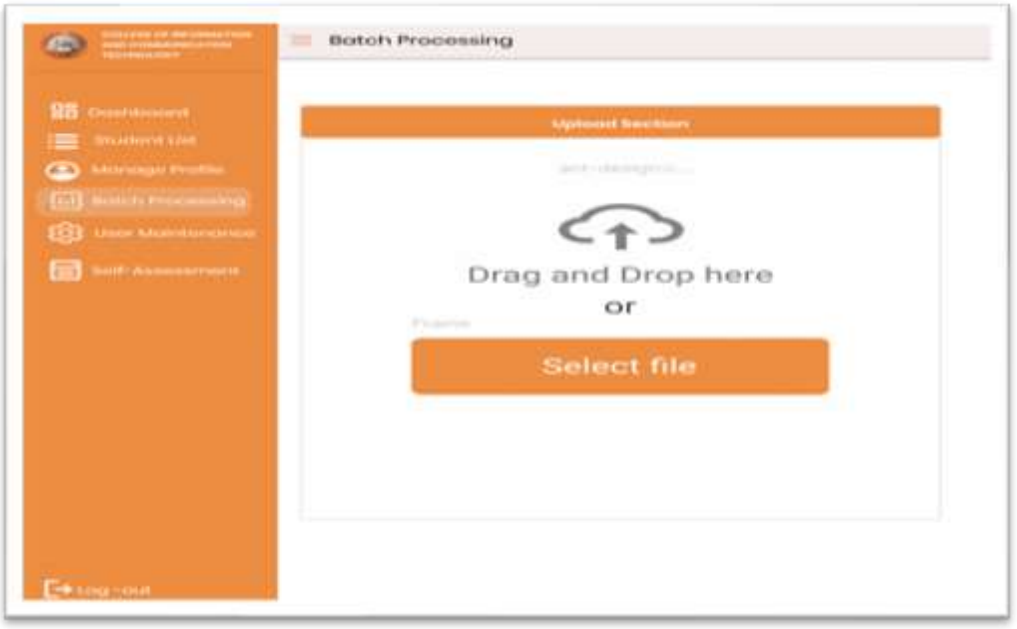

Figure 9 Batch Processing Module

The batch processing interface allows users to perform bulk data uploads, streamlining the system's assessment of multiple students simultaneously. Users can upload files for bulk processing through a drag-and-drop or file selection option, enabling efficient handling of large student datasets. This functionality supports administrators and educators in analyzing group performance and generating aggregate recommendations, enhancing decision-making for educational pathways on a larger scale.

4. Conclusion

The results of this study indicate that J48 is an effective and understandable machine learning model for predicting student specialization based on academic and demographic data. The pre-processing steps, including selecting significant attributes relevant to creating predictions; and applying SMOTE for balancing the dataset, significantly enhanced the model's accuracy and robustness. The pruned decision tree was also considered for a simpler model at the same time still demonstrated a solid predictive performance and with a high level of accuracy. This makes it a practical tool for educational institutions seeking to understand and support student success. Additionally, the web-based prototype system

developed in this study provides personalized and batch-processing options, helping students and advisors with track selection and improving overall educational guidance.

5. Recommendations

It is recommended that Bulacan State University integrate this predictive system into academic advising to support BSIT students' track selection, facilitating data-driven guidance. Regular updates to the model's training data are also recommended to maintain accuracy amid evolving academic and industry landscapes. Future research could incorporate additional predictors, such as extracurricular activities and internships, to improve the model's precision and applicability in guiding students toward successful educational and career outcomes.

References

1. Ali, J., Rehman, M., & Khan, F. A. (2021). Predictive analytics in higher education: An approach to enhancing student performance and retention. Journal of Educational Data Science, 8(3), 123-138.
2. Beck, K. (1999). Extreme Programming Explained: Embrace Change. Addison-Wesley. Retrieved from https://www.extremeprogramming.org
3. Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. Journal of Artificial Intelligence Research, 16, 321-357. Retrieved from https://www.jair.org/index.php/jair/article/view/10302
4. Commission on Higher Education (CHED). (2020). Higher Education Act and curriculum enhancement guidelines. CHED.
5. Department of Information and Communications Technology (DICT). (2022). Digital workforce report. DICT.
6. Haixiang, G., Yijing, L., Shang, J., Mingyun, G., Yuanyue, H., & Bing, G. (2020). Learning from class-imbalanced data: Review of methods and applications. IEEE Access, 5, 19651-19670. https://doi.org/10.1109/ACCESS.2020.3005677
7. Han, J., Pei, J., & Kamber, M. (2011). Data Mining: Concepts and Techniques (3rd ed.). Morgan Kaufmann. Retrieved from https://www.sciencedirect.com/book/9780123814791/data-mining
8. Hariri, N., Novick, D., & Goh, W. W. (2020). Machine learning models for student performance prediction: A systematic review. Educational Data Science, 5(2), 101-115.
9. He, H., & Ma, Y. (2019). Imbalanced learning: Foundations, algorithms, and applications. Wiley-IEEE Press.
10. Kotsiantis, S. B., Kanellopoulos, D., & Pintelas, P. E. (2022). Handling imbalanced datasets: A review. Artificial Intelligence Review, 30(1), 59-86. https://doi.org/10.1007/s10462-021-9232-8
11. Kumar, A., & Chadha, R. (2020). Educational data mining and learning analytics for student success prediction: A review. International Journal of Advanced Computer Science and Applications, 11(4), 145-154. https://doi.org/10.14569/IJACSA.2020.0110419
12. Li, F., Zhang, J., & Wang, P. (2021). Big data and machine learning in education: Enhancing student performance prediction. Computers & Education, 163, 104098. https://doi.org/10.1016/j.compedu.2020.104098
13. Padillo, M. G. (2023). Discretization techniques for socioeconomic data: A case study in the

Philippine educational context. International Journal of Data Science and Analytics, 5(2), 112-125. https://doi.org/10.1007/s41060-023-0020-3

14. Philippine Software Industry Association (PSIA). (2020). Industry skills assessment. PSIA.
15. Salman, M., Asghar, M., & Rafique, A. (2019). Predictive modeling for academic performance and course selection using machine learning techniques. IEEE Access, 7, 182082-182091. https://doi.org/10.1109/ACCESS.2019.2960368
16. Sole, F. (2023). Oversampling and its impact on imbalanced datasets: A detailed examination of SMOTE. Journal of Machine Learning Applications, 12(4), 110-122. https://doi.org/10.1016/j.jmla.2023.05.003
17. Sun, H., Xu, T., Wang, C., & Liu, X. (2019). Machine learning models for predicting student graduation: A case study in China. IEEE Access, 7, 160824-160832. Retrieved from https://ieeexplore.ieee.org/document/8875425
18. Tiwari, S., Shrivastava, A., & Mittal, A. (2019). Predictive analytics in higher education: Improving the learning experience of students. International Journal of Educational Technology in Higher Education, 16(1), 1-12. https://doi.org/10.1186/s41239-019-0170-8
19. Wickham, H., & Grolemund, G. (2021). R for data science: Import, tidy, transform, visualize, and model data. O'Reilly Media.
20. World Bank. (2021). Skills and jobs mismatch in the Philippines: Trends, policy considerations, and international experience. World Bank.
21. Zhao, X., Yang, Z., Lei, L., & Ma, X. (2020). Exploring student performance prediction in blended learning environments with learning behaviors and time management skills. IEEE Access, p. 8, 146619–146630. Retrieved from https://ieeexplore.ieee.org/document/9153617
22. Zhu, X., Quinlan, J. R., & Zhang, C. (2019). Data preprocessing and machine learning models: A critical survey of techniques. Journal of Machine Learning Research, 17(2), 112-134.