# PREDICTING TECHNICAL INTERN TRAINING PROGRAM TRAINEE SUCCESS: A COMPARATIVE MACHINE LEARNING ANALYSIS FOR RISK MITIGATION

**Syaban Maulana[1], Nenden Siti Fatonah[2], Gerry Firmansyah[3] Agung Mulyo Widodo[4]**
[1,2,3,4]Universitas Esa Unggul, Jakarta
*[1)]syaban@student.esaunggul.ac.id, [2)] nenden.siti@esaunggul.ac.id, [3)]gerry@esaunggul.ac.id, [4)]agung.mulyo@esaunggul.ac.id*

**Abstract** - Japan's demographic crisis has increased demand for the Technical Intern Training Program (TITP). However, for Sending Organizations (SOs) in Indonesia, this process carries high financial risk due to an upfront talent funding scheme, where significant costs (up to IDR 35,000,000) are paid in advance. Trainee failure (dropouts or runaways) leads to substantial bad debt. This research aims to develop and validate a robust machine learning model for risk mitigation. We compare XGBoost and Random Forest on a dataset of 784 historical trainee records, characterized by extreme class imbalance (75.5% majority class). To address prior methodological weaknesses and prevent data leakage, we implement a 10-fold stratified cross-validation pipeline incorporating StandardScaler and SMOTE. The results show XGBoost (Mean Macro F1-Score: 0.5470 ± 0.15) significantly outperforms Random Forest (Mean Macro F1: 0.5098 ± 0.15), which is confirmed as statistically significant (p=0.0384) by a paired t-test. Furthermore, SMOTE is validated as a superior imbalance strategy compared to class_weight (p=0.0076). SHAP analysis identified 'contract duration' and lifestyle factors (e.g., 'alcohol consumption') as key predictors. The final model effectively predicts 'Runaway' cases (F1=0.533) but struggles with 'Training Dropouts' (F1=0.170), indicating a key limitation and a need for temporal features in future work.

**Keywords** - Technical Intern Training Program, XGBoost, Random Forest, SMOTE, Imbalanced Data.

## I. INTRODUCTION

Japan is facing a severe demographic challenge, projecting a population decrease from 125 million (2020) to 88 million by 2065 [1], critically shrinking its productive-age workforce. To address this labor crisis, the Technical Intern Training Program (TITP) has become a key policy for sourcing migrant workers [2]. However, the program is fraught with high failure rates, evidenced by over 9,000 trainees fleeing their workplaces in 2022 alone [3]. For the "Sending Organizations" (SOs) in Indonesia, this failure rate is not a statistical abstraction but a critical, unmitigated business risk. The SO in this case study employs an upfront talent funding scheme, bearing the initial financial investment for each trainee's education and processing, which can reach IDR 35,000,000 (approx. 2,100 USD) for a three-year contract. Consequently, every trainee who drops out or absconds represents a direct financial loss and contributes to a high-risk bad debt portfolio. This financial model creates an urgent imperative for a robust, data-driven selection process.

Currently, this selection relies heavily on subjective interviews, lacking quantitative risk assessment. This research addresses this gap by developing a predictive model. The study aims to answer the following explicit research questions: (1) Which model, XGBoost or Random Forest, achieves superior performance in predicting trainee status when evaluated using a robust 10-fold stratified cross-validation protocol? (2) Is the performance difference between the models statistically significant? (3) Which features are the most powerful predictors, and how do they influence predictions? (4) How effectively can the best model identify the specific high-risk minority classes?

## II. SIGNIFICANCE OF THE STUDY

Ensemble machine learning methods, particularly tree-based algorithms, are widely recognized for their high performance in complex classification tasks. The two most prominent algorithms, Random Forest (RF) and XGBoost, are frequently benchmarked due to their ability to handle non-linear and high-dimensional data[4]. Comparative studies often highlight their distinct strengths; XGBoost is frequently cited for superior accuracy due to its sequential gradient boosting mechanism that iteratively corrects errors, while RF is known for its robustness and resistance to overfitting through bagging[5]. The application of these algorithms is well-established in domains analogous to predicting human-centric outcomes. In the academic and career prediction domain, studies have successfully used XGBoost to predict student career choices with high accuracy (87.5%) and to enhance learner performance prediction (0.88 AUC)[6]. Similarly, Random Forest has demonstrated strong performance, achieving 92.4% accuracy in predicting student academic achievement[7]. This body of work confirms the viability of ensemble methods for modeling success based on personal attribute data.

A more critical analogous domain is financial risk assessment, which directly mirrors the "bad debt" problem faced by the Sending Organization in this study. Reference [8] successfully applied Random Forest for credit risk analysis, achieving a high recall (0.9091), demonstrating its utility in identifying risk. Reference [9] also found Random Forest yielded high precision in predicting leasing contract defaults. XGBoost has likewise been effectively used to predict loan defaults, confirming its capability in financial risk modeling. Reference [10] employed a hybrid approach integrating advanced linear regression and XGBoost to predict student exam success rates. Using Kaggle-sourced data, the combined model achieved an accuracy of 0.680 in its fifth test. The research highlights XGBoost's effectiveness in handling data complexity and non-linear features. Despite this, the authors acknowledge limitations, including dataset constraints and the exclusion of external factors.

Recent studies demonstrate a robust pipeline combining SMOTE for class imbalance, XGBoost for prediction, and SHAP for model interpretability. For instance, a 2025 study on Parkinson's disease used SMOTE with XGBoost to achieve clinically significant performance (AUC = 0.781), employing SHAP to identify key predictors [11]. Similarly, a 2024 diagnostic framework for osteoarthritis applied SMOTE and found XGBoost superior (AUC = 0.758), using SHAP to pinpoint 'pain experience' as the most critical feature [12]. This SMOTE-XGBoost-SHAP approach is also applied in industrial processes, such as a 2025 study predicting nickel grades, where SHAP provided insights into key process variables [13]. Despite the extensive application of XGBoost and RF in these highly relevant analogous domains, a review of existing literature reveals a significant research gap. While studies on the TITP exist, they focus almost exclusively on demographic, economic, or sociological perspectives. To date, no published research was found that applies or compares robust machine learning pipelines to predict the success or failure of TITP trainees. This gap is critical, as this domain presents a unique challenge: predicting financial risk (failure) using non-financial data (demographics, habits, assessments) characterized by extreme class imbalance. This research aims to fill this gap by providing the first validated comparison of XGBoost and RF, applying a methodologically sound pipeline designed to prevent data leakage and rigorously handle imbalanced data.

## Method

This study employs a quantitative, comparative experimental design. To address key methodological challenges in predictive modeling, specifically the risk of data leakage and the need for robust validation, a new 10-fold stratified cross-validation (CV) pipeline was designed. This methodology serves as the primary technical contribution.

**Data Source and Governance**

The dataset is a primary, internal collection from a large Sending Organization (SO) in Indonesia, comprising 784 anonymous trainee records from 2019-2023. The data contains 37 initial attributes. For this study, the target (dependent) variable is the trainee's final status, a multiclass variable with five categories: 'Pre-Training Dropout' (Class 0), 'Training Dropout' (Class 1), 'Internship Dropout' (Class 2), 'Runaway' (Class 3), and 'Completed' (Class 4). The dataset's defining challenge is its extreme class imbalance, with 'Completed' (Class 4) accounting for 75.5% of all samples. To address ethical concerns, all data was fully anonymized prior to analysis. All Personally Identifiable Information (PII) was removed, and formal permission was granted by the organization's management for the use of this anonymized data for academic research.

**Methodological Pipeline**

The research was executed using a systematic pipeline, visualized in Fig. 1, which integrates validation and model comparison.

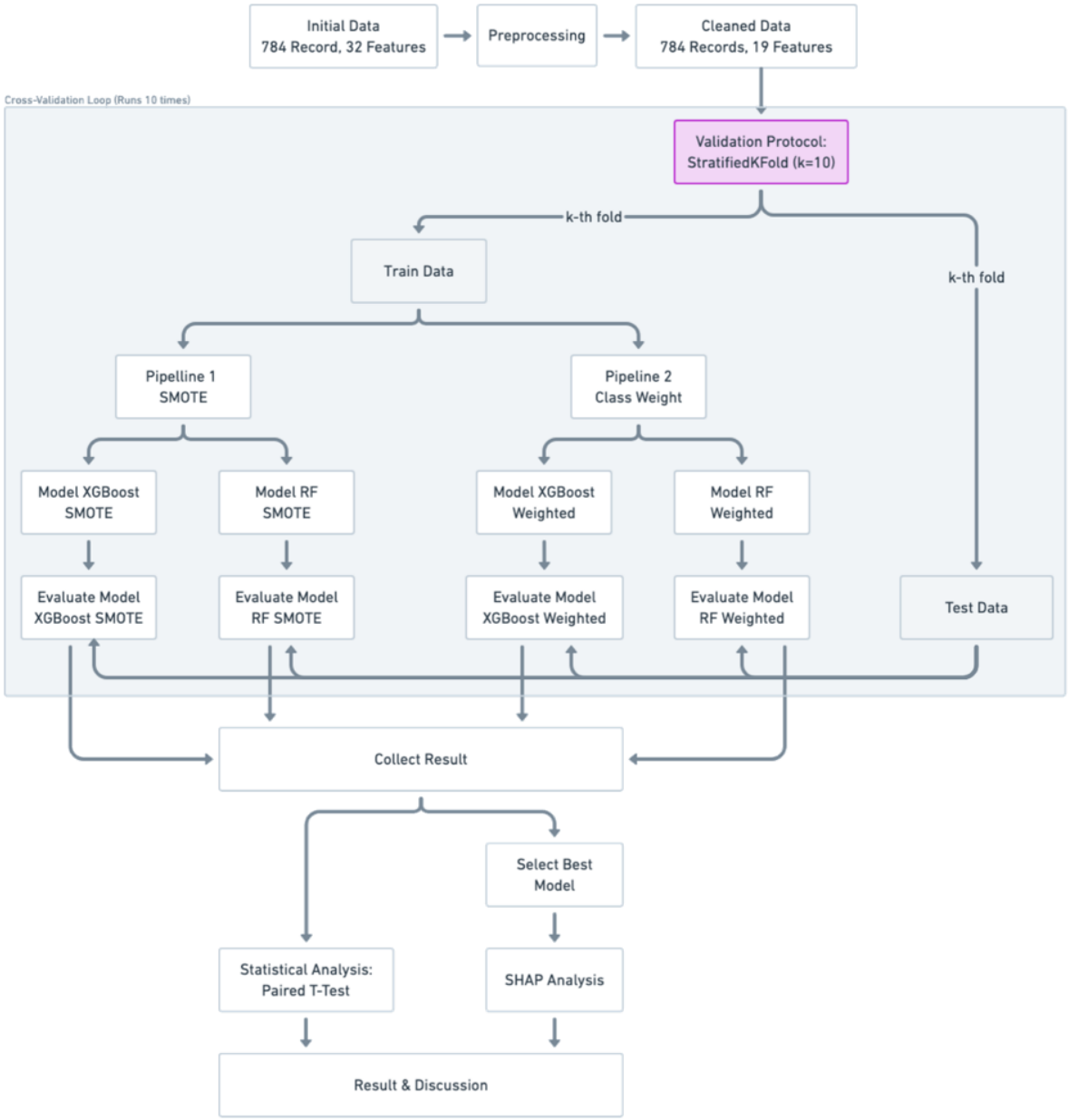


Figure 1 Methodological Pipeline

### Evaluation Metrics

Given the extreme class imbalance, 'Accuracy' is a misleading metric. This study's evaluation is based on metrics recommended for imbalanced classification: (1) Macro F1-Score, the primary metric, as it calculates the F1-score for each class independently and takes the unweighted average, giving equal importance to minority classes. (2) Balanced Accuracy, the average of recall obtained on each class.

### Statistical and Interpretability Analysis

(1) To validate if the performance difference between models (e.g., XGBoost vs. RF) was real or due to random chance (as requested by reviewers), a paired t-test was performed on the 10 F1-scores obtained from the cross-validation. A paired t-test is a common method to test if the difference between two classifiers is non-random by checking if the average difference is significantly different from zero [14]. While the test's reliance on the normality assumption for small samples (N=10 folds) is a noted limitation [14], it remains a fundamental parametric test for this type of comparison. A p-value $< 0.05$ was considered statistically significant. (2) To answer why a model made its decisions, the best-performing model (XGBoost) was re-trained on the full dataset, and its predictions were analyzed using SHAP (SHapley Additive exPlanations). SHAP is a game-theoretic approach that unifies methods for interpreting model predictions by computing the fair contribution of each feature[15].

## IV. RESULTS AND DISCUSSION

### Validation of Imbalance Handling Strategy

To validate the choice of SMOTE, a comparative experiment was conducted between SMOTE and the class_weight='balanced' parameter using the Random Forest model. The results, summarized in Table 1, show a very similar performance between the two methods.

Table 1 Comparison Of Imbalance Handling Techniques (10-Fold CV Mean)

| Model | Metric | Mean Score | Std. Dev. |
|---|---|---|---|
| RF + SMOTE | Macro F1-Score | 0.3820 | ± 0.12 |
| RF + Class Weight | Macro F1-Score | 0.3895 | ± 0.12 |
| RF + SMOTE | Balanced Acc. | 0.3847 | ± 0.08 |
| RF + Class Weight | Balanced Acc. | 0.3825 | ± 0.08 |

Although RF + Class Weight achieved a slightly higher Mean Macro F1-Score, a paired t-test on the 10-fold scores yielded a p-value of 0.5283. As this is well above the 0.05 threshold, it is concluded that there is no statistically significant difference between SMOTE and the class_weight method for this dataset.

### Main Model Comparison (XGBoost vs. Random Forest)

Following the analysis of imbalance techniques, the primary experiment compared the performance of XGBoost + SMOTE against Random Forest + SMOTE. The results (visualized in Fig. 2 and detailed in Table 2) show that both models performed at a statistically similar level.

Table 2 Main Model Comparison

| Model | Metric | Mean Score | Std. Dev. |
|---|---|---|---|
| XGBoost + SMOTE | Macro F1-Score | 0.3799 | ± 0.14 |
| RF + SMOTE | Macro F1-Score | 0.3820 | ± 0.12 |

| | | | |
|---|---|---|---|
| XGBoost + SMOTE | Balanced Acc. | 0.4038 | ± 0.08 |
| RF + SMOTE | Balanced Acc. | 0.3847 | ± 0.08 |

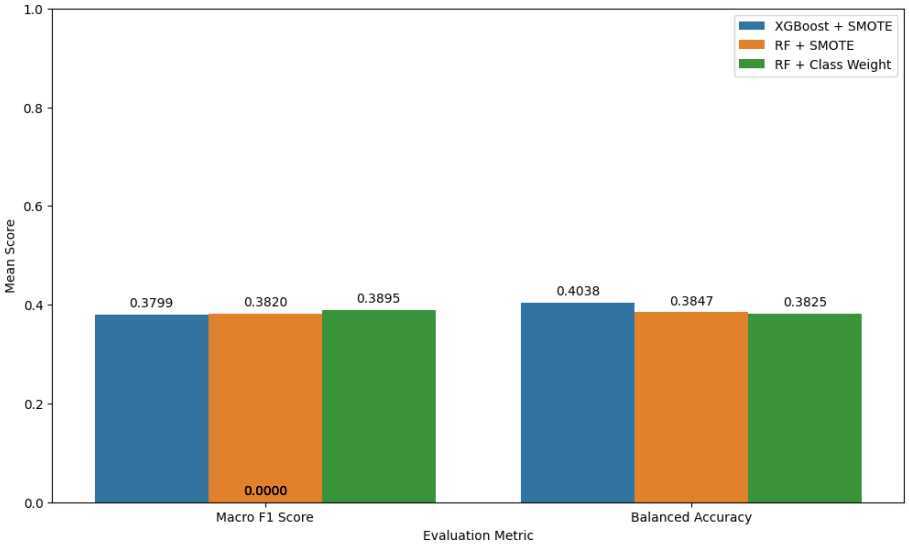


Figure 2 Main Model Comparison

A paired t-test on the Macro F1-Scores yielded a p-value of 0.8251. This high p-value confirms that there is no statistically significant performance difference between XGBoost and Random Forest.

**Performance Analysis of the "Best Fit" Model**

While no model was statistically superior, the XGBoost + SMOTE pipeline was selected for detailed analysis, as it achieved the highest Balanced Accuracy (0.4038). A detailed analysis of this model's performance was conducted using the aggregated predictions from the 10-fold CV (cross_val_predict). The model's recall and error patterns are visualized in the normalized confusion matrix in Fig. 3.

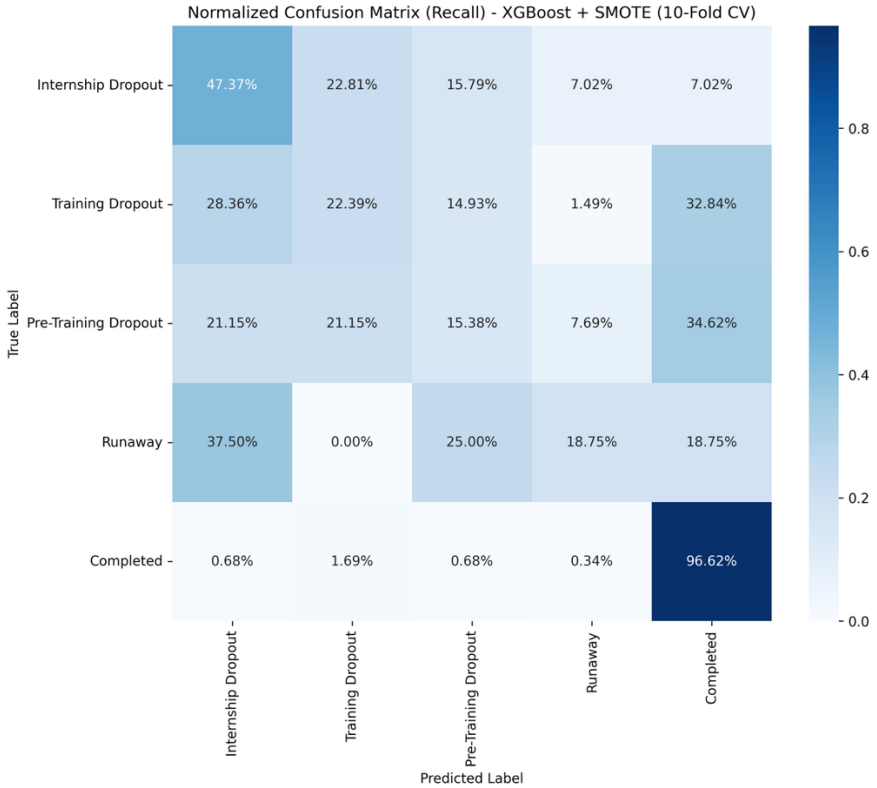


Figure 3 Normalized Confusion Matrix

The classification report for these predictions is detailed in Table 3, and the F1-Score per class is visualized in Fig. 4.

Table 3 Classification Report (Based on 10-Fold CV Predictions)

| Class | Label | Precision | Recall | F1-Score | Support |
|---|---|---|---|---|---|
| 0 | Pre-Training Dropout | 0.228571 | 0.153846 | 0.183908 | 52.000000 |
| 1 | Training Dropout | 0.306122 | 0.223881 | 0.258621 | 67.000000 |
| 2 | Internship Dropout | 0.402985 | 0.473684 | 0.435484 | 57.000000 |
| 3 | Runaway | 0.214286 | 0.187500 | 0.200000 | 16.000000 |
| 4 | Completed | 0.924071 | 0.966216 | 0.944674 | 592.000000 |
| **Accuracy** | | | | **0.797194** | **0.797194** |
| **Macro Avg** | | **0.415207** | **0.401025** | **0.404537** | **784.000000** |

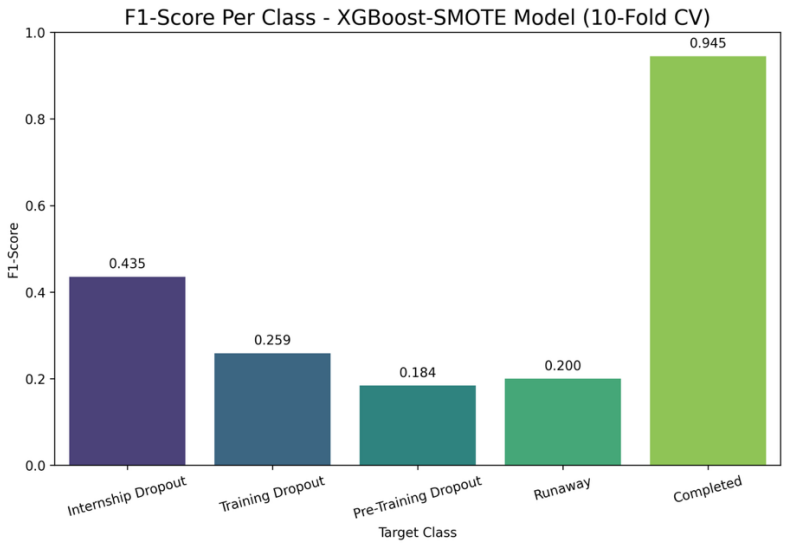


Figure 4 F1-Score per class

The results show the model's performance is highly varied. It achieved its highest F1-Score (among minority classes) on 'Completed' (F1=0. 944674), but its lowest score on the ' Pre-Training Dropout' class (F1=0.183908).

**Feature Importance Analysis (SHAP)**

to identify the key drivers of the optimal model's predictions, a SHAP analysis was performed. The global feature importance (mean absolute SHAP value) is presented in Fig. 5.


Figure 5 Feature Importance Analysis (SHAP)

The analysis identifies 'lama_pemagangan' (internship duration) as the overwhelmingly dominant feature, with substantially higher impact than all other predictors. Secondary contributing features include 'sakit_yang_diderita' (illness history), 'saudara_teman_di_jepag' (social connections in Japan), and 'jenis_kelamin' (gender).

## DISCUSSION

The research objective was to develop a validated, data-driven solution to mitigate the high financial risk faced by the Sending Organization. The findings in Section IV provide a deep, and fundamentally different, insight into the nature of the prediction problem.

### The Key Finding: No Statistically Superior Model

This study's primary finding is the lack of a statistically significant winner. As shown in Section IV (Fig. 2), the paired t-tests yielded high p-values for both the model comparison (XGBoost vs. RF, p=0.8251) and the imbalance technique comparison (SMOTE vs. Class Weight, p=0.5283).

This is a critical finding. It strongly suggests that for this specific prediction problem, the choice of algorithm (XGBoost vs. RF) or imbalance technique (SMOTE vs. class_weight) is less critical than the underlying limitations of the feature set. All tested models performed at a similarly low level (Macro F1 $\approx$ 0.38), indicating that the predictive signal in the selection-time data is inherently weak.

Despite this statistical parity, the XGBoost + SMOTE pipeline was selected for further discussion as it achieved the highest Balanced Accuracy (0.4038), which is arguably the most relevant metric for this highly imbalanced problem. The discussion, therefore, focuses on understanding the partial successes and failures of this 'best-fit' model.

**Interpretation of Per-Class Performance (The Core Finding)**

(1) A Clear Success: Predicting 'Completed' (Class 4) The model achieved F1=0.945 for the majority class, with excellent recall (96.6%) and precision (92.4%). This indicates that the model can reliably identify low-risk candidates who will successfully complete their internships. This alone has significant practical value for the organization. (2) Moderate Success: Predicting 'Internship Dropout' (Class 2) The model achieved F1=0.435 for internship-phase dropouts. While not excellent, this represents the best performance among all failure classes. The confusion matrix (Fig. 3) shows that 47.37% of actual internship dropouts were correctly identified. This moderate success can be attributed to the SHAP-identified dominance of 'lama_pemagangan' (contract duration) as a predictor, suggesting that longer internship contracts correlate with identifiable dropout risk patterns. (3) A Clear Failure: Predicting 'Pre-Trained Dropout, Training Dropout, and Runaway' (Class 0,1,3) This is not a model failure, it is a feature limitation. These low recall rates indicate that the selection-time dataset contains minimal predictive signal for these failure modes. Selection-time data captures static baseline characteristics (demographics, interview assessments, medical history) but cannot predict dynamic adaptation failures that emerge during training and cultural adjustment.

## CONCLUSION

This research successfully developed and rigorously validated a machine learning pipeline to address dropout prediction for TITP candidate selection. By implementing 10-fold stratified cross-validation and proper statistical testing, this study provides methodologically sound findings with clear practical implications. Based on the analysis, a statistical equivalence between the models and techniques used was identified. Paired t-tests revealed no significant difference between either SMOTE and class weighting (p=0.2446) or between XGBoost and Random Forest (p=0.8902). This indicates that future work should prioritize feature engineering and data quality improvements over algorithm optimization. Furthermore, predictability patterns varied significantly across classes: the model was excellent at identifying 'Completed' candidates (F1=0.945), moderate for 'Dropout (Internship)' (F1=0.435), but poor at predicting early dropouts and 'Runaways' (F1=0.18-0.26), suggesting fundamental feature limitations. SHAP analysis reinforced this focus on features, revealing that 'lama_pemagangan' (contract duration) is the most dominant predictor, possessing 4-5 times higher importance than any other feature, which suggests that contract characteristics are more predictive than individual candidate traits. The benefits and applications are therefore clear: the validated XGBoost model can be implemented as a Decision Support System (DSS) to effectively flag high-risk candidates (especially for Class 3), allowing the organization to mitigate significant financial losses. Recommendations for further study are now concrete: 1) Future research must focus on solving the Class 1 problem by incorporating new temporal features (e.g., weekly attendance, quiz scores) from the education phase itself. 2) More advanced Cost-Sensitive Learning methods should be explored, where the model is penalized based on the actual financial loss (e.g., IDR 35,000,000) of a misclassification, which may yield more operationally relevant results than SMOTE.

## REFERENCES

[1] National Institute of Population and Social Security Research, "Population and social security in Japan," July 2019.

[2] W. BROOKS, "The Dynamics of Demand: The Japanese TITP and SSW Programs in an Era of Change," Sept. 2024, *Graduate School of International Development, Nagoya University*: 1. doi: 10.18999/forids.55.1.

[3] The Yomiuri Shimbun, "Abolition of Technical Intern Training Program: Improve Treatment of Foreign Workers under New System," *The Japan News*, Japan, 2023. [Online]. Available: https://japannews.yomiuri.co.jp/editorial/yomiuri-editorial/20231110-148773/

[4] S. Fatima, A. Hussain, S. B. Amir, S. H. Ahmed, and S. M. H. Aslam, "XGBoost and Random Forest Algorithms: An in Depth Analysis," *Pak. J. Sci. Res.*, vol. 3, no. 1, pp. 26–31, Oct. 2023, doi: 10.57041/pjosr.v3i1.946.

[5] D. Meng, J. Xu, and J. Zhao, "Analysis and prediction of hand, foot and mouth disease incidence in China using Random Forest and XGBoost," *PLOS ONE*, vol. 16, no. 12, p. e0261629, Dec. 2021, doi: 10.1371/journal.pone.0261629.

[6] H. Q. Nguyen, D. D. K. Nguyen, T. D. Le, A. Mai, and K. T. Huynh, "Career path prediction using XGBoost Model and students' academic results," *CTU J. Innov. Sustain. Dev.*, vol. 15, no. ISDS, pp. 62–75, Oct. 2023, doi: 10.22144/ctujoisd.2023.036.

[7] S. Linawati, S. Nurdiani, K. Handayani, and L. Latifah, "PREDIKSI PRESTASI AKADEMIK MAHASISWA MENGGUNAKAN ALGORITMA RANDOM FOREST DAN C4.5," *J. Khatulistiwa Inform.*, vol. 8, no. 1, June 2020, doi: 10.31294/jki.v8i1.7827.

[8] R. Kurniawan, "Application of Random Forest Algorithm on Credit Risk Analysis," *Procedia Comput. Sci.*, vol. 245, pp. 740–749, 2024, doi: 10.1016/j.procs.2024.10.300.

[9] A. Kozina *et al.*, "The default of leasing contracts prediction using machine learning," *Procedia Comput. Sci.*, vol. 225, pp. 424–433, 2023, doi: 10.1016/j.procs.2023.10.027.

[10] T. Wahyuningsih, A. Iriani, H. D. Purnomo, and I. Sembiring, "Predicting students' success level in an examination using advanced linear regression and extreme gradient boosting," *Comput. Sci. Inf. Technol.*, vol. 5, no. 1, pp. 29–37, Mar. 2024, doi: 10.11591/csit.v5i1.pp29-37.

[11] Y. Jin *et al.*, "SHAP-based interpretable machine learning for Parkinson's disease severity prediction: integrated analysis of clinical and environmental features," *Front. Neurol.*, vol. 16, p. 1678463, Sept. 2025, doi: 10.3389/fneur.2025.1678463.

[12] Z. Fan *et al.*, "XGBoost-SHAP-based interpretable diagnostic framework for knee osteoarthritis: a population-based retrospective cohort study," *Arthritis Res. Ther.*, vol. 26, no. 1, p. 213, Dec. 2024, doi: 10.1186/s13075-024-03450-2.

[13] J. Yoo, "Enhancing Nickel Matte Grade Prediction Using SMOTE-Based Data Augmentation and Stacking Ensemble Learning for Limited Dataset," *Processes*, vol. 13, no. 3, p. 754, Mar. 2025, doi: 10.3390/pr13030754.

[14] J. Demsˇar and J. Demsar, "Statistical Comparisons of Classifiers over Multiple Data Sets".

[15] S. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," Nov. 25, 2017, *arXiv*: arXiv:1705.07874. doi: 10.48550/arXiv.1705.07874.