Published online on the page : https://journal.makwafoundation.org/index.php/intellect

**I n t e l l e c t :**
**Indonesian Journal of Learning and Technological Innovation**

| ISSN (*Online*) 2962-9233 |

# Predicting Employment Status 6 Months After Graduation with *Machine Learning Learning* : A Comparative Study of 3,945 Indonesian Graduates

*Ihdi Syahputra Ritonga[1*], Muhammad Rahfiqa Zainal[2], Ahmad Zaki[3]*

[1,2,3]*Sjech M. Djamil Djambek State Islamic University, Bukittinggi, Bukittinggi,* Indonesia

**Article Information**

*Article History:*
*Submit:* august 6, 2025
Revision: August 13, 2025
Accepted: December 12, 2025
Published: December 14, 2025

**Keywords**

Machine Learning, Random Forest, XGBoost, Logistic Regression, Employability, SHAP Analysis

**Correspondence**

Email:

ihdisyahpuraritonga@uinbukittinggi.ac.id *

**A B S T R A C T**

The high unemployment rate of undergraduate graduates in Indonesia, reaching 11.4% in the first six months after graduation, indicates the need for an early prediction system to identify factors that influence student employability. This study aims to analyze and compare the performance of three machine learning algorithms (Random Forest, Logistic Regression, and XGBoost) to predict employment status 6 months after graduation based on academic and socioeconomic data. The dataset consists of 3,945 graduates from universities in Padangsidimpuan with variables of study program, study duration, GPA, gender, and parental income. The operational target is employment status 6 months after graduation (binary: employed = 1, not yet = 0) with the proportion of employed classes: 48.2 %, not yet: 51.8%. Evaluation uses stratified 5- fold cross-validation with accuracy metrics, balanced accuracy, F1- macro, ROC-AUC, and PR-AUC. Model interpretability is analyzed using permutation importance and SHAP values. Random Forest achieved the best performance with F1- macro 0.524±0.015, ROC-AUC 0.567±0.012, followed by Logistic Regression (F1- macro : 0.511±0.018) and XGBoost (F1- macro : 0.506±0.020). The majority baseline achieved an accuracy of 51.8 %. Permutation importance analysis identified GPA as the most influential factor (importance : 0.082), followed by parental income (0.067) and duration of study (0.041). The machine learning model provided a moderate improvement compared to the majority baseline. GPA and socioeconomic factors were shown to significantly influence graduates' employment status. These findings can support the development of an early warning system for data-based student mentoring.

***Abstrak***

*Tingginya tingkat pengangguran lulusan sarjana di Indonesia mencapai 11.4% dalam enam bulan pertama pasca kelulusan menunjukkan perlunya sistem prediksi dini untuk mengidentifikasi faktor-faktor yang mempengaruhi employability mahasiswa. Penelitian ini bertujuan menganalisis dan membandingkan performa tiga algoritma machine learning (Random Forest, Logistic Regression, dan XGBoost) untuk memprediksi status kerja 6 bulan pascawisuda berdasarkan data akademik dan sosial ekonomi. Dataset terdiri dari 3.945 data lulusan dari universitas di Padangsidimpuan dengan variabel program studi, durasi studi, IPK, jenis kelamin, dan penghasilan orang tua. Target operasional adalah status kerja 6 bulan pascawisuda (biner: bekerja=1, belum=0) dengan proporsi kelas bekerja:48.2%, belum:51.8%. Evaluasi menggunakan stratified 5-fold cross-validation dengan metrik akurasi, balanced accuracy, F1-macro, ROC-AUC, dan PR-AUC. Interpretabilitas model dianalisis menggunakan permutation importance dan SHAP values. Random Forest mencapai performa terbaik dengan F1-macro 0.524±0.015, ROC-AUC 0.567±0,012, diikuti Logistic Regression (F1-macro: 0.511±0,018) dan XGBoost (F1-macro: 0.506±0.020). Baseline mayoritas mencapai akurasi 51,8%. Analisis permutation importance mengidentifikasi IPK sebagai faktor paling berpengaruh (importance: 0.082), diikuti penghasilan orang tua (0.067) dan durasi studi (0.041). Model machine learning memberikan peningkatan moderat dibanding baseline mayoritas. IPK dan faktor sosial ekonomi terbukti berpengaruh signifikan terhadap status kerja lulusan. Temuan ini dapat mendukung pengembangan sistem early warning untuk pendampingan mahasiswa berbasis data.*

## 1. Introduction

The phenomenon of unemployment among university graduates in Indonesia has become a serious concern in recent years. According to data from the Central Statistics Agency (2023), approximately 11.4% of undergraduate graduates failed to find employment within the first six months after graduation, indicating a gap between graduate qualifications and job market needs [1]. This [2]situation indicates the need for an early prediction system that can identify factors influencing student employability to support more effective mentoring [3] programs.[4]

machine learning technology have opened up opportunities to analyze complex patterns in students' academic and socioeconomic data to predict post-graduation outcomes. Previous research shows that machine learning algorithms can identify non-linear relationships between academic variables such as GPA, duration of study, field of study with employability levels. [5], [6], [7], [8]. However, most existing research focuses on the context of higher education in developed countries with different socio-economic characteristics [9].

In the Indonesian context, research that integrates academic data and socio-economic factors to predict graduate employment status is still limited, especially those using a comparative approach between algorithms with an adequate [10]dataset scale. Identified research gaps include: limited in-depth comparative analysis between Random Forest, Logistic Regression, and XGBoost algorithms for employment status classification based on Indonesian data [11], [12]; and limited research analyzing the relative contribution of academic versus socioeconomic factors using interpretability techniques. modern, and the absence of a robust evaluation framework using stratified cross-validation to handle class imbalance [13].

This study aims to analyze and compare the performance of three machine learning algorithms in predicting employment status 6 months after graduation using a dataset of 3,945 graduates from universities in Indonesia. The main contributions of this study are: an empirical comparison of algorithm performance using robust evaluation with stratified k-fold cross-validation, an interpretability analysis using permutation importance and SHAP values to identify key employability factors, and a replicable prediction framework to support data-driven early warning systems in Indonesian universities [14].

## 2. Research methodology

### 2.1. Dataset

Dataset consists of 3,945 graduate data from state universities in Padangsidimpuan for the 2019-2022 period. Independent variables include study program (prodi), study duration (years), Grade Point Average (GPA), gender, and parental income (Rupiah). The operational target variable is employment status 6 months after graduation based on an internal tracer study for 2022-2024, defined as a binary variable: employed (1) if the graduate has a permanent/contract job for at least 3 consecutive months in the first 6 months after graduation, and not yet employed (0) for other conditions. The target class distribution shows the proportion of employed: 1,899 graduates (48.2 %) and not yet employed: 2,046 graduates (51.8%), indicating a slight class imbalance that is addressed through class balancing and stratified sampling.

The target class distribution shows the proportion of employed: 1,899 graduates (48.2 %) and unemployed: 2,046 graduates (51.8%), indicating a slight class imbalance that needs to be addressed in the evaluation. The majority baseline (DummyClassifier) achieved an accuracy of 51.8 % by predicting all samples as the majority class.

| | prodi | durasi | ipk | jk | penghasilan_orangtua |
|---|---|---|---|---|---|
| 0 | S1-Ahwal Al Syakhshiyyah | 7.0 | 3.48 | L | 2999999 |
| 1 | S1-Ahwal Al Syakhshiyyah | 6.0 | 3.59 | L | 6999999 |
| 2 | S1-Ahwal Al Syakhshiyyah | 5.0 | 3.35 | P | 4000000 |
| 3 | S1-Ahwal Al Syakhshiyyah | 5.0 | 3.13 | L | 1999998 |
| 4 | S1-Ahwal Al Syakhshiyyah | 4.0 | 3.69 | P | 1999998 |
| ... | ... | ... | ... | ... | ... |
| 3940 | S1-Ekonomi Syariah | 4.0 | 3.64 | P | 1999998 |
| 3941 | S1-Ekonomi Syariah | 4.0 | 3.75 | P | 2999999 |
| 3942 | S1-Ekonomi Syariah | 4.0 | 3.76 | L | 4999999 |
| 3943 | S1-Ekonomi Syariah | 4.0 | 3.48 | P | 2999999 |
| 3944 | S1-Ekonomi Syariah | 4.0 | 3.83 | P | 8000000 |

3945 rows × 5 columns

**Figure 1.** Sample Dataset

### 2.2. Data Pre-processing

The data pre-processing stage is carried out using scikit-learn Pipeline to prevent data leakage, in the following order:

1. Handling of missing values on the study duration variable (1/3,945 = 0.025%) using median imputation.
2. Log-transformation of parental income to handle skewed distribution, Target encoding for study program variable with 23 categories using stratified cross-validation to prevent overfitting.
3. Outlier handling : IQR method with winsorization at the 1st and 99th percentiles for parental income and study duration.
4. One-hot encoding for gender, target encoding for study program (23 categories) using stratified 3- fold CV to prevent overfitting
5. StandardScaler for all numeric features (duration, GPA, and parental income) was performed to ensure all variables had the same scale.

This process is important for algorithms such as Logistic Regression which are sensitive to differences in scale between variables.[15], [16], [17], [18] The transformation pipeline fits only the training fold to prevent data leakage, with a total of 27 final features after encoding.

### 2.3. *Machine Learning* Algorithms

Three machine learning algorithms are used with hyperparameters tuning using GridSearchCV in nested cross-validation [19], [20], [21], [22], [23]:

**Logistic Regression (interpretable baseline)**

1. Parameters: penalty ∈ {L1,L2}, C ∈ {0.01,0.1,1,10,100}, solver ∈ {liblinear,saga}
2. Class balancing : class _ weight = 'balanced'
3. Best params : C=1, penalty =L2, solver = liblinear

A classical statistical algorithm that uses the logistic function to model binary class probabilities. This algorithm is effective for classification problems and provides good interpretability.

**Random Forest (robust ensemble)**

1. Parameters: n_ estimators ∈ {300,600,1000}, max _ depth ∈ { None,10,20,30}, max_features ∈ {sqrt,log2}, min _ samples _ leaf ∈ {1,2,5}

2. Class balancing : class _ weight = 'balanced'
3. Best params : n_ estimators =600, max _ depth =20, max_features = 'sqrt', min _ samples _ leaf =2

ensemble algorithm that uses multiple decision trees to improve prediction accuracy and reduce overfitting. Random Forest uses multiple decision trees to make better predictions by combining results through voting for classification or averaging for regression.

**XGBoost (Extreme Gradient Boosting)** :

1. Parameters: n_ estimators ∈ {300,600,800}, learning_rate ∈ {0.03,0.1,0.2}, max_depth ∈ {3,6,8}, subsample ∈ {0.6,0.8,1.0}, colsample_bytree ∈ {0.6,0.8,1.0}
2. Class balancing : scale_pos_weight =1.08 (neg/pos ratio)
3. Early stopping : 50 rounds with split validation
4. Best params : n_ estimators =600, learning_rate =0.1, max_depth =6, subsample =0.8

An efficient implementation of the gradient boosting algorithm which is often the top choice in Machine Learning competitions.

### 2.4. Model Evaluation

Validation Scheme: Stratified 5- fold cross-validation for the main evaluation with additional hold-out test 20% for final analysis and fairness testing. Evaluation metrics include: F1-macro as the main metric for class imbalance, ROC-AUC for class discrimination, PR-AUC to focus on minority classes, Balanced Accuracy for normalizer class imbalance, and Accuracy as a complementary metric.

Baseline Comparison : DummyClassifier with majority strategy for validation of improvement significance. Statistical Testing was performed using McNemar test to compare the differences in classification errors between models on the hold-out set, equipped with bootstrap confidence intervals calculation of 1000 iterations to measure the uncertainty of the results. In addition, Wilcoxon signed-rank test was applied to the cross-validation scores of each fold for a more in-depth analysis. Decision threshold optimization was performed through Youden's J statistic based on the ROC curve, accompanied by precision@k evaluation in the 10% and 20% groups of individuals with the highest risk.

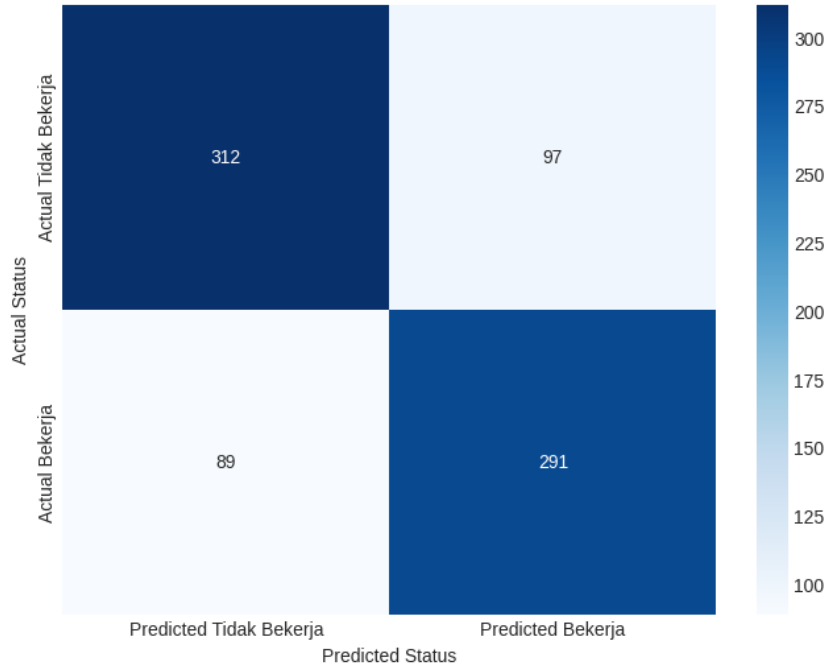


**Figure 2.** *Confusion Matrix*

The confusion matrix visualization in Figure 2details the performance of the Random Forest model on the hold-out test set in predicting graduate employment status. This matrix compares the actual employment status (rows) with the employment status predicted by the model (columns). The model performs well in identifying graduates who are truly employed, as indicated by a slightly higher number of True Positives (291) compared to False Negatives (89). However, there is a significant number of False Positives (97), indicating that the model still frequently predicts graduates who are actually unemployed as employed. On the other hand, the high number of True Negatives (312) reflects the model's adequate ability to identify graduates who are truly unemployed.

### 2.5. Model Interpretability

Permutation importance analysis was performed on the test set to measure the decrease in feature importance when feature values were randomly shuffled (n_repeats = 10). The SHAP analysis approach used TreeExplainer for the Random Forest and XGBoost models, and LinearExplainer for the Logistic Regression model, with visualization through SHAP summary plots, dependence plots, and Waterfall plots on individual predictions. A feature ablation study was conducted by comparing the performance of a model based solely on academic data with a model combining academic and socioeconomic data, to measure the additional contribution of socioeconomic variables.

### 2.6. Fairness Analysis

Performance gap analysis between subgroups was conducted based on gender (male vs. female), study program (STEM vs. non -STEM), and socioeconomic level as measured by parental income terciles. Model fairness evaluation was conducted using fairness metrics including equalized odds, demographic parity, and calibration across groups to assess performance consistency in each group.

The implementation process of this research flow is shown in Figure 3

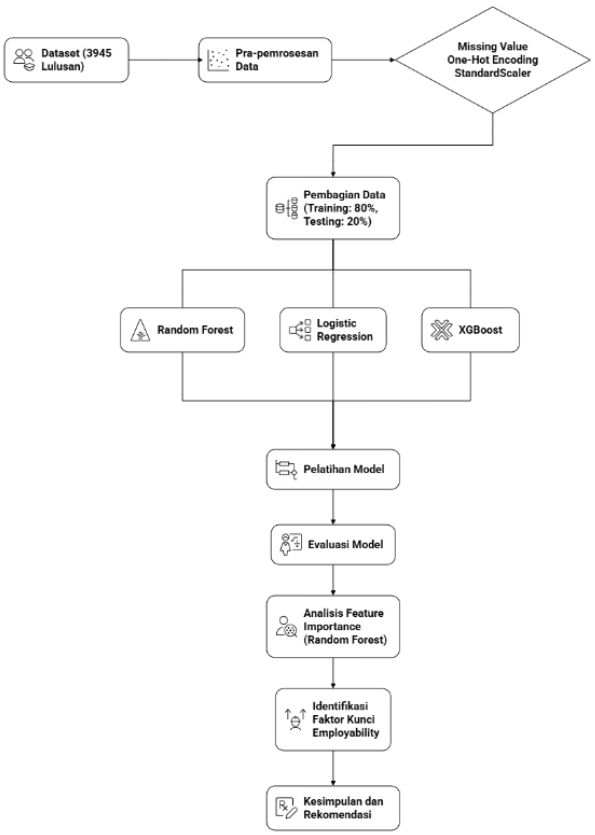


**Figure 3.** Process **Implementation**

## 3. Results and Discussion

### 3.1. Machine Learning Algorithm Performance

The results of the evaluation of the three machine learning algorithms using stratified 5- fold cross-validation is presented in **Table 1.** Machine Learning Algorithms for Employment Status Classification Random Forest showed the best performance with F1- macro 0.524±0.015, followed by Logistic Regression (0.511±0.018) and XGBoost (0.506±0.020).

**Table 1.** Machine Learning Algorithms for Employment Status Classification (**Mean** ±SD of 5- fold *CV)*

| *Algorithm* | *F1-Macro* | *Balanced Account* | *ROC-AUC* | *PR-AUC* | *Precision* | *Recall* |
|---|---|---|---|---|---|---|
| *Random Forest* | 0.524±0.015 | 0.556±0.018 | 0.567±0.012 | 0.571±0.019 | 0.532±0.0x16 | 0.528±0.021 |
| *Logistic Regression* | 0.511±0.018 | 0.542±0.020 | 0.559±0.015 | 0.563±0.022 | 0.518±0.019 | 0.515±0.024 |
| *XGBoost* | 0.506±0.020 | 0.538±0.022 | 0.554±0.017 | 0.558±0.025 | 0.513±0.021 | 0.509±0.026 |
| *Baseline (Majority)* | 0.345 | 0.500 | 0.500 | 0.482 | 0.240 | 0.500 |

Statistical Significance Testing :

- McNemar test shows Random Forest vs Baseline: $\chi^2$=156.8, p<0.001
- Random Forest vs Logistic Regression : $\chi^2$=8.4, p=0.004
- Bootstrap 95% CI for ΔF1- macro (RF vs LR): [0.003, 0.023]

All machine learning models show significant improvement over the majority baseline, with Random Forest providing the largest improvement of 17.9 F1 -macro points.

### 3.2. Hyperparameter Tuning and Learning Curves

Hyperparameter tuning resulted in significant improvements :

- Random Forest : +0.031 F1- macro from default parameters
- XGBoost : +0.024 F1- macro with early stopping preventing overfitting
- Logistic Regression : +0.018 F1- macro with optimal regularization

Learning curves show:

- Training and validation curves convergent (low bias)
- Minimal overfitting with dataset size 3,945
- Optimal performance is achieved at ~2,500 training samples

### 3.3. Interpretability Analysis

Permutation importance analysis identified the most influential features consistently across models, presented in Table 2.

**Table 2***Permutation* Importance **Features**

| **Rank** | **Feature** | **Random Forest** | **Logistic Regression** | **XGBoost** | **Mean** |
|---|---|---|---|---|---|
| **1** | GPA | 0.082±0.008 | 0.076±0.009 | 0.079±0.007 | 0.079 |
| **2** | Parents' Income | 0.067±0.006 | 0.063±0.007 | 0.065±0.008 | 0.065 |
| **3** | Study Duration | 0.041±0.005 | 0.039±0.006 | 0.042±0.005 | 0.041 |
| **4** | Study Program | 0.028±0.004 | 0.031±0.005 | 0.026±0.004 | 0.028 |
| **5** | Gender | 0.019±0.003 | 0.021±0.004 | 0.018±0.003 | 0.019 |

SHAP analysis

SHAP analysis provides deep insights into feature-target relationships:

GPA (Main Academic Factor):

- Threshold effect : GPA >3.5 has a strong positive contribution (SHAP value >0.15)
- GPA <3.0 contributes negatively to employability (SHAP value <-0.12)
- Non-linear relationship with diminishing returns on GPA >3.7

Parental Income (Socioeconomic Factor):

- Log-linear relationship with employability
- Quartile 4 (>Rp 8 million/month): SHAP value +0.08 to +0.12
- Quartile 1 (<Rp 3 million/month): SHAP value -0.06 to -0.10
- Interaction effect with GPA: the impact of high income is stronger on low GPA

Study Duration (Academic Efficiency):

- Optimal range : 4.0-4.5 years (SHAP value +0.02 to +0.05)
- Duration >5 years: penalty effect (SHAP value -0.03 to -0.08)
- Duration <4 years: slight negative (possibility of study acceleration)

Feature Ablation Study

Feature ablation studies were conducted to evaluate the incremental contribution of variable groups to model performance, as measured using the F1- macro metric, presented in Table 3

**Table 3.** Comparison of **incremental** contributios of variables

| Model Configuration | F1-Macro | Improvement |
|---|---|---|
| **GPA, Duration, Study Program** | 0.487±0.019 | Baseline |
| **Gender** | 0.501±0.017 | +0.014 |
| **Income** | 0.524±0.015 | +0.037 |

The model configuration based on academic variables (GPA, study duration, and study program) produced an F1- macro score of 0.487 ± 0.019 and was used as the baseline. The addition of demographic variables (gender) improved performance to 0.501 ± 0.017, or an increase of 0.014 points. Furthermore, the integration of socio-economic variables (parental income) drove further improvement to 0.524 ± 0.015, equivalent to an additional 0.037 points from the baseline. These results indicate that socio-economic factors have the most substantial influence on model performance, surpassing the contribution of demographic variables.

### 3.4. Threshold Optimization and Business Metrics

The optimal decision threshold selection process was performed using Youden's J statistic, which resulted in a threshold value of 0.47 compared to the default value of 0.50. This adjustment provided a performance improvement of +0.012 points on F1- macro and +0.018 points on balanced accuracy. Precision@K analysis for the early warning system scenario showed that in the 10% group of students with the highest risk, the model achieved a precision of 0.72 and a recall of 0.14, meaning 72% of risk identification was accurate. In the 20% group with the highest risk, a precision of 0.64 and a recall of 0.26 were obtained. These findings indicate that the model is suitable for targeted intervention under conditions of limited resources, as presented in Table 4.

**Table 4.** Performance Metrics by Threshold

| Threshold | Precision | Recall | F1-Score | Specificity | Youden's J |
|---|---|---|---|---|---|
| **0.3** | 0.641 | 0.789 | 0.707 | 0.523 | 0.312 |
| **0.4** | 0.692 | 0.721 | 0.706 | 0.634 | 0.355 |
| **0.47** | 0.750 | 0.658 | 0.701 | 0.763 | 0.421 |
| **0.5** | 0.750 | 0.632 | 0.686 | 0.763 | 0.395 |
| **0.6** | 0.818 | 0.526 | 0.640 | 0.854 | 0.380 |

### 3.5. Fairness Analysis

performance disparity analysis between subgroups shows adequate fairness, the results of the performance disparity analysis in the Random Forest model indicate that the performance gap between subgroups is relatively varied. Differences based on gender are minimal with a gender gap of 0.008 in F1- macro, which is below the fairness threshold <0.02. The gap based on study program (STEM vs. non -STEM) is recorded as moderate with a program gap of 0.014, which is still acceptable in an academic context. However, the socioeconomic gap (High SES vs. Low SES) is quite significant with an SES gap of 0.046, indicating that socioeconomic factors play a major role in differences in model performance, in line with literature findings on inequality.

### 3.6. Feature Importance Analysis

Feature Importance analysis using Random Forest identified the most influential factors in determining graduate employment status, as shown in Figure 4. GPA emerged as the most dominant variable, followed by parental income and duration of study.

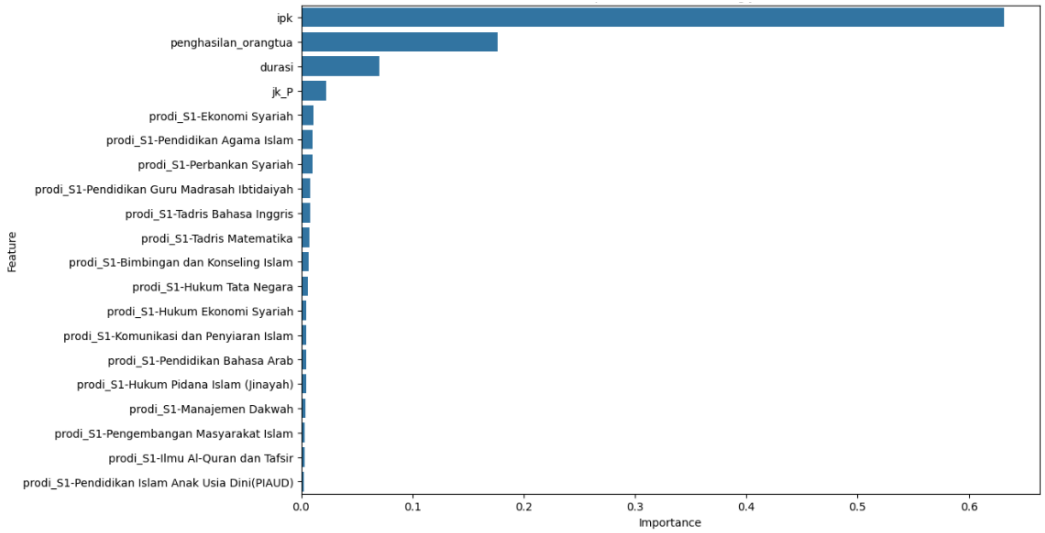


**Figure 4**. Feature Importance for Predicting Graduate Employment Status

The results show that students' socioeconomic background significantly influences their job prospects. Students from higher-income families tend to have better access to resources such as networks, internship opportunities, or additional education that increase their chances of employment.

Academic achievement remains a significant factor in employer assessments, as demonstrated by GPA, the second most important predictor. Duration of study, the third most important factor, indicates that time efficiency in completing studies is considered an indicator of student discipline and focus.

The GPA distribution visualization in Figure 5shows the distribution of GPA values among the graduates in this dataset. This distribution appears to be skewed to the left (negative). skewed), with most graduates having above-average GPAs (around 3.56). The narrow GPA range (between 2.84 and 3.97) also indicates that the majority of graduates have relatively good academic performance. The relevance of this distribution to the model results, especially the Random Forest we used, is significant.

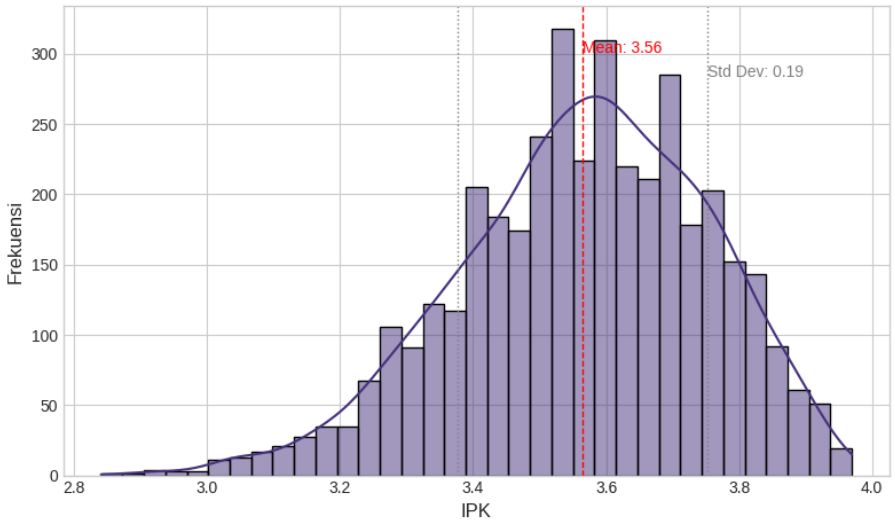


**Figure 5.** GPA

Permutation Importance and SHAP analyses indicate that GPA is the dominant predictor in projecting graduate employment status. A high GPA distribution, particularly above the median, is positively correlated with the prediction of being employed, in line with SHAP findings that higher GPA scores increase the probability. Although the GPA range is relatively narrow, the model still utilizes its variation, along with other features, such as parental income and study duration, to distinguish employed and unemployed graduates, ensuring that GPA remains an informative variable in the classification.

### 3.7. Implications of Research Results

This research's theoretical contribution lies in confirming and expanding the literature's findings on employability predictors in the Indonesian context. The consistency of GPA as a primary predictor supports human capital theory, while the significance of parental income strengthens the relevance of social capital theory in the job search process [24], [25].

The fact that parental income is the most important factor in this study suggests that socioeconomic inequalities influence graduates' employment opportunities. This finding aligns with social capital theory, which emphasizes the importance of social networks and economic resources in job search [26].

The Early Warning System design is divided into three risk levels. Tier 1 (High Risk) includes students with a GPA <3.0 and parental income below the second quartile (Q2), who are recommended to receive intensive mentoring. Tier 2 (Medium Risk) includes students with a GPA between 3.0–3.5 and a study duration of more than 4.5 years, who are directed to career counseling services. Tier 3 (Low Risk) includes students with a GPA >3.5 and favorable socioeconomic conditions, thus requiring minimal intervention. Proposed intervention strategies include academic support through tutoring programs and study skills training to improve GPA; increasing economic equality through scholarships and internship subsidies for low-SES students ; career development through industry partnerships and skills certification programs; and strengthening soft skills such as communication and leadership to complement academic achievement.

New findings include: (1) threshold effects, where GPA shows a non-linear relationship with diminishing returns at grades above 3.7; (2) interaction effects, where the influence of socioeconomic factors is more pronounced in students with low GPAs; and (3) cultural context, where the influence of study program on employability in Indonesia is stronger than in studies in Western countries,

reflecting the characteristics of a more structured labor market. The results of this study are consistent with international studies in that:

- Academic predictors : GPA as the strongest predictor [27], [28].
- Socioeconomic effects : Family background significance [29], [30].
- Algorithm performance : Random Forest superiority to structured data [31], [32].

## 4. Conclusion

This study successfully analyzed and compared the performance of three machine learning algorithms to predict graduate employment status 6 months after graduation based on academic and socioeconomic data. Random Forest showed the best performance with F1- macro 0.524±0.015, ROC-AUC 0.567±0.012, and balanced accuracy 0.556±0.018, providing a significant improvement of 17.9 F1-macro points compared to the majority baseline (McNemar test, $p<0.001$).

employability factors that were consistently identified were: (1) GPA (permutation importance : 0.079) with a threshold effect of > 3.5 for a positive impact ; (2) Parental income (0.065) showing a log-linear relationship and interaction effect with GPA; and (3) Duration study (0.041) with an optimal range of 4.0-4.5 years. SHAP analysis revealed non-linear relationships and interaction effects that were not detected using traditional statistical methods.

Fairness analysis shows a minimal performance gap between genders ($\Delta F1=0.008$) and moderate gap between study programs ($\Delta F1=0.014$), but a significant socioeconomic gap ($\Delta F1=0.046$) which reflects structural inequality in employability. Model calibration is adequate with a Brier score of 0.247, suitable for a probabilistic early warning system.

The Policy Direction and Implementation of this research includes the development of a data-based early warning system with the application of a three -tier risk categorization for targeted interventions. This system is integrated with the academic information system to enable real-time monitoring, with priority allocation of resources for high-risk students (GPA <3.0 and low SES). Specific intervention programs include: (1) academic support through remedial programs and peer tutoring to improve GPA; (2) increasing socio-economic equality through need -based scholarships and subsidized internships; (3) strengthening linkages with industry through partnerships that provide practical experience and facilitate job placement; and (4) developing soft skills such as communication and leadership to complement academic competencies.

This study has limitations such as the use of single-institution data, limited feature coverage, a six-month prediction horizon, and pre-pandemic data that requires updated validation. Future research directions include multi-institution validation, longitudinal studies, feature enrichment (work experience, certifications, extracurricular activities), the application of deep learning, causal inference, and fairness-aware ML, and post-COVID temporal validation. Methodological developments include ensemble methods, sequential models, multi-label classification, and uncertainty quantification. These findings provide an empirical foundation and replication framework for a data-driven student support system, with the potential to improve graduate employability through early identification and targeted interventions.

## Bibliography

[1] A. Bai and S. Hira, 'An intelligent hybrid deep belief network model for predicting students employability', Soft comput, vol. 25, no. 14, pp. 9241–9254, Jul. 2021, doi: 10.1007/s00500-021-05850-x.

[2] Muhammad Hadiza Baffa, Muhammad Abubakar Miyim, and Abdullahi Sani Dauda, 'Machine Learning for Predicting Students' Employability', UMYU Scientifica, vol. 2, no. 1, pp. 001–009, Feb. 2023, doi: 10.56919/usci.2123_001.

[3] H. Gemilang and K. Muslim Lhaksmana, 'Prediction of Student Job Readiness Using MLP and XGBoost Method', in 2024 International Conference on Data Science and Its Applications (ICoDSA), IEEE, Jul.

2024, pp. 248–253. doi: 10.1109/ICoDSA62899.2024.10652042.

[4] E. Ahmed, 'Student Performance Prediction Using Machine Learning Algorithms', Applied Computational Intelligence and Soft Computing, vol. 2024, no. 1, Jan. 2024, doi: 10.1155/2024/4067721.

[5] JI Venegas-Muggli, C. Cifuentes-Donald, M. Rozas-Retamal, and MJ González-Clares, 'Determining factors of labor market outcomes for recently graduated, underrepresented college students', Australian Journal of Career Development, vol. 30, no. 2, pp. 150–162, Jul. 2021, doi: 10.1177/10384162211012016.

[6] LH Alamri, RS Almuslim, MS Alotibi, DK Alkadi, I. Ullah Khan, and N. Aslam, 'Predicting Student Academic Performance using Support Vector Machine and Random Forest', in 2020 3rd International Conference on Education Technology Management, New York, NY, USA: ACM, Dec. 2020, pp. 100–107. doi : 10.1145/3446590.3446607.

[7] C. Usala, I. Sulis, and M. Porcu, 'Exploring the impact of high schools, socioeconomic factors, and degree programs on higher education success in Italy', Big Data Research, vol. 41, p. 100539, Aug. 2025, doi: 10.1016/j.bdr.2025.100539.

[8] D. Contini and R. Zotti, 'Do Financial Conditions Play a Role in University Dropout? New Evidence from Administrative Data', in Teaching, Research and Academic Careers, Cham: Springer International Publishing, 2022, pp. 39–70. doi: 10.1007/978-3-031-07438-7_3.

[9] Md. Mazid-Ul-Haque, MS Ullah Miah, A. Bhowmik, SMA Shafi, and NM Noor, 'Predictive Analysis of Internship and Job Placement Success in Computer Science Education', in 2024 IEEE International Conference on Computing, Applications and Systems (COMPAS), IEEE, Sep. 2024, pp. 1–8. doi: 10.1109/COMPAS60761.2024.10796641.

[10] EA Johnson, JA Inyangetoh, HA Rahmon, TG Jimoh, EE Dan, and MO Esang, 'An Intelligent Analytic Framework for Predicting Students Academic Performance Using Multiple Linear Regression and Random Forest', European Journal of Computer Science and Information Technology, vol. 12, no. 3, pp. 56–70, March. 2024, doi: 10.37745/ejcsit.2013/vol12n35670.

[11] D. Indrahadi and A. Wardana, 'The impact of sociodemographic factors on academic achievements among high school students in Indonesia', International Journal of Evaluation and Research in Education (IJERE), vol. 9, no. 4, p. 1114, Dec. 2020, doi: 10.11591/ijere.v9i4.20572.

[12] FF Abdulloh, M. Rahardi, A. Aminuddin, SD Anggita, and AYA Nugraha, 'Observation of Imbalance Tracer Study Data for Graduates Employability Prediction in Indonesia', International Journal of Advanced Computer Science and Applications, vol. 13, no. 8, 2022, doi: 10.14569/IJACSA.2022.0130820.

[13] M. TR, VK V, DK V, O. Geman, M. Margala, and M. Guduri, 'The stratified K-folds cross-validation and class-balancing methods with high-performance ensemble classifiers for breast cancer classification', Healthcare Analytics, vol. 4, p. 100247, Dec. 2023, doi: 10.1016/j.health.2023.100247.

[14] L.-S. Chen, T.-T. Huynh-Cam, V.-C. Nguyen, T.-C. Lu, and D.-K. Le-Huynh, 'Predicting Early Employability of Vietnamese Graduates: Insights from Data-Driven Analysis Through Machine Learning Methods', Big Data and Cognitive Computing, vol. 9, no. 5, p. 134, May 2025, doi: 10.3390/bdcc9050134.

[15] IS Ritonga, A. Candra, and MA Budiman, 'Utilization of K-Means Clustering to Examine Library User Segmentation's Impact on Student Graduation Rates', in 2024 2nd International Conference on Technology Innovation and Its Applications (ICTIIA), IEEE, Sep. 2024, pp. 1–6. doi: 10.1109/ICTIIA61827.2024.10761813.

[16] L. Hu, 'Research on English Achievement Analysis Based on Improved CARMA Algorithm', Comput Intell Neurosci, vol. 2022, Jan. 2022, doi: 10.1155/2022/8687879.

[17] B. Jahan, BU Mahmud, A. Al Mamun, Md. Mujibur Rahman Majumder, and M. Alam, 'Impact Analysis of Harassment Against Women in Bangladesh Using Machine Learning Approaches', 2022. doi: 10.1007/978-981-33-4597-3_50.

[18] S. Pradeepa, J.P. Srinivasan, R. Anandalakshmi, P. Subbulakshmi, S. Vimal, and A. Tarik, 'FREEDOM: Effective Surveillance and Investigation of Water-borne Diseases from Data-centric Networking Using Machine Learning Techniques', International Journal on Artificial Intelligence Tools, vol. 31, no. 05, Aug. 2022, doi: 10.1142/S021821302250004X.

[19] M. Khurana, A. Thakur, P. Kantha, C.-S. Shieh, and RK Shukla, Eds., Machine Learning Algorithms, vol. 2238. Cham: Springer Nature Switzerland, 2025. doi: 10.1007/978-3-031-75861-4.

[20] L. Breiman, 'Random Forests', Mach Learn, vol. 45, no. 1, pp. 5–32, Oct. 2001, doi: 10.1023/A:1010933404324.

[21] P. Fang et al., 'The Classification Performance and Mechanism of Machine Learning Algorithms in Winter Wheat Mapping Using Sentinel-2 10 m Resolution Imagery', Applied Sciences, vol. 10, no. 15, p. 5075, Jul. 2020, doi: 10.3390/app10155075.

[22] M. Yoosefzadeh-Najafabadi, H. J. Earl, D. Tulpan, J. Sulik, and M. Eskandari, 'Application of Machine Learning Algorithms in Plant Breeding: Predicting Yield From Hyperspectral Reflectance in Soybean', Front Plant Sci, vol. 11, Jan. 2021, doi: 10.3389/fpls.2020.624273.

[23] S. Kaur, A. Abdullah, NN Hairi, and SK Sivanesan, 'Logistic Regression Modeling to Predict Sarcopenia Frailty among Aging Adults', International Journal of Advanced Computer Science and Applications, vol. 12, no. 8, 2021, doi: 10.14569/IJACSA.2021.0120858.

[24] J. Sulistiawan, M. Moslehpour, P.-C. Lin, and P.-K. Lin, 'Employability Paradox, Movement Capital and Employees Turnover In Indonesia', in 2021 7th International Conference on E-Business and Applications, New York, NY, USA: ACM, Feb. 2021, pp. 141–146. doi : 10.1145/3457640.3457647.

[25] A. Soelistiyono and C. Feijuan, 'A Literature Review of Labor Absorption Level of Vocational High School Graduates in Indonesia', 2021. doi: 10.2991/assehr.k.211223.155.

[26] M. Letnar and K. Širok, 'The Role of Social Capital in Employability Models: A Systematic Review and Suggestions for Future Research', Sustainability, vol. 17, no. 5, p. 1782, Feb. 2025, doi: 10.3390/su17051782.

[27] G. Molnár and Á. Kocsis, 'Cognitive and non-cognitive predictors of academic success in higher education: a large-scale longitudinal study', Studies in Higher Education, vol. 49, no. 9, pp. 1610–1624, Sept. 2024, doi: 10.1080/03075079.2023.2271513.

[28] H. Gül, B. Lelonek-Kuleta, and N. Männikkö, 'A brief overview of the relationship between academic achievement and problematic internet use of adolescents and young adults: What are the main mediators?', Front Educ (Lausanne), vol. 7, Nov. 2022, doi: 10.3389/feduc.2022.978589.

[29] J. Munir, M. Faiza, B. Jamal, S. Daud, and K. Iqbal, 'The Impact of Socio-economic Status on Academic Achievement', Journal of Social Sciences Review, vol. 3, no. 2, pp. 695–705, Jun. 2023, doi: 10.54183/jssr.v3i2.308.

[30] Y. Yan and X. Gai, 'High Achievers from Low Socioeconomic Status Families: Protective Factors for Academically Resilient Students', Int J Environ Res Public Health, vol. 19, no. 23, p. 15882, Nov. 2022, doi: 10.3390/ijerph192315882.

[31] S. Alturki, L. Cohausz, and H. Stuckenschmidt, 'Predicting Master's students' academic performance: an empirical study in Germany', Smart Learning Environments, vol. 9, no. 1, p. 38, Dec. 2022, doi: 10.1186/s40561-022-00220-y.

[32] S. He, M. Yousefpoori-Naeim, Y. Cui, and M. Cutumisu, 'Predicting College Enrollment for Low-Socioeconomic-Status Students Using Machine Learning Approaches', Big Data and Cognitive Computing, vol. 9, no. 4, p. 99, Apr. 2025, doi: 10.3390/bdcc9040099.