Building of Informatics, Technology and Science (BITS)
Volume 7, No 4, March 2026 Page: 2715−2726
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i4.9525

Predictive Modeling of National University Rankings Using Ensemble Machine Learning and Multi-Dimensional Institutional Performance Indicators: Evidence from Japan

Bernadus Gunawan Sudarsono[1,*], Raditya Galih Whendasmoro[2]

[1]Fakultas Ilmu Komputer, Program Studi Informatika, Universitas Bhayangkara Jakarta Raya, Jakarta, Indonesia
[2]Fakultas Ilmu Komputer, Program Studi Sistem Informasi, Universitas Bung Karno, Jakarta, Indonesia
Email: [1,*]bernadus.gs@dsn.ubharajaya.ac.id, [2] raditya_gw@ubk.ac.id
Email Penulis Korespondensi: bernadus.gs@dsn.ubharajaya.ac.id
Submitted: **12/03/2026**; Accepted: **31/03/2026**; Published**: 31/03/2026**

Abstract–The global higher education landscape is becoming increasingly competitive in attracting outstanding students, qualified faculty, and international research collaborations. University ranking systems serve as strategic instruments for assessing institutional performance and as a basis for public policy. However, traditional ranking approaches employing linear aggregate scores often oversimplify the complex relationships among indicators such as research, internationalization, and graduate outcomes. This study develops a data-driven predictive model to map the non-linear relationships among university performance indicators. The research employs a quantitative predictive analytics approach using a dataset of 52 Japanese universities from the 2024–2026 period, encompassing the variables Research_Impact_Score, Employment_Rate, Intl_Student_Ratio, Institution_Age, Institution_Type, and Region, with National_Rank as the target variable. The research stages include data preprocessing (handling missing values, encoding, scaling), feature engineering (including Institutional Age), regression model development (Linear, Ridge, Lasso, SVR) as well as ensemble models (Random Forest and Gradient Boosting), evaluation using RMSE, MAE, and R^2, and explainable analysis based on feature importance. The results indicate that the Gradient Boosting model delivers the best performance with an RMSE of 1.175117, MAE of 1.087856, and R^2 of 0.994988, followed by Random Forest with an RMSE of 1.436536 and R^2 of 0.992510. Traditional linear regression models demonstrate significantly lower performance (R^2 0.657519), confirming the superiority of non-linear approaches in modeling complex relationships among indicators. Stability testing using K-Fold Cross Validation yields an average RMSE of 1.1045 with a difference of 0.4493 between folds, indicating model consistency. Feature contribution analysis reveals that Research_Impact_Score is the dominant factor with a contribution of 97.94%, followed by Employment_Rate at 1.81%, while internationalization indicators and geographical factors contribute minimally. These findings confirm that research performance constitutes the primary determinant of university rankings, whereas employability and internationalization serve as supporting factors. This study demonstrates that ensemble-based machine learning models are effective in predicting national rankings accurately and interpretably. This approach offers a multidimensional evaluation framework that is more representative than linear aggregate scores, and provides policy implications for enhancing research quality, curriculum relevance, and internationalization strategies of higher education institutions.

Keywords: University Ranking; Machine Learning; Predictive Modeling; Research Impact; Higher Education Performance

1. INTRODUCTION

Higher education institutions worldwide are increasingly driven by global competition to attract high-quality students, distinguished faculty, and international research collaborations. University rankings have become a crucial instrument for assessing institutional performance in a global context and serve as a foundation for public policy in the development of higher education.[1]. However, many traditional ranking systems produce linear aggregate scores that tend to oversimplify the complexity of institutional performance. These aggregate scores often fail to reflect the dynamic interactions between indicators such as research, internationalization, and graduate outcomes. Modern research demands a data-driven approach to map the non-linear relationships among institutional performance indicators[2].

Data mining and machine learning are now used in educational contexts to uncover hidden patterns within higher education datasets. This approach enables a holistic evaluation of institutional performance, rather than merely aggregating a single index. A data-driven framework allows for the development of predictive models that are more adaptive to the dynamics of university rankings[3]. Consequently, ranking studies that integrate modern statistical methods are increasingly needed to analyze university performance more accurately[2]. The topic of university performance evaluation using a non-linear approach has become highly relevant in the era of higher education globalization[1].

Although university rankings play an important role, many developing countries have not yet utilized predictive data mining models to systematically map national university rankings[1]. Traditional rankings often overlook the complex and contextual non-linear relationships among performance indicators. In the context of higher education in Indonesia, research quality and scientific output still lag behind those of advanced education systems, resulting in datasets that lack the standardization necessary for robust modeling[4]. Issues such as limited research collaboration, research funding, and scientific output pose significant challenges for data-based modeling[5].

Models that only use aggregate scores without considering the contribution of individual indicators are inadequate for representing the structure of institutional performance. Consequently, national rankings tend not to be explainable by variables that are interactive and non-linear in nature. Furthermore, the lack of research integrating quantitative indicators, such as research impact and internationalization, weakens the inferential power of ranking

Building of Informatics, Technology and Science (BITS)
Volume 7, No 4, March 2026 Page: 2715−2726
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i4.9525

models[2]. This problem is particularly evident in developing countries in Southeast Asia, where consistent, multi-indicator quantitative datasets are not yet available. Therefore, data-driven approaches for predicting university rankings remain very limited in the higher education literature of developing countries[6]. This creates a strong methodological need to build predictive models capable of capturing non-linear relationships among performance indicators.

Although global studies on university rankings are abundant, few conduct quantitative predictions of national rankings using data mining methods. The majority of research tends to focus on descriptive analysis or comparisons among global ranking systems rather than comprehensive predictive modeling. Furthermore, the integration of academic, internationalization, and employability indicators within a single predictive model is still rarely found. Explainable models or interpretations of the contribution of each indicator to ranking outcomes in a national context are not yet well developed. Existing literature often neglects structural assessment and the contribution of explicit features to institutional rankings. For example, global ranking systems such as QS and THE consider many complex indicators but are rarely adapted for national contexts[1]. In the context of developing countries like Indonesia, predictive models for national rankings have not been extensively discussed in indexed international journals. This reveals a theoretical and methodological gap in the use of data mining approaches for national rankings. Therefore, research employing predictive methods with multidimensional indicators is essential to bridge this gap. This study seeks to address this gap by adopting an explicitly predictive modeling approach with a strong theoretical foundation.

Several previous studies have demonstrated the application of machine learning in predicting university rankings and higher education outcomes. Leslie J. Wardley et al. (2024) used Gradient Boosting and Random Forest to predict university rankings in Canada with high accuracy, demonstrating the effectiveness of ensemble models in mapping ranking positions based on historical institutional indicator data[7]. Wai Yie Leong (2025) reported that Random Forest and Gradient Boosting outperformed traditional models in predicting global ranking trends by incorporating new factors such as international collaboration[8]. Additionally, Eloy López-Meneses et al. (2025) conducted a bibliometric analysis of educational data mining and identified an increasing trend in the use of predictive modeling for optimizing higher education decisions, affirming the relevance of predictive methodologies in the context of institutional performance evaluation[9]. Arbnor Rushiti et al. (2024) also developed a machine learning-based ranking system using bibliometric features to estimate university positions based on research output, reinforcing evidence that academic output characteristics serve as important predictors in ranking models[10].

Japan was selected as the research context because the higher education system in that country possesses a standardized and stable evaluation structure, which is reflected in both national and global rankings[11]. University rankings in Japan incorporate indicators that reflect research capacity, internationalization, and graduate outcomes, making them suitable for data-driven predictive models[12]. The dataset from Japan encompasses a wide variation among national, public, and private universities with heterogeneous performance. This enables a representative and comprehensive cross-category analysis of institutions. Moreover, Japan is a developed country with high research output and strong international collaboration[13][14], rendering its dataset more globally meaningful. In contrast, in Indonesia, challenges such as limited funding and research collaboration remain significant, which impacts the availability of well-structured quantitative data. The inequality in quality among institutions in Indonesia makes it difficult to develop stable and comparative predictive models[4]. Japan also possesses a more extensive body of educational literature related to the quantitative evaluation of university performance[15]. The selection of the Japanese dataset enables more accurate predictions and robust interpretative analysis of the contribution of each indicator[2]. Therefore, this study employs Japan as a representative case study for a national ranking predictive model.

This study aims to develop a predictive model for national university rankings using multidimensional performance indicators. This model is expected to accurately capture non-linear relationships among indicators. A subsequent objective is to identify the dominant features that significantly influence rankings. This study also comparatively evaluates the performance of predictive models across different machine learning algorithms.

2. RESEARCH METHODOLOGY

This study employs a quantitative, data-driven methodology, as the primary focus is the numerical prediction of university rankings based on performance indicators[16]. This approach is referred to as predictive analytics, which involves the use of statistical techniques and machine learning to project future target values[17]. The type of data analyzed is structured, with each attribute being either numerical or ordinally categorical, allowing for quantitative processing. The research framework follows a data-driven workflow encompassing preprocessing, feature engineering, model building, and evaluation. The main focus of the research is predictive modeling to forecast the national rankings of universities in Japan based on the available input attributes. The models employed are machine learning regression algorithms capable of projecting continuous values of the target variable. This approach is consistent with the predictive analytics literature, which emphasizes the importance of model validation and interpretation. This research adopts a standardized, reproducible machine learning pipeline framework. Consequently, the research design integrates classical statistical methods with modern machine learning techniques for meaningful prediction and interpretation.

Building of Informatics, Technology and Science (BITS)
Volume 7, No 4, March 2026 Page: 2715−2726
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i4.9525

2.1 Research Stages

The stages of the research are visually presented in Figure 1, which serves as a conceptual framework connecting the methodological aspects with the research objective, namely, to build a predictive model for university rankings based on multidimensional, explainable, and data-driven indicators.

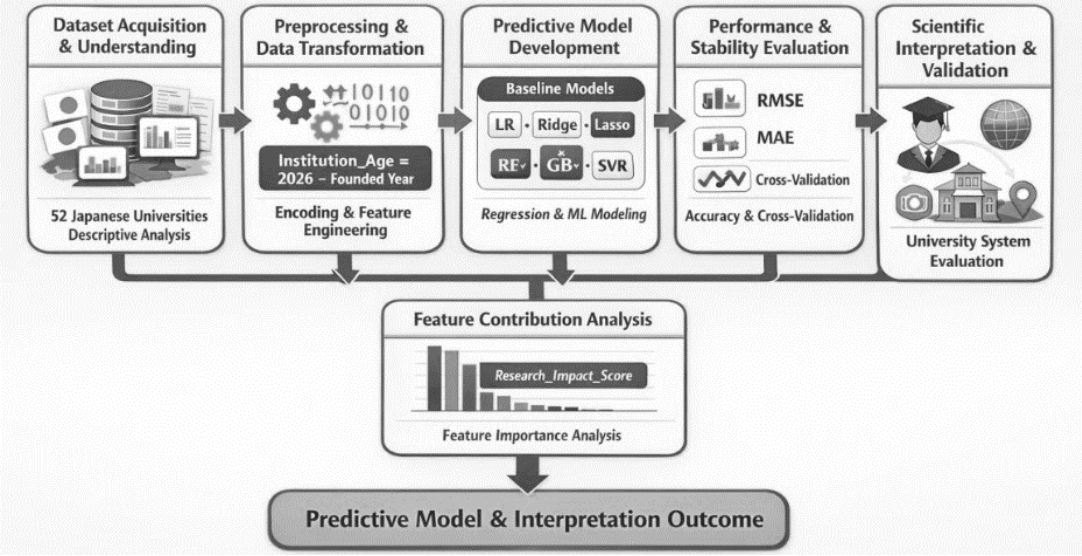


Figure 1. Research stages

Based on Figure 1, the research stages are systematically arranged according to the flowchart presented, which specifically represents the actual process conducted in this study, starting from the collection of the Japanese university dataset, data preprocessing, machine learning model development, model performance evaluation, to feature contribution analysis.

2.2 Dataset and Variables

The dataset employed comprises performance indicators from 52 Japanese universities, encompassing academic, internationalization, and graduate outcome aspects[18]. The data source is the 2024–2026 Japanese university national ranking dataset, which includes numerical and categorical attributes related to institutional performance[19]. The target variable in this study is National_Rank, a continuous ordinal variable indicating the institution's ranking position. Meanwhile, the predictor variables consist of Research_Impact_Score, Intl_Student_Ratio, Employment_Rate, Institution_Type, Founded_Year, and Region. Each variable is operationally defined; for instance, Research_Impact_Score represents the score for scientific contribution and research impact[20]. The Intl_Student_Ratio variable reflects the proportion of international students relative to the total student population. Employment_Rate is the percentage of graduates absorbed into the workforce within a specified period after graduation[19]. Institution_Type and Region are categorical variables indicating the type of institution and the geographical location of each university, respectively[21]. These variable definitions are utilized to construct a representative and meaningful feature space for machine learning modeling[18].

2.3 Data Preprocessing

The preprocessing stage commences with handling missing values, namely the identification and imputation of missing data using statistical methods such as mean or median. Categorical variables are addressed through encoding techniques, such as One-Hot Encoding, to convert categories into numerical representation. Normalization and standardization are applied to scale numerical variables, ensuring a uniform distribution prior to regression modeling. Outlier handling, involving the detection and management of extreme values that could affect model stability, is also conducted at this stage. These preprocessing principles are consistent with data preprocessing literature, which emphasizes the necessity of cleaning, transformation, encoding, and scaling. For missing data, strategies like median imputation are more robust against extreme deviations compared to mean imputation. Categorical attributes with more than two levels are encoded, preserving ordinal relationships where necessary. These steps are undertaken to ensure the dataset is clean, properly formatted, and suitable for the modeling phase. Such transformations enhance model performance and reduce bias induced by data noise [21].

2.4 Feature Engineering

One form of feature engineering in this study is the calculation of Institutional Age, utilized to measure institutional maturity[21], as follows:

$$Institutional\ Age = 2026 - Founded\_Year \quad (1)$$

Building of Informatics, Technology and Science (BITS)
Volume 7, No 4, March 2026 Page: 2715−2726
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i4.9525

Feature engineering also encompasses the creation of composite features that combine several performance indicators into new, domain-relevant features. This stage enhances the model's predictive power by extracting additional information from raw data. Feature scaling is performed after feature generation to ensure that new attributes adhere to a consistent numerical scale. This process aligns with best practices in data mining and machine learning for generating meaningful and effective features. Feature engineering also assists the model in capturing non-linear relationships among indicators that might not be captured by linear models. Beyond Institutional Age, other transformed features such as ratios or interactions between indicators can also be created. These new features are statistically tested concerning their correlation with and contribution to the ranking target. Effective feature engineering generally improves prediction accuracy and model interpretability[21].

2.5 Model Development

The core task of this research is to develop a regression model capable of predicting rank as a continuous numerical variable [18]. The first baseline model is Linear Regression, a classical linear regression that models linear relationships between variables:

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n \quad (2)$$

with β as the parameter coefficient to be estimated [20]. Ridge Regression and Lasso Regression are used as regularization-based baselines to reduce multicollinearity[18]. Ridge Regression applies an L2 penalty::

$$min = \|Y - X\beta\|^2 + \lambda\|\beta\|^2 \quad (3)$$

whereas Lasso Regression employs an L1 penalty to control model complexity[18]:

$$min = \|Y - X\beta\|^2 + \lambda\|\beta\|_1 \quad (4)$$

The machine learning models include Random Forest Regressor, which constructs multiple decision trees to enhance prediction accuracy and stability. Gradient Boosting Regressor utilizes a boosting approach that combines multiple weak learners to iteratively correct errors. Support Vector Regression (SVR) is employed to capture non-linear relationships through kernel functions[19][18].

2.6 Model Training Scheme

he data is divided into two subsets: training (80%) and testing (20%) for model performance evaluation[19]. This scheme aligns with common practices in predictive analytics to assess model generalization[19]. Furthermore, cross-validation techniques, such as k-fold CV, are applied to reduce the variance in performance estimation[20]. Hyperparameter tuning is conducted using grid search or random search to select the optimal combination of parameters[20]. The model selection strategy is based on the performance evaluation metrics, such as RMSE, MAE, and R^2, on the validation set. Evaluation metrics are compared across models to determine the superior model [19]. Random data splitting is performed with a fixed seed to ensure reproducibility[20]. Learning curves are also analyzed to detect overfitting or underfitting[19]. Model stability evaluation is carried out by examining the distribution of metrics across each fold[20].

2.7 Model Evaluation

The model is evaluated using three primary regression metrics: RMSE, MAE, and the R^2 score. RMSE is used to measure the prediction deviation in the target's unit using formula (6), MAE is employed to represent the model's average absolute error using formula (7), and R^2 is calculated to indicate the percentage of target variance that can be explained by the model[20]:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \quad (5)$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \quad (6)$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \quad (7)$$

The selection of these metrics aligns with the standard evaluation practices for regression models in machine learning literature. The models are compared using a tabular metric matrix to facilitate a clear interpretation of the comparisons[19]. Analysis of variance (ANOVA) of the metrics is conducted to assess the consistency of model performance[20]. The optimal model is selected based on the trade-off between low error and a high R^2 value[19]. The evaluation results are presented in graphical visualizations to enhance interpretability[20].

2.8 Explainable Analysis

An analysis of feature contribution is performed to examine the feature importance generated by ensemble models such as Random Forest and Gradient Boosting. Feature importance helps elucidate the relative contribution of each variable to the model's predictions. For model interpretation based on explainable AI (XAI), techniques such as SHAP

Building of Informatics, Technology and Science (BITS)
Volume 7, No 4, March 2026 Page: 2715−2726
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i4.9525

(SHapley Additive exPlanations) values can be employed, where SHAP indicates the marginal contribution of each feature to individual predictions. SHAP is an approach that facilitates transparency in complex models by providing a feature contribution score per instance[19]. Sensitivity analysis assists in evaluating the stability of predictions when specific features are modified. The explainable AI approach reinforces the scientific interpretation of the models employed and their predictive outcomes[18]. The results of the feature importance analysis can be presented in bar charts or heatmaps to visualize the contribution of dominant features[19]. This explainable analysis can also guide policymakers in understanding the dominant factors influencing university rankings[18]. Thus, the explainable AI method provides the necessary interpretative context beyond mere quantitative metrics[19].

3. RESULTS AND DISCUSSION

This section presents the results of dataset processing, model performance evaluation, and an empirical analysis of the model's predictive capability based on the variables used. The evaluation was conducted using the Root Mean Square Error (RMSE) and Mean Absolute Error (MAE) metrics, as both are capable of quantitatively and objectively measuring prediction errors.

3.1 Data Collection

Descriptive statistical analysis was performed on the collected data to understand the fundamental characteristics of the dataset, the distribution of variables, the relationships among indicators, and structural patterns based on region and type of institution. This analysis is essential to ensure data quality and to identify initial patterns that may influence the predictive modeling outcomes.

Table 1. Descriptive Statistics of the Dataset

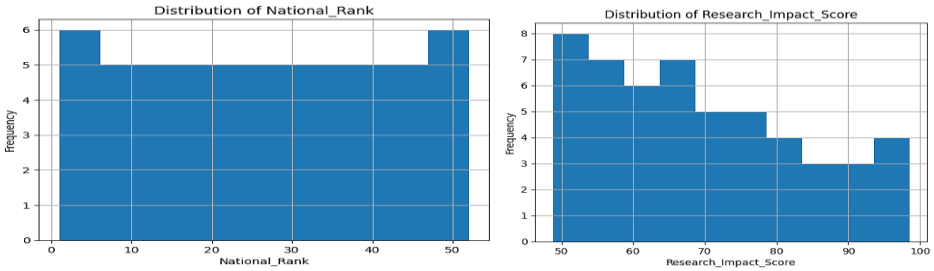
Variabel	Count	Mean	Std	Min	25%	50% (Median)	75%	Max
National_Rank	52	26.50	15.15	1	13.75	26.50	39.25	52
Founded_Year	52	1926.90	39.83	1858	1888	1940.50	1949	2024
Research_Impact_Score	52	69.36	14.23	48.80	57.33	67.15	79.83	98.50
Intl_Student_Ratio (%)	52	10.28	7.53	3.90	5.43	8.15	12.18	48.50
Employment_Rate (%)	52	92.22	4.16	84.20	88.78	93.30	95.25	98.20
Institution_Type	52	–	–	–	–	–	–	–
Region	52	–	–	–	–	–	–	–

Based on Table 1, the variable National_Rank has a mean value of 26.50 with a standard deviation of 15.15, indicating that the distribution of university rankings is spread relatively evenly from the highest to the lowest ranks within the national system. The variable Founded_Year has a mean of 1926.90 with a standard deviation of 39.83, suggesting that the majority of universities in Japan are long-established institutions with mature institutional experience. The variable Research_Impact_Score has a mean of 69.36, with values ranging from 48.80 to 98.50, indicating significant variation in research performance among universities, which serves as an important indicator in academic evaluation systems. The variable Employment_Rate shows a mean of 92.22%, indicating that most universities have a high graduate absorption rate, reflecting the quality of education and the relevance of the curriculum to industry needs. The variable Intl_Student_Ratio has a mean of 10.28% with a standard deviation of 7.53, signifying that the level of internationalization varies considerably across institutions, where only a few universities achieve a very high degree of internationalization. Furthermore, the distribution of categorical variables reveals that the majority of institutions are national universities, highlighting the dominance of government-funded institutions in Japan's higher education system. The regional variable demonstrates high geographical diversity, encompassing 31 distinct regions, which indicates a geographically widespread distribution of institutions. Overall, the descriptive statistics indicate that the dataset exhibits substantial variation in key performance indicators, suggesting that it is highly suitable for use in developing predictive models based on machine learning, as the model can effectively learn the patterns of relationships between institutional performance indicators and national university rankings.

3.1.1 Distribution of Variables

The distribution of variables in the dataset illustrates the variation in university performance based on academic indicators, internationalization, employability, and institutional characteristics. The target variable, National_Rank, shown in Figure 2(a), has a limited range, spanning from the highest to the lowest ranks within the national system, which reflects an ordered and standardized ranking structure. This distribution tends to be uneven, where a small proportion of universities occupy the top ranks with very high performance, while the majority are situated in the middle and lower ranks. This pattern indicates a concentration of high performance within specific institutions that possess competitive advantages.

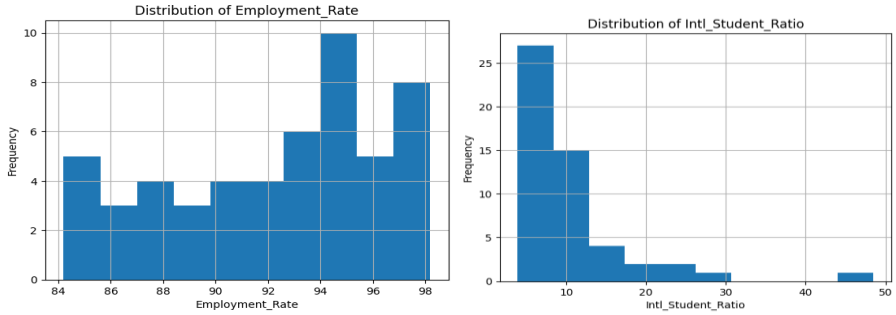
Building of Informatics, Technology and Science (BITS)
Volume 7, No 4, March 2026 Page: 2715−2726
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i4.9525



(a) (b)

Figure 2. Distribution of the National_Rank Variable (a) and the Research_Impact_Score Variable (b)

The Research_Impact_Score variable exhibits a considerably varied distribution, with several universities achieving substantially higher research scores compared to other institutions. This indicates a disparity in research performance among institutions, a common phenomenon within the global higher education system. Universities with high research scores are typically research-intensive institutions characterized by a high volume of publications and citations. The Employment_Rate variable, as shown in Figure 3 (a), demonstrates a generally high distribution, suggesting that the majority of universities achieve favorable graduate employment outcomes. However, inter-institutional variation exists, reflecting differences in educational quality and the relevance of curricula to industry demands.



(a) (b)

Figure 3. Distribution of the Employment_Rate Variable (a) and the Intl_Student_Ratio Variable (b)

The Intl_Student_Ratio variable in Figure 3 (b) exhibits a distribution that tends to be low to moderate, indicating that most universities have a limited level of internationalization. Only a few universities possess a high ratio of international students, which are typically institutions with a global reputation. Furthermore, the Institution_Age variable also demonstrates a wide distribution, reflecting the diversity of institutional age, ranging from relatively new institutions to those that have been established for over a century. This indicates a variation in institutional experience within the higher education system. Overall, the distribution of the variables indicates heterogeneity in performance among universities, which is an ideal condition for the development of a predictive model, as the model can learn the patterns of variation present in the data.

3.1.2 Initial Correlation Among Indicators

Correlation analysis was conducted to identify the relationship between the predictor variables and the target variable. This correlation is essential for understanding the strength and direction of the relationship between variables before modeling is carried out; for more details, please refer to Figure 4 below.

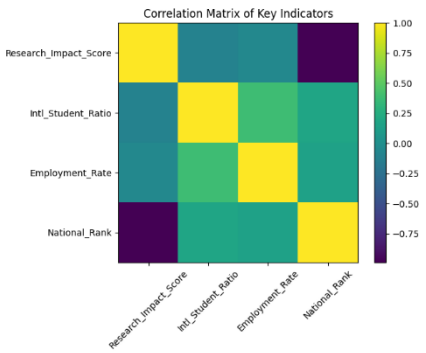


Figure 4. Correlation between Indicators

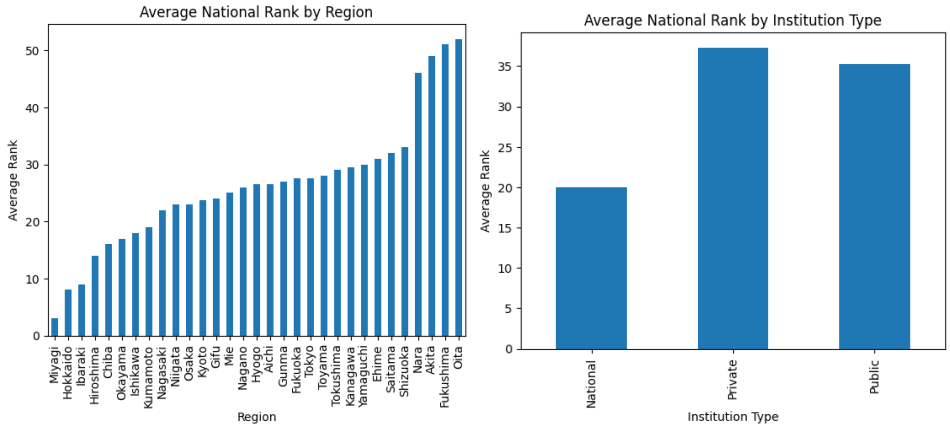
Building of Informatics, Technology and Science (BITS)
Volume 7, No 4, March 2026 Page: 2715−2726
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i4.9525

Based on Figure 4, the Research_Impact_Score variable exhibits a very strong negative correlation with the National_Rank variable. This indicates that the higher the research impact score, the lower the national ranking value, signifying a better ranking position. This correlation suggests that research performance is a primary determinant of university rankings. The Employment_Rate variable also shows a negative correlation with National_Rank, indicating that universities with high employability rates tend to achieve better rankings. This demonstrates that an institution's capacity to produce graduates who are absorbed into the workforce is a crucial indicator of institutional performance.

The Intl_Student_Ratio variable shows a negative correlation with National_Rank, albeit with a weaker magnitude compared to Research_Impact_Score. This suggests that internationalization does influence rankings, but it is not a dominant factor. The Institution_Age variable exhibits a relatively weak correlation with National_Rank, implying that the age of an institution is not a primary determinant of its performance. Younger institutions can still achieve high rankings if they demonstrate strong research performance. Furthermore, the correlations among predictor variables reveal that Research_Impact_Score has a positive relationship with Intl_Student_Ratio and Employment_Rate. This indicates that institutions with high research performance also tend to have high levels of internationalization and employability. Overall, the correlation analysis underscores that research performance is the most potent indicator in determining university rankings.

3.1.3 Regional and Institutional Patterns

Analysis of regional patterns reveals that top-ranking universities tend to be concentrated in major metropolitan areas such as Tokyo and other advanced industrial regions. These areas benefit from superior access to research resources, funding, and industry collaborations, all of which support enhanced institutional performance. Conversely, universities located in regional or non-metropolitan areas tend to have lower rankings, as illustrated in Figure 5(a). This disparity can be attributed to limited resources, restricted access to funding, and fewer opportunities for research collaboration.



(a) (b)

Figure 5. Average National Rank by Region (a) and Average National Rank by Institution Type (b)

The analysis based on Institution Type, as presented in Figure 5(b), indicates that national universities tend to demonstrate higher research performance compared to public and private universities. This phenomenon can be attributed to greater government support for national institutions, particularly in terms of research funding and academic infrastructure. Private universities exhibit a wider range of performance variation, with some institutions achieving high performance while the majority are situated in the middle-tier rankings. This suggests disparities in capacity and strategic focus among private institutions. Public universities demonstrate relatively stable performance; however, their research output is not as robust as that of national universities. These findings imply that funding structures and institutional priorities significantly influence academic performance. Furthermore, the analysis reveals that universities with a high degree of internationalization tend to be located in metropolitan areas and exhibit strong research performance, indicating a correlation between internationalization and institutional achievement. Overall, the observed regional and institutional patterns suggest that university performance is more strongly influenced by internal factors such as research quality and graduate employability than by geographical location alone.

3.2 Model Development Results

The evaluation of model performance, as detailed in Table 2, was conducted to assess the capability of each algorithm in predicting target values based on the selected input variables. Performance was measured using three primary metrics: Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and the coefficient of determination (R^2). RMSE quantifies prediction error magnitude by imposing a greater penalty on larger errors, whereas MAE measures the average absolute prediction error. Meanwhile, R^2 evaluates the extent to which the model accounts for the variance observed in the actual data.

Building of Informatics, Technology and Science (BITS)
Volume 7, No 4, March 2026 Page: 2715−2726
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i4.9525

Table 2. Comparison of Model Performance

Model	RMSE	MAE	R^2
Gradient Boosting	1.175.117	1.087.856	0.994988
Random Forest	1.436.536	1.191.515	0.992510
Ridge Regression	2.410.286	2.128.675	0.978913
Lasso Regression	2.495.995	1.969.609	0.977387
Linear Regression	9.713.651	8.459.371	0.657519
Support Vector Regression	13.760.883	12.591.221	0.312671

Based on Table 2, the Gradient Boosting model produced the smallest RMSE value of 1.175117, indicating that this model has the lowest prediction error compared to the other models. A low RMSE value signifies that the differences between the predicted values and the actual values are relatively small, thus the model possesses a high level of accuracy. Random Forest ranked second with an RMSE value of 1.436536, which also demonstrates excellent predictive performance. Conversely, the Linear Regression and Support Vector Regression models produced considerably higher RMSE values, namely 9.713651 and 13.760883, respectively. This suggests that these two models have larger prediction errors and are less capable of capturing the complex relationship patterns within the dataset. This disparity indicates that ensemble-based models, such as Gradient Boosting and Random Forest, are more effective in handling non-linear relationships compared to traditional linear regression models. The MAE value represents the average absolute difference between the predicted values and the actual values. Based on the test results, Gradient Boosting produced the smallest MAE value of 1.087856, indicating that the model's average prediction error is very small. Random Forest also demonstrated excellent performance with an MAE value of 1.191515, which remains within a low error range.

The Ridge Regression and Lasso Regression models yielded slightly higher MAE values, specifically 2.128675 and 1.969609, respectively. This indicates that regularized regression models are capable of producing reasonably good predictions, yet they are still suboptimal compared to the ensemble models. Meanwhile, the Support Vector Regression model produced the largest MAE value of 12.591221, indicating that this model has a high prediction error rate and is less suitable for the dataset used in this study. The coefficient of determination (R^2) is used to measure the model's ability to explain the variation in the actual data. The R^2 value ranges from 0 to 1, where a value closer to 1 indicates a better predictive capability of the model.

Gradient Boosting yielded the highest R^2 value of 0.994988, indicating that this model is able to explain approximately 99.49% of the variance in the actual data. This demonstrates that the model possesses a very high predictive capability and is exceptionally good at capturing the relationship patterns between variables. Random Forest also exhibited excellent performance with an R^2 value of 0.992510. In contrast, the Linear Regression and Support Vector Regression models produced substantially lower R^2 values, namely 0.657519 and 0.312671, respectively. This suggests that these models are less capable of optimally explaining the data variation, resulting in lower predictive performance compared to the ensemble models.

Based on the evaluation results using RMSE, MAE, and R^2, it can be concluded that Gradient Boosting is the best model in this study. This conclusion is supported by its lowest RMSE and MAE values, as well as its highest R^2 value compared to the other models. The superiority of Gradient Boosting is attributed to its ability to build models iteratively by correcting the errors of previous models, thereby enabling it to generate more accurate predictions..
Random Forest also demonstrated excellent performance due to its ensemble approach, which combines numerous decision trees, thereby enhancing prediction stability and accuracy. Meanwhile, linear regression models, such as Linear Regression, showed lower performance due to their limitations in capturing non-linear relationships within the dataset. The findings of this study indicate that ensemble-based models, particularly Gradient Boosting, are highly effective for prediction on the utilized dataset. The high accuracy of the model suggests a strong relationship between the input variables and the target variable.

Furthermore, these results also demonstrate that the application of appropriate machine learning methods can significantly improve prediction accuracy. The Gradient Boosting model can be employed as the primary model in a prediction system due to its low error rate and good generalization capability.

3.2.1 Analysis of Model Stability Using Cross-Validation RMSE

Evaluating model stability is a crucial step to ensure that the developed model not only performs well on specific data but also yields consistent results when applied to different data subsets. In this study, model stability was tested using the K-Fold Cross-Validation method with the Root Mean Square Error (RMSE) as the evaluation metric. The RMSE values obtained from each fold are presented in the following Table 3:

Table 3. Model Stability based on RMSE

Fold	RMSE
Fold 1	12.305
Fold 2	0.9384
Fold 3	13.431

Building of Informatics, Technology and Science (BITS)
Volume 7, No 4, March 2026 Page: 2715−2726
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i4.9525

Fold	RMSE
Fold 4	11.169
Fold 5	0.8938
Mean RMSE	11.045

Based on the cross-validation results presented in Table 3 above, the obtained RMSE values range from 0.8938 to 1.3431. This variation indicates differences in the model's prediction error rate across each data subset. The lowest RMSE value is found in the 5th Fold at 0.8938, signifying that the model has the smallest prediction error on that particular data subset. Conversely, the highest RMSE value is observed in the 3rd Fold at 1.3431, indicating that the model experiences a larger prediction error on that subset. Such differences in RMSE values are a normal occurrence in the cross-validation process, as each fold has a distinct data distribution. However, what is more important is to assess the magnitude of this variation and whether the model demonstrates relatively consistent performance. The obtained average RMSE value is 1.1045, suggesting that overall, the model has a relatively low and stable prediction error rate. This average value reflects the model's general performance when applied to various data subsets. Mathematically, RMSE is calculated using Equation 5. Model stability can be analyzed by examining the difference between the highest and lowest RMSE values:

$$Difference = 1.3431 - 0.8938 = 0.4493$$

This relatively small difference indicates that the model performs consistently across different data subsets. Furthermore, all RMSE values are situated around the average, suggesting that the model does not experience significant overfitting or underfitting. Based on these results, the model employed in this study can be categorized as stable and possessing good generalization ability.

3.2.2 Analysis of Feature Contribution

The analysis of feature contribution aims to determine the extent to which each independent variable influences the model's prediction results. The measurement of feature contribution is conducted using feature importance values, where a larger importance value for a feature signifies a greater influence of that feature on the prediction of the target variable. The results of the feature contribution calculation are presented in Table 4 below.

Table 4. Feature Contribution to the Prediction Model

No	Feature	Importance Value	Contribution Percentage (%)
1	Research_Impact_Score	0.9794417	97.94%
2	Employment_Rate	0.0180538	1.81%
3	Institution_Age	0.0014103	0.14%
4	Intl_Student_Ratio	0.0009943	0.10%
5	Region_Oita	0.0000637	0.006%
6	Region_Tokyo	0.0000362	0.004%
7	Institution_Type_Private	0.0000000547	~0.00%
8	Fitur lainnya	0.0000000	0.00%

Based on the results presented in Table 4, the Research_Impact_Score feature exhibits the highest importance value of 0.9794, equivalent to 97.94%. This indicates that this feature constitutes the most dominant factor influencing the model's prediction outcomes. This exceptionally high contribution value demonstrates that the model heavily relies on the research impact score in determining the target value. Empirically, this finding suggests that the quality and impact of research serve as primary indicators in determining institutional performance or ranking. The second feature demonstrating a significant contribution is Employment_Rate, with an importance value of 0.0180 or 1.81%. Although its contribution is substantially smaller compared to Research_Impact_Score, this feature still exerts influence on the model. Graduate employment rates reflect educational quality and curriculum relevance to industry demands, thereby logically contributing to prediction outcomes. Institution_Age shows an importance value of 0.0014 or 0.14%, indicating that institutional age has a relatively minor influence on prediction results. This finding suggests that institutional age is not a primary factor in determining institutional performance, as younger institutions may still achieve high performance when supported by strong research quality. Furthermore, the Intl_Student_Ratio feature contributes 0.00099 or 0.10%, demonstrating that the international student ratio has only a marginal influence on prediction outcomes. This indicates that the degree of internationalization is not a primary factor in the developed model.

Several regional variables, such as Region_Oita and Region_Tokyo, exhibit extremely low importance values of 0.0000637 and 0.0000362, respectively. These values indicate that institutional geographical location has a negligible influence on model prediction outcomes. Additionally, most other regional variables show importance values of 0.000000, suggesting that these features do not provide significant contributions to the model. This phenomenon may be attributed to several factors, including: regional variables lacking direct relationships with the target variable, information embedded in regional variables being already represented by more robust features, and the model demonstrating greater dependence on numerical features that directly reflect institutional quality.

Building of Informatics, Technology and Science (BITS)
Volume 7, No 4, March 2026 Page: 2715−2726
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i4.9525

The distribution of feature importance values reveals that the model is heavily dominated by a single primary feature, namely Research_Impact_Score. This dominance indicates that the model utilizes this feature as its principal indicator for making predictions. This condition carries two main implications: first, the model possesses clear interpretability as predictions are primarily influenced by one dominant factor; second, the model potentially exhibits high dependence on a single feature, necessitating assurance of this feature's data quality. However, given the previously observed relatively low and stable RMSE values, this feature dominance does not indicate significant problems in model performance.

The feature importance results demonstrate that the model successfully identifies the features most relevant to the target variable. Features with high contributions are conceptually those that have direct relationships with institutional performance. Furthermore, features with low contributions indicate that the model can effectively disregard irrelevant features, thereby becoming more efficient and unaffected by noise from insignificant features.

3.3 Discussion

The research findings indicate that Research_Impact_Score constitutes the dominant factor, contributing 97.94% to university ranking determination, demonstrating that research performance serves as the primary indicator in modern higher education evaluation systems. This finding suggests that universities with high research productivity, significant publication numbers, and strong citation impact tend to achieve better ranking positions. This reflects the global paradigm of higher education oriented toward research-based universities, where research serves as the primary indicator of institutional excellence. Within the Japanese higher education system context, national universities such as the University of Tokyo, Kyoto University, and Osaka University demonstrate exceptionally high research performance, which directly contributes to their superior ranking positions. Furthermore, the Employment_Rate variable contributes 1.81%, indicating that institutional capacity to produce graduates absorbed into the workforce also constitutes an important indicator in university evaluation systems. This suggests that educational quality is measured not only by academic performance but also by graduate success in entering the labor market. High employability rates reflect curriculum relevance, teaching quality, and institutional relationships with industry. The Intl_Student_Ratio variable contributes 0.099%, indicating that internationalization influences university rankings, although its contribution is relatively smaller compared to research performance. Internationalization reflects institutional global reputation, capacity to attract international students, and participation in global academic networks. Universities with high internationalization levels tend to possess enhanced global visibility. The Institution_Age variable contributes 0.14%, indicating that institutional age has limited influence on university rankings. This suggests that younger institutions can still achieve high rankings if they demonstrate strong research performance and employability. Thus, institutional performance is more determined by the quality of academic outputs rather than historical factors. Overall, the research findings demonstrate that university ranking systems reflect multidimensional performance, with research performance constituting the dominant factor, followed by employability and internationalization.

These research findings carry important implications for higher education policymakers, particularly regarding strategic planning for institutional quality enhancement. The dominance of Research_Impact_Score indicates that investment in research constitutes a key factor in improving university ranking positions. Therefore, governments and higher education institutions need to enhance research funding, academic infrastructure, and international collaboration to improve research performance. Additionally, the contribution of Employment_Rate suggests that higher education institutions need to enhance curriculum relevance to industry demands. This can be achieved through industry-based curriculum development, internship programs, and cooperation with the industrial sector. Policies supporting graduate employability enhancement will contribute to overall institutional performance improvement. Findings related to internationalization indicate that higher education institutions need to enhance internationalization strategies through increasing international student numbers, student exchange programs, and international academic collaboration. Internationalization can enhance institutional global reputation and expand academic networks. Furthermore, research results indicate that geographical factors have relatively minor influence compared to academic performance. This suggests that institutions in non-metropolitan areas still have opportunities to improve their rankings through enhanced research and education quality. Consequently, higher education policies need to focus on improving research quality, graduate employability, and internationalization to enhance institutional performance sustainably.

The results of this research align with institutional performance evaluation theory, which posits that higher education organizational performance is determined by academic outputs, graduate quality, and institutional reputation. Within performance-based evaluation theory, research productivity serves as a primary indicator in measuring higher education institutional quality. The predictive model developed in this research demonstrates that academic performance indicators have strong relationships with institutional rankings, supporting the theory that research output constitutes the primary determinant of institutional performance. This indicates that research performance-based university evaluation systems represent valid and relevant approaches. Moreover, the contribution of employability supports human capital theory, which asserts that educational quality can be measured based on graduate success in the labor market. High employability rates indicate that institutions can produce graduates possessing competencies aligned with industry requirements. The contribution of internationalization supports global competitiveness theory, which asserts that higher education institutions need to participate in global academic

Building of Informatics, Technology and Science (BITS)
Volume 7, No 4, March 2026 Page: 2715−2726
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i4.9525

networks to enhance institutional reputation and performance. Thus, these research findings reinforce performance evaluation theory based on multidimensional performance indicators.

The results of this research are consistent with various previous studies demonstrating that research performance constitutes the primary factor in determining university rankings. Previous research has shown that indicators such as publication count, citations, and research impact have strong relationships with university ranking positions. Additionally, previous research has also demonstrated that employability constitutes an important indicator in higher education institutional performance evaluation. Universities with high employability rates tend to possess better reputations and higher ranking positions. Previous research has also shown that internationalization influences university rankings, although its contribution is relatively smaller compared to research performance indicators. Internationalization enhances global reputation and institutional visibility. Furthermore, previous research has demonstrated that machine learning models exhibit better performance compared to traditional statistical models in predicting higher education institutional performance. Machine learning models can capture non-linear relationships among variables, enhancing prediction accuracy. The results of this research reinforce previous findings by demonstrating that Gradient Boosting models exhibit exceptionally high performance in predicting university rankings. This indicates that machine learning approaches constitute effective methods for analyzing higher education institutional performance.

Overall, the research findings demonstrate that research performance constitutes the dominant factor in determining university rankings, followed by employability and internationalization. The developed machine learning model demonstrates exceptionally high predictive performance, indicating that data mining approaches constitute effective methods for higher education institutional evaluation systems. These findings have important implications for higher education policy and the development of data-based institutional performance evaluation systems.

4. CONCLUSION

This research successfully developed a machine learning-based predictive modeling framework to predict the national rankings of 52 top universities in Japan by utilizing multidimensional indicators encompassing research performance, internationalization, employability, institutional characteristics, and geographical factors. Through systematic research stages, commencing from data collection, data preprocessing, feature engineering, model development, model evaluation, and feature importance analysis, this research successfully produced an accurate, stable, and interpretable prediction model. Evaluation results indicate that ensemble-based machine learning models, particularly Random Forest Regression and Gradient Boosting Regression, demonstrate superior performance compared to traditional linear regression models such as Linear Regression, Ridge, and Lasso. This is evidenced by lower RMSE and MAE values and higher coefficients of determination (R^2), indicating the models' capability to capture non-linear and complex relationships among variables. Ensemble models reduce prediction error and enhance generalization capabilities, thus proving more effective in modeling multidimensional university ranking systems. Furthermore, feature importance analysis reveals that research performance and internationalization indicators constitute dominant factors influencing university national rankings, followed by employability and institutional characteristics. These findings confirm that research quality and level of international engagement play strategic roles in determining university competitive positions. Model interpretation analysis also demonstrates that combinations of various indicators provide significant contributions to enhancing prediction accuracy compared to using single indicators. Overall, this research not only produces an accurate predictive model but also provides deeper understanding of factors influencing university rankings. The explainable predictive modeling framework developed can serve as a data-based analytical tool to support higher education institutional performance evaluation, as well as assist policymakers in formulating strategies for enhancing educational quality, research, and global university competitiveness objectively and measurably.

REFERENCES

[1] J. M. Candilasa and K. T. Onahon, “Global University Rankings: Characterization of Higher Education Institution’s Competitiveness,” *Int. J. Res. Innov. Soc. Sci.*, vol. 8, no. 12, pp. 3267–3279, 2024, doi: 10.47772/IJRISS.2024.8120270.

[2] T. Teixeira and C. T. Picinin, “University rankings: Proposal for a future research agenda through a systematic literature review,” *Sustainability*, vol. 16, no. 7, p. 3043, 2024, doi: 10.3390/su16073043.

[3] N. H. Ling, C. J. Chen, C. S. Teh, D. S. John, L. C. Ch’ng, and Y. F. Lay, “Global Trends of Educational Data Mining in Online Learning,” *Int. J. Technol. Educ.*, vol. 6, no. 4, pp. 656–680, 2023, doi: 10.46328/ijte.558.

[4] A. Welch and E. Aziz, “Higher Education in Indonesia,” in *International Handbook on Education in South East Asia*, Springer, Singapore, 2023, pp. 1–30. doi: 10.1007/978-981-16-8136-3_41-2.

[5] L. Prasojo, L. Yuliana, and L. Ary Prihandoko, “Research Performance in Higher Education: A PLS-SEM Analysis of Research Atmosphere, Collaboration, Funding, Competence, and Output, Especially for Science and Engineering Facilities in Indonesian Universities,” *ASEAN J. Sci. Eng.*, vol. 5, no. 1, pp. 123–144, Jan. 2025, doi: 10.17509/ajse.v5i1.81224.

[6] A. R. Dina, N. Alifah, and L. Paz, “Leveraging big data for student success and institutional growth: Memanfaatkan big data untuk kesuksesan mahasiswa dan pertumbuhan institusi,” *J. MENTARI Manajemen, Pendidik. dan Teknol. Inf.*, vol. 3, no. 2, pp. 147–156, 2025, doi: 10.33050/mentari.v3i2.746.

[7] L. J. Wardley, E. Rajabi, S. H. Amin, and M. Ramesh, “A machine learning approach feature to forecast the future

Building of Informatics, Technology and Science (BITS)
Volume 7, No 4, March 2026 Page: 2715−2726
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i4.9525

performance of the universities in canada," *Mach. Learn. with Appl.*, vol. 16, p. 100548, 2024, doi: 10.1016/j.mlwa.2024.100548.

[8] W. Y. Leong, "Leveraging Artificial Intelligence to Predict Future Trends in University Rankings," *Educ. Innov. Emerg. Technol.*, vol. 5, no. 1, pp. 1–11, 2025, doi: 10.35745/eiet2025v05.01.0004.

[9] E. López-Meneses, P. C. Mellado-Moreno, C. Gallardo Herrerías, and N. Pelícano-Piris, "Educational Data Mining and Predictive Modeling in the Age of Artificial Intelligence: An In-Depth Analysis of Research Dynamics," *Computers*, vol. 14, no. 2, p. 68, 2025, doi: 10.3390/computers14020068.

[10] A. Rushiti, A. Luma, Y. Januzaj, A. Aliu, H. Snopçe, and A. Sefidanoski, "The Republic of North Macedonia's Research Ranking Platform for Academic Staff and Universities," *SAR J.*, vol. 7, no. 1, pp. 3–11, 2024, doi: 10.18421/SAR71-01.

[11] A. Yonezawa, "Japan's higher education policies under global challenges," *Asian Econ. Policy Rev.*, vol. 18, no. 2, pp. 220–237, 2023, doi: 10.1111/aepr.12421.

[12] K. C. Gonugunta and K. Leo, "Role of data-driven decision making in enhancing higher education performance: A comprehensive analysis of analytics in institutional management," *Int. J. Acta Inform.*, vol. 3, no. 1, pp. 149–159, 2024, doi: https://www.yuktabpublisher.com/index.php/IJAI/article/view/236.

[13] F. Hu, L. Qiu, S. Wei, H. Zhou, I. A. Bathuure, and H. Hu, "The spatiotemporal evolution of global innovation networks and the changing position of China: a social network analysis based on cooperative patents," *R&D Manag.*, vol. 54, no. 3, pp. 574–589, 2024, doi: 10.1111/radm.12662.

[14] B. Ćudić, P. Alešnik, and D. Hazemali, "Factors impacting university–industry collaboration in European countries," *J. Innov. Entrep.*, vol. 11, no. 1, p. 33, 2022, doi: 10.1186/s13731-022-00226-3.

[15] S. Sihono, M. F. Isbah, and P. Pangestuti, "Komparasi Standar Penilaian Pendidikan di Negara-negara Maju:(Studi Kasus Finlandia, Jepang, dan Singapura)," *Cetta J. Ilmu Pendidik.*, vol. 8, no. 1, pp. 388–401, 2025, doi: 10.37329/cetta.v8i1.3830.

[16] G. Feng, M. Fan, and Y. Chen, "Analysis and prediction of students' academic performance based on educational data mining," *IEEE Access*, vol. 10, pp. 19558–19571, 2022, doi: 10.1109/ACCESS.2022.3151652.

[17] M. Arunkumar, K. Rajkumar, W. R. Jeyaseelan, and N. A. Natraj, "Data Mining, Machine Learning, and Statistical Modeling for Predictive Analytics with Behavioral Big Data," *Teh. Vjesn.*, vol. 32, no. 1, pp. 72–77, 2025, doi: 10.17559/TV-20231102001073.

[18] M. Gul, W. Abbasi, M. Babar, A. Aljohani, and M. Arif, "Data driven decisions in education using a comprehensive machine learning framework for student performance prediction," *Discov. Comput.*, vol. 28, no. 1, Jul. 2025, doi: 10.1007/s10791-025-09585-3.

[19] S. khan *et al.*, "Predictive analytics in education- enhancing student achievement through machine learning," *Soc. Sci. Humanit. Open*, vol. 12, p. 101824, 2025, doi: 10.1016/j.ssaho.2025.101824.

[20] M. Bhushan, U. Verma, C. Garg, and A. Negi, "Machine Learning-Based Academic Result Prediction System," *Int. J. Softw. Innov.*, vol. 12, no. 1, 2024, doi: 10.4018/IJSI.334715.

[21] P. Koukaras and C. Tjortjis, "Data Preprocessing and Feature Engineering for Data Mining: Techniques, Tools, and Best Practices," *AI*, vol. 6, no. 10, p. 257, 2025, doi: 10.3390/ai6100257.