Journal of

Higher Education Policy

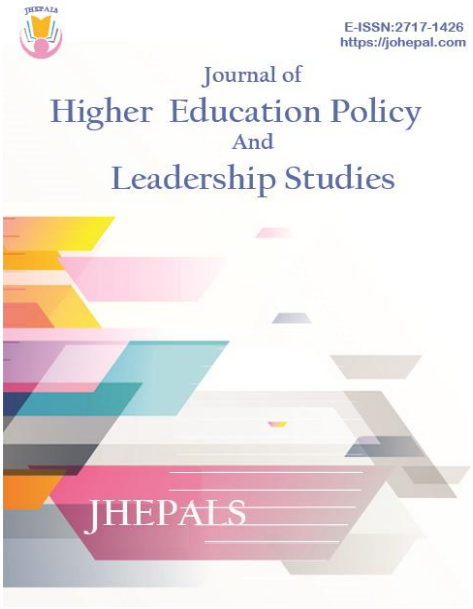
And

Leadership Studies

JHEPALS (E-ISSN: 2717-1426)

https://johepal.com

Prediction of Admission Decisions Using Machine Learning Models: An Analysis of the Holistic Undergraduate Admissions Review Process in Korea



Yousun Shin [*1]
Department of English Education, College of Education, Sunchon National University, SOUTH KOREA
Email: ysshin@scnu.ac.kr
https://orcid.org/0000-0001-5674-8593

Hee Sun Kang [2]
Department of Nursing Science, Sunchon National University, SOUTH KOREA
Email: kanghs@scnu.ac.kr
https://orcid.org/0000-0003-3808-306X

So Yun Park [3]
IR Center, Sunchon National University, SOUTH KOREA
Email: zera6840@scnu.ac.kr
http://orcid.org/0009-0000-7980-915X

Article Received **2026/02/19**

Article Accepted **2026/06/07**

Published Online **2026/06/30**

Cite article as:

Shin, Y., Kang, H. S., & Park, S. Y. (2026). Prediction of admission decisions using machine learning models: An analysis of the holistic undergraduate admissions review process in Korea. *Journal of Higher Education Policy and Leadership Studies*, 7(2), 92-110. https://dx.doi.org/10.66224/johepal.7.2.92

[Downloaded from johepal.com on 2026-11-15]

[DOI: 10.66224/johepal.7.2.92]

Prediction of Admission Decisions Using Machine Learning Models: An Analysis of the Holistic Undergraduate Admissions Review Process in Korea

Journal of Higher Education Policy And Leadership Studies (JHEPALS)

E-ISSN: 2717-1426
Volume: 7 Issue: 2
pp. 92-110
DOI:
10.66224/johepal.7.2.92

Yousun Shin *
Hee Sun Kang
So Yun Park

Abstract

This study aimed to examine and validate the consistency and predictive patterns of human-led undergraduate admissions decisions through the application of machine learning models. Unlike traditional holistic evaluation processes conducted by human assessors, this study compared five machine learning algorithms – Gradient Boosting, Random Forest, Support Vector Machine, Logistic Regression, and XGBoost – to identify the most accurate prediction model. The analysis utilized a dataset of 1,554 application records from the 2024 application cycle. To further improve prediction accuracy, Latent Dirichlet Allocation (LDA) was utilized to extract relevant features from unstructured textual data. The findings revealed that the XGBoost model performed best in predicting admission outcomes. This result is attributed to the learning mechanisms of tree-based ensemble models, which is capable of capturing the complex interactions between non-linear score patterns and various others variables. Major factors influencing admission decisions encompassed interview scores, type of application, and document evaluation scores, highlighting their significance in the selection process and validating the effectiveness of the XGBoost as a supportive tool. These findings not only provide practical recommendations for improving prediction accuracy but also inform future research directions in data-driven strategies for high-stakes educational assessment.

Keywords: Holistic Undergraduate Admissions Review Process; Latent Dirichlet Allocation (LDA); XGBoost; Machine Learning; Educational Data Mining; Prediction Accuracy

*Corresponding author's email: ysshin@scnu.ac.kr

[Downloaded from johepal.com on 2026-11-15]

[DOI: 10.66224/johepal.7.2.92]

Introduction

As Korea has experienced a sharp decline in birth rates, the number of students enrolling in higher education institutions (HEIs) has also dramatically decreased yearly (Anderson & Kohler, 2013; Ma, 2016; Yoo & Sobotka, 2018). This demographic crisis is likely to cause financial instability for universities and widen educational disparities between rural and urban areas. In this educational context, it is crucial for universities to attract potential applicants and support their successful academic progression to ensure institutional stability. High dropout rates and low graduation rates further emphasize the importance of ensuring that admission decisions are consistent and well-aligned with each institution's educational goals and expectations (Zafra & Ventura, 2009). Moreover, intensified competition among universities has increased the need for transparent and systematic admissions decision-making processes to maintain institutional stability (Romero & Ventura, 2007; Jia & Mareboyana, 2013; Yadav et al., 2012).

To address these issues, HEIs must implement rigorous admission standards that assess the academic compatibility between students and institutions (Kotsiantis, 2012). Furthermore, to reduce educational disparities, universities should adopt admission criteria that reliably predict students' future academic success, confirming fair access to higher education (Geiser & Santelices, 2007). Recent advances in technology suggest that machine learning models can serve as a supportive tool to validate complex decision patterns involved in undergraduate admissions decision-making processes, ultimately optimizing decision-making and improving educational outcomes.

Over the past few years, HEIs in Korea have embraced internal selection methods with distinct admission requirements (Kim & Kim, 2024). Despite these efforts, many universities continue to struggle with assigning available spots for first-year students. Applicants often favor programs considered prestigious or superior, leading to difficulty in filling quotas (Zafra & Ventura, 2009). Although researchers have investigated the efficiency of several markers of student preparation, there is no consensus among HEIs regarding the most effective admission criteria (Geiser & Santelices, 2007). Moreover, traditional admission processes often fail to fully capture applicants' potentials, sometimes reinforcing systemic biases. The challenges of predicting student admission success persist due to differing opinions on which indicators should be prioritized as admission criteria (Romero & Ventura, 2007). Consequently, HEIs face increasingly complex admission and registration issues, particularly in managing large-scale datasets (Nghe et al., 2007). To tackle these challenges, developing and empirically testing a comprehensive predictive model is necessary to examine and validate current admissions decision-making processes.

Well-designed admission criteria can increase successful enrollments by matching students with programs that fit their skills, thereby reducing dropout risks and improving degree completion rates (Lakkaraju et al., 2015). Machine learning and data mining should be considered as useful tools to reflect students' needs, minimize academic failure and achieve academic goals, as highlighted by Romero and Ventura (2007). Data mining algorithms can uncover hidden patterns in large datasets, which can lead to improvements in both student outcomes and the reputation of HEIs (Lakkaraju et al., 2015). Predictive machine learning models analyze data to identify trends and make informed predictions about future outcomes (Lantz, 2019).

[Downloaded from johepal.com on 2026-11-15]

[DOI: 10.66224/johepal.7.2.92]

Undergraduate Admission Decisions in South Korea

Existing research has extensively explored the effectiveness of data mining and machine learning applications in HEIs (Lantz, 2019; Lakkaraju et al., 2015; Romero & Ventura, 2007). However, few studies have examined how these tools are applied to examine and validate the undergraduate admissions decisions in South Korea's unique educational context, where demographic shifts and societal expectations place increasing pressure on HEIs (i.e., Jo, 2018; Kim, 2024; Kim & Kim, 2024). Therefore, this current study aimed to examine whether machine learning models can effectively analyze and replicate decision patterns in holistic undergraduate admissions, and to identify the key factors influencing admission decision-making. Specifically, this study focused on following research questions:

1. Which machine learning model, among Gradient Boosting, Random Forest, Support Vector Machine (SVM), Logistic Regression, and XGBoost, most effectively captures and replicates admission decision patterns?
2. What are the key factors influencing admission decisions?

Literature Review

Artificial Intelligence (AI) and Machine Learning (ML)

Technologies powered by artificial intelligence (AI), including data mining or ML, are increasingly utilized at universities for a variety of academic analyses. These educational technologies serve multiple purposes, such as predicting admission outcomes and student performance, as well as identifying at-risk students. For example, Ekowo and Palmer (2016) demonstrated the potential of AI-driven analytics in improving admission decisions and student support systems. They concluded that AI technologies could predict admission outcomes and students' academic performance with high accuracy.

Data mining refers to the process of extracting valuable insights from large datasets (Hussain et al., 2019; Taub & Azevedo, 2018). Dataset have accumulated over time within a institutional system, containing latent information which can be extracted to support evidence-based decision-making. In the educational context, educational data mining uniquely focuses on analyzing institutional data to identify learning challenges and academic performance trends among learners (Al-Alawi et al., 2023; Bucos & Drăgulescu, 2018), thus supporting data-driven decision-making (Baker et al., 2009; Bharara et al., 2018; Tair & El-Halees, 2012). As one of the key technologies in this area, Latent Dirichlet Allocation (LDA) – a widely used probabilistic topic modeling approach – is often applied to efficiently create and fit topic models to large e-learning corpus, enabling the analysis of huge amounts of textual data (Blei et al., 2003; Blei & Lafferty, 2007; Hussain et al., 2019). LDA was employed in this study due to institutional constraint regarding labeled training data and the need for interpretable topic structure to facilitate subsequent administrative reporting.

ML, on the other hand, involves the development of algorithms which enables computers to learn patterns from data with minimal explicit rule-based programming, hence enhancing machine intelligence while maintaining human oversight (Arora, 2024). ML has been extensively applied in HEIs to enhance various academic and administrative processes. These applications include estimating students' academic performance, evaluate learning practices, and increase administrative efficiency (Al-Alawi et al., 2023; Altabrawee et al., 2019; Rastrollo-Guerrero et al., 2020). Specifically, ML has been utilized to examine

[Downloaded from johepal.com on 2026-11-15]

[DOI: 10.66224/johepal.7.2.92]

academic performance and student achievement (Altabrawee et al., 2019; Fernandes et al. 2019; Rastrollo-Guerrero et al., 2020), predict dropout and graduation potential (Ahuja & Kankane, 2017), evaluate learning processes (Khan et al., 2020), and identify learning risks by analyzing students' textual feedback (Ibrahim, 2023). Although ML and data mining share similarities, their objectives differ: ML focuses on instructing machines to learn from defined parameters while data mining seeks to identify patterns or rules within large datasets (Arora, 2024). In summary, data mining and ML provide a methodological foundation for analyzing how complex institutional decisions are structured and for validating whether such decision patterns can be reliably replicated through data-driven models.

Comparison of ML Model Performance

This study focused on ML techniques to enhance prediction accuracy in the holistic undergraduate admissions review process. Specifically, the study evaluated models such as Gradient Boosting, Random Forest, Support Vector Machine (SVM), Logistic Regression, and XGBoost. These models' performance varies depending on data characteristics, selected factors, and evaluation criteria. Among these, XGBoost has emerged as one of the top-performing models for admission prediction tasks, particularly for complex datasets with high dimensionality (Chen & Guestrin, 2016). By handling non-linear relationships and interaction effects, XGBoost is particularly suitable for capturing the multi-dimensional criteria present in the holistic undergraduate admissions processes (Breiman, 2001; Hastie et al., 2009). However, its computational complexity may pose challenges for large datasets. Random Forest, an ensemble method known for interpretability and accuracy, often provides more accessible feature importance metrics than boosting algorithms. While it excels at managing structured data and features such as academic performance, interviews, or extracurricular activities, and efficiently handles overfitting, Random Forest may not perform as well as XGBoost in high-dimensional or complex datasets.

Logistic Regression is widely used as a baseline model due to its simplicity and interpretability, but it often fails to capture the complex, non-linear relationships common in holistic admissions data (Hosmer et al., 2013). Gradient Boosting, like XGBoost, delivers high accuracy and is effective in reducing bias and variance, particularly with balanced datasets. However, it can be less efficient on very large or high-dimensional data. Support Vector Machine (SVM) performs well when margins are clear and noise is minimal, but its effectiveness decreases in noisy, multi-dimensional admissions datasets compared to ensemble methods such as XGBoost or Random Forest (Obsie & Adem, 2018; Smola & Schölkopf, 2004). Because holistic undergraduate admissions involve nonlinear and interactive evaluation structures, comparing different ML models is essential to determine which approach most effectively captures these underlying decision patterns.

Previous Studies on ML Model Performance

Prior studies have shown that ML models effectively predict student enrollment success in university admissions (Alghamdi et al., 2020; Maulana et al., 2023; Mengash, 2020; Raghavendran et al., 2021; Walid et al., 2022;). For instance, Wu et al. (2023) proposed a method for admission committees to identify and select suitable applicants. Using a logistic regression analysis and publicly available datasets, they evaluated model performance through a confusion matrix, comparing predicted and actual data. Their analysis achieved

[Downloaded from johepal.com on 2026-11-15]

[DOI: 10.66224/johepal.7.2.92]

an 80% correlation between predicted and actual admission outcomes. The authors concluded that their approach could improve enrollment strategies by enabling more accurate applicant selection. Maulana et al. (2023) explored ML algorithms to address the limitations and systematic biases of conventional admission policies. They compared several ML models such as K-Nearest Neighbors, Random Forest, SVM, and XGBoost. Random Forest demonstrated the highest performance in terms of accuracy, consistency, and reliability. Notably, cumulative GPA emerged as the most influential predictor, with a feature importance score of 0.80, emphasizing the key role of academic performance in admissions. These results highlighted the potential of Random Forest in optimizing the admission decisions.

Similarly, Walid et al. (2022) examined six ML models using a realistic dataset and evaluated performance metrics including the Area Under the Curve (AUC), Precision, Recall, and F-Measure. They demonstrated the scalability of ML methods for international student admissions by utilizing diverse and multifaceted datasets. Their findings revealed that combining the Support Vector Machine (SVM) model with advanced resampling techniques, such as borderline SVM-based SMOTE, improved prediction accuracy by effectively handling imbalanced datasets.

Mengash (2020) demonstrated how ML algorithms and data mining techniques predicted applicants' academic success prior to admission. Using data from 2,039 students enrolled in a university between 2016 and 2019, the study found that Artificial Neural Networks (ANN) achieved a 79.22% accuracy rate in predicting academic performance. The study also identified the Scholastic Achievement Admission Test (SAAT) as the most reliable predictor of academic success, recommending its use in admission criteria. The study validated the effectiveness of prediction modeling in HEIs, suggesting that admission committees could use these models to optimize the allocation of limited institutional resources. In a similar context, Alghamdi et al. (2020) compared Linear Regression, Decision Tree, and Logistic Regression models to determine the most effective model for predicting graduate admission. The finding revealed that Logistic Regression demonstrated the highest precision, with the lowest error rate (7.2%).

Alyahyan and Düştegör (2020) conducted a comprehensive review of factors influencing academic success in universities, including academic and non-academic variables into predictive models. Their findings showed that ensemble models outperformed the conventional statistical methods. They also confirmed that cognitive factors, such as high school GPA and standardized test results, along with non-cognitive factors, such as motivation, were identified as key determinants of academic success. In a similar vein, Yağcı (2022) applied several algorithms, including Random Forests, K-Nearest Neighbors, Logistic Regression, SVM, and Naïve Bayes to predict students' final grades. Using data from 1,854 students at a university in Türkiye, their model achieved 70-75% accuracy, emphasizing the importance of data-driven approaches for early identification of at-risk students and timely interventions.

Despite these successes, challenges still persist in using ML models for admission decisions including model overfitting issues, and classification bias. Advanced models, such as Random Forest and Neural Networks, particularly prone to overfitting, while fluctuations in traditional admission records can cause classification bias. These challenges highlight the

[Downloaded from johepal.com on 2026-11-15]

[DOI: 10.66224/johepal.7.2.92]

need for more generalized evaluation criteria and multidisciplinary datasets to enhance model validity and reliability. Namoun and Alshanqiti (2020) noted that previous studies often relied on limited datasets from specific universities, which reduced generalizability of findings, though ML models demonstrated great potential for predicting university admissions and evaluate students' academic performance. Although research on model performance has been steadily accumulating, the application of these models to analyze and validate admission decision patterns within Korean university admissions remains limited. To address this gap, this study aims to examine and validate admission decisions using diverse datasets and various ML models. Our ultimate goal is to develop robust, replicable predictive models that offer practical insights for improving university admissions processes and policies. Rather than proposing that automated decision-making should replace the human evaluation, this study positions ML as a supportive tool designed to complement human-led holistic review admissions.

Research Methodology

Dataset

This study analyzed data collected from applicants for the 2024 academic year as part of the holistic undergraduate admissions review process at a regional university in Jeollanam-do, South Korea. The dataset was anonymized prior to analysis, and the study received approval from the Institutional Review Board (IRB) at the authors' institution.

A total of 1,554 student records were used in the final dataset. A total of 30 attributes extracted from application materials were used as the major inputs for the analysis, as shown in Table 1. These variables were categorized into two types of data: Structured data (binary, categorical, and numerical) and unstructured data (textual data including teacher comment or narrative records). The unstructured textual data were preprocessed and analyzed using Latent Dirichlet Allocation (LDA), a topic modeling method. Specifically, the variables include such as age, gender, type of high school, school size, residence area, academic major, the aggregate academic records, teachers' comprehensive evaluation comments (including academic activity, extracurricular activity, overall behavioral characteristics), a set of documentation evaluation scores across four areas (i.e., academic competence, major fit, potential for development, and community competence), along with a set of interview scores.

Table 1.
List of Features Used in Analysis

Category	Type	Details	
Structured Data	Binary	· Eligibility Status · Acceptance Status	
	Categorical	· Application type Applied College · Academic track · Recruitment Unit · Gender	· Province · City/district · High School Name · High School Type
	Numerical	· Number of Classes · Number of Students	

[Downloaded from johepal.com on 2026-11-15]

[DOI: 10.66224/johepal.7.2.92]

		· Converted Academic Scores · Document Evaluation: *Total average scores, Academic Competency, Major Suitability, Potential for Development, Community Competency* Interview Evaluation: *Total Average, Academic Competency, Major Suitability, Potential for Development Average, Community Competency*
Unstructured Data	Textual	· Creative Experiential Activity Record: *Autonomous Activities, Club Activities, Career Activities* · Behavior Characteristics and Comprehensive Opinion · Subject-Specific Details: *General Subjects, Arts and Physical Education Subjects*

The initial stage for admissions process involves reviewing and evaluating all submitted application materials. In the second stage, applicants who pass the initial review are invited for an interview. The final decision on acceptance or rejection is made by the admissions committee, based on the combined scores from both stages (Bornmann et al., 2006; Young et al., 2022).

Preprocessing Pipeline for LDA Topic Modeling of Unstructured Data

In this study, unstructured data consisted of teachers' comments and narrative descriptions of academic and extracurricular activities (e.g., academic/extracurricular activity, overall behavioral characteristics). To convert unstructured texts into analyzable data, the preprocessing stages consisted of the following steps:

a. **Text consolidation:** Various types of textual data, originally distributed across multiple columns, were consolidated into a single text corpus per applicant to ensure consistency in the downstream analysis (Siino et al., 2024).
b. **Lowercasing and punctuation removal:** The consolidated texts were converted to lowercase, and extraneous elements such as special characters and numbers were removed to prepare the data for further analysis (Wang et al., 2019).
c. **Stopword removal:** A custom stopword list (including both Korean and English stopwords such as "and," "but," etc.) was created to filter out commonly used words that add little analytical value. This is a standard noise-reducing step in text preprocessing (Bird et al., 2009; Kaur & Buttar, 2018; Pradana & Hayaty, 2019).
d. **Stemming:** Stemming was applied to standardize different word forms by reducing them to their root. For instance, terms like "participate," "participating," and "participated" were all converted to "participate," which helps to maintain consistency in the analysis (Ibrahim, 2023; Singh & Gupta, 2017).
e. **Tokenization:** The processed texts were tokenized into individual terms prior to vectorization.

Following preprocessing, a bag-of-words representation using count vectorization was constructed to create the document-term matrix required for LDA modeling. Count vectorization calculates word frequencies within each document and creates structured representations suitable for topic modeling (Yang et al., 2022). Latent Dirichlet Allocation (LDA) was implemented using the Gensim library with the following parameters: number of topics = 5, Alpha = symmetric, Eta = symmetric, iterations = 1,000, random_state = 42 (Baker & Yacef, 2009; Blei et al., 2003; Prihatini et al, 2018). Topic coherence scores were computed

[Downloaded from johepal.com on 2026-11-15]

[DOI: 10.66224/johepal.7.2.92]

to validate the optimal number of topics, ensuring that each extracted topic was meaningful and interpretable.

For each applicant, the probability distributions across the five topics were extracted from the LDA output and used as continuous input features in the predictive models (Prihatini et al., 2018). In Table 2, these topics included Overall Academic Achievement, Extracurricular Activities and Leadership, Performance-based Academic Activity, and Content-based Academic Activity, and Miscellaneous Factors such as Communicative Competence in English.

Table 2.
The Result of the Top 5 Topic Extraction Using LDA

Topic	Topic Classification	Top Keywords
Topic 1	Overall Academic Achievement (GPA)	Present, Understanding, Being, Appearance, Time, Explain, In Class, Participate
Topic 2	Extracurricular Activities and Leadership	In Activity, Content, Career Path, Together, Role, Become Aware, Participate
Topic 3	Performance-based Academic Activity	Music, Participation, Musical, Understanding, Opinion, Work, Composer, In Work, Activity, Actively
Topic 4	Content-based Academic Activity	Presentation, In Class, Social, World, Economy, Exploration, Small, Opinion, Family
Topic 5	Communicative Competence in English	English, Utilize, Have, Work, Literature, Effort, As a Student, Ability

These topic probability features were concatenated with structured variables and used as a model input (Bujang et al., 2021; Chen & Guestrin, 2016; Mengash, 2020)

Evaluation Metrics

Model performance was evaluated using accuracy, precision, recall, F1-scores, and AUC scores (Bowers & Zhou, 2019). Accuracy was calculated as (TN + TP) / (FN + TN + TP + FP). Precision and recall were calculated as TP / (TP + FP) and TP / (TP + FN), respectively. Two additional evaluation metrics are the F1 score and AUC (Area Under the Curve) score. The F1 score, calculated as the harmonic mean of precision and recall, provides a balanced evaluation when there is a trade-off between these two metrics (Chicco & Juman, 2020). The AUC score, derived from the Receiver Operating Characteristic (ROC) curve, measures the model's ability to discriminate between different classes. A higher AUC score indicates better overall performance in identifying true positives while minimizing false positives.

Model Training

In this study, five ML models – Gradient Boosting, Random Forest, XGBoost, Logistic Regression, and Support Vector Machine (SVM) – were selected to predict admission outcomes, implemented using the scikit-learn library for both training and evaluation. The dataset was divided into training and test sets using an 80:20 split (1,243 samples for training and 311 for testing), using stratified random sampling to preserve the class distribution of admitted and non-admitted applicants in both sets. The random state was fixed as 42 for reproducibility. Model performance was evaluated on the test set using predefined metrics. Prior to testing, five-fold stratified cross-validation with grid search was performed on the training set to conduct systematic hyperparameter tuning for each algorithm.

[Downloaded from johepal.com on 2026-11-15]

[DOI: 10.66224/johepal.7.2.92]

Data Analysis Procedure

Figure 1 below illustrates the overall data analysis procedure combining both structured and unstructured data. Unstructured data, including teacher comments, were preprocessed and underwent LDA topic modeling. The resulting topic features were then integrated with the preprocessed structured data to construct the final feature set for model training. The target variable was "Acceptance status" (binary: admitted=1, not admitted=0). This variable was excluded from input features in order to prevent data leakage. Prior to model training, the class distribution was examined and found that the numbers of admitted and non-admitted applicants were fairly balanced. As the data was relatively balanced, no resampling methods were applied. Missing values accounted for less than 3% of the dataset, and were imputed using the mean for numerical features and the mode for categorical features.

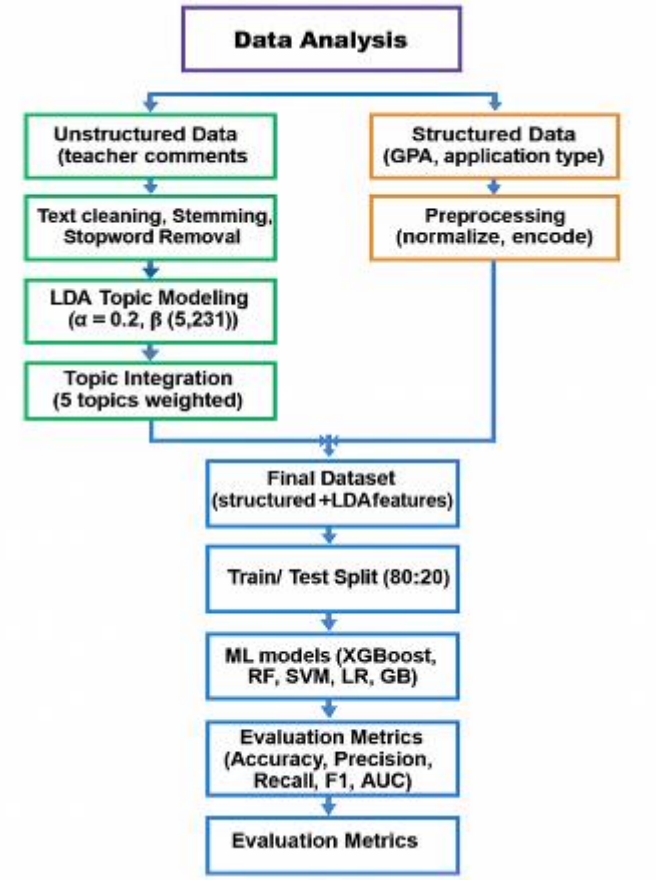


Figure 1. Data Analysis Procedure

Results & Discussions

Predictive Accuracy of ML Models

Table 3 summarizes the performance results of the five ML models focusing on their effectiveness.

[Downloaded from johepal.com on 2026-11-15]

[DOI: 10.66224/johepal.7.2.92]

Table 3.
Performance Comparison for the ML Models

Model	Accuracy	Precision	Recall	F1-score	AUC
Gradient Boosting	0.865	0.869	0.865	0.865	0.95
Random Forest	0.850	0.807	0.805	0.806	0.91
XGBoost	0.871	0.876	0.871	0.872	0.95
Logistic Regression	0.865	0.871	0.865	0.865	0.95
SVM	0.571	0.326	0.571	0.415	0.86

As presented in Table 3, XGBoost achieved the highest accuracy at 87.1%, followed by Logistic Regression and Gradient Boosting at 86.5%, Random Forest at 85.5%, and SVM at 57.1%. In addition to accuracy, XGBoost also demonstrated the highest precision (87.6%), recall (87.1%), and F1 score (87.2 %), indicating stable performance across multiple evaluation metrics. Similarly, AUC values for all models exceeded 0.85, with XGBoost and Gradient Boosting obtaining the highest AUC score (0.95), suggesting strong discriminative ability between admitted and non-admitted applicants. Figure 2 provides a visual comparison of model accuracy.

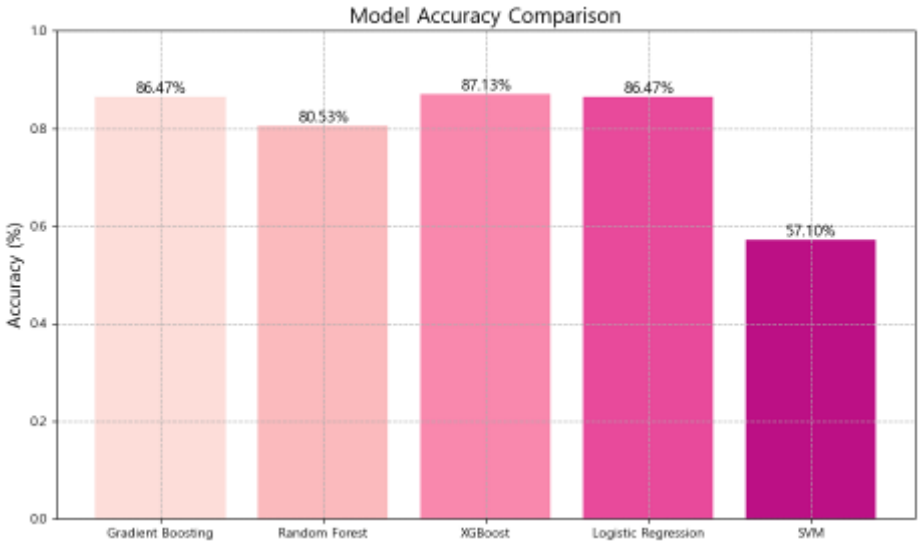


Figure 2. Model Accuracy Comparison

Tree-based ensemble models such as Random Forest, Gradient Boosting, and XGBoost, which evaluate the importance of each feature in the model's predictions, are particularly effective in capturing non-linear relationships within datasets. In contrast, linear models such as Logistic Regression and the SVM when used with the specific kernel or parameter settings adopted in this study, performed relatively poorly in capturing the non-linear relationships present in features derived from unstructured data.

[Downloaded from johepal.com on 2026-11-15]

[DOI: 10.66224/johepal.7.2.92]

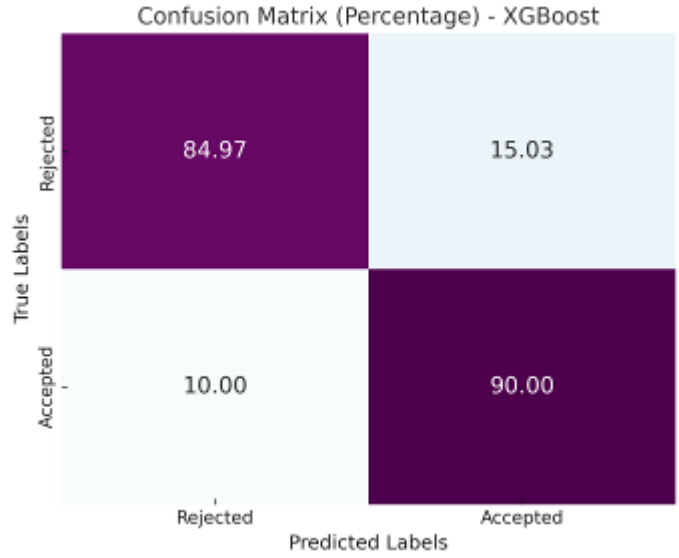


Figure 3. Confusion Matrix of XGBoost Model

Due to its outstanding performance among all evaluation metrics, XGBoost was selected for subsequent analysis. Figure 3 displays the confusion matrix for the XGBoost model, with a focus on its performance in predicting admission outcomes. The results are expressed as percentages. The model accurately predicted 90.0 % of students who were admitted (true positive) and 85.0 % of students who were not admitted (true negative). However, it also misclassified 15.0 % of students who were not admitted, predicting them as admitted (false positive), and 10.0 % of admitted students as not admitted (false negative). These results indicate that, despite the model's high overall predictive accuracy, classification errors persist, particularly in distinguishing borderline cases.

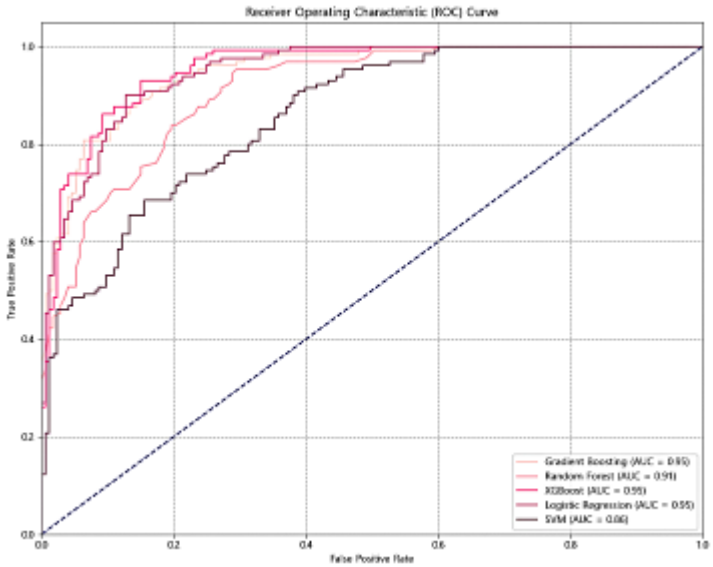


Figure 4. AUC Score Comparison

Figure 4 shows that all models achieved an AUC value greater than 0.85. This indicates strong performance in distinguishing between admitted and non-admitted applicants. Both XGBoost and Gradient Boosting achieved an AUC score of 0.95, consistent with prior

[Downloaded from johepal.com on 2026-11-15]

[DOI: 10.66224/johepal.7.2.92]

research that highlights the high AUC scores obtained by tree-based models in educational datasets (Deist et al., 2018; Ling et al., 2003; Ojajuni et al., 2021; Sahin, 2020; Walid et al., 2022).

Interpretation of Model Performance

The strong performance of XGBoost, Gradient Boosting, and Random Forest can be attributed to their tree-based ensemble models, which are capable of modeling nonlinear relationships and complex interactions among features. On the contrary, Logistic Regression relies on linear decision boundaries, which limits its effectiveness when dealing with topic-derived textual features that often demonstrate more complex patterns. Although SVM can capture non-linear patterns through the appropriate choice of kernel function, its lower performance in this study suggests that the data representation used here was not ideal for this method.

Influencing Factors in the Decision-Making Process

Feature importance analysis using XGBoost, which demonstrated superior predictive performance, was conducted to identify the five most influential variables among the thirty input features. These selected variables were further examined to assess their specific impact on admission decisions. Figure 5 illustrates the relative importance score of the top five features.

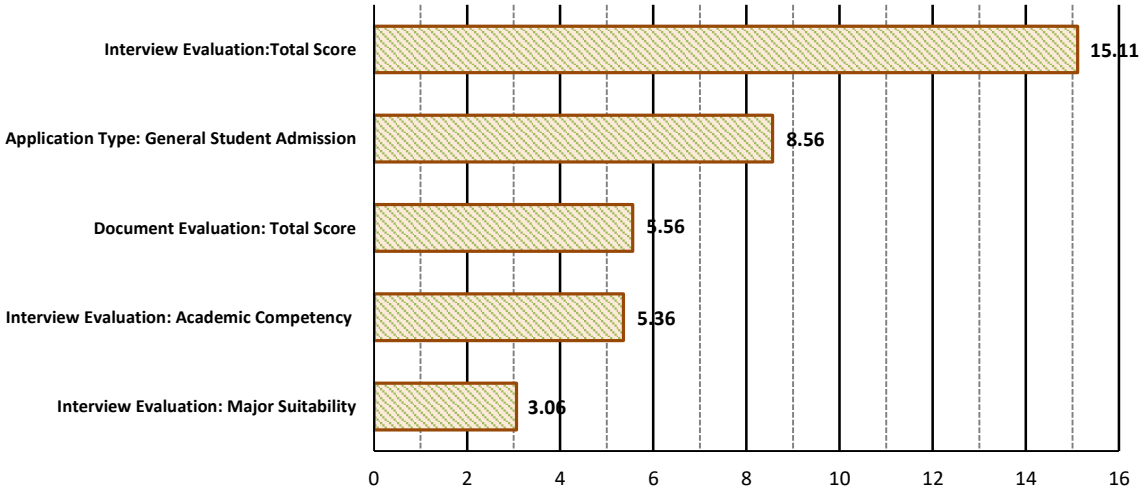


Figure 5. Top 5 Features Influencing Admission Decisions

Among these features, 'Interview Evaluation: Total Score,' emerged as the most influential variable, suggesting that interview-based evaluations play a substantial role in final admission decisions (Bastedo et al., 2018; Hossler et al., 2019). 'Application Type: General Student Admission' was also found to be a significant feature, indicating that the type of admission application may affect the decision-making process, likely because evaluation criteria differ among various application channels. Next, 'Document Evaluation: Total Score' underscores the role of academic and extracurricular achievements. This finding is supported by Hossler et al. (2019) research on comprehensive evaluations in the holistic admissions review process. Furthermore, 'Interview Evaluation: Academic Competency'

[Downloaded from johepal.com on 2026-11-15]

[DOI: 10.66224/johepal.7.2.92]

reflects applicants' readiness for the academic challenges of university life, while 'Interview Evaluation: Major Suitability' highlights aligning applicants' profiles and interests with their intended field of study, ultimately supporting improved long-term academic outcomes. Overall, the findings demonstrate that both academic performance in document-based evaluations and interview-based assessments (competency and suitability, in particular) influence holistic admission decisions.

Conclusion

This study applied various ML models to predict university admission outcomes. The analysis found that tree-based ML models generally demonstrated higher predictive power than traditional statistical models, which suggests that these models can effectively capture nonlinear relationships in the admission context where complex academic achievement indicators and background variables interact (Alghamdi et al., 2020; Xu, 2024).

Notably, the integration of Latent Dirichlet Allocation (LDA) with ML models enabled the inclusion of unstructured data into the analysis process. This approach facilitated the identification of non-academic factors such as field-specific interests and extracurricular activities alongside academic records (e.g., GPA) as significant contributors to admission decisions (Mengash, 2020). This study demonstrated that combining unstructured data enables the discovery of latent patterns that conventional methods might overlook. This finding highlights the importance of non-academic factors in admission decision-making, which aligns with the foundational rationale of the holistic undergraduate admissions review process.

The results of the study have several implications for HEIs. First, ensemble methods like XGBoost and Random Forest demonstrated superior predictive performance for admission outcomes. By employing advanced predictive models, HEIs can support admission evaluation processes while streamlining committee workflows. This extension beyond conventional variables addresses a critical gap in prior studies.

To maximize the impact of the study's findings, HEIs may consider applying these models as decision-support tools in real-world admission processes. For example, the admissions office could integrate factors such as interview scores and extracurricular activities into systematic evaluation tools, enabling a more comprehensive and focused review of shortlisted applicants. HEIs with limited evaluation resources may consider adopting AI-driven technologies to support evidence-based decision-making. Beyond enrollment accuracy, these models may support procedural consistency (Chen & Guestrin, 2016; Posselt, 2016).

However, this study makes it clear that improved predictive performance does not immediately justify fully automated admission decisions. Analyzing variable importance and interpretability helps to understand the relationships between the internal structure of models and the factors contributing to predictions, but this analysis is not intended to mechanically replace individual applicant judgments.

At the same time, this study acknowledges several limitations. First, the dataset used in this study was derived from a single regional university in South Korea, which may limit the generalizability of the findings to institutions with more diverse demographics and academic contexts. Second, external factors such as applicants' socioeconomic status or

[Downloaded from johepal.com on 2026-11-15]

[DOI: 10.66224/johepal.7.2.92]

geographic location, which might influence admission outcomes, were not considered in this study. Lastly, the complexity of advanced ML models could also pose challenges for HEIs, particularly those lacking sufficient technical expertise required for their implementation.

These limitations present opportunities for future research. For example, real-time data updates and adaptive learning mechanisms in ML models could improve their responsiveness to changes in institutional priorities or applicant pools. Conducting longitudinal studies on student success after admission would provide deeper insights into consequential validity and effectiveness of these predictive models. Refining predictive models in this manner would enhance their applicability to dynamic educational environments. However, ethical safeguards and institutional accountability must remain central to the implementation of these advanced technologies.

Declaration of Conflicting Interests

The authors have no conflicting interests to be cited here.

Funding

This research received no external funding.

Human Participants

This study was approved by the Institutional Review Board (IRB) at Sunchon National University (IRB Approval No: 1040173-202502-HR-007-02). The student data used fall within the scope of consent provided during the application process, which can be interpreted as including permission for research use and the publication of anonymized case details.

Originality Note

The authors confirm that this manuscript is their original work. The data will be available upon request.

Use of Generative AI/ AI-assisted Technologies Statement

During the revision of this work, the authors used [ChatGPT] in all sections for refining and polishing writing. After using this tool, the authors reviewed and edited the content as needed, taking full responsibility for the content of the publication.

[Downloaded from johepal.com on 2026-11-15]

[DOI: 10.66224/johepal.7.2.92]

References

Ahuja, R., & Kankane, Y. (2017). Predicting the probability of student's degree completion by using different data mining techniques. In *2017 Fourth International Conference on Image Information Processing (ICIIP)* (pp. 1-4). IEEE. https://doi.org/10.1109/ICIIP.2017.8313763

Al-Alawi, L., AL Shaqsi, J., Tarhini, A., & AL-Busaidi, A. S. (2023). Using machine learning to predict factors affecting academic performance: The case of college students on academic probation. *Education and Information Technologies, 28*(10), 12407-12432. https://doi.org/10.1007/s10639-023-11700-0

Alghamdi, A., Barsheed, A., AlMshjary, H., & AlGhamdi, H. (2020). A machine learning approach for graduate admission prediction. In *Proceedings of the 2020 2nd International Conference on Image, Video and Signal Processing* (pp. 155-158). https://doi.org/10.1145/3388818.3393716

Altabrawee, H., Ali, O. A. J., & Ajmi, S. Q. (2019). Predicting students' performance using machine learning techniques. *Journal of University of Babylon for Pure and Applied Sciences, 27*(1), 194-205. https://doi.org/10.29196/jubpas.v27i1.2108

Alyahyan, E., & Düştegör, D. (2020). Predicting academic success in higher education: Literature review and best practices. *International Journal of Educational Technology in Higher Education, 17*(1), 3-24. https://doi.org/10.1186/s41239-020-0177-7

Anderson, T., & Kohler, H. –P. (2013). Education fever and the east Asian fertility puzzle: A case study of low fertility in South Korea. *Asian Population Studies, 9*(2), 196-215. https://doi.org/10.1080/17441730.2013.797293

Arora, S. (2024, August 14). Data mining Vs. machine learning: The key difference. *Simplilearn.* https://www.simplilearn.com/data-mining-vs-machine-learning-article

Baker, R. S. J. D., Corbett, A. T., ROLL, I., & Koedinger, K. R. (2009). Developing a generalizable detector of when students game the system. *User Modeling and User-Adapted Interaction, 18*(3), 287-314. https://doi.org/10.1007/s11257-007-9045-6

Baker, R. S. J. D., & Yacef, K. (2009). The state of educational data mining in 2009: A review and future visions. *Journal of Educational Data Mining, 1*(1), 3-17. https://doi.org/10.5281/zenodo.3554658

Bharara, S., Sabitha, S., & Bansal, A. (2018). Application of learning analytics using clustering data mining for students' disposition analysis. *Education and Information Technologies, 23*(2), 957-984. https://doi.org/10.1007/s10639-017-9645-7

Bird, S., Klein, E., & Loper, E. (2009). *Natural language processing with Python: Analyzing text with the natural language toolkit*. O'Reilly Media, Inc.

Blei, D. M., & Lafferty, J. D. (2007). A correlated topic model of science. *The Annals of Applied Statistics, 1*(1), 17-35. https://doi.org/10.1214/07-AOAS114

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research, 3*, 993-1022. https://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf

Bornmann, L., Mittag, S., & Danie, H. -D. (2006). Quality assurance in higher education – meta-evaluation of multi-stage evaluation procedures in Germany. *Higher Education, 52*(4), 687-709. https://doi.org/10.1007/s10734-004-8306-0

Bowers, A. J., & Zhou, X. (2019). Receiver operating characteristic (ROC) area under the curve (AUC): A diagnostic measure for evaluating the accuracy of predictors of education outcomes. *Journal of Education for Students Placed at Risk (JESPAR), 24*(1), 20-46. https://doi.org/10.1080/10824669.2018.1523734

[Downloaded from johepal.com on 2026-11-15]

[DOI: 10.66224/johepal.7.2.92]

Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5-32. https://doi.org/10.1023/A:1010933404324

Bucos, M., & Drăgulescu, B. (2018). Predicting student success using data generated in traditional educational environments. *TEM Journal, 7*(3), 617-625. https://doi.org/10.18421/TEM73-19

Bujang, S. D. A., Selamat, A., Ibrahim, R., Krejcar, O., Herrera-Viedma, E., Fujita, H., & Ghani, N. A. Md. (2021). Multiclass prediction model for student grade prediction using machine learning. *IEEE Access, 9*, 95608–95621. https://doi.org/10.1109/ACCESS.2021.3093563

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785-794). https://doi.org/10.1145/2939672.2939785

Chen, X., Zou, D., Cheng, G., & Xie, H. (2020). Detecting latent topics and trends in educational technologies over four decades using structural topic modeling: A retrospective of all volumes of Computers & Education. *Computers & Education, 151*, 103855. https://doi.org/10.1016/j.compedu.2020.103855

Chicco, D., & Juman, G. (2020). The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics***,** *21*(1), 6. https://doi.org/10.1186/s12864-019-6413-7

Ekowo, M., & Palmer, I. (2016, October 24). The promise and peril of predictive analytics in higher education: A landscape analysis. *New America*. https://www.newamerica.org/education-policy/policy-papers/promise-and-peril-predictive-analytics-higher-education/

Fernandes, E., Holanda, M., Victorino, M., Borges, V., Carvalho, R., & Van Erven, G. (2019). Educational data mining: Predictive analysis of academic performance of public school students in the capital of Brazil. *Journal of Business Research*, *94*, 335-343. https://doi.org/10.1016/j.jbusres.2018.02.012

Geiser, S., & Santelices, M. V. (2007). Validity of high-school grades in predicting student success beyond the freshman year: High-school record vs. standardized tests as indicators of four-year college outcomes. *Research and Occasional Papers Series*. Center for Studies in Higher Education. https://cshe.berkeley.edu/sites/default/files/publications/rops.geiser._sat_6.13.07.pdf

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction*. Springer.

Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied logistic regression*. John Wiley & Sons.

Hossler, D., Chung, E., Kwon, J., Lucido, J., Bowman, N., & Bastedo, M. (2019). A study of the use of nonacademic factors in holistic undergraduate admissions reviews. *The Journal of Higher Education, 90*(6), 833-859. https://doi.org/10.1080/00221546.2019.1574694

Hussain, M., Zhu, W., Zhang, W., & Abidi, S. M. R. (2019). Student engagement predictions in an e-learning system and their impact on student course assessment scores. *Computational Intelligence and Neuroscience, 9*(4), 1-21. https://doi.org/10.1155/2018/6347186

Ibrahim, Z. M. (2023). *Text mining framework for detecting assessment and feedback issues using students' evaluation surveys,* [Doctoral dissertation*,* University of Portsmouth].

Jia, J. W., & Mareboyana, M. (2013). Machine learning algorithms and predictive models for undergraduate student retention. In *Proceedings of the World Congress on Engineering and Computer Science* (pp. 23-25). International Association of Engineers.

Jo, H. (2018). Changes and challenges in the rise of mass higher education in Korea. In A. Wu., & J. Hawkins (Eds.), *Higher education in Asia: Quality, excellence and governance* (pp. 39-56). Springer. https://doi.org/10.1007/978-981-13-0248-0_4

[Downloaded from johepal.com on 2026-11-15]

[DOI: 10.66224/johepal.7.2.92]

Khan, M. A., Nabi, M. K., Khojah, M., & Tahir, M. (2020). Students' perception towards e-learning during COVID-19 pandemic in India: An empirical study. *Sustainability*, *13*(1), 57. https://doi.org/10.3390/su13010057

Kaur, J., & Buttar, P. K. (2018). A systematic review on stopword removal algorithms. *International Journal on Future Revolution in Computer Science & Communication Engineering, 4*(4), 207-210. https://www.ijfrcsce.org/index.php/ijfrcsce/article/view/1499

Kim, H. (2024). A fad or the new norm for student access today? Evaluating enrollment outcomes of holistic admissions in South Korea. *Research in Higher Education*, *65*(5), 1040-1064. https://doi.org/10.1007/s11162-024-09776-9

Kim, S., & Kim, N. (2024). Unveiling the evolving educational inequality from upper secondary to higher education in South Korea: From effectively maintained inequality theory perspective. *Higher Education, 89*(6), 1637-1657. https://doi.org/10.1007/s10734-024-01301-2

Kotsiantis, S. B. (2012). Use of machine learning techniques for educational purposes: A decision support system for forecasting students' grades. *Artificial Intelligence Review, 37*(4), 331-344. https://doi.org/10.1007/s10462-011-9234-x

Lakkaraju, H., Aguiar, E., Shan, C., Miller, D., BHANPURI, N., GHANI, R., & ADDISON, K. L. (2015). A machine learning framework to identify students at risk of adverse academic outcomes. In *Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1909-1918). https://doi.org/10.1145/2783258.2788620

Lantz, B. (2019). *Machine learning with R: Expert techniques for predictive modeling*. Packt Publishing Ltd.

Ma, L. (2016). Female labour force participation and second birth rates in South Korea. *Journal of Population Research, 33*(2), 173-195. https://doi.org/10.1007/s12546-016-9166-z

Maulana, A., Noviandy, T. R., Sasmita, N. R., Paristiowati, M., Suhendra, R., Yandri, E., & Idroes, R. (2023). Optimizing university admissions: A machine learning perspective. *Journal of Educational Management and Learning, 1*(1), 1-7. https://doi.org/10.60084/jeml.v1i1.46

Nghe, N. T., Janecek, P., & Haddawy, P. (2007). A comparative analysis of techniques for predicting academic performance. In *Proceedings of the 37th Annual Frontiers in Education Conference* (pp. T2G7-T2G12). https://doi.org/10.1109/FIE.2007.4417993

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

Namoun, A., & Alshanqiti, A. (2020). Predicting student performance using data mining and learning analytics techniques: A systematic literature review. *Applied Sciences, 11*(1), 237-265. https://doi.org/10.3390/app11010237

Obsie, E. Y., & Adem, S. A. (2018). Prediction of student academic performance using neural network, linear regression and support vector regression: A case study. *International Journal of Computer Applications, 180*(40), 39-47. https://doi.org/10.5120/ijca2018917057

Posselt, J. R. (2016). *Inside graduate admissions: Merit, diversity, and faculty gatekeeping*. Harvard University Press.

Pradana, A. W., & Hayaty, M. (2019). The effect of stemming and removal of stopwords on the accuracy of sentiment analysis on Indonesian-language texts. *Kinetik: Game Technology, Information System, Computer Network, Computing, Electronics, and Control, 4*(4), 375-380. https://doi.org/10.22219/kinetik.v4i4.912

Prihatini, P. M., Suryawan, I. K., & Mandia, I. N. (2018). Feature extraction for document text using latent Dirichlet allocation. *The Journal of Physics: Conference Series, 953*(1), 012047. https://doi.org/10.1088/1742-6596/953/1/012047

[Downloaded from johepal.com on 2026-11-15]

[DOI: 10.66224/johepal.7.2.92]

Raghavendran, C. V., Pavan Venkata Vamsi, C., Veerraju, T., & Veluri, R. K. (2021). Predicting student admissions rate into university using machine learning models. In D. Bhattacharyya, & N. Thirupathi Rao (Eds.), *Machine Intelligence and Soft Computing: Proceedings of ICMISC 2020* (pp. 151-162). https://doi.org/10.1007/978-981-15-9516-5_13

Rastrollo-Guerrero, J. L., Gómez-Pulido, J. A., & Durán-Domínguez, A. (2020). Analyzing and predicting students' performance by means of machine learning: A review. *Applied Sciences,10*(3), 1-25. https://doi.org/10.3390/app10031042

Romero, C., & Ventura, S. (2007). Educational data mining: A survey from 1995 to 2005. *Expert Systems with Applications, 33*(1), 135-146. https://doi.org/10.1016/j.eswa.2006.04.005

Siino, M., Tinnirello, I., & La Cascia, M. (2024). Is text preprocessing still worth the time? A comparative survey on the influence of popular preprocessing methods on transformers and traditional classifiers. *Information Systems, 121*, 102342. https://doi.org/10.1016/j.is.2023.102342

Singh, J., & Gupta, V. (2017). A systematic review of text stemming techniques. *Artificial Intelligence Review*, *48*(2), 157-217. https://doi.org/10.1007/s10462-016-9498-2

Smola, A. J., & Schölkopf, B. (2004). A tutorial on support vector regression. *Statistics and Computing*, *14*(3), 199-222. https://doi.org/10.1023/B:STCO.0000035301.49549.88

Tair, M. M. A., & El-Halees, A. M. (2012). Mining educational data to improve students' performance: A case study. *International Journal of Information and Communication Technology Research, 2*(2), 140-146.

Taub, M., & Azevedo, R. (2018). Using sequence mining to analyze metacognitive monitoring and scientific inquiry based on levels of efficiency and emotions during game-based learning. *Journal of Educational Data Mining, 10*(3), 1-26. https://doi.org/10.5281/zenodo.3554712

Walid, Md. A. A., Ahmed, S. M. M., Zeyad, M., Galib, S. M. S., & Nesa, M. (2022). Analysis of machine learning strategies for prediction of passing undergraduate admission test. *International Journal of Information Management Data Insights, 2*(2), 100111. https://doi.org/10.1016/j.jjimei.2022.100111

Wang, Y., Sun, Z., Zhang, H., Cui, W., Xu, K., Ma, X., & Zhang, D. (2019). Datashot: Automatic generation of fact sheets from tabular data. *IEEE Transactions on Visualization and Computer Graphics, 26*(1), 895-905. https://doi.org/10.1109/TVCG.2019.2934398

Wu, J. -P., Lin, M. -S., & Tsai, C. -L. (2023). A predictive model that aligns admission offers with student enrollment probability. *Education Sciences, 13*(5), 440. https://doi.org/10.3390/educsci13050440

Xu, L. (2024). Prediction of college admission scores based on an XGBoost-LSTM hybrid model. In *Proceedings of the 3rd International Conference on Educational Innovation and Multimedia Technology, EIMT 2024, March 29-31*. http://dx.doi.org/10.4108/eai.29-3-2024.2347687

Yadav, S. K., Bharadwaj, B., & Pal, S. (2012). Mining education data to predict student's retention: A comparative study. *arXiv*. https://doi.org/10.48550/arXiv.1203.2987

Yağci, M. (2022). Educational data mining: Prediction of students' academic performance using machine learning algorithms. *Journal of Educational Management and Learning, 9*(1), 11-30. https://doi.org/10.1186/s40561-022-00192-z

Yang, X., Yang, K., Cui, T., Chen, M., & He, L. (2022). A study of text vectorization method combining topic model and transfer learning. *Processes, 10*(2), 350. https://doi.org/10.3390/pr10020350

Yoo, S. H., & Sobotka, T. (2018). Ultra-low fertility in South Korea: The role of the tempo effect. *Demographic Research, 38*, 549-576. https://doi.org/10.4054/DemRes.2018.38.22

[Downloaded from johepal.com on 2026-11-15]

[DOI: 10.66224/johepal.7.2.92]

Young, N. T., Tollefson, K., Zegers, R. G., & Caballero, M. D. (2022). Rubric-based holistic review: A promising route to equitable graduate admissions in physics. *Physical Review Physics Education Research, 18*(2), 020140. https://doi.org/10.1103/PhysRevPhysEducRes.18.020140

Zafra, A., & Ventura, S. (2009). Predicting student grades in learning management systems with multiple instance genetic programming. In *Proceedings of the 2009 9th International Working Group on Educational Data Mining* (pp. 307-314). https://files.eric.ed.gov/fulltext/ED539094.pdf

Dr. Yousun Shin is an Associate Professor in the Department of English Education at Sunchon National University, South Korea. Her recent work focuses on educational assessment, AI-assisted language education, and data-driven approaches to English language teaching and learning in EFL contexts.

Dr. Hee Sun Kang is a Professor in the Department of Nursing at Sunchon National University, South Korea. Her recent work focuses on community health nursing, regional healthcare systems, and health promotion in local communities.

Dr. So Yun Park is a Research Fellow at the Institutional Research (IR) Center, Office of Planning, Sunchon National University, South Korea. Her work focuses on institutional research, university ranking strategy, and data-driven performance analytics in higher education, with research interests in educational data mining and the application of machine learning to higher education administration.

[Downloaded from johepal.com on 2026-11-15]

[DOI: 10.66224/johepal.7.2.92]

This is an open access article distributed under the terms of the Creative Commons Attribution-NonCommercial 4.0 International (CC BY-NC 4.0) which allows reusers to distribute, remix, adapt, and build upon the material in any medium or format for noncommercial purposes only, and only so long as attribution is given to the creator.