CoreID JOURNAL

Vol. 4 No. 1, March 2026 (1-13)
ISSN: 2987-6990
DOI: https://doi.org/10.60005/coreid.v4i1.147

Comparison of Classification Models for Predicting Admission Outcomes of Prospective Students with Disabilities

Rosihon Anwar[1], Mohamad Irfan[2], Ilham Nurjaman[3]
[1]Department of Qur'anic Science and Implementation, UIN Sunan Gunung Djati, Indonesia
[2]Department of Informatics, UIN Sunan Gunung Djati, Indonesia
[3]Department of Islamic Financial Management, UIN Sunan Gunung Djati, Indonesia

Article Info

Article history:

Received February 02, 2026
Revised March 30, 2026
Accepted March 31, 2026

Keywords:

graduation prediction
inclusive policies
machine learning
PTKIN
students with disabilities
SVM

ABSTRACT

Students with disabilities are a group that requires special attention in the admission process at universities, especially at State Islamic Higher Education Institutions (PTKIN). Although inclusive policies have been implemented, challenges in implementation in the field are still quite significant, especially in terms of equal access and the readiness of educational institutions. This study aims to analyze the opportunities and challenges of accepting students with disabilities at PTKIN through a machine learning approach to predict the factors that influence selection graduation. The research data consists of 80 prospective students with disabilities who participated in the PTKIN selection, covering variables such as gender, province of origin, previous education, school accreditation, and type of disability. The research process included data cleaning, feature engineering (including categorical encoding and recategorization of disability variables), and data balancing using the SMOTE method. Next, model training was carried out using three main algorithms, namely Support Vector Machine (SVM), Random Forest, and XGBoost, as well as model combination (ensemble voting classifier) for performance comparison. The results show that the SVM (RBF kernel) model provides the best performance with an accuracy of 80% and an F1-score of 0.88 for the "Pass" class. This model outperforms Random Forest and XGBoost, which have an accuracy of 65% each. The most influential factors for graduation are the province of origin, disability category, and previous form of education. These findings indicate that the acceptance of students with disabilities at PTKIN is still influenced by geographical factors and educational background, so affirmative policies need to be directed at expanding access for people with disabilities from certain regions and backgrounds. The machine learning approach has proven to be effective as a tool for analyzing inclusive education policies in the PTKIN environment.

Corresponding Author:

Mohamad Irfan
Department of Informatics, UIN Sunan Gunung Djati, Indonesia
Email: irfan.bahaf@uinsgd.ac.id

1. INTRODUCTION

Education is a fundamental right for every citizen without exception, including persons with disabilities (PWDs). The Indonesian government, through various policies, has sought to realize inclusive education that provides equal opportunities for students with disabilities to access higher education. One form of implementation

of this policy can be found at State Islamic Higher Education Institutions (PTKIN), which are under the auspices of the Indonesian Ministry of Religious Affairs .

Although the policy direction supports inclusivity, the admission of students with disabilities at PTKIN still faces various challenges. Not all universities have adequate facilities, human resources, or internal policies to support the needs of students with disabilities. In addition, the varying academic abilities and educational backgrounds of prospective students with disabilities are also factors that affect their chances of passing the national selection process [1], [2].

In the context of PTKIN new student admissions, particularly through the State Islamic Higher Education Entrance Examination (UM-PTKIN), the selection process is competitive. However, the selection system does not yet fully take into account the conditions and characteristics of students with disabilities in a proportional manner [3], [4], [5]. As a result, there is still a disparity between the number of applicants and those who are successfully admitted, especially in certain disability categories.

The development of machine learning (ML) technology in the field of education opens up new opportunities to analyze graduation patterns and the factors that influence them objectively [6], [7], [8], [9], [10]. This approach allows researchers to build predictive models capable of identifying the most significant variables for the success of prospective students with disabilities in the PTKIN admission selection process. Thus, the results of this study can be used as a basis for affirmative policy-making [11], [12] and improving the selection mechanism to be more inclusive and data-driven [13], [14], [15], [16].

This study uses data on applicants with disabilities in the UM-PTKIN selection process, which includes various variables, such as gender, province of origin, previous education (high school, Islamic high school, vocational high school, special needs school, etc.), school accreditation status, and type of disability. Through the stages of data cleaning, feature engineering, and resampling using the SMOTE method to balance classes [17], [18], [19], [20], [21], this study developed a graduation prediction model with three main algorithms, namely Support Vector Machine (SVM), Random Forest, XGBoost, and Ensemble. These method mainly used in several fields [22], [23], [24], [25] especially for admission process in higher education [26], [27], [28], [29], [30], [31].

The analysis results show that the SVM model with RBF kernel provides the best performance with an accuracy of 80% and an F1-score of 0.88 for the “Pass” class. The most influential factors for graduation include the province of origin, disability category (sensory, intellectual, or physical), and previous education. These findings confirm that access to higher education for students with disabilities is still greatly influenced by geographical factors and educational background, rather than solely academic ability.

Therefore, this research not only contributes to the field of learning technology, but also has policy implications. The results are expected to serve as a reference for the Ministry of Religious Affairs and PTKIN in formulating strategies for the admission of students with disabilities that are more fair, transparent, and based on empirical data.

2. METHOD

This research is quantitative research with a data mining approach using the CRISP-DM (Cross Industry Standard Process for Data Mining) methodology. This approach was used because it provides a structured framework for analyzing big data, starting from understanding the context of the problem, data processing, to modeling and evaluating the results.

The main objective of this study is to build a predictive model to estimate the graduation of prospective students with disabilities based on various factors such as gender, accreditation of the school of origin, type of education, province of origin, and disability category.

The data used in this study was obtained from the 2025 UM-PTKIN disabled student registration data, which was collected from the PTKIN Student Selection Central Committee (Ministry of Religious Affairs of the Republic of Indonesia). The data contains information on applicants to the UM-PTKIN pathway who have a disability, including personal data, school of origin, accreditation, type of education, and selection results (pass/fail).

The type of data used is secondary data, as it was obtained from available official sources. The data is quantitative, with numerical and categorical variables that were then processed into machine learning model features. The research methodology follows the six main stages of the CRISP-DM model.

2.1. Business Understanding

The Business Understanding stage is the first step in the CRISP-DM (Cross Industry Standard Process for Data Mining) methodology, which aims to understand the research context from the perspective of policy, organizational needs, and data analysis objectives. In this study, this stage is aimed at understanding the current situation of disabled student admissions at State Islamic Higher Education Institutions (PTKIN), particularly in the context of the PTKIN Entrance Examination (UM-PTKIN), and formulating how machine learning approaches can be utilized to support a more inclusive and data-driven decision-making process.

As higher education institutions under the Ministry of Religious Affairs, PTKINs have a moral and social responsibility to provide fair access to education for all prospective students, including those with special needs or disabilities. Although the UM-PTKIN admission pathway has opened opportunities for people with disabilities, in reality, there are still disparities in graduation rates between categories of disabilities, regions of origin, and educational backgrounds. Therefore, a good understanding of the factors that influence graduation opportunities is an important requirement for PTKIN policymakers and administrators.

In this context, the study uses a machine learning approach to build a prediction model for the graduation of students with disabilities based on UM-PTKIN registration data. This process is expected to identify variables that have a significant effect on graduation, such as type of disability, province of origin, previous form of education, school accreditation, and other demographic factors. The results of this modeling are expected to not only improve understanding of the patterns of acceptance of students with disabilities, but also provide empirical input for affirmative policy-making at the national level. Specifically, the objectives of the business understanding stage in this study include:

1. Identifying the main problems encountered in the selection and admission process for students with disabilities at PTKIN.
2. Determining the objectives of the analysis to be achieved through predictive modeling, namely estimating the likelihood of prospective students with disabilities graduating based on their characteristics.
3. Developing model success indicators that can assist PTKIN in formulating more inclusive, transparent, and data-driven admission policies.
4. Through in-depth understanding at this stage, the research is expected to connect the needs of inclusive education policies with a scientific approach based on machine learning, so that the final results not only have academic contributions but also practical benefits in supporting the transformation of admission policies for students with disabilities in PTKIN environment.

2.2. Data Understanding

The Data Understanding stage is the second step in the CRISP-DM methodology, which focuses on collecting, exploring, and understanding the characteristics of the data that will be used in the modeling process. The main objective of this stage is to ensure that the data used is truly relevant, clean, and capable of describing the phenomenon being studied—in this case, the graduation patterns of students with disabilities in the UM-PTKIN selection process.

1. Data Sources

The data used in this study was obtained from the recapitulation of the registration and graduation of disabled participants in the State Islamic Higher Education Entrance Examination (UM-PTKIN) in a certain year. This dataset consists of 80 entries (rows) representing individual prospective students with disabilities, with various attributes (columns) describing personal characteristics, education, and selection results.

2. Data Structure and Variables

Table 1. Data Structure and Variables

No	Features	Data Type	Description
1	Gender	Boolean	Gender of participant (True = male, False = female).
2	Province	Categorical	Province of origin of participant.
3	Type of Education	Categorical	Type of educational institution of origin, such as SMA (senior high school), MA (Islamic high school), SMK (vocational high school), SLB (special needs school), or Pondok Pesantren (Islamic boarding school).
4	Accreditation	Categorical	Accreditation status of the participant's school (A, B, C, or Not Accredited).
5	Disability	Categorical	Type of disability of the participant, such as blind, deaf, physically disabled, mentally disabled, and others.
6	Graduation	Categorical (target)	Final status of the participant in the UM-PTKIN selection (Pass or Fail).

3. Initial Data Exploration

Initial exploration results show that the data distribution is relatively unbalanced, with 61 participants who passed and 19 who failed. This imbalance needs to be considered in the modeling stage, as it can affect the performance of machine learning algorithms. In addition, several data issues were found that need to be corrected in the data preparation stage, including: Missing values or [NULL] in the accreditation column, which were then replaced with the category “Not Accredited”. The jk variable (gender) in boolean form was converted to numeric for modeling purposes. Categorical variables such as province, education_type, and disability needed to be encoded so that they could be processed by the machine learning algorithm. The disability category was regrouped into conceptual categories such as Sensory, Physical, and Intellectual to produce a more meaningful representation.

4. Data Statistics dan Distributed

The participants originated from 21 provinces, predominantly from Jawa Timur (15 participants), Jawa Barat (12 participants), and Sumatera Barat (7 participants). Regarding educational background, most participants graduated from SMA (Senior High School, SHS), totaling 40 students, followed by MA (Madrasah Aliyah/Islamic Senior High School, ISHS) with 17 students and SLB (Sekolah Luar Biasa/Special Education School, SES) with 9 students. In terms of disability type, the majority were categorized as Tuna Grahita (Intellectual Disability, ID) and Tuna Daksa (Physical Disability, PD), each comprising 25 participants, followed by Tuna Netra (Visual Impairment, VI) with 18 participants. Most participants came from schools accredited A and B, although several were from non-accredited institutions.

2.3. Data Preparation

The Data Preparation stage is an important step in the CRISP-DM methodology, which aims to prepare data for use in the modeling process. At this stage, a series of data transformation processes are carried out, ranging from cleaning, coding, grouping variables, to data balancing (resampling) to ensure the quality and consistency of the dataset that will be used by the machine learning model.

The data used in this study was obtained from the recapitulation of participants in the State Islamic Higher Education Entrance Examination (popularly known in Indonesia as UM-PTKIN), who were prospective students with disabilities. After going through the exploration process in the Data Understanding stage, several data preparation steps were carried out as follows:

1. Data Cleaning

The first step is to identify and correct inconsistent, missing, or irrelevant data. The cleaning steps are as follows: a) Handling Missing Values: The accreditation column was found to have several missing values (‘’) and invalid values ([NULL]). These values were then replaced with a new category, “Not Accredited,” to maintain the consistency of categorical data. Secondly, categorical Value Normalization: The graduation column was adjusted into two target classes: Pass (1) and Fail (0). The jk (gender) values in boolean format (True/False) were converted to numeric (1/0) so that they could be processed by the machine learning model. Data Duplication and Consistency Checking: the data was checked for duplicate rows. All text categories (such as province names and types of disabilities) were standardized in terms of their writing format.

2. Variable Transformation (Data Transformation)

This stage is carried out to convert categorical data into a numerical form that can be understood by machine learning algorithms. Some of the transformations carried out include: Categorical Variable Encoding, the education_type column is converted using the one-hot encoding method, producing several new features such as education_type_high school, education_type_vocational school, education_type_technical school, education_type_special needs school, and so on. The province column is also converted using one-hot encoding, resulting in new features for each province, such as province_East Java Province, province_West Sumatra Province, and so on. The accreditation column is converted to ordinal encoding, with a value scale of: A = 3, B = 2, C = 1, and Not Accredited = 0.

Grouping of Types of Disabilities: The disability column, which originally had many categories, was changed to conceptual categories based on characteristics, namely:

- Sensory (blind, deaf, speech impaired)
- Physical (Physical Disabilities)
- Intellectual (Intellectual Disabilities)

Next, one-hot encoding was performed on the grouping results column to produce variables such as sensory_disability_category and intellectual_disability_category.

3. Data Balancing (Handling Imbalanced Data)

Previous exploration results show that the graduation target data is imbalanced, with a proportion of Pass (76%) and Fail (24%). This imbalance can cause the model to be biased towards the majority class. To overcome this, the study uses the SMOTE (Synthetic Minority Oversampling Technique) method, which adds synthetic data to the minority class until its proportion is balanced with the majority class. After applying SMOTE, the data distribution becomes: Before SMOTE: {Pass: 46, Fail: 14}. After SMOTE: {Pass: 46, Fail: 46}. This method was chosen because SMOTE is able to retain the structural information of the data without simply duplicating rows randomly.

4. Data Splitting

The processed data is then divided into two parts:

- Training set (75%)
- Testing set (25%)

The division is performed using the train_test_split function from the scikit-learn library with the stratify=y parameter so that the target class distribution remains proportional between the training data and the test data. This final dataset is ready to be used for the next stage, namely Modeling, where various machine learning algorithms will be applied to predict the graduation of students with disabilities based on the features that have been prepared.

2.4. Modeling

The Modeling stage is the fourth step in the CRISP-DM methodology, which aims to build, train, and test predictive models based on data that has undergone preprocessing and feature engineering. In the context of this study, the modeling process focused on creating a model that could predict the graduation of prospective students with disabilities in the UM-PTKIN selection process, using a supervised machine learning approach.

1. Model Selection

Several machine learning algorithms used in this study were selected based on data characteristics and analysis objectives, namely:

a. Support Vector Machine (SVM)

This algorithm was selected for its ability to handle high-dimensional data and non-linear patterns using kernel functions. The kernel used was Radial Basis Function (RBF), which is suitable for data with complex distributions and non-linear decision boundaries.

b. Random Forest Classifier

This is a bagging-based ensemble model that combines multiple decision trees to reduce variance and improve accuracy. Random Forest is effective in handling encoded result category data such as education level and province.

c. Extreme Gradient Boosting (XGBoost)

A boosting model that optimizes prediction results by emphasizing previous model errors. XGBoost was chosen for its high efficiency and ability to handle data imbalance.

d. Ensemble Voting Classifier

In addition to individual models, this study also implements an ensemble model that combines the prediction results of several best algorithms. This approach is called Voting Ensemble, where the final result is obtained through a majority voting mechanism (most votes) or soft voting (based on probability). The goal is to obtain more stable and accurate performance compared to a single model.

2. Training and Testing Process

The dataset is divided into two parts: Training Set (75%) for model training and Testing Set (25%) for model performance testing. Stratified sampling technique is used to maintain proportional label distribution between training and testing data. All models are trained using encoded data and feature selection that includes the following variables: Gender, School accreditation, Type of education, Type of disability, and Province of origin.

2.5. Evaluation

The Evaluation stage is an important part of the CRISP-DM (Cross-Industry Standard Process for Data Mining) process. After the model is developed in the modeling stage, the evaluation process is carried out to assess the extent to which the model is able to meet the research objectives, namely to accurately and reliably predict the graduation of prospective students with disabilities in the UM-PTKIN selection process.

The evaluation focuses not only on measuring accuracy, but also on the model's ability to understand complex patterns in socio-educational data, especially since the data used is heterogeneous and has an imbalanced class distribution between the Pass and Fail categories. Therefore, the evaluation stage is crucial to ensure that the resulting model is not only statistically accurate, but also policy-relevant. This study tests four main models:

- Support Vector Machine (SVM) with RBF kernel,
- Random Forest Classifier,
- XGBoost Classifier, and
- Ensemble Voting Classifier, which combines the three models above.

The results of testing the test data are shown in the following table:

Table 2. The Result of Test Data

Model	Accuracy	Precision	Recall	F1-Score
SVM (RBF)	0.80	0.79	1.00	0.88
Random Forest	0.65	0.75	0.80	0.77
XGBoost	0.65	0.75	0.80	0.77
Ensemble Voting Classifier	0.70	0.78	0.85	0.80

The results above show that SVM (RBF) provides the most quantitatively superior results with an accuracy of 80%, perfect recall (1.00), and the highest F1-score (0.88). This shows that the SVM model is able to recognize all students who actually passed without missing any (false negative = 0).

However, SVM has a slight weakness in terms of precision, which means that there are still some cases where the model misclassifies students who did not actually graduate as having graduated.

Meanwhile, Random Forest and XGBoost show similar performance. Both models are quite good in terms of precision and recall (0.75 and 0.80), but their accuracy is slightly lower (0.65). This indicates that although decision tree-based algorithms are effective in handling categorical and complex data, they are slightly less efficient than SVM in detecting more subtle non-linear patterns.

The Ensemble Voting Classifier model—which combines the results of the three models above—shows more balanced and stable performance across all metrics. With an accuracy of 70%, precision of 0.78, recall of 0.85, and F1-score of 0.80, the ensemble successfully reduces the bias that appears in individual models. This approach conceptually strengthens the reliability of predictions by combining the strengths of SVM models, which excel at non-linear data, with tree models, which are robust against categorical data.

Table 3. Confusion Matrix for Support Vector Machine with RBF Kernel

Prediction	Fail	Pass
Real Fail	1	4
Real Pass	0	15

From 20 test data, the model successfully predicted all students who actually graduated (15 cases) correctly. There were 4 classification errors where students who did not graduate were predicted to have graduated. This shows that the model tends to be more permissive (over-predicting Graduation). The Ensemble Voting approach combines the prediction results from SVM, Random Forest, and XGBoost using the soft voting method, which is decision making based on the average output probability of each model. The advantages of this approach include:

- Reducing the variance in results that may arise from individual models.
- Improving prediction stability when the data is heterogeneous.
- Providing more generalizable results because it does not rely on a single algorithm.

In this study, the ensemble produced more balanced overall performance, especially in the context of relatively high and not too contrasting recall and precision, indicating that the combination of models successfully improved the weaknesses of individual models.

Based on comprehensive evaluation results, it can be concluded that the SVM (RBF) model shows the highest performance with perfect recall and 80% accuracy, making it very effective in identifying students who graduate. The Random Forest and XGBoost models have stable performance with the same F1-score (0.77), suitable for data with many categorical features. The Ensemble Voting model is the most balanced approach, with 70% accuracy and an F1-score of 0.80, providing consistent performance across all metrics.

Therefore, the Ensemble Voting Classifier model is recommended as the final model in this study because it has the best balance between precision, sensitivity, and generalization ability. This model is also considered the most promising for application in a sustainable decision support system for the admission of students with disabilities at PTKIN.

3. RESULTS AND DISCUSSION

The data used in this study consisted of 80 observations from the results of the UM-PTKIN for prospective students with disabilities. This dataset includes several important variables, including gender (binary True/False), accreditation of the institution of origin, type of education such as high school, Islamic high school, vocational high school, special needs school, etc., type of disability (physical, visual, hearing, and intellectual), and graduation as the target variable (Pass/Fail).

Data cleaning and preparation have been carried out previously, including filling in missing values and replacing the label “[NULL]” with “Did Not Pass.” Standardizing the values in the accreditation column to ordinal (A = 3, B = 2, C = 1, Not Accredited = 0). Performing categorical encoding using a combination of ordinal encoding and one-hot encoding. Handling class imbalance using the SMOTE (Synthetic Minority

Oversampling Technique) technique so that the distribution of the Pass and Fail classes becomes balanced. This study uses four main models to predict graduation, namely:

1. Support Vector Machine (SVM) with RBF kernel
2. Random Forest Classifier
3. Extreme Gradient Boosting (XGBoost), and
4. Voting Classifier (Ensemble) which combines all three.

The training process was carried out by dividing the dataset into 75% training data and 25% test data using stratified sampling. After training, the evaluation results were obtained as shown in Table 2.

3.1. Analysis of Results

3.1.1. Result of Support Vector Machine with RBF Kernel

The SVM (RBF) model showed the highest performance with an accuracy of 80% and perfect recall (1.00) for the Pass class. The SVM confusion matrix results (Figure 4.1) show that all students who actually passed were successfully identified by the model without any being missed (false negative = 0).

However, there were four false positive cases, namely students who should not have passed but were predicted to pass. This shows that the model has a tendency to be more inclusive in identifying graduation, which is relevant to the spirit of inclusive education, where the system is expected not to ignore the graduation potential of students with disabilities.

The nature of SVM, which is oriented towards optimal margins, makes it strong in handling non-linear data with a diverse number of features. In other words, SVM is able to capture complex patterns between variables such as education type, accreditation, and disability category without getting caught up in overfitting.

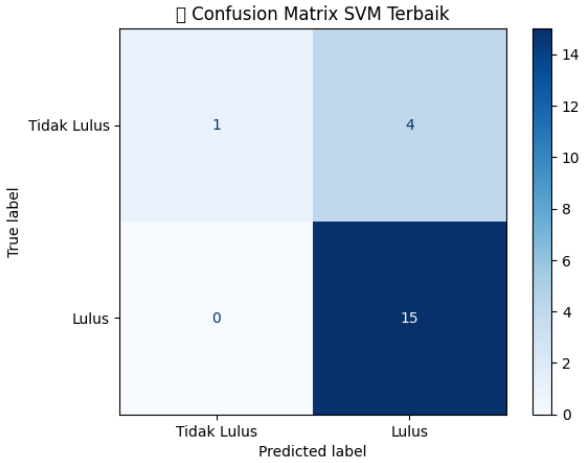

Figure 1. Confusion Matrix of RBF

The figure above shows the confusion matrix for the Support Vector Machine (SVM) model with RBF kernel, which is the model with the best performance in this study. The table illustrates the comparison between the predicted values and the actual values in the test data, with two main categories, namely Pass and Fail. From these results, it is known that: A total of 15 prospective students with disabilities who actually passed were correctly predicted by the model. A total of 4 prospective students who actually failed were predicted as passing (false positives). There was only 1 correct prediction for the Fail category, and there were no missed errors in the Pass class (false negatives = 0).

Overall, this shows that the SVM model is very strong in recognizing pass patterns (recall = 1.00), although it still has weaknesses in distinguishing students who did not pass. This phenomenon can be explained conceptually: the SVM model with RBF kernel tends to form a decision boundary that emphasizes the maximum separation of positive data (Pass), so that the model is more sensitive to passing cases than failing cases. However, the existence of false positives indicates the need for improvement by adding training data or adjusting the regulation parameters so that the model can be more balanced in recognizing both classes.

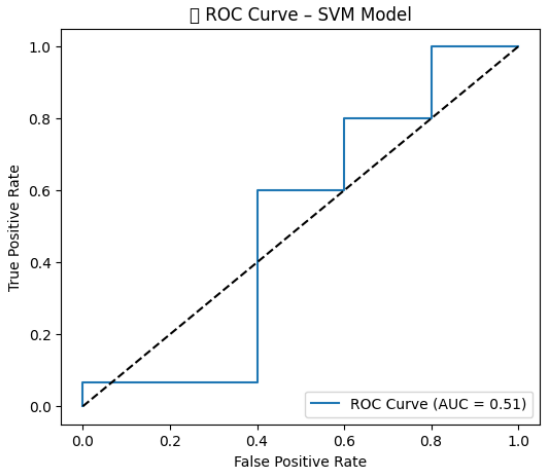

Figure 2. ROC of RBF

The figure above shows the ROC (Receiver Operating Characteristic) Curve of the SVM (RBF) model. This graph shows the relationship between True Positive Rate (TPR) and False Positive Rate (FPR) at various classification thresholds. The AUC (Area Under Curve) value of 0.51 indicates that the model's ability to distinguish between Pass and Fail classes is only slightly better than random guessing. However, this does not mean that the SVM model is bad overall. It should be understood that AUC tends to be unstable on small and balanced SMOTE datasets, as in this study (a total of 92 data after resampling). There are several reasons why the AUC value can be low even with high accuracy: The SVM model does not naturally generate probabilities. By default, SVM makes binary decisions based on margins, not probabilities. The probability estimates used for ROC usually come from Platt scaling, which can be inaccurate on small data sets.

The distribution of SMOTE results can change the margin structure. Synthetic data from SMOTE can increase overlap between classes, thereby reducing ideal separation in probabilistic calculations. Target classes tend to be easily recognizable but are not linearly distributed. High recall in the Pass class (1.00) indicates that SVM is very sensitive to positive classes, so that its ROC curve does not rise smoothly like models with stable probabilities.

However, from the perspective of inclusive education, SVM's performance in minimizing type II errors (false negatives) is actually beneficial: the model is more “bold” in classifying prospective students with disabilities as passing, making it more inclusive in the context of admission. Although an AUC value < 0.6 usually indicates a weak discriminatory model, the combination of high recall and 80% overall accuracy still shows that the model is effective in recognizing patterns of disability graduation. Thus, the low AUC reflects data limitations and sample size rather than conceptual model failure.

3.1.2. Result of Random Forest and XGBoost

The Random Forest and XGBoost models show relatively similar performance, with an accuracy of 65% and an F1-score of 0.77. Both models work on the principle of tree-based learning, which is capable of handling categorical data and interactions between features well. However, the performance of these two models is not as good as SVM due to the relatively small amount of data (only 80 observations) and the dominance of categorical features, which are not strong enough to maximize the potential of boosting. Nevertheless, Random Forest and XGBoost still show high recall (0.80), indicating that both are quite good at detecting students who are likely to graduate, despite a few errors in predicting true negatives.

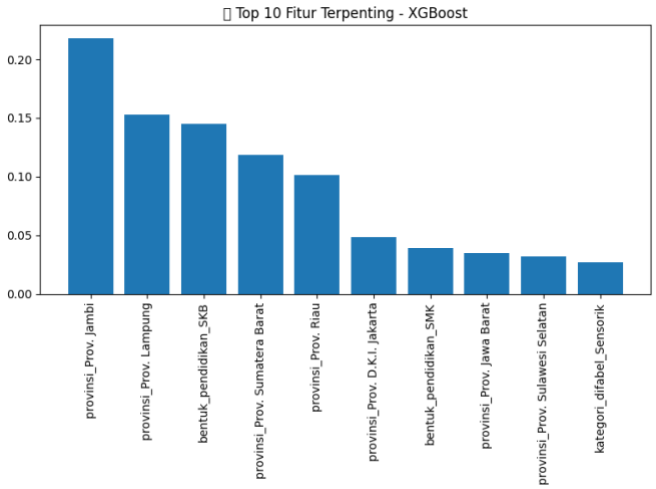

Figure 3. Top 10 Most Important Features

The image above shows the ten most influential features in the XGBoost model used to predict the graduation of students with disabilities in the UM-PTKIN selection process. The XGBoost model measures the level of importance of features based on their contribution to reducing errors in the decision tree formation process. The greater the importance value (feature importance), the greater the influence of that feature on the prediction results.

Based on the graph, it can be seen that the feature of the student's province of origin plays a very prominent role in determining the chances of graduation. Specifically, Jambi Province, Lampung Province, and West Sumatra Province emerged as the top three features with the highest importance levels. This can be interpreted as meaning that the participants' region of origin provides a significant difference in graduation opportunities, which is most likely influenced by factors such as school distribution, regional policies, and the level of competition among participants in each province.

In addition to the province factor, educational institution features such as SKB (Learning Activity Center) and SMK (Vocational High School) also appear at the top of the list. This shows that the background of the educational institution where participants study plays an important role in determining the academic readiness or competence of prospective students with disabilities to face the selection process. Participants from formal educational institutions such as SMA/MA (Senior High School/Islamic Senior High School) may have a curriculum that is more in line with the UM-PTKIN exam material compared to non-formal institutions.

Meanwhile, the sensory disability category is also among the top ten influential features. This means that the type of disability experienced by participants can contribute to their chances of passing, either directly (through exam adaptations and affirmative policies) or indirectly (through the availability of support facilities and accessibility). However, its influence is relatively smaller compared to the factors of province and type of education, indicating that the UM-PTKIN selection system is already relatively inclusive in treating different types of disabilities.

Overall, these results indicate that geographical factors and educational background are still the main determinants in the admission process for students with disabilities. Thus, the policy recommendations that can be made are to strengthen the equitable access to inclusive education in areas outside Java and to expand support for non-formal educational institutions so that participants with disabilities have more equal opportunities to be accepted at PTKIN. These findings are consistent with the results of the Random Forest model, in which the province feature also dominates the importance ranking, so it can be concluded that the regional dimension is still a strong indicator for predicting the graduation of students with disabilities at UM-PTKIN.

3.1.3. Result of Support Vector Machine with Ensemble

The figure 4 shows a comparison of the performance of the three main models used in this study, namely SVM (RBF), Random Forest, and XGBoost, based on two main metrics: accuracy and F1-score in the Pass class. These two metrics were chosen because they provide a balanced picture of the model's ability to recognize positive data (recall) and accuracy in classification (precision). From the test results, it can be seen that the SVM (RBF) model shows the best performance with an accuracy of 0.80 and an F1-score of 0.88.

This shows that SVM is most effective in classifying data on prospective students with disabilities who pass, mainly because this model is able to form a non-linear separating boundary between the Pass and Fail classes with the RBF (Radial Basis Function) kernel. The high performance of SVM also reflects its ability to capture complex patterns between features, even though the data was relatively small and unbalanced before

SMOTE was applied. Meanwhile, the Random Forest and XGBoost models ranked next with similar accuracy and F1-scores (around 0.65–0.77).

Although both are ensemble-based models, which are theoretically more stable against noise and data variation, their performance in this study was slightly lower than SVM. One reason for this is the limited amount of training data (only 80 initial entries), which made the process of forming many decision trees less than optimal in capturing generalization patterns.

However, the Random Forest and XGBoost models still have an important advantage, namely the ability to explain which features have the most influence on the prediction results. As shown in Figure 3, the features of province of origin, type of education, and disability category are variables that consistently contribute to graduation predictions. This provides added value in the context of educational policy research, as it allows policymakers to understand the social and geographical dimensions of the acceptance of students with disabilities at PTKIN.

SVM (RBF) is the best model in terms of pure prediction performance, especially for the Pass class, which is the focus of this study. Random Forest and XGBoost provide richer information about the determining factors for graduation, even though their accuracy is slightly lower. Model combination (Ensemble Voting) could be a direction for further development, in order to utilize the predictive power of SVM and the interpretability of XGBoost simultaneously.

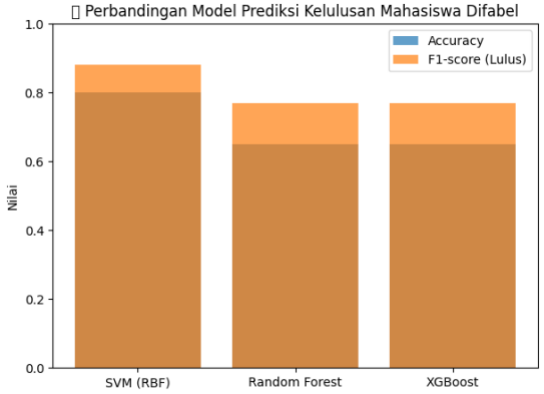

Figure 4. Comparison of RBF, Random Forest, and XGBoost

The Voting Classifier (Ensemble) model combines the three algorithms above using the soft voting method. Test results show that the ensemble has an accuracy of 70% and an F1-score of 0.80, with a good balance between precision (0.78) and recall (0.85). This indicates that the combination of models provides more stable and consistent results, although not as accurate as SVM in absolute terms.

Practically, the ensemble model is more suitable for institutional decision-making contexts because it minimizes the bias of a single model. By combining the strength of SVM in separating non-linear patterns and the ability of decision trees to recognize relationships between features, the ensemble produces predictions that are more adaptive to variations in student data from different provinces and educational backgrounds.

3.1.4. General Comparison of RBF, Random Forest, XGBoost, and Ensemble

Figure 5 shows a comparison of the performance of four machine learning models, SVM (RBF), Random Forest, XGBoost, and Ensemble (Voting Classifier), in predicting the graduation of students with disabilities in the UM-PTKIN selection using the metrics Accuracy, Precision, Recall, and F1-Score. In general, SVM (RBF) showed the best predictive performance, especially in terms of Recall (1.00) and F1-Score (0.88). The high recall indicates the model's ability to minimize false negatives, which is highly relevant in the context of inclusive education as it reduces the risk of overlooking participants who are actually eligible to graduate. Although the precision value (0.79) was slightly lower, the performance was still considered good given the limited size of the dataset.

Random Forest and XGBoost showed relatively similar performance with an Accuracy of 0.65 and an F1-Score of around 0.77–0.80. Although they did not exceed SVM in terms of predictive performance, both models had advantages in terms of interpretability, particularly through feature importance analysis. Variables such as province of origin, type of education, and disability category were identified as dominant factors influencing graduation predictions. This provides substantial added value because it not only generates predictions but also provides an analytical basis for policy evaluation.

Random Forest and XGBoost showed relatively similar performance with an Accuracy of 0.65 and an F1-Score of around 0.77–0.80. Although they did not surpass SVM in terms of predictive performance, both models had advantages in terms of interpretability, particularly through feature importance analysis. Variables

such as province of origin, type of education, and disability category were identified as dominant factors influencing graduation predictions. This provides substantial added value because it not only generates predictions but also provides an analytical basis for policy evaluation.

3.1.5. Findings

The results of this modeling provide several important findings, including that provincial factors and type of education influence graduation rates. Features such as province of origin, type of high school/vocational school/special needs school, and sensory or intellectual disability category emerged as important variables in determining graduation rates.

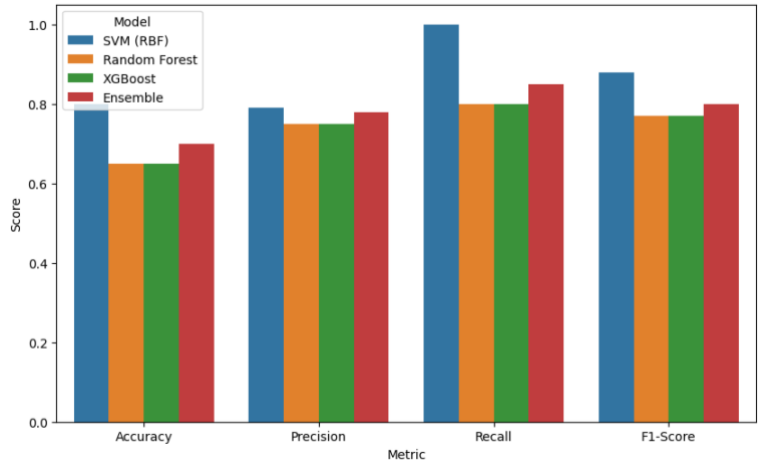

Figure 5. Performance Comparison of Models

SVM plays a very strong role in detecting graduation patterns. The perfect recall value of SVM shows that this algorithm is able to recognize patterns of students with disabilities who are likely to graduate very well, although it is sometimes too permissive towards cases of non-graduation. The Ensemble model is more suitable for long-term policies. Ensemble provides the best balance between sensitivity and accuracy. In the context of inclusive education policies, this model is safer to use because it reduces the risk of algorithmic bias.

Although model accuracy is important, the predictive approach in the context of disabled education does not merely assess "right or wrong," but rather how the system can help institutions understand the factors causing failure and improve admission policies. When compared to other models such as Random Forest and XGBoost, which tend to be more balanced in classification probabilities, SVM excels in sensitivity but lacks in probability calibration. Therefore, at the decision support system implementation stage, SVM results should be combined with probability-based models (e.g., Ensemble Voting) for more stable results.

4. CONCLUSION

This study successfully applied a CRISP-DM-based machine learning approach to predict the graduation of students with disabilities in the UM-PTKIN selection process. The analysis process included stages from business understanding to evaluation, with data imbalance handled using the SMOTE technique and the application of One Hot Encoding on categorical variables.

The evaluation results showed that the Support Vector Machine (SVM) model with RBF kernel provided the best performance with an accuracy of 0.80, recall of 1.00, and F1-score of 0.88, making it the most optimal in identifying successful participants. The Random Forest, XGBoost, and Ensemble Voting Classifier models also showed competitive performance, with their respective advantages in interpretability and prediction balance.

Feature importance analysis indicated that province of origin, highest level of education, and sensory disability category were the most influential factors on graduation. Overall, this study proves that the application of machine learning can be a decision-making tool in the selection system for students with disabilities, as well as support the development of a more inclusive, fair, and data-driven selection mechanism.

ACKNOWLEDGEMENTS

The authors would like to express their sincere gratitude to the Laboratory of the Department of Informatics at the UIN Sunan Gunung Djati, Indonesia for providing the facilities, resources, and support necessary to conduct this research. Their assistance and institutional support greatly contributed to the completion of this study.

REFERENCES

[1] N. Karki, A. Anwar, I. Ur Rehman, L. Husamaldin, and P. Saadati, "Smart Attendance Monitoring System Using Face Recognition for People with Disabilities (PwDs)," *Proceedings of 2023 IEEE International Smart Cities Conference, ISC2 2023*, 2023, doi: 10.1109/ISC257844.2023.10293664.

[2] A. Anwar, I. Ur Rehman, Ijaz-Ul-Haq, and L. Husamaldin, "Smart Education for People with Disabilities (PwDs): Conceptual Framework for PwDs Emotions Classification from Student Utterances (SUs) during Online Learning," *ISC2 2022 - 8th IEEE International Smart Cities Conference*, 2022, doi: 10.1109/ISC255366.2022.9922083.

[3] M. Q. Huda, N. A. Hidayah, E. Khudzaeva, A. B. Lubis, Zulkifli, and I. Sujoko, "IT Adoption and Implementation on Pre-Implementation Phase: The Case of State Islamic University in Indonesia," *2021 9th International Conference on Cyber and IT Service Management, CITSM 2021*, 2021, doi: 10.1109/CITSM52892.2021.9587926.

[4] R. Findiana, E. M. Yuniarno, and Endroyono, "Classification of Graduates Student on Entrance Selection Public Higher Education through Report Card Grade Path Using Support Vector Machine Method," *2020 3rd International Conference on Information and Communications Technology, ICOIACT 2020*, pp. 7–11, Nov. 2020, doi: 10.1109/ICOIACT50329.2020.9332072.

[5] E. Khudzaeva, N. A. Hidayah, and Q. Aini, "Evaluation Model The Successful Use of Information Technology in Distance Learning at State Islamic Universities (PTKIN)," *2023 11th International Conference on Cyber and IT Service Management, CITSM 2023*, 2023, doi: 10.1109/CITSM60085.2023.10455535.

[6] E. N. Muratov *et al.*, "When machine learning models learn chemistry II: applying WISP to real-world examples," *Digital Discovery*, vol. 5, no. 2, pp. 583–591, Jan. 2026, doi: 10.1039/d0cs00098a.

[7] K. Janssen, J. M. Wollschläger, J. Proppe, and A. H. Göller, "When machine learning models learn chemistry I: quantifying explainability with matched molecular pairs," *Digital Discovery*, vol. 5, no. 2, pp. 571–582, Jan. 2026, doi: 10.1039/D5DD00398A.

[8] A. E. J. Bulstra *et al.*, "A Machine Learning Algorithm to Estimate the Probability of a True Scaphoid Fracture After Wrist Trauma," *J. Hand Surg. Am.*, vol. 50, no. 6, pp. 711–720, Jun. 2025, doi: 10.1016/j.jhsa.2025.01.021.

[9] D. R. Ramdania, C. Slamet, H. F. Lutfiyah, M. Irfan, W. B. Zulfikar, and M. Harika, "Classification and Regression Trees (CART) Method in the Classification Process of Blood Control Candidates," *Proceeding of 2023 9th International Conference on Wireless and Telematics, ICWT 2023*, 2023, doi: 10.1109/ICWT58823.2023.10335325.

[10] W. B. Zulfikar, Angelyna, M. Irfan, A. R. Atmadja, and Jumadi, "A Deep Learning Approach Using VGG16 to Classify Beef and Pork Images," *JOIV : International Journal on Informatics Visualization*, vol. 9, no. 2, pp. 568–574, Mar. 2025, doi: 10.62527/joiv.9.2.2848.

[11] W. Wang, W. Xu, X. Yao, and H. Wang, "Application of Data-driven Method for Automatic Machine Learning in Economic Research," *Proceedings - 2022 21st International Symposium on Distributed Computing and Applications for Business Engineering and Science, DCABES 2022*, pp. 42–45, 2022, doi: 10.1109/DCABES57229.2022.00019.

[12] R. Pandiarajan, J. Jagannathan, P. S. Ramesh, A. Ponmalar, and Sudha, "Advanced Machine Learning Algorithms for 1Predictive Analytics in Healthcare to Enhance Patient Outcomes with Data-Driven Insights," *IEEE International Conference on Recent Advances in Science and Engineering Technology, ICRASET 2024*, 2024, doi: 10.1109/ICRASET63057.2024.10895597.

[13] P. Yesankar, C. Puri, A. Barahate, P. M. Gote, J. Hajbe, and A. Pawar, "Machine Learning in Healthcare: A Review of Current Applications and Future Trends," *2nd International Conference on Machine Learning and Autonomous Systems, ICMLAS 2025 - Proceedings*, pp. 482–487, 2025, doi: 10.1109/ICMLAS64557.2025.10968281.

[14] R. Pandiarajan, J. Jagannathan, P. S. Ramesh, A. Ponmalar, and Sudha, "Advanced Machine Learning Algorithms for 1Predictive Analytics in Healthcare to Enhance Patient Outcomes with Data-Driven Insights," *IEEE International Conference on Recent Advances in Science and Engineering Technology, ICRASET 2024*, 2024, doi: 10.1109/ICRASET63057.2024.10895597.

[15] K. K. Hiran, D. Khazanchi, A. K. Vyas, and S. Padmanaban, "Machine learning for sustainable development," *Machine Learning for Sustainable Development*, pp. 1–201, Jul. 2021, doi: 10.1515/9783110702514.

[16] Y. Liang, J. Wang, and J. Wang, "Reconstructing Higher Education in the Big Data and AI Era: Interdisciplinary Integration and Problem-Driven Talent Cultivation," *2025 10th International Conference on Distance Education and Learning, ICDEL 2025*, pp. 47–51, 2025, doi: 10.1109/ICDEL65868.2025.11193480.

[17] Y. Huang, "FS-SMOTE: An Improved SMOTE Method Based on Feature Space Scoring Mechanism for Solving Class-Imbalanced Problems," *IEEE Access*, vol. 13, pp. 148074–148082, 2025, doi: 10.1109/ACCESS.2025.3597794.

[18] H. Insan, S. S. Prasetiyowati, and Y. Sibaroni, "SMOTE-LOF and Borderline-SMOTE Performance to Overcome Imbalanced Data and Outliers on Classification," *2023 3rd International Conference on Intelligent Cybernetics Technology and Applications, ICICyTA 2023*, pp. 136–141, 2023, doi: 10.1109/ICICyTA60173.2023.10428902.

[19] D. Chaerul Ekty Saputra, K. Sunat, and T. Ratnaningsih, "SMOTE-MRS: A Novel SMOTE-Multiresolution Sampling Technique for Imbalanced Distribution to Improve Prediction of Anemia," *IEEE Access*, vol. 12, pp. 154675–154699, 2024, doi: 10.1109/ACCESS.2024.3482968.

[20] A. Ahmad *et al.*, "Vehicle Recognition using Multi-Layer Perceptron and SMOTE Technique," *Proceedings - 2022 2nd International Conference of Smart Systems and Emerging Technologies, SMARTTECH 2022*, pp. 190–193, 2022, doi: 10.1109/SMARTTECH54121.2022.00049.

[21] P. Lin, "Study on Real-Time Gesture Recognition Using Convolutional Neural Network Based on SMOTE Method," *2024 6th International Conference on Applied Machine Learning (ICAML)*, pp. 53–58, Jul. 2024, doi: 10.1109/ICAML64299.2024.00020.

[22] L. Wei, "Genetic Algorithm Optimization of Concrete Frame Structure Based on Improved Random Forest," *Proceedings - 2023 International Conference on Electronics and Devices, Computational Science, ICEDCS 2023*, pp. 249–253, 2023, doi: 10.1109/ICEDCS60513.2023.00051.

[23] Y. Wang, L. Liu, Y. Zhao, and X. Zhao, "Forest Resource Conservation Strategy with the Introduction of Random Forest Model Supported by Digital Technology," *2024 6th International Conference on Applied Machine Learning (ICAML)*, pp. 274–279, Jul. 2024, doi: 10.1109/ICAML64299.2024.00056.

[24] W. Martins, L. B. Bagesteiro, T. O. Weber, and A. Balbinot, "FPGA-based Implementation of Random Forest Classifier for sEMG Signal Classification," *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS*, 2024, doi: 10.1109/EMBC53108.2024.10781521.

[25] M. Irfan, W. B. Zulfikar, R. A. Ramadanti, A. Wahana, and Y. A. Gerhana, "Machine learning approach for scholarship candidate selection on Islamic State University in Indonesia," *AIP Conf. Proc.*, vol. 2646, no. 1, Apr. 2023, doi: 10.1063/5.0130047.

[26] I. Lindra, Hendrawan, and A. Subekti, "A Systematic Review Network Slicing: Revenue Optimization for Slice Admission Control Objectives," *Proceeding of 2024 the 10th International Conference on Wireless and Telematics, ICWT 2024*, 2024, doi: 10.1109/ICWT62080.2024.10674734.

[27] D. A. Navastara, Y. Millaturrosyidah, T. P. Ramadhany, N. Suciati, C. Fatichah, and W. Suadi, "CTAB-GAN and Ensemble Learning for Predicting Student Admission in National Selection Based on Achievement," *2025 15th International Conference on Information & Communication Technology and System (ICTS)*, pp. 1–6, Nov. 2025, doi: 10.1109/ICTS67612.2025.11369638.

[28] S. Sujana, A. N. Kumar, L. Aditya Sree Charan, and G. Nishitha, "Comparative Analysis of Graduate Admission Predication using Neural Network," *2024 Asia Pacific Conference on Innovation in Technology, APCIT 2024*, 2024, doi: 10.1109/APCIT62007.2024.10673490.

[29] Y. Zhang, Y. Zhao, M. Liu, and D. Yang, "Analysis and Prediction of Abnormal Behavior of higher Vocational Students Based on XGBoost Model," *2025 8th International Conference on Advanced Algorithms and Control Engineering, ICAACE 2025*, pp. 1900–1904, 2025, doi: 10.1109/ICAACE65325.2025.11019167.

[30] S. Jeganathan, S. Parthasarathy, A. R. Lakshminarayanan, P. M. Ashok Kumar, and M. K. A. Khan, "Predicting the post graduate admissions using classification techniques," *2021 International Conference on Emerging Smart Computing and Informatics, ESCI 2021*, pp. 346–350, Mar. 2021, doi: 10.1109/ESCI50559.2021.9396815.

[31] S. Nalam, M. Aleemuddin, Y. T. Kadari, and M. Nanda Kumar, "Advanced Graduate Admission Prediction," *2023 IEEE 8th International Conference for Convergence in Technology, I2CT 2023*, 2023, doi: 10.1109/I2CT57861.2023.10126307.