Journal of Applied Informatics and Computing (JAIC)
Vol.10, No.2, April 2026, pp. 1369~1382
e-ISSN: 2548-6861

A Probabilistic Ensemble-Based Decision Support Framework for Teacher Promotion Assessment

Ray Eka Novasani [1]*, MY Teguh Sulistyono [2]*
* Faculty of Computer Science, Universitas Dian Nuswantoro
112202206857@mhs.dinus.ac.id [1], teguh.sulistyono@dsn.dinus.ac.id [2]

Article Info

Article history:

Received 2026-01-26
Revised 2026-03-07
Accepted 2026-04-10

Keyword:

Brier Score,
Ensemble Learning,
Probabilistic Prediction,
Teacher Promotion,
Machine Learning.

ABSTRACT

This study proposes a probabilistic ensemble-based decision support framework for analyzing teacher promotion eligibility within the institutional Credit Point Assessment system. The dataset consists of 20 finalized teacher promotion records collected retrospectively from the institutional personnel administration unit covering the 2022–2024 assessment period. All personal identifiers were removed prior to analysis to ensure ethical compliance and data confidentiality. Data preprocessing included categorical variable transformation using One-Hot Encoding and numerical feature standardization through Min–Max normalization. The dataset was divided using stratified sampling to preserve class distribution, and preprocessing procedures were applied exclusively to the training data to prevent data leakage. Probabilistic predictions were generated using Random Forest and Extreme Gradient Boosting (XGBoost), and combined through a soft voting ensemble strategy to enhance robustness. Model performance was evaluated using confusion-matrix-based metrics, ROC-AUC, and probability calibration analysis through the Brier Score. Among the evaluated models, XGBoost achieved the lowest Brier Score (0.2034), indicating superior probability calibration, while the ensemble model demonstrated more stable classification behavior. Feature importance analysis identified cumulative credit points and professional development activities as dominant predictors, whereas demographic attributes showed minimal influence. Rather than serving as an automated decision-making mechanism, the proposed framework functions as a decision-support tool by providing interpretable probability estimates of promotion eligibility. Given the limited sample size and institutional data constraints, findings are intended to support analytical interpretation within a specific organizational context rather than broad predictive generalization.

I. INTRODUCTION

Teacher promotion is a critical component of human resource management in the education sector, as it reflects professional development, performance evaluation, and institutional accountability[1]. In Indonesia, teacher promotion is regulated through the Credit Point Assessment system, which evaluates administrative achievements, performance indicators, and continuous professional development activities[1]. This system is designed to ensure that promotion decisions are based on measurable and standardized criteria.

Despite its structured framework, the implementation of the Credit Point Assessment system presents several practical challenges. The assessment process involves multiple indicators with varying levels of completeness across teachers, leading to complex evaluation conditions. In many cases, promotion decisions are expressed as binary outcomes, such as eligible or not eligible, without explicitly representing how close a teacher is to meeting the promotion criteria. This limitation reduces the transparency of the evaluation process and restricts its usefulness for decision support[2]. In institutional promotion settings, reliance solely on manual administrative review may introduce subjectivity,

interpretative inconsistency, and limited visibility into how close candidates are to meeting eligibility criteria. A structured data-driven approach can enhance objectivity by quantifying promotion likelihood while maintaining administrative oversight.

Recent developments in machine learning have facilitated the adoption of data-driven analytical methods in educational decision-making. By extracting patterns from complex administrative and performance-related data, machine learning models can support evaluators in interpreting relationships among multiple assessment variables[3]. However, many existing studies focus primarily on deterministic classification and accuracy-based evaluation metrics, which may be unsuitable for small-scale and imbalanced datasets commonly found in institutional promotion records[4].

In the context of teacher promotion analysis, probabilistic prediction offers a more appropriate alternative[1][5]. Probability-based outputs allow promotion eligibility to be interpreted as a likelihood rather than a definitive decision[6]. Such an approach aligns with administrative practices that require flexibility, transparency, and professional judgment. Probabilistic modeling can therefore function as a decision-support mechanism rather than an automated replacement for formal evaluation procedures[7][8]. However, limited attention has been given to probability-calibrated ensemble modeling for teacher promotion assessment using small-scale, private administrative datasets, particularly within the context of the Indonesian Credit Point Assessment system.

Based on these considerations, this study applies probabilistic machine learning methods to analyze teacher promotion eligibility using administrative and performance data. Random Forest and Extreme Gradient Boosting (XGBoost) models are employed to generate probability estimates of promotion eligibility[5][9]. To enhance prediction stability and reduce model-specific bias, an ensemble approach using soft voting is implemented by averaging probabilistic outputs from both models[9][10].

Model evaluation emphasizes probability calibration rather than classification accuracy. The Brier Score is utilized to assess the quality of predicted probabilities, while classification behavior was examined through confusion matrix analysis to observe prediction tendencies across promotion categories. In addition, feature importance analysis is conducted to identify the most influential factors contributing to promotion eligibility, supporting interpretability and transparency in the analytical process[6][4].

This study utilizes a private institutional dataset obtained with formal authorization. All personal identifiers, including names and identification numbers, were removed prior to analysis. The dataset was used exclusively for academic research purposes and was not shared or published in raw form. The analysis focuses on aggregated patterns and model interpretation rather than individual-level evaluation.

The primary contribution of this study is the development of a probabilistic and interpretable analytical framework for supporting teacher promotion assessment[11]. Rather than claiming generalizable predictive performance, this research emphasizes decision support, probability interpretation, and policy relevance within the limitations of a small and institution-specific dataset[5].

II. METHODS

This section describes the main research stages adopted in this study, as presented in Figure 1, to describe the overall methodological framework applied throughout the analysis. This study adopts a quantitative research design to predict the likelihood of teacher promotion based on Credit Point Assessment using probabilistic machine learning methods[12]. The research procedure is depicted through a flow diagram that summarizes the main analytical stages, including data acquisition, preprocessing, probabilistic model development, ensemble-based prediction, and evaluation processes[1][5][9]. Random Forest and XGBoost models are employed to generate probabilistic outputs, which are combined using a soft voting ensemble approach. Model performance is evaluated using Brier Score, and confusion-matrix-based metrics, while feature importance analysis is conducted to enhance interpretability[9][13].

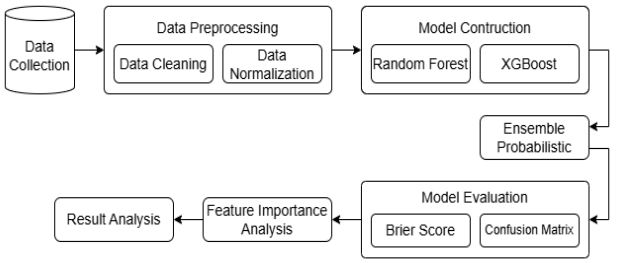

Figure 1. Research Flow Diagram

A. Data Collection

The dataset was obtained retrospectively from the institutional personnel administration unit following formal authorization for academic research purposes. The records cover teacher promotion assessments conducted between 2022 and 2024 within a single institutional cohort.

The dataset consists of 20 teacher promotion assessment records obtained from finalized administrative and performance evaluation documents used in the institutional Credit Point Assessment process. The variables include accumulated credit points, professional development activities, SKP scores, and years of service.

The class distribution is imbalanced, with 17 cases categorized as eligible for promotion and 3 cases categorized as non-eligible. This imbalance reflects the practical characteristics of institutional promotion records, where most candidates typically meet minimum administrative

requirements. All predictor variables were recorded prior to the official promotion decision to prevent data leakage. The target variable (Promotion_Label) was derived from the documented final promotion outcome. The dataset represents a single promotion cycle and does not include longitudinal tracking across multiple years.[14].

TABLE I
DATA COLLECTION

No	Teacher Code	Years of Service	Last Rank	SKP 2023	SKP 2024	Total Training (Hours)	Total Credit Points	Promotion Status
1	GURU1	11	III/c	87	87	32	204.5	Considered
2	GURU2	16	III/c	87	86	293	295.8	Considered
3	GURU3	15	III/d	87	87	64	168.5	Considered
4	GURU4	23	IV/a	25	25	264	57.1	Not Considered
5	GURU5	20	III/d	86	86	64	227	Considered
…	…	…	…	…	…	…	…	…
…	…	…	…	…	…	…	…	…
18	GURU18	21	III/c	86	87	128	275.4	Considered
19	GURU19	28	III/c	83	84	96	260.2	Considered
20	GURU20	22	III/d	80	81	32	145.6	Not Considered

B. Data Preprocessing

To improve data suitability for machine learning analysis, a preprocessing stage was conducted. The process involved data cleaning to ensure dataset consistency, transformation of categorical attributes into numerical representations, and normalization of numerical features using Min–Max scaling to standardize value ranges. These preprocessing procedures help stabilize model learning and facilitate reliable predictive analysis[15]. Categorical variables were transformed using One-Hot Encoding, where each category level was converted into a binary indicator feature. This approach prevents ordinal misinterpretation and ensures compatibility with tree-based machine learning algorithms. Numerical features were standardized using Min–Max normalization, which rescales each variable into a fixed range of [0,1] using the transformation:

$$x^1 = (x - \min(x))/(\max(x) - \min(x))$$

To avoid data leakage, all preprocessing procedures including encoding and normalization were fitted exclusively on the training subset and subsequently applied to the testing subset. It is commonly acknowledged that these actions are crucial for enhancing predictive performance and stabilizing model learning.

1) *Data Cleaning*: The dataset is carefully examined to detect and remove missing values and duplicate entries. This step is critical to prevent distortions and biases that may compromise the accuracy and reliability of model training[15].

2) *Data Normalization*: All numerical features are preprocessed through Min-Max normalization, effectively transforming the raw values into a normalized range between 0 and 1, thereby preserving relative relationships while preventing the dominance of variables with larger magnitudes[15].

C. Feature Selection

Feature selection was conducted based on domain relevance to the institutional Credit Point Assessment framework. Variables representing key evaluation components such as accumulated credit points, professional development activities, SKP performance indicators, years of service, and administrative responsibilities were selected because they directly correspond to official promotion assessment criteria. Non-essential identifiers and descriptive attributes, including teacher names and administrative identifiers, were removed prior to modeling. This domain-driven feature selection strategy ensures that the predictive model reflects institutional evaluation logic rather than arbitrary statistical correlations.

D. Model Construction

Two supervised machine learning algorithms, Random Forest and Extreme Gradient Boosting (XGBoost), were trained using the preprocessed dataset during the model construction phase[16]. Comparative studies in classification and regression tasks supported the selection of these algorithms because of their excellent performance in managing nonlinear relationships and complex feature interactions. As is customary in machine learning to evaluate generalization performance, the dataset was split into training and testing subsets with a ratio of 80% for training and 20% for testing. Both models were set up to generate probabilistic outputs, which indicate the probability of a teacher being promoted based on the provided input features, in contrast to traditional binary classification techniques[17][18].

1). Estimating Random Forest Probabilities: An ensemble learning algorithm called Random Forest builds several decision trees and compiles their predictions. The average probability generated by each individual tree is used to determine the promotion probability in probabilistic prediction. Random Forest's aggregation technique enhances robustness and lowers variance[19]. The probability of promotion is expressed as follows if the Random Forest model contains T decision trees:

$$P_{RF}(y = 1|x) = \frac{1}{T}\sum_{t=1}^{T} P_t\,(y = 1|x)$$

This method uses ensemble averaging to reduce variance, allowing Random Forest to produce stable probability estimates[20].

2). Probability Estimation with XGBoost: An ensemble of decision trees is constructed sequentially by the gradient boosting algorithm XGBoost. A well-known method for probabilistic outputs in boosting frameworks, XGBoost uses a logistic (sigmoid) function to convert the model output into a probability value for binary classification problems[21]. The following is the definition of the promotion probability that XGBoost predicts:

$$P_{XGB}(y = 1|x) = \frac{1}{1 + e^{-f(x)}}$$

where:

$$f(x) = \sum_{k=1}^{K} f_k\,(x)$$

The predicted probability is guaranteed to fall within the interval [0,1] by the sigmoid function.

E. System Architecture and Workflow

The overall analytical workflow of the proposed probabilistic decision-support framework is illustrated in Figure 2. The architecture summarizes the sequential stages of the analytical process, beginning with institutional promotion data acquisition and ending with probability-based promotion decision support.

The process starts with the collection of teacher promotion assessment records obtained from institutional administrative documentation. The collected data undergo preprocessing procedures including data cleaning, categorical transformation, and Min–Max normalization to ensure consistent feature representation for machine learning analysis. Subsequently, feature selection is performed based on relevance to the institutional Credit Point Assessment framework. Variables representing official evaluation indicators such as accumulated credit points, professional development activities, performance scores, and years of service are retained as predictive features for model development. The processed dataset is then divided into training and testing subsets using stratified sampling to preserve the original class distribution. Two machine learning models, Random Forest and Extreme Gradient Boosting (XGBoost), are trained to generate probabilistic predictions of teacher promotion eligibility. The probability outputs generated by both models are combined using a soft voting ensemble strategy, where the final promotion probability is calculated as the average of the predicted probabilities produced by the individual models.

Finally, the ensemble probability outputs are evaluated using both discrimination-based and calibration-based evaluation metrics, including confusion matrix analysis, ROC-AUC, and the Brier Score. These probabilistic estimates are then interpreted using adjustable probability thresholds to support institutional promotion decision-making while maintaining human oversight.

The diagram illustrates the analytical pipeline from institutional promotion data collection, data preprocessing and feature selection, probabilistic model training using Random Forest and XGBoost, ensemble probability estimation through soft voting, model evaluation using calibration and discrimination metrics, and threshold-based decision support for promotion assessment.

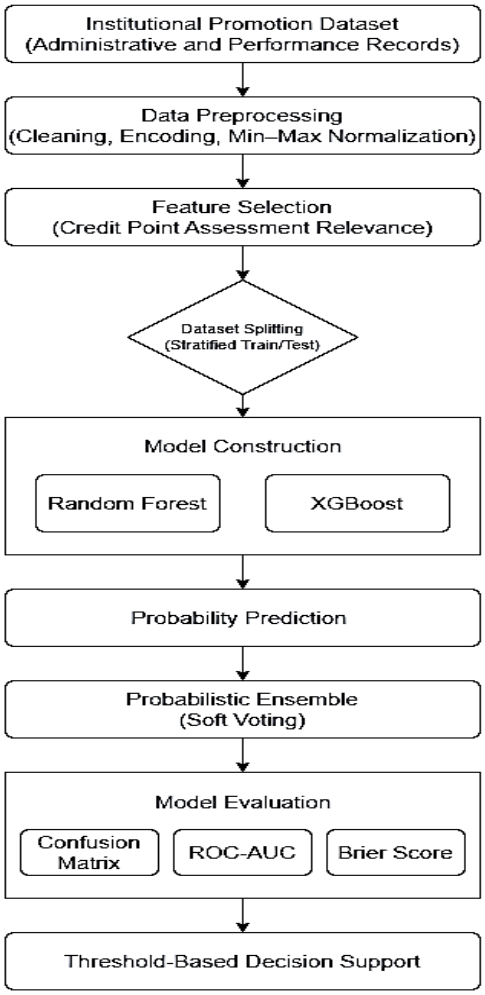

Figure 2. System Architecture and Analytical Workflow for Probabilistic Teacher Promotion Assessment

F. Model Validation Strategy

To reduce the risk of overfitting under limited sample conditions, the dataset was divided into training and testing subsets using stratified sampling with an 80:20 split ratio. Stratification was applied to preserve proportional representation of the minority class (17 eligible vs 3 non-eligible cases) in both subsets. Given the extremely limited minority class representation, full k-fold cross-validation was explored during preliminary experimentation. However, due to the scarcity of non-eligible cases, several folds produced statistically unstable discrimination metrics, particularly for ROC-AUC, when minority instances were absent in certain validation splits. Under such conditions, cross-validation results may become unreliable and sensitive to minor sample variations. Therefore, the final evaluation strategy emphasizes stratified hold-out validation combined with probability calibration assessment using the Brier Score. Calibration-based evaluation is particularly suitable for small and imbalanced datasets, as it provides a more stable measure of probabilistic reliability compared to discrimination-

focused metrics. This validation design ensures methodological transparency while acknowledging statistical constraints inherent in limited institutional datasets.

G. Probabilistic Ensemble Prediction

The probabilistic outputs of Random Forest and XGBoost were combined using the soft voting method in an ensemble approach to increase prediction robustness and decrease model bias[20]. When compared to single models, ensemble models that average predictions frequently produce better generalization and stability.

The average of the probabilities produced by the two models is used to calculate the final probability of teacher promotion:

$$P_{final}(y = 1|x) = \frac{P_{RF}(y = 1|x) + P_{XGB}(y = 1|x)}{2}$$

H. Model Evaluation

The performance of the probabilistic classification model was evaluated using a set of well-established evaluation metrics frequently applied in binary classification and probabilistic forecasting studies. This evaluation framework includes confusion-matrix-derived measures to quantify classification accuracy and error patterns. Model performance evaluation focused on probabilistic calibration using the Brier Score, complemented by confusion-matrix-based analysis[22]. Given the probabilistic nature of the proposed approach and the limited sample size, calibration-based evaluation using the Brier Score was emphasized. Collectively, these evaluation measures offer a multidimensional perspective on model performance, enabling both the differentiation and probabilistic interpretation of teacher promotion outcomes[23].

1). Confusion Matrix: Based on the final probability value, a decision is applied to determine the predicted class:

$$\hat{y} = \begin{cases} 1, & \text{if } P_{final} \geq 0.5 \\ 0, & \text{if } P_{final} < 0.5 \end{cases}$$

A confusion matrix, which shows how well a classifier classifies a subset of the test data, is used to summarize the classification results. The confusion matrix is a two-dimensional matrix where the actual class labels are represented by one dimension and the classifier's predicted class labels are represented by the other[24]. Four possible outcomes make up the matrix In this study, classification outcomes for teacher promotion eligibility were analyzed using standard binary indicators, namely True Positive, False Positive, True Negative, and False Negative, to examine prediction behavior under institutional data constraints. One class is referred to as the positive class, and the other as the negative class. In machine learning research, this evaluation tool is frequently used to assess classification performance and obtain performance metrics like recall, accuracy, and precision[25].

2). Brier Score for Probability Evaluation: Since this study emphasizes probabilistic prediction, the Brier Score was employed to evaluate the accuracy of predicted probability estimates. To assess the reliability of promotion likelihood estimates, this study employed the Brier Score, which evaluates the squared deviation between predicted probabilities and observed promotion outcomes, thereby assessing the calibration quality of the model. Lower Brier Score values indicate more accurate and well-calibrated probability predictions, which are essential in probability-based decision support systems[26].

$$BS = \frac{1}{N}\sum_{i=1}^{N}(p_i - y_i)^2$$

Where p_i denotes the predicted probability and y_i represents the actual class label. The Brier Score is widely used in probabilistic forecasting and classification studies.

3). ROC-AUC: To evaluate the discrimination capability of the predictive models, Receiver Operating Characteristic (ROC) analysis was employed. The ROC curve illustrates the trade-off between the True Positive Rate (TPR) and False Positive Rate (FPR) across varying probability thresholds. The Area Under the ROC Curve (ROC-AUC) summarizes the overall discrimination performance of the model, where higher values indicate better separability between eligible and non-eligible promotion classes.

$$Positive\ Rate(TPR) = \frac{TP}{TP + FN}$$

TPR, also known as sensitivity or recall, measures the proportion of correctly identified eligible promotion cases among all actual eligible cases.

$$False\ Positive\ Rate\ (FPR) = \frac{FP}{FP + TN}$$

FPR measures the proportion of incorrectly predicted eligible cases among all actual non-eligible cases.

$$Area\ Under\ the\ Curve\ (AUC) = \int_0^1 TPR(FPR)\ d(FPR)$$

AUC represents the integral of the ROC curve and indicates the overall discrimination performance across all threshold values. An AUC value closer to 1 indicates stronger separability between classes, while a value near 0.5 suggests random classification behavior[10].

4). Decision Threshold Rule: To convert probabilistic predictions into categorical promotion decisions, a probability threshold is applied. Instances with predicted probabilities

greater than or equal to the specified threshold are classified as eligible for promotion, while those below the threshold are categorized as non-eligible.

$$\hat{y} = \begin{cases} 1, & \text{if } P(y=1|x) \geq \tau \\ 0, & \text{if } P(y=1|x) < \tau \end{cases}$$

I. Performance Measurement and Analysis

Three standard performance metrics were used: accuracy, precision, and recall, based on these results. Accuracy measures the proportion of correctly classified instances relative to the total number of observations and is defined as:

Accuracy represents the proportion of correctly classified instances, both positive and negative, relative to the total number of observations[27]. It provides a general measure of the model's overall classification performance and is calculated as:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Precision reflects the reliability of positive predictions by indicating how many of the instances predicted as positive are actually positive[27]. This metric is particularly important when the cost of false positive predictions needs to be minimized:

$$Precision = \frac{TP}{TP + FP}$$

Recall, also known as sensitivity, measures the model's capability to correctly identify all actual positive instances[27]. A higher recall value indicates that fewer positive cases are incorrectly classified as negative:

$$Recall = \frac{TP}{TP + FN}$$

Collectively, these metrics offer a comprehensive evaluation of classification performance and are widely applied to assess accuracy, prediction reliability, and sensitivity in probabilistic classification models[28].

J. Feature Importance Analysis

To enhance model interpretability, feature importance analysis was conducted to identify the most influential variables affecting the probability of teacher promotion. Feature importance analysis provides insights into the relative contribution of each predictor to the model's probabilistic outputs, thereby supporting transparent and explainable decision-making[29]. Model interpretability has become an essential aspect of modern machine learning research, particularly in applications involving policy and educational decision support.

III. Result and Discussion

A. Overview of Experimental Results

This chapter presents the results obtained from the implementation of the proposed research framework. The discussion follows the same sequence as the methodological flow described in Chapter II, starting from descriptive analysis, exploratory visualization, model construction, evaluation, and ending with interpretative analysis. The objective of this chapter is not to claim predictive superiority, but to analyze probability-based promotion eligibility and identify key influencing factors based on model outputs.

B. Dataset Characteristics and Descriptive Analysis

1. Dataset Structure and Feature Composition:
After preprocessing, the final dataset consists of 20 teacher records with 15 selected features relevant to promotion assessment. The features represent administrative status, performance evaluation, professional development, and accumulated credit points.

TABLE II
Dataset Structure After Preprocessing

Category	Description
Number of records	20 teachers
Target variable	Promotion_Label (binary)
Eligible class	17 (Can be Considered)
Non-eligible class	3 (Not Yet Considered)
Total features	15
Feature types	Administrative, performance, training, credit points
Missing values	None (after preprocessing)

2. Descriptive Statistics of Numerical Variables:
Descriptive statistics were calculated to summarize the numerical attributes. The results indicate that most teachers have long working experience and relatively high performance scores, while variability is mainly observed in training-related variables and accumulated credit points.

TABLE III
Descriptive Statistics of Numerical Variables

Variable	Mean	Std. Dev	Min	Max
Length of Work (years)	21.55	6.21	11	37
SKP Grade 2023	78.2	22.91	0	87
SKP Grade 2024	78.5	22.99	0	88
Total Training (hours)	103.45	96.32	0	293
Total Credit Points	272.48	116.71	57.12	454.79

C. Target Variable Distribution

1. Construction of Promotion_Label:
The target variable, referred to as *Promotion_Label*, was constructed to represent promotion eligibility. Teachers categorized as "Can be Considered" were labeled as eligible (1), while others were labeled as not eligible (0).

2. Distribution Analysis:
The distribution of the target variable shows a clear imbalance, with 17 eligible cases and 3 non-eligible cases. This imbalance supports the use of probability-based evaluation metrics rather than accuracy-focused measures.

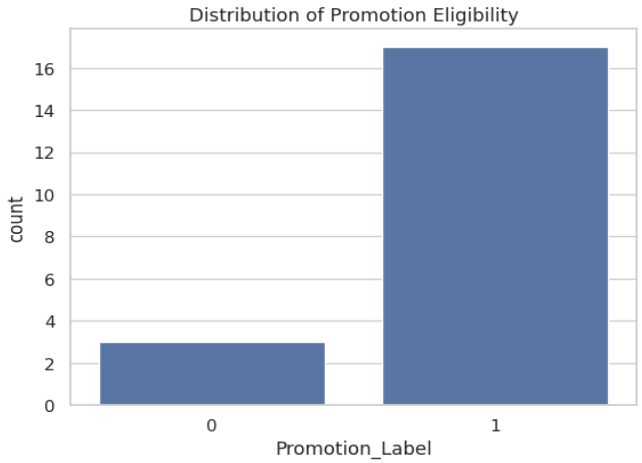

Figure 3.1 Distribution of promotion eligibility

D. Exploratory Data Analysis

Total Credit Points and Promotion Status:
Exploratory visualization indicates that teachers with higher Total Credit Points tend to fall into the eligible promotion category. However, overlapping ranges between eligible and non-eligible groups suggest that credit points alone are not sufficient to determine promotion outcomes.

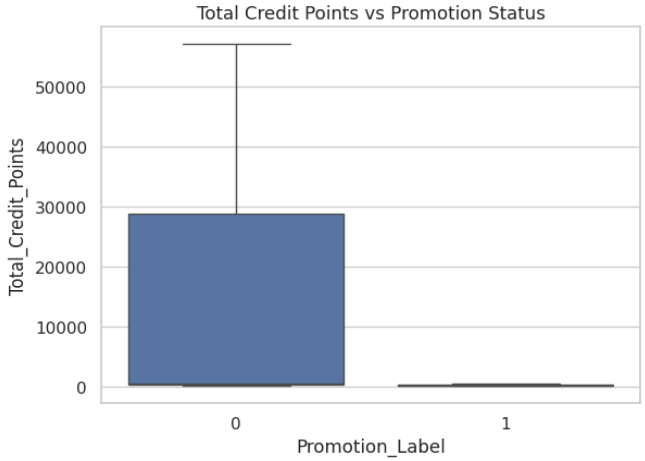

Figure 3.2 Total Credit Points vs Promotion Status

Training Hours and Credit Accumulation:
The relationship between training hours and total credit points shows a positive association, indicating that professional development activities contribute to credit accumulation. Nevertheless, similar training durations may result in different promotion outcomes, highlighting the influence of additional factors.

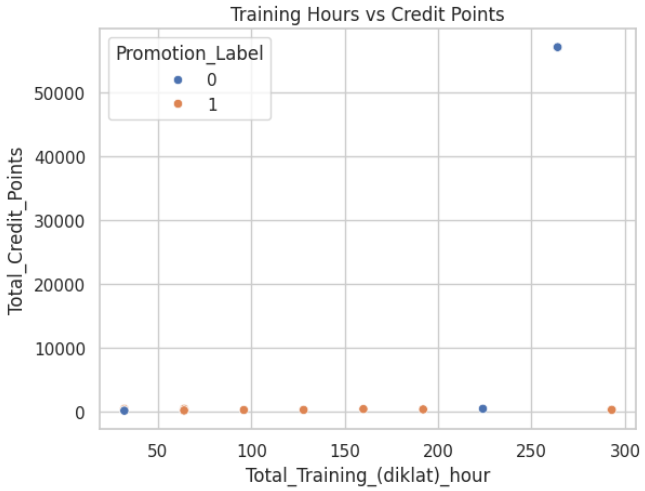

Figure 3.3 Training Hours vs Total Credit Points

1. *SKP 2024 and Promotion Eligibility:*
The comparison of SKP scores in 2024 shows that eligible teachers generally obtain higher scores. However, score dispersion across both groups indicates that SKP performance contributes moderately and does not solely determine promotion eligibility.

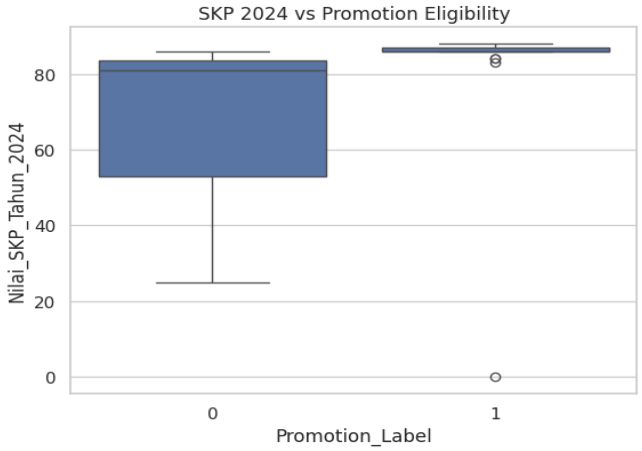

Figure 3.4 SKP 2024 vs Promotion Eligibility

E. Model Construction Results

1. Model Implementation:
This study employed three modeling strategies: Random Forest, XGBoost, and a probabilistic ensemble framework. To preserve the original distribution of promotion outcomes, the dataset was divided into training and testing subsets using an 80:20 ratio with stratified sampling. This approach ensured consistent class proportions throughout the model training and evaluation stages.

2. *Ensemble Probability Modeling:*
The ensemble approach combines the probability outputs generated by the Random Forest and XGBoost models through averaging, after which a threshold-based decision rule is applied to facilitate probability-focused interpretation.

F. Confusion Matrix Analysis

1. Random Forest Classification Results:
The confusion matrix results indicate that the Random Forest model correctly classified the majority of eligible cases, while a small number of misclassifications can be attributed to the limited size of the testing subset.

TABLE IV
CONFUSION MATRIX ON RANDOM FOREST

Actual \ Predicted	Not Eligible	Eligible
Not Eligible	0	1
Eligible	0	3

3. *XGBoost Classification Results:*
The XGBoost model demonstrates more balanced predictions, indicating improved handling of probability estimation compared to Random Forest.

TABLE V
CONFUSION MATRIX ON XGBOOST

Actual \ Predicted	Not Eligible	Eligible
Not Eligible	1	0
Eligible	0	3

4. *Ensemble Classification Results:*
The ensemble model produces more stable classification outcomes by reducing extreme predictions from individual models.

TABLE VI
CONFUSION MATRIX ON ENSEMBLE PROBABILITY

Actual \ Predicted	Not Eligible	Eligible
Not Eligible	1	0
Eligible	0	3

G. Probability Calibration Evaluation

1. Brier Score Comparison:
Model performance was assessed using the Brier Score as an indicator of probability prediction accuracy. Among the evaluated models, XGBoost produced the lowest Brier Score, reflecting the highest reliability in probability estimation, followed by the ensemble approach and the Random Forest model.

TABLE VII
COMPARISON OF BRIER SCORE VALUES

Model	Brier Score
Random Forest	0.2559
XGBoost	**0.2034**
Ensemble Probability	0.2269

The Brier Score ranges from 0 to 1, where lower values indicate better probabilistic calibration. A value of 0 represents perfect probability estimation, meaning that predicted probabilities exactly match observed outcomes. Conversely, higher values indicate larger deviations between predicted probabilities and actual class labels. In this study, XGBoost achieved the lowest Brier Score (0.2034), followed by the ensemble model (0.2269) and Random Forest (0.2559). These results indicate that XGBoost produced the most reliable probability estimates under the given dataset conditions. However, the interpretation of absolute Brier Score values must consider the pronounced class imbalance (17 eligible and 3 non-eligible cases). In imbalanced datasets, baseline probability estimation may inherently favor the majority class, potentially lowering the Brier Score without necessarily improving discriminatory capacity. Therefore, the evaluation emphasizes relative comparison across models rather than absolute threshold interpretation.

2. Calibration Curve Analysis:

To further examine probability reliability, a calibration curve was generated. The calibration curve compares predicted

probabilities with observed outcome frequencies across probability intervals. A perfectly calibrated model should follow the diagonal reference line, indicating that predicted probabilities match actual outcome frequencies.

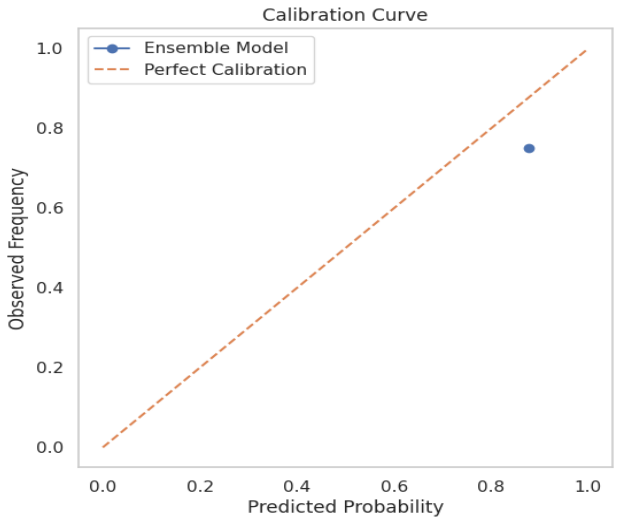

Figure 3.5 Calibration Curve of Ensemble Promotion Probability

Comparison between predicted promotion probabilities and observed outcome frequencies. The dashed diagonal line represents perfect probability calibration.

Based on Figure X, the ensemble model produces predicted promotion probabilities concentrated around high probability values (approximately 0.8–0.9). The observed outcome frequency in this region is approximately 0.75, indicating that the model slightly overestimates promotion likelihood but still maintains relatively consistent probability predictions. Because the test subset contains only four observations, the calibration curve contains a limited number of probability bins. Therefore, the visualization should be interpreted as an indicative assessment of calibration behavior rather than a statistically robust estimate.

3. Reliability Diagram:

To complement the calibration curve, a reliability diagram was constructed to illustrate the distribution of predicted probabilities produced by the ensemble model. The reliability diagram is shown in Figure 3.6

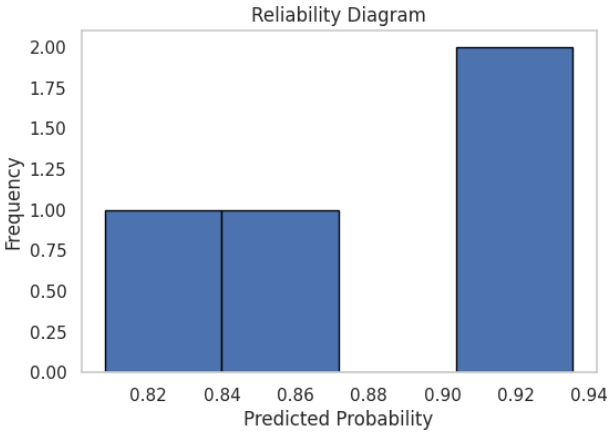

Figure 3.6 Reliability Diagram of Ensemble Promotion Probability

Histogram showing the distribution of predicted promotion probabilities across calibration bins.

As shown in Figure Y, the predicted probabilities are primarily concentrated in the 0.80–0.95 range, indicating that the ensemble model consistently assigns relatively high promotion likelihood scores to the evaluated cases. This distribution suggests that the model identifies strong promotion signals within the feature space, particularly for variables related to accumulated credit points and professional development indicators.

However, because the dataset contains a small number of test instances, the probability distribution should be interpreted cautiously. The reliability diagram mainly illustrates the tendency of the model to produce high-confidence predictions rather than providing a comprehensive calibration assessment.

H. Discrimination Performance Evaluation

In addition to calibration analysis, model discrimination capability was assessed using the Receiver Operating Characteristic Area Under the Curve (ROC-AUC). ROC-AUC evaluates a model's ability to rank positive instances higher than negative ones across varying probability thresholds.The obtained ROC-AUC values were 0.33 for Random Forest, 0.50 for XGBoost, and 0.33 for the ensemble model. Although XGBoost demonstrated slightly improved ranking performance, interpretation of ROC-AUC values must be approached cautiously. Due to the extremely limited minority class representation (n = 3) and small test subset size, ROC-AUC becomes highly sensitive to individual prediction variations and may appear discrete rather than continuous.
Under such small-sample conditions, discrimination metrics may not fully capture generalization capacity. Therefore, ROC-AUC is reported as a complementary evaluation measure, while calibration-based assessment using the Brier Score remains the principal analytical focus of this study.

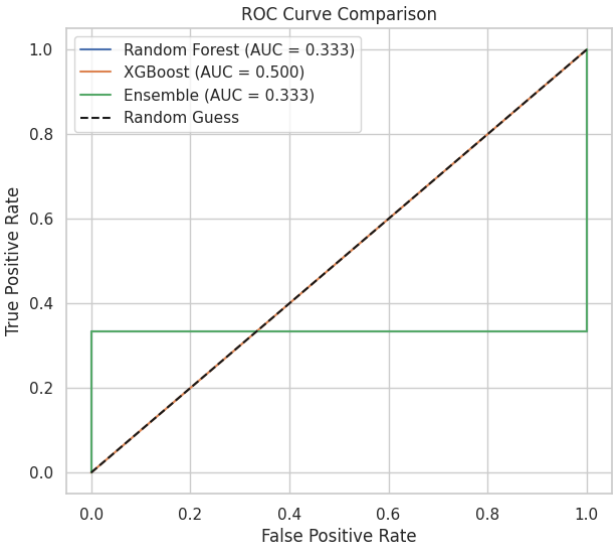

Figure 3.7 ROC Curve Comparison of Random Forest, XGBoost, and Ensemble Models

The ROC curves illustrate the discrimination behavior of each model. The stepwise pattern observed in the curves reflects the limited minority class samples and restricted test subset size. Differences between models remain modest under these data constraints.

I. Feature Importance Analysis

1. Key Influencing Factors:

Feature importance analysis was conducted using the Random Forest model to identify variables that most strongly contribute to prediction performance. As presented in Table VIII, Total Credit Points emerges as the most influential predictor (0.2401), followed by training-related variables such as Total Training Hours (0.2257) and Total Training Count (0.1194). Performance-based indicators, including SKP Grade 2023 and SKP Grade 2024, also demonstrate substantial predictive contribution.

TABLE VIII
FEATURE IMPORTANCE RANKING

Rank	Feature	Importance
1	Total Credit Points	0.2401
2	Total Training Hours	0.2257
3	Total Training Count	0.1194
4	SKP Grade 2023	0.1034
5	SKP Grade 2024	0.0716
6	Length of Work	0.0693
7	Service Orientation	0.0551

The prominence of credit accumulation and professional development variables reflects the structural logic embedded within the institutional Credit Point Assessment system. The model appears to capture established evaluation patterns rather than introducing independent predictive determinants. In contrast, demographic and secondary attributes contribute minimally to prediction performance, indicating that institutional performance metrics dominate eligibility estimation.

However, it is important to emphasize that feature importance values represent statistical contribution to prediction accuracy and should not be interpreted as evidence of causal influence. Tree-based ensemble models quantify how much each variable reduces prediction error within the observed dataset, but they do not establish cause-and-effect relationships.

For instance, while Total Credit Points demonstrates the highest importance score, this does not imply that credit accumulation independently causes promotion. Rather, it indicates that historical institutional decisions strongly align with accumulated credit metrics. The predictive framework therefore reflects correlational patterns embedded in administrative records, not experimentally validated causal mechanisms. Distinguishing between correlation and causation would require longitudinal or causal inference approaches beyond the scope of this study.

2. Visualization of Feature Importance:

The feature importance visualization further illustrates the relative contribution of each variable to promotion eligibility prediction.

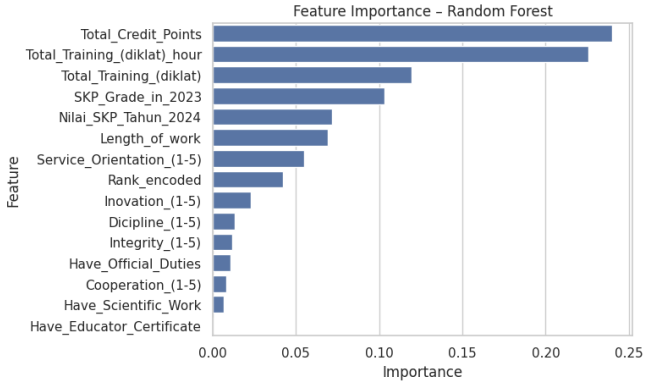

Figure 3.8 Feature Importance Visualization

The visualization further confirms the dominance of cumulative credit accumulation and professional development engagement within the predictive structure. The substantial gap between Total Credit Points and lower-ranked variables indicates that institutional eligibility decisions are primarily aligned with formal credit-based evaluation mechanisms. Notably, demographic and secondary behavioral indicators exhibit comparatively limited contribution. This pattern suggests that the predictive model reflects institutional performance criteria rather than demographic differentiation. However, as previously emphasized, importance values represent correlational

contribution within historical records and should not be interpreted as causal determinants of promotion outcomes.

J. Probability-Based Prediction Results

Prediction Output Analysis:
Predicted probabilities were calculated for each test sample using Random Forest, XGBoost, and the ensemble model. The ensemble probabilities provide a clearer representation of promotion likelihood by combining the strengths of individual models.

TABLE IX
PREDICTED PROMOTION PROBABILITIES

Teacher Code	RF Prob. (%)	XGB Prob. (%)	Ensemble Prob. (%)	Actual Status
GURU-A	71.2	78.5	74.9	Eligible
GURU-B	83.4	85.1	84.3	Eligible
GURU-C	42.7	38.9	40.8	Not Eligible
GURU-D	69.5	72.4	71	Eligible

K. Exploratory Fairness Analysis

To assess potential hidden bias within the predictive framework, an exploratory fairness analysis was conducted by examining predicted promotion probabilities across gender categories. The ensemble model was used as the reference predictor due to its probabilistic stability. Table X presents the average predicted promotion probabilities by gender.

TABLE X
PREDICTED PROMOTION PROBABILITIES

Gender	Mean Ensemble Probability
Female	0.8663
Male	0.8234

The results indicate that the average predicted probability for female teachers (0.8663) is slightly higher than that of male teachers (0.8234). However, the difference remains modest and does not indicate systematic exclusion of either group within the observed dataset.
To further visualize distributional differences, a probability distribution comparison is presented in Figure 3.9

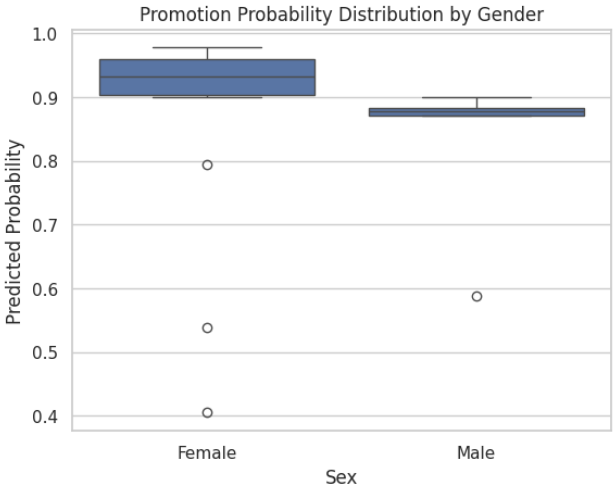

Figure 3.9 Promotion Probability Distribution by Gender

The boxplot illustrates the distributional range of predicted probabilities across gender categories. Minor variation is observed, with no consistent pattern of structural disadvantage detected. While female teachers exhibit a slightly higher mean predicted probability, this variation may reflect underlying differences in accumulated credit points and professional development records rather than gender-based discrimination. Feature importance analysis shows that demographic variables contribute minimally relative to institutional performance indicators. It is important to acknowledge that fairness evaluation remains statistically constrained due to the limited minority class representation (n = 3 non-eligible cases). Therefore, this exploratory assessment does not constitute a definitive fairness guarantee but rather an initial bias screening step. Future research involving larger and multi-institutional datasets is necessary to conduct formal fairness auditing using established metrics such as demographic parity or equalized odds.

L. Threshold Sensitivity Analysis

In probabilistic classification systems, the decision threshold determines how predicted probabilities are converted into binary classification outcomes. While the model produces continuous probability estimates, institutional decisions require categorical determination. Therefore, threshold sensitivity analysis was conducted to evaluate how varying probability cutoffs influence promotion classification. The baseline classification applied a standard probability threshold of 0.50. Additional policy scenarios were examined at 0.60, 0.70, and 0.90 to simulate moderate and highly conservative institutional decision settings. The impact of threshold variation is presented in Table XI.

TABLE XI
IMPACT OF PROBABILITY THRESHOLD VARIATION ON PROMOTION CLASSIFICATION

Threshold	Eligible Count	Non-Eligible Count
0.50	4	0
0.60	4	0
0.70	4	0
0.90	2	2

As shown in Table XI, moderate threshold adjustments between 0.50 and 0.70 do not alter classification outcomes within the test subset. All cases remain categorized as eligible under these conditions, indicating consistently high predicted probabilities across evaluated instances. However, under a highly conservative threshold scenario (0.90), classification outcomes shift substantially, with only two cases remaining eligible. This change demonstrates how stricter institutional policies may affect borderline candidates whose predicted probabilities fall below elevated cutoff levels. To visually illustrate this relationship, the number of eligible classifications across varying thresholds is presented in Figure 3.10

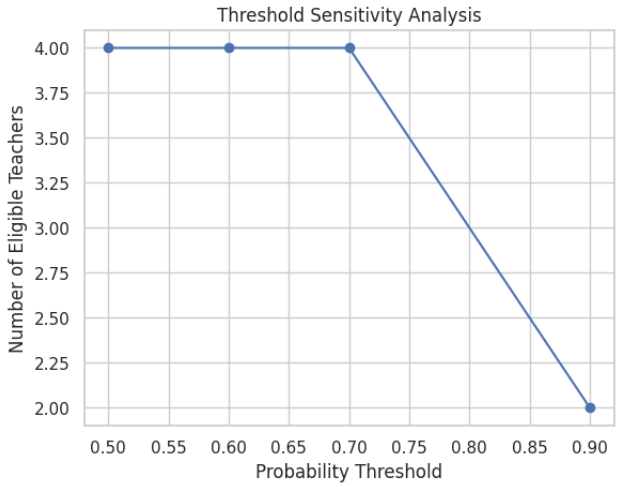

Figure 3.10 Sensitivity of Promotion Classification to Probability Threshold Variation

The figure shows that classification results remain unchanged when the threshold increases from 0.50 to 0.70, indicating consistently high predicted probabilities for all evaluated cases. A noticeable change occurs only at the 0.90 threshold, where two cases shift to non-eligible status. This demonstrates that moderate threshold adjustments do not substantially affect outcomes, while extreme cutoffs produce stricter eligibility decisions.

Threshold variation does not modify the model itself but only adjusts the decision boundary applied to predicted probabilities. Given the small test subset (n = 4), the results should be interpreted as an illustration of policy flexibility rather than a generalized performance claim. Overall, the analysis confirms that the framework supports adjustable decision calibration consistent with its role as a decision-support system.

M. Policy-Oriented Threshold Analysis

1. Threshold-Based Classification:
A probability threshold of 70% was applied to the ensemble predictions. Teachers with probabilities equal to or above this threshold were categorized as Highly Eligible, while others were categorized as Borderline.

TABLE XII
POLICY-BASED PROMOTION CLASSIFICATION

Ensemble Probability (%)	Policy Category
≥ 70%	Highly Eligible
< 70%	Borderline

N. Discussion

The findings of this study provide structured insight into the application of probabilistic ensemble modeling for institutional promotion assessment. The feature importance structure indicates that cumulative credit points and professional development indicators dominate predictive contribution. This pattern reflects the operational logic embedded within the Credit Point Assessment framework. Rather than redefining promotion standards, the model formalizes existing administrative evaluation patterns into a probabilistic analytical structure. In this sense, the predictive system functions as a structured reflection of institutional evaluation practices.

The calibration analysis demonstrates that XGBoost achieves superior probability reliability under limited and imbalanced data conditions, as reflected by the lowest Brier Score. Although the ensemble approach improves prediction stability by reducing variance between individual model outputs, calibration performance remains dependent on the characteristics of the underlying base learner. These findings reinforce the methodological emphasis on probability calibration rather than classification accuracy, particularly in institutional contexts where reliable likelihood estimation is more informative than deterministic decisions.

In addition to calibration evaluation, classification behavior was examined using confusion matrices to observe prediction tendencies across promotion categories. In addition to classification behavior analysis using confusion matrices, the inclusion of ROC-AUC further enriches model discrimination evaluation by summarizing the model's ability to distinguish between eligible and non-eligible cases across probability thresholds.

However, the limited minority class representation constrains the interpretability of discrimination-based

metrics. Under extreme class imbalance conditions, measures such as ROC-AUC may fluctuate considerably and provide limited inferential robustness. Consequently, calibration-oriented evaluation becomes particularly valuable for interpreting prediction reliability within small institutional datasets.

From a governance perspective, the probabilistic framework introduces adjustable decision thresholds that allow administrators to align promotion categorization with contextual policy considerations and acceptable risk levels. By modifying probability cutoffs, institutions can calibrate eligibility decisions according to operational priorities while maintaining transparency in evaluation procedures. This flexibility strengthens the positioning of the system as a decision-support mechanism rather than an automated decision authority.

It is important to emphasize that feature importance values represent correlational associations derived from historical administrative data and do not establish causal relationships. The predictive model captures statistical regularities embedded in prior institutional decisions and may therefore reflect structural characteristics of those processes. Because the model is trained on historical institutional records, it may implicitly reflect institutional evaluation patterns or policy-driven biases embedded in past promotion decisions.

Although exploratory fairness analysis did not identify substantial gender-based disparities, the limited dataset restricts definitive fairness assessment. In addition, the small and institution-specific dataset limits the extent to which the trained models can be generalized to broader educational institutions or different promotion systems.

Therefore, continuous monitoring, expanded datasets, and multi-institutional validation are necessary to ensure that predictive systems remain aligned with institutional policy objectives while minimizing unintended bias.

Overall, this study contributes to responsible machine learning deployment in policy-sensitive environments by demonstrating that calibrated probabilistic ensemble modeling can function as a transparent and interpretable analytical instrument within institutional promotion systems.

IV. CONCLUSION

This study proposed a probabilistic ensemble-based framework to support teacher promotion assessment within the institutional Credit Point system. By integrating Random Forest and XGBoost through a soft voting mechanism, the framework generates interpretable probability estimates that assist institutional decision-making rather than replacing it.

Empirical results indicate that cumulative credit accumulation and professional development indicators constitute the most influential predictors of promotion eligibility. XGBoost demonstrated superior probability calibration performance, while ensemble averaging improved predictive stability. The framework further enables adjustable threshold calibration, supporting policy-sensitive decision alignment.

However, several limitations must be explicitly acknowledged. First, the study is based on a small dataset (n = 20) with extreme class imbalance (3 non-eligible cases), which constrains statistical generalizability and limits the stability of discrimination-based evaluation metrics. Second, the dataset originates from a single institution and is restricted due to privacy considerations, limiting external replication and cross-institutional validation.

Accordingly, the findings should be interpreted as context-specific and exploratory rather than universally generalizable. Future research should incorporate larger, multi-institutional datasets, apply more robust validation strategies, and conduct formal fairness auditing to strengthen statistical reliability and ethical assurance.

Despite these constraints, the study demonstrates that calibrated probabilistic ensemble modeling can function as a structured and transparent decision-support instrument within institutional promotion systems.

REFERENCES

[1] A. Widayati, J. MacCallum, and A. Woods-McConney, "Teachers' perceptions of continuing professional development: a study of vocational high school teachers in Indonesia," *Teach. Dev.*, vol. 25, no. 5, pp. 604–621, 2021, doi: 10.1080/13664530.2021.1933159.

[2] I. Rahmi and S. Rassanjani, "Enhancing teacher quality in Indonesia: The impact of teacher professional development on achieving sustainable development goal 4.c," *Soc. Sci. Humanit. Open*, vol. 12, no. October, p. 102123, 2025, doi: 10.1016/j.ssaho.2025.102123.

[3] V. Pasupuleti, B. Thuraka, C. S. Kodete, and S. Malisetty, "Enhancing Supply Chain Agility and Sustainability through Machine Learning: Optimization Techniques for Logistics and Inventory Management," *Logistics*, vol. 8, no. 3, 2024, doi: 10.3390/logistics8030073.

[4] D. Chicco and G. Jurman, "The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation," *BMC Genomics*, vol. 21, no. 1, pp. 1–13, 2020, doi: 10.1186/s12864-019-6413-7.

[5] A. Whata, K. Dibeco, K. Madzima, and I. Obagbuwa, "Uncertainty quantification in multi-class image classification using chest X-ray images of COVID-19 and pneumonia," *Front. Artif. Intell.*, vol. 7, no. Ml, 2024, doi: 10.3389/frai.2024.1410841.

[6] A. J. Zeleke, P. Palumbo, P. Tubertini, R. Miglio, and L. Chiari, "Machine learning-based prediction of hospital prolonged length of stay admission at emergency department: a Gradient Boosting algorithm analysis," *Front. Artif. Intell.*, vol. 6, 2023, doi: 10.3389/frai.2023.1179226.

[7] C. C. Bey Lirna, T. Trimono, and A. T. Damaliana, "Employee Voluntary Attrition Prediction At Pt.Xyz: Ensemble Machine Learning Approach With Soft Voting Classifier," *J. Tek. Inform.*, vol. 5, no. 5, pp. 1231–1239, 2024, doi: 10.52436/1.jutif.2024.5.5.2007.

[8] A. T. Wibowo, M. Y. Teguh Sulistyono, and M. Hariadi, "Cryptospatial Coordinate Using The Rpca Based On A Point In Polygon Test For Cultural Heritage Tourism," *Commun. - Sci. Lett. Univ. Žilina*, vol. 22, no. 4, pp. 211–217, 2020, doi: 10.26552/com.C.2020.4.211-217.

[9] T. Kavzoglu and A. Teke, "Predictive Performances of Ensemble Machine Learning Algorithms in Landslide Susceptibility Mapping Using Random Forest, Extreme Gradient Boosting (XGBoost) and Natural Gradient Boosting (NGBoost)," *Arab. J. Sci. Eng.*, vol. 47, pp. 7367–7385, Jun. 2022, doi: 10.1007/s13369-022-06560-8.

[10] A. D. Kayit and M. T. Ismail, "Advancing stock price prediction through the development of hybrid ensembles: a comprehensive comparative analysis of machine learning approaches," *J. Big Data*, vol. 12, Dec. 2025, doi: 10.1186/s40537-025-01185-8.

[11] T. K. Robby and Noviyanti, "Analisis Proses Kenaikan Jenjang pada Jabatan Fungsional Pranata Humas di DPRD Kabupaten Sidoarjo," https://journal.unesa.ac.id/index.php/innovant/article/view/27339.

[12] M. N. Ambarita, M. Nasution, and R. Mutiah, "Analisis Prediksi Prestasi Siswa UPTD SD Negeri 30 Aek Batu dalam Machine Learning dengan Metode Naive Bayes," *J. Syntax Admiration*, vol. 5, no. 8, pp. 3167–3177, 2024, doi: 10.46799/jsa.v5i8.1493.

[13] A. Thakur *et al.*, "Product Length Predictions with Machine Learning: An Integrated Approach Using Extreme Gradient Boosting," *SN Comput. Sci.*, vol. 5, Aug. 2024, doi: 10.1007/s42979-024-02999-8.

[14] J. H. Hasugian and J. E. Situmorang, "Sosialisasi Perhitungan Dan Penilaian Angka Kredit Berdasarkan Permenpanrb Nomor 1 Tahun 2023bagi Guru - Guru TK, SD. SMP Se Kota Pematang SIANTAR," *J. Pengabdi. Masy. Sapangambei Manoktok Hitei*, vol. 3, pp. 166–174, Oct. 2023, doi: 10.36985/jy890y05.

[15] P. Koukaras and C. Tjortjis, "Data Preprocessing and Feature Engineering for Data Mining: Techniques, Tools, and Best Practices," Oct. 2025, *Multidisciplinary Digital Publishing Institute (MDPI)*. doi: 10.3390/ai6100257.

[16] O. Alshboul, A. Shehadeh, G. Almasabha, and A. S. Almuflih, "Extreme Gradient Boosting-Based Machine Learning Approach for Green Building Cost Prediction," *Sustain.*, vol. 14, Jun. 2022, doi: 10.3390/su14116651.

[17] Z. Zhou, C. Qiu, and Y. Zhang, "A comparative analysis of linear regression, neural networks and random forest regression for predicting air ozone employing soft sensor models," *Sci. Rep.*, vol. 13, Dec. 2023, doi: 10.1038/s41598-023-49899-0.

[18] J. Yang, A. A. S. Soltan, and D. A. Clifton, "Machine learning generalizability across healthcare settings: insights from multi-site COVID-19 screening," *npj Digit. Med.*, vol. 5, Dec. 2022, doi: 10.1038/s41746-022-00614-9.

[19] I. D. Mienye and Y. Sun, "A Survey of Ensemble Learning: Concepts, Algorithms, Applications, and Prospects," 2022, *Institute of Electrical and Electronics Engineers Inc.* doi: 10.1109/ACCESS.2022.3207287.

[20] Gullam Almuzadid and Egia Rosi Subhiyakto, "Stroke Risk Classification Using the Ensemble Learning Method of XGBoost and Random Forest," *J. Appl. Informatics Comput.*, vol. 9, pp. 828–837, Jun. 2025, doi: 10.30871/jaic.v9i3.9528.

[21] S. Demir and E. K. Sahin, "An investigation of feature selection methods for soil liquefaction prediction based on tree-based ensemble algorithms using AdaBoost, gradient boosting, and XGBoost," *Neural Comput. Appl.*, vol. 35, pp. 3173–3190, Feb. 2023, doi: 10.1007/s00521-022-07856-4.

[22] A. S. Hairani, R. Cahyono, and A. A. Balta, "Sistem Pendidikan Kinerja Siswa Berbasis Web Menggunakan Algoritma Decision Tree dan XGBost," vol. 5, no. 2, pp. 323–330, 2025.

[23] R. Ahmadian, M. Ghatee, and J. Wahlström, "Superior Scoring Rules for Probabilistic Evaluation of Single-Label Multi-Class Classification Tasks," Jul. 2024, doi: 10.1016/j.ijar.2025.109421.

[24] I. N. Bhakti, A. Z. Sholikhin, M. Abi, E. Daniati, and A. Ristyawan, "inotek,+1155-1164+S24-0030+Klasifikasi+Kategori+Berita+Menggunakan+Naive+Bayes, " pp. 1155–1164, 2024.

[25] A. S. Alfath, A. K. Wardhana, and R. Rumini, "Hypertension Risk Prediction Using Stacking Ensemble of CatBoost, XGBoost, and LightGBM: A Machine Learning Approach," *J. Appl. Informatics Comput.*, vol. 9, pp. 3146–3156, Dec. 2025, doi: 10.30871/jaic.v9i6.10370.

[26] D. B. M. Siahaan, E. C. Bagre, J. I. Wanda, G. Silahooy, and H. Sutejo, "Implementation Of Naïve Bayes Algorithm On The Eligibility Of Kartu Indonesia Pintar Scholarship (Case Study: University Of Sepuluh Nopember Papua)," *J. Ilm. Sist. Inf.*, vol. 4, pp. 191–204, Dec. 2025, doi: 10.51903/rdzdm469.

[27] Z. Z. Hulaifah Al Abrori and E. R. Subhiyakto, "Analisis Komparatif Akurasi Prediksi Kanker Payudara Menggunakan Algoritma Random Forest dan Logistic Regression," *J. Algoritm.*, vol. 22, pp. 300–311, May 2025, doi: 10.33364/algoritma/v.22-1.2164.

[28] M. Salsabila, "Pendekatan visual analytics dalam pemodelan prediksi cacat perangkat lunak menggunakan kombinasi pca dan smote," https://repository.uinjkt.ac.id/dspace/handle/123456789/65279.

[29] R. D. Yuniarsyih R.A, R. A. Muhadi, A. Fitrianto, and P. Silvianti, "Analisis Regresi Logistik Biner dan Random Forest untuk Prediksi Faktor-Faktor Stunting di Pulau Jawa," *Euler J. Ilm. Mat. Sains dan Teknol.*, vol. 13, no. 2, pp. 147–156, 2025, doi: 10.37905/euler.v13i2.31680.