

Article

Not peer-reviewed version

---

# Machine Learning Models for Predicting Student Enrollment Decisions in Higher Education

---

[Lazar Krstić](#)\*, [Dragan Soleša](#), [Marija Krstić](#)

Posted Date: 22 May 2026

doi: 10.20944/preprints202605.1507.v1

Keywords: machine learning; student enrollment prediction; educational analytics; stacking ensemble; classification models; imbalanced data; Matthews correlation coefficient (MCC)



Preprints.org is a free multidisciplinary platform providing preprint service that is dedicated to making early versions of research outputs permanently available and citable. Preprints posted at Preprints.org appear in Web of Science, Crossref, Google Scholar, Scilit, Europe PMC, OpenAlex.

Copyright: This open access article is published under a [Creative Commons CC BY 4.0 license](#), which permit the free download, distribution, and reuse, provided that the author and preprint are cited in any reuse.

Disclaimer/Publisher's Note: The statements, opinions, and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions, or products referred to in the content.

Article

# Machine Learning Models for Predicting Student Enrollment Decisions in Higher Education

Lazar Krstić<sup>1,\*</sup>, Dragan Soleša<sup>2</sup> and Marija Krstić<sup>1</sup>

<sup>1</sup> Department of Higher Business School Leskovac, Academy of Applied Studies Southern Serbia, Partizanska 7, 16000 Leskovac, Serbia

<sup>2</sup> Faculty of Economics and Engineering Management, University Business Academy in Novi Sad, Cvečarska 2, 21000 Novi Sad, Serbia

\* Correspondence: krstic.lazar@vpsle.edu.rs

## Abstract

An increasing number of higher education institutions in the Republic of Serbia are experiencing a decline in first-year enrollment, posing a significant challenge to their sustainability and effective resource planning. Timely identification of factors influencing candidates' enrollment decisions, as well as those at risk of not enrolling, is crucial for implementing appropriate institutional measures. This study aims to build and evaluate a machine learning model to predict candidates' decisions to enroll in a higher education institution based on relevant educational, administrative, geographic, and other characteristics. Various classification models were applied and compared in this study, including ensemble approaches, with a special focus on the Stacking Ensemble model. Experimental results show that the Stacking Ensemble model achieves slightly better performance than the other models, with an Area Under the ROC Curve (AUC) of 0.759 and a Matthews Correlation Coefficient (MCC) of 0.382, indicating its ability to provide reliable, balanced predictions under imbalanced data. However, the statistical analysis did not indicate statistically significant differences in performance among the models. The results suggest potential advantages of ensemble methods over individual models, particularly for complex classification problems. The application of the proposed model may contribute to improving the decision-making process at higher education institutions, enabling more efficient enrollment policy planning and more optimal resource management.

**Keywords:** machine learning; student enrollment prediction; educational analytics; stacking ensemble; classification models; imbalanced data; Matthews correlation coefficient (MCC)

## 1. Introduction

In an era of increasing competition in higher education, as reflected in the number of higher education institutions and study programs, predicting candidates' final enrollment decisions for a given institution is a significant challenge. This challenge is particularly pronounced in smaller countries, where candidates often apply to multiple higher education institutions. This is also the case in the Republic of Serbia, which is why higher education institutions need to estimate as accurately as possible the number of candidates who will actually enroll.

Higher enrollment levels are often associated with greater student satisfaction and a stronger institutional reputation. In contrast, low enrollment rates can pose serious financial and reputational challenges for higher education institutions, potentially leading to reduced funding and diminished public trust [1].

Predicting candidates' final decisions is complex because of the influence of numerous variables, including demographic changes, shifts in government policies, economic conditions, and institutional enrollment strategies. In this context, higher education institutions have increasingly relied on data analytics in recent years [2]. Data analytics are used to identify patterns in historical applicant data, which inform future predictions [3].

Machine learning algorithms are well-suited to analyzing such problems because they can model complex, nonlinear relationships among predictor variables that conventional statistical methods often cannot capture adequately [1]. Therefore, this study focuses on comparing the effectiveness of machine learning algorithms for predicting candidates' decisions to enroll in a higher education institution and on building a stable, accurate model to make those predictions.

A key contribution of this study is the application of a robust evaluation framework based on the Matthews Correlation Coefficient (MCC). Additionally, the MCC values of the two most successful models were compared using the Wilcoxon signed-rank test to provide a more reliable assessment of their relative performance.

## 2. Literature Review

To systematize relevant research on similar issues, Table 1 presents an overview of selected papers in educational analytics and the applications of machine learning methods. To identify these papers, open-access electronic databases were used as an efficient source for searching relevant literature. The analysis focuses on freely available papers published between 2021 and 2025 that apply classification models and ensemble methods to predict enrollment decisions at higher education institutions. Although some studies analyze the prediction of candidate admission rather than direct enrollment, they are considered relevant because of the similarity of the problem and the methodological approaches applied.

**Table 1.** Review of selected papers in educational analytics and the application of machine learning methods.

Author, Year	Problem	Aim of the Study	Data (Dataset)	Algorithms/ Model	Evaluation Metrics	Key Results
Soltys et al., 2021 [4]	Student enrollment prediction	Predict student enrollment under competitive and budget constraints	CSUCI admissions data (2018–2020)	XGBoost	Accuracy, Precision, Recall, Specificity	Optimal threshold = 0.09; FN = 25%, FP = 39%
Yang et al., 2021 [5]	Freshman enrollment prediction	Predict freshman enrollment using Machine Learning methods	Guangzhou University data (2009–2012), 10,382 records	Decision Tree, Random Forest, BP Neural Network	Accuracy, Precision, Recall, F1-score	Decision Tree: ACC 62.94% (F1 0.6963); BP Neural Network: ACC 62.07% (F1 0.6967); Random Forest: ACC 60.60% (F1 0.6660)
Shao et al., 2022 [6]	Course enrollment prediction	Improve course enrollment prediction using CART and Random Forest	SDSU student data (2010–2019), ~83,000 records	CART, Random Forest, Conditional Probability Analysis	Error rate	Random Forest: 0.8% (best); CART: 15.5%; Flowchart: 21.2%
Ayasi et al., 2023 [7]	Course enrollment prediction	Evaluate Machine Learning and Neural	AAUP data (2018–2021), ~9,000 students,	Logistic Regression, Stochastic Gradient	Accuracy, Precision, Recall, F1-score	Random Forest: ACC 94%, F1 86% (best); MLP:

		Networks for course enrollment prediction	137,000 records, imbalanced (10:1)	Descent, k-Nearest Neighbors, CART, Gradient Boosting, Bagging, Support Vector Machine, Random Forest, MLP		ACC 91%, F1 79%
Brianorman & Sucipto, 2024 [8]	Admission prediction	Predict admission likelihood using classification models	Pontianak University data (2020), 1,892 records	Decision Tree, Random Forest, Naive Bayes	Accuracy, Precision, Recall, F1-score	Random Forest: ACC 59.2%, F1 0.574 (best); Decision Tree: ACC 59.1%; Naive Bayes: ACC 58.1%
Raftopoulos, Davrazos & Kotsiantis, 2024 [9]	Admission prediction	Predict admission outcomes with fair and interpretable Machine Learning models	MBA (synthetic, 6,194 cases), LSAC Law School (1,991 cases, 12 features), College Admission (7 features)	Logistic Regression, Decision Tree, Naive Bayes, Ensemble methods	Accuracy, AUC, Precision, Recall, F1-score, Kappa, MCC; Selection Rate, Disparate Impact, TPR, FPR	Logistic Regression: ACC 0.9097, F1 0.9516 (Law, best); Recall 0.9722 (MBA); Naive Bayes: high precision (consistent)
Ramos, 2024 [10]	Enrollment prediction	Develop a machine learning model for enrollment prediction and key factor analysis Optimize freshman enrollment prediction using a voting ensemble	BSIT data (2019–2023), 76 students survey	Linear Regression	Accuracy	Predicted: increase in 1st-year (up to 74) and 3rd/4th-year; decrease in 2nd-year (24 to 8)
Yang et al., 2024 [11]	Freshman enrollment prediction	Identify key factors influencing enrollment (low-SES students)	Chinese university data (6 years), 17,652 records (9,839 / 7,813), 18 features	Decision Tree, Random Forest, BP Neural Network (ensemble)	Accuracy, Precision, Recall, F1-score	Soft Voting: ACC 64.67%, F1 0.69 (best; outperforms single models)
He et al., 2025 [12]	Enrollment factor analysis		HSLs:09 (5,223 low-SES 9th-grade students), 28 features	Logistic Regression, k-Nearest Neighbors, Support Vector	Accuracy, Macro Precision, Macro Recall, Macro F1-	Random Forest: ACC 67.73%, F1 0.6999 (best); no overfitting

---

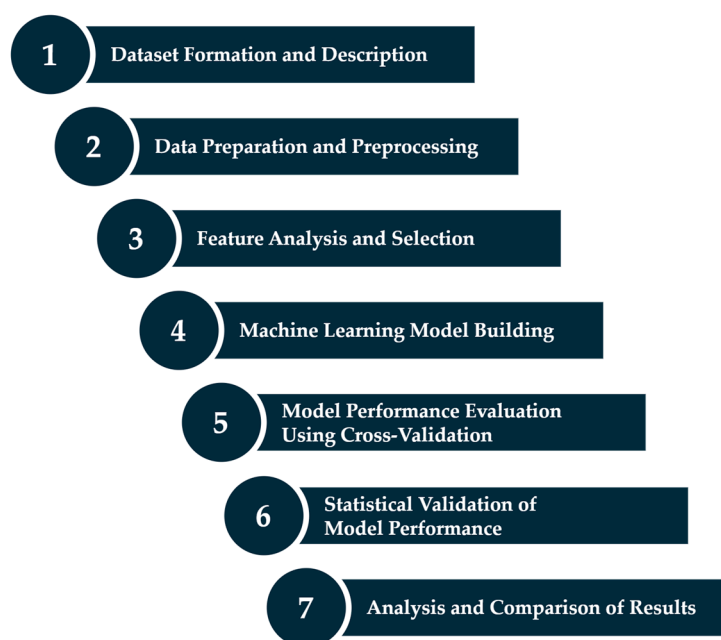
Machine, score, ROC-  
Decision Tree, AUC  
Random  
Forest

---

The studies presented in Table 1 show that machine learning methods are widely applied in educational analytics. In addition, there is notable diversity in the selection of evaluation metrics, which indicates the complexity of the problem and the need for robust performance measures, that is, metrics that provide a reliable and balanced assessment of model quality, especially when dealing with imbalanced datasets. Unlike most studies analyzed, which predominantly use standard metrics for model performance evaluation, this study places particular emphasis on MCC as a more robust metric. This approach enables a more comprehensive and balanced evaluation of performance, particularly in the context of imbalanced data. In this regard, this study stands out for employing a more methodologically grounded evaluation approach, thereby ensuring a more reliable assessment of model performance.

### 3. Materials and Methods

This section presents the research's methodological framework, covering all stages of building the prediction model, from describing and preparing the dataset to selecting and applying machine learning algorithms and evaluating and statistically analyzing their performance. The aim is to provide a clear and systematic overview of the steps involved in building and evaluating classification models to predict candidates' decisions to enroll in higher education institutions. Figure 1 shows the research's methodological framework, with clearly defined stages of machine learning model building.



**Figure 1.** Methodological framework of the research with clearly defined phases for building a machine learning model.

#### 3.1. Dataset Description

The dataset used in this study was compiled from historical data obtained from the information system of a public higher education institution in the Republic of Serbia and included 1,597 records with 15 attributes. Of the total number of attributes, 14 are input variables (features) describing the candidates' characteristics, while 1 attribute is the target variable indicating the outcome—the

candidates' enrollment decision. The target variable had two values, "enrolled" and "not enrolled," making the classification problem binary.

### 3.2. Attribute Description

The attributes used as input variables for the machine learning model capture characteristics of candidates that may influence their enrollment decisions. The selection of attributes is based on an analysis of data available from the higher education institution's information system and on their relevance for predicting candidates' enrollment decisions. As previously emphasized, this study used 14 input attributes, of which 11 were categorical and 3 were numerical. These attributes include demographic, geographic, educational, administrative, and social information about the candidates and serve as the basis for building a prediction model. Table 2 presents the dataset attributes, including their type, category, role in the model, and a brief description. In addition to the input attributes, the table displays a specially designated target variable indicating whether the candidates have completed enrollment.

**Table 2.** Attributes of the dataset used in the study.

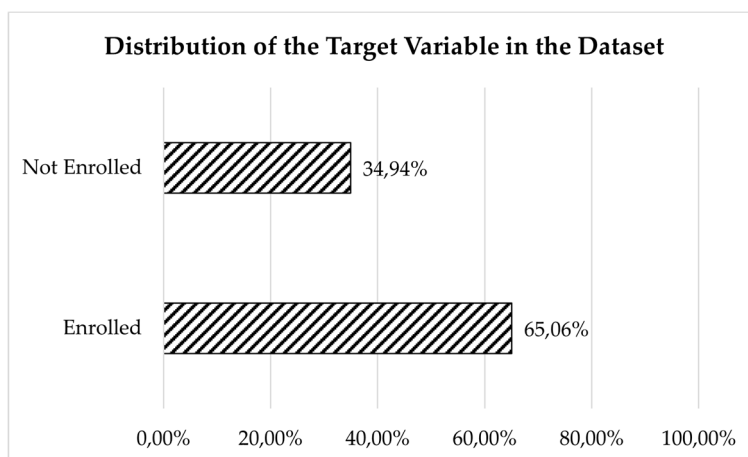
Name	Type	Category	Role in the Model	Description
Gender	categorical	demographic	input	Gender of the candidate applying for enrollment
Place of Residence	categorical	geographic	input	District of the candidate's permanent residence
Residence in the Institution's City	categorical	geographic	input	Information indicating whether the candidate resides in the same city as the higher education institution
Distance from Residence to Higher Education Institution	categorical	geographic	input	Discretized distance between the candidate's place of residence and the higher education institution, divided into categories
Secondary School Type	categorical	educational	input	Type of secondary school completed by the candidate
Average Secondary School Grade	numeric	educational	input	Average grade achieved by the candidate during secondary education
Attendance of Preparatory Classes	categorical	educational	input	Information indicating whether the candidate attended preparatory classes for the entrance examination
Entrance Exam Score	numeric	educational	input	Number of points achieved by the candidate on the entrance examination
Total Enrollment Score	numeric	educational	input	Total number of points used for candidate ranking
Enrollment Period	categorical	administrative	input	The enrollment period in which the candidate applies

Application to Multiple Higher Education Institutions	categorical	administrative	input	Information indicating whether the candidate applied to multiple higher education institutions
Study Program	categorical	administrative	input	Study program for which the candidate applies
Tuition Funding Status	categorical	social	input	Information indicating whether the candidate is funded through the state budget or self-financed
Eligibility for Student Housing	categorical	social	input	Information indicating whether the candidate is eligible for student housing
<b>Enrollment status</b>	<b>categorical</b>	<b>Candidate enrollment decision</b>	<b>target</b>	<b>Indicates whether the candidate completed the enrollment process or not</b>

Some attributes in Table 2 require further clarification. The attribute “Residence in the Institution’s City” is a binary indicator of whether the candidate resides in the city where the higher education institution is located. In contrast, the attribute “Distance from Residence to Higher Education Institution” provides broader context on the candidate’s geographic distance from the institution. “Total Enrollment Score” is a key criterion for ranking candidates in the enrollment process; therefore, it is included as a significant input attribute in machine learning models.

### 3.3. Data Preparation

The dataset was constructed by aggregating student records across multiple academic years. The core part of the dataset includes information on applicants for admission spanning five consecutive academic years (from 2020/2021 to 2024/2025). To improve the representativeness of the minority class and statistical reliability, rather than applying synthetic balancing techniques, additional records from previous academic cohorts about the “not enrolled” class (from 2015/2016 to 2019/2020) were included. Records from previous academic cohorts were collected from the same information system and structured according to the same administrative and record-keeping rules as the data in the core part of the dataset, ensuring complete data comparability. This approach reduced the initial class imbalance and created a more suitable dataset for training and evaluating machine learning models. After expanding the dataset, the final class distribution was 65.06% instances of the “enrolled” class and 34.94% instances of the “not enrolled” class, indicating a moderately imbalanced dataset suitable for standard classification methods. The distribution of the target variable in the dataset is shown in Figure 2.



**Figure 2.** Distribution of the target variable in the dataset.

The dataset contains no missing values because candidate records are maintained through standardized administrative procedures. This data collection and validation approach enabled the direct application of machine learning algorithms without additional imputation or data cleaning.

Categorical attributes were used in their original form, and their transformation into a numerical format—required by certain algorithms—was performed automatically within the tool. The attribute “Distance from Residence to Higher Education Institution” was originally numerical. During the data preparation phase, it was discretized into predefined categories to enable its use in models that require categorical input variables. In addition, numerical attributes were not scaled, as most of the applied algorithms are insensitive to the range of input data values.

Not all attributes were used to build the prediction models. Attribute selection was based on importance, and only those with the greatest predictive impact were included. This approach reduced model complexity while maintaining performance, improving interpretability, and reducing the risk of overfitting.

### 3.4. Algorithm Selection

In this study, various machine learning algorithms were applied to build models to predict candidates’ enrollment decisions at a higher education institution, including Logistic Regression, Random Forest, Gradient Boosting, Neural Network, and an Ensemble Model based on the Stacking Method. The selected algorithms encompassed different machine learning approaches, including linear models, ensemble methods, and neural networks. These algorithms are frequently used in classification problems, as previous research has demonstrated their strong performance on predictive analytics tasks in educational analytics. Table 3 describes the machine learning algorithms.

**Table 3.** Description of the algorithms used in the study.

Name	Description
Logistic Regression (LR)	LR is a widely used statistical method for analyzing and predicting binary outcomes. In LR, the dependent variable takes two values, such as “yes” or “no.” The model estimates the probability of an event using a logistic function, which maps a linear combination of predictors to the interval [0, 1]. Although the relationship between the predictors and the outcome probability is nonlinear, LR assumes a linear relationship in the log-odds space. The model coefficients are interpreted as changes in log-odds and are often transformed into odds ratios for easier interpretation [13,14].
Random Forest (RF)	RF is a popular machine learning algorithm that uses an ensemble of decision trees for classification and regression tasks. This ensemble

---

Gradient Boosting (GB)	<p>method combines predictions from multiple trees trained on different data subsets and averages them for regression, thereby achieving significantly higher accuracy than a single tree. The model is recognized for its ability to handle complex, nonlinear relationships while minimizing the risk of overfitting. In addition to prediction, RFs are extremely useful for identifying the most important factors in the data through variable importance measures [15,16].</p> <p>GB is an efficient ensemble machine learning method that sequentially builds a set of weak models, most commonly decision trees, to form a single strong predictive model. The key mechanism of this technique is an iterative process in which each new model attempts to correct the errors (residuals) of previous models by using information about the gradient, i.e., the slope of the loss function. In this manner, the algorithm gradually minimizes the overall loss and improves prediction accuracy through an additive process, making it well-suited for solving complex classification and regression problems [17,18].</p>
Neural Network (NN)	<p>NN is a mathematical model that processes input data to generate accurate predictions. Its structure consists of layers of interconnected neurons—input, hidden, and output layers—that serve as universal approximators for solving complex problems. Each neuron computes a weighted sum of its inputs and applies an activation function, enabling the model to recognize and interpret complex relationships in the data. Learning occurs through iterative adjustment of connection weights (most commonly via the backpropagation algorithm), thereby gradually minimizing prediction error [19,20].</p>
Ensemble Model Based on the Stacking Method	<p>The Stacking method is an advanced ensemble learning technique that combines multiple base models to achieve better predictive performance than any individual model. The process occurs at two levels: first, a set of base models is trained, and their predictions or probabilities serve as input features for a new dataset. In this new data space, a meta-model is trained to learn how to optimally integrate the base models' outputs into the final result. The advantage of this method lies in its ability to synthesize complementary patterns and reduce variance and bias, making the final model more robust and accurate across a wide range of applications [21,22].</p>

---

### 3.5. Experimental Setup

The models were built and tested using Orange Data Mining version 3.40.0. Orange Data Mining is an open-source tool for data analysis and machine learning based on visual programming. Users can create workflows by connecting components (widgets) that represent the individual steps in the data analysis process. In addition, the tool includes a special “Python Script” component that enables the writing of Python scripts within the workflow [23]. This tool allows for the transparent construction and evaluation of machine learning models through clearly defined workflows, thereby facilitating the reproducibility and interpretation of research results.

As part of the research, an appropriate workflow was created, encompassing all phases of the modeling process—from data loading and preparation to applying various classification algorithms and generating and analyzing results. In addition to the individual models, a Stacking Ensemble was employed, combining the predictions of multiple base models to improve overall performance. The Stacking Ensemble used Logistic Regression as the meta-model, with the same parameters as those used in the individual Logistic Regression model. The models were trained and tested using the available tool's functionalities, employing standard evaluation components. For certain models, a

class-balancing option was applied to improve recognition of the minority class. The workflow used in this study is shown in Figure 3.

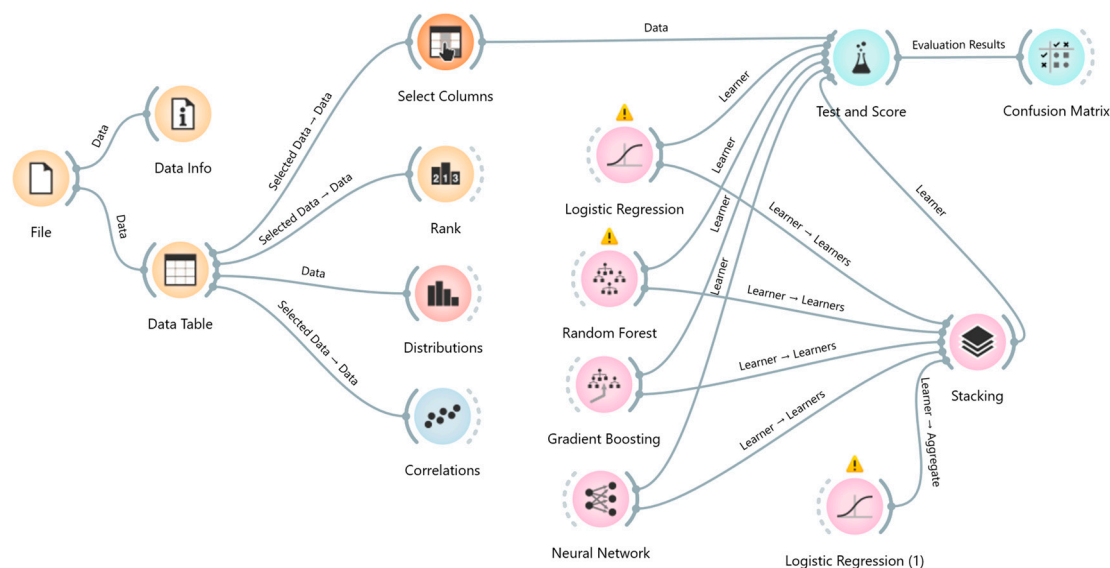


Figure 3. Workflow used in the study.

Fine-tuning of the model was performed to improve stability and prediction accuracy and to reduce the risk of overfitting. Table 4 presents the hyperparameter values used in the experiment, including all relevant model settings; parameters that were not explicitly modified retained their default values.

Table 4. Hyperparameter values used in the experiment.

Model	Hyperparameter	Value
Logistic Regression (LR)	Regularization strength (C)	0.900
	Class balancing	Enabled
Random Forest (RF)	Number of trees	300
	Class balancing	Enabled
	Maximum tree depth	3
	Minimum number of instances for splitting	5
Gradient Boosting (GB)	Number of trees	300
	Learning rate	0.050
	Reproducible training	Enabled
	Maximum tree depth	3
Neural Network (NN)	Number of neurons in hidden layers	100
	Maximum number of iterations	200
	Reproducible training	Enabled

### 3.6. Model Evaluation and Performance Metrics

Stratified k-fold cross-validation is one of the most commonly used approaches for model evaluation in machine learning research. It is widely regarded as an efficient and robust method for assessing models and is often considered the gold standard for working with imbalanced datasets. Stratified k-fold cross-validation divides the dataset into k folds while preserving the original class distribution, which is crucial for reliable evaluation under imbalanced conditions. The process unfolds over k iterations, with each iteration using one fold for testing and the remaining folds for

training the model. This approach ensures that each data point is used for both training and validation, yielding a more stable estimate of generalization performance and reducing the risk of biased outcomes. The final model performance is determined by aggregating the results across all iterations, most often by calculating the mean [24,25].

In the experimental section, the model was evaluated using stratified 10-fold cross-validation. The aim was to ensure a reliable, objective assessment of model performance, maximize the use of available data, and reduce the risk of overfitting. This approach provides a more accurate and objective evaluation of model performance than a simple split of the dataset into training and test sets. Additional experiments were conducted using 5-fold and 20-fold cross-validation to assess the model's stability further.

The confusion matrix underlies the calculation of various model performance metrics. In binary classification, it is a fundamental tool for evaluating models with two classes: positive and negative. Its structure is a 2x2 matrix, with rows representing the actual classes and columns representing the model's predicted classes. It consists of four key elements that describe the relationship between the predicted and actual values [26,27]:

- True Positive (TP) – the model correctly predicts the positive class.
- True Negative (TN) – the model correctly predicts the negative class.
- False Positive (FP) – the model predicts the positive class even though the actual value is negative (a Type I error).
- False Negative (FN) – the model predicts the negative class when the true value is positive (a Type II error).

The model's performance was evaluated using multiple metrics to provide a comprehensive assessment of prediction quality. These metrics include:

- Area Under the ROC Curve (AUC),
- Accuracy,
- F1 score,
- Precision,
- Recall and
- Matthews Correlation Coefficient (MCC).

### 3.6.1. Area Under the ROC Curve (AUC)

The Area Under the ROC Curve (AUC) quantifies a model's overall ability to distinguish between classes. It is computed by integrating the ROC curve, which plots recall against the false-positive rate (1 - specificity) across all thresholds. Unlike many other metrics, the AUC is independent of a specific classification threshold, making it robust for general assessment of model quality. The AUC ranges from 0 to 1, with 1 indicating perfect prediction and 0.5 corresponding to the performance of a random classifier [28,29].

### 3.6.2. Accuracy

Accuracy is the ratio of the total number of correctly classified instances (positive and negative) to the total number of observed cases. Although it is one of the most commonly used metrics, it is considered unreliable for imbalanced datasets because it can yield deceptively high values that favor the majority class. Accuracy is calculated using the following formula [28,29]:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

### 3.6.3. F1 Score

The F1 score is the harmonic mean of precision and recall, offering a balanced assessment of both metrics. This is particularly useful for imbalanced datasets, as it prevents over-optimizing one metric

at the expense of another. As a result, it is often used as a reliable indicator of overall model performance in real-world applications. The F1 score is calculated using the following formula [28,29]:

$$F1\ score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (2)$$

#### 3.6.4. Precision

Precision is the proportion of correct positive predictions. This metric is particularly important in applications where false positives are costly, such as fraud detection or marketing campaigns that aim to avoid unnecessary user targeting. Precision is calculated using the following formula [28,29]:

$$Precision = \frac{TP}{TP + FP} \quad (3)$$

#### 3.6.5. Recall

Recall is the proportion of actual positive cases that the model correctly identifies. This metric is especially important when missing a positive case (a false negative) is critical, such as in medical diagnoses or loan default prediction. Recall is calculated as follows [28,29]:

$$Recall = \frac{TP}{TP + FN} \quad (4)$$

#### 3.6.6. Matthews Correlation Coefficient (MCC)

The Matthews Correlation Coefficient (MCC) is a robust metric that accounts for all four elements of the confusion matrix (TP, TN, FP, and FN), providing a balanced assessment of model performance. It is considered superior to Accuracy and the F1 score on imbalanced datasets because it remains reliable even when class distributions are highly skewed. The MCC measures the correlation between the actual and predicted values, with values ranging from -1 to 1. It is calculated using the following formula [28,29]:

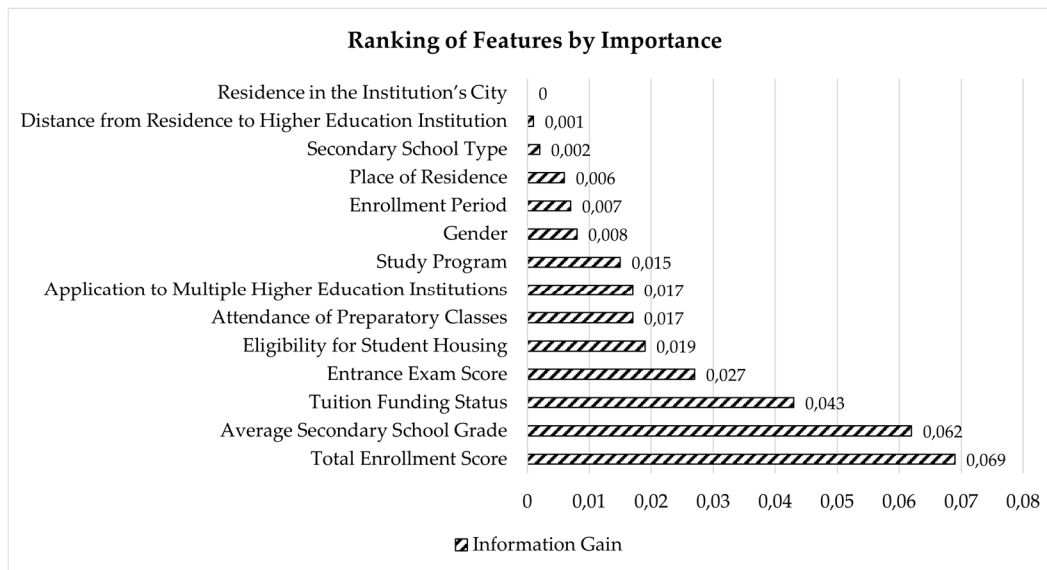
$$MCC = \frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \quad (5)$$

Identifying candidates who will not enroll is an important part of the analysis and can provide significant support for decision-making. To ensure consistent interpretation of the performance metrics, the “not enrolled” class was defined as the positive class.

## 4. Results

### 4.1. Feature Importance Analysis

To increase the transparency and interpretability of the machine learning models, a feature importance analysis was conducted, enabling the identification of the key factors that most strongly influenced candidates’ enrollment decisions. The Information Gain metric was used for feature ranking because it is among the most widely used methods for feature evaluation and ranking in machine learning. This metric quantifies the information an individual feature provides about the target variable, thereby enabling the identification of the most relevant features for prediction [30]. Although it may exhibit a bias towards features with more values, its application in this study enabled the identification of the features that contribute the most information. Figure 4 shows the ranking of features by importance.



**Figure 4.** Feature Importance Ranking.

Figure 4 shows that certain features achieved significantly higher Information Gain values than others, indicating their greater importance in the prediction process. The results of the analysis show that features essential to the enrollment process, such as the total enrollment score and average secondary school grade, have the greatest influence on model predictions, thereby confirming their alignment with real-world decision-making. These features represent important criteria for candidates' enrollment decisions at higher education institutions. This further confirms the model's interpretability and its ability to identify the key factors influencing candidates' enrollment decisions.

Based on the results, feature selection was performed, and the 8 features with the highest importance were used for further model development. This approach reduces model complexity while maintaining or even improving performance. In addition, identifying key features improves model interpretability and deepens understanding of the factors that influence candidates' decisions to enroll in higher education institutions.

#### 4.2. Results of the Models

In this section, the results of the machine learning models trained on the 8 most significant features are presented and analyzed. Table 5 summarizes the performance metrics for all models, computed via stratified 10-fold cross-validation.

**Table 5.** Overview of performance metrics for all models (10-fold cross-validation).

Model	AUC	Accuracy	F1	Precision	Recall	MCC
Logistic Regression (LR)	0.754	0.693	0.699	0.716	0.693	0.368
Random Forest (RF)	0.738	0.675	0.682	0.710	0.675	0.351
Gradient Boosting (GB)	0.739	0.719	0.707	0.708	0.719	0.350
Neural Network (NN)	0.751	0.719	0.708	0.709	0.719	0.352
Stacking Ensemble	0.759	0.700	0.706	0.722	0.700	0.382

Based on the results presented in Table 5, the Stacking Ensemble model achieved slightly better overall performance than the other models. It stands out particularly in terms of the AUC and MCC metrics, with MCC, a robust metric for imbalanced classes, further indicating the Stacking Ensemble model's somewhat better performance relative to the other models.

#### 4.3. Stability of the Models

To assess the model's stability, evaluation was conducted using 5-fold and 20-fold cross-validation. The results of this analysis are shown in Table 6.

**Table 6.** Overview of performance metrics for all models (varied cross-validation configurations).

Model	AUC	Accuracy	F1	Precision	Recall	MCC
5-fold cross-validation						
Logistic Regression (LR)	0.755	0.693	0.699	0.717	0.693	0.371
Random Forest (RF)	0.740	0.673	0.681	0.710	0.673	0.349
Gradient Boosting (GB)	0.736	0.716	0.704	0.705	0.716	0.343
Neural Network (NN)	0.752	0.721	0.712	0.712	0.721	0.361
Stacking Ensemble	0.760	0.699	0.705	0.722	0.699	0.382
10-fold cross-validation						
Logistic Regression (LR)	0.754	0.693	0.699	0.716	0.693	0.368
Random Forest (RF)	0.738	0.675	0.682	0.710	0.675	0.351
Gradient Boosting (GB)	0.739	0.719	0.707	0.708	0.719	0.350
Neural Network (NN)	0.751	0.719	0.708	0.709	0.719	0.352
Stacking Ensemble	0.759	0.700	0.706	0.722	0.700	0.382
20-fold cross-validation						
Logistic Regression (LR)	0.754	0.693	0.699	0.716	0.693	0.368
Random Forest (RF)	0.739	0.676	0.683	0.713	0.676	0.355
Gradient Boosting (GB)	0.737	0.710	0.697	0.698	0.710	0.328
Neural Network (NN)	0.751	0.718	0.708	0.708	0.718	0.351
Stacking Ensemble	0.759	0.700	0.706	0.722	0.700	0.382

The results in Table 6 indicate the model's stability across different cross-validation configurations. Varying the number of folds did not significantly affect model performance. The absence of significant variation in the metric values confirmed their consistency and reliability.

#### 4.4. Confusion Matrix

For a more detailed analysis of model performance, Table 7 presents the confusion matrix for the Stacking Ensemble model. Class "0" represents candidates who do not enroll and is treated as the positive class in this study, whereas class "1" represents those who do enroll.

**Table 7.** Confusion matrix for the Stacking Ensemble model.

		Predicted		
		0	1	$\Sigma$
Actual	0	387	171	558
	1	308	731	1039
	$\Sigma$	695	902	1597

Based on the results presented in Table 7, the model correctly classified 387 candidates who did not enroll (TP) and 731 candidates who did enroll (TN). Simultaneously, 171 candidates who did not enroll were incorrectly classified as enrolled (FN), whereas 308 candidates who did enroll were incorrectly classified as not enrolled (FP).

#### 4.5. Statistical Analysis of the Models

Given the similar performance of the Logistic Regression and Stacking Ensemble models, an additional statistical comparison was conducted to assess their performance. For both models, the evaluation metrics were computed using 10-fold cross-validation. The MCC values for the individual folds, which served as inputs for statistical testing, are presented in Table 8.

**Table 8.** MCC values across 10 cross-validation folds for the Logistic Regression and Stacking Ensemble models.

Fold	Logistic Regression (MCC)	Stacking Ensemble (MCC)
1	0.443	0.443
2	0.367	0.430
3	0.337	0.376
4	0.417	0.387
5	0.281	0.316
6	0.357	0.373
7	0.350	0.297
8	0.394	0.469
9	0.352	0.397
10	0.404	0.343

Using the values in Table 8, the means and standard deviations of performance were calculated and are presented in Table 9. Subsequently, the statistical significance of the differences between the models was evaluated using the Wilcoxon signed-rank test on the MCC values obtained for individual folds. The Wilcoxon signed-rank test is a non-parametric statistical test used to compare two dependent groups or two related samples. The test ranks the absolute differences between paired observations and is used when the assumption of normality is not met. It serves as an alternative to the paired t-test and allows testing the statistical significance of differences in model performance based on paired results [31].

**Table 9.** Mean and standard deviation of the MCC values (10-fold cross-validation).

Model	MCC Mean	MCC Standard Deviation
Logistic Regression (LR)	0.370	0.044
Stacking Ensemble	0.383	0.052

The results of the statistical analysis in Table 9 show that the Stacking Ensemble model achieves a higher mean MCC (0.383), indicating somewhat better classification performance than Logistic Regression (0.370). However, Logistic Regression exhibited a lower standard deviation (0.044 compared to 0.052), indicating greater stability across cross-validation folds.

The Wilcoxon signed-rank test results did not indicate a statistically significant difference in performance between the Logistic Regression and Stacking Ensemble models ( $p=0.426$ ). Although the Stacking Ensemble model achieved a slightly higher average MCC, the difference was not statistically significant.

Both models achieved similar performance. The Stacking Ensemble model showed a slight advantage in average MCC, whereas Logistic Regression provided greater stability in results.

## 5. Discussion

Based on the results, the built models show different performance, with the Stacking Ensemble model exhibiting better overall performance. Metrics such as AUC and MCC are key indicators of a model's ability to distinguish classes and provide reliable predictions in the context of imbalanced data. The achieved AUC (0.759) and MCC (0.382) indicate a slight advantage for this model over the others. The MCC metric is particularly important because it considers all elements of the confusion matrix and provides a balanced assessment of performance. The value indicates the model's ability to balance correct and incorrect classifications. The other models, despite having similar accuracy values (approximately 0.70), did not show significant differences in overall performance balance.

It is important to compare the performance of the Logistic Regression and Stacking Ensemble models, given their similar values across most evaluation metrics. Logistic Regression achieved an AUC of 0.754, an accuracy of 0.693, and an F1 score of 0.699, which were very close to those of the Stacking Ensemble model. However, the MCC (0.368 for Logistic Regression versus 0.382 for the

Stacking Ensemble model) indicates a slightly better overall performance balance for the ensemble model. This difference suggests that although individual models can achieve competitive results, ensemble approaches may yield more stable and balanced classifications under imbalanced data conditions.

The results also show that the individual models achieved competitive performance but with certain limitations. However, their MCC values remained lower than those of the Stacking Ensemble model, indicating less balanced classification performance.

The analysis of model stability across different cross-validation configurations showed no significant deviations in the performance metrics. These results confirm that the performance is consistent and reliable, regardless of the number of subsets (5, 10, or 20).

Based on the confusion matrix for the Stacking Ensemble model, the model achieved a balanced ratio of correct to incorrect classifications. Although there were 308 false positives and 171 false negatives, the model maintained stable performance across both classes. This is particularly significant for identifying candidates who will not enroll in a higher education institution.

Statistical analysis of model performance and stability indicates that the Stacking Ensemble model achieves a slightly higher average MCC, whereas Logistic Regression shows greater stability across cross-validation folds. However, the Wilcoxon signed-rank test did not detect a statistically significant difference in performance, suggesting that the models perform comparably on this problem.

The advantage of the Stacking Ensemble model is its ability to combine predictions from several models, thereby mitigating each model's individual weaknesses. Thus, ensemble approaches can lead to better generalization and more balanced performance than individual models, such as Logistic Regression, Random Forest, Gradient Boosting, and Neural Network.

## 6. Conclusions

Research in machine learning, especially in educational analytics, indicates that ensemble methods can outperform individual models, particularly for complex problems with imbalanced datasets. By combining multiple models, ensemble approaches such as stacking can achieve more balanced performance than single classifiers. This study analyzes the application of machine learning methods to predict candidates' decisions to enroll in higher education institutions, with a particular focus on identifying candidates at risk of not enrolling.

The experimental results suggest that different machine learning models achieve competitive performance, with some variation in classification performance. The Stacking Ensemble model was the best-performing, achieving the highest AUC and MCC values, indicating its potential to provide more balanced predictions in the presence of imbalanced data.

The MCC metric is particularly important in model evaluation because it is well-suited to assessing performance in this context by considering all elements of the confusion matrix and providing a balanced assessment of model quality. Additionally, analysis of model stability across different folds of cross-validation showed consistent results, further confirming the model's reliability.

The results indicate that the models perform comparably, with the Stacking Ensemble model showing a slight advantage, while Logistic Regression offers greater stability. Statistical analysis did not reveal any significant differences among the models.

The built model is practical and may serve as a basis for developing a prototype decision support system for enrollment planning. In this way, the research results go beyond a purely theoretical framework and suggest the possibility of integrating the model into the existing information environment of higher education institutions.

The research in this study demonstrates that ensemble methods, such as stacking, can improve the performance of classification models in educational analytics. However, they do not necessarily outperform individual models, as their effectiveness largely depends on the nature of the problem, data structure, and the selection of input features.

Despite the results achieved, this research has certain limitations, primarily related to the dataset's size and structure and the limited set of algorithms used. Future research should focus on expanding the dataset, improving the attribute selection process, and applying alternative ensemble approaches.

**Author Contributions:** All authors designed the study, developed the methodology, conducted the analyses, and wrote the manuscript. All authors have read and agreed to the published version of this manuscript.

**Funding:** This research received no external funding.

**Institutional Review Board Statement:** Ethical review and approval were waived because anonymized student data were used in accordance with institutional guidelines.

**Informed Consent Statement:** Informed consent was waived because the study used anonymized data and did not include any identifiable personal information.

**Data Availability Statement:** The data presented in this study are not publicly available due to the sensitive nature of student-related information and privacy protection requirements. The data were used exclusively for this research.

**Conflicts of Interest:** The authors declare no conflicts of interest.

## Abbreviations

The following abbreviations are used in this manuscript:

AUC	Area Under the ROC Curve
FN	False Negative
FP	False Positive
GB	Gradient Boosting
LR	Logistic Regression
MCC	Matthews Correlation Coefficient
NN	Neural Network
RF	Random Forest
TN	True Negative
TP	True Positive

## References

1. Sharma, S.; Panda, S.P.; Verma, S. Modeling and Predicting Student Enrolment: A Meta-Learning Perspective. *Int. J. Inf. Technol.* **2026**, doi:10.1007/s41870-025-03054-2.
2. Abiola-Adams, O.; Otokiti, B.O.; Olinmah, F.I.; Abutu, D.E.; Okoli, I.; Imohiosen, C. Building Performance Forecasting Models for University Enrollment Using Historical and Transfer Data Analytics. *J. Front. Multidiscip. Res.* **2021**, *2*, 162–168, doi:10.54660/JFMR.2021.2.1.162-168.
3. APatungan, A.J.; Francia, M.L.M. A Machine Learning Modeling Prediction of Enrollment among Admitted College Applicants at University of Santo Tomas.; Kedah, Malaysia, 2022; p. 020004, doi:10.1063/5.0100174.
4. Soltys, M.; Dang, H.D.; Reilly, G.R.; Soltys, K. Enrollment Predictions with Machine Learning. *Strateg. Enroll. Manag. Q.* **2021**, *9*, 11–18. Available online: <https://eric.ed.gov/?id=EJ1311204> (accessed on 3 February 2026).
5. Yang, L.; Feng, L.; Zhang, L.; Tian, L. Predicting Freshmen Enrollment Based on Machine Learning. *J. Supercomput.* **2021**, *77*, 11853–11865, doi:10.1007/s11227-021-03763-y.
6. Shao, L.; Jeong, M.; Levine, R.A.; Stronach, J.; Fan, J. Machine Learning Methods for Course Enrollment Prediction. *Strateg. Enroll. Manag. Q.* **2022**, *10*, 11–29. Available online: <https://eric.ed.gov/?id=EJ1356234> (accessed on 4 February 2026).
7. Ayasi, B.; Saleh, M.; García-Vico, A.M.; Carmona, C. Predicting Course Enrollment with Machine Learning and Neural Networks: A Comparative Study of Algorithms. In *Studies on Social and Education Sciences*

- 2023; Bilen, Ö., Shaaban, E., Eds.; ISTES Organization: Monument, CO, USA, 2023; pp. 157–182. Available online: <https://www.aaup.edu/about-university/faculty-members/bahgat-waleed-deeb-ayasi/publications/predicting-course-enrollment> (accessed on 4 February 2026).
8. Brianorman, Y.; Sucipto, S. PREDICTION OF PROSPECTIVE NEW STUDENTS USING DECISION TREE, RANDOM FOREST, AND NAIVE BAYES. *BAREKENG J. Ilmu Mat. Dan Terap.* **2024**, *18*, 1433–1446, doi:10.30598/barekengvol18iss3pp1433-1446.
  9. Raftopoulos, G.; Davrazos, G.; Kotsiantis, S. Fair and Transparent Student Admission Prediction Using Machine Learning Models. *Algorithms* **2024**, *17*, 572, doi:10.3390/a17120572.
  10. Ramos, M.C. Machine Learning-Based Enrollment Prediction for a Higher Education Institution. *Asia Pac. J. Manag. Sustain. Dev.* **2024**, *12*, 15–29, doi:10.70979/KFTM3818.
  11. Yang, L.; Du, S.; Chen, Y.; Zhang, L. Research on Prediction of College Students' Registration Based on Machine Learning and Voting Model. *Membr. Technol.* **2024**, *5*, 185–195, doi:10.52710/mt.96.
  12. He, S.; Yousefpoori-Naeim, M.; Cui, Y.; Cutumisu, M. Predicting College Enrollment for Low-Socioeconomic-Status Students Using Machine Learning Approaches. *Big Data Cogn. Comput.* **2025**, *9*, 99, doi:10.3390/bdcc9040099.
  13. Dey, D.; Haque, Md.S.; Islam, Md.M.; Aishi, U.I.; Shammy, S.S.; Mayen, Md.S.A.; Noor, S.T.A.; Uddin, Md.J. The Proper Application of Logistic Regression Model in Complex Survey Data: A Systematic Review. *BMC Med. Res. Methodol.* **2025**, *25*, 15, doi:10.1186/s12874-024-02454-5.
  14. Margalina, V.-M.; Kreienbaum, C.; Hair, J.F.; Becker, J.-M.; Ringle, C.M. Multiple Linear and Logistic Regression Analysis: A SmartPLS 4 Software Tutorial. *J. Mark. Anal.* **2026**, doi:10.1057/s41270-026-00466-2.
  15. O'Connell, N.S.; Jaeger, B.C.; Bullock, G.S.; Speiser, J.L. A Comparison of Random Forest Variable Selection Methods for Regression Modeling of Continuous Outcomes. *Brief. Bioinform.* **2025**, *26*, bbaf096, doi:10.1093/bib/bbaf096.
  16. Van Jaarsveld, B.; Hauswirth, S.M.; Wanders, N. Machine Learning and Global Vegetation: Random Forests for Downscaling and Gap Filling. *Hydrol. Earth Syst. Sci.* **2024**, *28*, 2357–2374, doi:10.5194/hess-28-2357-2024.
  17. Rizkallah, L.W. Enhancing the Performance of Gradient Boosting Trees on Regression Problems. *J. Big Data* **2025**, *12*, 35, doi:10.1186/s40537-025-01071-3.
  18. Gunasekara, N.; Pfahringer, B.; Gomes, H.; Bifet, A. Gradient Boosted Trees for Evolving Data Streams. *Mach. Learn.* **2024**, *113*, 3325–3352, doi:10.1007/s10994-024-06517-y.
  19. Ganesh Viswanathan; Gaurav Samdani; Yawal Dixit; Ranjith Gopalan Deep Learning. *World J. Adv. Eng. Technol. Sci.* **2025**, *14*, 512–527, doi:10.30574/wjaets.2025.14.3.0149.
  20. Eshraghian, J.K.; Ward, M.; Neftci, E.O.; Wang, X.; Lenz, G.; Dwivedi, G.; Bennamoun, M.; Jeong, D.S.; Lu, W.D. Training Spiking Neural Networks Using Lessons From Deep Learning. *Proc. IEEE* **2023**, *111*, 1016–1054, doi:10.1109/JPROC.2023.3308088.
  21. Garouani, M.; Barhrhouj, A.; Teste, O. XStacking : An Effective and Inherently Explainable Framework for Stacked Ensemble Learning. *Inf. Fusion* **2025**, *124*, 103358, doi:10.1016/j.inffus.2025.103358.
  22. Wu, W.; Tang, L.; Zhao, Z.; Teo, C.-P. Enhancing Binary Classification: A New Stacking Method via Leveraging Computational Geometry 2024, doi:10.48550/arXiv.2410.22722.
  23. Orange Data Mining. Orange Data Mining Toolbox. Available online: <https://orangedatamining.com/> (accessed on 8 March 2026).
  24. Xu, J.; Liu, X.; Gu, Z.; Xiao, G. A Rapid Cross-Validation Computing for Three-Way Decisions in Imbalanced Data. *Inf. Sci.* **2025**, *707*, 122016, doi:10.1016/j.ins.2025.122016.
  25. Qiu, J. An Analysis of Model Evaluation with Cross-Validation: Techniques, Applications, and Recent Advances. *Adv. Econ. Manag. Polit. Sci.* **2024**, *99*, 69–72, doi:10.54254/2754-1169/99/2024OX0213.
  26. Zeng, G. Invariance Properties and Evaluation Metrics Derived from the Confusion Matrix in Multiclass Classification. *Mathematics* **2025**, *13*, 2609, doi:10.3390/math13162609.
  27. Sathyanarayanan, S. Confusion Matrix-Based Performance Evaluation Metrics. *Afr. J. Biomed. Res.* **2024**, *4023–4031*, doi:10.53555/AJBR.v27i4S.4345.
  28. Rainio, O.; Teuhon, J.; Klén, R. Evaluation Metrics and Statistical Tests for Machine Learning. *Sci. Rep.* **2024**, *14*, 6086, doi:10.1038/s41598-024-56706-x.

29. Sujon, K.M.; Hassan, R.; Choi, K.; Samad, M.A. Accuracy, Precision, Recall, F1-Score, or MCC? Empirical Evidence from Advanced Statistics, ML, and XAI for Evaluating Business Predictive Models. *J. Big Data* **2025**, *12*, 268, doi:10.1186/s40537-025-01313-4.
30. Silwattananusarn, T.; Kanarkard, W.; Tuamsuk, K. Enhanced Classification Accuracy for Cardiotocogram Data with Ensemble Feature Selection and Classifier Ensemble. *J. Comput. Commun.* **2016**, *04*, 20–35, doi:10.4236/jcc.2016.44003.
31. StatsKingdom. Wilcoxon Signed-Rank Test Calculator. Available online: [https://www.statskingdom.com/175wilcoxon\\_signed\\_ranks.html](https://www.statskingdom.com/175wilcoxon_signed_ranks.html) (accessed on 31 March 2026).

**Disclaimer/Publisher's Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.