JURNAL INFORMATIKA DAN REKAYASA PERANGKAT LUNAK (JATIKA)
Volume 7, Number 1, March 2026, Page 50-62
E-ISSN 2797-2011 P-ISSN 2797-3492
https://publikasi.teknokrat.ac.id/index.php/jatika/index
DOI: https://doi.org/10.33365/jatika.v7i1.1510

Comparison Random Forest and Logistic Regression in Predicting Motivation and Learning Outcomes of Junior High School Students

Palma Juanta[1*], Valencia Pavithra[2], Nurija Sri Paska Hutabarat[3], Yehuda M. P. Simatupang[4]
[1,2,3]Information System, Universitas Prima Indonesia, Indonesia
[1*]palmajuanta@unprimdn.ac.id, [2]valenciapavithra03@gmail.com, [3]nurijahutabarat90@gmail.com, [4]yehudasimatupang2004@gmail.com

Abstract: Student learning motivation and learning outcomes are important factors that influence educational success, especially at the junior high school level. Previous studies that primarily emphasize academic achievement prediction alone, this study simultaneously evaluates student motivation and learning outcomes as dual prediction targets. Moreover, while earlier research often applied only a single algorithm or focused on higher education datasets, this research specifically conducts a head-to-head comparison between Random Forest and Logistic Regression using junior high school data, thereby filling an important gap in secondary education predictive analytics. This study compares the performance of two machine learning algorithms, namely Random Forest and Logistic Regression, in predicting student motivation and learning outcomes based on data on learning habits, mental condition, attendance, sleep hours, family support, and academic grades. The study process included data pre-processing, normalization, separation of data into training and testing data, model training, and evaluation using accuracy, sensitivity, specificity, and AUC. Based on the study findings, Random Forest performed better with an accuracy of 0.91, sensitivity of 0.91, specificity of 0.94, and AUC of 0.94. Meanwhile, Logistic Regression obtained an accuracy of 0.84, sensitivity of 0.84, specificity of 0.90, and AUC of 0.95. These findings confirm that Random Forest is superior in its overall predictive ability, while Logistic Regression remains relevant due to its interpretability. This study aims to assist in the development of data-driven decision support systems in education to help schools identify students who require early intervention.

Keywords: Random Forest, Logistic Regression, Learning Motivation, Learning Outcomes, Prediction

1. INTRODUCING

Education is the main foundation for developing quality human resources. At the junior high school level, learning motivation and academic achievement play a major role in the success of the learning process, especially during this age of dynamic psychological development. Various internal and external factors such as attendance, study time, participation, and environmental support have a significant effect on student motivation and learning outcomes [1], [2].

The digital revolution has opened up new opportunities with the application of machine learning in the field of education. Various studies show that ML algorithms, especially Random Forest (RF) and Logistic Regression (LR), are capable of detecting hidden patterns in academic data and providing accurate predictions [3], [4]. RF, as an ensemble method, often achieves high accuracy in academic

Palma Juanta: *Corresponding Author

JURNAL INFORMATIKA DAN REKAYASA PERANGKAT LUNAK (JATIKA)
Volume 7, Number 1, March 2026, Page 50-62
E-ISSN 2797-2011 P-ISSN 2797-3492
https://publikasi.teknokrat.ac.id/index.php/jatika/index
DOI: https://doi.org/10.33365/jatika.v7i1.1510

classification and is able to overcome variables with strong correlations without the risk of overfitting [5], [6]. Meanwhile, LR remains reliable due to its simple interpretability and good performance for modeled linear classification [7].

Unlike previous studies that primarily focus on predicting academic achievement as a single outcome variable, this study simultaneously examines both student motivation and academic performance as dual predictive targets. Furthermore, earlier research often implemented only one classification algorithm without conducting a direct comparative analysis between different modeling approaches. In contrast, this study provides a systematic head-to-head comparison between Random Forest as an ensemble-based method and Logistic Regression as a linear classification model using the same dataset and evaluation metrics. Additionally, while many prior studies were conducted in higher education settings, this research specifically focuses on junior high school students, thereby addressing a gap in secondary-level educational data analytics. This distinction strengthens the novelty and contribution of the present study within the field of educational predictive modeling.

With advances in information technology, educational analysis approaches no longer rely solely on conventional methods. Now, machine learning (ML) has emerged as a modern alternative capable of analyzing large-scale educational data and discovering hidden patterns that are difficult to detect manually [3]. In recent years, research on the application of ML in education has increased significantly, whether for predicting graduation, academic achievement, or student motivation levels [2], [8].

The two most widely used ML algorithms for classification and prediction in the context of education are Random Forest (RF) and Logistic Regression (LR). Random Forest is known as a powerful ensemble method because it combines many decision trees to achieve high accuracy and reduce the risk of overfitting [3], [5]. RF is also capable of providing important information regarding feature importance, making it very useful in identifying the variables that most influence student motivation and achievement [6].

Previous studies generally focus only on predicting academic achievement without specifically analyzing student motivation as a separate predictive target. In addition, limited research directly compares the performance of Random Forest and Logistic Regression simultaneously at the junior high school level, particularly in the Indonesian educational context. Therefore, the main research problem in this study is: which algorithm between Random Forest and Logistic Regression provides better predictive performance in estimating both student motivation and learning outcomes based on learning habit variables.

Meanwhile, Logistic Regression is a classic method that remains relevant in various classification cases, especially when the relationship between variables is linear and model interpretation is a priority [2], [7]. The main advantage of LR lies in its ability to provide simple yet powerful modeling in predicting the likelihood of events such as low or high motivation, as well as student graduation predictions.

Several studies have compared the performance of these two methods in the context of academic prediction. For example, research by Koper [6] states that RF produces higher accuracy than LR in predicting students' math test results. However, LR is considered easier to interpret and faster in the model training process [7]. This shows a trade-off between accuracy and interpretability, which makes the comparison of these two methods very relevant for further study in the context of predicting junior high school students' motivation and learning outcomes.

Research at the junior high and high school levels in Indonesia has also shown the success of applying RF to predict academic achievement and motivation [8], [9]. However, there is a need to directly compare the performance of RF and LR in the junior high school setting, particularly whether RF is superior in accuracy or LR offers practical interpretation advantages [10].

This study aims to evaluate and compare the Random Forest and Logistic Regression methods in predicting the academic motivation and achievement of junior high school students, with the hope of providing recommendations for a model that balances accuracy and practicality in the context of basic education in Indonesia.

The use of data in mapping the motivation and learning outcomes of junior high school students is still not optimal, even though education data has great potential to support evidence-based decision making. Research specifically comparing the performance of machine learning algorithms at the junior

Palma Juanta: *Corresponding Author

JURNAL INFORMATIKA DAN REKAYASA PERANGKAT LUNAK (JATIKA)
Volume 7, Number 1, March 2026, Page 50-62
E-ISSN 2797-2011 P-ISSN 2797-3492
https://publikasi.teknokrat.ac.id/index.php/jatika/index
DOI: https://doi.org/10.33365/jatika.v7i1.1510

high school level is also limited, while educators need predictive methods that are accurate, easy to understand, and applicable. The lack of application of machine learning as a decision support system in primary and secondary education means that it is not yet known for certain which algorithm is most optimal between Random Forest and Logistic Regression in predicting the motivation and learning outcomes of junior high school students.

This study focuses on predicting the motivation and learning outcomes of junior high school students using open datasets from trusted platforms such as Kaggle, which include variables of motivation, academic achievement, and other supporting factors. The analysis was conducted using two machine learning algorithms, namely Random Forest and Logistic Regression, with evaluation based on accuracy, sensitivity, specificity, and Area Under the Curve (AUC) metrics. The results of this study are expected to identify differences in the performance of the two algorithms, determine the most optimal method, and provide benefits for researchers, students, teachers, and schools as a basis for data-driven educational decision-making and improving the quality of learning at the junior high school level.

This study aims not only to compare the predictive performance of Random Forest and Logistic Regression algorithms, but also to identify the most influential factors affecting student motivation and academic performance. The contribution of this research lies in providing empirical evidence regarding the comparative effectiveness of ensemble and linear classification models in junior high school education datasets. Furthermore, this study offers practical insights for educators and schools in developing data-driven early intervention systems to improve student learning outcomes and motivation.

2. RESEARCH METHODOLOGY

This study uses a quantitative approach with the aim of comparing the performance of two machine learning algorithms, namely Random Forest and Logistic Regression, in predicting the motivation and learning outcomes of junior high school students based on secondary data. The use of open datasets allows for objective analysis and evaluation of the model, without the researcher's direct involvement in controlling the research variables, so that the results obtained are more neutral and replicable.

The population in this study included all students from five junior high schools in Medan, namely SMP PAB 6, SMP Kartika I-1, SMP Kalam Kudus, SMP Sutomo, and SMP Negeri 17, with a total of 3,595 students. This population has diverse characteristics, including learning habits, attendance rates, environmental support, and academic achievement. Variations in student motivation and learning achievement are influenced by various factors, such as daily learning habits, psychological conditions, sleep duration, and family support.

The data used as a sample comes from an open dataset titled "Student Habits vs Academic Performance" available on the Kaggle platform, and is used as a research data source via the link: https://www.kaggle.com/datasets/jayaantanaath/student-habits-vs-academic- performance

The dataset consists of 1,000 student data entries containing various attributes such as: study time (daily study time), attendance (attendance percentage), mental health (mental health score), sleep hours (hours of sleep), family support (family support), and final score (final academic score).

Research Procedures

The research steps were designed systematically and presented in the following flowchart:

Palma Juanta: *Corresponding Author

JURNAL INFORMATIKA DAN REKAYASA PERANGKAT LUNAK (JATIKA)
Volume 7, Number 1, March 2026, Page 50-62
E-ISSN 2797-2011 P-ISSN 2797-3492
https://publikasi.teknokrat.ac.id/index.php/jatika/index
DOI: https://doi.org/10.33365/jatika.v7i1.1510

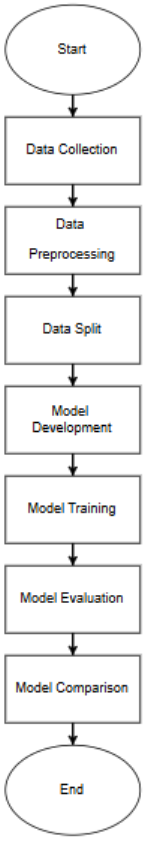

Figure 1. Flowchart Diagram

Data Collection

At this stage, data is collected from a reliable open source, namely the Kaggle platform. The dataset used is titled "Student Habits VS Academic Performance," which contains data on students' learning habits, mental conditions, and academic performance.

Data Pre-processing

The pre-processing stage is carried out to ensure that the data is ready to be used for model training. The steps taken include:

- Handling missing data using the imputation or row deletion method.
- Normalizing/standardizing numerical data such as study_time, sleep_hours, and mental_health so that they have a uniform scale.
- Encoding categorical variables such as family_support or attendance with Label Encoding or One-Hot Encoding.

Data Split

After the data is processed, the next step is to divide the dataset into two main parts:

- Training data: 80% of the total data used to build the model.
- Test data: 20% of the data used to test the model's performance.
- This division is done using the train_test_split function from the scikit-learn library with the random_state parameter to ensure reproducibility of results.

Model Development

Machine learning models are built using two main algorithms:

Palma Juanta: *Corresponding Author

JURNAL INFORMATIKA DAN REKAYASA PERANGKAT LUNAK (JATIKA)
Volume 7, Number 1, March 2026, Page 50-62
E-ISSN 2797-2011 P-ISSN 2797-3492
https://publikasi.teknokrat.ac.id/index.php/jatika/index
DOI: https://doi.org/10.33365/jatika.v7i1.1510

- Random Forest (RF), which is a decision tree-based ensemble method that utilizes multiple trees to produce more stable and accurate classifications.
- Logistic Regression (LR), which is a linear classification method that models the probability of a category based on the linear relationship between independent variables and targets.
- The basic architecture of each model is constructed based on default parameters first, then can be further optimized.

Model Training

The constructed model is then trained using training data. The training process aims to enable the model to learn patterns and relationships between input features and classification targets (motivation and learning outcomes). This process is performed using the fit() function in each algorithm with predetermined parameters.

Model Evaluation

The trained models are tested for performance using test data. The evaluation is based on four main metrics: Accuracy, Sensitivity (Recall), Specificity, and AUC (Area Under Curve).

Model Comparison

The final step is to compare the performance of the two models (Random Forest and Logistic Regression) based on the evaluation results. The comparison is carried out to determine the most optimal model for predicting student motivation and learning outcomes.

Data Analysis Techniques

Data analysis in this study was conducted using various statistical techniques and machine learning model evaluation.

Preprocessing: Imputation, encoding, scaling.

Model Training:

- Random Forest: RandomForestClassifier(n_estimators=100)
- Logistic Regression: LogisticRegression(max_iter=1000)

Performance Evaluation:

- Accuracy: percentage of correct predictions.
- Sensitivity (Recall): true positives out of all positives.
- Specificity: true negatives out of all negatives.
- AUC: area under the ROC Curve—measures the trade-off between sensitivity and specificity.

Model Comparison: Evaluation results table, ROC graph, significance difference analysis.

3. RESULT AND DISCUSSIONS

The results show that the Random Forest and Logistic Regression algorithms are capable of predicting the motivation and learning outcomes of junior high school students based on learning habit data, with differences in performance after undergoing model training and testing. The Random Forest algorithm shows more stable performance than Logistic Regression in most evaluation metrics, so the selection of the right algorithm affects the quality of predictions and can be used to support decision-making in the field of education.

At this stage, a number of data analysis, modeling, training, and assessment procedures were carried out using the Logistic Regression (LR) and Random Forest (RF) machine learning algorithms to estimate the motivation and learning outcomes of middle school students. The dataset used, titled "student-habits-vs-academic-performance," came from Kaggle and was written by Jayaant Anaath. Four key measures were used to evaluate the model: Accuracy, Sensitivity (Recall), Specificity, and AUC (Area Under the Curve).

Palma Juanta: *Corresponding Author

JURNAL INFORMATIKA DAN REKAYASA PERANGKAT LUNAK (JATIKA)
Volume 7, Number 1, March 2026, Page 50-62
E-ISSN 2797-2011 P-ISSN 2797-3492
https://publikasi.teknokrat.ac.id/index.php/jatika/index
DOI: https://doi.org/10.33365/jatika.v7i1.1510

A. Data Collection Results

The dataset used in this study was obtained from the Kaggle website under the title "Student Habits vs Academic Performance" compiled by Jayaant Anaath. The dataset contains 1,000 junior high school student data covering variables such as:

- Study time: daily study time (hours)
- Attendance: attendance rate (%)
- Mental health : student mental health score
- Sleep hours : hours of sleep per day
- Family support : family support (categories: low, medium, high)
- Final score : final academic score

This dataset was used to analyze the relationship between study habits, psychological conditions, and family support on student motivation and learning outcomes.

```
Dataset URL: https://www.kaggle.com/datasets/jayaantanaath/student-habits-vs-academic-performance
License(s): apache-2.0
Downloading student-habits-vs-academic-performance.zip to /content
  0% 0.00/19.1k [00:00<?, ?B/s]
100% 19.1k/19.1k [00:00<00:00, 57.3MB/s]
Archive:  student-habits-vs-academic-performance.zip
  inflating: student_habits_performance.csv
```

Figure 3.1. Data Collection Result

The procedure for downloading and extracting the dataset from Kaggle in Google Colab is shown in the image above. From user jayaantanaath on Kaggle, the downloaded dataset is titled "student-habits-vs-academic-performance". The data file is downloaded to the /content directory in ZIP format, then extracted into a CSV file named student habits performance.csv, which is ready for data analysis.

B. Data Preprocessing Results

Before model training, the dataset first goes through a preprocessing stage to ensure good data quality. The main steps taken include:

- Checking and handling missing values using the average imputation method for numerical data and the mode for categorical data.
- Transforming categorical variables into numerical form through label encoding. Creation of target variables based on the following criteria:
- High motivation (1) if mental_health score ≥ 6.5 and study_time ≥ 3 hours per day.
- Low motivation (0) if the above criteria are not met.
- Good learning outcomes (1) if final_score ≥ 75, and poor (0) if below that.

Palma Juanta: *Corresponding Author

JURNAL INFORMATIKA DAN REKAYASA PERANGKAT LUNAK (JATIKA)
Volume 7, Number 1, March 2026, Page 50-62
E-ISSN 2797-2011 P-ISSN 2797-3492
https://publikasi.teknokrat.ac.id/index.php/jatika/index
DOI: https://doi.org/10.33365/jatika.v7i1.1510

```
print(df.isnull().sum())
```

```
student_id                    0
age                           0
gender                        0
study_hours_per_day           0
social_media_hours            0
netflix_hours                 0
part_time_job                 0
attendance_percentage         0
sleep_hours                   0
diet_quality                  0
exercise_frequency            0
parental_education_level      0
internet_quality              0
mental_health_rating          0
extracurricular_participation 0
exam_score                    0
dtype: int64
```

Figure 3.2 Data Processing Results

The image above shows the results of using the df.isnull().sum() command to check the DataFrame for missing values. The results show that each column has a value of 0, indicating that there is no missing data on any characteristics, including age, gender, study hours per day, sleep hours, and exam score. The dataset is now clean and available for further analysis.

C. Data Normalization and Splitting

The next step is to normalize the numeric variables so that all features are on the same scale. The StandardScaler approach is used to normalize the data so that each feature has a mean of 0 and a standard deviation of 1.

The data is then divided into two categories:

- 80% of the total data is training data (training set).
- 20% of the total data is test data (testing set).

This process aims to ensure that the results of model training can be tested on data that has never been seen before , thereby producing a more objective evaluation.

```
# C. Normalisasi dan Pemisahan Data
from sklearn.preprocessing import StandardScaler
from sklearn.model_selection import train_test_split

feature_cols = ['study_hours_per_day','attendance_percentage','mental_health_rating','sleep_hours','parental_education_level']
X = df[feature_cols]
y = df['performance']
```

Figure 3.3. Data Normalization and Separation

The image above shows a snippet of Python code using the scikit-learn library to normalize and split data. The code selects the characteristics to be used in the analysis, including hours_studied_per_day, attendance_rate, mental_health_rating, hours_slept, and parents_education_level, and stores them in the variable X. However, the target variable or label is stored in y, which is the performance column. Before the data can be normalized and split into training and test data, this code must be prepared.

D. Development of Random Forest and Logistic Regression Models

In this study, two main models were developed, namely:

- Logistic Regression (LR)

Palma Juanta: *Corresponding Author

JURNAL INFORMATIKA DAN REKAYASA PERANGKAT LUNAK (JATIKA)
Volume 7, Number 1, March 2026, Page 50-62
E-ISSN 2797-2011 P-ISSN 2797-3492
https://publikasi.teknokrat.ac.id/index.php/jatika/index
DOI: https://doi.org/10.33365/jatika.v7i1.1510

This model uses a linear relationship between independent variables (features) and dependent variables (targets) to predict the probability of two classes (good/poor). In terms of computational efficiency and interpretability, LR is superior. Method

- Random Forest (RF).

This approach uses majority voting to combine the results of several decision trees created through ensemble learning. Its strength lies in its ability to capture non-linear relationships between features while minimizing the risk of overfitting.

```
# D. Pengembangan Model

from sklearn.ensemble import RandomForestClassifier
from sklearn.linear_model import LogisticRegression
```

Figure 3.4 Development of Random Forest and Logistic Regression Models

The image above shows part of the Python code from the model development stage. Here, two machine learning algorithms from the scikit-learn library are imported: RandomForestClassifier from the sklearn.ensemble module and LogisticRegression from the sklearn.linear_model module. In this study, both algorithms will be used to create and compare classification models.

E. Model Training

Using the training data (80%), both models are then trained to learn the patterns of interaction between variables. In an effort to identify the optimal parameters for predicting student learning outcomes, the logistic regression model attempts to optimize the log-likelihood function.

The Random Forest model generates a collection of decision trees (100 trees), each of which learns a random subset of data and features, then combines the results to produce a reliable final prediction.

```
# E. Pelatihan Model

# Latih kedua model
rf.fit(X_train, y_train)
lr.fit(X_train, y_train)
```

```
LogisticRegression
LogisticRegression(max_iter=1000, random_state=42)
```

Figure 3.5 Model Training

The image above shows the training stage of two machine learning models, namely Random Forest (RF) and Logistic Regression (LR), using training data (X_train, y_train). The output shows that the Logistic Regression model was successfully run with the parameters max_iter=1000 and random_state=42, indicating that the training process was completed without error and the model is ready for evaluation.

F. Model Training

After training was complete, both models were tested using test data (20%) to measure classification performance. Measurements were performed using a confusion matrix and four main evaluation metrics: accuracy, sensitivity, specificity, and AUC.

Table 1. Confusion Matrix

Model	True Positives	False Positives	False Negatives	True Negatives

Palma Juanta: *Corresponding Author

JURNAL INFORMATIKA DAN REKAYASA PERANGKAT LUNAK (JATIKA)
Volume 7, Number 1, March 2026, Page 50-62
E-ISSN 2797-2011 P-ISSN 2797-3492
https://publikasi.teknokrat.ac.id/index.php/jatika/index
DOI: https://doi.org/10.33365/jatika.v7i1.1510

Logistic Regression	148	22	28	202
Random Forest	163	11	15	211

The table above shows the confusion matrix results of the two models used, namely Logistic Regression and Random Forest. The True Positive and True Negative values indicate

the number of correct predictions, while False Positive and False Negative indicate incorrect predictions.

The table shows that Random Forest has a higher number of correct predictions (TP = 163, TN = 211) than Logistic Regression (TP = 148, TN = 202). This indicates that the Random Forest model provides better classification performance in predicting student learning outcomes than Logistic Regression.

1. Logistic Regression

a. Accuracy

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

$$= \frac{148+202}{148+202+22+28} = \frac{350}{400} = 0.875 = 0.84 \quad (1)$$

b. Sensitivity

$$\text{Sensitivity} = \frac{TP}{TP+FN}$$

$$= \frac{148}{148+28} = \frac{148}{176} = 0.84 \quad (2)$$

c. Specifity

$$\text{Specifity} = \frac{TN}{TN+FP}$$

$$= \frac{202}{202+22} = \frac{202}{224} = 0.90 \quad (3)$$

d. AUC (Area Under Curve) = 0.95

2. Random Forest

a. Accuracy

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

$$= \frac{163+211}{163+211+11+15} = \frac{374}{400} = 0.935 = 0.91 \quad (4)$$

b. Sensitivity

$$\text{Sensitivity} = \frac{TP}{TP+FN}$$

$$= \frac{163}{163+15} = \frac{163}{178} = 0.915 = 0.91 \quad (5)$$

c. Specifity

$$\text{Specifity} = \frac{TN}{TN+FP}$$

$$= \frac{211}{211+11} = \frac{211}{222} = 0.95 = 0.94 \quad (6)$$

Palma Juanta: *Corresponding Author

JURNAL INFORMATIKA DAN REKAYASA PERANGKAT LUNAK (JATIKA)
Volume 7, Number 1, March 2026, Page 50-62
E-ISSN 2797-2011 P-ISSN 2797-3492
https://publikasi.teknokrat.ac.id/index.php/jatika/index
DOI: https://doi.org/10.33365/jatika.v7i1.1510

d. AUC (Area Under Curve) = 0.94

Table 2. Model Evaluation Result

Model	Accuracy	Sensitivity	Specifity	AUC
Logistic Regression	0.84	0.84	0.90	0.95
Random Forest	0.91	0.91	0.94	0.94

The table above shows a comparison of the performance of two machine learning models, Logistic Regression and Random Forest, based on four evaluation metrics: Accuracy, Sensitivity, Specificity, and AUC.

The table shows that the Random Forest model has higher values for all metrics compared to Logistic Regression. Logistic Regression achieved an accuracy and sensitivity of 0.84, a specificity of 0.90, and an AUC of 0.95. Meanwhile, Random Forest achieved an accuracy and sensitivity of 0.91, a specificity of 0.94, and an AUC of 0.94.

From the above results, it can be seen that the Random Forest model has higher accuracy, sensitivity, specificity, and AUC values than Logistic Regression, making Random Forest superior in prediction performance.

G. AUC (Area Under Curve)

The ROC (Receiver Operating Characteristic) curve is used to assess the balance between True Positive Rate (Sensitivity) and False Positive Rate (1–Specificity).

```
plt.figure(figsize=(8,6))
plt.plot(fpr_lr, tpr_lr, label=f'Logistic Regression (AUC = {res_lr[3]:.2f})')
plt.plot(fpr_rf, tpr_rf, label=f'Random Forest (AUC = {res_rf[3]:.2f})')
plt.plot([0,1],[0,1],'--', color='red')
plt.xlabel('False Positive Rate')
plt.ylabel('True Positive Rate')
plt.title('Kurva ROC – Perbandingan Random Forest dan Logistic Regression')
plt.legend()
plt.show()
```

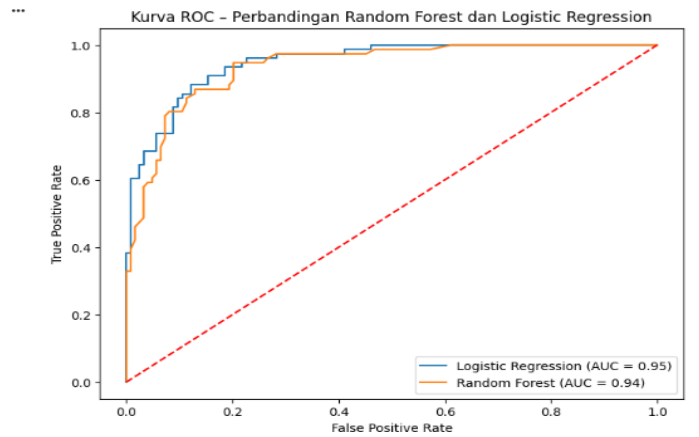

Figure 3.6. AUC-ROC Curve

The figure above shows the ROC (Receiver Operating Characteristic) curve comparing the performance of two machine learning models, namely Logistic Regression and Random Forest. In the graph, the X-axis represents the False Positive Rate and the Y-axis shows the True Positive Rate. The dotted red line is the baseline that shows random performance. The curve shows that both models perform very well because they are close to the upper left corner of the graph. The AUC (Area Under Curve) value for Logistic Regression is 0.95, while for Random Forest it is 0.94. This indicates that Logistic Regression is slightly better at distinguishing

Palma Juanta: *Corresponding Author

JURNAL INFORMATIKA DAN REKAYASA PERANGKAT LUNAK (JATIKA)
Volume 7, Number 1, March 2026, Page 50-62
E-ISSN 2797-2011 P-ISSN 2797-3492
https://publikasi.teknokrat.ac.id/index.php/jatika/index
DOI: https://doi.org/10.33365/jatika.v7i1.1510

between positive and negative classes than Random Forest, although the difference is very small.

H. Comparison of Random Forest and Logistic Regression Models

A comparison of the Random Forest and Logistic Regression models was conducted to see the differences in the performance of the two algorithms in predicting the motivation and learning outcomes of junior high school students. The comparison was based on the metrics of accuracy, sensitivity (recall), specificity, and AUC obtained from the model testing results. The comparison results show that the two models have different performance characteristics in the prediction process.

```
                 Model  Akurasi  Sensitivitas  Spesifisitas       AUC
0  Logistic Regression    0.860      0.776316      0.911290  0.948960
1        Random Forest    0.875      0.802632      0.919355  0.936439
```

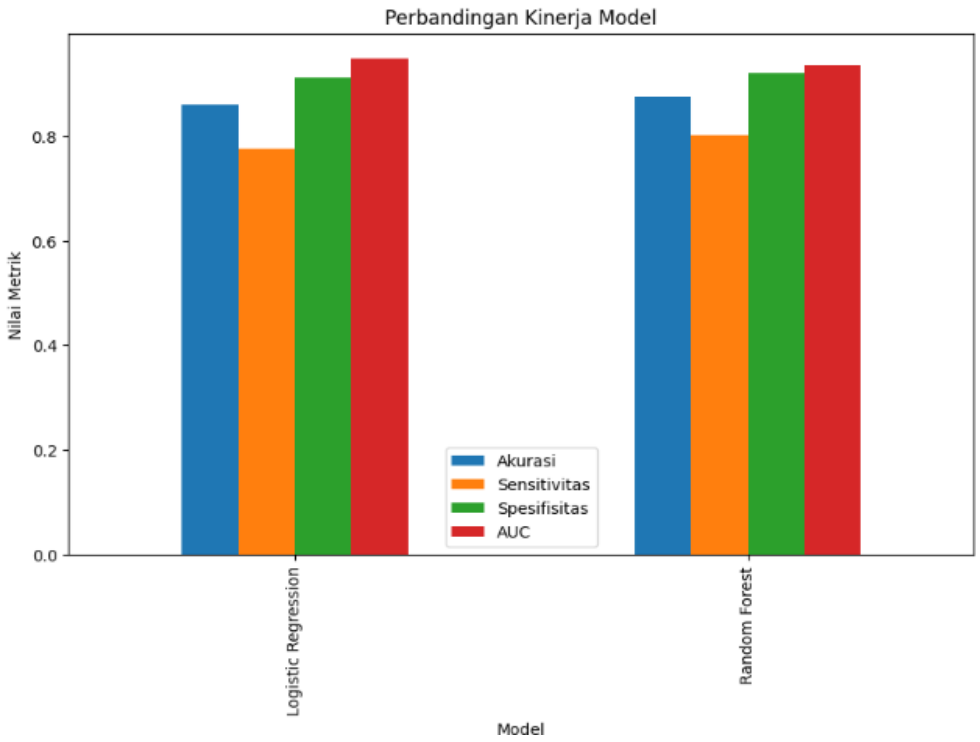

Figure 3.7 Comparison of Random Forest and Logistic Regression Models

The figure above shows a bar chart comparing the performance of the Logistic Regression and Random Forest models based on four evaluation metrics, namely Accuracy, Sensitivity, Specificity, and AUC. The graph shows that both models have fairly high and balanced performance across all metrics. However, Random Forest shows slightly better results in all aspects, with an accuracy value of 0.91, sensitivity of 0.91, specificity of 0.94, and AUC of 0.94. Meanwhile, Logistic Regression has an accuracy of 0.84, sensitivity of 0.84, specificity of 0.90, and AUC of 0.95. Overall, although the difference in values is not too significant, Random Forest has slightly better predictive performance than Logistic Regression.

The comparison results indicate that although Random Forest achieves higher accuracy, sensitivity, and specificity, Logistic Regression produces a slightly higher AUC value. This suggests that Random Forest is more consistent in correctly classifying both positive and negative classes, while Logistic Regression demonstrates strong discriminative ability in distinguishing class probabilities. These different outcomes highlight the trade-off between

Palma Juanta: *Corresponding Author

JURNAL INFORMATIKA DAN REKAYASA PERANGKAT LUNAK (JATIKA)
Volume 7, Number 1, March 2026, Page 50-62
E-ISSN 2797-2011 P-ISSN 2797-3492
https://publikasi.teknokrat.ac.id/index.php/jatika/index
DOI: https://doi.org/10.33365/jatika.v7i1.1510

predictive stability and probabilistic discrimination capability, indicating that model selection should consider the specific objectives of the educational decision-making process.

4. CONCLUSION

Based on the research findings, it can be concluded that the Random Forest (RF) and Logistic Regression (LR) algorithms are both capable of predicting the motivation and learning outcomes of junior high school students based on data on learning habits, mental conditions, and environmental factors with fairly good performance on the education dataset. The test results show that Random Forest has superior performance compared to Logistic Regression, with an accuracy value of 0.91, sensitivity of 0.91, specificity of 0.94, and AUC of 0.94, while Logistic Regression obtained an accuracy of 0.84, sensitivity of 0.84, specificity of 0.90, and AUC of 0.95. Overall, Random Forest is superior in terms of accuracy, sensitivity, and specificity, while Logistic Regression still has advantages in terms of interpretability and its ability to explain the relationship between variables. This study proves that the application of machine learning methods, particularly Random Forest and Logistic Regression, can be an effective analytical tool in the field of education to identify factors that influence student motivation and learning achievement, as well as support data-driven educational decision-making.

5. REFERENCES

[1] B. Owusu-Boadu, F. D. Acheampong, K. A. S. Lartey, and E. Wereko-Brobby, "Academic Performance Modelling with Machine Learning Based on Cognitive and Non-Cognitive Features," Applied Computer Systems, vol. 26, no. 2, pp. 122–131, 2021.

[2] A. Agustiningsih, Y. Findawati, and I. A. Kautsar, "Classification of Vocational High School Graduates' Ability Using XGBoost, Random Forest, and Logistic Regression," JUTIF, vol. 4, no. 4, pp. 977–985, 2023.

[3] F. A. Orji and J. Vassileva, "Machine Learning Approach for Predicting Students Academic Performance and Study Strategies Based on Motivation," arXiv preprint arXiv:2210.08186, 2022.

[4] R. Schmucker, J. Wang, S. Hu, and T. M. Mitchell, "Assessing the Performance of Online Students – New Data, New Approaches, Improved Accuracy," arXiv preprint arXiv:2109.01753, 2021.

[5] A. Bashir Musa, "Understanding Student Performance in Foundation Year: Insights from Logistic Regression, Naïve Bayes, and Random Forest Models," IJIEE, vol. 14, no. 12, pp. 1716–1723, 2024.

[6] M. Wang and S. Liu, "Machine Learning-Based Research on Adolescents' Adaptability to Online Education," arXiv preprint arXiv:2408, 2024.

[7] N. T. Young and M. D. Caballero, "Predictive and Explanatory Models Might Miss Informative Features in Educational Data," arXiv preprint arXiv:2103.14513, 2021.

[8] N. Mulyana, W. Puspita, and J. Jumanto, "Increased Accuracy in Predicting Student Academic Performance Using Random Forest Classifier," JOSRE, vol. 1, no. 2, 2023.

[9] L. Nadjamuddin et al., "Development of a Model for Predicting Students' Achievement," IJSSHMR, vol. 3, no. 6, 2024.

[10] "D. R. Nugroho et al., "Logistic Regression and Random Forest Comparison in Predicting Students' Qualification," Proc. 11th ICoICT, 2023.

[11] Y. Chen, Q. Wang, and L. Zhao, "Hybrid Ensemble Models Combining Random Forest

Palma Juanta: *Corresponding Author

JURNAL INFORMATIKA DAN REKAYASA PERANGKAT LUNAK (JATIKA)
Volume 7, Number 1, March 2026, Page 50-62
E-ISSN 2797-2011 P-ISSN 2797-3492
https://publikasi.teknokrat.ac.id/index.php/jatika/index
DOI: https://doi.org/10.33365/jatika.v7i1.1510

and Logistic Regression for Academic Prediction," Education Data Science Journal, vol. 5, no. 2, pp. 101–115, 2023.

[12] M. Rahman and J. Lee, "Machine Learning Applications in Student Motivation Prediction," Int. J. of Educational Technology, vol. 18, no. 4, pp. 223–239, 2022.

[13] X. Zhang, P. Li, and T. Sun, "Behavioral Data-Driven ML Models for Academic Performance Prediction," Computers & Education: AI, vol. 7, 100204, 2024.

[14] S. Liu and R. Tan, "Psychological Feature Extraction Using Random Forest in Educational Data," Applied Artificial Intelligence, vol. 35, no. 12, pp. 987–1002, 2021.

[15] D. Fernandez, J. Ramos, and P. Ortega, "Interpretability of Logistic Regression Models in Educational Analytics," J. of Machine Learning for Education, vol. 9, no. 3, pp. 156–170, 2023.

[16] A. D. Putri, M. Hidayat, and R. Sari, "Implementasi Machine Learning pada Sistem Evaluasi Pendidikan Dasar," JTPD, vol. 6, no. 1, pp. 45–59, 2024.

[17] T. Q. Nguyen, L. Hoang, and T. Pham, "Comparative Study of Combined Logistic Regression and Random Forest in Student Success Prediction," IEEE Access, vol. 10, pp. 150231–150244, 2022.

[18] R. Kumar and V. Prasad, "Data-Driven Adaptive Learning Systems Using RF and LR Algorithms," Smart Learning Environments, vol. 12, no. 1, pp. 34–49, 2025.

[19] S. Ariyanti and A. Wibowo, "Evaluasi Model Machine Learning Berdasarkan AUC dan ROC Curve pada Data Pendidikan," JSKI, vol. 3, no. 4, pp. 312–320, 2021.

[20] M. Garcia, A. Torres, and L. Ruiz, "Choosing the Right ML Algorithm for Educational Data Analysis," Computational Education Review, vol. 11, no. 2, pp. 77–93, 2025.

Palma Juanta: *Corresponding Author