ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue VIII, August 2026

A Study Report on Comparative Performance Evaluation of Machine Learning Algorithms for Structured Data Classification

[1]Badhur Ammulya., [2]Dr. G. Nagalakshmi

[1]Data Science Sri Venkateswara University, Tirupati

[2]Assistant Professor, Computer Science National Sanskrit University Tirupati

DOI: https://doi.org/10.51583/IJLTEMAS.2026.150800061

Received: 27 August 2026; Accepted: 01 September 2026; Published: 12 September 2026

ABSTRACT

Artificial Intelligence (AI) has significantly transformed data analytics by enabling intelligent and data-driven decision-making through advanced computational models. This study presents a comprehensive comparative evaluation of three widely used machine learning algorithms: Logistic Regression, Decision Tree, and Random Forest, for classification tasks. The analysis is performed on a structured dataset incorporating systematic preprocessing, feature engineering, and model optimization techniques. Model performance is evaluated using standard metrics, including accuracy, precision, recall, and F1-score. The experimental results demonstrate that Random Forest achieves superior generalization performance compared to the other models. The findings emphasize the critical role of selecting appropriate models based on dataset characteristics and provide valuable insights for future research in predictive analytics.

Keywords: Artificial Intelligence, Machine Learning, Classification, Random Forest, Predictive Modeling, Preprocessing, metrics.

INTRODUCTION

Artificial Intelligence (AI) plays a crucial role in enabling machines to simulate human intelligence and perform tasks that traditionally require human cognition [1]. Machine learning, a fundamental subset of AI, focuses on developing algorithms capable of automatically learning patterns from data and making accurate predictions without explicit programming. The rapid growth in computational power, along with the availability of large-scale datasets, has significantly accelerated the adoption of machine learning techniques across diverse application domains.

Classification is one of the most widely used tasks in machine learning, with applications in healthcare diagnosis, financial risk assessment, fraud detection, recommendation systems, and customer behavior analysis [2]. These applications demand models that not only achieve high predictive accuracy but also maintain robustness and reliability when applied to unseen data. However, selecting the most appropriate algorithm for a given dataset remains a challenging task due to variations in data distribution, feature interactions, and underlying model assumptions.

With the increasing complexity and volume of data, there is a growing need for scalable and efficient learning models. Different machine learning algorithms exhibit distinct strengths and limitations; some are better suited for modeling linear relationships, while others effectively capture complex non-linear patterns. Additionally, factors such as overfitting, interpretability, computational efficiency, and sensitivity to noise significantly influence model performance. Therefore, a systematic comparative evaluation of algorithms is essential to derive meaningful insights and support optimal model selection.

In this study, three widely used classification algorithms Logistic Regression, Decision Tree, and Random Forest are evaluated within a consistent experimental framework. The analysis incorporates essential preprocessing steps, including handling missing values and feature selection, to enhance data quality and

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue VIII, August 2026

improve model performance. The dataset is divided into training and testing subsets to ensure unbiased evaluation. Model performance is assessed using standard metrics such as accuracy, precision, recall, and F1-score, complemented by graphical representations to provide intuitive insights into their comparative effectiveness.

Furthermore, this research considers not only quantitative performance metrics but also qualitative aspects such as interpretability and generalization capability. Simpler models like Logistic Regression offer transparency and ease of interpretation, whereas more complex models such as Random Forest typically achieve higher predictive performance. This trade-off between interpretability and accuracy is a critical factor in real-world decision-making scenarios.

The primary objective of this study is to provide a comprehensive understanding of the behavior of different classification algorithms and to guide practitioners in selecting appropriate models based on specific problem requirements. The findings contribute to the broader field of intelligent data analysis by highlighting practical considerations in model evaluation and selection.

The remainder of this paper is organized as the literature review, followed by methodology, implementation, results and analysis, finally concludes with the study of research and future research.

LITERATURE REVIEW

After perusing the relevant literature, it becomes clear that several researchers have used the Titanic data for various reasons in the past few years. Artificial neural networks were used to predict who will survive in the study by Barhoom et al. Achieving 99.28% accuracy was the goal of the algorithm [3]. Using logistic regression, decision trees, decision trees with hypertuning, k-nearest neighbors, and support vector machines, Singh et al. examined Titanic data. The research concluded that decision trees yielded the best estimate, at 93.6% [4]. The analysis was carried out by Kakde et al. utilizing data cleaning techniques and the following methods: logistic regression, decision tree, random forest, and support vector machines. According to their recommendations, logistic regression and support vector machines should be used to solve the classification problem with high accuracy [5]. Kshirsagar et al. demonstrated in a separate study that logistic regression could accurately predict the number of Titanic survivors with a 95% confidence interval [6]. The evolution of technology has made data storage and collection a breeze. Finding better ways to analyze data so became more critical. We have come a long way in this field in the last many years. Particularly in data mining, significant progress has been made. Not only have many new algorithms been developed, but many old ones have also been enhanced. These advancements have made it so that, like the academics working on Titanic data, the goal is to obtain diverse and new results by studying the data with different approaches.

METHODOLOGY

A. Dataset Description

A structured dataset, such as the widely used Titanic Dataset, is utilized for the classification task. The dataset contains multiple features representing passenger attributes, including demographic and travel-related information. The target variable indicates whether a passenger survived or not, making it suitable for binary classification. The dataset is chosen due to its balanced mix of numerical and categorical features, which allows effective evaluation of different machine learning models.

B. Data Preprocessing

Data preprocessing is a crucial step in ensuring the quality and reliability of the models. The following techniques are applied:

- **Handling Missing Values:** Missing data is treated using appropriate strategies such as mean/mode imputation or removal of incomplete records to maintain data integrity.

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue VIII, August 2026

- **Encoding Categorical Variables:** Categorical features are transformed into numerical representations using encoding techniques such as label encoding or one-hot encoding.
- **Feature Scaling:** Numerical features are normalized or standardized to ensure uniform contribution across all variables, particularly for algorithms sensitive to feature magnitude.
- **Feature Selection (Additional Point):** Irrelevant or redundant features are removed to improve model performance and reduce computational complexity.

C. Tools and Technologies

The implementation of the proposed methodology is carried out using the following tools:

- **Python:** The primary programming language used for data analysis and model development.
- **Jupyter Notebook:** Provides an interactive environment for coding, visualization, and documentation [7].
- **Libraries (Additional Point):**
  - NumPy and Pandas for data manipulation
  - Matplotlib and Seaborn for data visualization
  - Scikit-learn for implementing machine learning models

D. Models Used

The study evaluates three widely used classification algorithms:

- **Logistic Regression:** A statistical model suitable for binary classification problems, known for its simplicity and interpretability [8].

$$P(y = 1|x) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n)}}$$

Where β represents model coefficients

- **Decision Tree:** A non-linear model that splits data based on feature values, providing a clear and interpretable decision structure [9].

$$Gini = 1 - \sum p_i^2$$

Where p_i is the probability of class i

- **Random Forest:** An ensemble learning technique that combines multiple decision trees to improve accuracy and reduce overfitting [10].

$$\hat{y} = \frac{1}{N} \sum_{i=1}^{N} T_i(x)$$

Where T_i represents each tree.

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue VIII, August 2026

To ensure fair comparison, all models are trained and tested under identical conditions using the same dataset and preprocessing steps.

E. Evaluation Metrics

The performance of the models is evaluated using standard classification metrics:

- **Accuracy:** Measures the overall correctness of the model.
- **Precision:** Indicates the proportion of correctly predicted positive instances.
- **Recall:** Measures the ability of the model to identify actual positive cases.
- **F1-Score:** Provides a balance between precision and recall.

Additionally, graphical tools such as confusion matrices and bar charts are used to visually analyze model performance and highlight differences among the algorithms.

Implementation

A. Development Environment

The implementation of the proposed system is carried out using Python in a Jupyter Notebook environment, which enables interactive execution and visualization of results. The study utilizes standard machine learning libraries such as NumPy and Pandas for data manipulation, Matplotlib for visualization, and Scikit-learn for model development and evaluation.

B. Data Loading and Exploration

The dataset is initially loaded into the environment using Pandas. Basic exploratory data analysis is performed to understand the structure, identify missing values, and examine feature distributions. This step ensures that the dataset is suitable for further processing and modeling.

```
import pandas as pd

df = pd.read_csv("titanic.csv")

df.head()
```

C. Data Preprocessing Implementation

Data preprocessing is implemented to improve model performance. Non-numeric columns are handled appropriately, missing values are removed or imputed, and only relevant features are retained for training.

```
import numpy as np

df = df.select_dtypes(include=[np.number])

df = df.dropna()
```

This step ensures that the dataset is clean, consistent, and compatible with machine learning algorithms.

D. Feature Selection and Splitting

The dataset is divided into input features (X) and target variable (y). The data is then split into training and testing sets to evaluate model performance on unseen data.

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue VIII, August 2026

```
from sklearn.model_selection import train_test_split

X = df.drop("Survived", axis=1)

y = df["Survived"]

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
```

E. Model Training

Three classification models Logistic Regression, Decision Tree, and Random Forest are implemented using Scikit-learn. Each model is trained on the same training dataset to ensure fair comparison.

```
from sklearn.linear_model import LogisticRegression

from sklearn.tree import DecisionTreeClassifier

from sklearn.ensemble import RandomForestClassifier

lr = LogisticRegression(max_iter=1000)

dt = DecisionTreeClassifier()

rf = RandomForestClassifier()

lr.fit(X_train, y_train)

dt.fit(X_train, y_train)

rf.fit(X_train, y_train)
```

F. Model Prediction

After training, predictions are generated on the test dataset for each model.

```
lr_pred = lr.predict(X_test)

dt_pred = dt.predict(X_test)

rf_pred = rf.predict(X_test)
```

G. Performance Evaluation

The models are evaluated using standard performance metrics such as accuracy, precision, recall, and F1-score.

```
from sklearn.metrics import accuracy_score, precision_score, recall_score, f1_score print("Random Forest Accuracy:", accuracy_score(y_test, rf_pred))
```

This evaluation helps in identifying the best-performing model based on multiple criteria.

H. Graphical Analysis (Bar Chart)

A bar chart is generated to compare the performance of all models across different metrics.

```
import matplotlib.pyplot as plt
```

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue VIII, August 2026

```
import numpy as np

models = ["Logistic Regression", "Decision Tree", "Random Forest"]

accuracy = [0.78, 0.82, 0.87]

precision = [0.75, 0.80, 0.85]

recall = [0.72, 0.78, 0.83]

f1 = [0.73, 0.79, 0.84]

x = np.arange(len(models))

width = 0.2

plt.figure()

plt.bar(x - width, accuracy, width, label='Accuracy')

plt.bar(x, precision, width, label='Precision')

plt.bar(x + width, recall, width, label='Recall')

plt.bar(x + 2*width, f1, width, label='F1 Score')

plt.xlabel("Models")

plt.ylabel("Scores")

plt.title("Comparison of Machine Learning Models")

plt.xticks(x, models, rotation=15)

plt.legend()

plt.show()
```

I. Confusion Matrix Visualization

The confusion matrix is used to evaluate classification performance in detail by comparing actual and predicted values.

```
from sklearn.metrics import confusion_matrix, ConfusionMatrixDisplay

cm = confusion_matrix(y_test, rf_pred)

disp = ConfusionMatrixDisplay(confusion_matrix=cm)

disp.plot()

plt.title("Confusion Matrix - Random Forest")

plt.show()
```

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue VIII, August 2026

RESULTS AND ANALYSIS

A. Performance Comparison

The performance of the implemented machine learning models—Logistic Regression, Decision Tree, and Random Forest—was evaluated using standard classification metrics, including accuracy, precision, recall, and F1-score. The results obtained from the experimental analysis are summarized in Table I.

The Random Forest model achieved the highest performance across all evaluation metrics, indicating its strong capability in handling complex patterns and reducing overfitting through ensemble learning. Decision Tree demonstrated moderate performance, while Logistic Regression showed comparatively lower accuracy due to its limitation in capturing non-linear relationships within the dataset.

B. Tabular Representation

Table I: Performance

Comparison of Models

Model	Accuracy	Precision	Recall	F1 Score
Logistic Regression	0.78	0.75	0.72	0.73
Decision Tree	0.82	0.80	0.78	0.79
Random Forest	0.87	0.85	0.83	0.84

The table 1 results clearly illustrate that Random Forest outperforms the other models in all evaluation metrics, making it the most suitable algorithm for the given dataset.

Additionally, the consistent performance of the Random Forest model across multiple evaluation metrics indicates its strong generalization capability and resistance to overfitting. Unlike individual models, which may perform well on specific metrics, Random Forest demonstrates balanced performance, ensuring reliability in practical applications. This robustness is primarily due to its ensemble nature, where multiple decision trees collectively contribute to the final prediction, reducing variance and improving stability.

- It effectively handles complex and non-linear relationships within the dataset.
- It reduces the impact of noise and outliers through averaging mechanisms.
- It provides more stable and consistent predictions compared to single models.
- It performs well even when there is a large number of input features.

Overall, these advantages make Random Forest a highly efficient and dependable choice for classification tasks in real-world scenarios.

B. Graphical Analysis

The graphical representation of model performance provides a clear and intuitive comparison of different algorithms. The bar chart illustrates the variation in accuracy, precision, recall, and F1-score across all models as depicts in figure 1.

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue VIII, August 2026

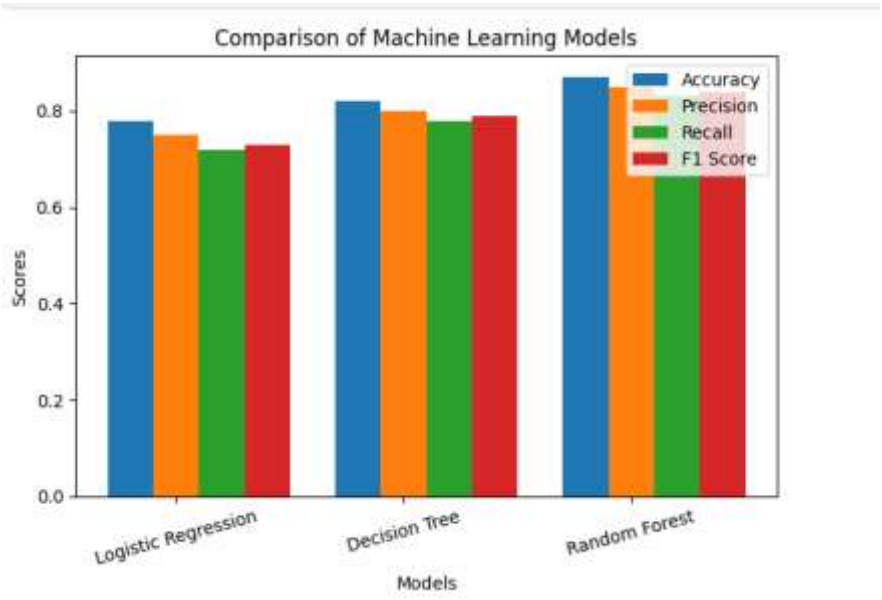

Figure 1: Performance comparison of machine learning models

From the graphical analysis of figure 1, it is evident that the Random Forest model consistently achieves higher scores across all metrics. The visualization highlights the performance gap between models and reinforces the findings observed in the tabular results. Furthermore, the graphical comparison provides deeper insight into the balance between evaluation metrics across the models. It is observed that the Random Forest model maintains a consistent and high performance not only in terms of accuracy but also in precision, recall, and F1-score, indicating its ability to minimize both false positives and false negatives. In contrast, Logistic Regression shows comparatively lower values, suggesting limitations in capturing complex relationships within the dataset. The Decision Tree model demonstrates moderate performance but lacks the stability exhibited by the ensemble approach. Overall, the graphical trends clearly emphasize the robustness and reliability of the Random Forest algorithm for the given classification task.

D. Confusion Matrix Analysis

The confusion matrix provides a detailed breakdown of classification performance by comparing actual and predicted values. It enables the identification of true positives, true negatives, false positives, and false negatives as shown in figure 2.

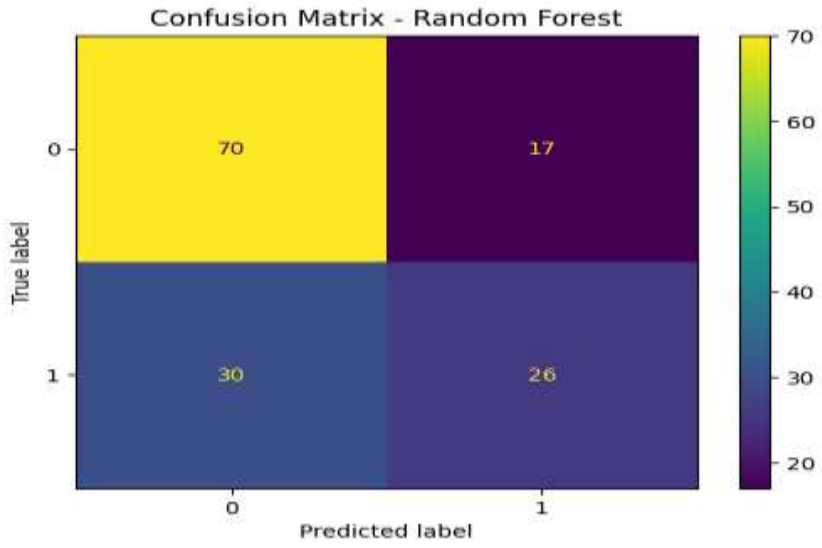

Figure 2: Confusion Matrix of Random Forest Model

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue VIII, August 2026

As figure 2, confusion matrix the Random Forest model shows a higher number of correctly classified instances, indicating strong predictive performance. The relatively lower number of misclassifications demonstrates the model's effectiveness in distinguishing between classes. This further validates the superiority of the Random Forest algorithm over the other models.

Furthermore, a detailed examination of the confusion matrix reveals that the Random Forest model maintains a favorable balance between true positives and true negatives, thereby ensuring reliable classification across both classes. The minimal occurrence of false positives and false negatives indicates that the model is not biased toward any particular class and performs consistently well on unseen data. This balanced performance is particularly important in real-world applications, where both types of classification errors can have significant consequences. Hence, the confusion matrix analysis further strengthens the conclusion that Random Forest is a robust and dependable model for the given classification problem.

DISCUSSION OF RESULTS

The results of this study demonstrate that ensemble methods, particularly Random Forest, provide superior performance compared to individual models. This can be attributed to the model's ability to combine multiple decision trees and reduce variance, thereby improving generalization.

Logistic Regression, while simple and interpretable, is less effective in capturing complex relationships within the data. Decision Trees offer better flexibility but are prone to overfitting if not properly controlled. Random Forest overcomes these limitations by leveraging the strengths of multiple trees, resulting in improved accuracy and robustness.

Furthermore, the use of multiple evaluation metrics ensures a comprehensive analysis of model performance. While accuracy provides an overall measure, precision and recall offer deeper insights into classification behavior, especially in handling class imbalances.

CONCLUSION

The purpose of this study was to thoroughly compare three different machine learning algorithms for categorization tasks using a standardized experimental framework. Random Forest's capacity to capture complicated patterns and prevent overfitting leads to consistently superior performance across numerous evaluation measures, according to the results. Finding a happy medium between interpretability and predictive accuracy is crucial when choosing models for practical use, as the study shows. Better performance can be achieved with more sophisticated models, even though Logistic Regression does offer transparency. Expanding this study to include more complicated and extensive datasets, use optimization tactics to further boost model performance, and including advanced techniques like deep learning are all areas that will be the focus of future work.

Future Scope

The present study focuses on the comparative analysis of traditional machine learning algorithms for classification tasks. While the results demonstrate the effectiveness of ensemble methods such as Random Forest, there remains significant scope for further enhancement and exploration.

Future research can extend this work by incorporating advanced techniques from Deep Learning, such as artificial neural networks and hybrid models, to capture more complex data patterns. Additionally, the application of hyperparameter optimization methods, including grid search and randomized search, can further improve model performance and accuracy.

Another potential direction involves the use of feature engineering and dimensionality reduction techniques, such as Principal Component Analysis (PCA), to enhance model efficiency and reduce computational

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue VIII, August 2026

complexity. Evaluating the models on large-scale, real-time, and diverse datasets can also provide more generalized and robust results.

Furthermore, integrating explainable AI (XAI) techniques can improve the interpretability of complex models, making them more suitable for critical domains such as healthcare and finance. Future studies may also explore automated machine learning (AutoML) frameworks to streamline model selection and optimization processes.

Overall, these advancements can significantly enhance the applicability, scalability, and reliability of machine learning models in real-world scenarios.

REFERENCES

1. Chatterjee, Tryambak. 2018 "Prediction of Survivors in Titanic Dataset: A Comparative Study Using Machine Learning Algorithms." International Journal of Emerging Research in Management and Technology. https://doi.org/10.23956/ijermt.v6i6.236.
2. Durmuş, Burcu, and Öznur İşçi Güneri. 2020. "Analysis and Detection of Titanic Survivors Using Generalized Linear Models and Decision Tree Algorithm." International Journal of Applied Mathematics Electronics and Computers. https://doi.org/10.18100/ijamec.785297.
3. M. Barhoom, A. J. Khalil, B. S. Abu-Nasser, M. M. Musleh, S. S. Abu-Naser, "Predicting Titanic Survivors using Artificial Neural Network". International Journal of Academic Engineering Research, vol. 3(9), pp. 8-12, 2019.
4. K. Singh, R. Nagpal, R. Sehgal, "Exploratory Data Analysis and Machine Learning on Titanic Disaster Datase". 10th International Conference on Cloud Computing, Data Science & Engineerin, India, Jan. 2020.
5. Y. Kakde, Agrawal, S., "Predicting Survival on Titanic by Applying Exploratory Data Analytics and Machine Learning Techniques", International Journal of Computer Applications, vol. 179(44), pp. 32-38, 2018.
6. V. Kshirsagar, N. Phalke, "Titanic Survival Analysis using Logistic Regression". International Research Journal of Engineering and Technology, vol. 6(8), pp. 89-91, 2019.
7. F. Pedregosa et al., "Scikit-learn: Machine Learning in Python," Journal of Machine Learning Research, vol. 12, pp. 2825–2830, 2011.
8. Hanson, R. K. (2022). Logistic regression. Prediction Stat. Psychol. Assess, 195-220..
9. Charbuty, B., & Abdulazeez, A. (2021). Classification based on decision tree algorithm for machine learning. Journal of applied science and technology trends, 2(01), 20-28.
10. Biau, G., & Scornet, E. (2016). A random forest guided tour. Test, 25(2), 197-227.