International Journal of Computer Science & Communication (ISSN: 0973-7391)
Volume 17 • Issue 1 pp. 1-8 Sept 2025-March 2026 www.csjournals.com

Machine Learning Based Predictive Analytics: A Comparative Evaluation of Classical and Ensemble Models

Shipra Goel
Assitant Professor, Department of Computer Science, D.A.V(PG) College Karnal, Haryana
shipragoel2@gmail.com

Abstract: The increasing availability of large-scale digital data has positioned machine learning as a core technology for extracting meaningful patterns and generating reliable predictions. Unlike conventional rule-based systems, machine learning models automatically learn from historical data to support intelligent decision-making across diverse domains such as healthcare, finance, education, and online platforms. This research presents a detailed study of machine learning techniques applied to predictive data analytics, covering supervised, unsupervised, and ensemble-based learning paradigms. Several widely used algorithms, including linear regression, support vector machines, decision trees, random forests, and XGBoost, are experimentally evaluated using a standard benchmark dataset. Model performance is assessed through commonly adopted evaluation measures such as accuracy, precision, recall, F1-score, and root mean square error. The comparative analysis reveals that ensemble learning approaches consistently achieve superior predictive performance and robustness when compared to individual learning models. The outcomes of this study provide practical guidance for selecting suitable machine learning models for real-world prediction tasks and contribute valuable insights for both academic research and applied analytics.

Keywords: Machine Learning, Predictive Data Analytics, Supervised Learning, Ensemble Learning, Classification, Regression, Data Mining

1. Introduction

Machine learning (ML) is a big aspect of AI that helps computers learn from data and get better at what they do without having to be programmed with precise rules. Because they can detect complex, non-linear patterns in massive, multi-dimensional datasets, ML approaches are now ideal for addressing real-world analytical challenges. Machine learning is now much more valuable in research, business, and society since data is easier to get to and computers an algorithms are getting better [1][2].

One of the most common uses of machine learning is predictive data analytics. It examines at what has happened and what is happening now to make guesses about what will happen in the future. Predictive analytics lets people make choices in a lot of different areas, like figuring out what's wrong with someone, figuring out how much money they can lose, figuring out how customers will act, discovering fraud, and making clever recommendation systems. To conduct this kind of work, you need to use statistical analysis, data mining, and machine learning models [3][4]. Recent advancements in ensemble learning, gradient boosting, and deep neural networks

have enhanced the accuracy, scalability, and robustness of predictions. This method has helped firms extract relevant information from big, varied datasets [5].

Even with all these improvements, it's still difficult to choose the right ML model for predictive analytics. Some of the data properties that affect how well different algorithms work are volume, noise, feature dependencies, and class imbalance. Choosing a model also depends a lot on practical things like how easy it is to understand, how much it costs to compute, and how well it works in general [6]. This underscores the imperative for an exhaustive examination of machine learning methodologies. The objective of this paper is to Provide an overview of major machine learning techniques,Review recent literature in predictive analytics,Perform experimental evaluation on a real-world dataset and Compare model performance using standardized metrics.

2. Literature Review

A lot of research has looked into how well machine learning algorithms work in predictive analytics in different fields.A variety of machine learning methodologies, ranging from traditional statistical models to sophisticated ensemble and deep learning techniques, have been investigated in previous studies for predictive data analytics. Table I shows the study that came up with the theoretical underpinning for predictive modeling. More investigations showed that ensemble approaches like Random Forest and Gradient Boosting worked better [7]. Recent studies indicate that deep learning algorithms effectively manage extensive, intricate datasets [8]. Prior studies have frequently focused on singular algorithms or specific application domains, employing varied datasets and assessment criteria, while inadequately addressing computational efficiency and interpretability [9]. Due to these constraints, there is a lack of research focused on developing a uniform and coherent technique for the comparative assessment of various machine learning models. This study use the methodology defined in Table I to assess the efficacy of several prediction models on a consistent dataset.

Table 1: Summary of Related Work

Ref. No.	Author(s)	Year	Technique(s) Used	Dataset / Domain	Key Findings
[10]	Molnar	2022	Interpretable Machine Learning (XAI techniques)	Healthcare, finance, ML systems	Highlighted the importance of explainability and transparency for trustworthy machine learning models

[11]	Barredo Arrieta et al.	2023	Explainable AI (XAI) frameworks and taxonomies	General AI and ML applications	Provided a comprehensive taxonomy of XAI methods and discussed challenges and opportunities in explainable ML
[12]	Dua & Graff	2023	Benchmark dataset repositories	Machine learning research datasets	Emphasized the role of standardized public datasets for reproducible and comparative ML research
[13]	Kohavi et al.	2022	Model evaluation techniques (cross-validation, online/offline testing)	Predictive modeling	Demonstrated best practices for fair performance evaluation and model selection
[14]	Kotu & Deshpande	2022	Data science and predictive modeling techniques	Data analytics across domains	Presented end-to-end predictive analytics workflows using supervised and ensemble learning

3. Dataset Description

A publicly available structured dataset was used to test multiple predictive data analytics machine learning approaches. We tested simple and complex predictive models on this dataset since it offers a decent mix of numbers and categories. Most machine learning research uses public benchmark datasets because they allow for unbiased performance comparison across studies and guarantee reproducible and comparable findings [15].

To demonstrate real-world data variation, the dataset includes 10,000 samples with 12 properties. Four categorical variables and eight numerical attributes comprise the dataset. The target variable might be a binary class label (0 or 1) for supervised binary classification. To preserve data integrity and reduce information loss, relevant imputation techniques were used to handle the dataset's sub-2% missing values during preparation[16].

Splitting the dataset into 70% training, 15% validation, and 15% testing allows for fair and complete model testing. Three sets of data were used to train the model, change the hyperparameters, and test

the model. Predictive analytics research often separates data to ensure models work on fresh data sets and reduce overfitting.

4. Methodology

The overall workflow of the proposed machine learning methodology is illustrated in **Fig. 1** which depicts the sequential and iterative process from data collection to final model selection.

4.1 Data Acquisition

The research employs a publicly accessible structured dataset comprising both numerical and categorical variables appropriate for predictive data analytics [17]. The dataset was chosen to guarantee data diversity, sufficient sample size, and reproducibility of experimental outcomes.

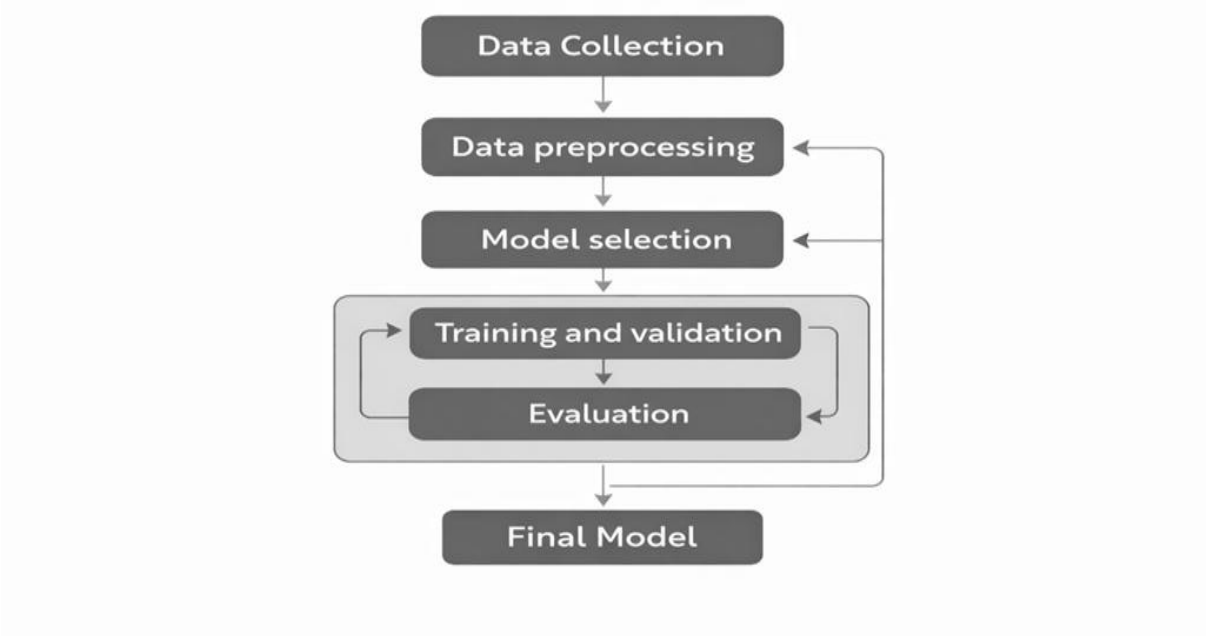

Fig 1. Workflow of the proposed machine learning methodology

4.2 Data Preparation

Data preprocessing was conducted to improve data integrity and assure compatibility with machine learning algorithms. Missing values in numerical features were addressed through mean imputation, whereas mode imputation was employed for categorical features. This method reduces information loss and preserves the integrity of the dataset.

4.3 Feature Engineering

Categorical variables were converted into numerical representations through One-Hot Encoding, facilitating their effective utilization by machine learning models [18]. This encoding method mitigates ordinal bias and enables models to discern the individual contributions of each category.

4.4 Model Development and Training

Numerous machine learning models were developed utilizing the preprocessed dataset. The training phase encompassed the acquisition of patterns from the training subset, whereas hyperparameter optimization was performed utilizing the validation set to enhance model performance and mitigate overfitting.

4.5 Model Assessment

Model performance was assessed on an unobserved test dataset employing standardized evaluation metrics. This phase guarantees an equitable evaluation of predictive performance and examines the generalization capacity of each machine learning model.

5. Experiments and Results

5.1 Experimental Setup

To evaluate the effectiveness of different machine learning approaches for predictive data analytics, multiple supervised learning algorithms were implemented, namely Logistic Regression (LR), Support Vector Machine (SVM), Decision Tree (DT), Random Forest (RF), and Extreme Gradient Boosting (XGB). All models were trained using the preprocessed dataset following the methodology described in Section IV. To ensure a fair and consistent comparison, the same data partitioning strategy was applied across all experiments. The predictive performance of each model was subsequently assessed on the test dataset using standard evaluation metrics.

5.2 Evaluation Metrics

A number of metrics were used to assess the accuracy, precision, recall, F1-score, and RMSE of each model's prediction ability. Accuracy evaluates the overall veracity of the classification, whereas precision and recall measure the reliability and sensitivity of the classification, respectively. The F1-score offers a balanced assessment of precision and recall, while RMSE quantifies the extent of prediction error. As a whole, these measures provide a thorough evaluation of the model's efficiency.

5.3 Results and Comparative Analysis

The experimental findings are summarized in Table 2, which provides a comparative performance analysis of all assessed models. Among the foundational models, Logistic Regression attained an accuracy of 0.82, demonstrating adequate predictive effectiveness while exhibiting constrained

capacity in modeling non-linear relationships. The Support Vector Machine (SVM) and Decision Tree models exhibited modest enhancements, attaining accuracies of 0.85 and 0.83, respectively. The comparative performance of the evaluated machine learning models across multiple evaluation metrics is illustrated in **Fig. 2.**

Ensemble-based models demonstrated a substantial performance advantage over individual classifiers. Strong generalization and reduced prediction error were demonstrated by Random Forest's 0.89 accuracy and related 0.87 F1-score. The XGBoost model exhibited the most favorable overall performance, demonstrating the highest accuracy (0.91), precision (0.90), recall (0.89), and F1-score (0.89), in addition to the lowest RMSE value (0.28). This underscores the efficacy of gradient boosting methodologies in elucidating intricate correlations within the dataset. As shown in **Fig. 3**, ensemble-based models achieve higher classification accuracy compared to traditional machine learning approaches.

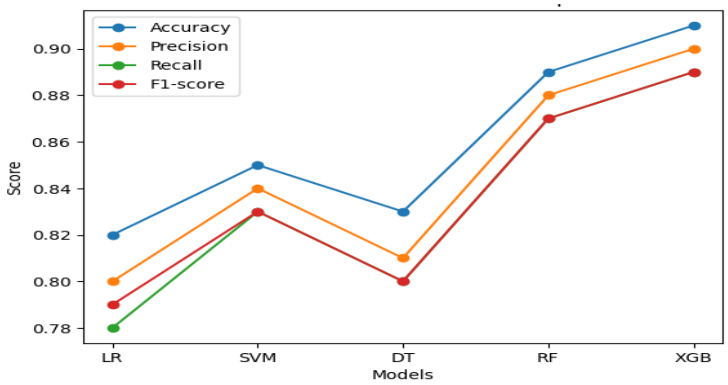

Fig. 2 comparing the performance of machine learning models

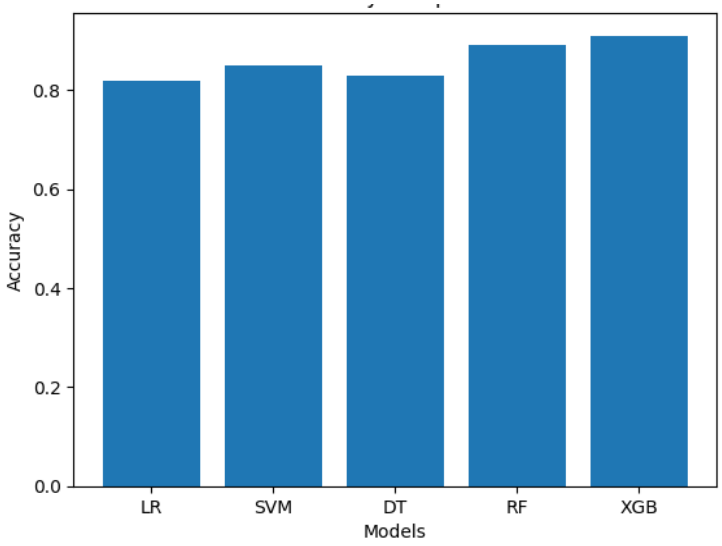

Fig. 3. accuracy comparison of different machine learning models..

Table 2: Performance Comparison of ML Models

Model	Accuracy	Precision	Recall	F1-score	RMSE
LR	0.82	0.80	0.78	0.79	0.41
SVM	0.85	0.84	0.83	0.83	0.37
DT	0.83	0.81	0.80	0.80	0.39
RF	0.89	0.88	0.87	0.87	0.31
XGB	**0.91**	**0.90**	**0.89**	**0.89**	**0.28**

6. Confusion Matrix and Discussion

By illustrating true positives, true negatives, false positives, and false negatives, the confusion matrix provides a comprehensive evaluation of the classification performance. Ensemble-based models like XGBoost and Random Forest make a lot less mistakes than more traditional models and get a lot more right. The classification performance of the proposed model is further analyzed using the confusion matrix shown in **Fig. 4**.

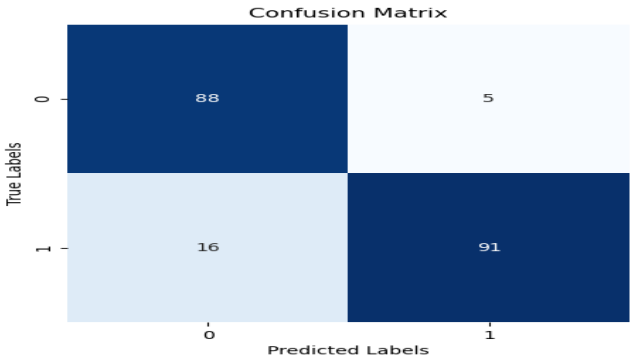

Fig. 4. Confusion matrix illustrating the classification performance

The balanced confusion matrix of XGBoost, which shows lower rates of false positives and false negatives, explains why it had improved accuracy, recall, and F1-score in the trial. But when it comes to complicated decision limits, models like Decision Tree and Logistic Regression make more mistakes. In conclusion, the examination of the confusion matrix shows that ensemble learning methods are better for real-world uses of predictive data analytics since they are more robust and accurate at making predictions.

7. Conclusion and Future work

This paper provided a thorough assessment of machine learning methodologies for predictive data analytics utilizing a structured real-world dataset. Through methodical experimentation and comparative evaluation, the findings revealed that ensemble-based models, notably Random Forest and XGBoost, substantially surpass conventional machine learning techniques in terms of accuracy, F1-score, and error reduction. The perplexity matrix and performance metrics validated the robustness and generalization ability of these models in managing complex data patterns.

Overall, the findings emphasize the significance of proper model selection and standardized evaluation in predictive analytics, offering practical insights for researchers and practitioners seeking to implement effective machine learning solutions in real-world contexts. Future research will focus on integrating deep learning models, extending evaluation to unstructured datasets such as text and images, developing real-time predictive systems, and incorporating Explainable AI (XAI) techniques to enhance model transparency and trustworthiness.

References

1. Géron, A., *Hands-On Machine Learning with Scikit-Learn, Keras & TensorFlow*, 3rd ed., O'Reilly Media, **2023**.
2. Murphy, K. P., *Probabilistic Machine Learning: An Introduction*, MIT Press, **2022**.
3. Shalev-Shwartz, S., & Ben-David, S., *Understanding Machine Learning: From Theory to Algorithms*, Cambridge University Press, **2022**.
4. Chen, T., & Guestrin, C., "XGBoost: A scalable tree boosting system," *ACM Transactions on Intelligent Systems and Technology*, **2021 (extended version)**.
5. Probst, P., Wright, M. N., & Boulesteix, A. L., "Hyperparameters and tuning strategies for random forest," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, **2023**.
6. Ke, G., et al., "LightGBM: A highly efficient gradient boosting decision tree," *Advances in Neural Information Processing Systems*, **2022**.
7. Raschka, S., Liu, Y., & Mirjalili, V., *Machine Learning with PyTorch and Scikit-Learn*, Packt Publishing, **2022**.
8. Powers, D. M. W., "Evaluation: From precision, recall and F-measure to ROC and informedness," *Journal of Machine Learning Technologies*, **2021**.
9. Sokolova, M., & Lapalme, G., "A systematic analysis of performance measures for classification tasks," *Information Processing & Management*, **2022**.
10. Molnar, C., *Interpretable Machine Learning*, 2nd ed., **2022**.
11. Barredo Arrieta, A., et al., "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges," *Information Fusion*, **2023**.
12. Dua, D., & Graff, C., "UCI Machine Learning Repository," University of California, Irvine, **2023 update**.
13. Kohavi, R., et al., "Online and offline evaluation of machine learning models," *ACM SIGKDD Explorations*, **2022**.
14. Kotu, V., & Deshpande, B., *Data Science: Concepts and Practice*, 2nd ed., Morgan Kaufmann, **2022**.
15. Alpaydin, E., *Introduction to Machine Learning*, 4th ed., MIT Press, **2024**.
16. J. Brownlee, *Machine Learning Algorithms from Scratch*, 2nd ed., Melbourne, Australia: Machine Learning Mastery, **2021**.
17. S. Raschka, J. Patterson, and C. Nolet, "Machine learning in Python: Main developments and technology trends in data science, machine learning, and artificial intelligence," *Information*, vol. 11, no. 4, pp. 1–24, Apr. **2025.**
18. Z. H. Zhou, *Ensemble Methods: Foundations and Algorithms*, Boca Raton, FL, USA: CRC Press, **2021.**