Advanced Data Processing using Machine Learning: A Comparative Study of Predictive Models for Large-Scale Data Analytics

Jerald Nirmal Kumar S[1], Pawan Kumar Verma[2]
[1,2]Lincoln University College, Petaling Jaya, Malaysia
pdf.jerald@lincoln.edu.my, jeraldcse@gmail.com, abes.pawan@gmail.com

Abstract: As the amount of data collected in various fields continues to grow, the challenge of deriving meaningful information for making sound decisions is growing apace. The technology of advanced data processing, combined with machine learning (ML), has proven to be an effective approach to analytics, pattern recognition and accurate predictions. In this paper, some of the popular ML models are compared in the context of predictive analytics to large volume data environment. The proposed workflow is as follows: Data preparation, feature selection and transformation, model training, and evaluation of the results to determine the effectiveness of the five supervised classification methods that are used in the study (Logistic Regression, Decision Tree (DT), Random Forest (RF), Support Vector Machine (SVM), and Extreme Gradient Boosting (XGBoost)). They are evaluated based on accuracy, precision, recall, F1 score, and training time of the model for model performance, as well as predictive power, and scalability in terms of efficiency. Moreover, the study examines the validity of the models for real life scenarios, such as trend forecasting, recommendation systems, predicting customer greed and decision support systems. The results have practical suggestions for selection of suitable machine learning algorithms based on your output data volume, and based on your application requirements. The comparative study supports the development of intelligent, data-driven systems, by specifying the advantages and disadvantages of the available predictive models, and by pointing towards future development of scalable and efficient machine learning based data analysis.
Keywords: Predictive Analytics, Large-Scale Data Analytics, Feature Engineering, Support Vector Machine, Predictive Modeling.

Introduction

Recent advances in digital technologies, cloud platforms, IoT, and social networking have resulted in the generation of large volumes of structured as well as unstructured data. The process of obtaining useful information from the massive amount of data has presented organizations with difficulties in many fields such as healthcare, finance, retail, education, and smart cities. Due to advanced data processing methods combined with machine learning (ML), data processing became effective and ensured efficient decision-making and predictive analytics [1]. Machine learning can support pattern discovery, classification of data, and prediction of future outcomes from complex datasets. The study considers Logistic Regression, Decision Tree, Random Forest, Support Vector Machine (SVM), and Extreme Gradient Boosting (XGBoost) as representative predictive models and assessment of the effectiveness of various real-life cases. Nevertheless, their performance depends on such aspects as data preprocessing, feature engineering, data set properties, and computational efficiency [2], [3].

Recently, studies have shown that machine learning has been effectively used in retail analytics, advertisements, recommendation systems, healthcare, and crime forecasting [4 - 10]. While the findings are encouraging, most of these studies deal with specific areas of application rather than broad data mining approaches. There has not been any recent empirical research that involves comparing the most widely used machine learning techniques and approaches based on common

evaluation criteria. In this research paper, five machine learning techniques—LR, DT, RF, SVM, and XGBoost—are analyzed and compared with respect to predictive analytic techniques. In comparison to one another, these models would be evaluated in terms of the quality of their work based on common performance criteria—the accuracy of the work done by the model and its effectiveness.

The major contributions of the present study can be summarized as follows:

- Developing a machine-learning-based framework for processing and analyzing large-scale data.
- A comparative evaluation is performed across five widely used predictive algorithms.
- The selected models are compared using accuracy, precision, recall, F1-score, and overall computational performance.
- The study identifies the model that provides the most suitable performance for large-scale predictive analytics.

The remainder of this paper is organized into six sections. Section II discusses previous research, while Section III presents the proposed methodology. Section IV describes the experimental setup, Section V discusses the obtained results, and Section VI concludes the study and presents possible future directions.

Related Work

Big datasets have been processed across various application areas using machine learning techniques that are increasingly employed towards predictive analytics. Rodrigues and Givigi [1] discussed optimization-based approach for predictive analytics and highlighted the selection of the right machine-learning models to get better predictions while consuming less time and resources. They explored the emergence of requirement for scalable predictive systems where data-rich environments exist.

Fatima and her co-authors compared the performance of Random Forest and XGBoost algorithms and found that ensemble-based algorithms showed better classification performance and robustness than the traditional ones. In another case, Santoso and Priyadi [3] researched the concept of feature engineering and concluded that their research found that feature selection and preprocessing can improve the performance of the models and minimize the computation time.

There have been multiple studies discussing the use of predictive analytics solutions in their respective domains. Dahake et al. [4] developed a machine-learning approach for analyzing and predicting customer behaviour in the retail domain. Satheeskumar et al. [5] investigated the application of machine learning to predictive marketing analytics, particularly for improving customer segmentation. Singh et al. [6] proposed a machine-learning-based recommendation system for hotel-related applications.

Predictive analytics has also been explored extensively in healthcare. Previous research has applied machine-learning methods to the prediction of cardiovascular conditions. Charan et al. [9] compared machine-learning and deep-learning approaches in healthcare. Krishna et al. [10] evaluated different machine-learning methods and reported improved prediction performance, while Rajareddy et al. [8] demonstrated the application of machine learning to crime-pattern prediction using historical data.

There have been a number of studies showing the efficacy of machine learning in various fields, but most of them apply to a certain application or to a limited set of algorithms. There is still a lack of comparisons made between many commonly used widely applied predictive models within a single

framework. To overcome this drawback, this paper compares the Logistic Regression, Decision Tree, Random Forest, Support Vector Machine (SVM), and XGBoost models in the same experimental setup, with the same performance metrics.

Table 1. Summary of Related Work

Ref.	ML Technique	Application	Key Contribution	Limitation
[1]	Optimization-based ML	Predictive Analytics	Improved predictive decision-making using optimization techniques.	No comparative evaluation of ML models.
[2]	Random Forest, XGBoost	Classification	Compared ensemble algorithms and achieved high prediction accuracy.	Limited to two algorithms.
[3]	Feature Engineering	Predictive Analytics	Enhanced prediction performance through feature selection and preprocessing.	No comparison of ML algorithms.
[4]	Supervised ML	Retail Analytics	Predicted customer buying behavior for retail applications.	Domain-specific evaluation.
[5]	ML Models	Marketing Analytics	Improved customer segmentation and marketing prediction.	Scalability not analyzed.
[6]	Recommendation Model	Hotel Recommendation	Developed a personalized recommendation system.	Evaluated on a small dataset.
[7]	Machine Learning	Healthcare	Predicted cardiovascular disease effectively.	Limited performance comparison.
[8]	Machine Learning	Crime Prediction	Forecasted crime patterns using historical data.	Focused on a single domain.
[9]	ML and Deep Learning	Healthcare Analytics	Compared ML and DL models for disease prediction.	High computational cost.
[10]	Comparative ML	Cardiology	Evaluated multiple ML algorithms for diagnosis.	Scalability not considered.

Proposed Methodology

This paper introduces a comprehensive machine-learning framework for large-scale data processing and predictive analysis purposefully designed with the use of CICIDS2017 benchmark dataset. The workflow involves data preprocessing, feature extraction, model construction, and performance evaluation, which are implemented under the same experimental conditions. Five supervised algorithms—Logistic Regression (LR), Decision Tree (DT), Random Forest (RF), Support Vector Machine (SVM), and Extreme Gradient Boosting (XGBoost)—are evaluated as part of the framework. The overall workflow is shown in Figure 1.

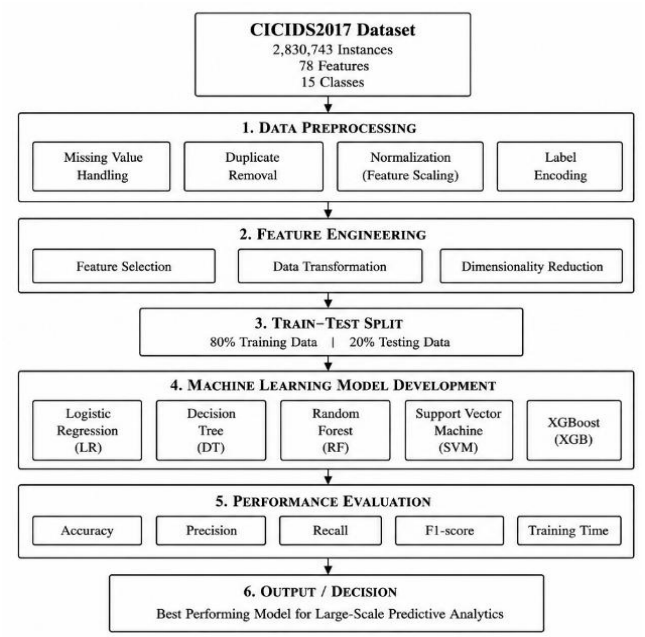

Figure 1. Overall Framework of the Proposed Methodology

a. Dataset Description

The experiments are conducted on the CICIDS2017 dataset, which is a benchmark dataset for network intrusion detection research. It includes both good and bad network traffic data from realistic network environments as well as several attack categories. It is relatively large and has a variety of features making it appropriate for testing machine-learning methods for intrusion and traffic analysis.

Table 2. Characteristics of the CICIDS2017 Dataset

Parameter	Description
Dataset	CICIDS2017
Domain	Network Intrusion Detection
Number of Instances	2,830,743
Number of Features	78
Number of Classes	15
Target Attribute	Traffic Class
Training–Testing Split	80:20

b. Data Preprocessing

The data set is pre-processed in advance of model training to ensure its quality and consistency as input. The cleaning process eliminates redundant or missing information by duplicating entries and completing missing data. Numerical attributes are scaled so that the differences in the magnitude of the attributes do not significantly impact the model training. In the process of encoding, categorical features are given numerical values. The records are split into 80% Training and 20% Testing after the preprocessing. A dataset for training is used for creating the models while the dataset meant for testing is going to be used for checking performance.

c. Feature Engineering

Feature engineering is crucial in increasing the efficiency of predictions as well as helping avoid computational complexities. The extracted features that are obtained from the network traffic are analyzed to uncover useful features that can be used while writing off redundant features which may have a high correlation. The normalization of data as well as feature selection techniques are implemented to improve the representation of the data attributes and assure dimensionality reduction. This process is crucial in making machine learning models able to learn and improve their operations.

d. Machine Learning Model Development

Five supervised-learning models were implemented and evaluated using identical experimental conditions. Logistic Regression (LR) was included because it offers relatively low computational complexity and straightforward interpretation. The Decision Tree (DT) model represents predictions through a hierarchical sequence of decision rules, making its outputs comparatively easy to interpret. Random Forest is an ensemble method that makes predictions by aggregating the output of multiple decision trees. Support Vector Machine (SVM) defines a separating decision boundary as the one with the largest margin between classes. XGBoost is a gradient-boosting method, where future learners are added together to enhance the predictive power.

Table 3. Machine Learning Algorithms Considered

Algorithm	Learning Type	Major Advantage
Logistic Regression	Linear Classification	Fast training and high interpretability
Decision Tree	Tree-based Learning	Simple rule-based classification
Random Forest	Ensemble Learning	Improved robustness and prediction accuracy
Support Vector Machine	Margin-based Classification	Effective for high-dimensional datasets
XGBoost	Gradient Boosting	High predictive accuracy and scalability

e. Performance Evaluation

Standard classification measures are used to compare model performance. Accuracy is the overall percentage of correct predictions, whilst precision is the percentage of the positive predicted instances that are positive. Recall is a measure of the model's ability to find relevant positive examples, and F1 score is a combined measure of precision and recall. Training time is also a measure of computational efficiency and scalability.

Experimental Setup

The proposed approach was implemented in Python 3.11 with the help of the Scikit-learn library. The data preprocessing process was carried out by using Pandas and NumPy modules; also, Matplotlib was used for visualizing the information collected. The five selected supervised-learning models were tested on the CICIDS2017 benchmark data set. Records and incomplete records were duplicated and removed, numerical features were normalized and categorical features were encoded before training. The data was divided as 80/20 in order to train and test the data. To make the comparison between

models developed as fair as possible, they were pre-processed in the same way, and the experiments were run under identical conditions.

Table 4. Experimental Configuration

Parameter	Specification
Dataset	CICIDS2017
Programming Language	Python 3.11
ML Library	Scikit-learn
Data Processing	Pandas, NumPy
Training–Testing Split	80:20
Validation	10-Fold Cross Validation
Evaluation Metrics	Accuracy, Precision, Recall, F1-score, Training Time
Compared Models	LR, DT, RF, SVM, XGBoost

All experiments were performed on an Intel Core i7 system with 16 GB RAM, providing a consistent computing environment for evaluating model performance and scalability.

5. RESULTS AND DISCUSSION

The selected machine-learning models were evaluated on the CICIDS2017 dataset under identical experimental conditions. Logistic Regression, Decision Tree, Random Forest, Support Vector Machine, and XGBoost were compared using accuracy, precision, recall, F1-score, and training time. The results indicate that ensemble-based approaches provide strong performance for predictive analytics.

Table 5. Performance Comparison of Machine Learning Models

Algorithm	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	Training Time (s)
Logistic Regression	94.12	93.84	93.58	93.71	38.4
Decision Tree	96.35	96.10	95.92	96.01	24.8
Random Forest	98.42	98.25	98.18	98.21	82.6
Support Vector Machine	97.16	96.88	96.71	96.79	145.3
XGBoost	99.08	98.97	98.91	98.94	67.9

Among the evaluated approaches, XGBoost achieved the highest accuracy at 99.08%, followed by Random Forest at 98.42%. The two ensemble methods produced stronger predictive results than Logistic Regression and Decision Tree on the evaluated dataset. Although Support Vector Machine performed well in terms of classification efficiency, it took much longer than other techniques because of its complexity.

In terms of time taken to develop models, Logistic Regression took the minimum amount of time; however, it performed poorly in terms of classification efficiency due to its linearity. Decision Tree managed to do a

much better job than Logistic Regression and still retained the tendency toward lower complexity; however, it had more propensity to overfitting than the ensemble learning algorithms.

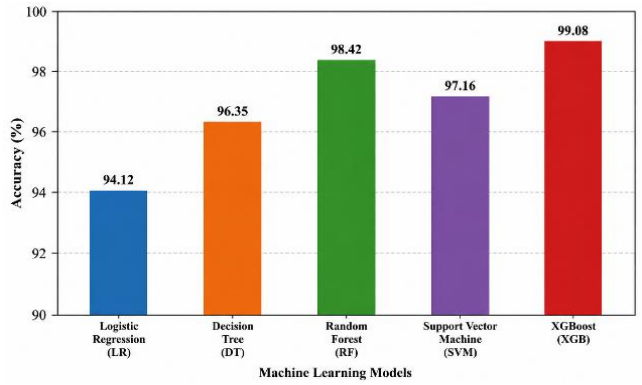

Figure 2. Accuracy Comparison of Machine Learning Models

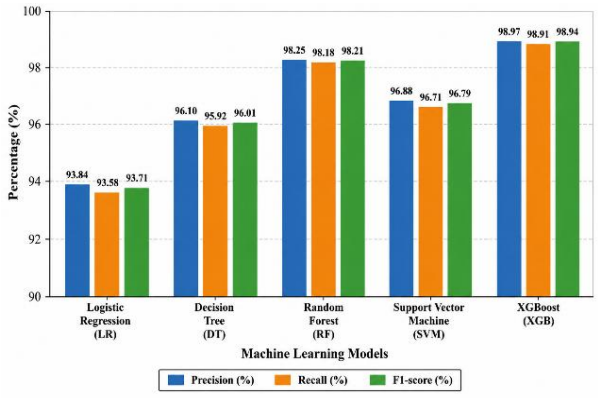

Figure 3. Precision, Recall, and F1-score Comparison

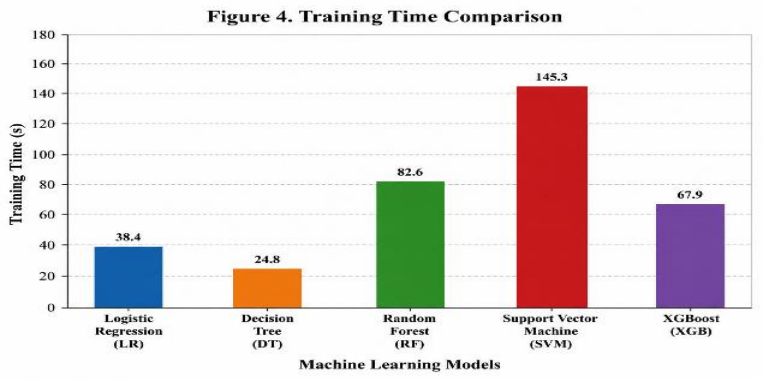

Figure 4. Training Time Comparison

Based on the experimental comparison, it is concluded that the ensemble-based methods are suitable for large scale predictive analysis. XGBoost outperformed the other models in terms of all the metrics reported here and had a fairly low training cost. The other random forest also showed good predictive ability and resistance to overfitting. The Logistic Regression and Decision Tree models were the other models that took less computational time, but their predictive power was not as high as that of the ensemble models. SVM had good classification accuracy but it is relatively time consuming which may be a problem if the data is very large.

Overall, the experiments conducted demonstrate the effectiveness of XGBoost as the algorithm for processing and predictive analysis, offering reliable performance in predictive quality, risk resistance, and scalability.

Conclusion

In this study, large-scale predictive analysis was evaluated and compared for supervised machine-learning techniques with the CICIDS2017 benchmark dataset. The proposed workflow involved data preprocessing, feature engineering, model development, and performance evaluation. Five algorithms were tested—Logistic Regression, Decision Tree, Random Forest, Support Vector Machine and XGBoost—with accuracy, precision, recall, F1 score and training time. It can be seen from the experimental results that ensemble based methods yield better predictive performance for the data set considered. The XGBoost model performed best in terms

of accuracy in classification and F1 score, while at the same time having an acceptable computational overhead. The results underscore the need to carefully choose machine-learning algorithms, not only for their predictive power, but also for their computational efficiency. The results can be used to create data-driven applications in various fields like Cyber Security, Health Care, Recommendation Systems, and Business Analytics. Future research would explore deep-learning and explainable-AI methodologies to enhance predictive accuracy, model interpretability and scale to large data and real-time analytics.

References

1. L. Rodrigues and S. N. Givigi, "Predictive Analytics: An Optimization Perspective," in *IEEE Access*, vol. 12, pp. 106983-106995, 2024, doi: 10.1109/ACCESS.2024.3434617.
2. XGBoost and Random Forest Algorithms: An In-Depth Analysis. (2023). *Pakistan Journal of Scientific Research*, *3*(1), 26-31. https://doi.org/10.57041/vol3iss1pp26-31
3. Comparative Study of Feature Engineering Techniques for Predictive Data Analytics. (2024). *Journal of Technology Informatics and Engineering*, *3*(2), 417-435. https://doi.org/10.51903/jtie.v3i2.225
4. P. Suresh Dahake and N. S. Dahake, "Predicting Buyer Behaviour: A Reconnaissance of Retail Predictive Analytics Model," *2024 Intelligent Systems and Machine Learning Conference (ISML)*, Hyderabad, India, 2024, pp. 238-245, doi: 10.1109/ISML60050.2024.11007411.
5. R. Satheeskumar, E. Vandana Dutt, C. S. Reddy, Z. Gupta, S. Natarajan and M. L B, "Machine Learning Models for Predictive Marketing Analytics," *2024 Second International Conference Computational and Characterization Techniques in Engineering & Sciences (IC3TES)*, Lucknow, India, 2024, pp. 1-5, doi: 10.1109/IC3TES62412.2024.10877468.
6. S. Singh, S. Dey, S. Sharma and B. P. Prathaban, "Predictive Analytics in Hotel Recommendations Using Machine Learning," *2024 International Conference on IoT Based Control Networks and Intelligent Systems (ICICNIS)*, Bengaluru, India, 2024, pp. 1072-1076, doi: 10.1109/ICICNIS64247.2024.10823282.
7. P. S. Ingle and S. Deshpande, "Predictive Analytics of Cardiovascular Disease Using Machine Learning," *2024 IEEE Pune Section International Conference (PuneCon)*, Pune, India, 2024, pp. 1-8, doi: 10.1109/PuneCon63413.2024.10895205.
8. G. N. Rajareddy, G. Lakshmeeswari and P. P. Pradhan, "Predictive Analytics in Crime: A Machine Learning Approach," *2024 OITS International Conference on Information Technology (OCIT)*, Vijayawada, India, 2024, pp. 223-227, doi: 10.1109/OCIT65031.2024.00047.
9. B. Charan, D. Jaswanth, E. Hemanth and M. S. Naidu, "Machine Learning and Deep Learning Approaches for Healthcare Predictive Analytics," *2024 5th International Conference on Electronics and Sustainable Communication Systems (ICESC)*, Coimbatore, India, 2024, pp. 1698-1707, doi: 10.1109/ICESC60852.2024.10689833.
10. S. D. K. K, A. R, N. G, R. V. K and D. K. V, "A Predictive Analytics in Cardiology: Evaluating Machine Learning Algorithms," *2025 International Conference on Computing and Communication Technologies (ICCCT)*, Chennai, India, 2025, pp. 1-6, doi: 10.1109/ICCCT63501.2025.11019319.