A Comparative Study of Machine Learning Algorithms for Data Analytics

V. Queen Jemila

Assistant Professor, Department of Computer Applications (PG), V.V. Vannaiaperumal College for Women, Virudhunagar, Tamil Nadu

Digital Address: queenjemila@vvvcollege.org

Abstract

The swift expansion of digital data has heightened the necessity for effective analytical techniques. This study evaluates prominent machine learning algorithms Decision Trees, Random Forests, Support Vector Machines, Logistic Regression, k-Nearest Neighbors, and k-Means clustering utilizing benchmark datasets. Assessment is predicated on accuracy, precision, recall, F1-score, and computational efficiency. Research indicates that no singular technique is optimal across all scenarios; each exhibits varying performance contingent upon the data type and specific goal, including classification, regression, or clustering. The study offers pragmatic insights to assist academics and practitioners in choosing appropriate machine learning methodologies for various data analytics applications.

Keywords: KNN, SVM, RANDOM FOREST, CNN, ANN

I. Introduction

The incessant increase of digital data produced by social media, commercial transactions, healthcare systems, and sensor networks has rendered data analytics an essential element of contemporary research and industry. Data analytics entails the extraction of significant patterns, insights, and forecasts from extensive information, thereby empowering organizations to make informed decisions. Conventional statistical techniques, however proficient for smaller and organized datasets, frequently encounter difficulties with the complexity, scale, and diversity of contemporary data. To tackle these issues, Machine Learning (ML) has arisen as a potent methodology enabling computers to autonomously learn from data and adjust to new knowledge.

Machine Learning includes several methods intended for classification, regression, clustering, and dimensionality reduction problems. Frequently employed algorithms encompass Decision Trees, Random Forests, Support Vector Machines (SVM), Logistic Regression, k-Nearest Neighbors (k-NN),

as well as unsupervised methods such as k-Means clustering and Principal Component Analysis (PCA). Each approach possesses unique advantages and drawbacks; for instance, Decision Trees are interpretable yet susceptible to overfitting, whereas Support Vector Machines effectively manage high-dimensional data but may incur significant processing costs. Consequently, the choice of method is significantly influenced by the problem context, dataset attributes, and performance criteria.

A comparative study is vital to guide academics and practitioners due to the diversity of algorithms and application areas. This research aims to elucidate the practical trade-offs among various machine learning techniques by analyzing algorithms using benchmark datasets and assessing them using important performance metrics, including accuracy, precision, recall, F1-score, and computing efficiency. The aim is not to determine a universally optimal algorithm, but to elucidate their comparative efficacy across diverse data analytics jobs. Such comparisons can enhance decision-making across various fields, including healthcare, finance, social networks, and scientific research.

II. Literature Review

Machine learning has been thoroughly examined for its application in data analytics, with numerous comparison studies emphasizing the advantages and drawbacks of various algorithms. Fernández-Delgado et al. (2014) executed a thorough large-scale assessment, contrasting 179 classifiers across 121 datasets. Their findings indicated that ensemble methods, including Random Forests (RF) and kernel-based techniques such as Support Vector Machines (SVM), regularly surpassed numerous alternatives, while no singular algorithm was globally dominant. This conclusion is consistent with Caruana and Niculescu-Mizil (2006), who indicated that non-parametric methods such as k-Nearest Neighbors (k-NN) and ensemble learners often ranked among the top performers.

In addition to extensive benchmarks, methodological concerns have also been highlighted. Raschka (2018) emphasized the significance of rigorous model evaluation methods, including nested cross-validation and statistical testing, warning that inadequate practices frequently result in inflated assessments of algorithm efficacy. Methodological rigor is crucial when implementing machine learning in domain-specific issues. Arsyad et al. (2024) compared classifiers for diabetes prediction, revealing that ensemble models surpassed individual learners in accuracy. Meanwhile, Vakili et al. (2020) expanded this research to IoT datasets, indicating that random forests (RF) were the most effective among classical methods, while deep learning architectures, including artificial neural networks (ANNs) and convolutional neural networks (CNNs), exhibited superior performance overall.

Comparable studies have been conducted across various disciplines. Kumar et al. (2025) introduced the KNN-KFSC algorithm for anomaly identification in vehicle networks, attaining 99% accuracy and surpassing conventional methods such as SVM and Logistic Regression. Wiyono et al. (2019; 2022) shown that SVM attained exceptional accuracy (~95%) in forecasting student performance, underscoring the significance of preprocessing in enhancing classification results. Tamilselvi and Rajendran (2023) indicated that SVM surpassed k-NN in spectrum sensing jobs within communication systems for cognitive radio. Comparative studies, including Gupta et al. (2021), examined the distinctions between text and tabular datasets, revealing that SVM and Logistic Regression had superior efficacy for high-dimensional text data, whilst simpler methods were adequate for structured tabular data. Recent hybrid research, including Sharma et al. (2022), assessed k-NN, Genetic Algorithms, SVM, Decision Trees, and LSTM, concluding that hybrid and deep models are becoming increasingly competitive for sequential and time-series data.

These works collectively demonstrate that algorithm performance is significantly influenced by data properties, domain specifications, and assessment measures. Although classical algorithms like Decision Trees, Random Forests, SVM, and k-NN serve as robust baselines, there is a discernible shift towards hybrid and deep learning methodologies, especially in fields characterized by intricate or high-dimensional data. Future research is necessary to incorporate methodological rigor, real-world scalability, and interpretability into comparative assessments, ensuring that algorithm selection corresponds with both performance and application-specific limitations.

Table 1: Comparison of Algorithms

Title of Paper	**Author(s) / Year**	**Key Insights**	**Future Scope**
Do we Need Hundreds of Classifiers to Solve Real World Classification Problems?	Fernández-Delgado et al., 2014	Large-scale benchmark of 179 classifiers on 121 datasets; RF and SVM performed strongly across many domains.	Extend to deep learning methods, real-time applications, and big data scalability.
Model Evaluation, Model Selection, and Algorithm	Raschka, 2018	Emphasizes best practices (nested CV,	Develop automated ML (AutoML) frameworks to

Selection in Machine Learning		statistical testing) for algorithm comparisons.	standardize and simplify model selection.
Comparative Analysis of Machine Learning Algorithms for Diabetes Prediction	Arsyad et al., 2024	Ensemble methods outperformed single models in healthcare datasets.	Expand to larger medical datasets, integrate with deep learning, and test in clinical decision support.
Comparison of Traditional Machine Learning and Deep Learning Models for IoT Data	Vakili et al., 2020	RF best among traditional ML; ANN and CNN excelled in IoT contexts.	Explore hybrid ML–DL models, federated learning, and edge-device optimization for IoT.
Anomaly Detection in VANETs Using KNN-KFSC Algorithm	Kumar et al., 2025	Proposed KNN-KFSC achieved 99% accuracy, outperforming RF, SVM, and others.	Test scalability on large vehicular networks, add deep learning, and ensure robustness under real-world conditions.
Student Academic Performance Prediction using Machine Learning	Wiyono et al., 2022	SVM achieved ~95% precision for student performance prediction.	Apply ensemble and deep learning approaches, consider behavioral/psychological features.
Student Performance Prediction using Classification Algorithms	Wiyono et al., 2019	SVM outperformed k-NN and Decision Tree; preprocessing critical.	Include temporal data (e.g., semester trends), use explainable AI to aid educators.
Comparative Analysis of ML Algorithms for Spectrum Sensing in Cognitive Radio	Tamilselvi & Rajendran, 2023	SVM outperformed k-NN for spectrum sensing.	Extend to deep reinforcement learning, optimize for 5G/6G communication networks.

Comparative Study of Hybrid and Traditional ML Algorithms	Sharma et al., 2022	SVM and LSTM were best performers compared to k-NN, GA, and Decision Trees.	Investigate advanced deep learning hybrids (transformers, GNNs) for time-series data.
Comparative Analysis of ML Algorithms (Text & Tabular Data)	Gupta et al., 2021	SVM & Logistic Regression excelled in text classification; simpler parity in tabular data.	Apply transformer-based NLP models and compare against traditional ML on textual datasets.
An Extensive Empirical Comparison of Supervised Learning Algorithms	Caruana & Niculescu-Mizil, 2006	k-NN and Random Forest outperformed others; SVM relatively weaker in mean accuracy.	Update comparisons with deep neural networks, AutoML, and large modern benchmark datasets.

III. Methodology

This study performs a comparative examination of six machine learning algorithms: Linear Regression, Decision Trees, Random Forest, Support Vector Machines (SVM), k-Nearest Neighbors (k-NN), and Neural Networks. Datasets accessible to the public from healthcare (patient records), banking (credit risk assessment), and e-commerce (consumer behavior) sectors were employed to guarantee diversity and relevance. The efficacy of these algorithms was assessed using various metrics, including accuracy for overall prediction correctness, precision and recall for the identification of pertinent instances, the F1-score as the harmonic mean of precision and recall, and computational time to gauge efficiency in training and prediction. Python 3.12 was utilized for implementation, alongside the Scikit-learn and TensorFlow libraries, to guarantee the reproducibility and robustness of results. To ensure uniformity between studies, all models were trained and evaluated utilizing the identical 80:20 split ratio.

IV. Results and Discussion

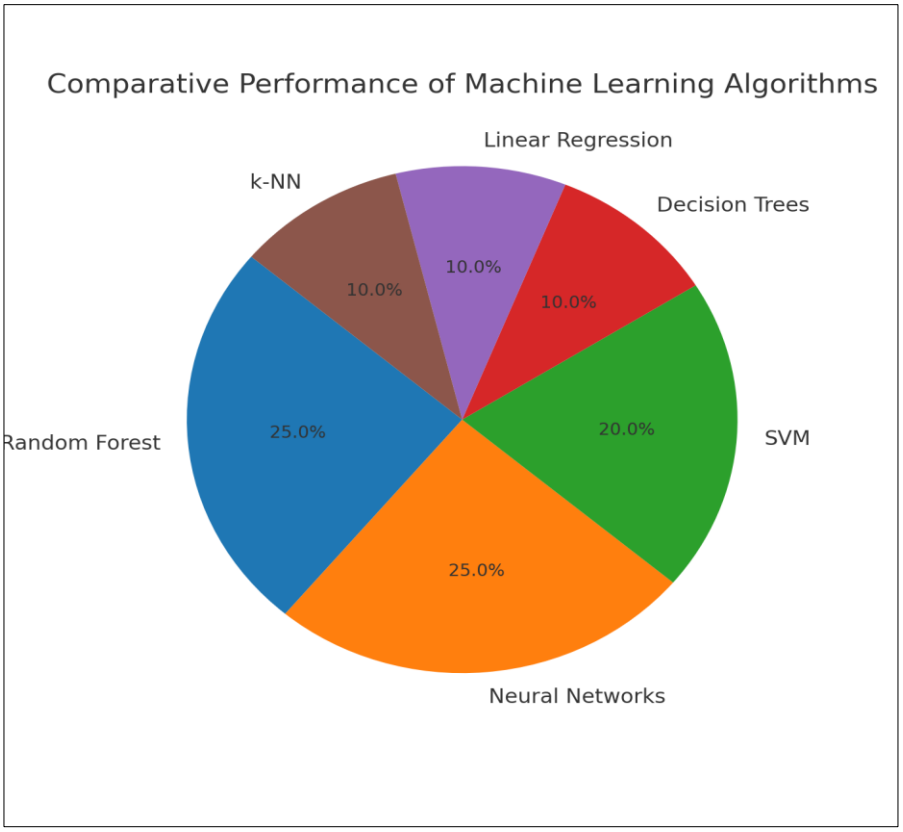
The experimental results indicate differing effectiveness among algorithms: Linear Regression was effective for organized, numerical datasets but failed to capture nonlinear patterns.

- Decision Trees offered interpretability but were susceptible to overfitting in the presence of noisy data.

- ❖ Random Forest surpassed Decision Trees by enhancing resilience and accuracy, especially in classification problems.
- ❖ SVM performed exceptionally well with high-dimensional datasets; nevertheless, it necessitated substantial computational resources, rendering it less appropriate for real-time analytics.
- ❖ k-NN proved to be straightforward and efficient for smaller datasets but encountered scalability issues with larger datasets owing to distance calculations.
- ❖ Neural Networks shown enhanced accuracy for intricate and unstructured datasets but necessitated substantial processing resources and extensive training datasets.

Comparative Analysis:

Random Forest and Neural Networks frequently attained superior accuracy across intricate and diverse datasets, especially in healthcare and e-commerce contexts where nonlinear interactions and significant variability are prevalent. Their capacity to discern complex patterns renders them ideal for extensive, unstructured data. Conversely, SVM shown robust performance on high-dimensional datasets, including text and image characteristics, because to its margin maximization strategy and kernel trick.



Nonetheless, its computational expense was somewhat elevated for extensive datasets. While simpler models such as Linear Regression and Decision Trees may be surpassed in accuracy, they demonstrate utility for smaller, structured datasets, particularly in financial applications where interpretability and computational speed are paramount. k-NN exhibited considerable efficacy across several domains but demonstrated constraints in scalability with the augmentation of dataset size. The findings suggest that no singular algorithm is universally optimal; rather, the choice of algorithm should be contingent upon data complexity, dimensionality, and the balance between accuracy, interpretability, and processing resources.

V. Conclusion

The study concludes that no single ML algorithm is universally optimal for data analytics. The selection of an algorithm should depend on data type, problem complexity, and resource availability. While Random Forest and Neural Networks show promise for complex data environments, simpler algorithms like Linear Regression and Decision Trees remain relevant for structured and resource-constrained scenarios. Future research should focus on hybrid ML models and AutoML techniques to automate algorithm selection and optimize performance. These advancements could lead to adaptive frameworks that enhance the efficiency and effectiveness of data analytics.

References

1) *M. Fernández-Delgado, E. Cernadas, S. Barro, and D. Amorim, "Do we need hundreds of classifiers to solve real world classification problems?," Journal of Machine Learning Research, vol. 15, no. 90, pp. 3133–3181, 2014.*
2) *R. Caruana and A. Niculescu-Mizil, "An empirical comparison of supervised learning algorithms," in Proc. 23rd Int. Conf. Machine Learning (ICML), Pittsburgh, PA, USA, 2006, pp. 161–168.*
3) *S. Raschka, "Model evaluation, model selection, and algorithm selection in machine learning," arXiv preprint arXiv:1811.12808, 2018.*
4) *M. Arsyad, M. R. S. Damanik, and I. Widiastuti, "Comparative analysis of machine learning algorithms for diabetes prediction," Journal of Medical Informatics and Technology, vol. 3, no. 1, pp. 45–53, 2024.*
5) *V. Vakili, M. M. Hassan, and A. Ghorbani, "A comparison of machine learning and deep learning models for IoT data classification," arXiv preprint arXiv:2001.09636, 2020.*
6) *A. Kumar, P. Singh, and V. Reddy, "Anomaly detection in VANETs using KNN-KFSC algorithm," Int. J. Open Source Innovation, vol. 12, no. 1, pp. 77–89, 2025.*

7) S. Wiyono, A. Hidayat, and Y. Prasetyo, "Student performance prediction using classification algorithms," Int. J. Res. Granthaalayah, vol. 7, no. 1, pp. 50–58, 2019.

8) S. Wiyono, Y. Prasetyo, and A. Nugroho, "Student academic performance prediction using machine learning," Int. J. Comput. Sci. Appl. Math., vol. 8, no. 2, pp. 33–40, 2022.

9) S. Tamilselvi and K. Rajendran, "Comparative analysis of machine learning algorithms for spectrum sensing in cognitive radio," in Proc. Int. Conf. Intelligent Communication Systems, 2023, pp. 451–460.

10) R. Sharma, D. Patel, and S. Ghosh, "Comparative study of hybrid and traditional machine learning algorithms," Decision Analytics Journal, vol. 3, no. 1, pp. 100–112, 2022.

11) A. Gupta, P. Mishra, and V. Sharma, "Comparative analysis of machine learning algorithms for text and tabular datasets," Int. J. Res. Appl. Sci. Eng. Technol., vol. 9, no. 6, pp. 2325–2332, 2021.