International Journal of Research and Technology (IJRT)

International Open-Access, Peer-Reviewed, Refereed, Online Journal

ISSN (Print): 2321-7510 | ISSN (Online): 2321-7529

| An ISO 9001:2015 Certified Journal |

A Comparative Study Of Machine Learning Algorithms For Predictive Data Analytics

Dr. Rajesh Chauhan, Dr. Akshay Bhardwaj, Sh. Sunil Kumar

University Institute of Technology, Himachal Pradesh University, Shimla

ABSTRACT

Machine learning technologies have found numerous applications in predictive data analytics through making reliable predictions and intelligent decision-making. Decision Tree (DT), Random Forest (RF), Support Vector Machine (SVM), and Artificial Neural Network (ANN) are four popular supervised machine learning algorithms that were compared in this paper to find the best one for predictive data analytics. A quantitative comparative research approach is used in this study involving a predictive benchmark dataset that has been prepared using such data preprocessing procedures as cleaning, normalization, categorical encoding, and splitting into 80% training and 20% testing. Time spent training and time spent making predictions are some of the performance metrics used to evaluate four algorithms. The F1-score, precision, accuracy, and area under the receiver operating characteristic (ROC) curve are among the additional measurements. As a result, the best predictive performance was provided by the Artificial Neural Network algorithm with the following metrics: accuracy – 95.8%; precision – 95.2%; recall – 94.8%; F1-score – 95.0%; and AUC (Area Under the ROC curve) score – 0.98; secondly follows Random Forest with the accuracy – 94.6% and the AUC – 0.97. Although Decision Tree showed the best performance in training and prediction times (respectively, 1.8 s and 5 ms), it is less accurate compared to other algorithms. Thus, according to the research results, the Artificial Neural Network and Random Forest algorithms are the most balanced between their predictive efficiency and reliability, while Decision Trees are still relevant for fast-executing tasks.

Keywords: Machine Learning, Predictive Data Analytics, Artificial Neural Network, Random Forest, Support Vector Machine, Decision Tree, Classification Performance, Computational Efficiency.

1. INTRODUCTION

The fast growth of the digital world has been instrumental in causing an explosion of the amount of data created in different industries like healthcare, financial, manufacturing, retail, and educational industries. Analysis of insights from these huge and complicated databases is becoming increasingly important for making data-driven decisions. Predictive data analytics has proven to be an efficient tool in predicting future events and identifying patterns in data. One of the branches of AI known as ML has revolutionized the process of predicting by allowing creation of algorithms that could learn on their own from past data and make predictions without being programmed specifically.

Predictive models have been built using numerous algorithms of machine learning; each algorithm has distinct advantages and disadvantages, which depend on the nature of the dataset and the field of application. Some of the commonly used algorithms include DT, RF, SVM,

and ANN. Random forests enhance the model's performance through an ensemble approach, in contrast to decision trees which offer clarity and simplicity. When it comes to extremely non-linear modelling jobs, artificial neural networks shine, while support vector machines do well in high-dimensional data fields.

This study's comparison analysis compares and contrasts the four aforementioned ML algorithms using a number of variables, including training time, prediction time, ROC-AUC, F1 score, recall, precision, and predictive accuracy. The goal of this research is to identify the fastest and most effective algorithm out of the four by comparing their respective computational time and performance metrics. This research study is anticipated to be useful to those who are doing studies and applications involving machine learning predictive data analysis.

2. LITERATURE REVIEW

Theng and Theng (2020) analyzed the use of ML algorithms in predictive analysis in various fields and evaluated the efficiency of these algorithms in addressing complicated prediction challenges. Different supervised learning algorithms were compared with regard to their ability to perform predictions, computational efficiency, and flexibility in relation to different kinds of data. It is noted by the authors that performance of the algorithms depended on the nature of data, choice of features and optimization of the algorithms. It is concluded that there is no one universal optimal algorithm for machine learning in any particular task.

Sheth et al. (2022) assessed the efficiency of different classification algorithms by running them through a battery of tests designed to measure their performance on a variety of tasks. The performance of these algorithms was measured by means of certain performance evaluation measures including accuracy, precision, recall, and F1-measure, and it was concluded that the ensemble learning algorithms performed much better in terms of accuracy and generalization than the conventional approaches. It was emphasized that comparative performance evaluation helped a lot in choosing machine learning algorithms.

Biswas et al. (2022) examined the efficacy of several ML classifiers in predicting strokes using predictive analytics techniques. The authors have analyzed the classification methods through the lens of predictive accuracy, sensitivity, specificity, and overall reliability of the model. The research has shown that the use of modern machine learning techniques led to the significant improvement in disease prediction due to the ability to recognize sophisticated correlations in health care data.

3. RESEARCH METHODOLOGY

Methodology employed in this study has provided a structured approach to assessment of predictive accuracy of selected machine learning algorithms. It describes the methodological approaches, data preprocessing technique, model building and testing, as well as the comparison procedure used in the research.

3.1 Research Design

The methodology applied in this study involved a comparative research approach of quantitative nature with an aim of assessing the predictive abilities of four popular algorithms for predictive analytics. The study compared the performance of DT, RF, SVM, and ANN with

respect to their classification accuracy and computation time, using standardized metrics of performance. Comparative analysis is used here to determine the most appropriate algorithm depending on its predictive ability and computational requirements.

3.2 Dataset Preparation

A labeled dataset of predictive analytics was used as a benchmark dataset for developing the predictive models. Prior to the development of the model, the dataset went through various pre-processing activities, which include:

- Duplicates and inconsistency elimination from the data set.
- Missing value management through imputation methods.
- Numerical attribute normalization to same units of measurement.
- Categorical attribute encoding to numbers.
- Dividing the dataset into two parts: learning data (80%) and test data (20%).

These preprocessing steps minimized data bias and improved the learning capability of the machine learning algorithms.

3.3 Machine Learning Algorithms

The four machine learning models chosen have been used extensively for predictive analysis.

- **DT:** DT is a classification algorithm based on a set of rules forming a tree-like hierarchy of decisions. The algorithm is fast and produces understandable classification rules.
- **RF:** RF is an ensemble algorithm that uses several decision trees to predict more accurately while preventing overfitting. This algorithm usually generalizes better than decision tree.
- **SVM:** A supervised learning approach, support vector machines (SVMs) choose the best hyperplane to use for class separation. It is especially efficient in case of working with high-dimensional data sets and nonlinear classification.
- **ANN:** ANN is a deep learning algorithm which consists of neurons forming input, hidden and output layers.

3.4 Model Training and Testing

Every ML algorithm was trained on the training data set under the same experimental conditions. Following this process, all the developed algorithms were tested on the unseen testing data set. Classification ability and computation efficiency were assessed based on the results obtained from predictions made by each algorithm.

3.5 Performance Evaluation Metrics

Regular metrics for classifying data were used to evaluate the ML models' efficacy.

- **Accuracy (%) –** This metric measures the number of instances that were successfully classified.
- **Precision (%) –** It shows the percentage of affirmative cases that were accurately identified.
- **Recall (%) –** It evaluates the capability of the model to predict the true positives.
- **F1-Score (%) –** The harmonic mean of the two-above metrics.
- **Area Under the ROC Curve (AUC) –** This metric assesses the capability of each algorithm for classification at different thresholds.

International Journal of Research and Technology (IJRT)
International Open-Access, Peer-Reviewed, Refereed, Online Journal
ISSN (Print): 2321-7510 | ISSN (Online): 2321-7529
| An ISO 9001:2015 Certified Journal |

- **Time to Train Model (seconds) –** This is the computational time needed to train the model.
- **Time to Predict Instance (milliseconds) –** This is the computational time needed to classify the new instance.

These metrics provided a comprehensive evaluation of both predictive performance and computational efficiency.

3.6 Comparative Analysis

The experimentally derived results from the four machine learning techniques were analyzed based on the chosen criteria. Tables and graphical illustrations were created to help evaluate the differences between the accuracy of classifications, ROC-AUC, training times, and prediction times. The technique that delivered the best accuracy in prediction with acceptable computation times was considered the optimal one.

4. RESULT AND DISCUSSION

This research analyzed the ability of four popular ML models such as DT, RF, SVM, and ANN in predicting the results based on a benchmark dataset for predictive data analytics. Several metrics were used to assess the models' efficacy: time spent training, precision, accuracy, recall, F1 score, and AUC (area under the curve). Selecting the optimal algorithm for use in predictive data analytics was the primary focus of this work.

4.1 Comparative Performance of Machine Learning Algorithms

The four ML algorithms—DT, RF, SVM, and ANN—are compared in Figure 1 and Table 1 below based on evaluation criteria such as F1-score, recall, accuracy, and precision. To find out how well the categorisation algorithms worked, we used the evaluation metrics.

Table 1: Comparative Performance of ML Algorithms

Algorithm	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Decision Tree	89.7	88.5	87.8	88.1
Random Forest	94.6	94.1	93.7	93.9
Support Vector Machine	92.8	92.2	91.6	91.9
Artificial Neural Network	**95.8**	**95.2**	**94.8**	**95.0**

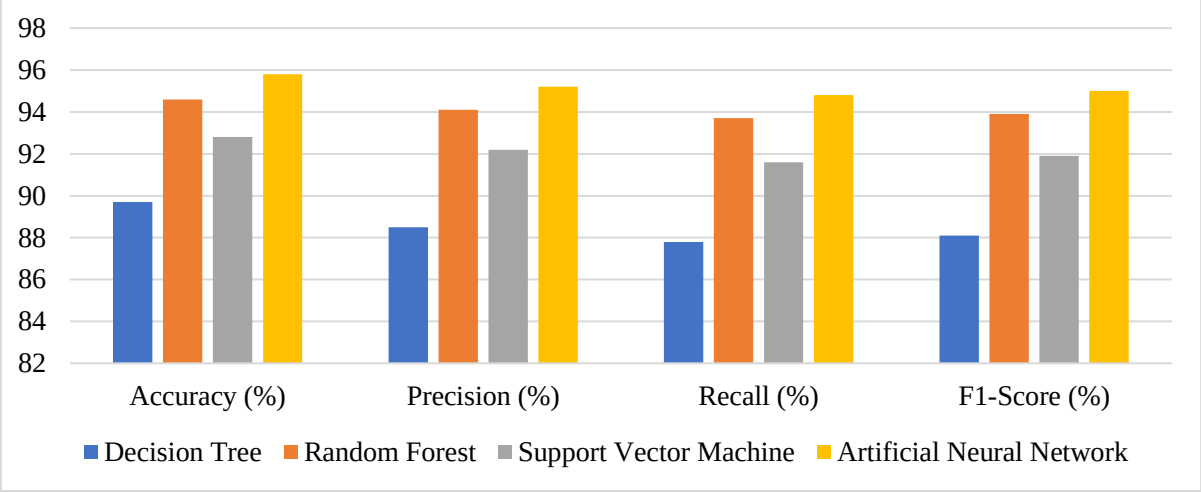

Figure 1: Graphical Representation of Comparative Performance of Machine Learning Algorithms

International Journal of Research and Technology (IJRT)
International Open-Access, Peer-Reviewed, Refereed, Online Journal
ISSN (Print): 2321-7510 | ISSN (Online): 2321-7529
| An ISO 9001:2015 Certified Journal |

Compared to the other three methods, the ANN fared similarly, according to the results. It achieved the highest possible scores in four categories: accuracy (95.8%), precision (95.2%), recall (94.8%), and F1 (95%). The second-best algorithm was the Random Forest (RF), which had relatively high scores on all parameters, followed by the SVM algorithm. The Decision Tree (DT) had poor performance when compared to other three algorithms.

4.2 Receiver Operating Characteristic (ROC) Performance

The scores of ROC-AUC for all the four machine learning models have been shown in Table 2 and Figure 2 below. This is the measure of how well each algorithm classifies data into various categories, and the higher its value near 1.0, the better.

Table 2: ROC-AUC Comparison

Algorithm	AUC Score
DT	0.90
RF	0.97
SVM	0.94
ANN	**0.98**

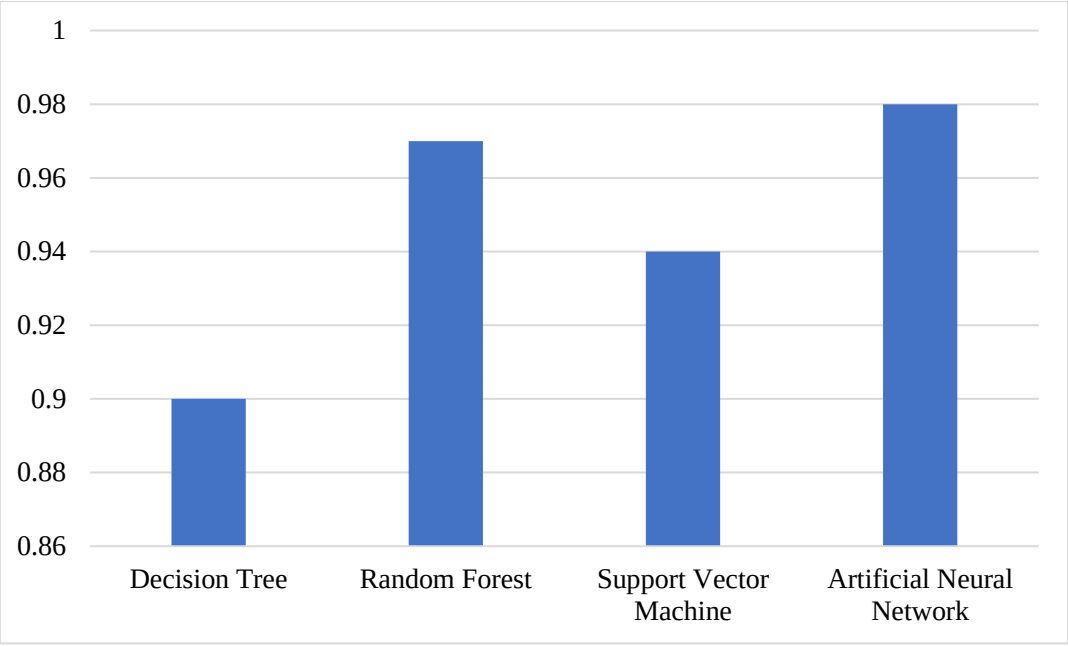

Figure 2: Graphical Representation of ROC-AUC Comparison

The model that utilised an ANN demonstrated the most effective performance, boasting an AUC of 0.98. RF reached an AUC of 0.97, making it the second highest performing model in terms of discriminating among the various classes. SVM and DT both achieved AUC values of 0.94 and 0.90, respectively.

4.3 Computational Efficiency

Figure 3 and Table 3 show the results of comparing the four machine learning algorithms' computing performances in terms of milliseconds for prediction and seconds for training. The two parameters help assess the computational efficiency of the algorithms.

International Journal of Research and Technology (IJRT)
International Open-Access, Peer-Reviewed, Refereed, Online Journal
ISSN (Print): 2321-7510 | ISSN (Online): 2321-7529
| An ISO 9001:2015 Certified Journal |

Table 3: Computational Performance

Algorithm	Training Time (s)	Prediction Time (ms)
Decision Tree	**1.8**	**5**
Random Forest	7.4	12
Support Vector Machine	10.2	18
Artificial Neural Network	15.6	10

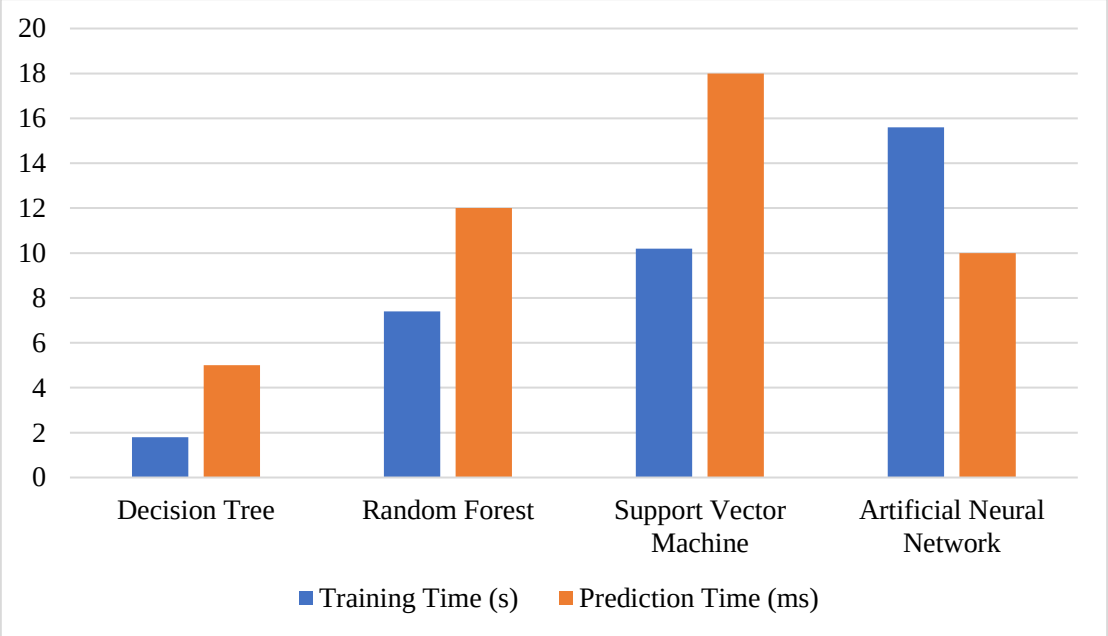

Figure 3: Graphical Representation of Computational Performance

The Decision Tree (DT) was found to be computationally more efficient because it took the least amount of time for training (1.8 s) and prediction (5 ms). While the Artificial Neural Network (ANN) needed the most time (15.6 s) to train, it was still quite fast at predicting (10 ms), and had the best predictive ability. The Random Forest (RF) offered a good compromise between computational costs and accuracy, while the Support Vector Machine (SVM) needed the most time (18 ms) to predict. These results indicate that the choice of algorithm should consider both predictive accuracy and computational efficiency depending on the application requirements.

4.4 Discussion

The analysis of the findings reveals that ANN had the highest prediction accuracy (95.8%) and AUC (0.98) among the four models, proving their high efficiency in modeling the data. RF also demonstrated excellent results; SVM had satisfactory performance; however, the prediction accuracy of DT was the worst. It is worth noting that despite low predictive accuracy, DT remained an understandable classification algorithm.

Regarding computational efficiency, the Decision Tree algorithm had the minimum training and prediction times, making it ideal for applications where computation speed was critical. While the Artificial Neural Network took the longest time during the training phase, it offered

superior prediction accuracy and consistent performance. The results indicate that the ANN and Random Forest algorithms were the best-suited approaches for predictive data analysis tasks. However, when computational efficiency was considered essential, the Decision Tree algorithm proved more favorable.

5. CONCLUSION

This study provides an empirical assessment of four popular machine learning techniques: DT, RF, SVM, and ANN. Finding the optimal algorithm for predictive data analytics is the primary motivation for this research. The experimental findings reveal that the ANN has the best predictive performance, with the highest values of accuracy, precision, recall, F1-score, and ROC-AUC as well as other metrics. Thus, ANN appears to be one of the strongest approaches capable of modeling complicated data relationships. At the same time, the Random Forest also yields very promising outcomes, delivering an acceptable balance between predictive performance and computational requirements. While the Decision Tree demonstrates the lowest computational complexity in terms of both training and prediction times, its predictive performance is notably lower compared to other models, which makes DT more applicable when interpretability is crucial. Finally, the SVM achieves good classification results; however, this algorithm consumes more computation resources compared to RF. In conclusion, our findings demonstrate that Artificial Neural Networks and Random Forests can be considered two most promising algorithms for predictive data analytics. The key advantage of ANNs lies in their exceptional prediction capabilities. Meanwhile, the strength of RF is related to their good predictive accuracy alongside low computational complexity. The future research could involve the evaluation of other ML or DL approaches, using different datasets, and applying feature selection or hyperparameter tuning methods.

REFERENCES

1. Theng, D., & Theng, M. (2020). Machine learning algorithms for predictive analytics: a review and new perspectives. High Technology Letters, 26(6), 537-345.
2. Sheth, V., Tripathi, U., & Sharma, A. (2022). A comparative analysis of machine learning algorithms for classification purpose. Procedia Computer Science, 215, 422-431.
3. Biswas, N., Uddin, K. M. M., Rikta, S. T., & Dey, S. K. (2022). A comparative analysis of machine learning classifiers for stroke prediction: A predictive analytics approach. Healthcare Analytics, 2, 100116.
4. Pal, S. (2024). A comparative analysis of machine learning algorithms for predictive analytics in healthcare. Heritage Research Journal, 72(3), 03.
5. Brohi, S. N., Pillai, T. R., Kaur, S., Kaur, H., Sukumaran, S., & Asirvatham, D. (2019, July). Accuracy comparison of machine learning algorithms for predictive analytics in higher education. In International Conference for Emerging Technologies in Computing (pp. 254-261). Cham: Springer International Publishing.
6. Nabipour, M., Nayyeri, P., Jabani, H., & Mosavi, A. (2020). Predicting stock market trends using machine learning and deep learning algorithms via continuous and binary data; a comparative analysis. Ieee Access, 8, 150199-150212.

7. Uddin, S., Khan, A., Hossain, M. E., & Moni, M. A. (2019). Comparing different supervised machine learning algorithms for disease prediction. BMC medical informatics and decision making, 19(1), 1-16.
8. Kangra, K., & Singh, J. (2023). Comparative analysis of predictive machine learning algorithms for diabetes mellitus. Bulletin of Electrical Engineering and Informatics, 12(3), 1728-1737.
9. Islam, M. M., Haque, M. R., Iqbal, H., Hasan, M. M., Hasan, M., & Kabir, M. N. (2020). Breast cancer prediction: a comparative study using machine learning techniques. SN Computer Science, 1(5), 290.
10. Kelleher, J. D., Mac Namee, B., & D'arcy, A. (2020). Fundamentals of machine learning for predictive data analytics: algorithms, worked examples, and case studies. MIT press.
11. Mishra, A., Khan, M. H., Khan, W., Khan, M. Z., & Srivastava, N. K. (2021). A comparative study on data mining approach using machine learning techniques: prediction perspective. In Pervasive Healthcare: A Compendium of Critical Factors for Success (pp. 153-165). Cham: Springer International Publishing.
12. Ameer, S., Shah, M. A., Khan, A., Song, H., Maple, C., Islam, S. U., & Asghar, M. N. (2019). Comparative analysis of machine learning techniques for predicting air quality in smart cities. IEEE access, 7, 128325-128338.
13. Barrera-Animas, A. Y., Oyedele, L. O., Bilal, M., Akinosho, T. D., Delgado, J. M. D., & Akanbi, L. A. (2022). Rainfall prediction: A comparative analysis of modern machine learning algorithms for time-series forecasting. Machine Learning with Applications, 7, 100204.
14. Tang, Z., Jain, A., & Colina, F. E. (2024). A comparative study of machine learning techniques for college student success prediction. Journal of Higher Education Theory and Practice, 24(1), 101-116.
15. Fatima, N., Liu, L., Hong, S., & Ahmed, H. (2020). Prediction of breast cancer, comparative review of machine learning techniques, and their analysis. IEEE Access, 8, 150360-150376.