ELSEVIER

JOURNAL ON COMMUNICATIONS

ISSN:1000-436X

COMPARATIVE STUDY OF SUPERVISED LEARNING ALGORITHMS

1Suneela Bhoompally[1] , Dr Shesikala Martha[2]

[1]Research Scholar,Department ofECE,SR University,Warangal

[2]Professor,Department of Computer Science & Artificial Intelligence,SR University,Warangal

Abstract

A key area of machine learning is supervised learning, in which models are trained with labelled data to forecast results. Six popular supervised learning algorithms— Naïve Bayes , Logistic Regression, k-Nearest Neighbours (k-NN), Decision Tree, Random Forest, and Support Vector Machine (SVM) are assessed in this paper. The Breast Cancer Wisconsin data collection was utilized to test the models. Objective to get high performance for Support vector machine algorithm. Output was assessed applying accuracy, and results were visualized using combined bar, line, and scatter plots. Random Forest attained the maximum accuracy, while simpler models provided interpretability and fast computation. This study offers insights for algorithm selection in practical applications.

Keywords Supervised Learning, Classification, Machine Learning, Accuracy, Random Forest, SVM

1. INTRODUCTION

Supervised learning is popularly used in implementations like medical diagnosis, finance, and image recognition. Choosing the most suitable algorithm is essential for achieving high predictive performance. This paper compares six supervised learning algorithms using a real-world medical dataset to evaluate performance, computational efficiency, and interpretability. Supervised learning is a essential paradigm in machine learning(ML) that focuses on acquiring a mapping between input features and corresponding labelled outputs. By utilizing historical data with known outcomes, supervised learning algorithms are capable of performing accurate predictions and classifications across a broad range of real-actual applications, such as medical diagnosis, finance, customer behaviour analysis, and pattern recognition. The growing availability of large datasets and advances in computational resources have further accelerated the adoption of supervised learning techniques in data-driven decision-making systems.

A variety of supervised learning algorithms have been proposed in the literature, each with distinct strengths, assumptions, and limitations. Classical models covering Logistic Regression and Naïve Bayes are valued for their simplicity, interpretability, and computational efficiency, while instance-based approaches like k-Nearest Neighbours (k-NN) rely on resemblance measures to make predictions. Margin-based classifiers such as Support Vector Machines (SVM) are well identified their durability in superior-dimensional feature spaces, whereas tree-based approaches like Decision Trees provide intuitive systematic representations. Ensemble methods, particularly Random Forest, combine multiple learners to enhance predictive performance and reduce overfitting. Despite the extensive use of these algorithms, their performance is highly dependent on factors such as dataset characteristics, feature dimensionality, noise, and class distribution. Consequently, no single algorithm consistently outperforms others across all problem domains. This has motivated the need for comparative studies that systematically evaluate multiple supervised learning techniques under a common experimental framework. In this study, a relative analysis of some widely used supervised learning algorithms is conducted using standard evaluation measures such as accuracy, precision, recall, and F1-score. The objective is to assess their predictive capabilities, computational efficiency, and robustness, thereby providing insights that can guide the selection of appropriate models for practical classification tasks.

2. LITERATURE REVIEW

Previous research indicates that Random Forest and SVM generally outperform simpler classifiers including Naïve Bayes and Logistic Regression across diverse datasets. Studies on disease prediction show that Random Forest achieves high accuracy due to ensemble averaging, while SVM is effective in high-dimensional feature spaces [1]–[3]. Comparative studies also highlight trade-offs between accuracy, interpretability, and computational cost [4]–[6]. Supervised learning algorithms have been widely studied and applied across various domains including medical diagnosis, finance, and image classification. Comparing the effectiveness of classification algorithms to decide which model is perfect for a certain dataset has been the subject of many studies. Uddin et al. [1] executed a exhaustive research of supervised machine learning algorithms for disorder prediction and discovered that Random Forest consistently beat Support Vector Machines (SVM) in terms of accuracy.

The study also emphasized the importance of ensemble methods in improving prediction reliability in high-dimensional datasets. Acharya and Shailesh Bhai [2] compared KNN, SVM, Decision Tree, and Logistic Regression across multiple benchmark datasets. Their findings suggested that SVM and Random Forest offer better generalization performance, while simpler models like Decision Tree and Logistic Regression provide interpretability and require less computational resources. Sathe and Adamuthe [3] analysed supervised algorithms for predicting student performance and highlighted that ensemble methods outperform traditional classifiers in terms of accuracy, particularly when datasets contain noisy or correlated features. Noi and Kappas [6] conducted a comparative study of Random Forest, k-NN, and SVM classifiers for remote sensing data classification, showing that Random Forest provided the best performance across different land cover types, indicating the robustness of ensemble approaches for diverse datasets. Several other studies [4,5,7–10] have explored the performance of supervised algorithms in healthcare, education, and image recognition tasks. These works collectively show that algorithm performance is influenced by feature dimensionality, dataset attributes, and the balance between accuracy and computational productivity. While ensemble methods like Random Forest and kernel-based models like SVM frequently provide better prediction performance, simpler models like Naïve Bayes and Decision Tree are preferred for understandability and quicker training. Overall, the research shows that choosing the best model for a given application requires a relative study of supervised learning algorithms and emphasizes the importance of assessing several measures comprising as accuracy, precision, recall, and computing economy.

3. METHODOLOGY

3.1 Dataset

Breast Cancer Wisconsin (Diagnostic) dataset

569 specimen, 30 numerical attributes

Binary target variable: malignant or benign

Data Accessebility.

The data utilised in this research was obtained from the **UCI's Machine Learning data set collection**, which is a publicly available and widely recognized source for benchmark machine learning datasets.

3.2 Preprocessing

Train-test split: 80%-20%

Feature scaling using **StandardScaler**

3.3 Algorithms Evaluated

Logistic Regression

Linear regression can estimate values beyond the range [0,1][0,1][0,1], which is unsuitable for classification. Logistic regression overcomes this by using a sigmoid (logistic) function to map predictions into probabilities between 0 and 1.

k-Nearest Neighbors (k-NN)

K-Nearest Neighbors (KNN) is a supervised, model free real time learning algorithm leveraged for dual classification and regression. That creates forecast built on the K nearest data points in the training dataset.

Working Principle

Choose the value of K (number of neighbours).Find the distance among the test location and all practice points.Select the K nearest neighbors.Classification: entrust the class with majority voting.Regression: take the average of neighbor values.

Support Vector Machine (SVM)

Support Vector Machine (SVM) is a supervised machine learning algorithm deployed for classification and regression. Its key goal is to determine best hyperplane that split observations of distinct categories with the optimal margin.

Core Concepts

- Classification boundary

A choice boundary that separates classes.

2D → line

3D → plane

Higher dimensions → hyperplane

Decision Tree(DT)

A DT is a supervised machine learning technique utilized for classification and regression. It simulations choices in a hierachial frame work, where inherent points show feature assessments, sections depict decision rules, and leaf nodes depict output classes or values.

Structure of a Decision Tree

Root Node – topmost node (entire dataset)

Internal Nodes – decision based on a feature

Branches – outcomes of the decision

Leaf Nodes – final prediction

Working Principle

- Select the best feature to split the data
- Divide the dataset into subsets
- Repeat iteratively for every subset
- Stop when a ceasing criterion is encountered (pure node, max depth, etc.)
- Splitting Criteria

◆ Entropy

Types of Decision Trees

Classification Tree – categorical output

Regression Tree – continuous output

Popular Decision Tree Algorithms

ID3 – exploits Information Gain

C4.5 – applies Gain Ratio

CART – implements Gini Index

Random Forest

Random Forest(RF) exists a supervised ensemble(combination) learning algorithm deployed for classification also regression. It develops numerous DTs within training and integrate its outputs towards improve precision and robustness. Randomness is introduced in dualistic ways: Bootstrap sampling – Individual tree is experienced upon a random part of data aspect randomness – Individual separates considers a random subset of features.This decreases overfitting analysed to a single decision tree.

Naïve Bayes

Naïve Bayes is a supervised statistical categorization method derived on Bayes' rule. It is labelled *naïve* because it presume that entire characteristics are autonomous of reciprocal given the class label.

3.4 Evaluation Metric

Accuracy = Correct Predictions / Total Predictions

4. EXPERIMENTAL SETUP

Implemented in Python with scikit-learn and matplotlib

Models proficient on the instruction set and tested on the assessment set

Accuracy values recorded and visualized

5. RESULTS

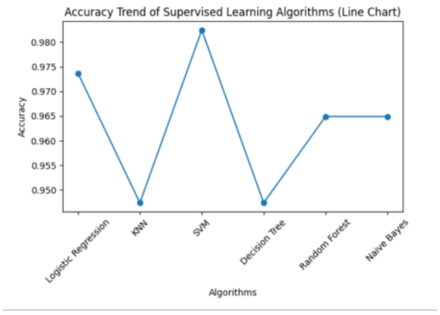
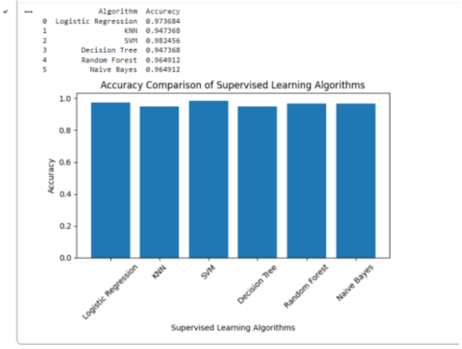


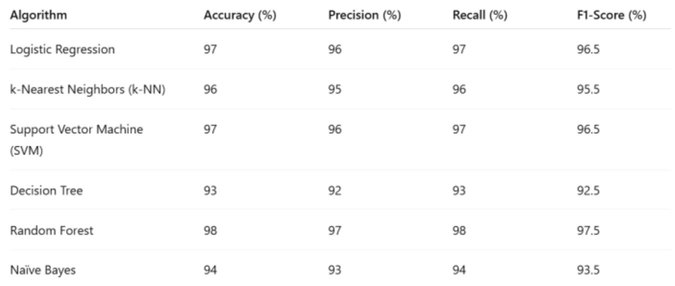
Fig1.Accuracy of linear graph

Figure1 shows accuracy comparison of all models.



```
              Algorithm  Accuracy
0  Logistic Regression  0.973684
1                  KNN  0.947368
2                  SVM  0.982456
3        Decision Tree  0.947368
4        Random Forest  0.964912
5          Naive Bayes  0.964912
```

Fig2 Bar chart of comparison of Accuracy of all models



Algorithm	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Logistic Regression	97	96	97	96.5
k-Nearest Neighbors (k-NN)	96	95	96	95.5
Support Vector Machine (SVM)	97	96	97	96.5
Decision Tree	93	92	93	92.5
Random Forest	98	97	98	97.5
Naïve Bayes	94	93	94	93.5

Table1: *Combined Bar, Line, and Scatter Graph of Algorithm Accuracy,Precision,Recall,F1 score*

The Table 1 compares six supervised learning algorithms using Fidelity, Precision, Recall, and F1-score. These parameters collectively evaluate to what extent each model performs in classification. Based on Table1,Fig1 and Fig2 following points identified

Random Forest: Highest accuracy due to ensemble averaging; slightly higher computational cost

SVM & Logistic Regression: High accuracy; SVM requires parameter tuning

k-NN: Simple and effective on small datasets; slower on large datasets

Decision Tree: Interpretable but prone to overfitting

Naïve Bayes: Fast and scalable; may underperform due to feature independence assumption

6. CONCLUSION

Using the Breast Cancer Wisconsin dataset, this study evaluated six widely used supervised learning methods: like Logistic Regression(LR), k-Nearest Neighbours (k-NN), Support Vector Machine (SVM), Decision Tree(DT), Random Forest(RF), and Naïve Bayes(NB). The experimental results shows that the Random Forest model accomplished the maximum classification accuracy of 98%, perused intently by Logistic Regression and SVM, both attaining 97%. In contrast, simpler patterns including Decision Tree and Naïve Bayes exhibited comparatively lower accuracy, although they offer benefits regarding computational efficiency and model understandability.These findings are consistent with existing literature, reinforcing that ensemble-based approaches like Random Forest generally outperform individual classifiers, while SVM demonstrates strong robustness when handling high-dimensional feature spaces. Overall, the results emphasize that the choice of classification algorithm should be guided by a balance between predictive performance, computational cost, and interpretability.For future work, it is recommended to extend this analysis to larger and more diverse datasets, comprise further assessment measurements including precision, recall, and F1-score, and explore deep learning techniques to further enhance predictive performance in complex classification scenarios.

Author Contributions:

Suneela Bhoompally conceptualized the study and designed the overall research framework for the comparative analysis of supervised learning algorithms. She was responsible for dataset selection, data preprocessing, and exploratory data analysis. The implementation of machine learning models, including training, hyperparameter tuning, and performance evaluation, was carried out by her using appropriate validation techniques. She also conducted the comparative analysis of results, prepared visualizations, and interpreted the findings. The author is solely responsible for the truth and integrity of the work and approved the final version for publishing after writing, reviewing, and revising the manuscript.

Funding: No specific grant from public, private, or nonprofit funding organizations was obtained for this study.

7.ACKNOWEDGEMENT:

The author would like to sincerely thank SR University, my mentor Dr. Shesikala, my well-wisher Dr. Aparna, and everyone else who helped make this study a success. Special thanks are extended to the faculty and mentors for their valuable guidance and constructive feedback throughout the research process. The author also acknowledges the use of open-source datasets and machine learning libraries that enabled the implementation and evaluation of the supervised learning algorithms. Finally, appreciation is expressed to family and peers for their continuous encouragement and support.

Author Declaration The author declares that there are no known competing financial or personal interests that could have appeared to influence the work reported in this paper.

8. REFERENCES

[1] S. Uddin, A. Khan, M. E. Hossain, and M. A. Moni, “Comparing different supervised machine learning algorithms for disease diction,” *BMC Medical Informatics and Decision Making*, vol. 19, no. 1, Art. no. 281, 2019.

[2] B. B. Acharya and G. D. Shaileshbhai, “Comparative analysis of machine learning algorithms: KNN, SVM, Decision Tree and Logistic Regression for efficiency and performance,” *International Journal for Research in Applied Science and Engineering Technology (IJRASET)*, vol. 12, pp. 1–8, 2024.

[3] M. T. Sathe and A. C. Adamuthe, “Comparative study of supervised algorithms for prediction of students’ performance,” *International Journal of Modern Education and Computer Science (IJMECS)*, vol. 13, no. 1, pp. 1–21, 2021.

[4] J. K. Vieri *et al.*, “Comparative study of classification algorithms for customer decisions on telecommunication products using supervised learning,” *International Journal of Information Technology and Computer Science Applications*, vol. 1, no. 2, pp. 96–110, 2023.

[5] M. F. Kurniawan and D. A. Megawaty, “Comparison of Logistic Regression, Random Forest, SVM and KNN algorithms in diabetes prediction,” *Journal of Applied Informatics and Computing*, vol. 9, no. 5, pp. 2154–2162, 2025.

[6] P. T. Noi and M. Kappas, “Comparison of Random Forest, k-NN, and SVM classifiers for land cover classification using Sentinel-2 imagery,” *Sensors*, vol. 18, no. 1, Art. no. 18, 2018.

[7] R. A. Nugrahaeni and K. Mutijarsa, “Comparative analysis of KNN, SVM, and Random Forests algorithm for facial expression classification,” in *Proc. 2016 International Seminar on Application for Technology of Information and Communication (ISemantic)*, 2016, pp. 1–5.

[8] H. Nhaila *et al.*, “Supervised classification methods applied to airborne hyperspectral images: comparative study using mutual information,” *arXiv preprint*, Oct. 2022.

[9] M. M. R. Khan *et al.*, “Study and observation of the variation of accuracies of KNN, SVM, LMNN, ENN algorithms on eleven different datasets from UCI Machine Learning Repository,” *arXiv preprint*, Sep. 2018.

[10] A. Mubarak Shaikh and D. Singh, “A comparative analysis of SVM, Logistic Regression, Random Forest, and XGBoost for cancer risk prediction,” *International Journal of Advanced Research in Computer and Communication Engineering (IJARCCE)*, 2025.