Journal of Advances in Science and Technology
Vol. 21, Issue No. 1, March-2024, ISSN 2230-9659

Performance Evaluation of Machine Learning Algorithms for Iris Dataset Classification

Arshad Iqbal*

K. A. Nizami Centre for Quranic Studies, Aligarh Muslim University, Aligarh,UP, India

Email: aiqbal.cqs@amu.ac.in

Abstract: ***The problem of classifying iris flowers is a well-known issue in the fields of machine learning and pattern recognition. The Iris dataset is widely recognized as a standard for evaluating the effectiveness of classification algorithms in machine learning. This paper offers a comparative assessment of different machine learning algorithms when applied to the Iris dataset. We implement and assess the performance of k-Nearest Neighbors, Logistic Regression, Decision Tree, Linear SVC methods, Random Forest, Gaussian Naïve Bayes and AdaBoost algorithms based on their accuracy in classification. We examine diverse classification methods, appraise their effectiveness, and discuss our findings. The outcomes reveal the accuracy of these models in correctly categorizing the Iris species. To achieve high accuracy, k-Nearest Neighbors, Logistic Regression, Linear SVC methods,Decision Tree,, Random Forest and AdaBoost classifiers are utilised.***

Keywords: Iris dataset, machine learning algorithms, k-NN,SVM,Logistic Regression, Gaussian Naive Bayes,Decision Trees, Random Forest, AdaBoost

----------X----------

1. INTRODUCTION

Machine learning algorithms improve their capabilities by learning from historical datasets. This technology enables us to leverage existing data to derive insightful answers and predictions. In the realm of machine learning, both continuous and discrete values are utilized. The applications of machine learning are vast and varied, encompassing fields such as computer vision, pattern recognition, disease diagnosis, spam detection, weather forecasting, biometric attendance systems, and sentiment analysis.An exemplary instance of a dataset utilized in machine learning is the iris flower dataset, containing 150 samples from three types of iris flowers: Iris versicolor, Iris setosa, and Iris virginica. Each sample is described by four attributes: petal length, petal width, sepal length, and sepal width. By training a predictive model on this dataset, it becomes possible to identify the species of an iris flower based on these features when presented with new data.The methodology for creating such a predictive model involves several key steps. The dataset is initially divided into training and testsets. The training set is then utilized to train the model.Lastly, a variety of machine learning methods are used for classification, such as Gaussian Naive Bayes, AdaBoost classifier, logistic regression, k-Nearest Neighbors, decision trees, Linear Support Vector Classifiers (Linear SVC), random forest classifiers, and decision trees. The objective of this study is to evaluate and compare the effectiveness of different machine learning algorithms in classifying the different species of iris flowers.By examining the performance of each algorithm, insights can be gained into their respective strengths and weaknesses, ultimately contributing to the development of more accurate and efficient classification models.

2. LITERATURE REVIEW

The Iris dataset was presented by Ronald A. and is available via the University of California, Irvine Machine Learning Repository.This dataset is widely recognized in the field of pattern recognition; Fisher first published it in 1936.[1]. Using this dataset, many machine learning approaches have been used to reliably categorize floral species: k-Nearest Neighbors, Support Vector Machine, Logistic Regression, and Neural Networks. Three primary steps are usually involved in implementing these techniques: segmentation, feature extraction, and classification[2].In paper [3], the authors highlighted the feature extraction and evaluation of flower species, demonstrating that the k-Nearest Neighbors and Random Forest algorithms achieved the highest accuracy.The method proposed by the authors aims to adjust a pre-trained convolutional neural network (CNN) in order to improve its ability to detect flower features[4]. Tests on four different datasets showed that the proposed CNN-based model could accurately recognize flowers and also explored the use of machine learning algorithms such as k-Nearest Neighbors, Linear SVC, Logistic Regression, Decision Trees, Gaussian Naive Bayes,

and Random Forest classifiers for Iris species identification [5].In paper [6], a convolutional neural network-based method for flower classification was proposed, utilizing the VGG19 architecture for feature extraction and softmax as the activation function in a transfer learning approach. This method achieved a training accuracy of 100% and a validation accuracy of 91.1%. Comparative studies using the Iris dataset showed that the Support Vector Machine (SVM) classifier achieved the highest accuracy at 95%, followed by Decision Trees at 93%, and k-Nearest Neighbors at 92% [7]. The authors in paper [8] applied the Gaussian Naive Bayes supervised learning algorithm to classify Iris species, achieving an accuracy of approximately 95%. A flower recognition system using image processing techniques was developed to classify flowers based on edge and color features. This system, tested on ten different flower species, achieved over 80% accuracy using the k-Nearest Neighbor algorithm [9].

3. METHODOLOGY

The Iris dataset, based on Fisher's model, is a cornerstone in machine learning classification tasks. The dataset comprises 50 instances each of three types of Iris flowers: Iris setosa, Iris versicolor, and Iris virginica.Each row represents one flower and has columns for the measurements of petal length, petal width, sepal length, and sepal width in centimeters.

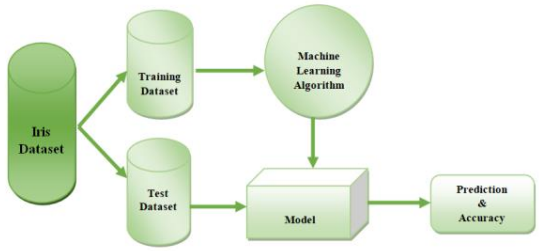

Figure 1: Block Diagram of Machine Learning System

Developing a machine learning system to recognize iris flowers using this dataset involves several critical steps as depicted in Fig 1. To begin with, it is crucial to prepare the data by addressing any missing information, normalizing features, and dividing the dataset into training and test sets. Following this, a variety of machine learning techniques are examined - including Logistic Regression, k-Nearest Neighbors, Decision Trees, Support Vector Machine (SVM), Gaussian Naive Bayes, Random Forest, and AdaBoost - in order to identify the most suitable one for classification purposes. Once chosen, the model is trained on the training set and then assessed on the testing set using measures like accuracy, precision, recall, and F1 score.Finally, the trained model is implemented to classify iris flowers effectively, ensuring practical applicability and robust performance. This comprehensive methodology not only facilitates accurate classification of iris flowers but also provides a structured framework for applying machine learning techniques to similar classification problems.

4. RESULTS AND DISCUSSION

In Iris dataset, 80% is made up of training dataset, while the remaining 20% is test dataset. Supervised learning algorithms likek-Nearest Neighbors,Logistic Regression, Support Vector Machines, Decision Trees, Random Forest, and AdaBoost demonstrate remarkable efficiency, achieving 100% classification accuracy, while Gaussian Naive Bayes achieves an accuracy of 98%. These results highlight the effectiveness of these algorithms for classification tasks. The performance indicators for this classification task, involving the target classes Iris versicolor,Iris setosa and Iris virginica, are detailed in Table I and Table II. These tables present the main classification metrics—recall, precision accuracy, and F1 score—for each class. These metrics are derived from the elements of the confusion matrix, illustrated in Fig 2 and 3, where positive and negative correspond to the predicted classes in the dataset. In summary, the use of supervised learning algorithms, including, k-Nearest Neighbors, Logistic Regression,Support Vector Machines, Decision Trees,Random Forest, and AdaBoost, has resulted in achieving the highest 100% classification accuracy for the Iris dataset. Gaussian Naive Bayes also performs admirably with an accuracy of 98%. The comprehensive evaluation metrics presented confirm the robustness of these models, providing valuable insights into their precision, recall, accuracy, and F1 score for each iris species.

Table I: Classification report of, K-NN=3, Linear (SVC), Logistic Regression,Random Forest, Decision Tree, and AdaBoost Classifiers

	Iris-setosa	Iris-versicolor	Iris-virginica	Accuracy
Precision	100%	100%	100%	100%
Recall	100%	100%	100%	
F1-score	100%	100%	100%	
Support	13	19	13	45

Table II: Classification report of Gaussian Naïve Bayes

	Iris-setosa	Iris-versicolor	Iris-virginica	Accuracy
Precision	100%	100%	100%	98%
Recall	92%	100%	96%	
F1-score	100%	100%	96%	
Support	13	19	13	45

Fig 2 presents the confusion matrix, which assesses the performance of the classifiers k-Nearest Neighbors (k=3),Linear SVC,Logistic Regression,Decision Tree, AdaBoostand Random Forest on the Iris dataset, demonstrating accurate species predictions. Figure 3, on the other hand, evaluates the performance of the Gaussian Naive Bayes classifier.

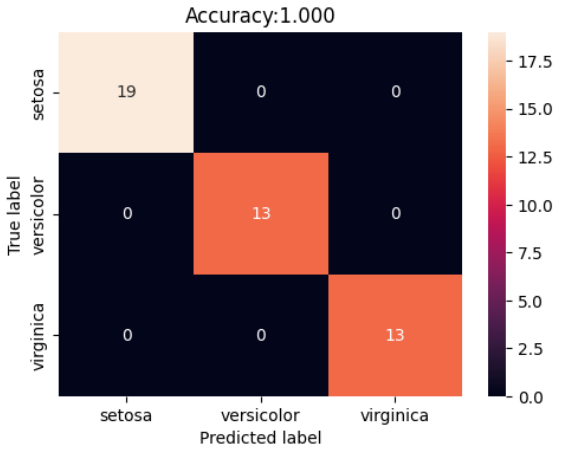

Figure 2: Confusion matrix of Logistic Regression, K-NN=3, Decision Tree, Linear (SVC), Random Forest and AdaBoost Classifiers

The confusion matrixof Fig 2 is interpreted as below:

True Positives (TP): The diagonal elements represent the count of accurately classified instances for each respective class.

- **Setosa**: There were 19 instances correctly classified as Setosa.
- **Versicolor**: There were 13 instances correctly classified as Versicolor.
- **Virginica**: There were 13instances correctly classified as Virginica.

False Positives (FP): The off-diagonal elements in each column show the number of instances that were incorrectly classified as belonging to a specific class.

- **Setosa**: No instances of Versicolor or Virginica were misclassified as Setosa (both are 0).
- **Versicolor**: No instancesSetosaor Virginicawere misclassified as Versicolor(0 for both)
- **Virginica**: No instancesSetosaor Versicolorwere misclassified as Virginica(0 for both)

False Negatives (FN): The off-diagonal elements in each row display the number of instances of a particular class that were misclassified as another class.

- **Setosa**: No Setosa instances were misclassified as Versicolor or Virginica (0 for both).
- **Versicolor**: No Versicolor instances were misclassified as Setosaor Virginica (0 for both).
- **Virginica**: No Virginica instances were misclassified as Setosa or Versicolor (0 for both).

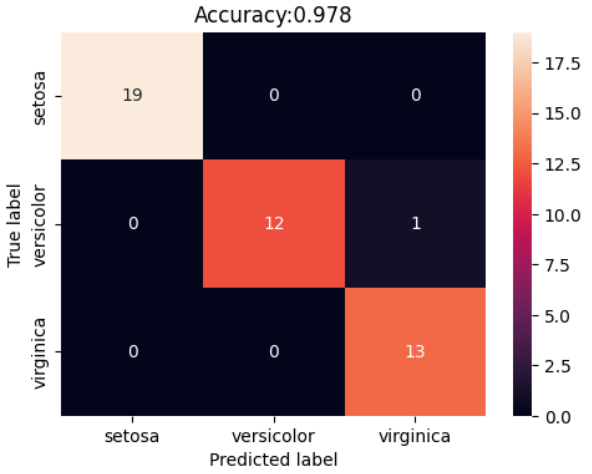

Figure 3: Confusion matrix of Gaussian Naïve Bayes Classifier

The confusion matrix of Fig 3 is interpreted as below:

True Positives (TP):

- **Setosa**: There were 19 instances correctly classified as Setosa.
- **Versicolor**: There were 12 instances correctly classified as Versicolor.
- **Virginica**: There were 13instances correctly classified as Virginica.

False Positives (FP):

- **Setosa**: No instances of Versicolor or Virginica were misclassified as Setosa (0 for both).
- **Versicolor**: No instancesSetosaor Virginicawere misclassifiedas Versicolor(0 for both)
- Virginica: 1 instance of Versicolor was misclassifiedas Virginica, and 0 instances of Setosa were **incorrectly** classified as Virginica.

False Negatives (FN):

- **Setosa**: No Setosa instances were misclassifiedas Versicolor or Virginica (0 for both).
- **Versicolor**: 1Versicolor instance wasmisclassifiedas VirginicaSetosa or 0 instances of Versicolor were misclassified as Setosa.
- **Virginica**: No Virginica instances were misclassifiedas Setosa or Versicolor (0 for both).

5. CONCLUSION

In addition to improving understanding of the fundamental ideas, classifying iris species using machine learning enables researchers to better understand of plant biodiversity, ecological studies, and the practical application of medical knowledge. k-Nearest Neighbors, Logistic Regression, Linear SVC techniques,Decision Tree, Random Forest, and AdaBoost classifiers are used to attain high accuracy.A nearly 100% accuracy rate has been attained, demonstrating the effectiveness of supervised learning-based techniques like Random Forest, AdaBoost, Decision Trees, Linear SVC methods, k-Nearest Neighbors, Logistic Regression, and Random Forest.Future work aims to expand this research by incorporating additional datasets and exploring more supervised learning algorithms. This will further advance the understanding of machine learning principles and improve the classification performance across diverse applications.

REFERENCES

1. Fisher, Ronald A. "UCI Machine Learning Repository: Iris Data." (1936): Available: http://archive.ics.uci.edu/ml/datasets/Iris.
2. Shukla, Asmita, Ankita Agarwal, Hemlata Pant, and Priyanka Mishra. "Flower classification using supervised learning." *Int. J. Eng. Res* 9, no. 05 (2020): 757-762.
3. Bhutada, Sunil, K. Tejaswi, and S. Vineela. "Flower recognition using machine learning." *International Journal of Researches In Biosciences, Agriculture And Technology* 4, no. 2 (2021): 67-73.
4. Dias, Philipe A., Amy Tabb, and Henry Medeiros. "Apple flower detection using deep convolutional networks." *Computers in Industry* 99 (2018): 17-28.
5. Ahuja, Gunjan, Muskan Aggarwal, Jashn Tyagi, and Onkar Mehra. "Identification Of Different Species Of Iris Flower Using Machine Learning Algorithms."*International Research Journal of Engineering and Technology*10, no. 1(2022): 434-438.
6. Desai, Sandip, Chetan Gode, and PunitFulzele. "Flower image classification using convolutional neural network." In *2022 First International Conference on Electrical, Electronics, Information and Communication Technologies (ICEEICT)*, pp. 1-4. IEEE, 2022.
7. Goel, Hridya, and MsPushpanjali. "Iris Flower Classification Using Machine Learning."*International Journal of Progressive Research In Engineering Management and Science* 4, no 5 (2024): 655-657.
8. Iqbal, Zainab, and Manoj Yadav. "Multiclass Classification with Iris Dataset using Gaussian Naive Bayes." *International Journal of Computer Science and Mobile Computing* 9 (2020): 27-35.
9. Tiay, Tanakorn, PipimphornBenyaphaichit, and PanomkhawnRiyamongkol. "Flower recognition system based on image processing." In *2014 Third ICT International Student Project Conference (ICT-ISPC)*, pp. 99-102. IEEE, 2014.

Corresponding Author

Arshad Iqbal*

K. A. Nizami Centre for Quranic Studies,
Aligarh Muslim University, Aligarh,UP, India

Email: aiqbal.cqs@amu.ac.in