Vidhyayana - ISSN 2454-8596
An International Multidisciplinary Peer-Reviewed E-Journal
www.vidhyayanaejournal.org
Indexed in: Crossref, ROAD & Google Scholar

28

Enhancing Sustainable AI with Efficient Machine Learning Models: An Iris Classification Perspective

Gulshan Hariyani

Student, Dept. of Computer Engineering and Technology

Dr. Vishwanath Karad MIT-World Peace University, Pune, India

Shamla Mantri

Associate Professor, Dept. of Computer Engineering and Technology

Dr. Vishwanath Karad MIT-World Peace University, Pune, India

Abstract

The earth's biodiversity is abundant. Approximately 360000 species contribute to the earth's ecology by forming a healthy biome. In terms of size, form, and colour, some of them are physically identical. As a result, identifying any species is challenging. Setosa is first, Versicolor is second, and Virginica is third, these are the three subspecies of the flower species known as Iris. As the Iris dataset is regularly accessible, we chose to use it. There are three classes of fifty examples each in the Iris flower dataset. Machine learning is used in the Iris dataset to determine the subspecies of iris blossoms. The work focuses on the automatic identification of floral classes by machine learning techniques with a high degree of accuracy as opposed to approximation. Pre-processing, dataset partitioning, and classification utilizing Random Forest, K-Nearest Neighbours, Logistic Regression, and Support Vector Machine are the four models involved in putting this strategy into practice.

Vidhyayana - ISSN 2454-8596
An International Multidisciplinary Peer-Reviewed E-Journal
www.vidhyayanaejournal.org
Indexed in: Crossref, ROAD & Google Scholar

Keywords: Logistic Regression, Classification, Support Vector Machines, Machine Learning, Random Forest, Iris dataset, K-Nearest Neighbours.

I. Introduction

Data-driven decision-making has been transformed by machine learning in several fields, including engineering, social sciences, healthcare, and finance. Classification, which entails grouping data into predetermined categories according to feature qualities, is one of the core tasks in machine learning. Ronald A. Fisher first presented the Iris dataset in 1936, and it is now one of the most popular benchmark dataset for classification tasks. This dataset's simplicity, well-balanced structure, and suitability for testing classification models make it one of the most well-known and widely used dataset in machine learning research.

150 samples from three different species of iris flowers— Iris-virginica is first, Iris-setosa is second and Iris-versicolor is third make up the Iris dataset. Length of sepal, width of sepal, length of petal, and width of petal all measured in centimetres are the four main characteristics that define each sample. Assigning an input flower specimen to one of these three species using the given feature set is the classification challenge in this case. The prevalence of overlapping feature distributions between the classes, especially for Iris-versicolor and Iris-virginica, makes the dataset an intriguing challenge for classification algorithms despite its small size.

This study's main goal is to evaluate the relative performance and efficacy of four popular classification methods for iris flower species categorization: Random Forest (RF), K-Nearest Neighbours (KNN), Logistic Regression (LR), and Support Vector Machine (SVM). These algorithms are all appropriate for a thorough comparison analysis because they each represent a unique categorization strategy and have unique benefits and drawbacks.

II. Literature Survey

A. Shukla, et al. [1] draw attention to how difficult it might be to distinguish between species when their physical characteristics are identical. They make use of the popular Iris dataset, which includes 50 samples from each of the three species—Virginica, Versicolor, and Setosa. The three stages of the process are classification, feature extraction, and segmentation. To

categorize the flower species with high accuracy, the study uses a variety of supervised learning methods, such as NN, LR, SVM, and KNN. The significance of machine learning in automating species recognition is emphasized throughout the paper.

T. S. Rao, et al. [2] look at various ML techniques for categorizing the Iris dataset. The historical context of the dataset and the importance of the features—sepal length, sepal width, petal length, and petal width—are covered in the study. The study describes how different algorithms are implemented and evaluates how well they perform in terms of accuracy for classification. The author highlights the significance of selecting the appropriate method for optimizing flower classification accuracy by presenting findings that demonstrate how each approach performs with this dataset. There is an additional discussion of the difficulties encountered when processing the dataset.

S. A. Mithy, et al. [3] examine some ML methods, such as SVM, RF, and DT. They offer a thorough evaluation of each algorithm's performance using accuracy metrics. The significance of model validation and selection in obtaining trustworthy classification results is emphasized in the paper. Before offering recommendations for further research into improving classification methods for botanical datasets, the authors also go into feature selection and preprocessing procedures that enhance model performance.

S. T. HalaKatti, et al. [4] classify several approaches, including hybrid approaches, supervised learning, and unsupervised learning, and evaluate their efficacy. They talk about how deep learning techniques and other developments in machine learning can be used for the Iris dataset. The paper identifies possible improvements while highlighting the drawbacks of the current approaches, such as their processing demands and data dependence. The authors offer suggestions for future lines of inquiry that may result in improved classification techniques for botanical study in their conclusion.

III. Dataset Description

Often used for classification problems, the Iris dataset is a classic and old in statistics and machine learning. Kaggle is an open-source dataset repository from which the dataset was

Vidhyayana - ISSN 2454-8596
An International Multidisciplinary Peer-Reviewed E-Journal
www.vidhyayanaejournal.org
Indexed in: Crossref, ROAD & Google Scholar

extracted. There is nothing that we have altered because the dataset itself has all the information needed for the study.

Detail description of each attribute in the final dataset are as follows:

1. ***SepalLengthCm:*** This indicates the flower's sepal length in centimeters. It is a feature that is numerical.
2. ***SepalWidthCm:*** This indicates the flower's sepal width in centimeters. It is a feature that is numerical.
3. ***PetalLengthCm:*** This indicates the flower's petal length in centimeters. It is a feature that is numerical.
4. ***PetalWidthCm:*** This indicates the flower's petal width in centimeters. It is a feature that is numerical.
5. ***Species:*** It shows the types of the iris flower species. It contains 3 species as Virginica, Versicolor and Setosa. It is a categorical target value.

	SepalLengthCm	SepalWidthCm	PetalLengthCm	PetalWidthCm	Species
0	5.1	3.5	1.4	0.2	Iris-setosa
1	4.9	3.0	1.4	0.2	Iris-setosa
2	4.7	3.2	1.3	0.2	Iris-setosa
3	4.6	3.1	1.5	0.2	Iris-setosa
4	5.0	3.6	1.4	0.2	Iris-setosa

Fig 1. Data of First 5 Records

	SepalLengthCm	SepalWidthCm	PetalLengthCm	PetalWidthCm
count	150.000000	150.000000	150.000000	150.000000
mean	5.843333	3.054000	3.758667	1.198667
std	0.828066	0.433594	1.764420	0.763161
min	4.300000	2.000000	1.000000	0.100000
25%	5.100000	2.800000	1.600000	0.300000
50%	5.800000	3.000000	4.350000	1.300000
75%	6.400000	3.300000	5.100000	1.800000
max	7.900000	4.400000	6.900000	2.500000

Fig 2. Statistical Information of Dataset

Vidhyayana - ISSN 2454-8596
An International Multidisciplinary Peer-Reviewed E-Journal
www.vidhyayanaejournal.org
Indexed in: Crossref, ROAD & Google Scholar

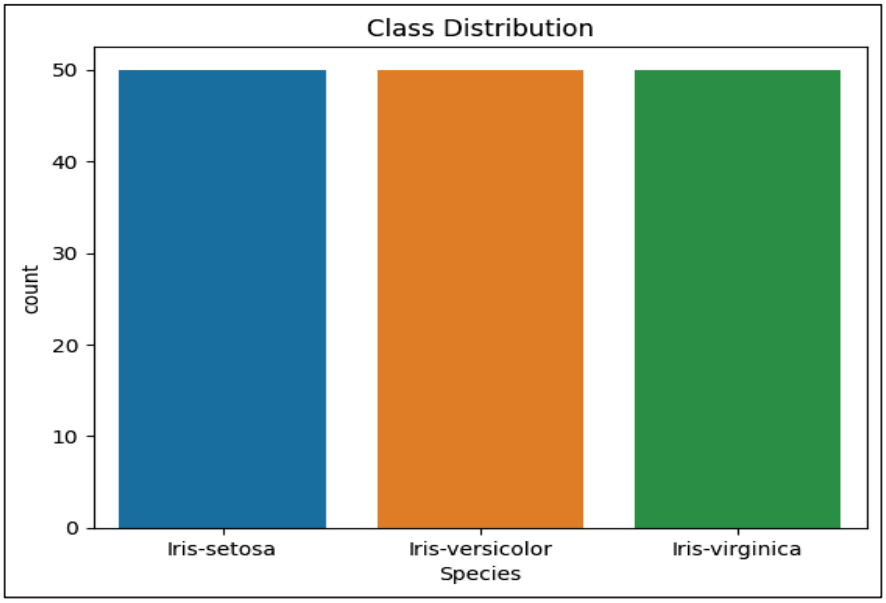

Fig 3. Class Distribution of Iris Flower

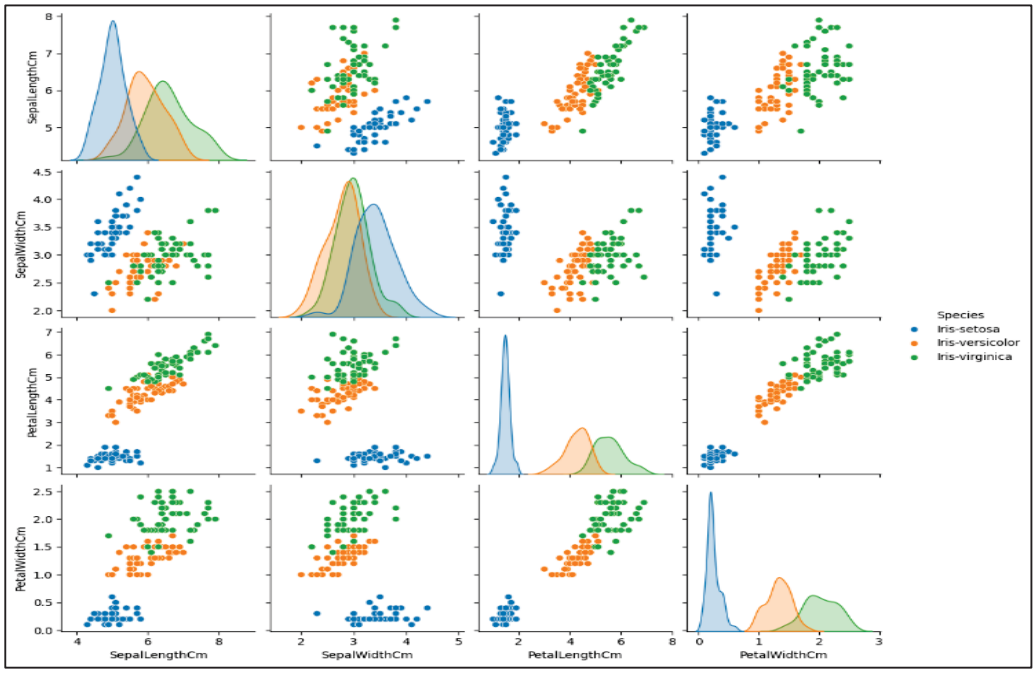

Fig 4. Pair Plot of Whole Dataset

Vidhyayana - ISSN 2454-8596
An International Multidisciplinary Peer-Reviewed E-Journal
www.vidhyayanaejournal.org
Indexed in: Crossref, ROAD & Google Scholar

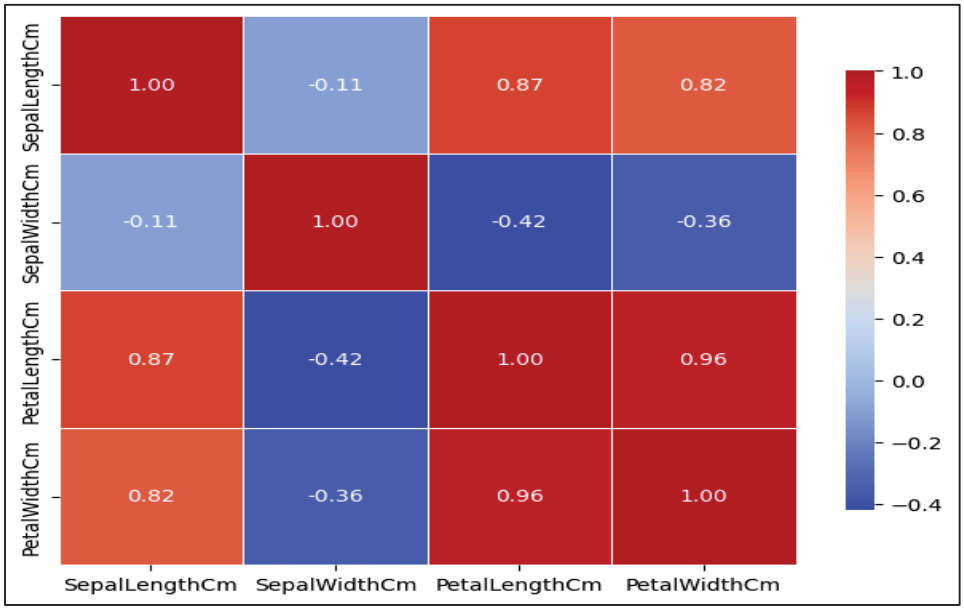

Fig 5. Correlation Heatmap of Dataset

Above all figures show all the visualizing information done on our dataset to carry on further calculations. We have plot different maps between features to observe which features are more related to each other or more dependent on one-another. By doing these visualizations we came to know which features contribute more in calculating the target value so we can give it more weightage.

IV. Methodology

Import Data: Data is taken from the Kaggle Repository.

Data Pre-processing: No Null values and no empty rows. Only two pretreatment techniques are used:

a) **Lable Encoding:** This encoding method is used to convert our target value which is species name to numerical value for better training and performance of models.

b) **Standard Scaling:** It is used to transform data features in order for their standard deviation to be one and their mean to be zero, essentially centering the data and ensuring all features contribute equally to a machine learning model.

Model Selection and Training: The research has used four machine learning models to train, they are:

a) **Random Forest:** A learning system which uses ensemble technique internally is called Random Forest (RF) constructs several decision trees and aggregates their results to generate predictions. It employs a method known as bagging (Bootstrap Aggregating), in which each tree is trained individually by using distinct subsets of the training data. It uses a majority voting system among the trees for classification tasks and averages their predictions for regression tasks. Compared to individual decision trees, Random Forest is much more resilient, less likely to overfit, and effective with both linear and non-linear data. However, compared to a single decision tree, it can be less interpretable and computationally costly for huge datasets.

b) **K-Nearest Neighbours:** A straightforward but efficient technique called K-Nearest Neighbours (KNN) classifies a data item according to the class having major of its K nearest neighbours. Although Euclidean distance is commonly used to quantify the linear distance between two random points, other metrics such as Manhattan or Minkowski can also be employed. Instead of requiring a training step, KNN maintains the complete dataset and uses runtime distance comparison to generate predictions. This makes it practical for small datasets and simple to apply. However, because KNN needs to calculate the distance for each new query, it is computationally costly for datasets having large size. Additionally, it is sensitive to unimportant traits, and for best results, the K value must be carefully chosen.

c) **Logistic Regression:** Using strategies like one-vs-rest, the statistical model known as logistic regression (LR), which is mainly utilized for binary classification, can be expanded to multi-class issues. It converts linear combinations of input information into probability values between 0 and 1 using a sigmoid function. It groups observations into different categories based on a threshold, usually it is 0.5. When the data is linearly separable, logistic regression performs well and is quick and easy to understand. It is sensitive to outliers and has trouble understanding intricate, non-linear interactions. To enhance its

Vidhyayana - ISSN 2454-8596
An International Multidisciplinary Peer-Reviewed E-Journal
www.vidhyayanaejournal.org
Indexed in: Crossref, ROAD & Google Scholar

performance and avoid overfitting, regularization techniques such as L1 and L2 which is also known as Lasso and Ridge regression are frequently employed.

d) **Support Vector Machines:** A potent classification system called Support Vector Machine (SVM) increases the margin and maximizes it between classes by determining the best hyperplane to divide them. If the datapoints are not linearly separable, SVM can convert it into a higher-dimensional space where it can be separable by using kernel functions (such as radial basis function (RBF) or polynomial). SVM performs well in high-dimensional environments and is quite efficient for small to medium-sized datasets. For large datasets, it is computationally costly, and it necessitates careful adjustment of hyperparameters like as the kernel type and the regularization parameter, C.

Model Evaluation: After training the above model, we try to predict the data based on the test data (new data) and find out the performance of our models using some model evaluation metrics.

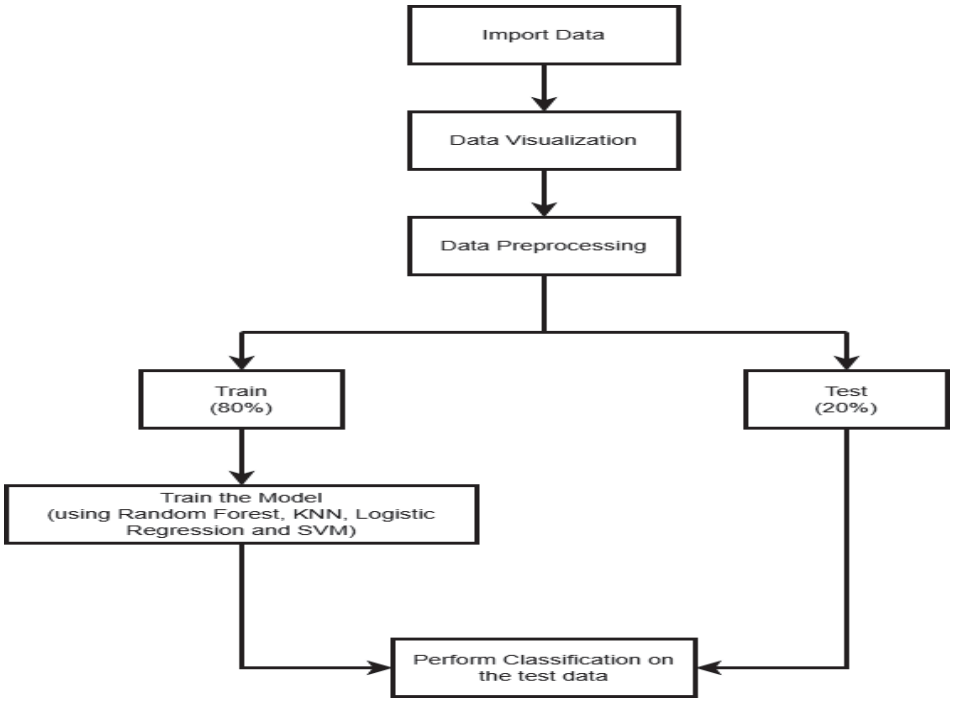

Fig 6. Flowchart of full process

Vidhyayana - ISSN 2454-8596
An International Multidisciplinary Peer-Reviewed E-Journal
www.vidhyayanaejournal.org
Indexed in: Crossref, ROAD & Google Scholar

Above Fig. 6., shows the full flowchart of the process which is followed in this research.

V. Results and Analysis

Table 1: Metrics table of All Models

Metrics	Random Forest	KNN	Logistic Regression	SVM
Accuracy	90%	93.34%	90%	96.67%
Precision	90.23%	94.44%	90.23%	96.96%
Recall	90%	93.34%	90%	96.67%
F1-Score	89.97%	93.26%	89.97%	96.65%

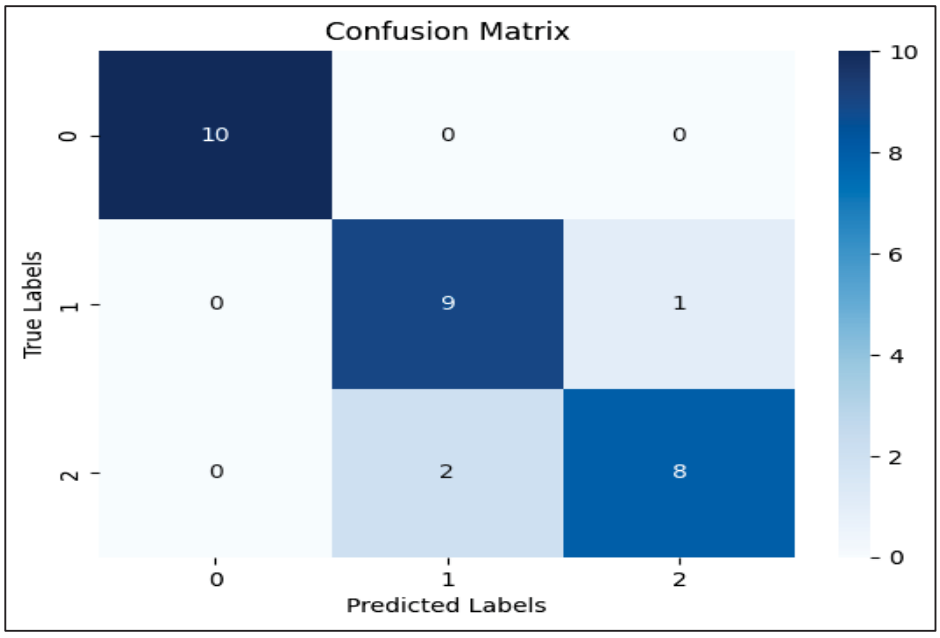

Fig 7. Confusion Matrix of RF

Vidhyayana - ISSN 2454-8596
An International Multidisciplinary Peer-Reviewed E-Journal
www.vidhyayanaejournal.org
Indexed in: Crossref, ROAD & Google Scholar

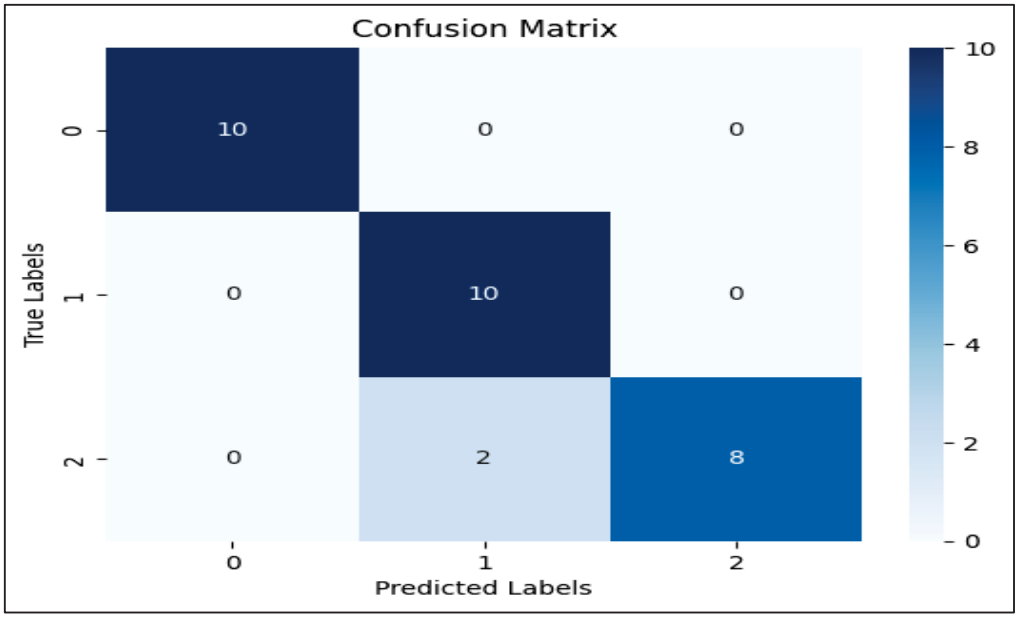

Fig 8. Confusion Matrix of KNN

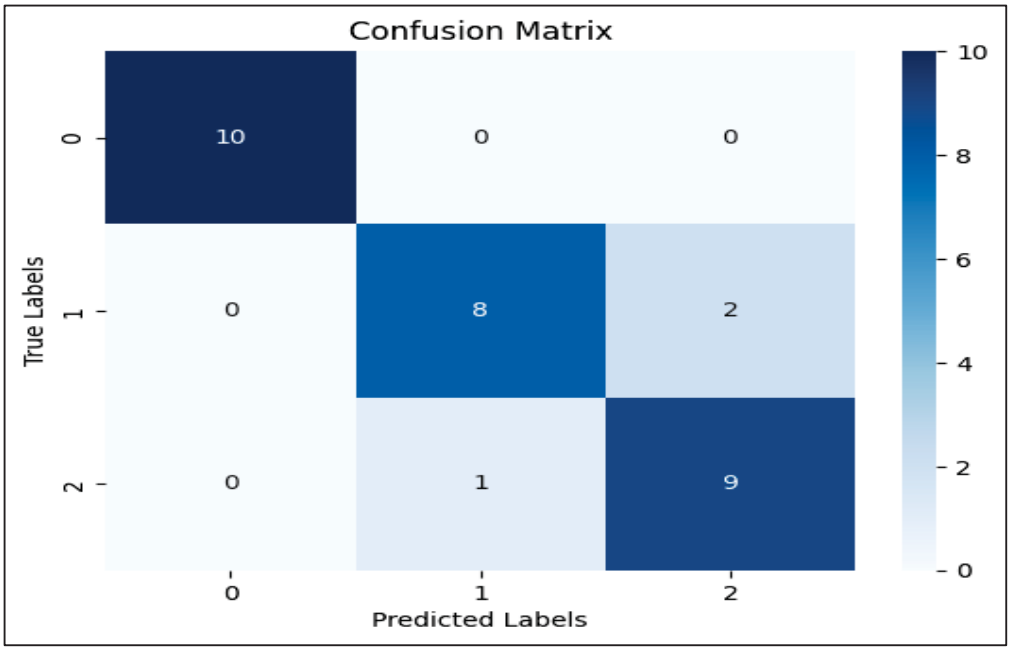

Fig 9. Confusion Matrix of LR

Vidhyayana - ISSN 2454-8596
An International Multidisciplinary Peer-Reviewed E-Journal
www.vidhyayanaejournal.org
Indexed in: Crossref, ROAD & Google Scholar

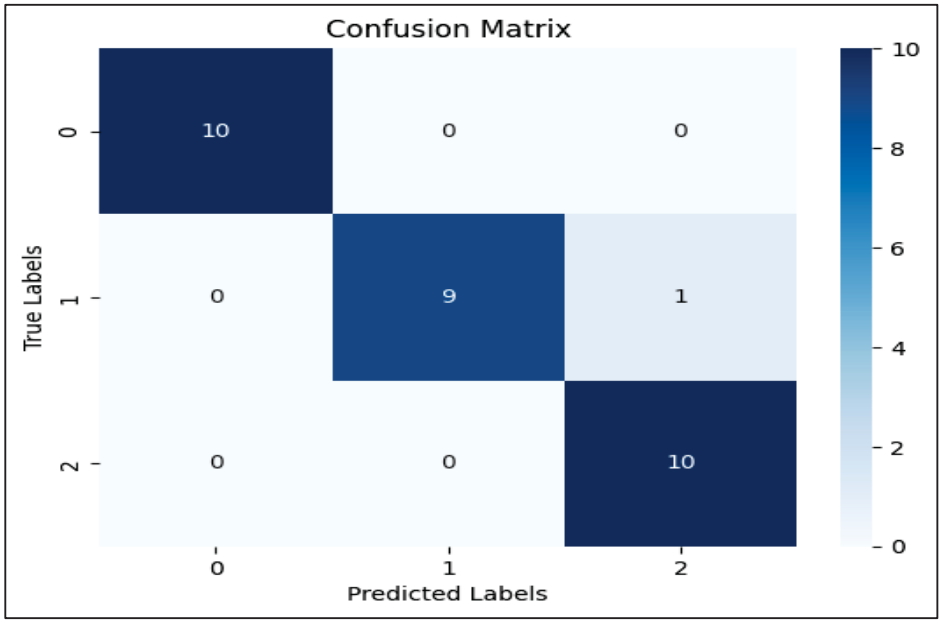

Fig 10. Confusion Matrix of SVM

The four figures from 7 to 10 above show the Confusion Matrices of all the four machine learning models that are used.

VI. Conclusion

We have successfully pre-processed, visualize, trained and compared RF, KNN, LR and SVM for the Classification of Iris Flower Species. Support Vector Machines shows more accurate and precise results than the other two because values of all the metrics like Accuracy (to be accurate), Precision (to be precise), Recall (to measure sensitivity) and F1-Score are more than other three. For Further research, we can change the species of flower or implement image processing using images captured of flowers and plants to classify. Also, different models from ML or DL can also indulge.

References

[1] Shukla, Asmita, et al. "Flower classification using supervised learning." Int. J. Eng. Res 9.05 (2020): 757-762.

[2] Rao, T. Srinivas, et al. "Iris Flower Classification Using Machine Learning." Network 9.6 (2021).

[3] Mithy, S. A., et al. "Classification of iris flower dataset using different algorithms." Int. J. Sci. Res. in (2022).

[4] Halakatti, Shashidhar T., and Shambulinga T. Halakatti. "Identification of iris flower species using machine learning." IPASJ International Journal of Computer Science (IIJCS) 5.8 (2017): 59-69.

[5] Pachipala, Yellamma, et al. "Iris flower classification by using random forest in aws." 2022 6th International Conference on Intelligent Computing and Control Systems (ICICCS). IEEE, 2022.

[6] Kholerdi, Hedyeh A., Nima TaheriNejad, and Axel Jantsch. "Enhancement of classification of small data sets using self-awareness—An iris flower case-study." 2018 IEEE international symposium on circuits and systems (ISCAS). IEEE, 2018.

[7] Bayrakçı, Hilmi, Abdullah Burak Keşkekçi, and Recep Arslan. "Classification of iris flower by random forest algorithm." Advances in Artificial Intelligence Research 2.1 (2022): 7-14.

[8] Mani, Nishanthini, et al. "Enhancing Accuracy: Iris Flower Classification with Ensemble Models." 2024 International Conference on Cognitive Robotics and Intelligent Systems (ICC-ROBINS). IEEE, 2024.

[9] Yang, Yu. "A study of pattern recognition of Iris flower based on Machine Learning." (2013).

[10] Nath, Rituparna, and Arunima Devi. "Machine Learning Algorithms Used for Iris Flower Classification." Critical Approaches to Data Engineering Systems and Analysis. IGI Global, 2024. 193-217.