RESEARCH ARTICLE OPEN ACCESS

Iris Dataset Classification Using Machine Learning

Uppalapati Gowtham[a], Uppunuti Ushasandhya[b], Adlapally Bharadwaj[c], M.Karthikeyan Reddy[d] , Mrs.D.Prathyusha [e,*]

[a,b,c,d] Student, Department of IT. Malla Reddy Engineering College, Maisammaguda, Hyderabad-500100

[e] Professor, Department of IT. Malla Reddy Engineering College, Maisammaguda, Hyderabad-500100

Abstract:

The well-known Iris dataset is used in this case study to use the K-Nearest Neighbors (KNN) method. The 150 iris flower observations in the Iris dataset include 50 observations of each of the three species—Setosa, Versicolor, and Virginica. This case study aims to identify the four characteristics of sepal length, sepal breadth, petal length, and petal width that may be used to categorize iris flowers into their respective species. The KNN method is a well-liked and straightforward classification technique that makes predictions by locating the nearest neighbors of each observation. To guarantee that all of the characteristics in this case study are on the same scale, the dataset is first divided into training and testing sets. The next step is to train a KNN model with k=3, which takes into account each observation's three nearest neighbours.This study utilizes multiple machine learning algorithms for flower classification, including K-Nearest Neighbors (KNN) for distance-based classification, Decision Tree for rule-based decision-making, Logistic Regression for probability estimation, Support Vector Machine (SVM) for margin-based classification, Naïve Bayes for probabilistic modeling, and Random Forest for ensemble learning to enhance accuracy and reduce overfitting, collectively improving the reliability of species identification.Lastly, the accuracy score is used to assess how well the model performed on the test set, Information may be obtained easily because the Iris dataset is readily available from a number of sources, including the Python Sci kit-learn library. Exploratory data analysis is done to see how the data are distributed, identify trends, and understand how the features relate to one another. Common visualization techniques including scatter plots, pair plots, and histograms are used to analyze the data. Preprocessing of the information entails addressing any missing attributes (which are absent from this dataset) and normalizing the component values to ensure that they have a zero mean and a one standard deviation. This is a crucial step for machine learning algorithms that are sensitive to feature scaling.

Keywords: *Iris flower, computer, Decision Tree, Support Vector Machine, flower size, identify..*

I. INTRODUCTION

The financial and healthcare sectors have changed significantly as a result of machine learning and data science's ability to extract meaningful information from vast, complex databases. Among the most significant and educationally significant datasets in pattern recognition and classification is the Iris flower dataset. Ever since it was first presented in the 1936 work "The Use of Multiple Measurements in Taxonomic Problems" by British statistician and biologist Ronald, the Iris dataset has established itself as a standard for illustrating the core ideas of machine learning. The 150 iris flower samples in the Iris Dataset are split evenly across the three species: setosa, versicolor, and virginica.. The length, width, length, and width of the petals, all expressed in centimeters, are the four unique characteristics that define each sample. To categorize a given iris flower into one of the three species using these measurements is the task of creating a predictive model. Because of its small size and the obvious patterns in the data, this classification problem is not only interesting from a botanical standpoint but also makes a perfect teaching tool.

Beyond its biological foundation, the Iris dataset is significant. Many machine learning topics, such as supervised learning, feature selection, model evaluation, and the trade-offs between various classification algorithms, have been shown with it on a large scale. The simplicity of the dataset makes it possible to use a variety of algorithms, from straightforward linear classifiers to more complex techniques like support vector machines and neural networks. Because of its adaptability, it serves as a priceless tool for both practitioners and students, offering a practical introduction to the process of developing and assessing

predictive models. It is critical to recognize the historical and technological advancements in the field of pattern recognition in order to comprehend the larger context.

Aiming to find patterns and base judgments on data, statistical approaches first appeared in the middle of the 20th century. Fisher's release of the Iris dataset was a significant development that demonstrated how these methods could be applied to solve actual classification issues. Machine learning's capabilities have been greatly enhanced over the years by developments in computer power and new algorithmic techniques, making it possible to analyze datasets that are considerably larger and more complex. A number of crucial phases in the machine learning process are shown by the Iris flower classification issue, including data collection, exploratory data analysis (EDA), data preparation, model selection, training, and evaluation[3]. To ensure the creation of a reliable and accurate model, each of these stages is essential. While EDA focuses on displaying and summarizing the data to identify underlying patterns, data collection include sourcing and interpreting the dataset. Data preprocessing, which frequently entails feature normalization or standardization, gets the data ready for analysis.

In conclusion, the Iris flower classification problem serves as a gateway to comprehending the core ideas of machine learning rather than merely being a benchmark dataset. Solving this issue will provide you with understanding of the full machine learning project life cycle.The knowledge gained from this experiment is broadly applicable and offers a solid basis for taking on increasingly difficult and varied categorization problems in the field of machine learning.

II. RELATED WORK

In the area of example acknowledgment and artificial intelligence, the 1936 presentation of the Iris blossom dataset by English scientist and analyst Ronald A. Fisher is a unique resource. Fisher's publication, "The Use of Multiple Measurements in Taxonomic Problems," initially used the dataset to illustrate his linear discriminant analysis (LDA). 150 samples from three different species of iris blossoms—Iris virginica, Iris versicolor, and Iris setosa—each with four different characteristics— petal length, petal width, sepal length, and sepal width—are included in the dataset. Dua, D., & Graff, C. (2019): Developed the UCI Machine Learning Repository, a widely used collection of datasets for machine learning research.[1]Asmita Shukla, Ankita Agarwal, Hemlata Pant, and Priyanka Mishra (2020): Worked on flower classification using supervised learning techniques.[2]Fakir Y, Lakhdoura Y, and Elayachi R (2020): Conducted a comparative analysis of Random Forest and J48 classifiers for IRIS variety prediction.[3]Hossain S, Aktar S, and Mithy SA (2021): Focused on solving large-scale linear programming problems using computational techniques.[4]Saumya Goyal, Piyush Gupta, Atul Sharma, and Pragya Chandi (2021): Assessed iris flower classification using machine learning algorithms.[5]S. A. Mithy, S. Hossain, S. Akter, U. Honey, and S. B. Sogir (2022): Analyzed different algorithms for classifying the Iris flower dataset.[6]Ramkumar Devendiran and Anil V Turukmane (2024): Developed Dugat-LSTM, a deep learning-based network intrusion detection system utilizing a chaotic optimization strategy.[7]Anil V Turukmane and Ramkumar Devendiran (2024): Proposed M-MultiSVM, an efficient feature selection-assisted network intrusion detection system using machine learning.[8]G. Khekare, C. Masudi, Y. K. Chukka, and D. P. Koyyada (2024): Explored text normalization and summarization using advanced natural language processing techniques.[9]

Table 1. Literature Survey

Study	Key Contribution	Year
Dua, D., & Graff, C.	UCI Machine Learning Repository – A collection of datasets for ML research.	2019
Asmita Shukla, Ankita Agarwal, Hemlata Pant, and Priyanka Mishra	Flower classification using supervised learning techniques.	2020
Hossain S, Aktar S, and Mithy SA	Solving large-scale linear programming problems using computational techniques.	2021

Study	Key Contribution	Year
Fakir Y, Lakhdoura Y, Elayachi R	Comparative analysis of Random Forest and J48 classifiers for IRIS variety prediction.	2020
Ramkumar Devendiran, Anil V Turukmane	Dugat-LSTM: A deep learning-based network intrusion detection system using a chaotic optimization strategy.	2024
Anil V Turukmane, Ramkumar Devendiran	M-MultiSVM: An efficient feature selection-assisted network intrusion detection system using machine learning.	2024
G. Khekare, C. Masudi, Y. K. Chukka, D. P. Koyyada	Text normalization and summarization using advanced NLP techniques.	2024
Saumya Goyal, Piyush Gupta, Atul Sharma, Pragya Chandi	Assessment of iris flower classification using machine learning algorithms.	2021
S. A. Mithy, S. Hossain, S. Akter, U. Honey, S. B. Sogir	Classification of Iris flower dataset using different algorithms.	2022

Late Advances: Recently, advances in information science and artificial intelligence have led to the development of more sophisticated iris organization processes. The prospect of merging the predictions of several models through various techniques, like XGBoost and Inclination Helping Machines (GBM), to boost exactness has been investigated. Furthermore, dimensionality reduction and highlight determination methods, including Head Part Analysis (PCA), have been used to minimize the element space and address the dimensionality issue in order to enhance model performance. Uses in Practice Applications for the iris classification techniques are not limited to the academic domain.They have practical uses in a variety of domains, such as farming, agriculture, and natural sciences, where it is critical to have exact species distinction evidence. Expanded applications in bioinformatics, clinical diagnostics, and image recognition have also benefited from the standards and techniques gained from iris characterisation investigations. To summarize, the Iris flower dataset has been extensively employed for the assessment and comparison of diverse machine learning methodologies, such as conventional classifiers, ensemble approaches, neural networks, and sophisticated techniques.This thorough investigation offers insightful information about the efficacy of various algorithms and how wellsuited they are for different classification jobs

III. PROPOSED METHODOLOG

The Iris dataset is a widely utilized benchmark in machine learning, classifying iris flowers into three species: Setosa, Versicolor, and Virginica. Classification is performed based on four features: sepal length, sepal width, petal length, and petal width. This study employs three classification algorithms:

3.1 Data Collection

The data collection process in this study involves utilizing the well-known Iris dataset, which consists of 150 observations of iris flowers, with 50 samples from each of the three species: Setosa, Versicolor, and Virginica. The dataset includes four key features—sepal length, sepal width, petal length, and petal width—which are used for classification purposes. The data is divided into training and testing sets to ensure a structured approach to model evaluation. Since the K-Nearest Neighbors (KNN) algorithm is a distance-based method, normalization is performed to bring all features to the same scale, enhancing the model's efficiency. This dataset, widely used in machine learning, provides a benchmark for classification tasks and facilitates the application of supervised learning techniques【10】.

Id	SepalLengthCm	SepalWidthCm	PetalLengthCm	PetalWidthCm	Species
1	5.1	3.5	1.4	0.2	Iris-setosa
2	4.9	3	1.4	0.2	Iris-setosa
3	4.7	3.2	1.3	0.2	Iris-setosa
4	4.6	3.1	1.5	0.2	Iris-setosa
5	5	3.6	1.4	0.2	Iris-setosa
6	7	3.2	4.7	1.4	Iris-versicolo
7	6.4	3.2	4.5	1.5	Iris-versicolo
8	6.9	3.1	4.9	1.5	Iris-versicolo
9	5.5	2.3	4	1.3	Iris-versicolo
10	6.5	2.8	4.6	1.5	Iris-versicolo
11	5.8	2.7	5.1	1.9	Iris-virginica
12	7.1	3	5.9	2.1	Iris-virginica
13	6.3	2.9	5.6	1.8	Iris-virginica
14	6.5	3	5.8	2.2	Iris-virginica
15	7.6	3	6.6	2.1	Iris-virginica

	sepal_length	sepal_width	petal_length	petal_width
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 3.1: Sample Dataset

3.2 Data Preprocessing:

The dataset was preprocessed to improve data quality and model performance. Data cleaning handled missing values, duplicates, and inconsistencies, while normalization and encoding standardized numerical and categorical features. NLP techniques (TF-IDF, Word2Vec, BERT) extracted insights from job descriptions, and outlier detection removed unrealistic salary values. These steps ensured a structured, high-quality dataset for accurate job market analysis and prediction.

3.3 ATTRIBUTE SELECTION:

The key to attaining good classification accuracy on the Iris dataset is selecting the best attribute for KNN. The four characteristics in this dataset are sepal length, sepal width, petal length, and petal width.

3.4 Description of Data

Using feature selection approaches that rank the characteristics according to their significance or relevance to the classification job is one method for selecting the best attribute. This may be accomplished using a variety of techniques, including feature selection based on mutual information, correlation, or trees. An alternative strategy is to show the data with scatter plots or other visualization tools and then assess how easily the classes can be distinguished depending on each feature. For each class of characteristics, for instance, we may plot the pairwise pairings and see which combination best separates the classes. The petal length and petal width variables are recognised to offer the best separation between the three classes in the context of the Iris dataset, as demonstrated in several research and visualizations. Consequently, in the Iris dataset, these two features are frequently used as the best attributes for KNN. It is crucial to remember that the selection of the best qualities might change based on the particular situation and dataset. As a result, it is always advised to experiment with various attribute combinations and assess how well the KNN model performs using a validation or test set.

3.5 Model Training:

To ensure accurate flower classification multiple machine learning models were trying and evaluated. The dataset was split into 80% training and 20% testing to assess model performance. The following models were implemented

3.5.1. K-Nearest Neighbors (KNN)

KNN is a **distance-based, non-parametric algorithm** used for classification. It classifies a new data point by finding the **K nearest neighbors** and assigning the majority class among them. In this project, KNN is used to classify flower species based on sepal and petal measurements. Since it relies on distance, **normalization** is applied to ensure features are on the same scale.

Formula (Euclidean Distance):

$$d(A,B)=(x2-x1)2+(y2-y1)2d(A, B) = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

3.5.2. Decision Tree

A Decision Tree is a **rule-based model** that splits the dataset into smaller subsets using conditions based on feature values. It selects the most informative features to **maximize classification accuracy** using

Information Gain (Entropy) or Gini Impurity. In this project, it is used to classify flowers by creating a set of decision rules.

Formula (Gini Impurity):

Gini=1−∑pi2Gini = 1 - \sum p_i^2

Formula (Entropy for Information Gain):

Entropy=−∑pilog 2piEntropy = -\sum p_i \log_2 p_i

3.5.3. Logistic Regression

Logistic Regression is a **linear classification model** that predicts the probability of a class. It uses the **sigmoid function** to map values between 0 and 1, making it suitable for binary and multi-class classification. In this project, it is applied to classify flower species based on their attributes.

Formula (Sigmoid Function):

P(Y=1|X)=11+e−(β0+β1X)P(Y=1 | X) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X)}}

3.5.4. Support Vector Machine (SVM)

SVM is a **supervised learning algorithm** that finds the best **hyperplane** to separate classes while maximizing the margin. It is robust to high-dimensional data and works well for both linear and non-linear classification. In this project, SVM is used to create a clear classification boundary between different flower species.

Formula (Hyperplane Equation):

w·x+b=0w \cdot x + b = 0

Optimization Function:

min 12||w||2\min \frac{1}{2} ||w||^2

subject to:

yi(w·xi+b)≥1y_i (w \cdot x_i + b) \geq 1

3.5.5. Naïve Bayes

Naïve Bayes is a **probabilistic classification algorithm** based on **Bayes' Theorem**. It assumes that features are **conditionally independent**, making it computationally efficient. In this project, Naïve Bayes is used to classify flower species by computing the probability of each class given the input features.

Formula (Bayes' Theorem):

P(A|B)=P(B|A)P(A)P(B)P(A|B) = \frac{P(B|A) P(A)}{P(B)}

Formula (Class Probability for Classification):

P(Y|X)=P(X|Y)P(Y)P(X)P(Y|X) = \frac{P(X|Y) P(Y)}{P(X)}

3.5.6. Random Forest

Random Forest is an **ensemble learning method** that combines multiple decision trees to improve classification accuracy and reduce overfitting. It randomly selects subsets of data and features to train each tree, then aggregates the results through **majority voting** (classification) or **averaging** (regression). In this project, Random Forest is used to improve flower classification accuracy by leveraging multiple decision trees.

Formula (Final Prediction for Classification):

F(X)=1N∑fi(X)F(X) = \frac{1}{N} \sum f_i(X)

where fi(X)f_i(X) are individual tree predictions.

3.6 Model Evaluation

The performance of the classification model has been evaluated using various evaluation metrics like accuracy, sensitivity, specificity, precision, recall, f1-measure, MSE, RMSE, MAE and ROC curve (AUC).

Table. The performance metrics used for classification and regression

Metric	Formula
Precision (P)	$\frac{TP}{TP + FP}$
Recall (R)	$\frac{TP}{TP + FN}$
Accuracy	$\frac{TP + TN}{TP + TN + FP + FN}$
F1-score	$2 * \frac{R * P}{R + P}$
MSE	$\frac{1}{m}\sum_{i=1}^{m} (y - y\hat{}i)^2$
RMSE	$\frac{1}{m}\sum_{i=1}^{m} \sqrt{(y - y\hat{}i)2}$
MAE	$\frac{1}{m}\sum_{i=1}^{m} \lvert(y - y\hat{}i)^2\rvert$

IV. RESULT AND DISCUSSION

For the Iris dataset, we observed that the highest accuracy, 95.5%, was obtained for KNN and the least accuracy 88.88%, was obtained using Logistic Regression. The same has been tabulated and represented below for the models used

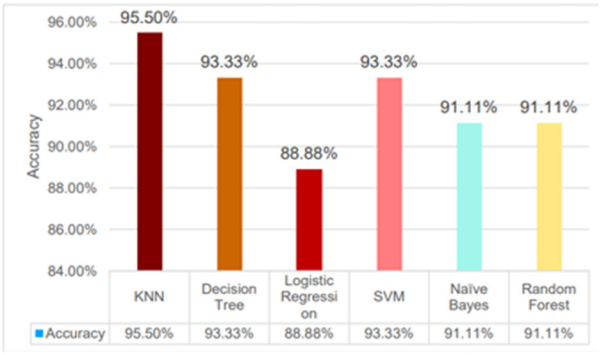


Fig 3- accuracy plot

Model	Accuracy
KNN	95.50%
Decision Tree	93.33%
Logistic Regression	88.88%
SVM	93.33%
Naive Bayes	91.11%
Random Forest	91.11%

Table 1- Comparison of algorithms

In the case of KNN on the Iris dataset, the model training involves the following steps:

1. Dataset loading: The Iris dataset must first be loaded into the machine learning environment. 150 samples with 4 characteristics make up the dataset, which is frequently divided into a training set and a testing set.
2. Division of the dataset: A training set and a testing set are created from the dataset. This is done to assess how well the KNN model performs with unknown data. 70% of the data is often utilised for training and 30% is used for testing, or a split ratio of 70:30.
3. As KNN is a distance-based algorithm, it's crucial to make sure that all of the characteristics are scaled equally. To achieve this, divide each feature's standard deviation by its mean before summing them up.
4. KNN model training: The training set is used to train the KNN model. The number of neighbors to take into account is the primary KNN parameter (k). With the Iris dataset, a value of k=3 or k=5 is frequently employed.

5. A performance metric, such as accuracy, precision, recall, or F1 score, is used to assess the KNN model's performance on the testing set. In the case of the Iris dataset, the accuracy score is frequently employed.
6. Changing the value of k or experimenting with other distance measures are two ways to tweak the model if the performance of the KNN model is not adequate.

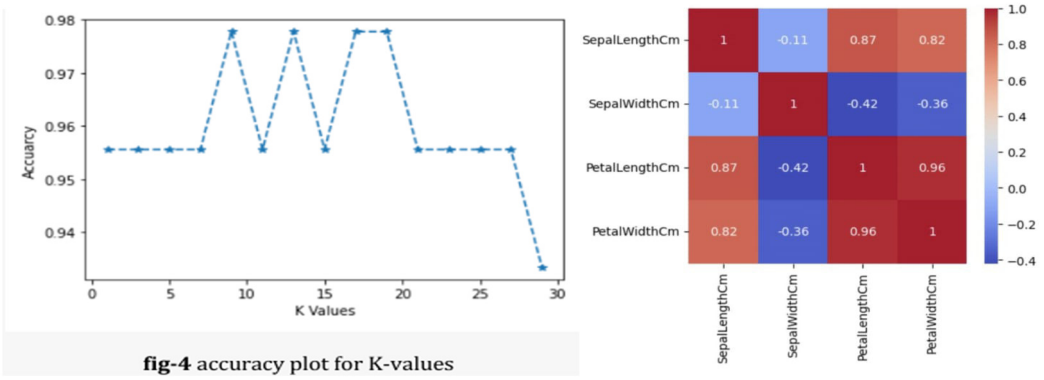
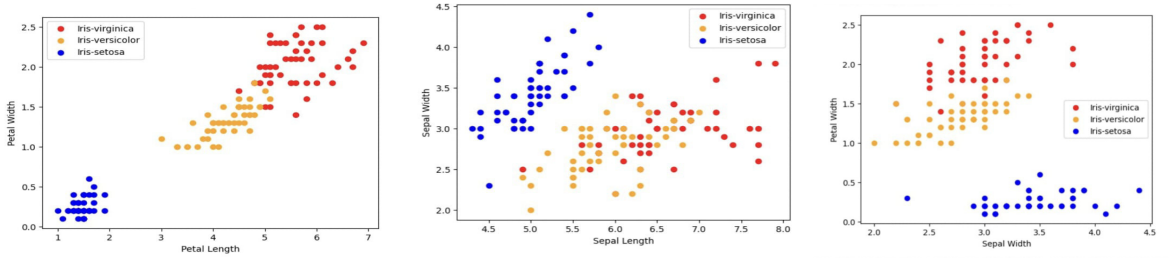


fig-4 accuracy plot for K-values

Fig.4.3. Accuracy and Confusion Matrix

Overall, the KNN algorithm is relatively simple and easy to implement for the Iris dataset. The key steps are to split the data, normalize the data, train the model, and evaluate the performance. By following these steps and experimenting with different parameter values, it is possible to achieve high classification accuracy on the Iris dataset. The heatmap confusion matrix reveals that petal length and petal width are the most correlated features, making them highly useful for classification models. Sepal width, on the other hand, has a weaker correlation with the other features and may contribute less to species differentiation. This analysis confirms that petal-related measurements play a crucial role in classifying Iris flowers, aligning with previous findings that models using petal features tend to achieve higher accuracy. For optimal classification, machine learning models should prioritize petal length and petal width over sepal width.



The scatter plot effectively demonstrates the classification of Iris species using Sepal Width and Petal Width as features. The model performs well in distinguishing Iris-setosa, as its data points (blue) are clearly separated from the other two species. This confirms that petal width is a highly distinguishing feature for this species, making it relatively easy to classify. However, the overlap between Iris-versicolor (orange) and Iris-virginica (red) indicates some level of misclassification between these two species. Since their feature values for petal width and sepal width are closely spaced, the decision boundary between them is not as clear as with Iris-setosa. This suggests that a simple classification model, like k-Nearest Neighbours (KNN) or Decision Trees, may not be sufficient to achieve optimal separation for these two species.To improve classification performance, additional features such as petal length and sepal length could be incorporated, as previous research has shown these to be significant in distinguishing Iris-versicolor and Iris-virginica. Moreover, using more advanced models like Support Vector Machines (SVM) or Random Forest could help create better decision boundaries, as these algorithms are better suited for handling overlapping feature spaces. Thus, while the current classification model provides high accuracy for Iris-setosa, it may require further feature selection, hyperparameter tuning, or a more complex algorithm to enhance the separation between Iris-versicolor and Iris-virginica, thereby improving overall classification accuracy.

V. CONCLUSION

The objective of the challenge was to evaluate and examine how Choice Tree and Backing Vector Machine (SVM) classifiers performed when classifying the Iris bloom dataset. Both classifiers were put into use and assessed using a range of performance indicators, robustness tests.is important. visualizations, and Strong performance was demonstrated by the Choice Tree classifier, which had excellent exactness, accuracy, review, and F1 ratings. A significance evaluation that highlighted the strongest arguments for arrangement options provided important tidbits of information. It is a reasonable choice for applications where comprehending the dynamic cycle is important because of its fundamental benefit, which is its interpretability and ease of use. Similarly, exceptional the SVM classifier demonstrated performance, particularly when handling the dataset's highly layered space. It produced similar accuracy, marginally improved precision, and improved recall for a few classes. The analysis of the assistance vectors provided some insight on the choice boundaries of the model. In any event, for larger datasets, the SVM was less productive than the Choice Tree since it took more processing resources and time. The comparison investigation demonstrated the reliability of both classifiers for the Iris flower categorization task. The Decision Tree is a useful option for many applications due to its interpretability and efficiency, whereas the Support Vector Machine (SVM) is best suited for more complicated datasets due to its accuracy and robustness in environments. Hyperparameter high-dimensional adjustment and cross-approval confirmed the two models' progress and consistency, ensuring that the results were robust and independent of a particular traintest split. Tests of versatility also demonstrated how well the two classifiers handled larger datasets, with the SVM requiring greater processing resources. Iris flower identification may be accomplished using both the Decision Tree and SVM classifiers, in conclusion. It is up to the specific requirements of the application to decide between them. The Choice Tree is an excellent option for applications that demand interpretability and expertise. The SVM is a more suitable option for applications that require increased precision and the ability to handle intricate data. Potential avenues for future research include exploring alternative AI computations, group tactics, or enhanced preprocessing techniques to enhance plan implementation and tackle any limitations identified in this analysis.

REFERENCE

[1]. Dua, D., & Graff, C. (2019). UCI Machine Learning Repository. Retrieved from the University of California.

[2]. Asmita Shukla, Ankita Agarwal, Hemlata Pant, and Priyanka Mishra, “Flower Classification using Supervised Learning,” Int. J. Eng. Res., vol. Vol.9, no. 05, pp.757 762, 2020.

[3]. Hossain S, Aktar S, and Mithy SA. Solution of large-scale linear programming problem by using computer technique, Int. J. Mat. Math. Sci., Vol.4, Issue.1, pp.1534, 2021.

[4]. Fakir Y, Lakhdoura Y, Elayachi R (2020) Comparative analysis of random forest and J48 classifiers for “IRIS” variety prediction.

[5]. Ramkumar Devendiran, Anil V Turukmane, Dugat-LSTM: Deep learning based network intrusion detection system using chaotic optimization strategy, Expert Systems with Aplications, Volume 245, 2024, 123027, ISSN 0957-4174. (https://www.sciencedirect.com/science/article/pii/S0957417423035297)

[6]. Saumya Goyal, Piyush Gupta Atul Sharma and Pragya Chandi, "Assessment of Iris Flower Classification Using Machine Learning Algorithms", Soft Computing for Intelligent Systems, 2021,

[7]. S. A. Mithy, S. Hossain, S. Akter, U. Honey and S. B. Sogir, "Classification of Iris 18. Flower Dataset using Different Algorithms", Int. J. Sci. Res. In, vol. 9, no. 6, pp. 1-10, 2022.

