Student name: Shashank Jha

Enrollment Number: 02219051623

Email ID: [jhashashank707@gmail.com](mailto:jhashashank707@gmail.com)

Batch : AIML – B1

Github Link: <https://github.com/shashankjha498/ML-project>

**Comparative Analysis of Machine Learning Classifiers on the Iris Dataset**

**Abstract:** This report presents a comparative analysis of five supervised machine learning models-Decision Tree, Logistic Regression, K-Nearest Neighbours, Support Vector Machine, and Neural Network-applied to the Iris dataset for multiclass classification. Each model was trained on normalized features, and their performance was evaluated using accuracy scores on a test set(outside training data). The results reveal significant variations in accuracy among the models, with most achieving high performance due to the simplicity and structure of the dataset. A bar chart visualizes the comparative accuracy, aiding in identifying the most effective model for Iris classification tasks under current preprocessing conditions.|

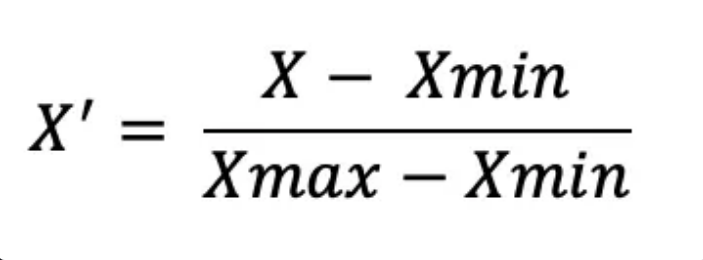
**INTRODUCTION**: The Iris dataset is among the most used benchmarking datasets for machine learning models. It comprises 150 iris flowers samples, split across three species: Setosa, Versicolor, and Virginica, for which four features have been observed per flower: sepal length, sepal width, petal length, and petal width. The features are numeric, so the dataset is ideal for classification problems. The aim of this research is to compare the performance of these machine learning classifiers, such as Decision Tree, Logistic Regression, K-Nearest Neighbors (KNN), Support Vector Machine (SVM), and Neural Network (MLP), on this data, with accuracy being the main evaluation metric.

**Proposed Methodology**: The methodology for this study involves several key steps:

1. **Data Collection** : The Iris dataset, sourced from the **UCI Machine Learning Repository**, was used for this classification task. It contains 150 samples across three flower species, with 4 features: **sepal length**, **sepal width**, **petal length**, and **petal width**.

2. **Data Preprocessing**:

**Feature Scaling**: Since many machine learning algorithms are sensitive to the scale of the input data, the features were scaled using **MinMax scaling**. This normalization ensures all features are within a consistent range [0, 1], improving model performance and convergence.



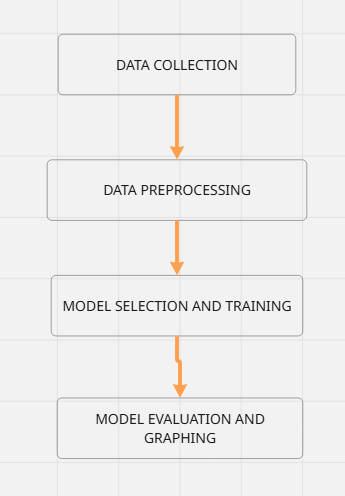
**Train-Test Split**: The data was split into 80% for training and 20% for testing. To prevent any bias, the split was done in a stratified manner, maintaining the same distribution of classes in both the training and testing sets.

**Label Encoding**: The target labels (flower species) were encoded numerically, where each class was assigned a unique integer value.

**3**. **Model Selection and Training**:  
 The following models were selected for comparison:

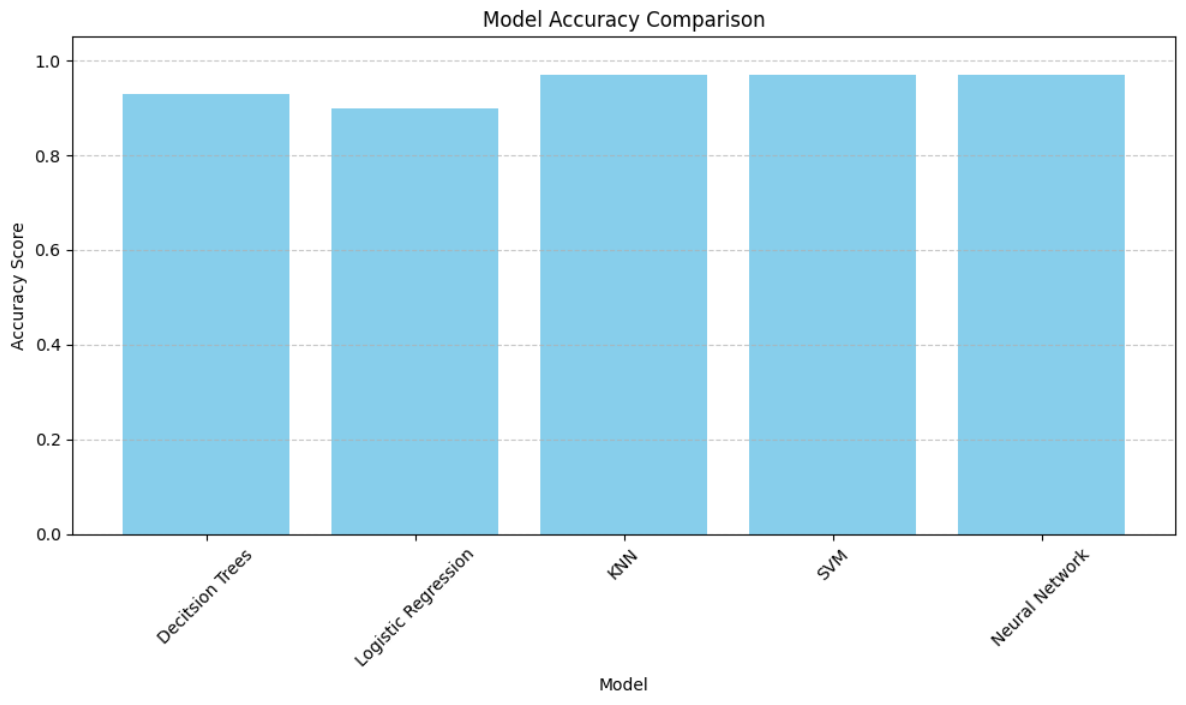
* **Decision Tree**: A simple yet interpretable model, which splits the data based on feature values to create a tree structure.
* **Logistic Regression**: A linear model often used for binary classification but adapted for multiclass problems using the softmax function.
* **K-Nearest Neighbors (KNN)**: A non-parametric method that classifies a sample based on the majority class among its k nearest neighbors in the feature space.
* **Support Vector Machine (SVM)**: A powerful classifier that works by finding the hyperplane that best separates the classes in the feature space.
* **Neural Network (MLP)**: A multi-layer perceptron (feedforward neural network) that can capture complex non-linear relationships between features.

4. **Evaluation**:  
The performance of each model was evaluated based on **accuracy**, which measures the proportion of correctly classified samples in the test set. The results were visualized using a bar chart to facilitate comparison of the models' effectiveness.

****

**RESULTS:**

|  |  |
| --- | --- |
| **MODEL** | **ACCURACY** |
| **Decision Tree** | **93%** |
| **Logistic Regression** | **90%** |
| **K- Nearest Neighbours** | **97%** |
| **Support vector Machine** | **97%** |
| **Neural Network** | **97%** |

****

**Conclusion:** The comparative analysis of five machine learning models on the Iris dataset demonstrated high classification accuracy across all approaches. The K-Nearest Neighbors, Support Vector Machine, and Neural Network models each achieved an accuracy of 97%, outperforming both the Decision Tree (93%) and Logistic Regression (90%). These results suggest that while simpler models like Decision Tree and Logistic Regression offer decent performance, more robust methods like KNN, SVM, and MLP are better suited for capturing the non-linear class boundaries present in the dataset. Given their comparable performance, model selection can be further guided by considerations such as training time, interpretability, or application context.