# Iris Species Classification Using Morphological Features

## Dataset Source

The dataset used in this project was obtained from **Kaggle** and accessed via the following Google Drive link:
https://drive.google.com/file/d/1VH9Hfj3cNwxlZw-QNX_CTxXoECjybaAg/view?usp=sharing

## 1. Problem Overview and Motivation

The objective of this project is to develop supervised machine learning models to classify iris flowers into three species: **Iris-setosa**, **Iris-versicolor**, and **Iris-virginica**. The classification is performed using morphological features such as sepal length, sepal width, petal length, and petal width.

This is a multi-class classification problem and is widely used as a benchmark dataset in machine learning due to its simplicity, interpretability, and presence of both linearly separable and overlapping classes. The problem helps in understanding the behavior of different classification algorithms and their ability to generalize.

## 2. Dataset Description and Preprocessing

The Iris dataset consists of **150 samples** and **4 numerical features**:

- Sepal Length (cm)
- Sepal Width (cm)
- Petal Length (cm)
- Petal Width (cm)

The target variable is the categorical feature **Species**, which contains three classes.
The dataset was preprocessed using the following steps:

- Selection of relevant morphological features
- Stratified train–test split with an 80:20 ratio to preserve class distribution

## 3. Mathematical Formulation of the Model

The classification task can be mathematically represented as learning a function:

$f(X) \rightarrow Y$

where $(X \in R^4)$ represents the feature vector and (Y) represents the class label.
Linear models attempt to find linear decision boundaries, while non-linear models learn complex boundaries capable of separating overlapping classes.

## 4. Loss Function and Training Process

- **Logistic Regression** minimizes the categorical cross-entropy loss function using iterative optimization.
- **K-Nearest Neighbors (KNN)** classifies samples based on distance-based majority voting.

- **Decision Tree** classifiers recursively split the feature space by minimizing impurity measures such as the Gini Index.

Each model is trained using the training dataset and evaluated on the unseen test dataset.

# 5. Model Architecture and Justification

Three supervised classification models were implemented:

- **Logistic Regression** (Linear model): Chosen as a baseline due to its simplicity and interpretability.
- **K-Nearest Neighbors** (Non-linear model): Used to capture local patterns in the feature space.
- **Decision Tree** (Non-linear model): Used to learn hierarchical decision rules.

# 6. Evaluation Methodology and Results

The models were evaluated using the following metrics:

- Accuracy
- Precision
- Recall
- F1-score
- Confusion Matrix

The observed accuracies were:

- **Logistic Regression:** 95.56%
- **K-Nearest Neighbors:** 95.56%
- **Decision Tree:** 100%

These results indicate strong classification performance across all models.

# 7. Interpretation of Results and Error Analysis

The confusion matrix analysis shows that **Iris-setosa** is classified with near-perfect accuracy due to its clear separability from the other species. Most misclassifications occur between **Iris-versicolor** and **Iris-virginica**, which have overlapping petal measurements.

Feature scaling resulted in a slight decrease in accuracy for some models, which is expected as scaling removes feature dominance and leads to more balanced and generalizable learning. Non-linear models demonstrated better performance in handling overlapping class distributions.

# 8. Limitations and Future Improvements

The dataset is relatively small, which limits the generalization capability of the models. Future improvements could include:

- Hyperparameter tuning
- Cross-validation
- Ensemble methods such as Random Forests

- Application to larger and more diverse datasets

# Conclusion

This project demonstrates the effectiveness of supervised machine learning techniques in classifying iris species using morphological features. The results show that petal-based features are highly discriminative and that non-linear models perform better when class boundaries overlap.

_BY: Anushri Maheshwari (250041005)