**Data Cleaning and Preprocessing**

The first step was to load the Iris dataset from the CSV file. I then examined the dataset using the `data.head()`, `data.shape`, and `data.info()` functions to get an understanding of the data.

The dataset contains 150 rows and 5 columns - Sepal\_Length, Sepal\_Width, Petal\_Length, Petal\_Width, and Class. All columns except Class are numeric. The Class column contains 3 unique values - Iris-setosa, Iris-versicolor, and Iris-virginica.

There were no missing values in the dataset, so no additional data cleaning was required. However, I did perform some standard preprocessing steps:

1. **Encoding the Target Variable**: Since the Class column is categorical, I encoded it using label encoding.

2. **Scaling the Features**: I scaled the numeric features (Sepal\_Length, Sepal\_Width, Petal\_Length, Petal\_Width) using StandardScaler to ensure they are on a similar scale.

3. **Splitting the Data**: I split the dataset into training and testing sets using the `train\_test\_split` function from scikit-learn.

**Feature Selection**

For this problem, all the available features (Sepal\_Length, Sepal\_Width, Petal\_Length, Petal\_Width) are relevant and important for classifying the iris species. These four morphological features of the iris plant are well-known to be highly predictive of the iris species. Therefore, I decided to use all the available features in my analysis.

**Machine Learning Algorithm**

We evaluate multiple algorithms: K-Nearest Neighbors (KNN), Random Forest, Logistic Regression, and Support Vector Machine (SVM).

**Model Evaluation**

I evaluated the performance of the Random Forest Classifier model using the following metrics:

1. **Accuracy Score:** This measures the overall accuracy of the model in correctly classifying the iris species.

2. **Precision, Recall, and F1-Score:** These metrics provide a more nuanced understanding of the model's performance, especially when dealing with imbalanced classes.

3. **Confusion Matrix:** The confusion matrix helps visualize the model's performance and identify any specific misclassifications.

**The results of the model evaluation are as follows:**

A black screen with white numbers

Description automatically generated

**Challenges and Overcome**

The main challenge in this project was the relatively small size of the dataset (150 samples). While the Iris dataset is a well-known and widely used benchmark, the limited number of samples can make it challenging to train more complex models and assess their generalization performance.

To mitigate this challenge, I focused on using a simpler and more robust algorithm (Random Forest) that is less prone to overfitting on small datasets. Additionally, I carefully split the data into training and testing sets to ensure a reliable evaluation of the model's performance.

Overall, the Iris dataset provided a great opportunity to demonstrate my data preprocessing, feature selection, and model selection skills. The excellent performance of the Random Forest Classifier on this dataset suggests that it is a suitable choice for this type of classification problem.