Final Report: Logistic Regression vs Decision Tree vs KNN on the Iris Dataset

Introduction

This report presents the implementation and evaluation of three machine learning classification algorithms – Logistic Regression, Decision Tree, and K-Nearest Neighbors (KNN) – using the Iris dataset. The dataset consists of 150 rows, with four feature columns representing sepal and petal measurements of the iris plant, and a target column indicating the species (Setosa, Versicolor, Virginica). The data is balanced, with 50 samples per class, and contains no missing or inconsistent values.

Dataset Insights

- **Features**: Four numerical measurements (sepal length, sepal width, petal length, petal width).
- **Target**: Three species of Iris plants (Setosa, Versicolor, Virginica).
- **Balance**: Equal representation of 50 samples for each species.
- **Data Quality**: No missing data, no outliers, and no inconsistencies.

Model Training and Evaluation

Each model was trained on 80% of the dataset and tested on the remaining 20%. Key metrics used for evaluation include accuracy, precision, recall, and F1-score. These metrics remain consistent due to the balanced nature of the dataset.

Metric	Logistic Regression	Decision Tree	K-Nearest Neighbors
Train Accuracy	0.9583	**0.9833**	0.9583
Test Accuracy	0.9333	**0.9667**	**0.9667**
Train F1 Score	0.9583	**0.9833**	0.9583
Test F1 Score	0.9333	**0.9666**	**0.9666**
Train Precision	0.9585	**0.9841**	0.9600
Test Precision	0.9333	**0.9697**	**0.9697**

Train Recall	0.9583	**0.9833**	0.9583
Test Recall	0.9333	**0.9667**	**0.9667**

Insights and Observations

1. **Logistic Regression**: Logistic Regression performed well with a train and test accuracy of 95.83% and 93.33%, respectively. It demonstrated consistent performance across F1-score, precision, and recall.
2. **Decision Tree**: The Decision Tree outperformed Logistic Regression slightly, achieving the highest training accuracy of 98.33% and a test accuracy of 96.67%. Its F1-score, precision, and recall also matched its strong accuracy.
3. **K-Nearest Neighbors (KNN)**: KNN matched Decision Tree in test accuracy (96.67%) and F1-score (96.66%), indicating its effectiveness in this balanced dataset. However, its training performance was slightly below Decision Tree's.
4. **Balanced Dataset Impact**: Since the target variable is perfectly balanced, there is no disparity between accuracy, F1-score, precision, and recall. This allows for straightforward comparisons among models without concerns of bias.

Conclusion

Based on the evaluation metrics, the **Decision Tree** classifier exhibited the best overall performance, with the highest train and test accuracy, as well as precision and recall. While KNN matched Decision Tree in test metrics, its training performance was marginally lower, indicating that Decision Tree is better at capturing the underlying patterns of the Iris dataset.

Logistic Regression performed well but slightly lagged behind the other two models in terms of accuracy and generalization. The results affirm that both Decision Tree and KNN are excellent choices for balanced datasets, but Decision Tree's superior performance and interpretability make it the top choice for this task.