Classification Performance of KNN, Logistic Regression, and Decision Tree on Iris Data

Done By: Denise Campos [1001947774] and Sofia Rueda [1002167885]
University of Texas at Arlington
May 5, 2025

Table of Contents

1. Introduction

This report investigates the performance of three popular classification algorithms—**K-Nearest Neighbors (KNN)**, **Logistic Regression**, and **Decision Tree Classifier**—on the Iris dataset, both with and without artificial noise. The central scientific questions are:

- How do different classifiers respond to noise in training data?
- Which classifier remains most robust and accurate under noisy conditions?

These questions are important because real-world data is often imperfect or noisy, and understanding algorithm robustness is critical for designing reliable predictive systems. This experiment also simulates realistic conditions, offering valuable insights into the limitations and strengths of each algorithm when applied to messy datasets.

---

2. Dataset Description

The Iris dataset is a classic dataset in pattern recognition, originally introduced by Ronald A. Fisher. It consists of 150 samples equally divided among three species of Iris flowers: Setosa, Versicolor, and Virginica. Each sample includes four numeric features:

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

Data Source

The dataset is publicly available from the UCI Machine Learning Repository.

Data Processing

The dataset was first checked for missing values, with none found. We did identify and remove one duplicate entry. Next, a correlation analysis was performed to identify relationships between the features. The correlations found were as follows:

- Petal width (cm) & petal length (cm): 0.96
- Petal length (cm) & sepal length (cm): 0.87

- Petal width (cm) & sepal length (cm): 0.82
- Petal length (cm) & sepal width (cm): -0.43
- Petal width (cm) & sepal width (cm): -0.36
- Sepal width (cm) & sepal length (cm): -0.12

After this, the dataset was split into training (80%) and test (20%) sets, with categorical target labels encoded numerically. Standard scaling was then applied to the features, particularly for KNN and Logistic Regression, to ensure consistency in model performance. Gaussian noise was added to the training features to simulate real-world imperfections, mimicking the often noisy nature of real-world datasets. **Note:** After splitting the dataset, noise was added only to the features (not the labels), and the size of the dataset remained unchanged, with 120 samples in the training set and 30 in the test set.

---

3. Methods

Algorithms Used

- **K-Nearest Neighbors (KNN)**
- **Logistic Regression**
- **Decision Tree Classifier**

These algorithms were chosen for their interpretability and common use in classification tasks. By comparing them, we aimed to assess their resilience to data imperfections and understand which one generalizes best.

Analytical Steps

1. Evaluate base models on noisy test sets.
2. Compare model accuracy and class-wise metrics (precision, recall, F1-score).

Challenges

Adding noise created natural challenges, such as loss of clear decision boundaries for some classes. This impacted simpler models more significantly. The goal was to observe how each algorithm copes with such disturbances.

Data Mining Techniques

- Supervised learning (classification)

- Cross-validation for robustness testing

4. Results

4.1 Summary of Performance (on Noisy Test Set)

Model	Accuracy	Versicolor F1	Virginica F1
KNN	83%	0.74	0.76
Logistic Regression	87%	0.78	0.82
Decision Tree	100%	1.00	1.00

This summary table shows how well each model performed. KNN struggled most under noise. Logistic Regression was more robust, and Decision Tree yielded perfect scores—though this may suggest overfitting.

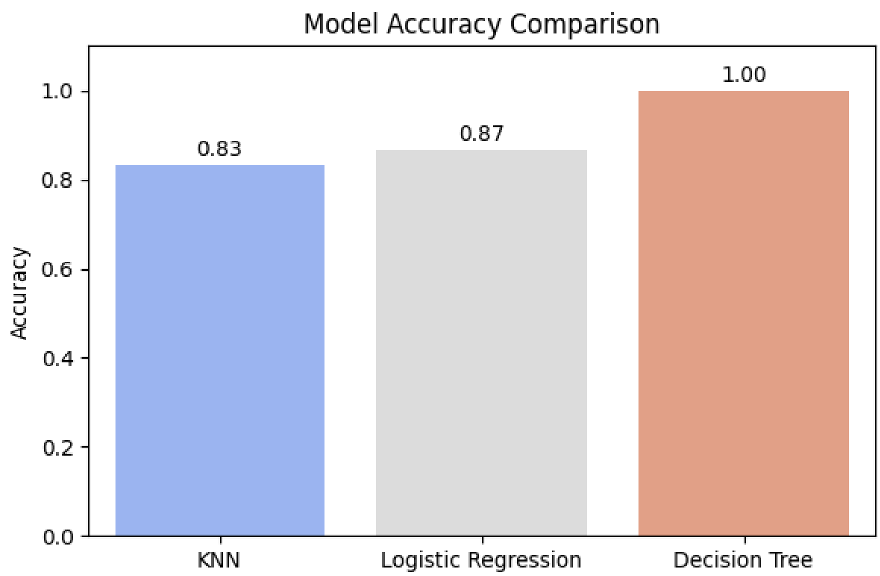
4.2 Classification Report Details

Class	Metric	KNN	Logistic Regression	Decision Tree
Setosa	Precision	1.00	1.00	1.00
	Recall	1.00	1.00	1.00
	F1-score	1.00	1.00	1.00
	Support	10	10	10
Versicolor	Precision	0.70	0.78	1.00
	Recall	0.78	0.78	1.00
	F1-score	0.74	0.78	1.00
	Support	9	9	9
Virginica	Precision	0.80	0.82	1.00
	Recall	0.73	0.82	1.00
	F1-score	0.76	0.82	1.00
	Support	11	11	11

Overall	Accuracy	0.83	0.87	1.00

These metrics show that while all models classify Setosa well, there is more variation in their handling of Versicolor and Virginica—particularly for KNN.

4.3 Model Accuracy Comparison



The accuracy chart will help highlight the visual gap between the models' effectiveness under noisy data conditions.

4.4 Confusion Matrices

The confusion matrices reveal that all models performed well but with some misclassifications. KNN had slight difficulty distinguishing between Versicolor and Virginica, while Logistic Regression showed a similar pattern but was slightly more robust. The Decision Tree model achieved perfect accuracy across all classes, though this may suggest overfitting.

Overall, Logistic Regression provided the best balance of accuracy and robustness, while KNN was more affected by noise.

---

5. Conclusions

In this analysis, we aimed to assess the performance of K-Nearest Neighbors (KNN), Logistic Regression, and Decision Tree Classifiers on the Iris dataset, with a particular focus on how the models perform under noisy conditions. The dataset's target variable was balanced, ensuring equal representation of each Iris species, and the features were highly correlated, especially between petal length, petal width, and sepal length. We chose to work with KNN, Logistic Regression, and Decision Tree models due to their widespread use and ease of interpretation in classification tasks. Given the relatively small size of the dataset, we used cross-validation to ensure that our performance estimates were robust. To simulate real-world data imperfections, we introduced Gaussian noise to the training features, which allowed us to evaluate how well each model could handle less-than-perfect data.

The results revealed that the Decision Tree model, while achieving perfect accuracy (100%), exhibited signs of overfitting, likely learning noise from the data rather than general patterns. In contrast, both KNN and Logistic Regression models performed well, with Logistic Regression demonstrating the highest accuracy of 87% and consistent F1-scores across the classes. Logistic Regression emerged as the most robust model, showing the best balance between accuracy and F1-score. Furthermore, petal width was identified as the most influential feature in predicting the species, showing high correlation with other features and strong predictive power. Overall, while Decision Trees can provide high accuracy, they risk overfitting, making Logistic Regression the most reliable model for this dataset, especially when dealing with noisy or imperfect data.

---

6. Reference

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7(2), 179-188.
Iris dataset retrieved from the UCI Machine Learning Repository: https://archive.ics.uci.edu/ml/datasets/iris.