Iris Flower Classification using K-Nearest Neighbors (KNN)

Author: *Mridul Goyal*

---

1. Problem Statement

The objective of this project is to build a supervised machine learning model capable of classifying Iris flowers into three species—**Setosa, Versicolor, and Virginica**—based on their physical measurements. This is formulated as a **multiclass classification problem**, where the goal is to learn decision boundaries that accurately separate the species using numerical features.

This project is designed as a **personal portfolio project** and is intended to demonstrate end-to-end machine learning workflow, including data understanding, visualization, model selection, evaluation, and interpretation.

---

2. Dataset Overview

The Iris dataset is a well-known benchmark dataset commonly used for learning classification techniques.

- **Total samples:** 150
- **Features:**
  - Sepal Length
  - Sepal Width
  - Petal Length
  - Petal Width
- **Target variable:** Iris species (Setosa, Versicolor, Virginica)
- **Data quality:**

    - No missing values
    - Clean and structured
    - Balanced class distribution

The dataset is small and idealized, making it suitable for conceptual understanding and algorithm behavior analysis.

---

3. Exploratory Data Analysis (EDA)

Exploratory analysis was performed to understand feature behavior and class separability before model selection.

- A **pairplot visualization** was used to analyze relationships between features.
- It was observed that **petal length and petal width provide strong visual separation** among species.
- Sepal-based features showed significant overlap and comparatively weaker discriminative power.
- Setosa formed a completely isolated cluster, while Versicolor and Virginica exhibited partial overlap.

Key insight:
Feature dominance analysis indicated that petal-based features play a crucial role in classification performance, guiding the subsequent modeling decisions.

---

4. Model Selection and Justification

The **K-Nearest Neighbors (KNN)** algorithm was selected as the primary model based on the following considerations:

- The dataset exhibits **cluster-like structure**, which aligns well with distance-based methods.
- KNN is non-parametric and adapts naturally to data geometry.
- Visualization results suggested that local neighborhood voting would be effective.

Since KNN relies on distance calculations, **feature scaling** was applied to ensure equal contribution from all features.

Choice of K:

- Multiple K values were tested.
- Lower and moderate K values performed optimally.
- Very high K values introduced slight divergence due to **decision boundary over-smoothening**.

A balanced K value was chosen to maintain generalization without overfitting.

---

5. Model Evaluation

The model was evaluated using **accuracy and confusion matrix analysis**.

- The model achieved **100% accuracy (1.0)** on the test set.
- A confusion matrix heatmap confirmed that predictions were **class-wise consistent**, with no systematic misclassification.
- All three classes were predicted correctly without dominance bias.

Despite the perfect accuracy, additional checks were performed across different K values to ensure that the result was not a consequence of overfitting. Minor performance divergence at higher K values was observed, which is expected behavior for KNN due to excessive smoothing.

Conclusion from evaluation:
The model performs accurately while maintaining stability, indicating genuine learning rather than memorization.

---

6. Final Conclusion

The KNN classifier successfully learned the underlying structure of the Iris dataset and achieved excellent classification performance. The results validate the insights obtained during exploratory analysis, particularly the dominance of petal-based features. While the dataset is idealized, the project effectively demonstrates a complete and correct machine learning workflow.

7. Limitations

- The dataset is **small and noise-free**, which does not reflect real-world data complexity.
- The absence of outliers and bias limits generalization conclusions.
- KNN performance showed slight degradation for larger K values due to excessive smoothing.
- Results obtained on this dataset should not be directly extrapolated to real-world scenarios.

8. Future Scope

- Apply cross-validation to further validate stability.
- Compare KNN performance with models such as Logistic Regression, SVM, and Random Forest.
- Test the approach on **larger, noisier, real-world datasets**.
- Enrich data with real-world biases to better simulate production-level conditions.

Overall Verdict

This project demonstrates a strong understanding of classification concepts, exploratory analysis, algorithm behavior, and evaluation methodology. While the dataset is introductory in nature, the analytical approach and justification align with industry-level best practices.