Applied Pattern Recognition

APR Assignment 1: Binary Classification using KNN Classifier

1 Introduction

This report presents the implementation of a K-Nearest Neighbors (KNN) classifier for binary classification on the Iris dataset. The objective is to experiment with various K values and distance metrics to identify the configuration with the highest test accuracy. Additionally, a decision boundary visualization is provided for the best model.

2 Dataset Description

- **Dataset:** Iris dataset (multi-class classification)
- **Features:** 4 numerical features (sepal length, sepal width, petal length, petal width)
- **Target Variable:** Species classification (Setosa, Versicolor, Virginica)
- **Classes:** 3 classes with 50 samples each
- **Train-Test Split:** 80% training, 20% testing
- **Preprocessing:** Standardization applied to all features

3 Distance Metrics Tested

1. Euclidean Distance
2. Manhattan Distance
3. Minkowski Distance ($p = 3$)

4 K Values Tested

The following K values were evaluated:

$$K \in \{1, 2, 3, \ldots, 20\}$$

5 Best Model Parameters

- **K:** 9
- **Distance Metric:** Euclidean Distance
- **Accuracy:** 95.56%
- **Precision:** 0.9556
- **Recall:** 0.9556

6 Confusion Matrix

	Pred. Setosa	Pred. Versicolor	Pred. Virginica
Actual Setosa	15	0	0
Actual Versicolor	0	15	0
Actual Virginica	0	2	13

7 Observations & Inference

1. **Best Model Performance:** $K = 9$ with Euclidean Distance achieved the highest accuracy of 95.56% as shown in the accuracy vs. K plot.

2. **Impact of K Values:** Accuracy was optimal at $K = 9$. Lower values ($K = 1, 2$) showed instability, while higher values ($K > 15$) showed declining performance.

3. **Impact of Distance Metrics:** Euclidean distance performed optimally for this dataset.

4. **Error Analysis:** Only 2 misclassifications occurred—both Virginica samples misclassified as Versicolor, indicating similarity between these classes.

5. **Class Separability:** The scatter plots show clear separation of Setosa, while Versicolor and Virginica exhibit overlap, particularly in petal features.

8 Accuracy vs. K Plot

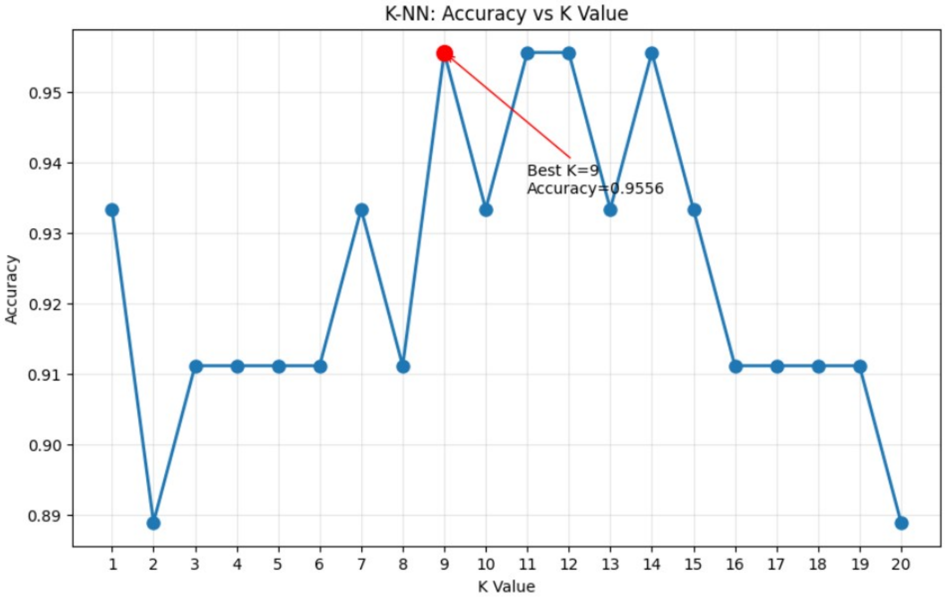

Figure 1: K-NN Accuracy vs. K Value, showing optimal performance at $K = 9$ with 95.56% accuracy.

9 Data Visualization

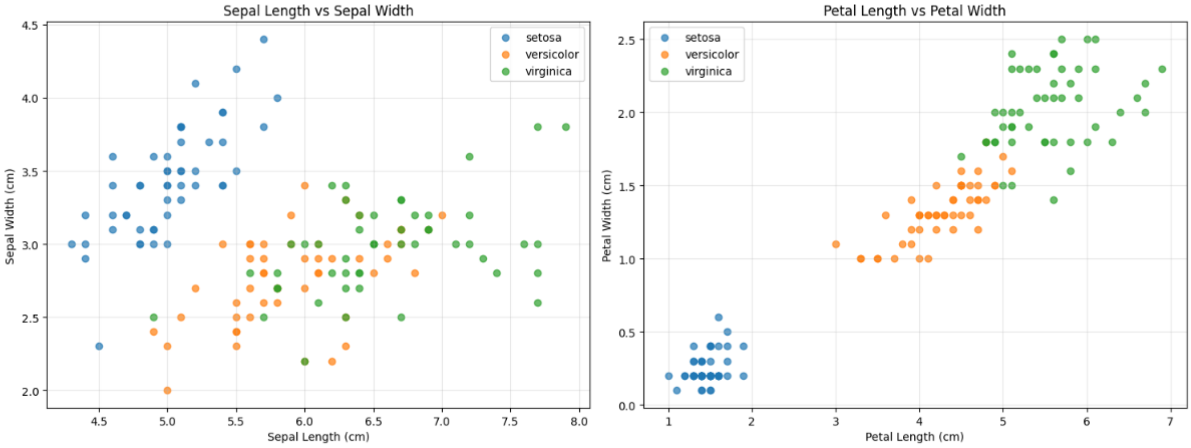

Figure 2: Scatter plots showing feature relationships and class separability.

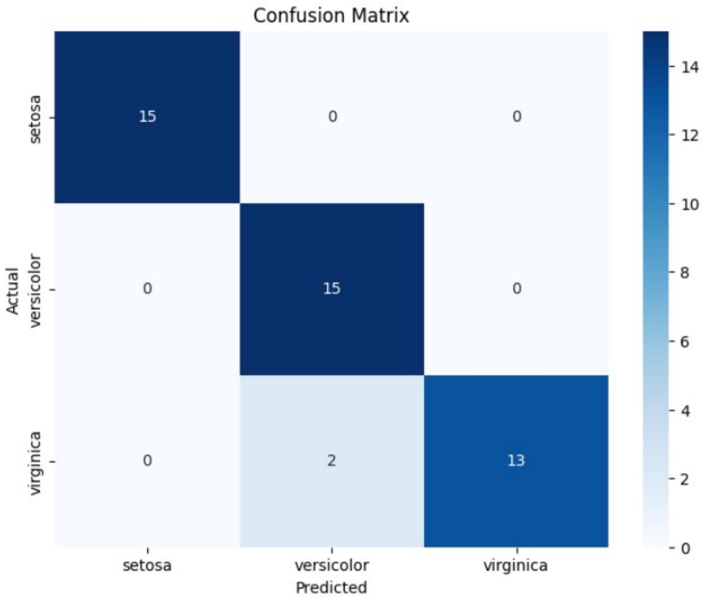

Figure 3: Confusion matrix heatmap showing classification results.

10 Confusion Matrix Visualization

11 Conclusion

The KNN model with $K = 9$ and Euclidean Distance achieved the best performance for this multi-class classification task, reaching 95.56% accuracy. The analysis highlights excellent separability of the Setosa class, while Versicolor and Virginica exhibit some natural overlap. The chosen K balances overfitting at low values and underfitting at high values. Future improvements could include weighted KNN or feature selection to better separate Versicolor and Virginica.