Lab Report: K-Nearest Neighbors from Scratch

Anushka, 23/CS/071

October 2025

1 Introduction

This experiment involves implementing the K-Nearest Neighbors (KNN) algorithm from scratch using Python. The objective is to understand the working of KNN, perform data preprocessing, visualize the dataset, train the classifier, and evaluate its performance on the Iris dataset. Additionally, hyperparameter tuning for different values of k is performed to analyze model performance.

2 Exploratory Data Analysis (EDA)

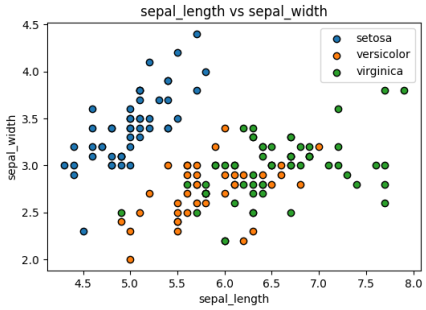
2.1 Scatter Plots

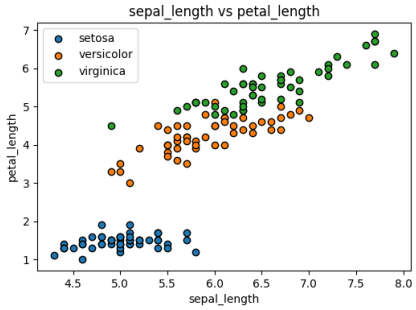
The dataset was visualized using scatter plots for every unique pair of features. These plots help in understanding the relationships between features and class separability.

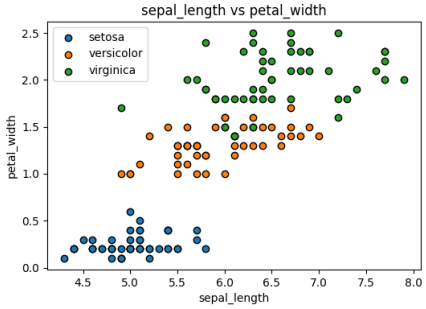
Analysis:

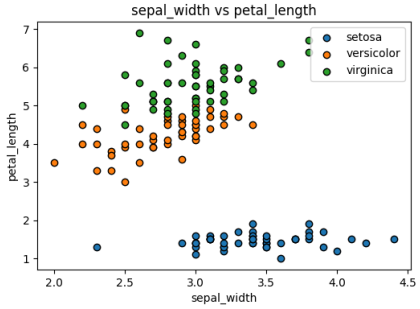
- Best feature pair for class separation: Petal length vs Petal Width
- Classes that are inherently easier to distinguish: setosa

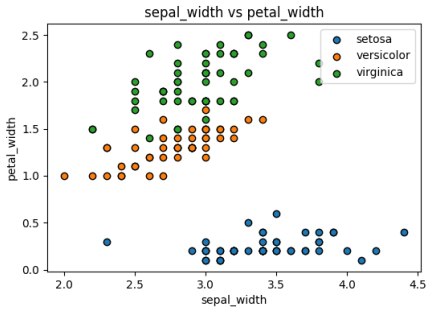
Feature Pair Plots

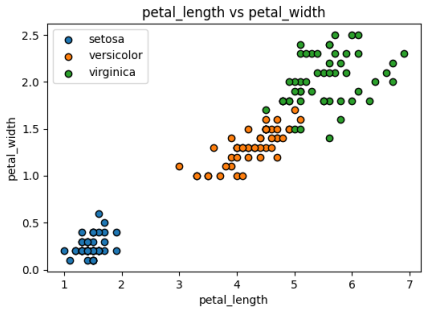



sepal_length vs petal_length
setosa
versicolor
virginica
petal_length
sepal_length


sepal_length vs petal_width
setosa
versicolor
virginica
petal_width
sepal_length


sepal_width vs petal_length
setosa
versicolor
virginica
petal_length
sepal_width


sepal_width vs petal_width
setosa
versicolor
virginica
petal_width
sepal_width


petal_length vs petal_width
setosa
versicolor
virginica
petal_width
petal_length

3 Model Implementation

3.1 KNN Classifier

The KNN classifier was implemented with a customizable number of neighbors k. The Euclidean distance metric was used to compute the distance between samples. The classifier was trained using the training set and predictions were made on the test set.

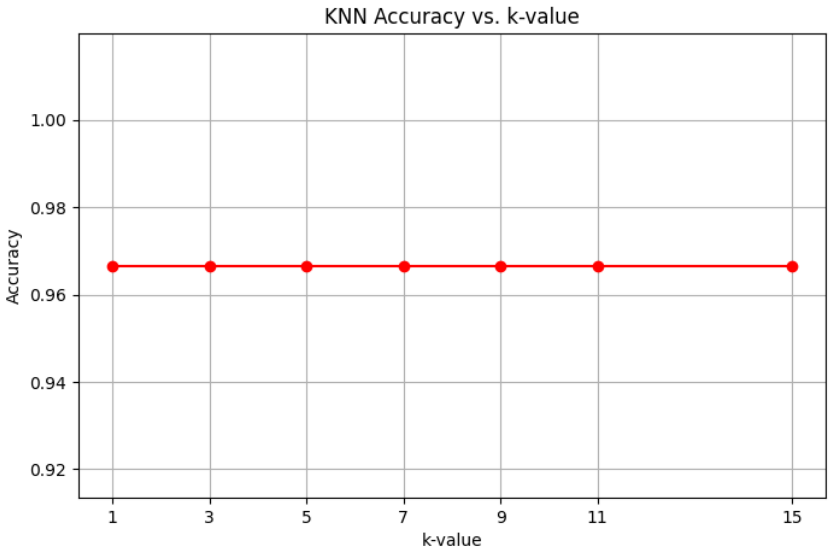
3.2 Performance on Iris Dataset

- Number of test samples: 30
- Number of correct predictions: 29
- Final classification accuracy (with $k = 3$): 96.67

4 Hyperparameter Tuning and Analysis

The model was trained for multiple k values: [1, 3, 5, 7, 9, 11, 15]. Accuracy for each k was recorded.

4.1 Accuracy vs. k-value Plot



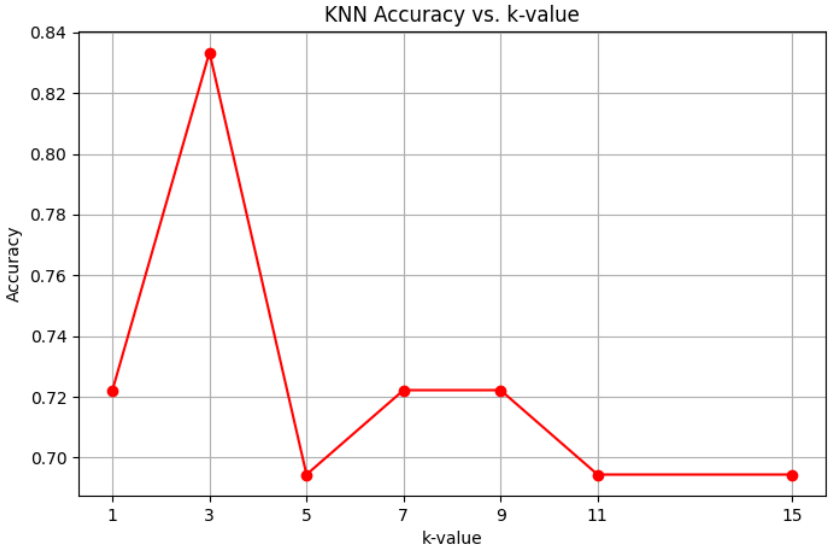
4.2 Analysis of k-values

- Best k value: 1, 3, 5, 7, 9, 11, 15 (all yielded the same accuracy in this experiment)
- Explanation: The dataset is small and classes are well-separated, so different values of k all resulted in the same performance. In general, very small k can overfit to noisy samples (high variance), while very large k can oversmooth the decision boundary (high bias), reducing accuracy. Here, the high separability of the classes makes the choice of k less critical.

5 Wine Dataset Performance

- Best k value: 3
- Best accuracy: 83.33

Accuracy vs. k-value Plot



6 Conclusion

Summarize your key learnings, observations, and any challenges faced during the experiment:

This experiment implemented KNN from scratch. Petal length and width best separate classes. The choice of k affects performance, and the exercise highlighted distance metrics and train-test splitting.