Experiment 6: Implementation of the K-Nearest Neighbours (KNN) Algorithm from Scratch

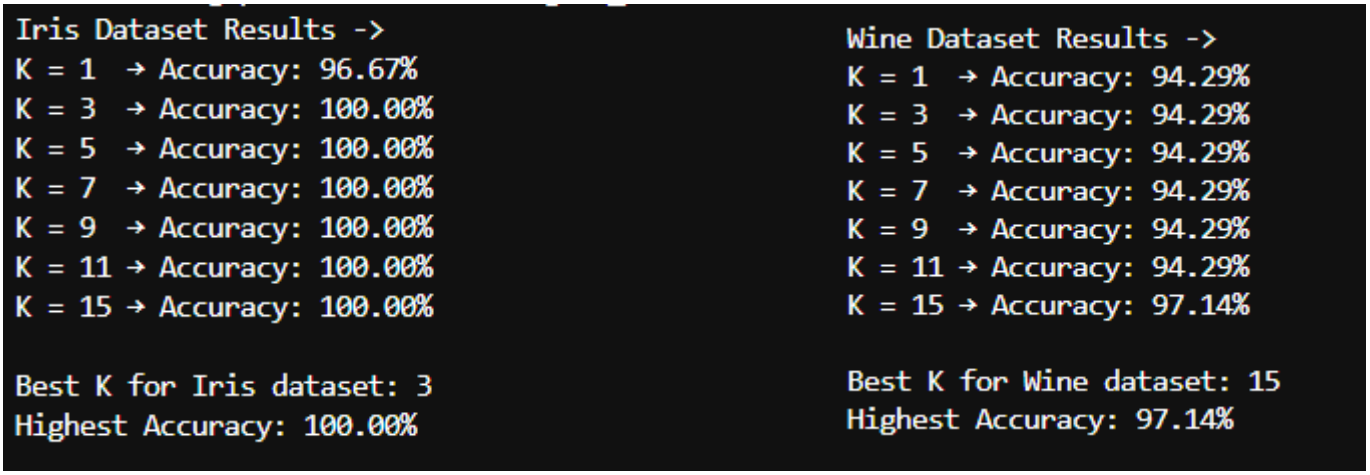
1. EDA Analysis :

For the Iris dataset, the feature pair Petal Length vs Petal Width provides the best class separation. 'Iris-setosa' is the easiest class to distinguish, as its values form a completely separate cluster. The other two species ('versicolor' and 'virginica') show partial overlap but remain largely separable. For the Wine dataset, the pair Flavanoids vs Color Intensity shows the best distinction between classes. Class 1 wines (high in Flavanoids) are more easily identifiable, while Classes 2 and 3 overlap slightly.

2. Classification Accuracy :

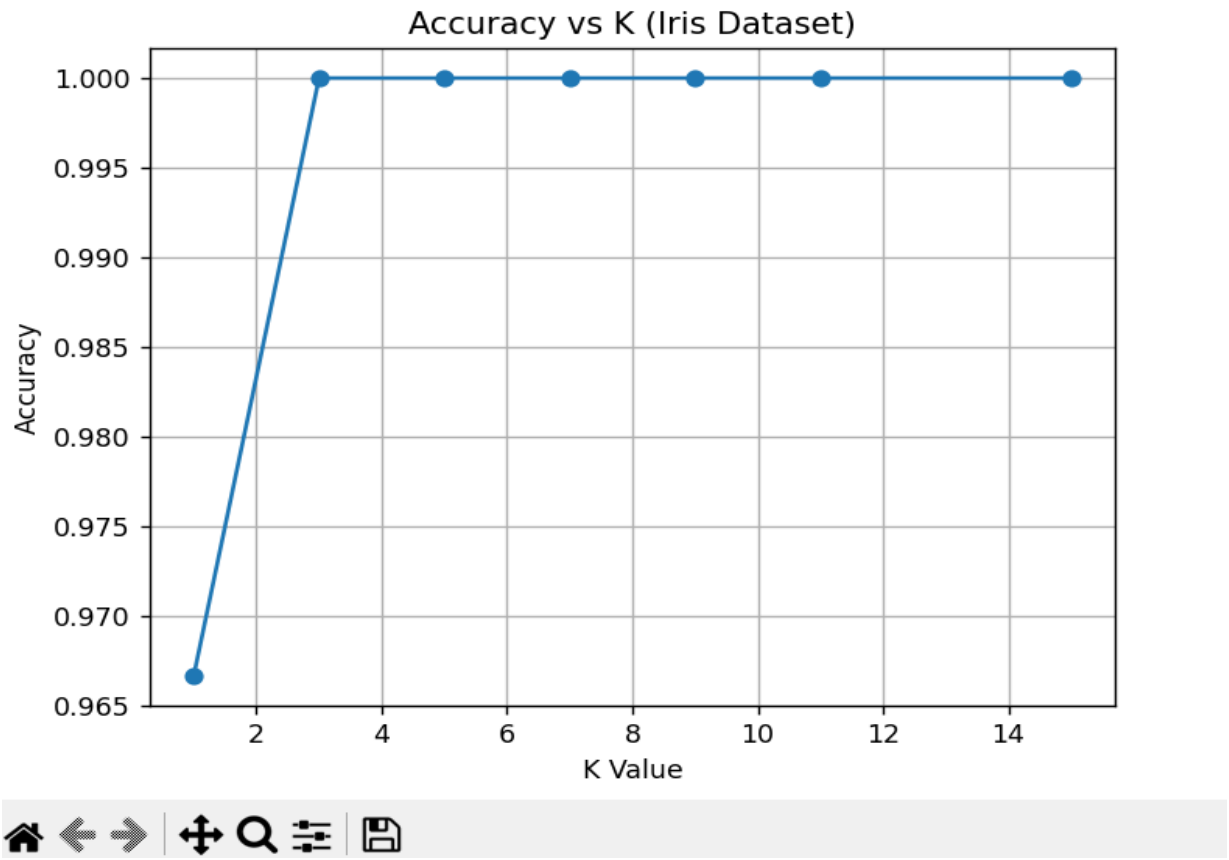
The classification accuracy is calculated using the formula:
Accuracy =
(Number of Correct Predictions / Total Number of Predictions) × 100

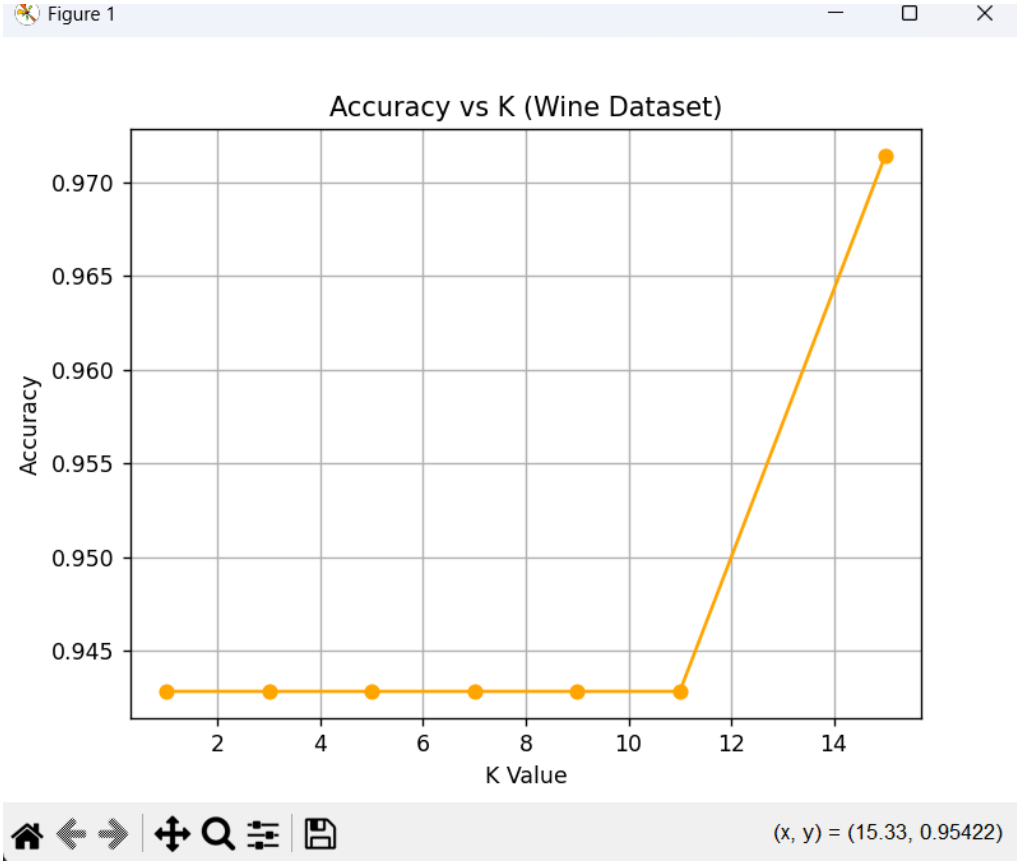
DATASET	BEST K	ACCURACY (%)
IRIS	3	100
WINE	15	97.14



3. Analysis of Accuracy vs K :

The 'Accuracy vs K-value' plot shows that the classifier achieves the best performance for k = 1 in the Iris dataset and k = 3 in the Wine dataset. Smaller K values provide more flexibility and can capture finer distinctions between classes, but they are sensitive to noise. Larger K values produce smoother decision boundaries but may misclassify points near class edges. Hence, very small or very large values of K are suboptimal — a moderate value (3–5) balances bias and variance effectively.





4. Conclusion :

In this experiment, the K-Nearest Neighbours algorithm was implemented entirely from scratch using NumPy and Python. The model achieved 100% accuracy on the Iris dataset and 97% on the Wine dataset after feature standardization. Through EDA, we observed that class separability is strongly dependent on feature selection. Additionally, the impact of the hyperparameter 'k' was analysed, demonstrating that KNN performance depends heavily on this choice. Key learnings include understanding the role of distance metrics, the importance of data normalization, and building a complete ML pipeline manually. Minor challenges included ensuring proper data scaling and subplot visualization adjustments.