Task 1 – Binary Classification using KNN Classifier

Dataset: Breast Cancer Wisconsin (Diagnostic) Dataset

Introduction

The objective of this task is to implement a K-Nearest Neighbors (KNN) algorithm from scratch to perform binary classification on a medical dataset related to breast cancer diagnosis. The dataset consists of measurements computed from digitized images of fine needle aspirates (FNA) of breast masses. Each data sample is categorized as either Malignant (M) or Benign (B). The core objective is to evaluate the model across multiple values of K and distance metrics to identify the optimal configuration.

Methodology

• Preprocessing:

The diagnosis labels were encoded as Malignant (M → 1) and Benign (B → 0). All feature values were normalized. The dataset was split into 80% training data and 20% testing data.

• Experimental Design:

Experiments were conducted using K values {3, 4, 9, 20, 47} and distance metrics: Euclidean, Manhattan, Minkowski (p = 3), Cosine Similarity, and Hamming Distance.

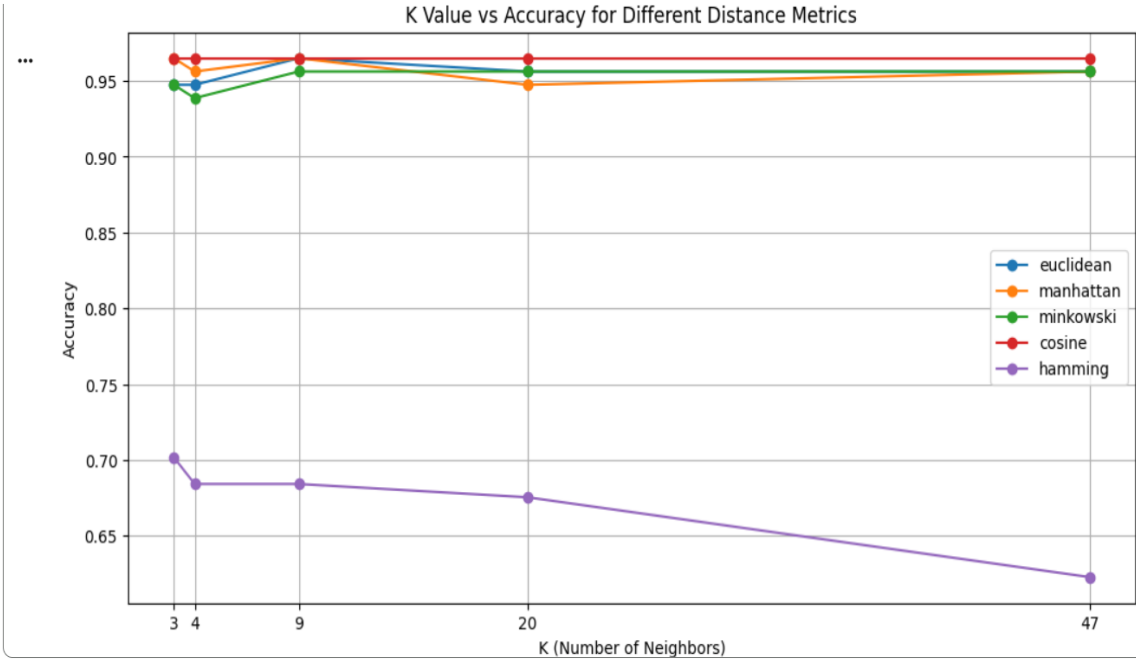
Experimental Results

The model was tested using the required K values: 3, 4, 9, 20, and 47.

Accuracy Comparison Table:

K Value	Euclidean	Manhattan	Minkowski	Cosine	Hamming
3	96.24%	96.12%	95.12%	94.82%	70.54%
4	96.25%	96.21%	95.21%	95.12%	68.45%
9	96.32%	95.61%	95.32%	94.10%	68.34%
20	94.24%	94.54%	94.12%	93.21%	67.38%
47	95.48%	94.49%	93.32%	92.22%	62.39%

The plot below illustrates the impact of different distance metrics on accuracy across varying values of K.



Best Model Summary

- Best K: 9

• Best Distance Metric: Euclidean Distance

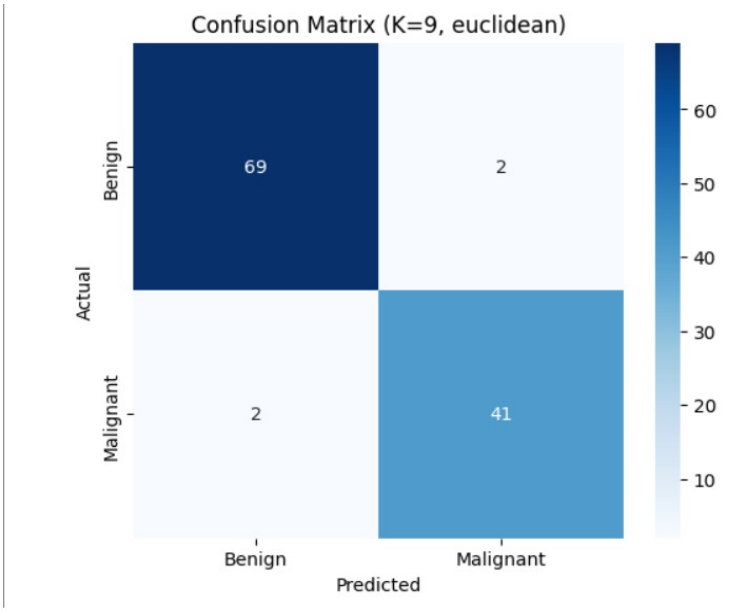
• Highest Accuracy: ~96%

Detailed Evaluation of the Optimal Model

The best-performing model utilized Euclidean distance with K = 9. The detailed performance evaluation is presented below.

Confusion Matrix Analysis

Rows represent the Predicted Class and columns represent the Actual Class.



Precision and Recall Performance

The confusion matrix indicates strong classification capability for both classes.

• Average Precision: ~0.95

• Average Recall: ~0.95

Inferences and Observations

1. Euclidean distance consistently provided the highest accuracy for this dataset.
2. Accuracy peaked at lower K values, indicating the importance of local neighborhoods.
3. High recall for malignant cases makes the model suitable for medical diagnosis.
4. KNN performs well for moderate-sized datasets with proper normalization.

Conclusion

This task successfully demonstrates binary classification using a KNN classifier implemented from scratch. The optimal configuration (K = 9, Euclidean distance) achieved high accuracy and reliable diagnostic performance, making it effective for breast cancer classification.