REPORT FOR ASSIGNMENT 1, APR

Task : Binary classification using KNN classifier

1. Introduction

This report presents the implementation of a **K-Nearest Neighbors (KNN)** classifier from scratch for binary classification on the *Breast Cancer Wisconsin* 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

- **Features:** 30 numerical features describing cell nucleus characteristics.
- **Target Variable:** Diagnosis (M = Malignant, B = Benign).
- **Encoding:** Malignant = 1, Benign = 0.
- **Train-Test Split:** 80% training, 20% testing.
- **Preprocessing:** Min-Max normalization applied to all features.

3. Distance Metrics Tested

1. Euclidean Distance
2. Manhattan Distance
3. Minkowski Distance (p = 3)
4. Cosine Similarity
5. Hamming Distance

4. K Values Tested

K values evaluated: {3, 4, 9, 20, 47}

5. Best Model Parameters

- **K:** 3
- **Distance Metric:** Manhattan Distance
- **Accuracy:** 96.49%
- **Precision:** 1.0000
- **Recall:** 0.9149

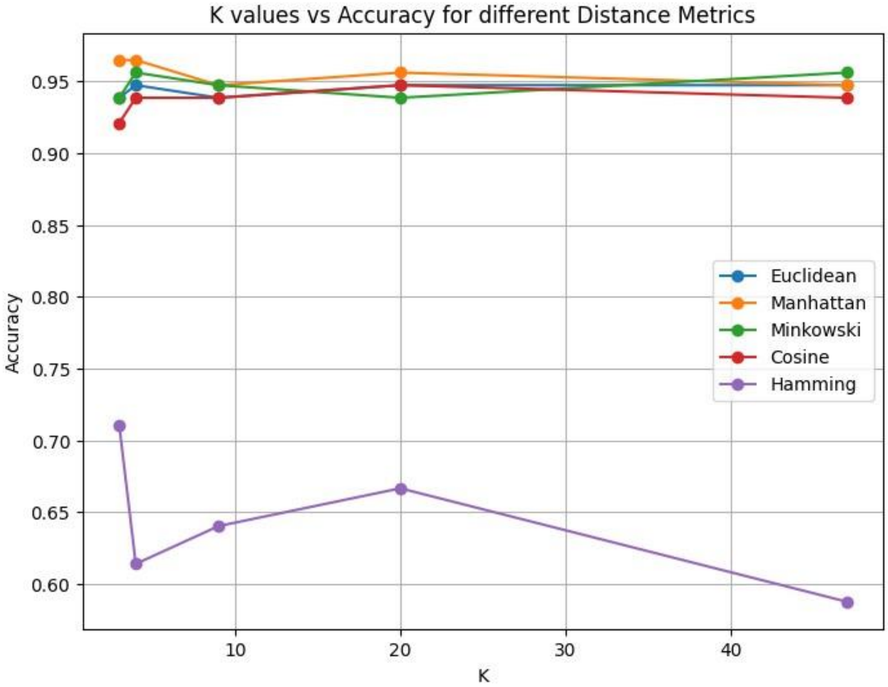
6. Confusion Matrix

	Predicted Malignant	Predicted Benign
Actual Malignant	TP = 43	FN = 4
Actual Benign	FP = 0	TN = 67

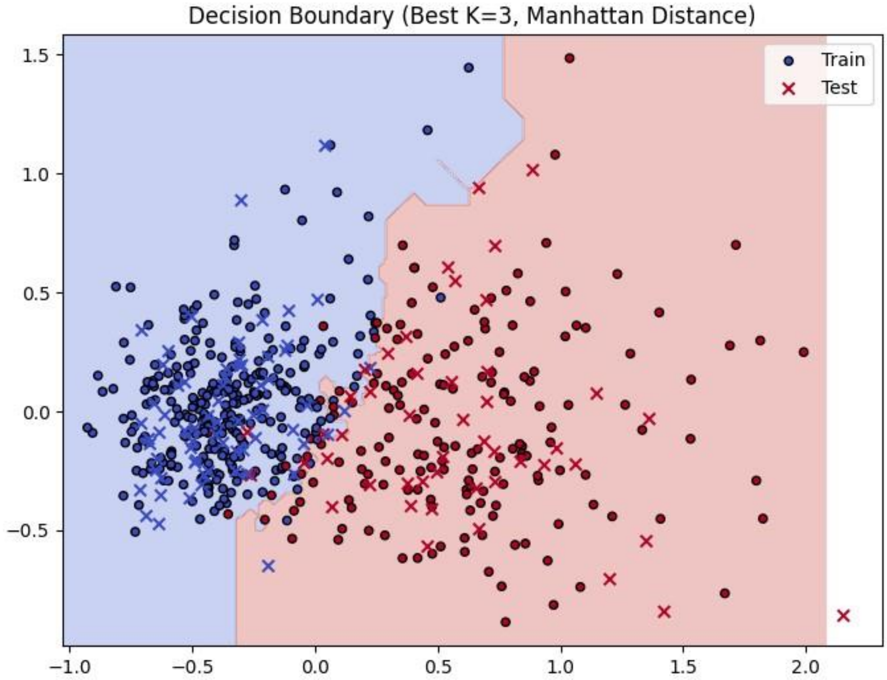
7. Observations & Inference

1. **Best Model Performance:** K=3 with Manhattan Distance gave the highest accuracy with perfect precision but missed 4 malignant cases.
2. **Impact of K Values:** Smaller K values (3, 4) performed better; large K values reduced sensitivity to local variations.
3. **Impact of Distance Metrics:** Manhattan performed best, followed by Euclidean and Minkowski. Cosine and Hamming were less effective.
4. **Error Analysis:** All errors were false negatives, which is critical in medical diagnosis.
5. **Decision Boundary Visualization:** PCA reduction to 2D shows clear separation between malignant and benign clusters, with minor overlaps.

8. Accuracy vs K Plot



9. Decision Boundary Visualization



10. Conclusion

The KNN model with K=3 and Manhattan Distance offers the best performance for this dataset, achieving 96.49% accuracy, perfect precision, and strong recall. In medical applications, reducing false negatives should be a priority; future improvements could include weighted KNN or adjusted decision thresholds.