K-Nearest Neighbors (KNN) Classification for Breast Cancer Diagnosis (Task 1)

[Group 4: 2101MC19, 2101MC41, 2101MC29]

August 17, 2024

1 Introduction

This report presents an analysis of the K-Nearest Neighbors (KNN) algorithm applied to breast cancer diagnosis. We implement KNN using various distance metrics and evaluate their performance across different k values.

2 Methodology

2.1 Data Preparation

We used a breast cancer dataset, partitioning it into training (80%) and testing (20%) sets. The features include various cell nucleus characteristics.

2.2 KNN Implementation

We implemented KNN with five distance metrics:

- Euclidean Distance
- Manhattan Distance
- Minkowski Distance (with p=3)
- Cosine Similarity
- Hamming Distance

For each metric, we tested k values of 3, 4, 9, 20, and 47.

3 Results

3.1 Accuracy Comparison

Table 1 shows the accuracy for each distance metric and k value.

Table 1: Accuracy (%) for Different Distance Metrics and k Values

k Value	Euclidean	Manhattan	Minkowski	Cosine	Hamming
3	90.35	92.98	90.35	22.81	70.18
4	92.11	91.23	91.23	14.04	70.18
9	92.98	92.11	92.98	11.40	67.54
20	91.23	92.11	91.23	7.89	66.67
47	89.47	89.47	89.47	6.14	66.67

3.2 Graphical Representation

Figure 1 illustrates the relationship between k values and accuracy for each distance metric.

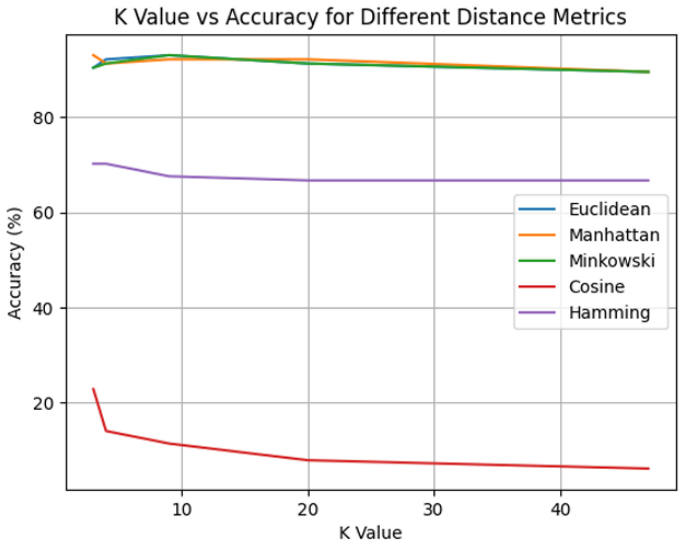

Figure 1: K Value vs Accuracy for Different Distance Metrics

3.3 Confusion Matrix

For the best performing model (Manhattan distance with k=47), we computed the confusion matrix:

Performance metrics:

- Recall: 0.7632
- Precision: 0.9063
- Accuracy: 0.8947

Table 2: Confusion Matrix for Manhattan Distance, k=47

	Actual M	**Actual B**
Predicted M	29	3
Predicted B	9	73

4 Discussion

The results demonstrate that Euclidean, Manhattan, and Minkowski distances generally outperform Cosine similarity and Hamming distance for this dataset. The highest accuracy (92.98%) was achieved by both Manhattan (k=3) and Minkowski (k=9) distances.

Cosine similarity performed poorly, suggesting that the angle between feature vectors is less informative than absolute distances for this classification task. Hamming distance also underperformed, likely due to its binary nature not capturing the nuances in continuous feature values.

The confusion matrix for the Manhattan distance (k=47) shows good overall performance, with high precision but slightly lower recall for malignant cases.

5 Conclusion

This study demonstrates the effectiveness of KNN with Euclidean, Manhattan, and Minkowski distances for breast cancer diagnosis. The choice of distance metric significantly impacts performance, with Manhattan and Minkowski distances slightly outperforming Euclidean distance. Future work could explore feature selection or scaling techniques to further improve classification accuracy.