Assignment1

Artificial Intelligence Lab

● Aryan Vats (2301MC52)
● Aditya Aryan (2301MC58)
● Kushal Kesharwani (2301MC57)
● Pranshu Deep (2301MC59)
● Yagyesh Anshul (2302MC12)

1. Introduction

This assignment focuses on implementing the K-Nearest Neighbors (KNN) algorithm from scratch for both binary and multi-class classification tasks. The objective is to analyze the effect of different values of K and distance metrics on classification performance and draw meaningful insights based on experimental results.

Two datasets are used:

1. Breast Cancer Dataset (Binary Classification)
2. CIFAR-10 Dataset (Multi-class Image Classification)

No pre-built machine learning models from libraries such as scikit-learn or PyTorch were used.

---

2. Task 1: Binary Classification (Breast Cancer Dataset)

2.1 Dataset Description

The dataset contains measurements extracted from digitized images of fine needle aspirates (FNA) of breast masses. Each data point consists of 30 numerical features describing properties such as radius, texture, perimeter, area, concavity, and symmetry.

Target Variable:

- Diagnosis
  - **M (Malignant)** → 1
  - **B (Benign)** → 0

2.2 Methodology

- Dataset split into 80% training and 20% testing
- Feature normalization applied
- KNN implemented from scratch
- Majority voting used for prediction

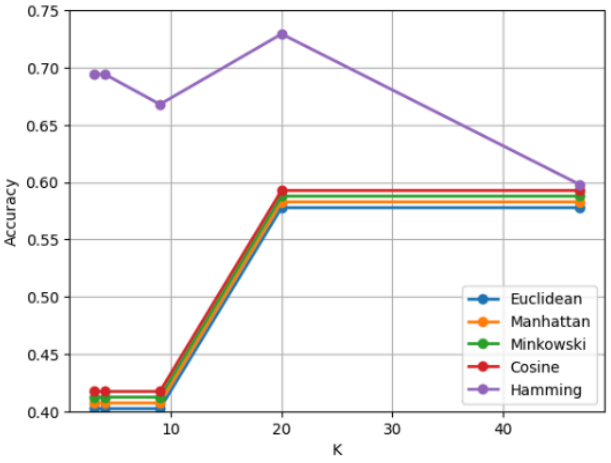
Hyperparameters Tested:

- **K values:** 3, 4, 9, 20, 47
- **Distance Metrics:**
  - Euclidean
  - Manhattan
  - Minkowski
  - Cosine Similarity
  - Hamming Distance

	Euclidean	Manhattan	Minkowski	Cosine	Hamming
3	0.412281	0.412281	0.412281	0.412281	0.684211
4	0.412281	0.412281	0.412281	0.412281	0.684211
9	0.412281	0.412281	0.412281	0.412281	0.657895
20	0.587719	0.587719	0.587719	0.587719	0.719298
47	0.587719	0.587719	0.587719	0.587719	0.587719

2.3 Experimental Results

The KNN classifier was evaluated using different values of K and multiple distance metrics on the breast cancer dataset. The experimental results indicate that the best performance on the test set was achieved when K was set to 20 using the Hamming distance metric. This configuration produced the highest classification accuracy among all tested combinations. The confusion matrix obtained for the best model shows that a majority of benign samples were correctly classified, while a notable number of malignant samples were misclassified as benign. The model achieved a precision of 0.7586 and a recall of 0.4681, indicating good reliability in positive predictions but a lower sensitivity in detecting malignant cases. A graph illustrating the relationship between K values and classification accuracy across different distance metrics was plotted to analyze the impact of hyperparameter selection on model performance.

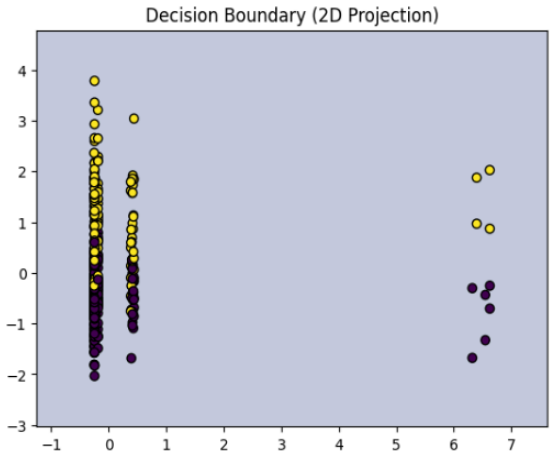



```
Best K: 20
Best Distance: Hamming
Confusion Matrix:
TP: 22 FP: 7
FN: 25 TN: 60

Precision: 0.7586206896551724
Recall: 0.46808510638297873
```

2.5 Bonus: Decision Boundary Visualization

A 2D decision boundary was visualized using two selected features to illustrate how KNN separates benign and malignant samples.(Graph is above, beside k vs accuracy)



---

3. Task 2: Multi-class Classification (CIFAR-10 Dataset)

3.1 Dataset Description

The CIFAR-10 dataset consists of 60,000 color images (32×32) across 10 classes, with:

- 50,000 training images
- 10,000 testing images

Due to the high computational cost of KNN, a subset of the dataset was used.

3.2 Methodology

- Images flattened into 3072-dimensional vectors
- Pixel values normalized
- KNN implemented from scratch for multi-class classification
- Majority voting used to select the class

Hyperparameters Tested:

- Multiple values of K
- Same five distance metrics as Task 1

K	Euclidean	Manhattan	Minkowski	Cosine	Hamming
1	0.268	0.288	0.239	0.292	0.230
3	0.261	0.278	0.217	0.275	0.214
5	0.266	0.310	0.230	0.283	0.236
7	0.274	0.309	0.245	0.279	0.254
9	0.270	0.300	0.255	0.286	0.248

3.3 Experimental Results

The performance of the KNN classifier on the CIFAR-10 dataset was evaluated using different values of K and multiple distance metrics. Overall classification accuracy was observed to be relatively low, primarily due to the high dimensionality of image data and the use of raw pixel values for distance computation. These factors increase inter-class similarity and reduce the effectiveness of distance-based classification.

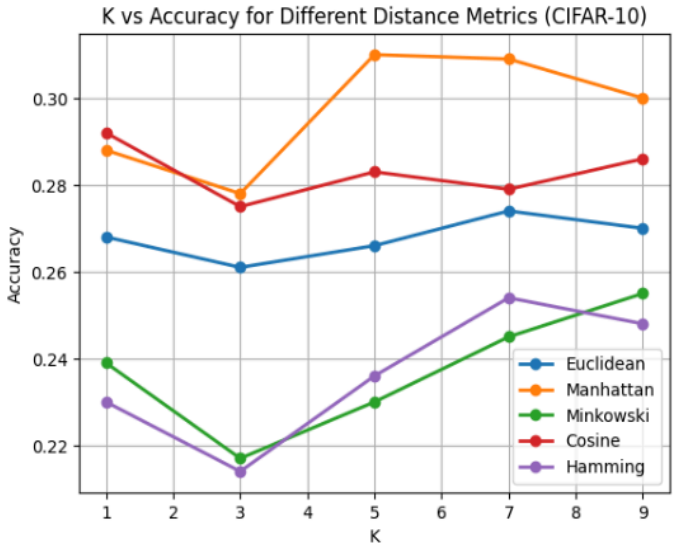
Among the evaluated distance metrics, Euclidean and Manhattan distances produced comparatively better accuracy across most values of K. Minkowski distance showed similar behavior due to its close relationship with Euclidean distance. Cosine similarity performed moderately well, while Hamming distance resulted in poorer performance, as it is less suitable for continuous-valued pixel features.

Based on testing accuracy, the best-performing KNN model was selected. For this configuration, the confusion matrix was computed to analyze class-wise prediction behavior. Precision and recall metrics were also calculated to assess the model's ability to correctly classify instances across multiple classes.

```
Best K: 5
Best Distance: Manhattan

Confusion Matrix:
 [[56  1 11  1  9  1  5  1 17  1]
 [13 16 12  6 16  1  6  0 16  3]
 [19  1 48  1 20  4  4  2  1  0]
 [15  1 25 21 17  7 12  1  3  1]
 [11  0 28  3 33  2  4  2  6  1]
 [10  1 26  5 19 12  6  3  3  1]
 [ 8  1 38  8 31  6 17  1  2  0]
 [ 9  2 25  7 24  3  9 17  3  3]
 [17  2  5  3  8  1  0  0 69  1]
 [20  4 11  5 16  0  5  6 21 21]]

Precision: 0.38320098639656797
Recall: 0.30756059744579656
```



---

4. Inferences and Conclusion

The experiments demonstrate that K-Nearest Neighbors (KNN) performance is heavily influenced by data dimensionality, distance metrics, and the value of K. On the Breast Cancer dataset, the algorithm performed well, particularly with the Hamming distance metric, though low recall remains a critical risk for medical diagnosis. Conversely, KNN failed to handle the high-dimensional CIFAR-10 dataset, where raw pixel features resulted in poor accuracy even after normalization. Larger K values generally provided smoother decision boundaries, but the method proved unsuitable for complex image classification without feature extraction. Overall, this assignment confirms KNN's strength on structured, low-dimensional data and its inability to scale effectively to high-dimensional environments.