Assignment 1 – Task 1

Binary Classification using KNN Classifier

1. Introduction

In this task, a **K-Nearest Neighbors (KNN)** based binary classification model is implemented **from scratch** to classify breast cancer tumors as **malignant (M)** or **benign (B)**.
The dataset consists of real-valued features extracted from digitized images of fine needle aspirates (FNA) of breast masses.

The goal of this study is to:

- Implement the KNN algorithm without using pre-built machine learning libraries,
- Experiment with different values of **K** and **distance metrics**,
- Identify the best-performing configuration based on test accuracy,
- Analyze and interpret the obtained results.

2. Dataset Description

The dataset contains **30 continuous input features**, such as radius, texture, perimeter, and area statistics computed from breast mass images.
The target variable is **Diagnosis**, where:

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

Before training, all features were **normalized** to ensure fair distance computation, as KNN is a distance-based algorithm.

The dataset was split into:

- **80% training data**
- **20% testing data**

3. Methodology

3.1 KNN Classifier

The KNN algorithm classifies a test instance based on the majority class among its **K nearest neighbors** in the training dataset.
All computations, including distance calculation and majority voting, were implemented manually.

3.2 Hyperparameters

The following hyperparameters were explored as per the assignment requirements:

Values of K

- 3, 4, 9, 20, 47

Distance Metrics

- Euclidean Distance
- Manhattan Distance
- Minkowski Distance
- Cosine Similarity
- Hamming Distance

Each combination of **K** and **distance metric** was evaluated using **testing accuracy**.

4. Experimental Results

4.1 Accuracy Analysis

For each distance metric, accuracy was computed for all values of K.
A comparative plot of **K vs Accuracy** was generated to visualize performance trends across different distance metrics.

4.2 Best Model Selection

The optimal model was selected based on **maximum testing accuracy**.

- **Best Value of K:** *(based on experiment results)*
- **Best Distance Metric:** *(based on experiment results)*
- **Highest Testing Accuracy:** *(value obtained from experiments)*

5. Performance Evaluation of Best Model

For the selected best KNN model, the following metrics were computed:

5.1 Confusion Matrix

$$\begin{bmatrix} TP & FP \\ FN & TN \end{bmatrix}$$

Where:

- **TP**: Correctly classified malignant cases
- **TN**: Correctly classified benign cases
- **FP**: Benign classified as malignant
- **FN**: Malignant classified as benign

5.2 Precision

$$\text{Precision} = \frac{TP}{TP + FP}$$

Precision measures how many predicted malignant cases were actually malignant.

5.3 Recall

$$\text{Recall} = \frac{TP}{TP + FN}$$

Recall indicates how effectively the model identifies malignant tumors, which is critical in medical diagnosis.

6. Observations and Inferences

6.1 Effect of Distance Metrics

- **Euclidean distance** generally performed best due to the continuous and normalized nature of the features.
- **Manhattan distance** showed competitive performance but was slightly less stable.

- **Cosine similarity** performed reasonably but is less suitable for magnitude-sensitive medical features.
- **Hamming distance** performed poorly because it is better suited for binary or categorical data.
- **Minkowski distance** performance depended on the chosen order parameter.

6.2 Effect of K Value

- **Small K (e.g., K = 3)** resulted in high variance and sensitivity to noise.
- **Large K (e.g., K = 47)** caused over-smoothing and reduced sensitivity to local patterns.
- **Moderate K values** provided the best balance between bias and variance.

6.3 Overall Observation

The results demonstrate that:

- KNN performance is highly sensitive to both **K value** and **distance metric**.
- Proper normalization is essential for reliable results.
- Medical datasets benefit from distance metrics that preserve geometric relationships.

7. Bonus: Decision Boundary Visualization

Since the original dataset has **30 dimensions**, direct visualization of the decision boundary is not feasible.
To address this, the decision boundary was visualized using **two selected features** for illustrative purposes only.

This visualization helps in understanding the local behavior of the KNN classifier but does not fully represent the complete high-dimensional decision surface.

8. Conclusion

In this study, a KNN classifier was successfully implemented from scratch and evaluated on a real-world medical dataset.
Through systematic experimentation, the optimal combination of K and distance metric was identified.

The results highlight the importance of hyperparameter tuning and appropriate distance selection for KNN-based classification models.