**Artificial Intelligence Assignment Report**

**Title:** K-Nearest Neighbors (KNN) Classification from Scratch

## Introduction

This assignment focuses on implementing the K-Nearest Neighbors (KNN) algorithm from scratch for both binary and multi-class classification problems. The objective was to understand how KNN works internally, how distance metrics affect performance, and how model evaluation is done.

## Task 1: Binary Classification (Breast Cancer Dataset)

The breast cancer dataset contains measurements obtained from Fine Needle Aspirate (FNA) tests. The target variable is diagnosis, where Malignant is labeled as 1 and Benign as 0. The dataset contains 30 numerical features.

Data preprocessing involved removing unnecessary columns, encoding class labels, and normalizing features. The dataset was split into 80% training data and 20% testing data. Multiple distance metrics such as Euclidean, Manhattan, Minkowski, Cosine, and Hamming were implemented. Different values of K were tested to find the best model based on accuracy.

For the best-performing model, a confusion matrix was computed along with precision and recall. Recall was considered especially important because missing a cancer case is critical in medical diagnosis. A K versus accuracy graph was plotted, and a 2D decision boundary was visualized using two features.

## Task 2: Multi-Class Classification (CIFAR-10 Dataset)

The CIFAR-10 dataset consists of 60,000 color images of size 32x32 belonging to 10 different classes. Due to high computational cost, a subset of 5,000 training images and 1,000 testing images was used. Images were flattened into vectors and normalized.

KNN was applied using multiple distance metrics and different values of K. Accuracy was calculated for each configuration. The best model was selected based on maximum accuracy. A 10x10 confusion matrix was created, and average precision and recall were calculated across all classes.

## Observations

KNN performed well on the breast cancer dataset but achieved lower accuracy on CIFAR-10 due to high dimensionality. Euclidean distance worked best for numerical medical data, while Cosine similarity performed better for image data. KNN is computationally expensive and not ideal for large datasets without optimization.

## Conclusion

This assignment provided a deep understanding of distance-based learning, hyperparameter tuning, and model evaluation. Implementing KNN from scratch helped in understanding the importance of feature scaling, distance metrics, and evaluation methods in machine learning.

**End of Report**