Student Name:	Fall 2024 Assignment 2
Student Number:	November 3, 2024

Assignment 2: Supervised Learning

K-Nearest Neighbor (kNN) Classifier is a supervised pattern classifier that determines the class of an input sample based on the distance to k nearest labeled neighbors. kNNs are considered as a type of lazy learning method where an evaluation is performed as needed. In this assignment, you are going to implement basic kNN algorithm and analyze the effect of normalization.

Implement basic kNN algorithm to classify the given data set. Do not use built-in function such as knn for clustering. You must implement it by yourself. This is an individual assignment.

Dataset: The sample data set to be used for this project is given below. The bcwdisc.data.mb.csv is a version of a breast cancer database published by the University of Wisconsin Hospitals. The original data set was obtained from the University of California Irvine (UCI) machine learning repository (http://archive.ics.uci.edu/ml/). The data set is in .csv format, with one sample / pattern vector per line of input. Each line will contain a series of attribute values, separated by commas. The file does not contain headers. The first 30 columns are attributes, and the last attribute indicates the class. The class labels are 1 (malignant) and -1 (benign).

A. What to DO:

Write a program to perform supervised pattern classification using kNN classifiers as described in class. Your program should have two components: distance calculation and class assignment.

1. Discuss whether you need to normalize the dataset or not. Check the ranges of values and that will give you a hint. Then, make the decision.
2. Separate the entire dataset (613 samples) into training and testing set such as 70% training, 30% testing.
3. For algorithm development, first implement a distance calculation module where a data set with labeled samples (training data) will be used to calculate distances to testing samples.
4. Second implement an assignment module where it will find k-Nearest Neighbors to classify samples with unknown class assignment. You must implement these two modules in one program.
5. Test your kNN for k=1, 3, 5, 7, and 9.
6. Show your testing results for different k values using their accuracy and confusion matrix.

B. What to turn in:

- A zip file with all the necessary SOURCE code
- A written report (in pdf format) using the template including the following contents:
  - cover page with your name, class title, class number, date, etc.
  - an "Abstract" (no more than 100 words) summarizing what this project is about (objective), what you did, and what you found out in this project.
  - a "Result and Discussion" section showing your results, discussing them, and summarizing lessons learned, your experience working on the project, potential future work if given time, etc.
  - source code printout.
- Upload the zip file and project report to D2L as separate files.

C. Due Date: It is announced on D2L.