Homework 2 *(due: Feb 03)*
Machine Learning - COSC 4360

Department of Computer Science and Electrical Engineering

Spring 2025

Exercises

*Create a **New Project** for every exercise. Take a screenshot of the source code along with its output and place the **source code** and the **screenshot** in a **zipped folder** named **LastNameFirstName_HW2***

Exercise 1

Given the following dataset: *hsbdemo.csv*, perform classification using the **KNN** supervised learning algorithm with **K=5** and **test_size=0.10**. Print the **accuracy** of the model as well as the **visual confusion matrix**. Additionally, print the labels of the **misclassified** data points (predicted vs actual). For features (i.e., predictor variables), use all columns apart from columns: *id, prog* and *cid*. For the target variable (i.e., y or response variable) use column *prog*.

Note 1: All features (i.e., predictor variables) must be converted from categorical to numerical.

Note 2: You may use **random_state=3** as the last parameter in the *train_test_split()* function to ensure that the same random samples are selected in every run.

Exercise 2

Given the following dataset: *hsbdemo.csv*, apply the **Principal Component Analysis** (PCA) unsupervised learning algorithm. Print the **variance ratio** and plot the **cumulative sum** of the **variance ratio** for all 10 features, as shown in Fig. 1 below. Use the same columns for **X** and **y** as in Exercise 1.

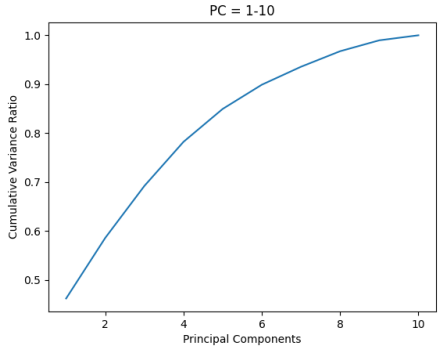

Figure 1: Cumulative variance ratio for 10 Principal Components

Exercise 3 *(Optional)*
Implement your **own** version of KNN in Lab 2, Ex. 1.

Exercise 4 *(Optional)*
Implement your **own** version of PCA using either the *hsbdemo.csv* or *iris.data.csv* dataset.

Note: Submit through **Canvas**