CS 250 Spring 2025

Homework Assignment 4

Due on Moodle: Before class on Wednesday, February 26. Mail submissions will not be accepted. Submit one single Jupyter Notebook for all questions. For this homework assignment, download the Possum Regression dataset from Kaggle(Link: https://www.kaggle.com/datasets/abrambeyer/openintro-possum) and do the following:

1.
    a. Read the data into a dataframe [1 points]
    b. Find if there are any cells containing null values, and remove those rows. [2 points]
    c. Drop the 'case', 'site' and 'Pop' columns. [2 points]
    d. Separate the sex column into a separate array/dataframe/series called labels, and the rest of the remaining columns into an array/ dataframe called features. Remember to drop the original sex column from here [5 points]

2. Now do the following:
    a. Split the features and labels into 75% training and 25% testing using train_test_split() (will be taught in class) [5 points]
    b. Predict the label (m or f) of each test data based on the K-nearest-neighbor technique. Use a value of K>1 and either Euclidean or Manhattan distance. [5 points]

3. Evaluate the performance on the test data using normalized confusion matrix (code given in class) [5 points]