Name: Sarah Huang

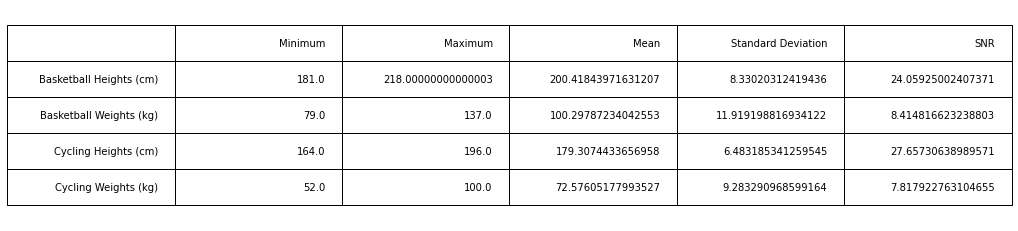
Date: 9/22/2023

# Lab 2: K-Nearest Neighbors

**Question 1 (15 points): Read in the data from the given CSV file. Down-select to create a dataframe that only includes three columns (sport, height, and weight) for all of the males who competed in either basketball or cycling, then perform dropna() to remove any missing rows. This should result in a dataframe with 3 columns and 450 rows. Convert the nonnumerical feature (sport) to numerical values. Further, convert the height values from meters to centimeters. Include in your report a brief description (1-2 sentences) of why it might be a good idea to use the data in centimeters rather than meters for k-nearest neighbors.**

It’s a good idea to use the data in centimeters rather than meters for k-nearest neighbors because the algorithm will be able to pick up on the small differences better when distinguishing between data points. For example, it feels like there’s a bigger difference between 181 cm and 172 cm rather than 1.81 m and 1.72 m even though the centimeter and meter numbers are equivalent.

**Question 2 (15 points): For each class (basketball, cycling) and feature (height, weight), compute the following statistics: minimum, maximum, mean, standard deviation, and signalto-noise ratio (SNR = mean/standard deviation). Create a table that shows the statistics for each class and feature. Do any of the statistics give rise to concern? Provide a brief description (1-2 sentences) in the report.**

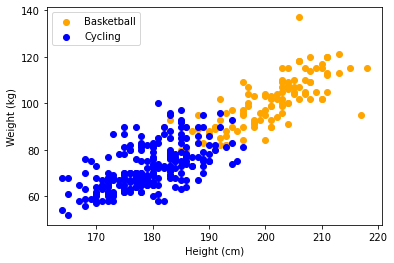


The basketball height and cycling height have a good signal-to-noise ratio at around 24 and 28 respectively. That means there is some noise but it is acceptable. However, basketball weight and cycling weight have very low signal-to-noise ratio, so there must be a lot of noise and unwanted behavior in the data.

**Question 3 (5 points): Divide the data into training and test data using the standard 80-20 ratio. Apply 10-fold cross-validation to the training data.**

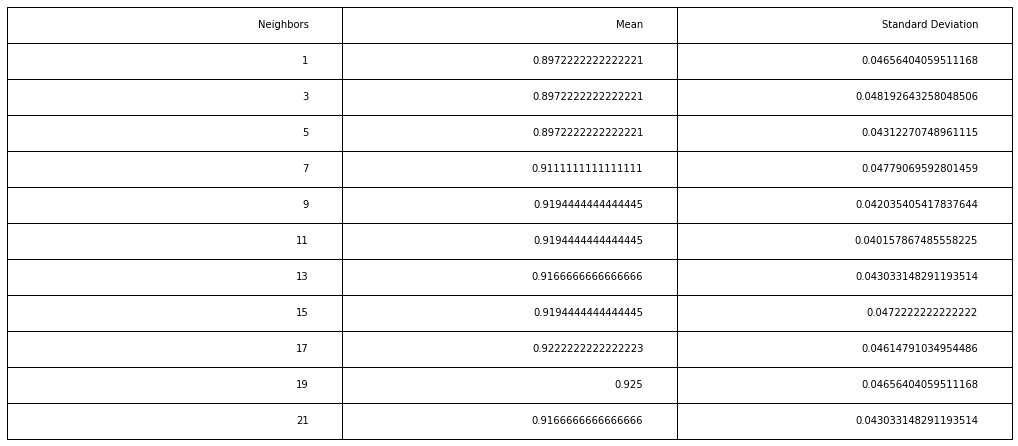
Check the code!

**Question 4 (10 points): Create a scatter plot of the training data, where you color the basketball data and cycling data differently. Make sure to label your axes and include a legend for the labels. Does the scatter plot indicate potential success or failure with respect to separating the two classes using KNN? Why? Provide a brief description (1-2 sentences) in the report.**

****

The scatter plot may have potential failure with separating the two classes using KNN. While most of it is clearly separated, they are intermingled at the middle.

**Question 5 (25 points): Create and iterate over different numbers of neighbors from 1 to 21 (inclusive, but odd numbers only). For each hyperparameter, calculate the mean and standard deviation of classification accuracy across the different folds of the data. Create a table 1 for your report showing the mean and standard deviation for each hyperparameter. Select which hyperparameter to use based on the mean classification accuracy on the validation data. Note which hyperparameter value is selected in the report.**



The best mean classification accuracy is at neighbor 19.

**Question 6 (10 points): ): Build a new K-nearest neighbor classifier using all your training data with the hyperparameter defined in Question 5. Use the resulting model to classify the test data (which haven’t used until now). Calculate and report overall training and test data performance.**

Training: 0.9277777777777778

Testing: 0.9777777777777777

**Question 7 (20 points): Create a 2D plot that visualizes the classifier’s performance (a decision boundary). Run a mesh of data through the classifier to determine basketball and cycling decision regions. Color them two different light colors (alpha=0.2) that are easy to visually separate. Then, overlay the test data using two different colors for basketball and cycling. Make sure all plots are labeled and that a legend is included.**

