MA103: Mathematical Modeling & Intro to Calculus
K-Nearest Neighbors 1

Lesson Objectives: Cadets will

1. create a visualization of standardized data.
2. explain the concept of the KNN algorithm.
3. use standardized distances to find nearest neighbors.
4. use KNN to make a prediction.

Admin Notes / Agenda

- Scavenger Hunt due now
- PSL 2 due on Canvas 10 2359 September 2025

Review:

- Explanatory variable
- Response variable
- Historical records
- Predictor record
- Sample mean: $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$
- Sample standard deviation: $s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$
- Euclidean distance between (x_1, y_1) and (x_2, y_2): $D = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$

k-Nearest Neighbors Summary: k-NN is a classification and prediction algorithm that uses the k historical records that are closest to the predictor record (the k-nearest neighbors) as the region to which the predictor record is classified.

k-Nearest Neighbors Steps:

1. Standardize all records, do not use the predictor record to calculate $\bar{x}$ and s
2. Find the Euclidean distance between the predictor record and each historical record
3. Select the k-nearest neighbors based on Euclidean distance

4. Make a classification prediction of a categorical outcome variable based on the majority of the k-nearest neighbors

5. Make a prediction of a numerical outcome variable based on the average value of the k-nearest neighbors

Why standardize data? SAT scores range from 0 to 1600 and the local high school has a GPA range of 0 to 4.0. Data for three students are shown below.

	SAT	GPA
Billy	1500	3.5
Sally	1490	1.0
Jane	1520	3.5

1. Which student's performance do you think is most similar to Billy's?

2. Calculate the Euclidean distance between Billy's performance and that of Sally and Jane without standardizing.

3. According to the distance formula, which student's performance is closer to Billy's? Do you agree?

4. The mean and standard deviation of SAT scores are $\bar{x} = 1503$ and $\sigma = 15.3$. The mean and standard deviation of GPA are $\bar{x} = 2.66$ and $\sigma = 1.44$. Standardize the SAT and GPA data and recalculate the distance between Billy's performance and that of Sally and Jane, which student's performance is closer?