Tram Le:

Homework 6:

1. When I load the dataset file, the summary of species showing “character” instead of the name “Adelie,…” and it cause my logistic model having an error. So instead, I use the built in Penguins dataset by library library(palmerpenguins)
2. Split data into training/testing using 80/20 split with sample function then use that split to select all row then select all other row but these row into train\_pen/test\_pen
3. Using glm() function with independent variables “species” and other dependent variables to compute the logistic regression then summarize it
4. Then use above logistic regression to predict on testing data
5. Then show the confusion matrix table on species in test\_pen dataset and the test prediction

predict\_tes

0 1 ( False True)

Adelie 26 1

Chinstrap 1 18

Gentoo 0 23

# tp is 1st row 1col: actual and predict shoule be the same

tp = confusion\_matrix[1] = 26

# fp is the sum of all value column except tp

fp = sum(confusion\_matrix[2],confusion\_matrix[3]) = 1

# fn is sum of value row excepts tp

fn = confusion\_matrix[4] = 1

# tn is sum of all col and row except values we calc

tn = sum(confusion\_matrix[5],confusion\_matrix[6]) = 41

1. Compute the Accuracy, Error Rate, Precision, and Recall:

accuracy = sum(tp,tn)/sum(tp,tn,fp,fn) = 0.971…

precision = tp/sum(tp,fp) = 0.963… (correct 96% of time)

recall = tp/sum(tp,fn) = 0.963 (correct 96% of time)

error rate = 1 – accuracy = 0.0289…

1. 3. I can only do the split. My knn() function for some reason it keeps getting an error saying that “no missing values are allowed” even I input all value in parameter.