Dear all,

Here are detailed instruction for Project-Report #3. Please combine ALL of Project #1, Project #2 and Project #3.

**First, in your weekly reports, numerate your possible output tables and graphs clearly with a title, so when you write your summary, you can refer to Table 1 or Figure 1, Table 1, et. al.**

***Next: You will need to report the running time for each classifier. If your dataset is large, pay extra attention to the running time as it may take up to hours to complete in the future.***

## Here are codes on how to get running time.

start.time <- Sys.time()

***[… your procedure here…]***

end.time <- Sys.time()

time.taken <- end.time - start.time

time.taken

##

## Here are codes to write out your outputs to a separated CSV file ##

Write.csv(KNN.OUT, “KNN\_OUT.csv”)

##

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

Part I: Regression

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

For simplicity, let’s only consider first 20 BIF feature variables for “MorphII\_BIF\_s7-37\_g0.1\_max\_partial.csv” dataset:

Please consider all possible Linear, Quadratic, and Polynomial (consider certain degree of polynomial, eg: degree=3, or degree=4) Regression Models. Check your model fits numerically and graphically. Draw conclusions respectively.

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

Part II: Classification

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

By using all BIF features for each individual, we are starting to get into the heart of statistical data mining, so projects are getting more and more computational intensive.

Consider all four classifiers among Logistic Regression, LDA, QDA and KNN: Use model selection with variable selection to find the best model fit.

Note: For **EACH of** four classifiers among Logistic Regression, LDA, QDA and KNN, you MUST follow two steps listed below respectively.

Step 1: Build models with *all your observation* (of size N). Then predict the outcomes for all your observation of size N. Finally find the confusion matrix, the overall prediction accuracy, sensitivity and specificity for all four models.

Step2: Take *first* 80% of all your observation of size N as a training set. All the *rest* 20% will be taken as a testing set. Next, build your model with the training set (with size of 80%\*N). Then predict the outcomes for your testing set (with size of 20%\*N). Finally find the confusion matrix, the overall prediction accuracy, sensitivity and specificity for your testing set.

Please let me know if you have any questions/concerns.