# Logistic Regression; Example in Python; AMS 580

# Name:\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_SBU ID:\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

# Name:\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_SBU ID:\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

# Name:\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_SBU ID:\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

# Name:\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_SBU ID:\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

# Please include (1) Python file; (2) Output from Python with answers to all the questions asked; (3) Comparison of the results to those using R; (4) Recommended websites for logistic regression using Python.

#### Logistic Regression (\* A type of Generalized Linear Model when the response is binary) with the bwt Data

The PimaIndiansDiabetes2 [in mlbench package] data is a built in R dataset containing 9 variables and 768 cases – please see the following link for detailed description:

<https://rdrr.io/github/quantide/qdata/man/pimaindiansdiabetes2.html>

Your task is to use all the other 8 variables to predict the binary dependent variable ‘diabetes’ telling us whether the subject is diabetic or not (factor with 2 levels: neg and pos). You will split the data into 80% training and 20% testing.

1. Please split the data into 80% training and 20% testing using seed =123.
2. Then you shall fit a logistic regression model with all the other 8 predictors using the training data.
3. Please use this fitted model based on the training data to predict the response variable ‘diabetes’ (whether the subject is diabetic or not) for the testing data. Please generate the confusion matrix, and report:
4. The overall accuracy;
5. The sensitivity (that is, the probability a subject is predicted to be diabetic given that he/she was in fact diabetic);
6. The specificity (that is, the probability a subject is predicted to be not diabetic given that he/she was in fact not diabetic).

