Data Mining
Georgia State University
Professor: Jaya Krishna Mandivarapu **Home Work 5**

<p style="color:red; text-align:center">Due Date: December 1st 2020 at 11.59 PM</p>

Submission Requirements

You must turn work at the SPECIFIED TIME so you can receive credit for Homework!

Please provide all the work in single html, jupyternotebook file. PLease don't submit multiple different files like word,pdf. We only accept two file .html,jupyternotebook file

Files Required for submission : One Jupyter Notebook and HTML file (Can be download from Jupyter notebook you are working with)
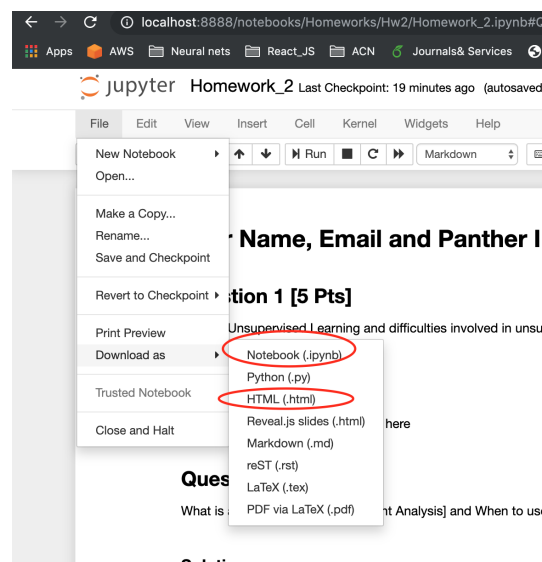

Figure 1: Download as Jupyter Notebook and HTML file

Homework 5 must be **submitted on icollege** by the due date and time. Late homework will be subject to a penalty of 50 percent for 1 day and 80 percent for two days and after 3 days no submission allowed, as stated in the course grading policy. No email or hard copies of homework will be accepted.

You may discuss the assignments with other students in the class, but (as stated in the academic honesty policy) your written answers **must be your own**, and you must list the names of other students you discussed the assignment with.

How to Submit

Log into **iCollege(iCollege)**, select the class to view its drop box folders, select the correct folder for the given assignment and upload the file there.

You will get a confirmation email. Please save the conformation email in the event something goes wrong, for example work was submitted to the wrong folder etc..

Data Mining
Georgia State University
Professor: Jaya Krishna Mandivarapu **Home Work 5**

1. Apply Linear Regression on give dataset and predict the output target variable ("quality"). Please plot all the necessary plots like input data, loss , predicted vs actual values in the plot [15 pts]

   Dataset Link : Dataset check Icollege

2. Explain some of the assumptions followed in the parametric methods and non-parametric methods? Please explain explain which method to use in which cases of the dataset. [10 Pts]

3. Explain in detail about how logistic regression has been derived from Bernoulli distribution and when to apply logistic regression instead of linear regression ? [10 Pts]

4. Do the required exploratory data analysis and appply Logistic regression on the given dataset ? The output should have your code, plots and below following values at the end (15 Pts for Exploratory data analysis and plots ,15 points for the rest of the algorithm)[30 Pts]

   The acuuracy of the model = TP+TN/(TP+TN+FP+FN) = _____
   The Missclassification = 1-Accuracy = _____
   Sensitivity or True Positive Rate = TP/(TP+FN) = _____
   Specificity or True Negative Rate = TN/(TN+FP) = _____
   Positive Predictive value = TP/(TP+FP) = _____
   Negative predictive Value = TN/(TN+FN) = _____
   Positive Likelihood Ratio = Sensitivity/(1-Specificity) = _____
   Negative likelihood Ratio = (1-Sensitivity)/Specificity = _____

   Dataset Link : Dataset check Icollege

5. Apply Cross validation on the above Logistic regression and comment on the change in accuracy [5 pts].

6. a) Appply Support Vector Classifier on the given dataset ? The output should have your code, plots, accuracy, precision, recall, f1-score.

   b) Use different kernel methods available (Signmoid, Gaussian, Polynomial) in the SVM and comment on the accuracy and performance using different kernels.

   Dataset Link : Dataset check Icollege

   (15 pts part a , 15 points for part b)[30 Pts]

   Hints:

   As per the data preprocessing step convert all the variables in the dataset into Numerical values as the algorithms only work with Numerical values