**Data Science with R- Project 4**

**Problem Statement**: “FlightDelays.csv” is a dataset given here. It shows that there have been 2201 airplane flights in January 2004, flying from the Washington DC area into the NYC area.

What’s interesting to note is that the main characteristic of interest (the response) is if a flight has been delayed by more than 15 min (coded as 0 for no delay, and 1 for delay) or not. The explanatory variables include three different arrival airports—John F Kennedy, Newark, and LaGuardia. The three different departure airports are Reagan, Dulles, and Baltimore, along with eight carriers, and a categorical variable for 16 different hours of departure (6 am to 10 pm). The other variables include weather conditions (0 = good/1 = bad) and day of the week (1 for Sunday and Monday and 0 for all other days).

Apply logistic regression analysis to identify the flights that are likely to be delayed. This can be done by selecting 60% of the cases of data set (1320 cases) for the fitting (training) data set, wherein the remaining 40% of the cases (881 cases) become the evaluation data set.

For classification, the logistic regression model can be used that includes all explanatory variables as covariates. On the basis of the available training data set, the logistic regression model can help you estimate and classify a new case as a success in case its predicted probability of success is greater than 0.5. Do not forget to provide the R-script file with all the necessary codes.