**Final Project Report - Individual Contribution**

**ALY 6015 – Intermediate Analytics**

**Name: Nalini Macharla**

**NUID: 001534506**

**MPS Analytics, College of Professional Studies**

I and my teammate together selected this dataset from the UCI Machine Learning Repository which is available for public research. In this group assignment, I have contributed in various things. Firstly, I have performed the EDA (Exploratory Data Analysis) on the selected dataset.

Let me explain in detail, I have found out the unwanted columns, for example, let’s assume if there are any serial number/ID sequence out there. After analysis, there are no unwanted columns present in the dataset. Further I have located the missing values. If you find any sort of missing values, we try and find the relation that is there between the targeted variable and the missing values. However, in our code there are no missing values. We further check the values with one by running the code and we realize that with one value there is no feature present.

In our dataset there are 9 categorical features and the features like “job” and “month” seems to be having the highest value. Further for these categorical features we will find the distribution. To find the distribution I have plotted few graphs. We find the relation between the labels and the features after we find the categorical features. After finding the categorical values, I found the numerical values in the dataset for example: day, age and so on. We found both discrete continuous numerical features and also, we will find the distribution for these features and then find the relationship that these features have with the labels. Next, I tried to find the outliers of the features that I found and tried to explore the correlation factors between those outliers. Further with mean values we have replaced the unwanted values that are present in the dataset. I have found a pair plot and we have checked for the targeted values from the classification to see if they are balanced or not.

I have next performed logistic regression analysis to find the dependent variable is dichotomous or binary. It is used to make classifications into various categories like 1/0, Yes/No, True/False etc. It predicts the probability of occurrence of an incident by fitting data to a logit function. Logistic regression will efficiently compute a maximum likelihood estimate assuming that all the inputs are independent. So, we chose logistic regression model.

Finally, as a part of this assignment, my contributions involves individual R codes for the above and I wrote my part in the draft and final report of the project. Also, helped the teammate to prepare the first four slides in the final PPT that we presented in the class. I thoroughly enjoyed by doing this assignment with my teammate and we are happy with the results.