Logistic Regression Assignment

### **Problem Statement:**

Sam is done building linear regression models. But, Sam realizes that Linear Regression is only for continuous variables. Hence, this time, Sam decides to build logistic regression models on categorical variables.

Sam starts by building a simple logistic regression model where there is only one independent variable.

## Questions on Simple logistic regression:

1. Build a simple logistic regression model on the ‘customer\_churn’ dataframe, where the dependent variable is ‘Churn’ & the independent variable is ‘TechSupport’. Store the result in ‘log\_mod\_1’
   1. Have a glance at the summary of the model built
   2. Predict the result when the value of ‘TechSupport’ is ‘Yes’
   3. Predict the result when the value of ‘TechSupport’ is ‘No’
   4. Predict the result when the value of ‘TechSupport’ is ‘No internet service’
2. Build a simple logistic regression model on the ‘customer\_churn’ dataframe, where the dependent variable is ‘Dependents’ & the independent variable is ‘tenure’. Store the result in ‘log\_mod\_2’
   1. Have a glance at the summary of the model built
   2. Predict the result when the value of ‘tenure’ is 10
   3. Predict the result when the value of ‘tenure’ is 50
   4. Predict the result when the value of ‘tenure’ is 70

After building logistic regression models with just one independent variable, Sam decides to build multiple logistic regression models with multiple independent variables.

## Questions on Multiple Logistic regression:

1. Build a multiple logistic regression model:
   1. Start off by dividing the data-set into ‘train’ & ‘test’ sets in 65:35 ratio, with the split-criteria being determined by ‘gender’ column
   2. Build a logistic regression model on the train set where the dependent variable is ‘gender’ & the independent variables are ‘Dependents’, ‘InternetService’ & ‘Contract’ & store the result in ‘log\_mod\_multi’
   3. Predict the values on top of the test set & store the result in ‘result\_log\_multi’
   4. Have a look at the range of ‘result\_log\_multi’ & build a confusion matrix where the threshold of predicted values is greater than ‘0.49’
   5. Calculate the accuracy of the model from the confusion matrix
2. Build second logistic regression model on the same ‘train’ & ‘test’ sets
   1. In this case dependent variable is ‘gender’ & the independent variables are ‘tenure’, ‘MonthlyCharges’ & ‘PaymentMethod
   2. Predict the values on top of the test set & store the result in ‘result\_log\_multi2’
   3. Have a look at the range of ‘result\_log\_multi2’ & build a confusion matrix where the threshold of predicted values is greater than 0.49
   4. Calculate the accuracy of the model from the confusion matrix

Finally, Sam decides to do some performance analysis by using the ROCR package

## Questions on ROCR package:

1. Build a logistic regression model:
   1. Start off by dividing the data-set into ‘train’ & ‘test’ sets in 80:20 ratio, with the split-criteria being determined by ‘Churn’ column
   2. Build a logistic regression model on the train set where the dependent variable is ‘Churn’ & the independent variables are ‘MonthlyCharges’, ‘tenure’ & ‘TechSupport’ & store the result in ‘log\_mod\_roc’
   3. Predict the values on top of the test set & store the result in ‘result\_log\_roc’
   4. Use the performance() function from the ROCR package & build the ‘Accuracy vs cut-off’ plot
   5. Plot the ‘ROC’ curve
   6. Find out the “area under the curve”