# DATA MINNING - Project 1 Milestone

**Team 16** –

Akshaya Sharma 1001707344

***Overall Status:***

We read the concept of Decision Tree. Understood the Information and Entropy concept. Then we went through the Rpart package information given to us by Sir in the link as well as R tutorial. Understood the methods of rpart and the arguments given to the method. Then completed the tasks required for the MileStone.

**Status:** Completed data cleaning, data sampling with 30 as seed, Split the data into Training Set and test data set with 60% for training set to analyze and train the module and 40% data for test data to perform prediction based on training data. Then performed information gain with rpart method. Then we performed withholding of all the columns one by on to find the most accurate.

***Detailed Analysis:***

**Steps: -**

1. At first, we took the text data file and converted the same to xlsx.
2. Then we removed the missing values by converting all “?” to NA (the notation to indicate missing in R) and after cleaned the NA from data set.
3. Our sample size for data is 2K so we chose 2K dataset with 30 seed.
4. After that we trained 60% of data using **rpart** library and tested remaining 40%.
5. During this we encountered two issues as our dataset was too small so native country = Poland and relationship = Married-AF-Spouse were having new levels, so we removed the same.
6. Then we plotted the decision tree and created the confusion matrix.
7. From matrix we calculated accuracy, precision, recall and F Score.
8. Now we tested the model by withholding one column and found that withholding **fnlwgt** column we are getting 85% accuracy which is 0.375% more than regular model.
9. Even the variables used in tree construction are 6 making new model more descriptive and same can be analyzed in the plotting.

***Problems encountered and how you handled them;***

Took time to learn R programming language. Found difficult to determine the split column.

References:

[1] <https://www.edureka.co/blog/decision-tree-algorithm/>

[2] <https://www.gormanalysis.com/blog/decision-trees-in-r-using-rpart/>

[3] <http://www.sthda.com/english/wiki/reading-data-from-excel-files-xls-xlsx-into-r>