# Corporate Termination Assessment

## Team Members

**Shrey Kshatriya**

Course: CS 513-B
Department: MIS
Level: Graduate student

**Kavit Shah**

Course: CS 513-B
Department: MIS
Level: Graduate student

**Avirat Belekar**

Course: CS 513-B
Department: CS
Level: Graduate student

## Problem Statement

Employees are an integral part of an organization. They make up the organization and are the ones responsible for smooth functioning of the company.

However, employees are not always loyal, or the company might fire them due to a financial crisis. Either way, it is always a loss for the company.

We aim to develop a classification model(s) to predict the potential of employees leaving the company (become terminated).

## Dataset Explanation

We have the attrition dataset of the companies which has 27 columns and 9612 entries which also include the missing values. The dataset has information about the employer such as Employer ID, Annual Rate, Hourly rate etc.

## Pre-processing

1. **All categorical values to numeric:**
   We have converted all the categorical values to the numeric values to apply the algorithms by using the code line **'as.numeric'** inbuilt function given by R. This was done because the classification model requires numeric values as input.

2. **Removed all missing values:**
   We removed the termination column because it contained a lot of missing values and it contributed the least towards the target column ‘Status’.

3. **Normalize the values:**
   We normalized the values before applying any algorithms using ‘min-max scaler’. This was done because change the values of numeric columns in the dataset to a common scale, without distorting differences in the ranges of values.

4. **Data Split:**
   We have split the data into 80% training which is used to train the model. The model accuracy is evaluated on 20% testing data.

# Feature Selection

The complete dataset was used for feature selection and only the important columns that were highlighted by these algorithms were selected. The algorithms used were:

1. **Forward and Backward Selection**
2. **Boruta**
3. **Earth**

**Conclusion of Feature Selection**

The columns that were selected were All of these algorithms gave several important columns. However, all these algorithms had a few common columns and those were the columns that we have selected. We selected a total of 8 independent variables and 1 dependent variable ‘STATUS’. These independent variables are:
Annual Rate, Job Code, Job Group, Hourly Rate, Previous Year 1, Previous Year 3, Previous Year 4, Previous Year 5.

# Classification Algorithms

1. **Logistic Regression:** Accuracy : 57.142 %

2. **Random Forest:** Accuracy :75.589%

3. **Artificial Neural Network:** Accuracy : 56.221%

4. **Decision Tree:** Accuracy : 68.49%

5. **C50 Algorithm:** Accuracy : 75.614%

6. **K Nearest Neighbour with 3 neighbours:** Accuracy : 71.24%

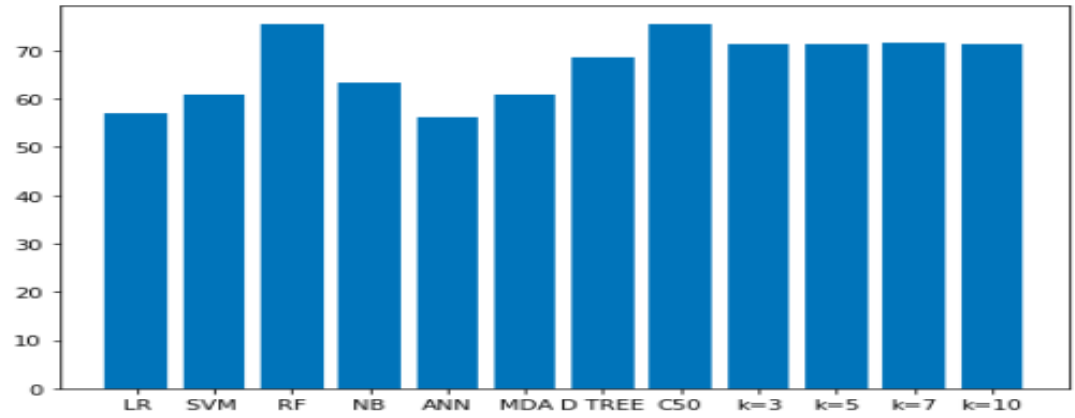
7. **K Nearest Neighbour with 5 neighbours:** Accuracy : 71.32%

8. **K Nearest Neighbour with 7 neighbours:** Accuracy : 71.53%

9. **K Nearest Neighbour with 10 neighbours:** Accuracy : 71.41%

10. **Mean Discriminant Analysis:** Accuracy : 60.923%

# Conclusion




So, after performing the multiple Classification Algorithms we conclude that Random Forest and C.50 algorithms are the best performing algorithms to classify whether the employee is leaving by choice or is terminated. They have the best accuracy as compared to the other models.