**Group Info:** Kalyan Kumar Paladugula (kpalad4, 679025059)

Michael Ybarra (mybarr3, 659036727)

Zoheb Mohammed (zmoham2, 654090066)

**Problem Statement:** Predict the unemployment rates of the counties for the year 2017 based on the unemployment rates of counties in the year 2015.

**Solution flowchart:**

1. First, we will perform exploratory data analysis to clean the data and to identify the irrelevant variables and redundant variables
2. We will remove the irrelevant variables and redundant variables. Then, we will standardize data.
3. Next, we will partition the dataset as training dataset and the test dataset using hold-out method with 80% percent of data as training and 20% as test data.
4. We will build linear regression models on the 2015 data set using multiple combinations of variables with unemployment rate as the response variable.
5. We pick the best model based on the adjusted R2 values and the root mean square error on the test dataset.
6. Lastly, we use that model to predict the unemployment rate of the counties for the year 2017. And we will calculate the mean squared error from the actual unemployment rates.
7. We also want to classify the counties into 3 classes in terms of unemployment:
   * + - 1. 12% and over (Worst)
         2. 8% to 12% (Moderate)
         3. 8% and below (Best)

and we will build a classifier to predict the classes of the counties in terms of unemployment for the year 2017.

**Data Sources:**

We will use two datasets:

1. US Census Demographic dataset for the year 2015 from Kaggle.com
2. US Census Demographic dataset for the year 2017 from Kaggle.com

Reference: <https://www.kaggle.com/muonneutrino/us-census-demographic-data>