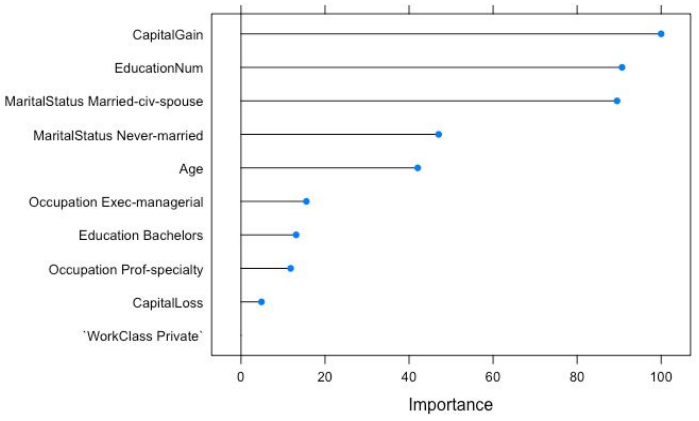
PCS5024
Angel Felipe Magnossão de Paula - 11030561
Resume of The Unreasonable Effectiveness of Data

In this project we analyze a U.S. census data taken from the UCI (University of California at Irvine) Machine Learning Repository. The project is divided into Three parts: **Cleaning and Preprocessing the Data**, **Exploratory Data Analysis** and **Predictive Analysis**. Our final goal is to build a KNN (K Nearest Neighbors model to predict whether the income of a random adult American citizen is less or greater than 50000$ a year based on given features.

A) Cleaning and Preprocessing the Data

1. Set the work directory.
2. Load libraries.
3. Load all the datasets (train and test).
4. Read the training and test data.
5. Cleaning the missing data:
   - Original data: (train=32561, test=16281)
   - Clean data: (train=30162, test=15060)
6. Add title to the all columns (train and test).
7. Standardise the column “IncomeLevel” features as “ <=50K" " >50K" for train and test.



B) Exploratory Data Analysis

1. Apply a Chi-squared test and get p-value for all the features.
2. All of the Pearson’s chi-square tests give very small p-values, which means that it is very likely for the considered categorical variables to be related with “IncomeLevel”.
3. Reduce the number of observations to half and set the other half of observations aside for a validation set.
4. Apply feature selection to decide which predictors to keep and which to throw away (otherwise the train will take a very long time to finish).
5. Apply varImp() to determine which variables I want to include.
6. Plot varImp()
7. Select the 5 top features to train/validation/test (CapitalGain, MaritalStatus, EducationNum, Age, and Occupation).

C) Predictive Analysis

1. Training the model varying k (the process selected k = 9 as it gave the highest accuracy on bootstrapped resamples of the test data).
2. We train your model on your training data set and then use the validation set to estimate out-of-sample accuracy and then retrain a new model.
3. The model correctly predicted the validation set outcome 86.04% of the time.
4. The model correctly predicted the test outcome 84.01% of the time.
5. Plot KNN