**Assignment-1**

**By.: Szabados Dávid**

**Introduction**

While building models for prediction analysis, individuals may unconsciously use too many predictor variables, which may lead to overestimations in the actual data. Using too few predictor variables may lead to underestimations, then which one is better? During this assignment, the *cps-earnings* dataset will be used, to build four basic prediction models, from very simple ones to more complex ones. The models will be compared based on their performance and their complexity, to get a general understanding about the pros and cons of simple and complex models.

**Data & Data Cleaning**

Personal Care and Service occupations were selected from the dataset as the base of the models. Most variables, while being represented as numbers, are qualitative variables, with many possible values. These will be transformed into dummy, True/False variables:

* grade92: True if the value is 41 or more *(has an associate degree or higher)*
* unionmme: True if the person is a union member.
* Private: True if the person works at a private firm.
* race: True if the person’s race is white.
* marital: True if the person has been married at least once.
* prcitshp: True if the person is a native USA citizen.
* sex: True if the person’s gender is male.

Other variables that’s been used for the models:

* age: The age of the person.
* ownchild: The numbers of own child in the primary family.