**Part 1 – Python Test**

This US Census dataset contains detailed but anonymized information for ~300,000 persons.

The archive contains 3 files:

* A large learning .csv file
* Another test .csv file
* A metadata file describing the columns of the two above mentioned files (identical for both)

The goal of this exercise is to analyze and predict the information contained in the last column (42nd), i.e., a person makes over $50,000 / year, from the information contained in the other columns.

Ideally, you’ll work with Python (3.6 or 3.7) to carry out the following steps:

* Import the learning and text files
* Based on the learning file:
  + Make a quick statistic based and univariate audit of the different columns’ content and produce the results in visual / graphic format.
  + This audit should describe the variable distribution, the % of missing values, the extreme values, and so on.
  + Create a model using these variables (you can use whichever variables you want, or even create you own; for example, you could find the ratio or relationship between different variables, the binarization of “categorical” variables, etc.) to determine an income >= $50,000 / year (binary target).
  + You are free to choose any models as long as there’s more than 2 different types.
  + Choose the model that appears to have the highest performance based on a comparison between reality (the 42nd variable) and the model’s prediction.
  + Apply your model to the test file and measure its real performance on it (same method as above).

The goal of this exercise **is not to create the best or the purest model**, but rather to describe the steps you’ll take to accomplish it.

Explain the places that may have been the most challenging for you.

Find clear insights on the profiles of the people that make more than $50,000 / year. For example, which variables seem to be the most correlated with this phenomenon?

**Part 2 – General Questions**

Answer the following questions. Make sure to explain your responses.

1. A, B and C are events and P(X) denotes the probability of X. Also, A implies B and B implies C.
   1. P(A) ≤ P(C).
   2. P(A) ≥ P(C).
   3. P(B ⋂ C) ≥ P(B).
   4. P(B ⋂ C) ≤ P(B).
   5. More information is needed to decide whether c) and d) hold or not.
2. What can you say about this formula ?
   1. .
3. A magician has two cards: one is white on both faces and the other is black on one side and white on the other. The magician picks one card at random (with probability 1/2 for each) and throws it on the floor in a way that does not favor a card side. The card on the floor is white on its visible side. **What is the probability that the other side is white as well?**
4. For each of these statements about binary classification and the various measures related to confusion matrices, indicate if you agree or not, and explain why:
   1. If two models have different accuracy values, the one with a higher accuracy is better.
   2. Type II errors are less problematic than type I errors.
5. As part of a game’s development process, we invite 32 subjects randomly selected from a list of volunteers to play the game in its current state and ask them to score the game on a scale of 0 to 10 on 20 different criteria (movement, camera, shooting, story, difficulty, etc.). We repeat the same process one month later with the same subjects and the new current state of the game. Among the t-tests performed independently on each of the 20 score differences, one of them (the movement score) detected a positive significant difference in score at the 0.05 level. We report that the new version of the game displays significant improvements regarding character movement appreciation. **What are the potential flaws affecting the internal and external validity of this study?**
6. A model for detecting cheater accounts in an online game is tested on 1000 thoroughly verified accounts containing 20 cheaters and 980 non-cheaters. The model flags 25 cheater accounts, 5 of which are actual cheaters. **Compute the following values associated with the confusion matrix**: Accuracy, Precision, Recall.
7. Write a pseudo-code chunk that finds the maximum of two numbers. You cannot use if-else or any other comparison operator.
8. Enumerate the tunable hyper-parameters of a feed-forward classification neural network with fully connected layers, a cross-entropy loss function with an elastic net regularization term using stochastic gradient descent with momentum and linearly decreasing learning rate.