Assignment 3

Applied Machine Learning

From the Kaggle web site (https://www.kaggle.com/datasets) download the `Suicide Rates Overview 1985 to 2016` dataset. This dataset has 12 features and 27820 data points. In this assignment we would like to develop a machine learned model to predict, given some feature vectors, if the outcome would be suicide or not, as a binary dependent variable. The binary categories could be {"low suicide rate", "high suicide rate"}. (Note that a different approach could seek to generate a numerical value by solving a regression problem.)

A machine learning solution would require us to pre-process the dataset and prepare/design our experimentation.

Load the dataset in your model development framework (Jupyter notebook) and examine the features. Note that the Kaggle website also has histograms that you can inspect. However, you might want to look at the data grouped by some other features. For example, what does the `'number of suicides / 100k'` histogram look like from country to country?

To answer the following questions, you have to think thoroughly, and possibly attempt some pilot experiments. There is no one right or wrong answer to some questions below, but you will always need to work from the data to build a convincing argument for your audience.

1. [10 pts] Due to the severity of this real-world crisis, what information would be the most important to "machine learn"? Can it be learned? (Note that this is asking you to define the big-picture question that we want to answer from this dataset. This is not asking you to conjecture which feature is going to turn out being important.)

2. [10 pts] Explain in detail how one should set up the problem. Would it be a regression or a classification problem? Is any unsupervised approach, to look for patterns, worthwhile?

3. [20 pts] What should be the dependent variable?

4. [20 pts] Find some strong correlations between the independent variables and the dependent variable you decided and use them to rank the independent variables.

5. [20 pts] Pre-process the dataset and list the major features you want to use. Note that not all features are crucial. For example, country-year variable is a derived feature and for a classifier it would not be necessary to include the year, the country and the country-year together. In fact, one must avoid adding a derived feature and the original at the same time.
   List the independent features you want to use.

6. [20 pts] Devise a classification problem and present a working prototype model. (It does not have to perform great, but it has to be functional.)
   Note that we will continue with this problem in the following modules.