Problem Set 2 - Data preparation

NB

1) Which programming languages to use?

You can use Python, R or both of them (but with one limitation: one dataset - one language).

2) What libraries/packages to use?

You are free to choose any appropriate data processing libraries (good choice would be **pandas**, **numpy** or **pyspark** for Python and **dplyr** or **tidyr** for R).

3) How to summarize my homework?

The best way is to create an individual Jupyter/R notebook with code and explanations for each dataset. In case you are not familiar with these tools, you can create a Python/R scripts and write explanations as comments. However, we strongly recommend you to use Jupyter/R notebooks, as those are #1 tools in applied data analysis nowadays.

Tasks

1) Data integration (Ukrainian Vehicles Data Set)

1.1. Download the dataset from Ukrainian Open Data Portal **(dataset)**.

1.2. Read the data, remove quotes, trailing whitespaces and tabs.

1.3. Find and remove duplicates (if any),

1.4. Save a dictionary of unique operation codes and its meanings in external file (op_codes.txt). Column **oper_name** is redundant and we probably won't need it in our data analysis, but it could be a good reference to better understand our data. Drop this column and think about any other redundant columns. If you find any, explain why you think of them being redundant?

1.5. Translate at least 3 attributes' values to english (color, kind, etc.).

1.6. Think about which column names are not clear enough and change them.

1.7. What is the possible usage of this data? Give 2-3 examples?

1.8. Save the result of your work as a regular .csv file.

1.9. What is the most popular car and car color in Ukraine?

2) Data cleaning (Census-Income Data Set)

2.1. Download the dataset from UCI Machine Learning Repository **(dataset)**.

2.2. Read the data and add attribute headers (they are absent in the original .data file).

2.3. Investigate the attributes for wrong data, fix the typos in classes' names, standartidize it's representation (so later it would be easier to convert them to boolean cagetorical attributes).

2.4. Fill the missing values in categorical attributes (missing values are denoted by '?') and explain which method you have used and why for each column.

2.5. There are **a lot of** missing values in columns **capital-gain** and **capital-loss**. Investigate what is the best way to deal with them, and apply it.

2.6. Check continuous attributes for outliers, and if you find any - propose the way to deal with them.

2.7. Chose some ML model you would like to use to solve the classification problem with this dataset and explain the next steps you would take to make your data ready for modeling (3-4 sentences).

3) Data normalization (Pima Indians Diabetes Database)

3.1. Download the dataset from Kaggle **(dataset)**.

3.2. Investigate the attributes and propose which of them need to be normalized? Explain your choice.

3.3. Perform normalization using the most appropriate method. Why did you choose it? How can you check that it works better than others?

3.4. Why do we need to normalize numerical attributes?

3.5. What is standartization and how it differs from normalization?