Applied Probability and Statistics

Software and Data Engineering

Singidunum University, Belgrade

Instructions for the Final Project

1. Each student should find, according to the individual interests, a dataset on which she/he will apply adequate methods of statistical reasoning, learnt during the course (all the students should choose **different datasets**). Datasets can be found e.g. using the new google dataset search engine: https://toolbox.google.com/datasetsearch, or directly from the popular websites e.g.: https://www.kaggle.com/, https://archive.ics.uci.edu/ml/datasets.html, https://datahack.analyticsvidhya.com/contest/all/ ...

   The chosen dataset should have **at least 3 dependent variables** (columns in *Pandas DataFrame* structure).

2. Based on the inital analysis of data, the students should make an assesment of which statistical characteristics of the data would be the most interesting and useful to obtain. Based on that, each student should choose at least 5 of the following statistical methods (covered during the course) to extract the wanted information:

   Application of regression analysis (linear, multiple, as well as nonlinear (e.g polynomial)) is mandatory!

   The rest (as least 4) methods should be chosen from the following list (other methods, not covered during the course (e.g. more advanced machine learning methods), are also acceptable):

   - Estimation of the unconditional and conditional probabilities,
   - Estimation of the parameters of distributions of single variables (mathematical expectation, variance, standard deviation, median, quartiles, quantiles,...),
   - Estimation of the parameters of the joint distributions (covariance, correlation coefficient),
   - Estimation of PDF based on histogram, and nonparametric hypothesis testing of distributions using Kolmogorov-Smirnov and/or chi-squared test,
   - Estimation of confidence intervals of certain parameters,
   - Parametric hypothesis testing.

3. Results of the project should be contained in one Jupyter Notebook (.ipynb) file, and will be orally defended (5-10 min).