Instruction Guide

(Data Visualization and Decision Trees)

This instruction is following what you have learned during lectures 3 (Data Visualization) and lecture 4 (Decision Trees). The instruction includes two parts:

- **Technical (Python)**
- **Theoretical (Paper & Pen)**

Technical (Python)

In the technical part, you will learn how to use python as a powerful tool for data visualization and making decision trees. Before starting this part, you should make sure that you have installed following libraries:

- Pandas
- Seaborn
- Numpy
- Matplotlib
- Scipy
- Sklearn
- Subprocess
- Graphviz

If you have installed “Anaconda”, all these libraries should be installed on your system along with “Anaconda”. You can check availability of a library by the following command:

- python -c "import <name of library>"

For example to check if “Numpy” is installed or not you can use the following command:

- python -c "import numpy"

In case there is any missing library, you can install it by the following command:

- pip install <name of library>

If you have not installed “Graphviz” package, in the following link you can find the right package with respect to your operation system.

- https://www.graphviz.org/download/

Data visualization includes following topics:

- Basic statistical analysis
- Simple plots by Numpy/Pandas
- Generating sample data
- Box plots
- Distributions

- Plotting univariate distributions
- Plotting bivariate distributions
- Pair plot

And decision tree part includes following topics:

- Loading data
- Identifying descriptive and target features
- Identifying parameters to make the desired tree (algorithm, pruning, etc)
- Using “Graphviz” to visualize the resulted tree

Theoretical (Paper & Pen)

In this part we will solve some theoretical questions together. This part includes some decision tree questions, and answers. For example, we will make a decision tree based on ID3 algorithm, and we will see some exercises about how one should deal with numerical data as descriptive or target feature.