UNIVERSITY OF MANNHEIM

Data and Web Science Group
Prof. Dr. Christian Bizer
B6 – B1.15
68159 Mannheim

Data Mining – FSS 2020

Exercise 1: Simple Pre-processing and Visualization

1.1. RapidMiner Installation

Download RapidMiner and install the software on your laptop using the educational license:
https://rapidminer.com/get-started-educational/

1.2. Load and Pre-process the Students Dataset

Import the *students* data set into RapidMiner. The *students* data set is provided in ILIAS as an Excel file. Use different pre-processing operators and plotters to answer the following questions:

1. What is the most common mark that has been given in FSS2010? To find the answer filter the examples using a `Filter Examples Operator` and draw a histogram afterwards.
   Solution: Filter the data using the *Filter Examples* Operator with the Filter String "Semester=FSS2010" and visualize it, using a histogram.

   Answer: 2.0

2. Is there a correlation between the mark and the number of attended classes? Find the answer using a scatter plot.
   Solution: Display the filtered data set using a scatter plot (x-axis = attended, y-axis = mark).

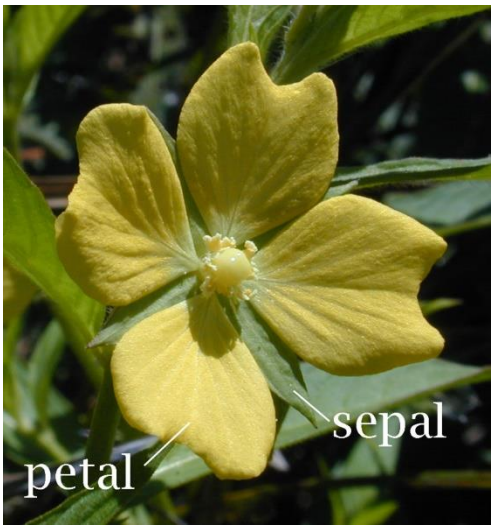
   Answer: Yes, students that attend a lot of classes usually get better marks.

3. Does this correlation hold for all students? Find the answer by aggregating the examples by student and use a scatter plot afterwards.
   Solution: Use the *Aggregate* Operator. Aggregate on *attended* and *grade*. Group the data based on *name.* Reuse the scatter plot from 1.2.2 and put the student name on the colour-axis.

   Answer: There is one outlier. Mariano Selina is getting very good marks without attending too many classes. All other students follow almost the rule "the more attended classes, the better the grade."

UNIVERSITY OF MANNHEIM

1.3. Visual Exploration of the Iris Dataset



The data set describes three types of Iris flowers:

- Setosa
- Virginica
- Versicolour

There are four (non-class) attributes

- Sepal width and length
- Petal width and length

Retrieve the Iris data set from the Samples repository. Use different plotters to visualize and explore the data set.

1. Which attribute combination and (approximate) value ranges determine the type of Iris flower?
   Solution: Use plot Scatter Multiple with x-axis = label and y-axis = a3, a4

   Or: use plot Scatter with x-axis=a3, y-axis=a4 and colour column=label

Answer:

Type of Iris Flower	Attribute combination and value ranges
Setosa	A4 < 0,8
Virginica	A3 > 4,75 A4 > 1,5
Versicolour	2,75 < A3 < 4,8 0,9 < A4 < 1,5