Data Clustering:
Grouping Higher Education Students

In this case, you have to cluster higher education students based on their academic performance, and use the clustering results to provide some recommendations for improvement of the study process.

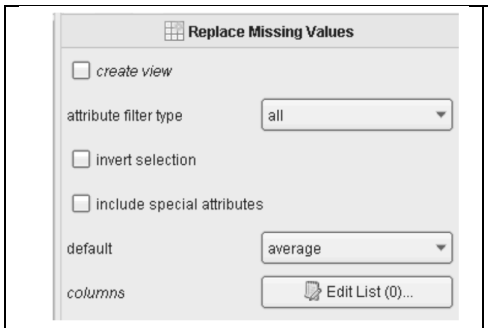
Data-exploration: The student database consists of 366 records about graduated students and their performance.

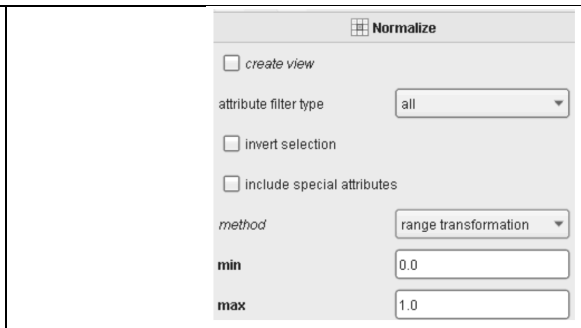
Attribute	Attribute Description
Sex	Student's sex Value: male/female type: binominal
Region	Region of the country from which student comes type: nominal
Score on entrance qualification exam	Points achieved in the entrance qualification exam for Faculty Values: 40–100 type: real
Grades at the first year of studies	Grades: marks on each of 11 exams in the first year of studies Values: 6–10 type: integer
Average grade	Average grade of the student after graduation Values: 6–10 type: continuous
Students' academic performance	Student's academic performance at the end of studies Values: "Bad", "Good", and "Excellent" type: polynomial

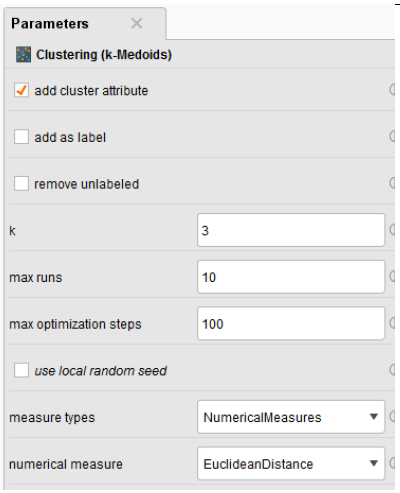
- Attributes include the performance on the first-year exams (11 attributes), average grade of the student after graduation, number of points on entrance examination, sex, and the region of the country from which the student originates.
- Students' academic performance: attribute is discretized from the "Average grade" where all students that had an average grade below 8 are classified as "Bad", between 8 and 9, "Good", and between 9 and 10 "Excellent.

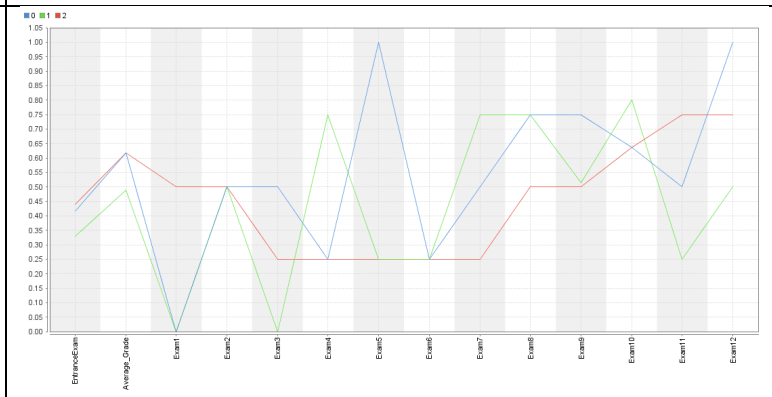
Data pre-processing

a) Use Set-role (batch role) for attribute students' academic performance => this attribute will not be used in the clustering process, but used for the analyses and the discussion of the clustering results.
b) Use Select Attribute operator to remove "Sex", "Region" from the original dataset. The "Invert selection" option is activated and so all the selected attributes are removed from further process.









c) Replace missing values, Normalize Attributes

Data clustering

d) Cluster students using best clustering methods

e) Explain each cluster

f) Suggest strategies for improving performance of students