Guidelines for the task on Data Understanding

- Data understanding (30 points)
  1. Data semantics (3 points)
  2. Distribution of the variables and statistics (7 points)
  3. Assessing data quality (missing values, outliers) (7 points)
  4. Variables transformations (6 points)
  5. Pairwise correlations and eventual elimination of redundant variables (7 points)

Guidelines for the task on clustering

- Clustering Analysis by K-means: (13 points)
  1. Choice of attributes and distance function (1 points)
  2. Identification of the best value of k (5 points)
  3. Characterization of the obtained clusters by using both analysis of the k centroids and comparison of the distribution of variables within the clusters and that in the whole dataset (7 points)
- Analysis by density-based clustering (9 points)
  1. Choice of attributes and distance function (2 points)
  2. Study of the clustering parameters (2 points)
  3. Characterization and interpretation of the obtained clusters (5 points)
- Analysis by hierarchical clustering (5 points)
  1. Choice of attributes and distance function (2 points)
  2. Show and discuss different dendograms using different algorithms (3 points)
- Final evaluation of the best clustering approach and comparison of the clustering obtained (3 points)

Guidelines for the task on Association Rules Mining

- Frequent patterns extraction with different values of support and different types (i.e. frequent, close, maximal), (6 points)
- Discussion of the most interesting frequent patterns and analyze how changes the number of patterns w.r.t. the min_sup parameter (7 points)
- Association rules extraction with different values of confidence (6 points)
- Discussion of the most interesting rules and analyze how changes the number of rules w.r.t. the min_conf parameter, histogram of rules' confidence and lift (7 points)
- Use the most meaningful rules to replace missing values and evaluate the accuracy (2 points)
- Use the most meaningful rules to predict the target variable and evaluate the accuracy (2 points)

Guidelines for the task on Classification

- Learning of different decision trees/classification algorithms with different parameters and gain formulas with the object of maximizing the performances (12 points)
- Decision trees interpretation (6 points)
- Decision trees validation with test and training set (6 points)
- Discussion of the best prediction model (6 points)

Guidelines for the Project

- Title page is not counted in the 20 page limits, i.e., you can have 20 pages + 1 title page, the page limit is strict: additional pages will not be considered for the final evaluation, i.e., pages 21,22,23 etc. will not be read and evaluated.
- The project size must not exceed 25Mb, i.e. you must be able to send it by email without compression.

- Only PDF file are allowed, you do not have to submit python code or the knime workflows.
- The final paper must be easily readable, i.e., it is better to use font size higher than 9pt.
- Use a readable font type and size, e.g. Arial, Times New Romans
- You can use multiple columns and change the margin size but the project must be readable.
- It is NOT required to put python code, knime flows, or theoretical descriptions of the algorithm in the final paper.
- You must justify every choice you make with respect to the features used and selected for each algorithm and the parameters you tune. Discuss every result. Plots without any comment are useless. Even if you find a top configuration for your algorithm (e.g. K-Means with k=5) you MUST list which are the different parameters you tested and justify your choice.
- You can get 3 additional extra points in the final mark with respect to the following criteria:
  1. Innovation (0.5 points)
  2. Experimentation (0.5 points)
  3. Performance (0.5 points)
  4. Appearance (0.5 points)
  5. Organization (0.5 points)
  6. Summary (0.5 points)