Tampere University
Faculty of Information Technology and Communication Sciences

Data.ml.320 Knowledge Mining and Big Data (5 cr)

Exam 8.5.2025 / Ari Visa

Please, do not forget the Kaiku-feedback!

1. Define terms: Data warehouse, Data Lake, and Knowledge Mining An example of a definition: **Data:** Facts and things certainly known. Data are any facts, numbers, or text that can be processed by a computer.	6p
2. Why do you use HADOOP? Are there alternatives to HADOOP? What is HADOOP? How do you use HADOOP when solving a clustering problem? What is the relation between cloud computing and HADOOP?	6p
3.What do you know about associative analysis?	6p
4. What is the difference between classification and prediction? You have only 2 labeled samples of type $\mathbf{X} = (x_1, x_2, x_3)^T$, consisting of real numbers. You should make a predictive model. What kind of model do you use? Motivate your answer! What is the robustness of your solution?	6p
5. How do you define cluster analysis? How can you estimate the number of clusters? Motivate your answer. You have 1T ($=10^{12}$) samples of high dimensional data (dimension > 100) available. What kind of clustering method do you use? Motivate your answer! How does computer architecture influence your proposal?	6p