YLIOPISTOTENTTI - UNIVERSITY EXAM

521156S Towards Data Mining

Answer all the questions. Each question is worth 6 points.

1. Explain terms briefly.
   a. Predictive model (1p)
   b. Data pollution (1p)
   c. Numerosity reduction (1p)
   d. Accuracy paradox (1p)
   e. Validation set (1p)
   f. Missing not at random (MNAR) missingness mechanism (1p)

2. Information privacy
   a. Explain the concept of **information privacy** and how it differs from the classical notion of privacy. (2p)
   b. How is data mining a potential **threat** to information privacy? (2p)
   c. What can be done to **protect** information privacy in data mining? (2p)

3. You want to collect data from 100 persons about shopping habits.
   a. How should you choose the persons to make a model that works for everyone? (2p)
   b. Make a data collection plan. (4p)

4. How would you divide a dataset for training and testing to ensure model generalizability in the following cases?
   a. The dataset available is small (2p)
   b. Some of the variables are time-dependent (2p)
   c. There is data from multiple individuals (2p)

5. What are the characteristics that define data quality? (6p)