Assignment 1 CIS 4523/5523 (Knowledge Discovery and Data Mining)

Number of problems/points: Ten problems for total of 100 points
Out: January 23, 2020
Due: January 30, 2020 in class (hard copy).

Homework Policies (applicable for all assignments):
1. You are required to do the homework problems in order to pass.
2. Understandability of the solution is as desired as correctness.
3. Penalty for late homework assignments submissions is 20% per day. So, do it on time.
4. Solutions are expected to be your own work. Group work is not allowed unless explicitly approved for a particular problem. If you obtained a hint with help (e.g., through library work, discussion with another person, etc.) acknowledge your source, and write up the solution on your own. Plagiarism and other anti-intellectual behavior will be dealt with severely.

Read sections 1 and 2 in the textbook and answer the following problems.

Problem 1: (10 points). An analyst sets up a sensor network in order to measure the temperature of different locations over a period. What is the data type of the data collected?

Problem 2 (10 points). The same analyst as discussed in problem 1 above finds another database from a different source containing pressure readings. She decides to create a single database containing her own readings and the pressure readings. What is the process of creating such a single database called?

Problem 3 (10 points). An analyst processes Web logs in order to create records with the ordering information for Web page accesses from different users. What is the type of this data?

Problem 4 (10 points). Consider a data object corresponding to a set of nucleotides arranged in a certain order. What is this type of data?

Problem 5 (10 points). It is desired to partition customers into similar groups on the basis of their demographic profile. Which data mining problem is best suited to this task?

Problem 6 (10 points). Suppose in problem 5, the merchant already knows for some of the customers whether or not they have bought widgets. Which data mining problem would be suited to the task of identifying groups among the remaining customers, who might buy widgets in the future?
Problem 7 (10 points). Suppose in problem 6, the merchant also has information for other items bought by the customers (beyond widgets). Which data mining problem would be best suited to finding sets of items which are often bought together with widgets?

Problem 8 (10 points). An analyst collects surveys from different participants about their likes and dislikes. Subsequently, the analyst uploads the data to a database, corrects erroneous or missing entries, and designs a recommendation algorithm on this basis. Which of the following actions represent data collection, data preprocessing, and data analysis? (a) conducting surveys and uploading to database, (b) correcting missing entries, and (c) designing a recommendation algorithm.

Problem 9 (10 points). What is the data type of each of the following kinds of attributes (a) Age , (b) Salary , (c) Zip Code , (d) State of Residence , (e) Height , (f) Weight .

Problem 10 (10 points). An analyst obtains medical notes from a physician for data mining purposes, and then transforms them into a table containing the medicines prescribed for each patient. What is the data type of (a) the original data, and (b) the transformed data. (c) What is the process of transforming the data to the new format called?