Bài tập Lý thuyết môn Khai phá Dữ liệu

Họ và tên: Huỳnh Nguyễn Thế Dân

MSSV: 21110256

Lớp: 21TTH1

Questions

1. An analyst collects surveys from different participants abouts their likes and dislikes. Subsequently, the analyst uploads the data to a database, corrects errorneous or missing entries, and designs a recommendation algorithm on this basis. Which of the following questions actions represent data collection, data preprocessing, and data analysis?

- (a) Conducting surveys and uploading to database;
- (b) Correcting missing entries;
- (c) Designing a recommendation algorithm.

2. What is the data type of each of the following kind of attributes?

- (a) Age;
- (b) Salary;
- (c) ZIP code;
- (d) State of residence;
- (e) Height;
- (f) Weight.

3. 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 processing of transforming the data to the new format called?

Answers

1.

- **Data Collection:** This involves gathering data from various sources. In this question, it includes conducting surveys and uploading the collecteds data to a database. So, the correct answer is **(a) Conducting serveys and uploading to database.**

- **Data Preprocessing:** This involves cleaning and preparing the data for analysis. Correcting missing intries falls under data preprocessing, ass it is a step to ensure the data is accurate and complete before analysis. So, the correct answer is **(b) Correcting missing entries.**
- **Data Analysis:** This involves apply algorithms and techniques to extract insights or make predictions from the data. Designing a recommendation algorithm is part of data analysis, as iot involves creating a model to make recommendations based on the collected and preprocessed data. So, the correct answer is **(c) Designing a recommendation algorithm.**

2. The data type of each attribute can vary depending on how the data is represented and used in the context of analysis.

- **(a) Age:** Integer or numerical data type. Age is typically represented as a whole number.
- **(b) Salary:** Floating-point or numberical data type. Salary can have decimal values and is usually represented as a number.
- **(c) ZIP code:** Categorical or nominal data type. ZIP codes are typically represented as integers, but they are categorial in nature and do not have numerical significance beyond identification.
- **(d) State of residence:** Categorical or nominal data type. State of categorial variables and are usually represented as string or codes.
- **(e) Height:** Floating-point or numerical data type. Height can have decimal values and is typically represented as a number.
- **(f) Weight:** Floating-point or numerical data type. Weight can have decimal values and is typically represented as a number.

3. Based on the question provided:

- **(a) Original Data:** The original data obtained from the physician's medical notes is likely to be unstructured text data, as it consists of free-form notes written by the physician. Unstructured text data does not fit neatly into rows and columns like structured data does.
- **(b) Transformed Data:** The transformed data is in the form of a table containing the medicines prescribed for each patient. This data is likely to be structured data, where each row represents a patient and each column represents a variable (such as patient ID, medicine prescribed, dosage, etc.). The medicines prescribed would likely be categorical data.
- **(c) Processing to Transform Data:** The processing of transforming the unstructured text data into a structured table format is called *text mining or text preprocessing*. This involves techniques such as text parsing, extracting relevant information, and organizing it into a structured format suitable for analysis. In this case, specifically, the transformation process involves extracting the medicines prescribed from the medical notes and organizing them into a table format.

End.