Assignment - 6

Transforming Different Data Types

HARSH SIDDHAPURA

1230169813

02/11/2024

PROBLEM - 1

Definitions and characteristics of each type of attribute commonly encountered in datasets are:

1. **Categorical (Nominal):**
    - **Definition:** Categorical attributes represent distinct categories or labels. They do not have any inherent order or magnitude.
    - **Examples:** Eye color (e.g., brown, green, blue), gender (e.g., male, female, non-binary), country of origin (e.g., USA, Canada, Japan).
    - **Operations:**
        - Equality: We can compare whether two values are the same or different (e.g., "brown" vs. "green").
        - Counting: We can count the frequency of each category.
        - Mode: We can find the most frequent category.

2. **Ordinal:**
    - **Definition:** Ordinal attributes have ordered categories with meaningful relative positions. However, the intervals between categories are not uniform.
    - **Examples:** Education level (e.g., high school, bachelor's, master's), customer satisfaction ratings (e.g., low, medium, high).
    - **Operations:**
        - Ordering: We can arrange the categories from lowest to highest (e.g., "low" < "medium" < "high").
        - Median: We can find the middle value (median) within the ordered categories.

3. **Interval:**
    - **Definition:** Interval attributes have ordered categories with uniform intervals between them. However, they lack a true zero point.
    - **Examples:** Time, Date
    - **Operations:**
        - Addition and Subtraction: We can perform arithmetic operations (e.g., adding or subtracting temperatures).
        - Mean and Standard Deviation: We can calculate the mean and standard deviation.

4. **Ratio:**
    - **Definition:** Ratio attributes have ordered categories with uniform intervals and a true zero point.
    - **Examples:** Height (in centimeters), weight (in kilograms), income (in dollars).

- **Operations:**
  - All Arithmetic Operations: We can add, subtract, multiply, and divide values.
  - Mean, Median, and Standard Deviation: All statistical measures are applicable.

++

PROBLEM - 2

Classifying each of the given attributes based on their types are as follows:

1. **Height of a person:**
   - **Type: Ratio attribute**
   - **Explanation:** Height is a continuous measurement with a true zero point (i.e., a height of 0 cm is meaningful). We can perform all arithmetic operations (addition, subtraction, multiplication, division) on height values.

2. **Disk space in Giga Bytes:**
   - **Type: Ratio attribute**
   - **Explanation:** Disk space is a continuous measurement with a true zero point (0 GB represents no disk space). We can perform all arithmetic operations on disk space values.

3. **Traffic lights (Red, Green, Yellow):**
   - **Type: Categorical (Nominal) attribute**
   - **Explanation:** Traffic lights represent distinct categories (colors) without any inherent order. We can only compare whether two values are the same or different (e.g., "Red" vs. "Green").

4. **Ethnicity of a person:**
   - **Type: Categorical (Nominal) attribute**
   - **Explanation:** Ethnicity is a categorical attribute with distinct labels (e.g., "Asian," "African American," "Hispanic"). We can only compare whether two ethnicities are the same or different.

5. **Location, expressed using Zip Codes:**
   - **Type: Categorical (Nominal) attribute**

- **Explanation:** Zip codes represent distinct geographic areas without any inherent order. We can only compare whether two zip codes are the same or different.

6. **Location, expressed using Longitude and Latitude:**
   - **Type:**
     - **Longitude: Interval attribute**
     - **Latitude: Interval attribute**
   - **Explanation:** Longitude and latitude are continuous measurements representing positions on the Earth's surface. While they have an ordered scale, the intervals between values are not uniform. We can perform ordering and calculate differences (e.g., distance between two points), but we cannot perform meaningful arithmetic operations like addition or multiplication.

++

PROBLEM - 3

Transformation of raw format to a processing format based on their attribute types are as follows:

1. **Patient Date of Birth (Interval Attribute):**
   - **Raw Format: The patient's date of birth (e.g., "1990-05-15").**
   - **Processing Format:**
     - Convert the raw date to a standardized format (e.g., "May 15, 1990").
     - Set the earliest/smallest date in the dataset as day 0
     - Each other date in the dataset is represented as the number of days away from day 0.
     - This way date differences (in days/time) are preserved.

2. **Patient Allergy (Categorical Attribute):**
   - **Raw Format: The patient's allergy type (e.g., "Medicine Allergy").**
   - **Processing Format:**
     - Check the unique number of values, x.
     - Create x binary attributes, one for each value.
     - Create binary indicator variables (dummy variables) for each allergy type
     - For each data point, set the binary variable corresponding to this data point's value to 1, and set the other variables to 0.

3. **Number of Training Laps per Week (Ratio Attribute):**
    - **Raw Format: The count of training laps (e.g., 5 laps).**
    - **Processing Format:**
        - Ratio attributes are numbers with meaningful differences and ratios.
        - Ensure that the data is in a consistent unit.
        - Ratio attributes are usually kept as they are (no transformation needed).

4. **User Evaluation of a Product (Ordinal Attribute):**
    - **Raw Format: The user's evaluation (e.g., "Good," "Neutral," "Bad").**
    - **Processing Format:**
        - Find unique values from the dataset.
        - Create a dictionary and assign a numerical value as per ranking/order.
        - Since there's an order between the values, then it can be transformed into numeric values provided that the value order is preserved.
        - Domain knowledge can be leveraged for selecting meaningful values.
        - Assign and rank numerical values to each evaluation category (e.g., "Good" → 3, "Neutral" → 2, "Bad" → 1).