**Name: Greendee Roper B. Panogalon BSIT 3B**

1. What is data preprocessing in data mining?

Answer: The data preprocessing in data mining is very important in terms of cleaning, organizing, and transforming data raw. It is fixing missing info, getting rid of repeats, handling odd values, and making sure everything is in a good shape for analysis. This clean-up job is important because it helps computer programs, using data mining algorithms, work more effectively and gives us better and more reliable results.

2. How to handle missing values?

Answer: To handle missing values, we can either remove incomplete entries or fill gaps with averages, adjacent values, or predicted values using techniques like mean imputation or predictive modeling. We can choose the method based on our data's characteristics and specific requirements.

3. What is outlier? what are the factors affected by outliers?

Answer: An outlier is an observation that stands out as unusually far from the other values in a random sample. Outliers can affect various factors such as statistical measures (like mean and standard deviation), regression models (changing coefficients), clustering algorithms (disturbing cluster formation), and certain machine learning models (potentially causing biased predictions).

4. Normalize the following group of data

500,1500,2000,3000,9000, 10500

a. **using min-max normalization by setting min:0 and max:1**

$$V = \frac{x - min}{max - min}$$

$$V = \frac{500 - 500}{10500 - 500} = 0$$

$$V = \frac{1500 - 500}{10500 - 500} = 0.1$$

$$V = \frac{2000 - 500}{10500 - 500} = 0.15$$

$$V = \frac{3000 - 500}{10500 - 500} = 0.25$$

$$V = \frac{9000 - 500}{10500 - 500} = 0.85$$

$$V = \frac{10500 - 500}{10500 - 500} = 1$$

| DATA | NORMALIZED DATA |
|---|---|
| 500 | 0 |
| 1500 | 0.1 |
| 2000 | 0.15 |
| 3000 | 0.25 |
| 9000 | 0.85 |
| 10500 | 1 |

b. **Z-score normalization**

$$Z = \frac{x - \mu}{\sigma}$$

**Mean:**

$$\mu = \frac{(500 + 1500 + 2000 + 3000 + 9000 + 10500)}{6} = \mathbf{4416.67}$$

**Standard Deviation:**

$$\sigma = \frac{(500 + 4416.67)^2 + (1500 + 4416.67)^2 + (2000 + 4416.67)^2 + (3000 + 4416.67)^2 + (9000 + 4416.67)^2 + (10500 + 4416.67)^2}{6 - 1}$$

$$= \mathbf{4212.09}$$

$$Z = \frac{500 - 4416.67}{4212.09} = -0.93$$

$$Z = \frac{1500 - 4416.67}{4212.09} = -0.69$$

$$Z = \frac{2000 - 4416.67}{4212.09} = -0.57$$

$$Z = \frac{3000 - 4416.67}{4212.09} = -0.34$$

$$Z = \frac{9000 - 4416.67}{4212.09} = 1.09$$

$$Z = \frac{10500 - 4416.67}{4212.09} = 1.44$$

| DATA | NORMALIZED DATA |
|---|---|
| 500 | -0.93 |
| 1500 | -0.69 |
| 2000 | -0.57 |
| 3000 | -0.34 |
| 9000 | 1.09 |
| 10500 | 1.44 |