Experiment 7: Data Pre-processing on Customer Dataset

Aim:

To perform **pre-processing** on the Customer dataset, including handling missing values, normalization, discretization, standardization, removing unnecessary attributes, and encoding categorical attributes.

Theory:

- **Data pre-processing** improves dataset quality for data mining tasks.
- Common steps:
  - Handle **missing values**
  - **Normalize** numerical attributes
  - **Discretize** continuous values into categories
  - **Standardize** attributes for uniform scale
  - Remove **irrelevant attributes**
  - Encode **categorical attributes** to numerical values

Dataset (customer.arff)

```
@relation customer

@attribute CustomerID numeric
@attribute Age numeric
@attribute Gender {Male, Female}
@attribute AnnualIncome numeric
@attribute SpendingScore numeric
@attribute Segment {High, Medium, Low}

@data
101,25,Male,50000,70,Medium
102,30,Female,60000,60,High
103,22,Male,35000,40,Low
104,28,Female,58000,80,High
```

```
105,35,Male,45000,50,Medium

106,40,Female,62000,30,Low

...
```

Procedure (Using WEKA):

1. Open **WEKA → Explorer**.
2. Click **Open File** → select **customer.arff**.
3. Go to **Preprocess tab**.
4. **Handle missing values:** Use **Filter → unsupervised → attribute → ReplaceMissingValues**.
5. **Normalize numerical attributes:** Filter → unsupervised → attribute → **Normalize**.
6. **Discretization:** Filter → unsupervised → attribute → **Discretize**.
7. **Standardization:** Filter → unsupervised → attribute → **Standardize**.
8. **Remove unnecessary attributes:** Filter → unsupervised → attribute → **Remove** (e.g., CustomerID).
9. **Encode categorical attributes:** Filter → unsupervised → attribute → **NominalToBinary**.
10. Apply filters step by step and **save processed dataset** if needed.

Result (Sample / Expected):

- All **missing values handled**.
- Numerical attributes (**Age, AnnualIncome, SpendingScore**) **normalized and standardized**.
- Continuous attributes **discretized** into categories.
- **CustomerID removed**.
- Categorical attributes (**Gender, Segment**) encoded numerically.

Conclusion:

- Pre-processing ensures **clean and consistent dataset**.
- Improves **model accuracy and performance**.
- WEKA provides **easy tools** for all pre-processing tasks.