**Experiment 8: Data Pre-processing on Iris Dataset**

---

**Aim:**

To apply **data pre-processing techniques** on the Iris dataset, including handling missing values, normalization, and encoding categorical data.

---

**Theory:**

- **Data pre-processing** improves dataset quality for data mining and machine learning.
- Steps include:
  - **Handling missing values** to avoid errors
  - **Normalizing** numerical attributes for uniform scale
  - **Encoding categorical attributes** (e.g., Species) into numerical values

---

**Dataset (iris.arff)**

```
@relation iris


@attribute SepalLength numeric

@attribute SepalWidth numeric

@attribute PetalLength numeric

@attribute PetalWidth numeric

@attribute Species {Setosa, Versicolor, Virginica}


@data

5.1,3.5,1.4,0.2,Setosa

4.9,3.0,1.4,0.2,Setosa

6.2,3.4,5.4,2.3,Virginica

5.9,3.0,5.1,1.8,Virginica

6.0,2.2,4.0,1.0,Versicolor

5.5,2.3,4.0,1.3,Versicolor

...
```

---

**Procedure (Using WEKA):**

1. Open **WEKA → Explorer**.
2. Click **Open File** → select **iris.arff**.
3. Go to **Preprocess tab**.
4. **Handle missing values:** Filter → unsupervised → attribute → **ReplaceMissingValues**.
5. **Normalize numerical attributes:** Filter → unsupervised → attribute → **Normalize**.
6. **Encode categorical attribute (Species):** Filter → unsupervised → attribute → **NominalToBinary**.
7. Apply filters step by step and **save the pre-processed dataset**.

---

**Result (Sample / Expected):**

- All missing values **handled**.
- Numerical attributes (**SepalLength, SepalWidth, PetalLength, PetalWidth**) **normalized**.
- Categorical attribute (**Species**) **encoded numerically**.

---

**Conclusion:**

- Pre-processing improves **dataset quality and model performance**.
- WEKA makes it **easy to handle missing values, normalize data, and encode categories**.
- Processed dataset is ready for **classification or clustering** tasks.