Synopsis of Dataset and Description of High Dimensionality

Kavish Nag
24070126085
AIML B1

Dataset 1: Iris Dataset
The Iris dataset contains:

- 150 samples
- 4 numerical features
    - Sepal length
    - Sepal width
    - Petal length
    - Petal width
- 3 target classes
    - Setosa
    - Versicolor
    - Virginica

Although Iris is not truly high dimensional, it is commonly used to demonstrate dimensionality reduction because it allows clear visualization before and after transformation.

Dataset 2: Higher Dimensional Dataset

This dataset contains:

- Larger number of features compared to Iris
- Multiple classes
- Possibly correlated features

High dimensionality refers to datasets where:

- Number of features is large relative to number of samples
- Features may be redundant or highly correlated
- Model training becomes computationally expensive
- Risk of overfitting increases
- Visualization becomes difficult

Problems caused by high dimensionality:

- Curse of dimensionality
- Increased variance

- Poor generalization
- Increased training time

Dimensionality reduction techniques like LDA help reduce features while preserving class separability.