Synopsis of Datasets

Kavish Nag
24070126085
AIML B1

1. Iris Dataset

The Iris dataset is a low dimensional dataset often used as a baseline to explain feature spaces and classification behaviour.

Dataset overview

- Total samples: 150
- Total features (dimensions): 4 numerical attributes
  - Sepal length
  - Sepal width
  - Petal length
  - Petal width
- Target classes: 3 species

High dimensionality perspective

- Iris is not considered high dimensional because the number of features is very small compared to the number of observations.
- Feature space is easily visualizable in 2D or 3D projections.
- Algorithms rarely suffer from the curse of dimensionality here.
- It is commonly used to demonstrate dimensionality reduction techniques like PCA since relationships between features are strong and interpretable.

Implication

- Low risk of overfitting due to dimensionality.
- Feature interactions are simple and computational cost is minimal.

---

2. Heart Attack Dataset

Dataset overview

Samples: 1319

- Features: 8 input variables + 1 target label
- Columns:

    - age
    - gender
    - impluse
    - pressurehight
    - pressurelow
    - glucose
    - kcm
    - troponin
    - class (target)

High dimensionality perspective

- This dataset is moderate dimensional, not truly high dimensional.
- The number of features is small relative to sample size, which reduces risk of sparsity problems.

Feature space characteristics

- Contains physiological measurements that may be correlated:
    - pressurehight and pressurelow
    - kcm and troponin as cardiac biomarkers
- Mixed feature distributions increase effective dimensional complexity even with few columns.

High dimensionality considerations
Even with only 8 features, challenges similar to higher dimensional spaces can arise:

1. Scaling sensitivity
   Variables exist on different ranges, affecting distance-based models.
2. Risk of overfitting with complex models
   Using high-capacity models like deep neural networks may simulate high dimensional behaviour.

Implication

- Dimensionality reduction may still help through PCA or feature selection.
- Model interpretability remains feasible compared to truly high dimensional datasets like genomics or text embeddings.

Summary Comparison

Dataset	**Number of Features**	**High Dimensional?**	**Key Insight**
Iris	4	No	Classic low dimensional dataset
Heart Attack	8	Moderate	Biomedical feature interactions increase effective complexity