Report: Comparison of Dimensionality Reduction and Classifier Performance on Breast Cancer Dataset

Dataset Features

1. **Number of Instances (Samples):**
   - **569 samples**
2. **Number of Features:**
   - **30 numeric features (all float type), describing the characteristics of the cell nuclei.**
3. **Target Variable:**
   - **Binary classification:**
     - **0: Malignant (Cancerous)**
     - **1: Benign (Non - Cancerous)**

Objective:

The goal of this experiment was to apply three dimensionality reduction techniques—Self-Organizing Maps (SOM), Restricted Boltzmann Machines (RBM), and Autoencoders—and compare their performance against the original dataset using three classifiers: XGBoost, LightGBM, and Cat Boost. Performance was measured in terms of classification accuracy and execution time.

Result

	Dataset	Classifier	Accuracy	Time (s)
0	Original	XGBoost	0.980952	0.096054
1	Original	LightGBM	0.976190	0.216141
2	Original	CatBoost	0.985714	3.448707
3	SOM	XGBoost	0.966667	0.063982
4	SOM	LightGBM	0.957143	0.027295
5	SOM	CatBoost	0.957143	1.013347
6	RBM	XGBoost	0.976190	0.046851
7	RBM	LightGBM	0.961905	0.062937
8	RBM	CatBoost	0.966667	1.964469
9	Autoencoder	XGBoost	0.966667	0.046875
10	Autoencoder	LightGBM	0.976190	0.068369
11	Autoencoder	CatBoost	0.971429	1.919503

Observations:

1. **Accuracy:**
    - Original Dataset achieved the highest accuracy for all classifiers, with CatBoost performing best at 98.57%.
    - **Dimensionality-Reduced Datasets:**
        - RBM and Autoencoder consistently outperformed SOM in terms of accuracy.
        - Autoencoder-based reduction achieved competitive accuracy, closely matching the original dataset.
2. **Execution Time:**
    - XGBoost and LightGBM demonstrated faster training times compared to CatBoost across all datasets.
    - SOM was the fastest dimensionality reduction technique due to its simplicity but slightly lagged in classification accuracy.
    - Autoencoders and RBMs showed moderate training times, balancing complexity and accuracy effectively.
3. **Dimensionality Reduction Techniques:**
    - **SOM:** Efficient but limited in maintaining feature importance, resulting in lower accuracy compared to RBM and Autoencoder.
    - **RBM:** Achieved higher accuracy with moderately faster training times.
    - **Autoencoder:** Delivered a balance between accuracy and training time, making it a suitable choice for classification tasks.

Conclusion:

- For high accuracy and acceptable training time, the **original dataset** with CatBoost performed best, albeit at a higher computational cost.
- Among dimensionality reduction techniques, **Autoencoder** emerged as the most effective, providing a trade-off between accuracy and speed.
- For scenarios prioritizing speed, **LightGBM with SOM** provided the quickest solution, although with slightly lower accuracy.

This experiment highlights the trade-offs between dimensionality reduction methods and classifiers, emphasizing the importance of selecting techniques based on task-specific requirements such as accuracy and computational efficiency.