# Breast Cancer Classification and Dimensionality Reduction Project

This project explores the application of dimensionality reduction techniques to improve the classification performance for the Breast Cancer Wisconsin (Diagnostic) dataset. We evaluate Principal Component Analysis (PCA), Linear Discriminant Analysis (LDA), and t-Distributed Stochastic Neighbor Embedding (t-SNE) on five classification algorithms. Both pre- and post-dimensionality reduction results are analyzed in detail.

## Dataset Description

The dataset consists of 569 samples with 30 continuous features. The target variable is binary: Malignant (M) or Benign (B). Features are derived from cell nucleus measurements, such as radius, texture, perimeter, and area. This dataset is sourced from the UCI Machine Learning Repository.

## Dimensionality Reduction Techniques

1. PCA: Principal Component Analysis reduces the feature dimensions by transforming the data into a lower-dimensional space while retaining maximum variance.  
2. LDA: Linear Discriminant Analysis projects data to maximize class separability.  
3. t-SNE: t-Distributed Stochastic Neighbor Embedding is a non-linear technique for preserving local relationships in lower dimensions.

## Methodology

The methodology involves the following steps:  
1. Preprocessing: Data scaling and encoding.  
2. Dimensionality Reduction: Applying PCA, LDA, and t-SNE.  
3. Classification: Evaluating Logistic Regression, Random Forest, SVM, KNN, and Naive Bayes.  
4. Performance Metrics: Accuracy, precision, recall, and F1 score.  
5. Cross-validation: Using Stratified K-Fold cross-validation for robust evaluation.

## Results and Analysis

The performance of each dimensionality reduction technique is compared against the baseline (no reduction). The results are visualized using scatter plots and tabulated for clarity. The metrics show the impact of dimensionality reduction on model accuracy.

## Conclusions

Dimensionality reduction techniques can significantly improve classification performance while reducing computational complexity. PCA and LDA generally perform well for linear models, while t-SNE is useful for exploratory analysis.

## References

1. UCI Machine Learning Repository: Breast Cancer Wisconsin (Diagnostic) dataset.  
2. Scikit-learn documentation: https://scikit-learn.org/stable/  
3. Dimensionality Reduction: A Review by Van der Maaten et al.