Phase 2 - Design Document

Objective

To design an ensemble-based classification system using multiple machine learning models to improve prediction accuracy on the Breast Cancer dataset.

Input

- Features from the Breast Cancer dataset (sklearn.datasets.load_breast_cancer)
- Target: Binary classification (Malignant / Benign)

Workflow Design

1. Load and explore data
2. Preprocess and split into training/testing sets
3. Train models:
   - Random Forest
   - Gradient Boosting
   - Bagging
   - Stacking (with SVC + RF as base estimators)
4. Evaluate using accuracy and classification metrics
5. Visualize comparison with plotly and confusion matrix with seaborn

Model Architecture (Stacking)

- Base Models:
  - Random Forest
  - Support Vector Classifier
- Final Estimator:
  - Logistic Regression

Libraries Used

- scikit-learn

- matplotlib, seaborn, plotly

- pandas, numpy

Tools for Visualization

- plotly.express for bar charts

- seaborn for confusion matrix