Breast Cancer Classification Using Ensemble Learning

Project Summary

This project focuses on accurate classification of breast cancer (malignant vs benign) using machine learning and ensemble techniques. The dataset used is the Wisconsin Breast Cancer Diagnostic (WBCD) dataset. The project applies feature selection, model tuning, and ensemble stacking to maximize classification accuracy.

Methodology

1. Feature Selection: Stepwise Linear Discriminant Analysis (LDA) was used to select the top 16 most relevant features.

2. Preprocessing: StandardScaler was used to normalize feature values.

3. Models Used:
   - Logistic Regression
   - Random Forest
   - XGBoost
   - LightGBM
   - CatBoost

4. Ensemble Techniques:
   - StackingClassifier: Combines all 5 models using Logistic Regression as the meta-learner.
   - VotingClassifier: Soft voting was used to average predictions from top models.

5. Model Tuning: GridSearchCV was applied to optimize hyperparameters for LR, XGBoost, and LightGBM.

Results

The individual model accuracies were all above 97%, with Logistic Regression, LightGBM, and CatBoost achieving up to 98.25%. The Stacked Ensemble model also reached an accuracy of 98.25% on the test set.

10-Fold Cross-Validation confirmed stable performance with a mean accuracy near 98.5%.

The optimized VotingClassifier with tuned models produced the highest cross-validated accuracy: ~99%.

Scope for Improvement

Potential improvements include:

- Use of SHAP for model explainability
- Feature selection using SHAP or RFECV
- External dataset testing for generalization
- Integration of a neural network into the ensemble
- Deployment of the model in a web app using Flask or Streamlit
- Use of AutoML (TPOT) to explore additional pipelines