Breast Cancer Classification Using Ensemble Learning

Step-by-Step Project Pipeline: Breast Cancer Classification

1. Dataset Collection

The dataset used in this project is the Wisconsin Breast Cancer Diagnostic (WBCD) dataset, which contains 569 samples with 30 numerical features and a binary label: 'M' (malignant) or 'B' (benign).

2. Data Preprocessing

- Removed unnecessary columns such as 'id' and 'Unnamed: 32'.
- Mapped target labels: 'M' to 1, 'B' to 0.
- Checked and confirmed no missing values.
- Applied StandardScaler to normalize feature values.

3. Feature Selection

Used Stepwise Linear Discriminant Analysis (LDA) to select the top 16 most relevant features that contribute significantly to class separation.

4. Train/Test Split

Split the data into training and testing sets using an 80-20 split ratio with stratification to maintain label balance.

5. Model Training

Trained the following machine learning models on the training data:

- Logistic Regression
- Random Forest
- XGBoost
- LightGBM
- CatBoost

6. Hyperparameter Tuning

Applied GridSearchCV to tune hyperparameters for Logistic Regression, XGBoost, and LightGBM, improving their performance using 5-fold cross-validation.

7. Ensemble Methods

- VotingClassifier (Soft Voting): Averaged the predicted probabilities of top-performing models.
- StackingClassifier: Combined all five models using Logistic Regression as the meta-learner.

8. Model Evaluation

Evaluated models using accuracy score and 10-fold cross-validation.

- Individual models achieved 97 to 98.25 percent accuracy.
- Ensemble models reached up to 99 percent accuracy.

Also evaluated using confusion matrix and classification report.

9. Conclusion

The stacked and voting ensembles produced the most accurate and robust results. The model pipeline is modular and can be extended with explainability, neural networks, or deployment.