ML ASSIGNMENT-2

REPORT

Name: Kandhuri Sai Rasagna
Roll Number: 160123737151
Date: October 4, 2025
Github Repository: Machine-Learning-Practice/Assignment-2 at main · rasagna2320/Machine-Learning-Practice

Overview:

This assignment explores ensemble learning techniques applied to disease prediction, specifically breast cancer tumor classification. The goal is to fill a research gap identified in the paper “Enhancing Disease Prediction through Ensemble Learning Techniques” (IEEE), which lacked detailed analysis of hyperparameter tuning effects. Utilizing the Breast Cancer Wisconsin dataset, ensemble models were trained and tuned to evaluate improvements in classification accuracy and robustness.

Technologies & Tools Used:

- Dataset: Breast Cancer Wisconsin dataset (scikit-learn)
- Programming Language: Python 3.x
- Libraries: scikit-learn, pandas, numpy, matplotlib, seaborn
- Modeling: Random Forest, Gradient Boosting, Stacking Classifier
- Hyperparameter Tuning: GridSearchCV
- Visualization: Matplotlib, Seaborn

System Architecture:

1. Data loading and preprocessing (encoding, scaling) of breast cancer tumor features.
2. Baseline ensemble models trained with default hyperparameters.
3. Hyperparameter tuning implemented for Random Forest and Gradient Boosting models.
4. Tuned models combined in a stacking classifier.
5. Model evaluation with accuracy, confusion matrix, classification report, and feature importance visualization.

Folder Structure:

Assignment-2/

```
├—Notebooks                # Folder
    ├—ml-assignment-2.ipynb   #The actual code
├— requirements.txt        # Dependencies list
├— README.md               # Description
└— Report.pdf              # Final Report
```

Testing & Results:

- Tuned ensemble models improved accuracy up to 95.6% on test data.
- Hyperparameter tuning significantly optimized Random Forest and Gradient Boosting classifiers.
- Stacking classifier with tuned base models performed comparably, demonstrating ensemble synergy.
- Feature importance highlighted key tumor metrics like mean radius and texture influencing prediction.

Deliverables:

- GitHub Repository with code, data preprocessing, modeling, and results and report.
- Kaggle Notebook demonstrating stepwise data analysis, model training, tuning, and evaluation.
- This report documenting research gap, methodology, results, and conclusions.

Conclusion:

This project successfully addressed the gap by rigorously analyzing hyperparameter tuning effects across ensemble classifiers for disease prediction. The Breast Cancer Wisconsin dataset served as a suitable testbed, with tuned models achieving strong predictive accuracy. The approach lays groundwork for applying tuned ensembles in other healthcare prediction tasks.