Task 1: ML Model Building – Detailed Report

1. Problem Overview

The objective of Task 1 is to build a complete end-to-end classification pipeline using Scikit-learn or TensorFlow on a public dataset. For this assignment, the Breast Cancer Wisconsin dataset from Scikit-learn was selected due to its popularity, real-world relevance, and suitability for demonstrating classification and explainability techniques.

2. End-to-End ML Workflow

The following workflow was implemented:

- Data loading using sklearn.datasets
- Train-test split (80% training, 20% testing)
- Feature scaling using StandardScaler
- Baseline model training (Logistic Regression, Random Forest, Gradient Boosting, XGBoost)
- Performance evaluation using Accuracy, Precision, Recall, F1 Score, ROC-AUC
- Hyperparameter tuning for XGBoost using RandomizedSearchCV
- SHAP Explainability to interpret feature impact
- Model and scaler saved using joblib
- Confusion Matrix, ROC Curve, and SHAP Summary Plot generated

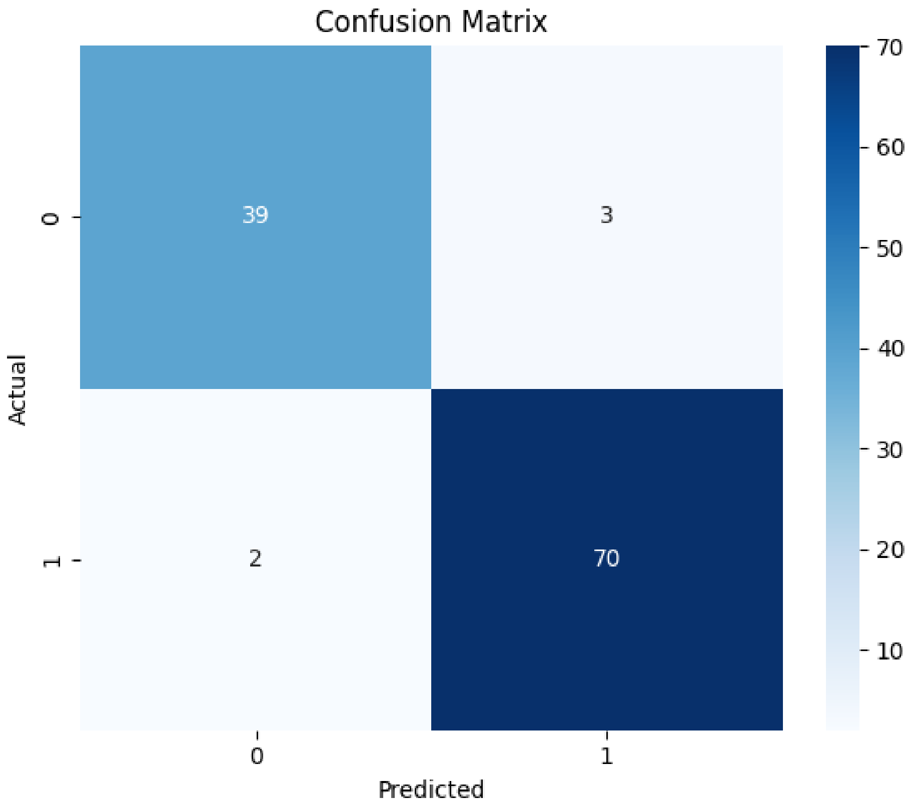
3. Model Performance Summary

After evaluating multiple baseline models, the tuned XGBoost model achieved the highest performance.

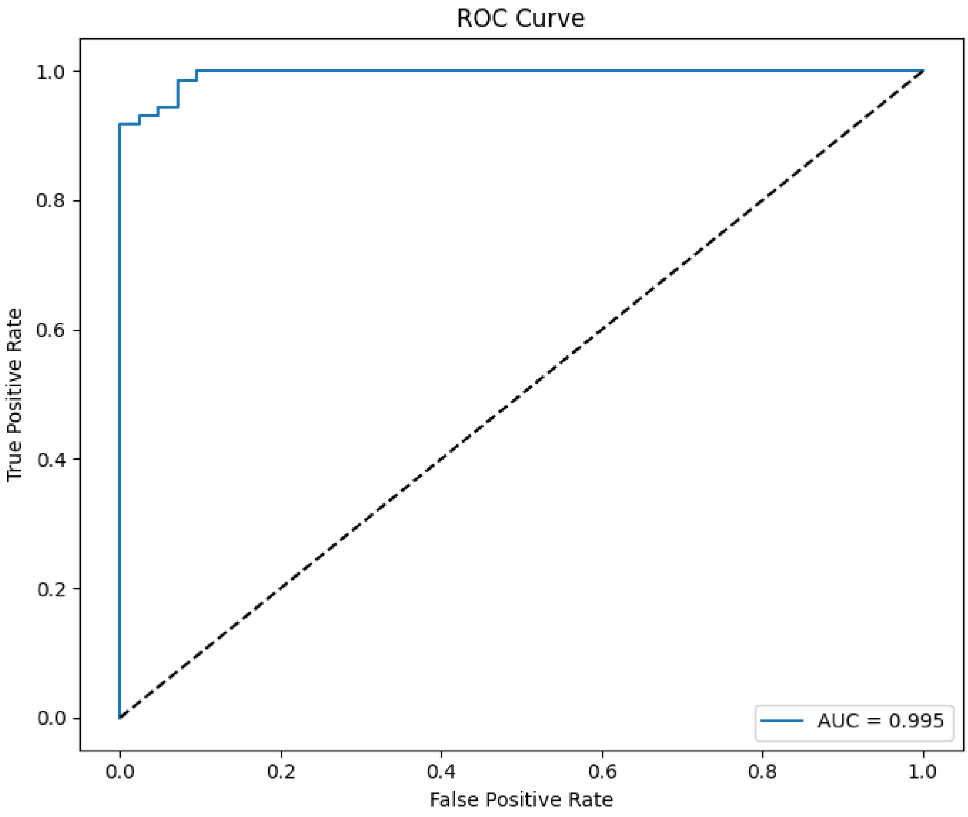
Final Evaluation Metrics:

- Accuracy: 0.9561
- Precision: 0.9589
- Recall: 0.9722
- F1 Score: 0.9655
- ROC-AUC: 0.9947

4. Confusion Matrix

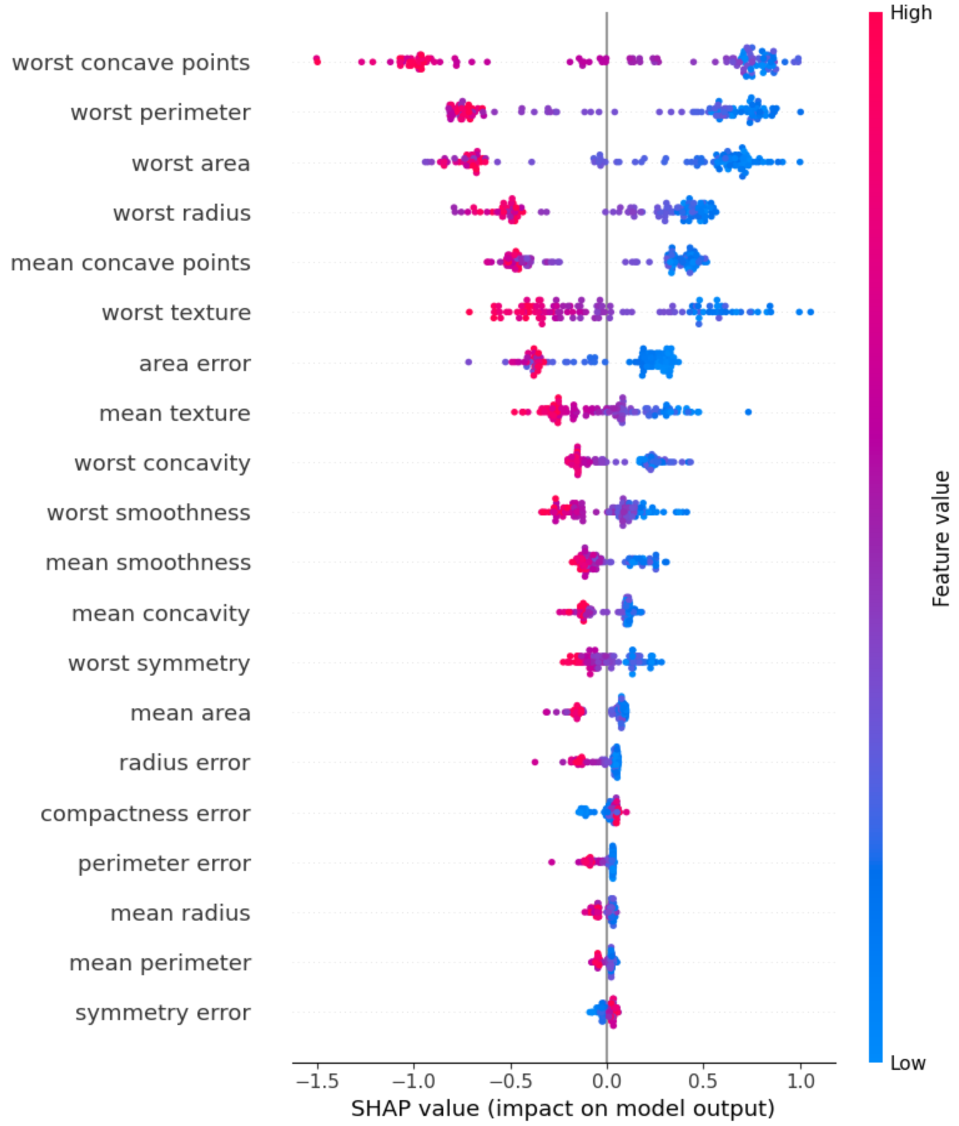

Confusion Matrix
Actual
0
1
Predicted
0
1
39
3
2
70
70
60
50
40
30
20
10

5. ROC Curve


ROC Curve
True Positive Rate
False Positive Rate
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
AUC = 0.995

6. SHAP Explainability

The SHAP summary plot highlights the most important features influencing the model. High-impact features such as 'worst concave points', 'worst perimeter', and 'worst radius' play major roles in classifying malignant vs benign tumors.



7. Conclusion

This ML pipeline demonstrates a complete classification workflow with hyperparameter tuning and explainability. The final XGBoost model achieves high accuracy and generalization capability, supported by visual diagnostics such as ROC and SHAP. This satisfies the requirements of Task 1 for technical accuracy, structured problem solving, and presentation quality.