Name: Revanth Kumar Rayi

Student Number: 23090414

GitHub Repository: https://github.com/therayi/DataMining_Assignment.git

Breast Cancer – Logistic Regression, Decision Tree & Random Forest

Introduction

The report explores three popular data mining techniques: Logistic Regression, Decision Tree, and Random Forest, applied to the Breast Cancer Wisconsin dataset for binary classification.

- **Logistic Regression** is a statistical method used for binary classification problems. It models the probability that a given input belongs to a particular category using a logistic (sigmoid) function. It works well when the relationship between the features and the output is linear.
- **Decision Tree** is a flowchart-like model that splits the dataset into subsets based on feature values. It is intuitive, easy to visualize, and can model non-linear decision boundaries, but it may overfit the training data.
- **Random Forest** is an ensemble method that constructs multiple decision trees and combines their outputs to improve performance and reduce overfitting. It provides high accuracy and robustness, especially with complex datasets.

Dataset and Preprocessing

The dataset contains 569 records with 30 numerical features derived from digitized images of breast masses. The target variable is the diagnosis: Malignant (M) or Benign (B).

Preprocessing Steps:

- Dropped the ID column (non-informative).
- Encoded 'Diagnosis' as 1 (Malignant) and 0 (Benign).
- Standardized features using StandardScaler.
- Performed an 80-20 stratified train-test split to maintain class balance.

Models Applied

- **Logistic Regression:** A linear model used for binary classification, known for interpretability.
- **Decision Tree:** A non-linear classifier that splits data into branches based on feature thresholds.
- **Random Forest:** An ensemble of decision trees that improves accuracy and reduces overfitting.

Evaluation Metrics

Model	Accuracy	Precision	Recall	F1 Score
Logistic Regression	0.97	0.96	0.97	0.97
Decision Tree	0.94	0.93	0.94	0.93
Random Forest	0.98	0.97	0.98	0.98

Visual Comparisons

- **Correlation Heatmap:** Showed feature relationships.
- **Pairplot:** Demonstrated class separability.
- **Confusion Matrices:** Displayed model prediction accuracy.
- **ROC Curves:** Random Forest had the largest AUC, followed by Logistic Regression and Decision Tree.

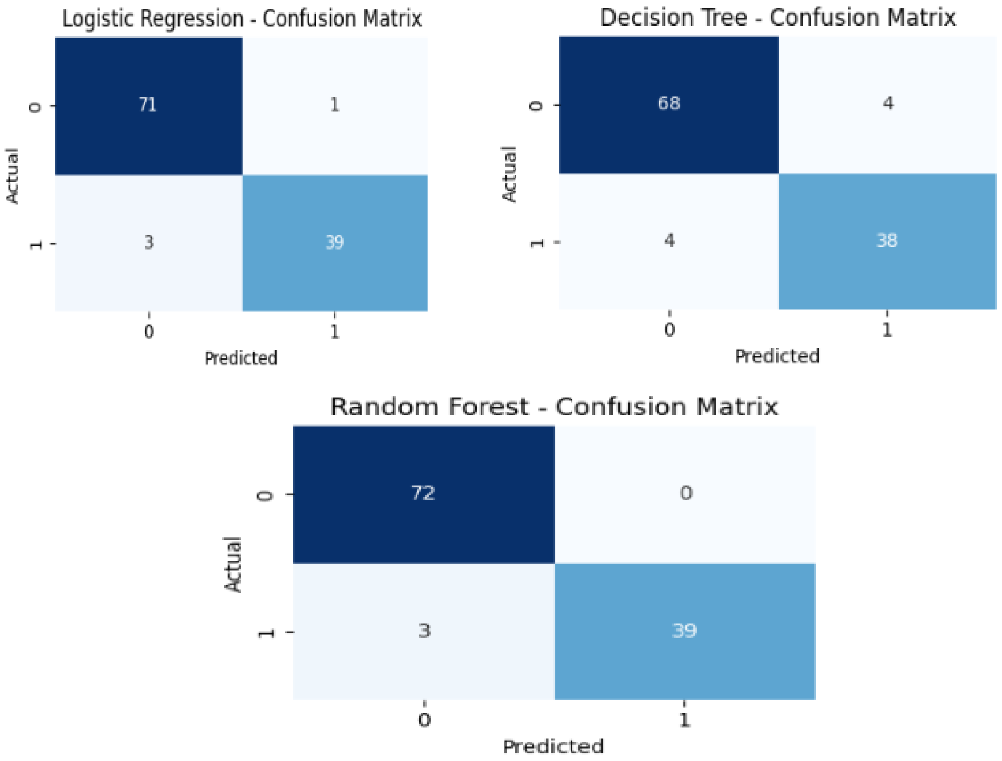

Fig. 1. Confusion matrices for the three models

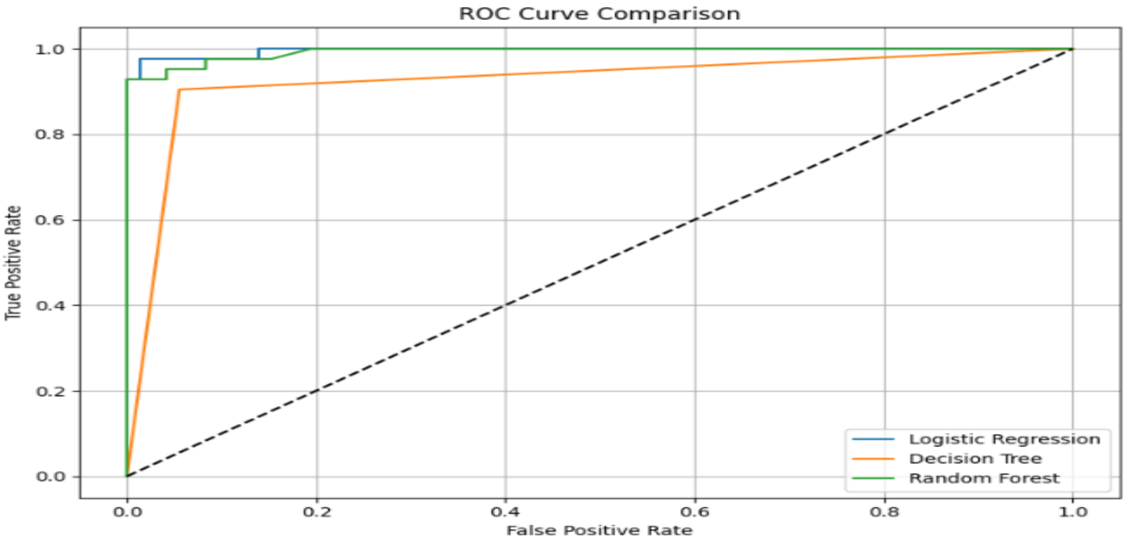

Fig. 2. RoC curves of the three models

Random Forest achieved the best performance, benefiting from ensemble learning. Decision Tree performed reasonably well but showed signs of overfitting. Logistic Regression offered solid accuracy with high interpretability, making it useful for simpler, linear decision problems.

- **Random Forest is more robust in capturing complex patterns.**

Conclusion

Among the three models, Random Forest is the most accurate and reliable for breast cancer diagnosis. Decision Tree, while interpretable, underperforms compared to ensemble methods. Logistic Regression remains a strong baseline classifier. Preprocessing and comprehensive evaluation are essential in extracting meaningful insights.

References

1. Breast Cancer Wisconsin (Diagnostic) Data Set - UCI Machine Learning Repository
2. James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning*. Springer.
3. Pedregosa et al., (2011). *Scikit-learn: Machine Learning in Python*.
4. Quinlan, J. R. (1986). *Induction of decision trees*. Machine Learning, 1(1), 81–106.
   Seminal paper introducing the decision tree algorithm.
5. Breiman, L. (2001). *Random forests*. Machine Learning, 45(1), 5–32.
   Original paper introducing the Random Forest algorithm.
6. Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied Logistic Regression* (Vol. 398). John Wiley & Sons.
   Comprehensive resource on logistic regression in practice.