A Comparative Analysis of Support Vector Machines and Random Forests on the Breast Cancer Dataset

Introduction

This study examines the performances of Support Vector Machines and Random Forests in classifying a tumour as malignant or benign on the Breast Cancer dataset. It compares two modern and popular machine learning models: Support Vector Machines and Random Forests. The support vector machine performs a binary classification problem that finds the best hyperplane, whereas RF uses a set of decision trees for effective prediction. The following sections present data preprocessing, model implementations, and key performance metrics.

Data Preprocessing

- **Removing Invalid Data**: No invalid rows were found in the Breast Cancer dataset.
- **Feature Scaling**: Normalized features to [0, 1] for SVM to ensure improved performance with distance-based calculations.
- **Feature Selection**: All 10 numerical features, such as Clump Thickness and Bare Nuclei, were retained for both models.
- **Target Variable**: Transformed class labels into binary form (0 for benign, 1 for malignant).
- **Dataset Split**: 80% of data was used for training and 20% for testing to ensure robust evaluation.

Comparison Table:

Aspect	**SVM (RBF Kernel)**	**Random Forest**
Objective	Classify tumours as malignant or benign	Classify tumours as malignant or benign
Features Used	All 10 numerical features	All 10 numerical features
Performance Metrics	Accuracy: 98%, ROC-AUC: 99.7%	Accuracy: 96%, ROC-AUC: 99.5%
Strengths	Handles high-dimensional data effectively	Provides feature importance insights
Weaknesses	Lacks interpretability	Computationally expensive

Model Implementations and Results

Subtopic 1: Support Vector Machines (SVMs)

Description: SVMs find the optimal hyperplane for classification by maximizing the margin between data points of different classes. They use kernel functions, such as radial basis function (RBF), for non-linear decision boundaries.

Objective: To classify tumours as malignant or benign based on 10 features.

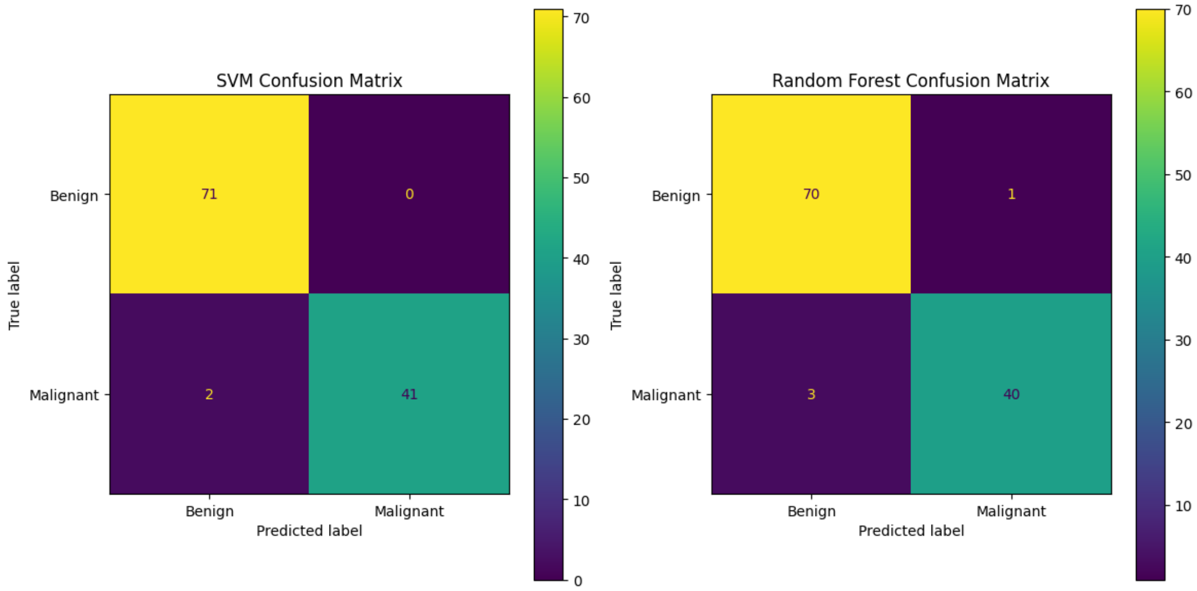
Metrics:

Accuracy: 98%, ROC-AUC: 99.7%, Classification Report: Precision (0.97), Recall (1.00), F1-Score (0.98)

Key Observations:

- The model achieved high accuracy with minimal misclassifications.
- **Figure 1** illustrates the confusion matrix, showing 71 true benign predictions, 41 true malignant predictions, and only 2 false negatives.

Figure 1: Confusion Matrix for SVM and Random Forest



Subtopic 2: Random Forests (RF)

Description: RF is an ensemble learning method that combines multiple decision trees to reduce overfitting and improve accuracy. Each tree is trained on a bootstrap sample, and predictions are aggregated using majority voting.

Objective: To classify tumours as malignant or benign based on 10 features.

Metrics:

Accuracy: 96%, ROC-AUC: 99.5%, Classification Report: Precision (0.97), Recall (0.95), F1-Score (0.96)

Key Observations:

- The model showed slight misclassifications with 3 false negatives and 1 false positive.
- **Figure 2** represents the confusion matrix, highlighting most samples classified correctly.
- Feature importance analysis revealed "Feature 24" (Importance: 15.4%) and "Feature 28" (Importance: 14.5%) as significant predictors.

Conclusion

The models developed using SVM and Random Forest performed well on the Breast Cancer dataset, yielding good accuracy and reliability. The SVM marginally outperformed RF regarding ROC-AUC and classification accuracy, thus making it more suitable for precision-focused applications like medical diagnostics. However, the importance of the features given by Random Forest makes it a perfect choice for performing exploratory analysis and feature engineering.

Code: https://github.com/sravanth-space/dm_23001152

DataSet: https://archive.ics.uci.edu/dataset/15/breast+cancer+wisconsin+original

Student Register Number: 23001152

Name: Sravanth Baratam

References

1. Tan, Pang-Ning, et al. *Introduction to Data Mining.* Pearson Education, 2019.
2. UCI Machine Learning Repository: Breast Cancer Dataset. https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(original).
3. Scikit-learn Documentation. https://scikit-learn.org.