Evaluation of Naive Bayes, Decision Tree, and Random Forest Classifiers on Breast Cancer Diagnosis Data

Name: VISHAL KUMAR SENTHIL KUMAR

STUDENT ID: 23076841

GitHub: https://github.com/vishalkumar041298/report

Introduction

This report presents a comparative analysis of three supervised classification algorithms: Naive Bayes, Decision Tree, and Random Forest, applied to the Breast Cancer Wisconsin dataset. The primary objective is to assess and contrast their performance in predicting tumor malignancy. Accurate classification in medical contexts is crucial, as misdiagnosis can have severe consequences.

Dataset and Preprocessing

The dataset used in this analysis is the Breast Cancer Wisconsin (Original) dataset from the UCI Machine Learning Repository. It includes 699 samples, each described by 10 numerical features related to cell characteristics and a class label indicating whether a tumor is benign or malignant. Preprocessing steps were as follows:

- Replaced missing values in the "Bare Nuclei" column with NaN and removed affected rows.
- Dropped the non-predictive "Sample Code Number" column.
- Encoded target values: benign = 0, malignant = 1.
- Divided the dataset into 80% training and 20% testing sets.

Algorithms Used

- **Naive Bayes:** A probabilistic classifier assuming feature independence and normally distributed features. It is known for its simplicity and speed.
- **Decision Tree:** Builds a tree-like structure of decision rules based on feature values, capable of capturing non-linear relationships.
- **Random Forest:** An ensemble method combining multiple decision trees with randomized subsets of data and features, aiming to improve generalization and reduce overfitting.

[117]:

Performance Comparison of Classifiers

	Accuracy	Precision	Recall	F1 Score
Naive Bayes	0.9562	0.9483	0.9483	0.9483
Decision Tree	0.9343	0.9623	0.8793	0.9189
Random Forest	0.9489	0.9811	0.8966	0.9369

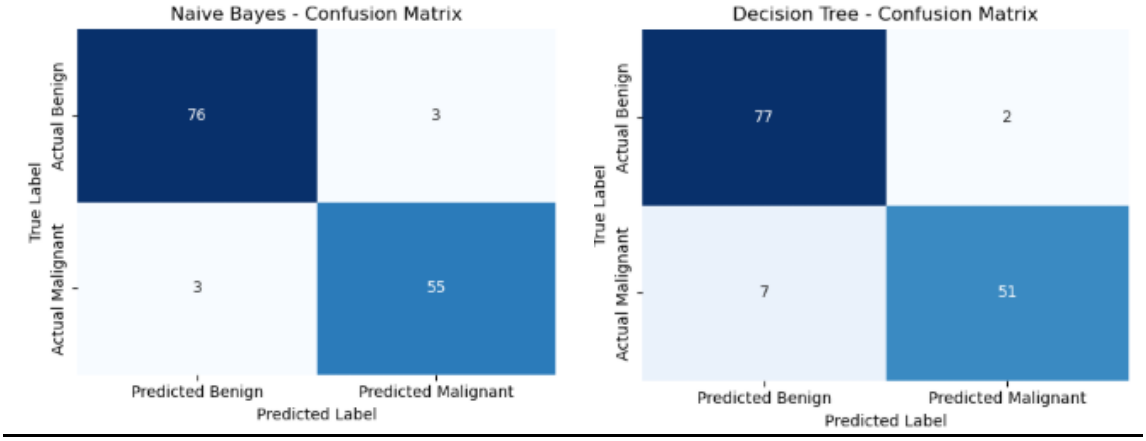
Interpretation & Comparison:

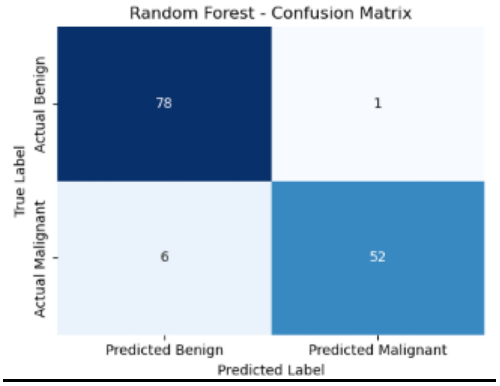
Naive Bayes achieved the highest overall accuracy, which is notable given its strong assumptions. However, its performance is more consistent when features are relatively independent. Decision Tree delivered the highest precision but had a lower recall, indicating a tendency to miss some malignant cases.

Evaluation of Naive Bayes, Decision Tree, and Random Forest Classifiers on Breast Cancer Diagnosis Data

Random Forest offered the best F1 score, reflecting a strong balance between precision and recall. Its ensemble nature helps in reducing overfitting and improving generalization, especially evident in medical classification tasks where both false positives and false negatives have high stakes.

Confusion Matrix





Confusion Matrices for Classifier Predictions Each confusion matrix illustrates the classifier's predictions:

- **True Positives (TP):** Malignant tumors correctly classified as malignant.
- **True Negatives (TN):** Benign tumors correctly classified as benign.
- **False Positives (FP):** Benign tumors incorrectly classified as malignant.
- **False Negatives (FN):** Malignant tumors incorrectly classified as benign (most critical error in medical settings).

Conclusion

While all models demonstrated good performance, Random Forest emerged as the most balanced and reliable classifier for this dataset. Naive Bayes is efficient and accurate for smaller, simpler datasets. Decision Trees provide interpretability but can be prone to overfitting. Random Forests combine strength and robustness, making them well-suited for healthcare applications where predictive accuracy is critical.

References

Evaluation of Naive Bayes, Decision Tree, and Random Forest Classifiers on Breast Cancer Diagnosis Data

- Dua, D. and Graff, C. (2019). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.
- Tan, P.-N., Steinbach, M., & Kumar, V. (2018). Introduction to Data Mining (2nd ed.). Pearson.
- scikit-learn documentation: https://scikit-learn.org/stable/modules/classes.html