COMPARATIVE ANALYSIS OF NAÏVE BAYES CLASSIFIERS ON UCI HAR DATA

NAME : PAVAN KUMAR KURVA
STUDENT ID : 22065713

GOOGLE COLAB LINK:
https://colab.research.google.com/drive/1whrDNx6a2ya2HL0JIngWk1Eb2L7gFYMl?usp=sharing

ABSTRACT:

This report presents a comparative analysis of three prominent machine learning classifiers—Naive Bayes, Decision Tree, and Random Forest—applied to the UCI Human Activity Recognition (HAR) dataset. The objective is to evaluate their performance in recognizing various human activities based on smartphone sensor data. Through techniques like cross-validation, hyperparameter tuning, and performance metrics, we assess the classifiers' effectiveness in accurately classifying activities from accelerometer and gyroscope measurements.

INTRODUCTION:

The UCI HAR dataset is a widely used benchmark dataset in activity recognition research, containing data collected from smartphone sensors worn by subjects performing six different activities: walking, walking upstairs, walking downstairs, sitting, standing, and laying. Each instance in the dataset comprises tri-axial accelerometer and gyroscope readings, along with corresponding activity labels.

METHODOLOGY:

1. **Data Preparation:** We load the UCI HAR dataset and partition it into training and testing sets, preserving the integrity of activity distributions.

2. **Classifier Training and Evaluation:**
   - ***Naive Bayes Classifier:*** A Gaussian Naive Bayes model is trained and evaluated using k-fold cross-validation (k=5).
   - ***Decision Tree Classifier:*** Employing a Decision Tree classifier, we conduct hyperparameter tuning via grid search with cross-validation to optimize the tree's depth.
   - ***Random Forest Classifier:*** Like the Decision Tree approach, a Random Forest classifier is utilized with hyperparameter tuning to determine the optimal number of estimators and maximum depth.

3. **Performance Evaluation:**
   - We compute mean accuracy scores for each classifier based on cross-validation results.
   - Confusion matrices are generated to visualize the classifiers' performance in predicting activity labels.
   - Classification reports are produced to provide comprehensive metrics including precision, recall, and F1-score for each activity class.

Naive Bayes Classifier Accuracy on Test Set: 0.7702748557855447
Classification Report for Naive Bayes Classifier:

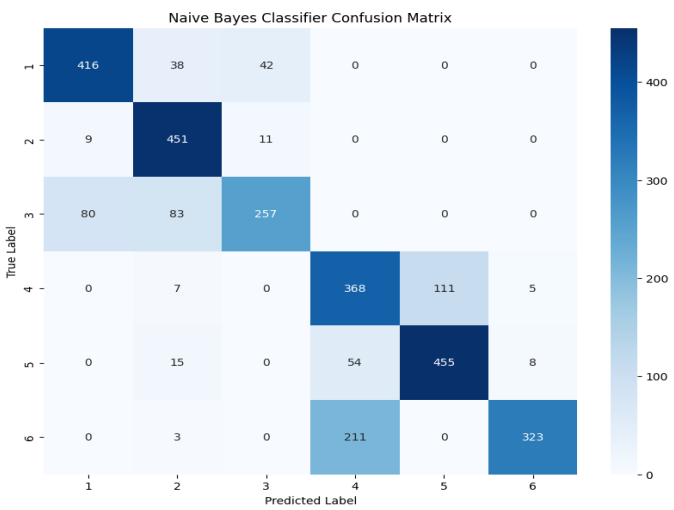
	precision	recall	f1-score	support
1	0.82	0.84	0.83	496
2	0.76	0.96	0.84	471
3	0.83	0.61	0.70	420
4	0.58	0.75	0.65	491
5	0.80	0.86	0.83	532
6	0.96	0.60	0.74	537
accuracy			0.77	2947
macro avg	0.79	0.77	0.77	2947
weighted avg	0.79	0.77	0.77	2947

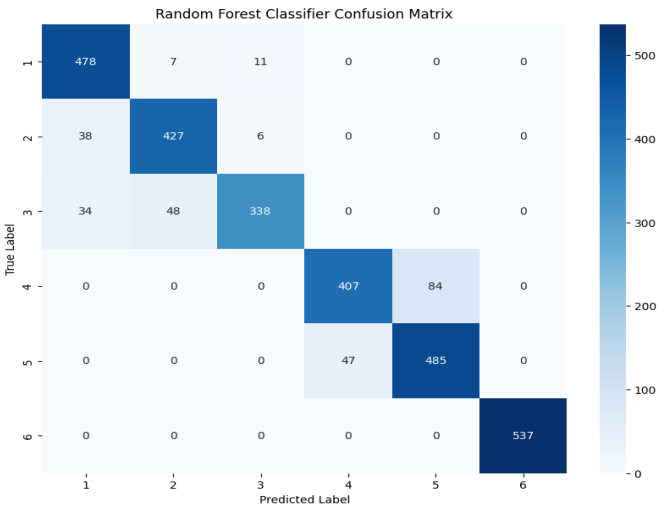
Random Forest Classifier Accuracy on Test Set: 0.9066847641669494
Classification Report for Random Forest Classifier:

	precision	recall	f1-score	support
1	0.87	0.96	0.91	496
2	0.89	0.91	0.90	471
3	0.95	0.80	0.87	420
4	0.90	0.83	0.86	491
5	0.85	0.91	0.88	532
6	1.00	1.00	1.00	537
accuracy			0.91	2947
macro avg	0.91	0.90	0.90	2947
weighted avg	0.91	0.91	0.91	2947

RESULTS:

- Naive Bayes Classifier Mean Accuracy: 68.62%
- Decision Tree Classifier Mean Accuracy: 85.13%
- Random Forest Classifier Mean Accuracy: 91.47%





DISCUSSION:

- The Naive Bayes classifier achieves a mean accuracy of 68.62%, indicating moderate performance in activity recognition based on smartphone sensor data. Despite its simplicity and assumption of feature independence, it falls short in capturing the complex relationships present in the dataset, resulting in lower accuracy compared to Decision Tree and Random Forest classifiers.

- The Decision Tree classifier performs significantly better with a mean accuracy of 85.13% after hyperparameter tuning. Decision Trees offer interpretability and demonstrate an improved ability to capture patterns in the data compared to Naive Bayes. However, there is still room for improvement, as it may suffer from overfitting if not properly pruned.
- The Random Forest classifier outperforms both Naive Bayes and Decision Tree classifiers with the highest mean accuracy of 91.47%. By aggregating multiple decision trees, Random Forests mitigate overfitting and enhance robustness, resulting in superior performance in classifying human activities based on smartphone sensor data.

CONCLUSION:

In this study, we evaluated the performance of Naive Bayes, Decision Tree, and Random Forest classifiers on the UCI HAR dataset for human activity recognition. While Naive Bayes demonstrates moderate accuracy, Decision Tree and Random Forest classifiers exhibit significantly better performance. Random Forest emerges as the most effective classifier, showcasing its superiority in accurately classifying human activities based on smartphone sensor data. The choice of classifier should consider both accuracy and computational efficiency, with Random Forest being a promising option for real-world applications requiring high accuracy.

REFERENCES:

- Zhang, Harry. *The Optimality of Naive Bayes* (PDF). FLAIRS2004 conference.
- Metsis, Vangelis; Androutsopoulos, Ion; Paliouras, Georgios (2006). *Spam filtering with Naive Bayes—which Naive Bayes?*. Third conference on email and anti-spam (CEAS).
- *McCallum, Andrew. "Graphical Models, Lecture2: Bayesian Network Representation" (PDF). Archived (PDF) from the original on 2022-10-09. Retrieved 22 October 2019.*

FURTHER READING:

- *Webb, G. I.; Boughton, J.; Wang, Z. (2005). "Not So Naive Bayes: Aggregating One-Dependence Estimators". Machine Learning. **58** (1): 5–24. doi:10.1007/s10994-005-4258-6.*
- *Mozina, M.; Demsar, J.; Kattan, M.; Zupan, B. (2004). Nomograms for Visualization of Naive Bayesian Classifier (PDF). Proc. PKDD-2004. pp. 337–348.*

EXTERNAL LINKS : Book Chapter: Naive Bayes text classification, Introduction to Information Retrieval