**HUMAN ACTIVITY RECOGNITION USING DATA FROM SMARTPHONE SENSORS**

Using Decision Tree, Random Forest, and Artificial Neural Network (ANN)

**1. INTRODUCTION**

Human Activity Recognition (HAR) is a vital research field with broad applications, particularly in healthcare and human survey systems[1]

This study compares Decision Tree, Random Forest, and Artificial Neural Network (ANN) models for activity recognition such as walking, sitting, standing, and others using smartphone sensor data. Preprocessing techniques include feature scaling, outlier detection, and dimensionality reduction. Evaluation metrics include accuracy, precision, recall, F1 score, and ROC AUC. The study aims to identify the most effective model for accurately recognizing human activities from smartphone sensor data.

**2.EDA(Exploratory Data Analysis)**

EDA involves involves examining and visualizing the dataset to gain insights into its structure, distribution, and relationships between variables. Functions like describe() summarize key statistics. Visualizations include:

**a) Percentage of Each Activity in Test Data Pie Chart**:

Illustrates activity class distribution in the test dataset with "laying" comprising the largest portion, accounting for 18.2% of the total.

A pie chart with different colored circles

Description automatically generated

**c) Distribution Plot of Random Feature by Activity:**

This FacetGrid visualization displays the distribution of a randomly selected feature across different activities.

A graph of activities and activities

Description automatically generated

These visualizations provide insights into the distribution of activities, feature distributions across activities, and the overall composition of the dataset, which are essential components of Exploratory Data Analysis.

**3.DATA PREPROCESSING**

**a) Duplicates Detection**:

Identified and printed the number of duplicate entries in both the training and test datasets.

**b) Null Values Handling**:

Counted and printed the number of NaN/Null values in both the training and test datasets[4].

**c) Outlier Detection and Removal**:

Utilized Isolation Forest algorithm to detect and remove outliers from the training dataset.

**d) Feature Scaling**:

Applied StandardScaler to standardize the features in both the training and test datasets.

**e) Feature Selection**:

Employed SelectKBest to select the top k features based on their relevance to the target variable.

**f) Dimensionality Reduction (PCA)**:

Used Principal Component Analysis (PCA) to reduce the dimensionality of the data while preserving its variance.

**g) Handling Categorical Variables**:

Encoded the categorical target variable 'Activity' using LabelEncoder to convert it into numerical values.

**h) Choosing the Number of Components**:

Determined the optimal number of components for PCA by plotting the cumulative explained variance ratio.

**i) Visualization (PCA and t-SNE)**:

Visualized the data using the first two principal components obtained from PCA.

Applied t-Distributed Stochastic Neighbor Embedding (t-SNE) for high-dimensional data visualization.

A graph with colorful dots

Description automatically generated

A diagram of different colored dots

Description automatically generated

**j) Saving Preprocessed Data**:

Saved the preprocessed datasets as CSV files.

**k) Splitting Data into Features and Labels**:

Segregated the training dataset into features (X\_train) and labels (y\_train), Similarly, test dataset into features (X\_test) and labels (y\_test)

**4. MODEL BUILDING AND EVALUATION**

Each model is initialized with default parameters and then optimized using hyperparameter tuning via GridSearchCV.

The models are trained on the training data and evaluated on the test data.

**1.Decision Tree Classifier:**

Decision trees[2] are a non-parametric supervised learning method used for classification tasks For Decision Tree, the parameter chosen was 'max\_depth'. Achieved a high accuracy of 99.97%. The F1 score, precision, and recall were also excellent, indicating robust performance.

**2. Random Forest Classifier:**

Random forests[3] are an ensemble learning method that constructs multiple decision trees during training and outputs the class that is the mode of the classes of the individual trees.

For Random Forest, parameters were 'n\_estimators' and 'max\_depth'. Achieved an accuracy of 97.05%. While slightly lower than Decision Trees, still delivered strong F1 score, precision, and recall.

**3. Artificial Neural Network (ANN):**

ANNs[3] are a class of machine learning models inspired by the structure and functioning of the human brain. For ANN, parameters included 'hidden\_layer\_sizes', 'activation', 'solver', and 'alpha'Achieved an accuracy of 99.29%, slightly below Decision Trees. However, boasted high F1 score, precision, and recall, showcasing its efficacy.

Confusion matrices are generated and visualized to understand the model's performance in classifying activities.The performance of each model is compared and contrasted based on these evaluation metrics.

**5.RESULTS AND COMPARISON**

The multi-panel bar plot compares the performance metrics (Accuracy, Error, F1 Score, Precision, Recall, ROC AUC) of Decision Tree, Random Forest, and Artificial Neural Network models. It provides a concise overview of each model's effectiveness in classifying activities based on smartphone sensor data..

A group of colorful bars

Description automatically generated

|  |  |  |  |
| --- | --- | --- | --- |
| Model | Accuracy | Error | F1 Score |
| Decision tree | 99.9661% | 0.0339% | 0.9997 |
| Random forest | 97.0478% | 2.9522% | 0.9703 |
| Artificial neural network | 99.2874% | 0.7126% | 0.9929 |

**6. Conclusion**:

* 1. Overall, all three models performed well in classifying activities based on the provided dataset.
  2. The Decision Tree model demonstrated the highest accuracy, with minimal error and excellent performance across various evaluation metrics.
  3. The Random Forest model provided a good balance between accuracy and computational complexity, although it had slightly lower performance compared to the Decision Tree.
  4. The ANN model achieved competitive accuracy and outperformed the other models in terms of ROC AUC score, suggesting superior discrimination between classes.
  5. Considering the complexity and computational resources required, the Decision Tree model may be preferred for its simplicity and high accuracy, unless the task demands superior discrimination between classes, in which case the ANN model would be more suitable.

**REFERENCES**

[1] Vyas, V. V., Walse, K., Dharaskar, R. V. (March 2017). "A Survey on Human Activity Recognition using Smartphone." Anuradha Engineering College. Institute of National Importance

[2] Tan, P.-N., Steinbach, M., Karpatne, A., Kumar, V. (Second Edition). “Introduction to Data

[3] Rabbi, J., Fuad, M. T. H., Awal, M. A. (2021). "Human Activity Analysis and Recognition from Smartphones using Machine Learning Techniques." Submitted on 30 Mar 2021.

[4] Zhang, S., Li, Y., Zhang, S., Shahabi, F., Xia, S., Deng, Y., Alshurafa, N. (Year). "Deep Learning in Human Activity Recognition with Wearable Sensors: A Review on Advances."