Physical Activity Prediction

Prerequisite :

Create a python virtual environment, activate the environment and execute the following command to install the dependencies:

```
pip install -r requirements.txt
```

Create a folder in the root directory called `data` to store the dataset that is needed for the training.

The following notebooks are the primary files to be used for the end-to-end Model building pipeline for this physical activity prediction dataset.

- `Data_Collection.ipynb`
- `Data_EDA_Cleanup.ipynb`
- `Model_Building_Evaluation_Combined.ipynb`

Other files prefixed with `Train_` are the draft notebook files that we used to experiment the individual algorithms on the dataset.

Data collection :

The data that will be used is to be stored in the folder 'data' from the dataset

These files should be dowloaded from the above link and stored in the `data` folder - `subject101.dat, subject102.dat, subject103.dat, subject104.dat, subject105.dat, subject106.dat, subject107.dat, subject108.dat, subject109.dat`

Please refer to the readme.pdf and other docs in the 'data' folder for the data collection related information that was provided with the dataset.

Using this informatation, The raw data from various subjects are collected by running the `Data_Collection.ipynb` python notebook. This collects the data from the various subject files and stores into a single .csv file (`compiled_raw_data.csv`) in the same folder for further analysis.

Exploratory Data Analysis :

We use the `compiled_raw_data.csv` data file to perform exploratory data analysis using `Data_EDA_Cleanup.ipynb` python notebook.

Here, We perform EDA and collect useful information from the data, and perform feature extraction using PCA (Principle Component Analysis) and other techniques. We remove the unnecessary columns and rows based on the exploratory analysis and store the final data as `final_data.csv` in the same folder

This final data is normalized (standard) and the target data is label encoded.

Activity IDs are encoded and the mappings are :

```
{0:'lying',1:'sitting',2:'standing',3:'walking',4:'running',5:'cycling',6:'nordic_walking',7:'ascending_stairs',8:'descending_stairs',9:'vacuum_cleaning',10:'ironing',11:'rope_jumping'}
```

Training and Evaluation:

We use the `final_data.csv` from the folder `data` , which has the final set of features and cleaned data to perform training on various machine learning algorithms.

Run the `Model_Building_Evaluation_Combined.ipynb` python notebook to train and evaluate the final set of models that we explored with the dataset.

The machine learning algorithms that we train the data on includes :

- Logistic Regression
- Decision Trees
- Random Forests
- Support Vector Machines (SVM)
- Deep Neural Networks

Results

We collected the model performance from each of the algorithms and picked the best model to be used with it's best hyperparameters. Below are the performances for each models :

Models	Testing Accuracies with best hyperparameters
Logistic Regression	82.142 %
Support Vector Machines	99.53 %
Decision Trees	86.29 %
Random Forest	99.986 %
XGBoost	99.984 %
Neural Network	99.31 %

We can finally train the whole dataset (Train + Test) on this best model to get the final model to be deployed (`deployed_model`).