Human Activity Recognition Using Smartphone's Sensor Data
Phase - II Report

Instructor: Dr. Tanmay Basu

Name:	**Amartyaraj Kumar**
Registration No./Roll No.:	19029
Institute/University Name:	IISER Bhopal
Program/Stream:	EECS
Problem Release date:	February 02, 2022
Date of Submission:	April 24, 2022

Introduction

Human Activity Recognition (HAR) has become one of the most sought-after processes to run different human surveys, health care & fitness applications and smart home appliances. As most of the smartphones in today's world are built-in with different sensors like accelerometer, gyroscope, GPS and etc. we can set up a model with appropriate machine learning techniques using these sensor data to predict a human activity.

Related Work

Attal et al. tried to classify activity depending on Four supervised classification techniques namely, K-NN, SVM, Gaussian Mixture Models, and Random Forest (RF) as well as unsupervised classification techniques like k-Means and Hidden Markov Model (HMM) using from wearable inertial sensor data [1].Bayat et al. studied on Recognition of human physical activities using acceleration data generated by a user's cell phone [2] . Ronao et al. structured a convolutional artificial neural network in order to recognize user activity using smartphones accelerometer and gyroscope [3]. To meantion at last, Bulbul et al. have used several supervised learning classifiers to get the predictive HAR model with the best accuracy [4].

Dataset description

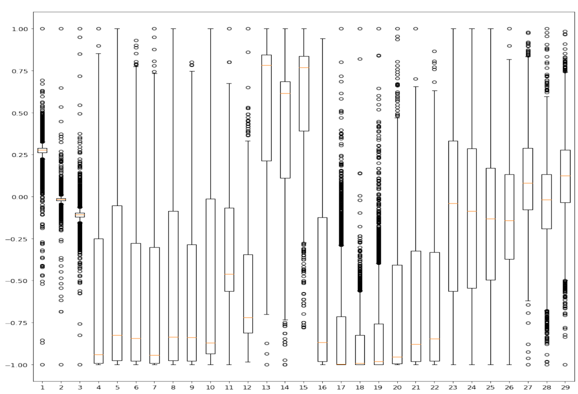

Figure 1: Value plot of 30 feature vectors

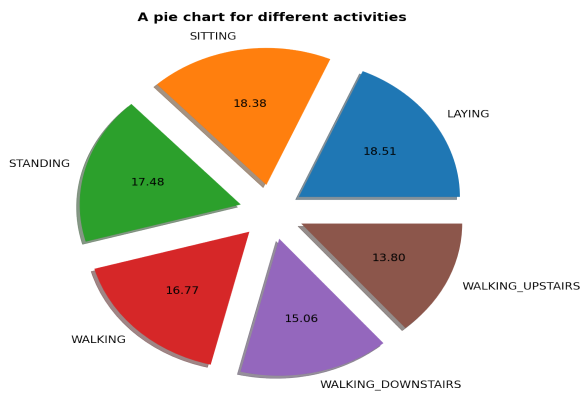

Figure 2: Pie plot of different activity

We have here 30 volunteers within an age bracket of 19-48 years performing six locomotive motions (`WALKING, WALKING_UPSTAIRS, WALKING_DOWNSTAIRS, SITTING, STANDING, LAYING`), **Figure** 8. In this study, the dataset consists of signals from accelerometer and gyroscope of a smartphone carried by those volunteers while doing those activities. We have a total of 561 feature vectors, one subject identification column and one index number column for a total of 8239 entries in the training data set.

Also, We have the activity labels for this training dataset. The given data is scaled between [-1,1], **Figure 1**, and also it is noise-filtered as the data is performed under Butterworth filter using median and 20 Hz frequency.

Methods

Here, given the dataset, mainly four supervised learning classifiers are considered for training and building the model. Those are k-NN Classification, Support Vector Machine (SVM), Logistic Regression, and Adaboost Ensemble Classifier.

At first, we preprocessed our training data by removing 'subject' and 'index' column. Then we used `StandardScaler` to scale the data, and then proceeded with `train_test_split`.We kept a training size of 75%.

KNN Classifier

We started with k-NN classifier using our preprocessed & split training data which contains all the 561 feature vectors. With the speedrun and no parameter tuning we got an accuracy of **95.82%**.

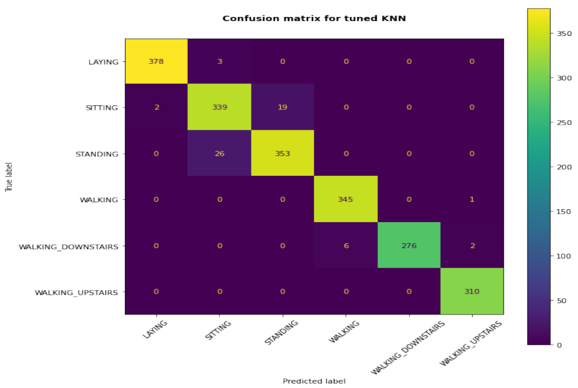

Figure 3: Confusion Matrix

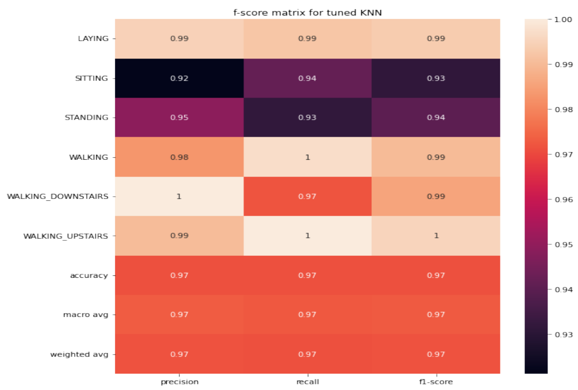

Figure 4: f-score and Accuracy

Then We opted for hyperparameter tuning with 10 cross validation of the classifier parameters and found the best k-NN fit with accuracy = **97.14%**.

SVM Classifier

With Support Vector Machine classifier, firstly we had our speedrun with all the scaled feature vectors and no parameter tuning. and as a result it produced an accuracy of **97.57%**.

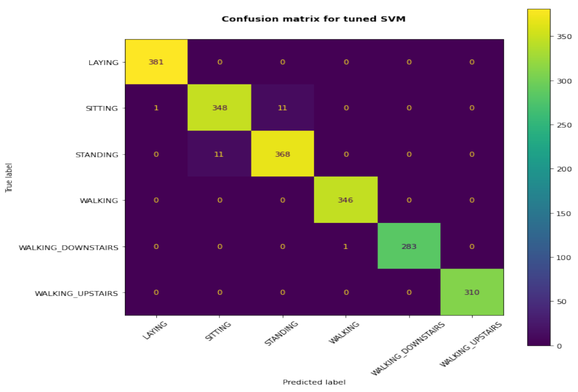

Figure 5: Confusion Matrix

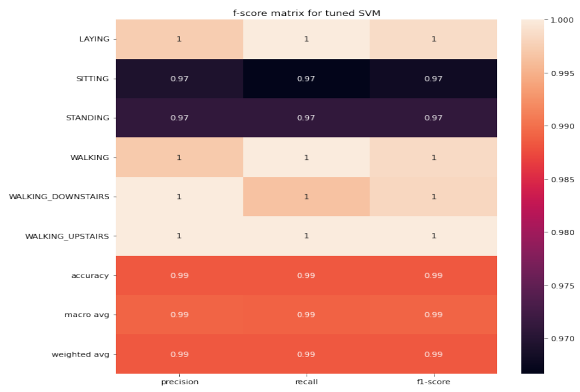

Figure 6: f-score and Accuracy

Then we opted for hyperparameter tuning with 10 cross validation of the classifier parameters including `'linear'`, `'rbf'` & `'poly'` kernels, and found the best SVM fit with accuracy = **98.83%**.

Logistic Regression

Using our preprocessed & split training data which contains all the 561 feature vectors, we first had our no-tuning speedrun accuracy for Logistic Regression = **98.40%**.

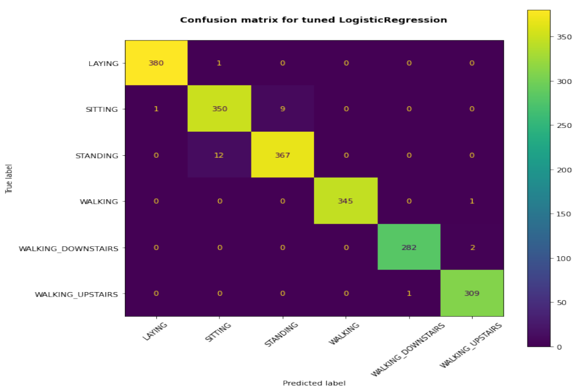

Figure 7: Confusion Matrix

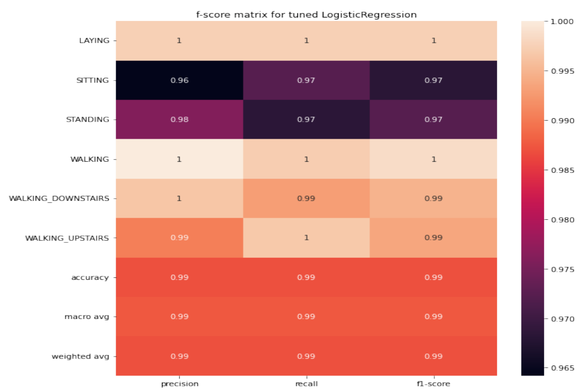

Figure 8: f-score and Accuracy

Then We opted for hyperparameter tuning of the classifier parameters with 10 cross validation and found the best Logistic Regression fit with accuracy = **98.69%**. We can clearly see that despite having a better accuracy in no-tuning condition, Logistic Regression classifier under-performs with respect to SVM after parameter tuning.

Adaboost Ensemble

Using our preprocessed & split training data which contains all the 561 feature vectors, we first had our no-tuning speedrun accuracy with the Adaboost Ensemble Classifier = **53.69%**. So, We opted to drop this classifier as we were already getting far better results without tuning with other classifiers.

Accuracy Comparison		
Classifier Name	Speedrun accuracy	Accuracy after tuning
K -Nearest Neighbour	95.82%	97.14%
Support Vector Machine	97.57%	98.83%
Logistic Regression	98.40%	98.69%
AdaBoost	53.69%	—

So, our best baseline model after hyperparameter tuning is **SVM**.

GitHub Link

https://github.com/Amartyaraj47/Project_HAR_DSML_ECS308

Experimental Analysis

After deciding the best Baseline model for our dataset, we tried to perform some feature engineering techniques and then run the engineered dataset into our selected classifier to see how the accuracy varies with respect to the baseline model.

We first performed **KBest** feature selection technique and tried to fit the classifiers with in 9 best features of the scaled training dataset. In this feature engineering technique as we did lose 551 other feature vectors our maximum accuracy without tuning happens to be 68.93% by SVM.

Then we performed **Correlation Elimination** feature selection by removing one of two features that were correlated more than 0.9. In this method we could achieve our best accuracy of 97.33% by SVM.

Lastly we opted to perform Principal Component Analysis (PCA) while selecting 10 features with highest variances. In this method we could cover a total variance of 0.7; then we cropped the data as to remove outliers of the last principal component. In this method we could achieve our best accuracy of 87.64% without any parameter tuning.

Accuracy Comparison after Feature engineering			
Classifier Name	KBest Selection	Correlation Elimination	PCA
K -Nearest Neighbour	68.85%	93.45%	87.64%
Support Vector Machine	68.93%	97.33%	87.59%
Logistic Regression	67.64%	97.16%	86.47%

After evaluating all the models before and after parameter tuning and also before and after implementing different feature engineering techniques, We can clearly see that **SVM** (Support Vector Machine) is the best fit for this dataset with the best accuracy after parameter tuning of **98.83%**. So, we will continue to predict our test data using this tuned SVM model only.

Discussions

We have performed some supervised learning techniques on the HAR dataset and got a maximum of 98.83% of accuracy in our best model. The whole dataset is noise free but the main issue is the astronomical number of feature vectors to work with. So, selecting some feature vectors would definitely result in loss of data. Still, Using Correlation elimination we could reach 97.33% accuracy without any parameter tuning. It can be made more accurate following the same feature selection method while tuning the parameters of the classifier.

References

[1] Ferhat Attal, Samer Mohammed, Mariam Dedabrishvili, Faicel Chamroukhi, Latifa Oukhellou, and Yacine Amirat. Physical human activity recognition using wearable sensors. *Sensors*, 15(12):31314–31338, 2015.

[2] Akram Bayat, Marc Pomplun, and Duc A. Tran. A study on human activity recognition using accelerometer data from smartphones. *Procedia Computer Science*, 34:450–457, 2014. The 9th International Conference on Future Networks and Communications (FNC'14)/The 11th International Conference on Mobile Systems and Pervasive Computing (MobiSPC'14)/Affiliated Workshops.

[3] Charissa Ann Ronao and Sung Bae Cho. Human activity recognition with smartphone sensors using deep learning neural networks. *Expert Systems with Applications*, 59:235–244, October 2016. Funding Information: This research was supported by the MSIP (Ministry of Science, ICT and Future Planning), Korea, under the ITRC (Information Technology Research Center) support program (IITP-2016-R0992-15-1011) supervised by the IITP (Institute for Information communications Technology Promotion). Publisher Copyright: © 2016 Published by Elsevier Ltd.

[4] Erhan Bulbul, Aydin Cetin, and Ibrahim Alper Dogru. Human activity recognition using smartphones. In *2018 2nd International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT)*, pages 1–6, 2018.