Recognition of physical activities using smartphone sensors

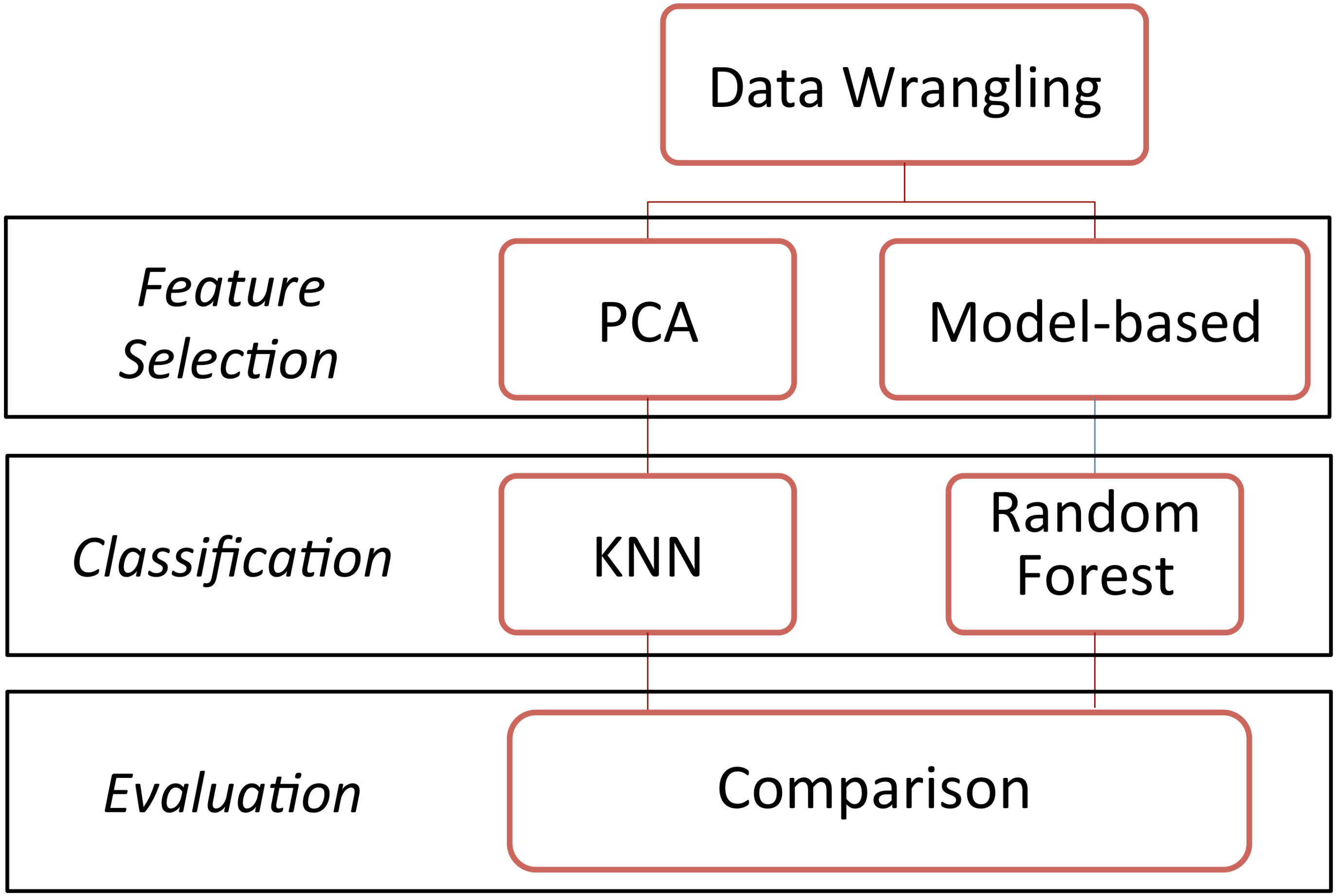
Mehmet Solmaz
Student at Springboard Data Science Intensive Course
July 2017

Problem definition

Anguito, Ghio et al. 2012

- Smartphone accelerometer and gyroscope continuously provide sensory data that vary from activity to activity.
- Can we accurately predict Human Activities and Activity Transitions from wearable sensors?

Methodology



Dataset

- UCI Maching Learning Repository
  - (https://archive.ics.uci.edu/ml/datasets/Smartphone-Based+Recognition+of+Human+Activities+and+Postural+Transitions)
- The original dataset is in multivariate form, is in TXT format and is already split into 70-30% Train-Test split. Feature names, Target Activities and Subject ID (The volunteer who performs the activities) are provided in separate TXT files: *X_test.txt, X_train.txt, y_test.txt, y_train.txt, features.txt, activity_labels.txt, subject_id_test.txt, subject_id_train.txt*
- There are 6 major and 6 transitional activities: '**WALKING**', '**WALKING_UPSTAIRS**', '**WALKING_DOWNSTAIRS**', '**SITTING**', '**STANDING**', '**LAYING**', 'STAND_TO_SIT', 'SIT_TO_STAND', 'SIT_TO_LIE', 'LIE_TO_SIT', 'STAND_TO_LIE', 'LIE_TO_STAND'
- This dataset was obtained from 30 individuals or subjects wearing their smartphones that record the sensor data.

Dataset

Total of **561** features

Name	Time	Freq.
Body Acc	1	1
Gravity Acc	1	0
Body Acc Jerk	1	1
Body Angular Speed	1	1
Body Angular Acc	1	0
Body Acc Magnitude	1	1
Gravity Acc Mag	1	0
Body Acc Jerk Mag	1	1
Body Angular Speed Mag	1	1
Body Angular Acc Mag	1	1

(a)

Function	Description
mean	Mean value
std	Standard deviation
mad	Median absolute value
max	Largest values in array
min	Smallest value in array
sma	Signal magnitude area
energy	Average sum of the squares
iqr	Interquartile range
entropy	Signal Entropy
arCoeff	Autorregresion coefficients
correlation	Correlation coefficient
maxFreqInd	Largest frequency component
meanFreq	Frequency signal weighted average
skewness	Frequency signal Skewness
kurtosis	Frequency signal Kurtosis
energyBand	Energy of a frequency interval
angle	Angle between two vectors

(b)

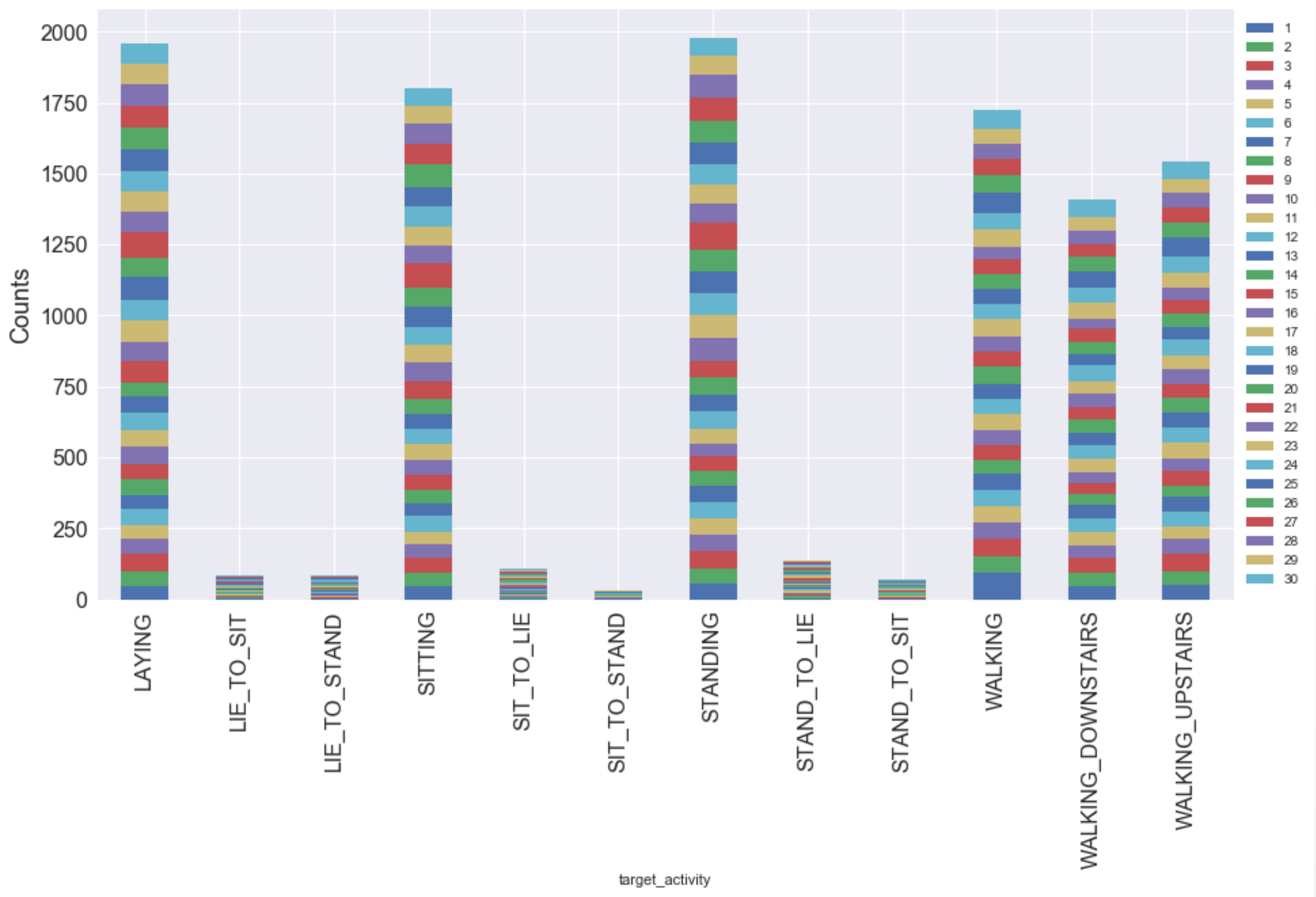
a) The signals obtained from accelerometer and gyroscope sensors; b) Each signal on (a) has the measures on the right for computing the feature vectors.

Data Wrangling

- Loading TXT files.
- Renaming column names in feature dataset
- Mapping activity numbers to names in target dataset
- Concatenating train-test feature dataset to form a dataframe with size 10929x561.
- Concatenating train-test target dataset to form a dataframe with size 10929x1.
- Extracting and comparing activity counts per subject (Volunteer who performs the activities)

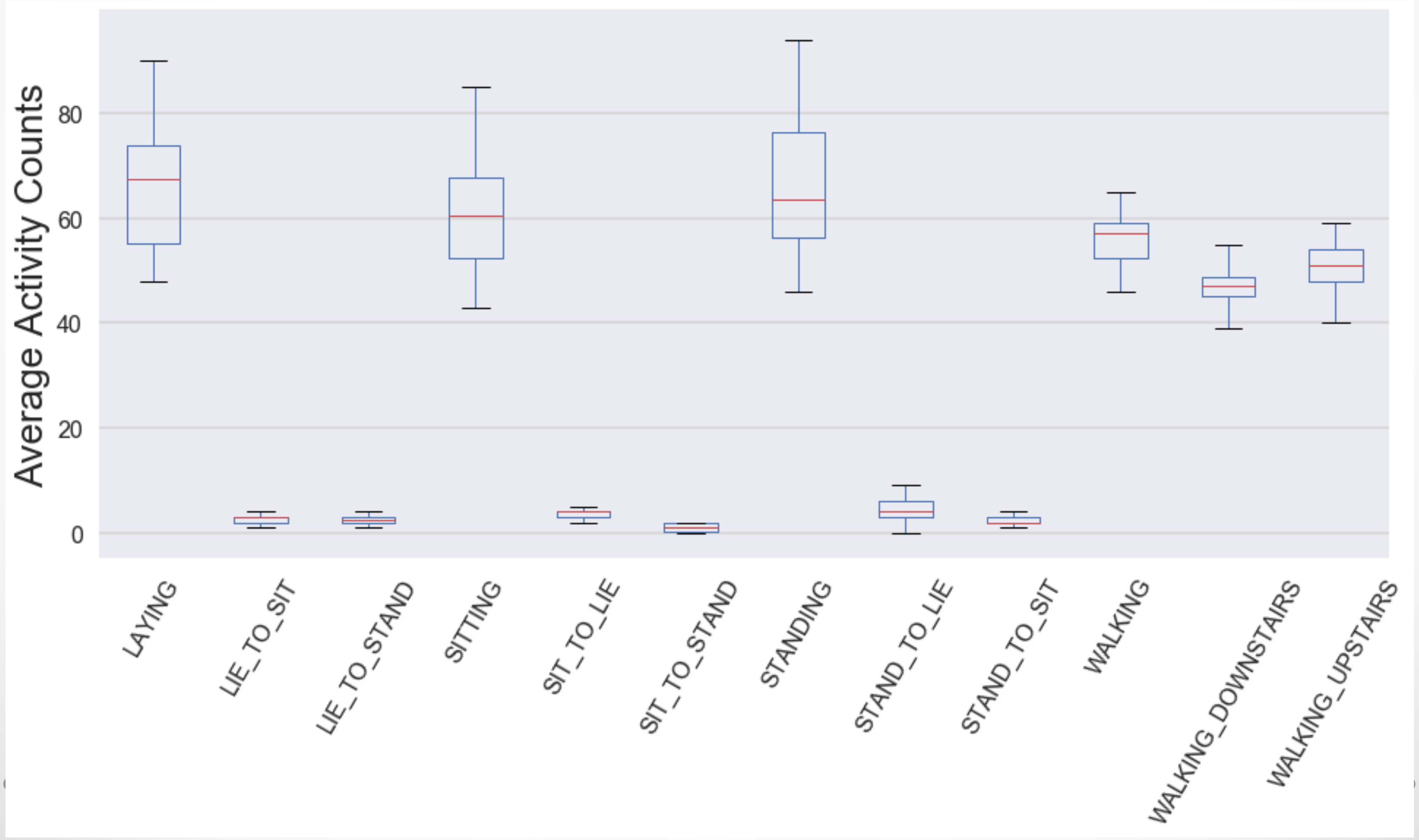
Data Visualization

- Activity counts per subject



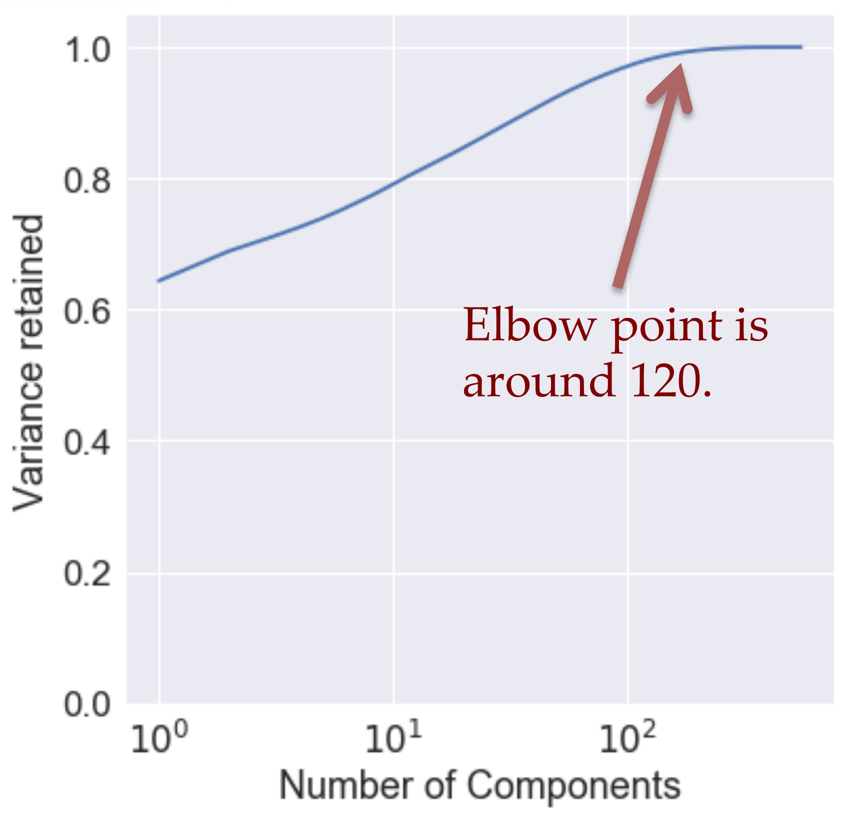
Data Visualization

- Average activity counts (Box Plot)



Feature Selection - PCA

- Scree Plot: # of components vs. explained variance ratio as a cumulative sum to find out how much data we are throwing away.



To be exact, we can explain 98% of the data with 123 principal components. Reduced from 561 to 123.

KNN

➢ n_neighbors = 3 gave the optimum performance.
➢ Kfold=5 cross validation on the whole and reduced datasets were performed.
➢ Avg. Cross-val-score is 92.48 ± 1.14%

Confusion Matrix on Reduced dataset

True Label \ Predicted Label	WALKING	WALKING_UPSTAIRS	WALKING_DOWNSTAIRS	SITTING	STANDING	LAYING	STAND_TO_SIT	SIT_TO_STAND	SIT_TO_LIE	LIE_TO_SIT	STAND_TO_LIE	LIE_TO_STAND
WALKING	1715	3	32	0	0	0	1	0	0	0	0	0
WALKING_UPSTAIRS	1	1523	21	0	2	0	1	0	1	0	2	1
WALKING_DOWNSTAIRS	2	2	1264	0	0	0	1	0	0	0	1	0
SITTING	1	10	47	1635	163	63	4	4	18	2	24	3
STANDING	3	6	37	152	1808	24	25	0	2	0	19	0
LAYING	0	0	6	14	6	1870	0	0	4	3	5	6
STAND_TO_SIT	0	0	0	0	0	0	35	0	1	0	1	0
SIT_TO_STAND	0	0	0	0	0	0	1	29	0	0	1	0
SIT_TO_LIE	0	0	0	0	0	0	0	0	72	0	18	0
LIE_TO_SIT	0	0	0	0	0	1	0	0	0	61	0	46
STAND_TO_LIE	0	0	0	0	0	0	2	0	9	0	67	0
LIE_TO_STAND	0	0	0	0	0	0	0	0	0	19	1	28

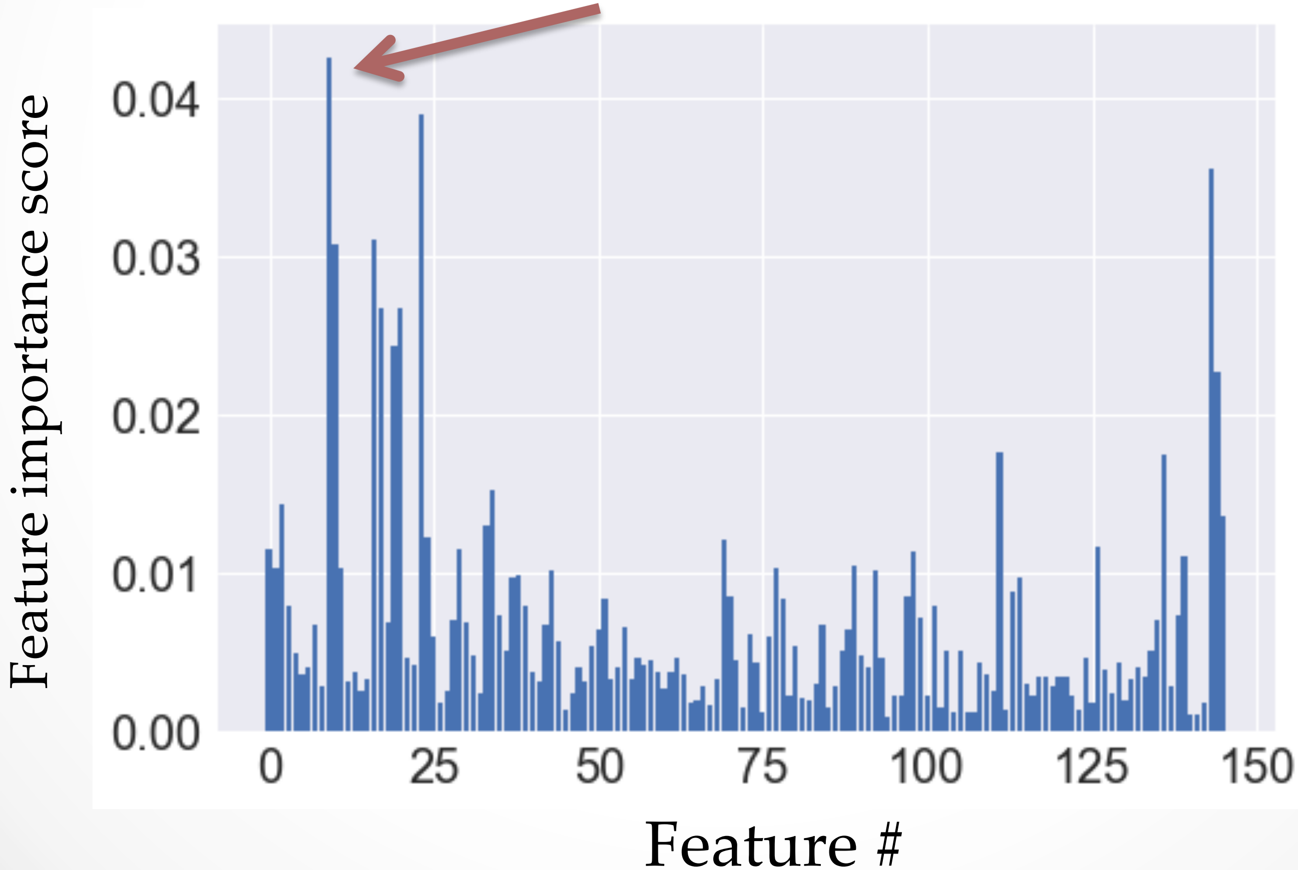
Feature Selection – Model Based

- Tree based classifiers provide the important of features. Features whose importance are greater than 0.75*mean (threshold value) were kept, the rest wasdiscarded.
- 146 features were selected. The barcode like figure shows selected (white) and discarded (black) features.

Feature Selection – Model Based

- Importance level of 146 features.

Most important feature: tBodyAcc-max X-axis.



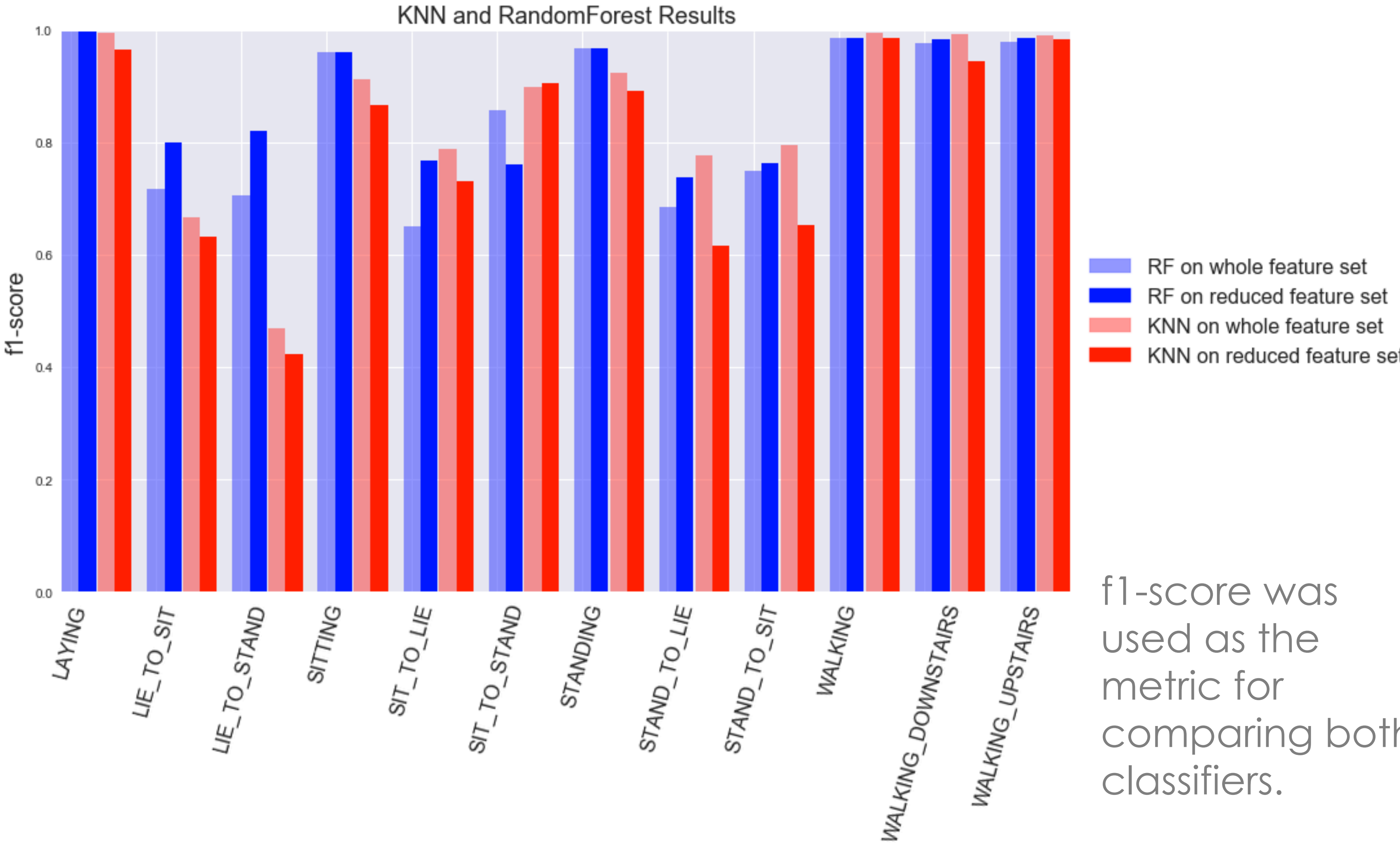
Random Forest

- GridSearchCV was performed on 3 tree parameters using 75% of the data. Optimum parameters were recorded.
- The model was tested with never-before-seen Test data.
- Best. Cross-val-score is 95.77%

Confusion Matrix on Reduced dataset

True Label \ Predicted Label	WALKING	WALKING_UPSTAIRS	WALKING_DOWNSTAIRS	SITTING	STANDING	LAYING	STAND_TO_SIT	SIT_TO_STAND	SIT_TO_LIE	LIE_TO_SIT	STAND_TO_LIE	LIE_TO_STAND
WALKING	420	0	3	0	0	0	0	0	0	0	1	0
WALKING_UPSTAIRS	3	353	2	0	0	0	1	0	0	0	2	0
WALKING_DOWNSTAIRS	4	2	340	0	0	0	0	0	0	0	0	0
SITTING	0	0	0	455	14	0	0	1	1	0	0	0
STANDING	0	0	0	19	487	0	0	0	0	0	0	0
LAYING	0	0	0	0	0	504	0	0	0	0	0	0
STAND_TO_SIT	0	0	0	1	0	0	13	3	0	0	0	0
SIT_TO_STAND	0	0	0	0	0	0	1	8	0	0	0	0
SIT_TO_LIE	0	0	0	0	0	0	0	0	20	1	6	0
LIE_TO_SIT	0	0	0	0	0	0	0	0	0	14	1	2
STAND_TO_LIE	0	0	0	0	0	0	2	0	4	0	24	1
LIE_TO_STAND	0	0	0	0	0	1	0	0	0	3	0	16

Model Comparison



f1-score was used as the metric for comparing both classifiers.

Model Comparison

Comparison tables for f1-scores

Table 1. f1-score for Postural Transitional Activity Classes

	KNN	RF
Original Dataset	0.74	0.71
Reduced Dataset	0.65	0.77

Table 2. f1-score of Complete Classes

	KNN	RF
Original Dataset	0.96	0.97
Reduced Dataset	0.93	0.97

Results

- There is no perfect recall and precision. There are confusions about all of the classes in both classifiers. The perfect 1.00 score is due to rounding.
- **The average precision and recall for Major Activities (Laying, Sitting, Standing, Walking, Walking_Downstairs, and Walking_Upstairs) are over 0.95 for both classifiers. There are not many false positives for these classes except for Standing and for Sitting classes, which are confused with each other.**
- The Transitional Activities show worse f1-scores than the Major Activities (Tables 1,2).
- **The Transitional Activities for KNN classifier show worse average f1-score for Reduced Dataset compared to original dataset.**
- As for RF classifier, the classification scores of Major and Transitional Activities marginally increased compared to KNN and the confusion rate between Standing and for Sitting classes decreased.
- **RF Classifier is able to predict better in a data imbalance situation, i.e. between Major and Transitional Activities. A tuned RF performs better than KNN for HAPT classification problem.**
- KNN performed worse with Reduced dataset when the Whitening option of PCA was used.

Potential Clients

- Continuous monitoring of patients (Healthcare Management).
  - Healthcare monitoring, lifecare assessment
- Smart environments (Retail).
  - Tracking individuals movements could help create unique shopping plans for shoppers.
- Public surveillance (Security).
  - Activity tracking could be upgraded to trajectory tracking by including more classes and using physical characteristics of persons.

Client Recommendations

- For healthcare management, predicting major activities could be enough for following patient's mobility and calorie tracking. Along with other information such as age, height and weight, it would be easy to track an individual's life style.
- **Companies in smart environments and public surveillance should focus on collecting more data on postural transitions to increase the detection ability.**
- Companies in public surveillance should invest in tracking an individual's trajectory using Human Activity Dataset. It is possible to build a trajectory for an individual if more classes are added as targets. For instance, the data to be collected when turning right and left, and changing lanes could help with drawing a trajectory.
- **The company should look at the most important features that help classify an activity. Using less number of features will help energy effective real-time tracking due to less intensive CPU use.**