WISDM - Activity Prediction using Sensor Data

Presented by: Mai Nguyen

About the dataset

Source: UCI Machine Learning Repository

Members of the WISDM (Wireless Sensor Data Mining) Lab in the Department of Computer and Information Science of Fordham Unversity collected data from the accelerometer and gyroscope sensors of a smartphone and smartwatch as **51 subjects** performed **18 diverse activities** of daily living. Each activity was performed for 3 minutes, so that each subject contributed 54 minutes of data.

Being able to sense/ recognize human activity can be a huge asset to many fields including but not limited to healthcare, construction, etc...

There are a total of 51 files for 2 types of sensor datas from phone and watch, so a total of 51*4 = **204 arff data files**. For each, there are a total of **93 features** (2 categorical; 91 continuous). Overall, there are approx. **75000~ data points.**

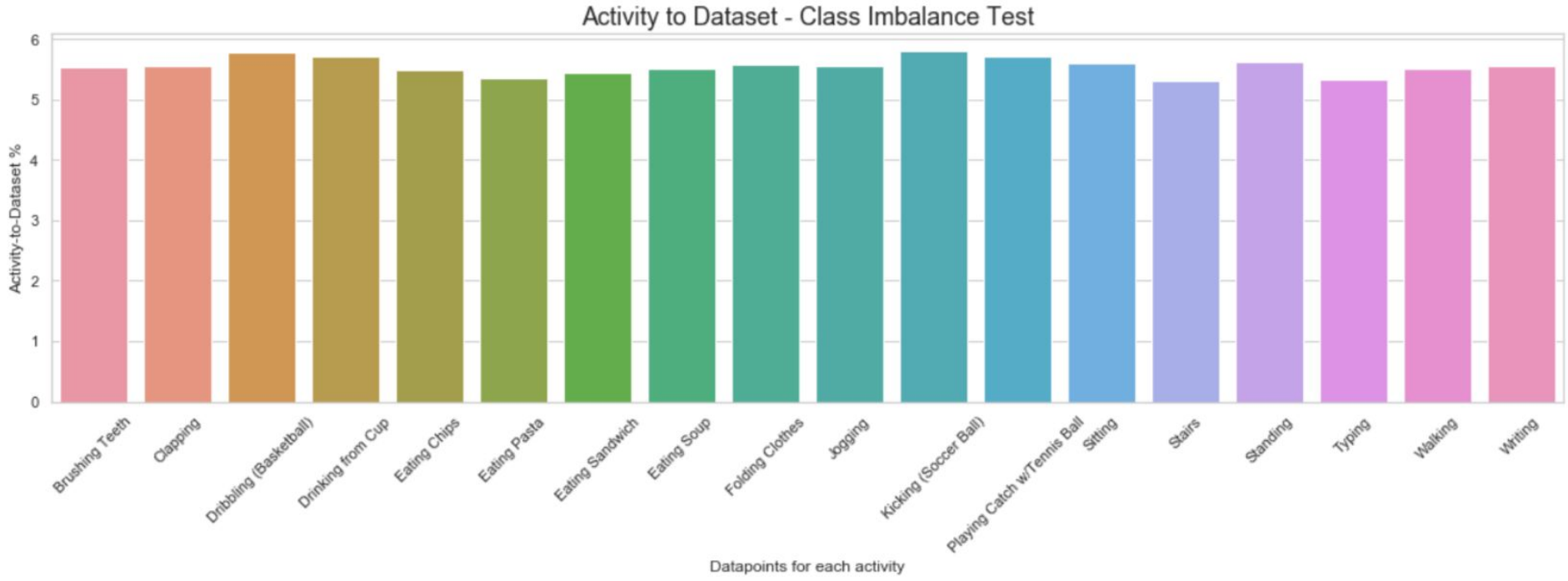
18 Categories

- Non-hand-oriented activities: walking, jogging, stairs, standing, kicking
- Hand-oriented activities (General): dribbling, playing catch, typing, writing, clapping, brushing teeth, folding clothes
- Hand-oriented activities (eating): eating pasta, eating soup, eating sandwich, eating chips, drinking

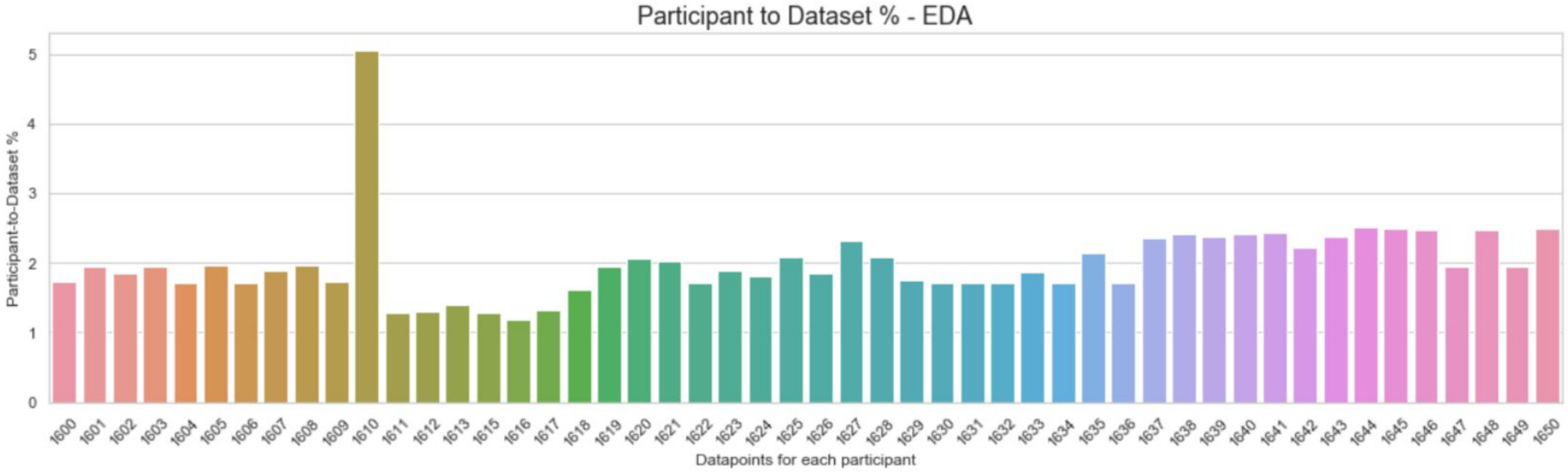
Inspiration/ Research Questions

- Can this data be used to predict the human activity?
- What are some of the most important variables for a model to predict activities?
- Which model is best for the task? (fatest, most accurate)

EDA - Categories balance



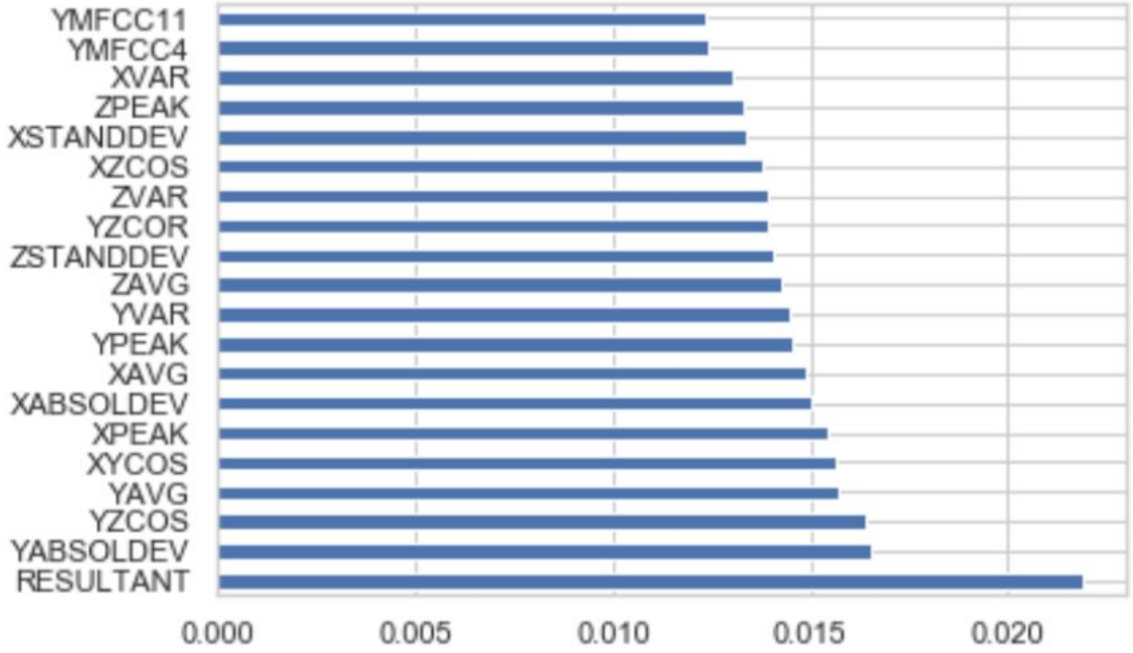
EDA - Categories imbalance (cont)



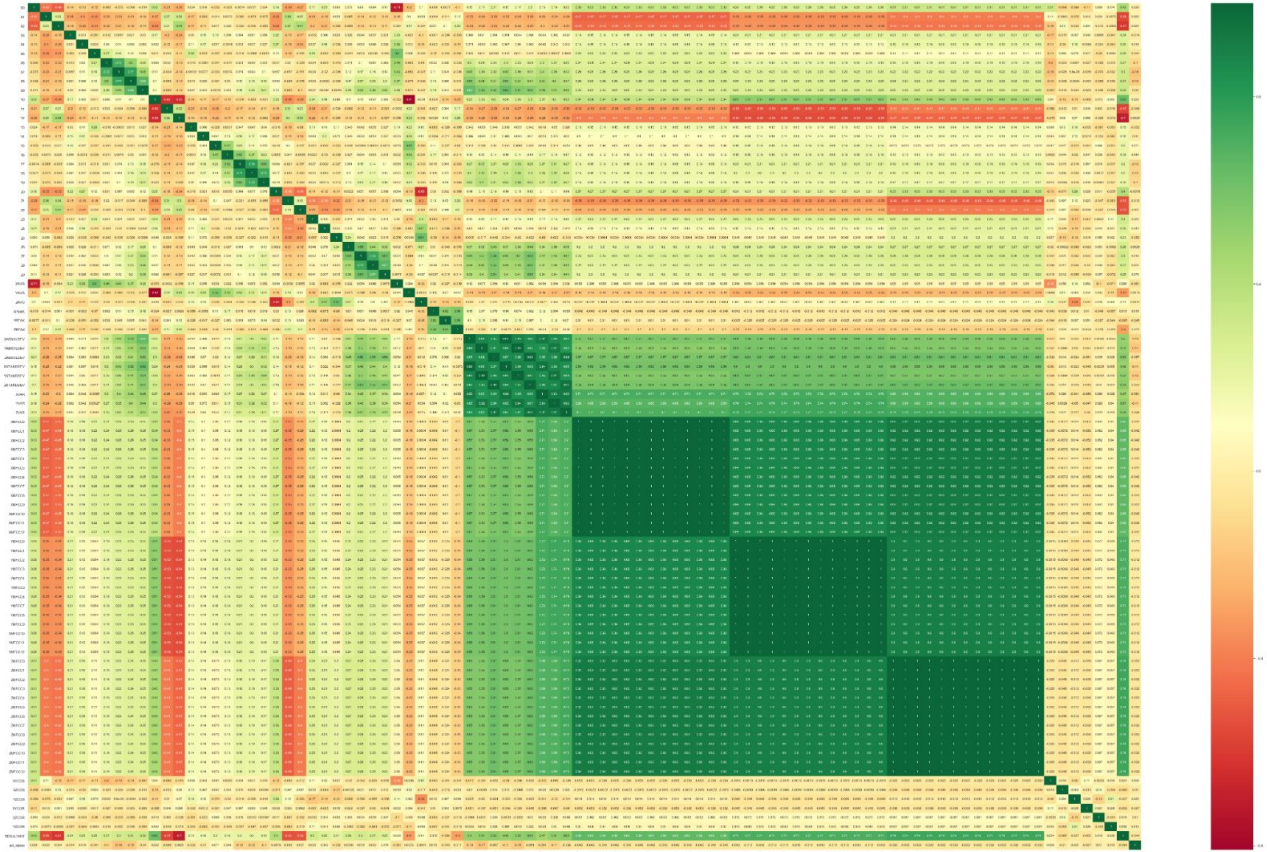
The data collected for different participants are not as well-balanced but are not horrible. Some participants' data is only 1% of the dataset length, while there's one participant with 5% length of the whole dataset.

EDA - Feature importance

*Feature importance gives a score for each feature of our data: the **higher** the score **more important** or relevant is the feature towards the output variable.*



EDA - Correlation Matrix



Building Models - Evaluation Metrics

Before building our models, let's look at the metrics we'll use to evaluate them. Here, I use Classification Reports for all models, which include:

1. **Accuracy Score (Precision):**
   - Equation: TP/ (TP + FP)
   - Measures the ability of a classifier to identify only the correct instances for each class
2. **Recall**:
   - Equation: TP/ (TP + FN)
   - Measures the ability of a classifier to find all correct instances per class
3. **f1-score:**
   - Is a weighted harmonic mean of precision and recall normalized between 0 and 1.
   - F-score of 1 indicates a perfect balance as precision and the recall are inversely related.
4. **Support:**
   - is the number of actual occurrences of the class in the test data set.
   - Imbalanced support in the training data may indicate the need for stratified sampling or rebalancing.

Phase 1: PCA=3

Activities	Gbm 1	Knn 1
Brushing Teeth	0.1700	0.1200
Claaping	0.1700	0.1600
Dribbling (Basketball)	0.1300	0.2200
Drinking from cup	0.1200	0.1000
Eating Chips	0.0900	0.1200
Eating Pasta	0.0800	0.1200
Eating Sandwich	0.0600	0.1200
Eating Soup	0.0800	0.0900
Folding Clothes	0.1800	0.1300
Jogging	0.3700	0.2700
Kicking (Soccer ball)	0.2200	0.1700
Playing Catch w/ tennis b..	0.2400	0.1900
Sitting	0.1400	0.1100
Stairs	0.2000	0.1200
Standing	0.0400	0.0800
Typing	0.1900	0.1200
Walking	0.2200	0.1700
Writing	0.2500	0.1700

With only 3 PCA variables, approximately **92%** of the dataset variance is already explained.

Average F1-scores for all models:

- ➔ For KNN: 0.15
- ➔ For GBM: 0.19 <before and after tuning>

Evaluation - Phase 1:

The GBM model outperforms KNN by a small 2%, and improved slightly after tuning.

Phase 2: Include PCA =1 in original features

Activities	Svc 2	Svc Ovr	Gbm 2	Knn 2	Rfc 2
Jogging	0.6500	0.5700	0.9000	0.9500	0.9400
Walking	0.4000	0.5300	0.8500	0.8700	0.9000
Kicking (Soccer ball)	0.3800	0.5600	0.7400	0.7400	0.7900
Dribbling (Basketball)	0.4000	0.3900	0.7000	0.7900	0.8100
Claaping	0.2800	0.4000	0.7000	0.7700	0.8500
Folding Clothes	0.2700	0.4800	0.6500	0.7200	0.7800
Stairs	0.2300	0.3800	0.7400	0.7400	0.8000
Writing	0.2600	0.4700	0.6700	0.6600	0.7900
Brushing Teeth	0.2700	0.4300	0.6600	0.7000	0.7600
Playing Catch w/ tennis b..	0.2000	0.2800	0.6900	0.7600	0.7800
Typing	0.2600	0.3800	0.6300	0.6500	0.7600
Standing	0.3600	0.4200	0.5300	0.5700	0.6700
Sitting	0.2400	0.3500	0.5900	0.5800	0.7000
Drinking from cup	0.2400	0.3900	0.5500	0.6100	0.6600
Eating Soup	0.2100	0.3100	0.5100	0.5700	0.6700
Eating Pasta	0.1300	0.3400	0.5000	0.5700	0.6400
Eating Chips	0.2400	0.2400	0.4800	0.5300	0.6100
Eating Sandwich	0.1600	0.2000	0.4300	0.5200	0.5800
Average	0.2878	0.3956	0.6400	0.6833	0.7494

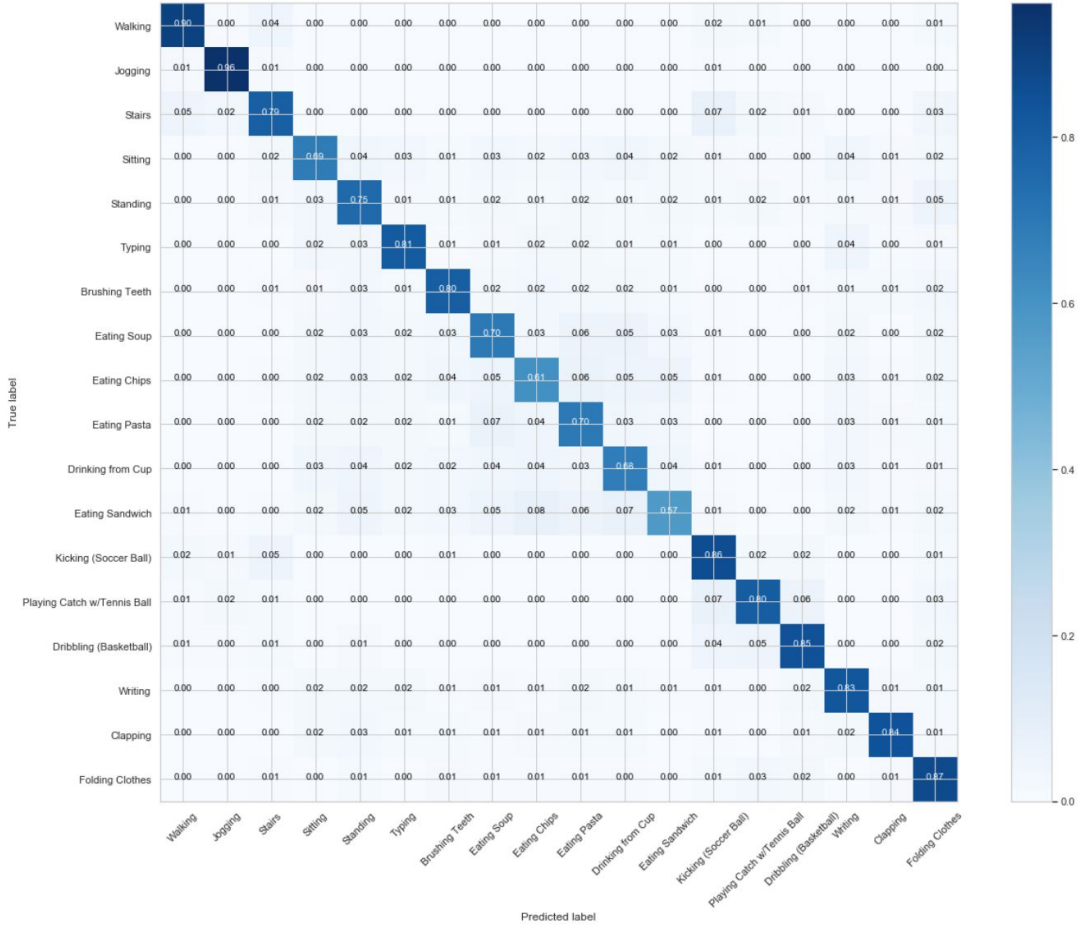
Phase 3: PCA = 30

Activities	Gbm 3	Knn 3	Rfc 3
Jogging	0.8900	0.9500	0.9500
Walking	0.7700	0.8900	0.9000
Kicking (Soccer ball)	0.6900	0.8000	0.8000
Dribbling (Basketball)	0.6700	0.8600	0.8400
Claaping	0.6700	0.8500	0.8700
Folding Clothes	0.6500	0.8200	0.8000
Stairs	0.6500	0.8000	0.8000
Writing	0.6400	0.7900	0.8000
Brushing Teeth	0.5900	0.7600	0.8000
Playing Catch w/ tennis b..	0.6500	0.8300	0.8100
Typing	0.5800	0.7900	0.8100
Standing	0.5500	0.7500	0.7200
Sitting	0.5800	0.7200	0.7300
Drinking from cup	0.5000	0.6800	0.6700
Eating Soup	0.4900	0.7100	0.7000
Eating Pasta	0.5100	0.6900	0.6900
Eating Chips	0.4500	0.6300	0.6400
Eating Sandwich	0.4200	0.6300	0.6400
Average	0.6083	0.7750	0.7761

Evaluation - Final Phase 3

Activities	Gbm 1	Gbm 2	Gbm 3	Knn 1	Knn 2	Knn 3	Rfc 2	Rfc 3	Svc 2	Svc Ovr
Brushing Teeth	0.1700	0.6600	0.5900	0.1200	0.7000	0.7600	0.7600	0.8000	0.2700	0.4300
Claaping	0.1700	0.7000	0.6700	0.1600	0.7700	0.8500	0.8500	0.8700	0.2800	0.4000
Dribbling (Bask..	0.1300	0.7000	0.6700	0.2200	0.7900	0.8600	0.8100	0.8400	0.4000	0.3900
Drinking from c..	0.1200	0.5500	0.5000	0.1000	0.6100	0.6800	0.6600	0.6700	0.2400	0.3900
Eating Chips	0.0900	0.4800	0.4500	0.1200	0.5300	0.6300	0.6100	0.6400	0.2400	0.2400
Eating Pasta	0.0800	0.5000	0.5100	0.1200	0.5700	0.6900	0.6400	0.6900	0.1300	0.3400
Eating Sandwich	0.0600	0.4300	0.4200	0.1200	0.5200	0.6300	0.5800	0.6400	0.1600	0.2000
Eating Soup	0.0800	0.5100	0.4900	0.0900	0.5700	0.7100	0.6700	0.7000	0.2100	0.3100
Folding Clothes	0.1800	0.6500	0.6500	0.1300	0.7200	0.8200	0.7800	0.8000	0.2700	0.4800
Jogging	0.3700	0.9000	0.8900	0.2700	0.9500	0.9500	0.9400	0.9500	0.6500	0.5700
Kicking (Soccer..	0.2200	0.7400	0.6900	0.1700	0.7400	0.8000	0.7900	0.8000	0.3800	0.5600
Playing Catch ..	0.2400	0.6900	0.6500	0.1900	0.7600	0.8300	0.7800	0.8100	0.2000	0.2800
Sitting	0.1400	0.5900	0.5800	0.1100	0.5800	0.7200	0.7000	0.7300	0.2400	0.3500
Stairs	0.2000	0.7400	0.6500	0.1200	0.7400	0.8000	0.8000	0.8000	0.2300	0.3800
Standing	0.0400	0.5300	0.5500	0.0800	0.5700	0.7500	0.6700	0.7200	0.3600	0.4200
Typing	0.1900	0.6300	0.5800	0.1200	0.6500	0.7900	0.7600	0.8100	0.2600	0.3800
Walking	0.2200	0.8500	0.7700	0.1700	0.8700	0.8900	0.9000	0.9000	0.4000	0.5300
Writing	0.2500	0.6700	0.6400	0.1700	0.6600	0.7900	0.7900	0.8000	0.2600	0.4700

Best Model - KNN 3 Confusion matrix



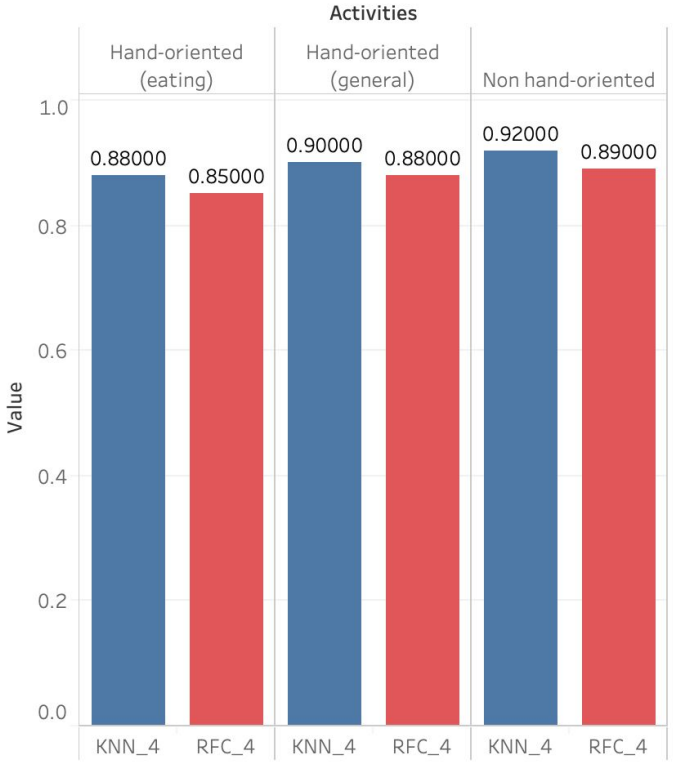
Reshape & Redefine the problem

So far, my best models have yielded **78% accuracy score** (81% if 90 pca components are to be used). I can't help but wonder, would being able to identify exactly each of these 18 activities help with any health research studies or businesses?

I've made peace with the answer that it... might. But maybe being able to predict the general group of activities the participant might be doing is enough. Hence, in the next part of this project, I'm reshaping the problem question at hand, diving these 18 activities into 3 broader groups and attempt to classify these **3 groups.**

- Non-hand-oriented activities
- Daily activities (General)
- Hand-oriented activities (eating)

Phase 4: No PCA



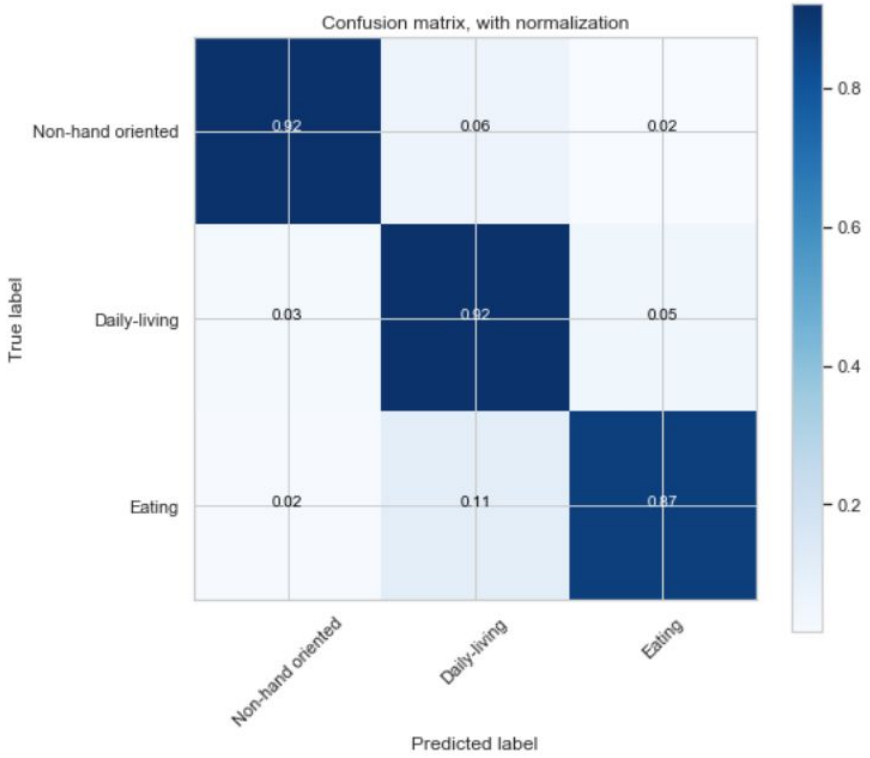
Great. We now see much better performing models. Suprisingly, KNN outperforms our RFC model with an average accuracy score of 90% (KNN) over 87% (RFC).

Phase 5: PCA = 30

Evaluation - Final Phase 5

Activities	KNN_4	KNN_5	RFC_4	RFC_5	SVC_5
Hand-oriented (eating)	0.8800	0.9100	0.8500	0.8700	0.5900
Hand-oriented (general)	0.9000	0.9100	0.8800	0.8900	0.6300
Non hand-oriented	0.9200	0.9100	0.8900	0.9100	0.6900

Best Model 2 - KNN 5 Confusion Matrix



Final Recommendations

- Considering the following metrics when choosing the final model:
  - Speed (Wall time & CPU time)
  - Computational resources
  - Performance
- My final recommendation would be to use KNN models for both: Categorizing 18 activities individually & Categorizing bigger groups of activities. The tradeoff for quick processing time is huge; yet KNN had been much faster to run than any of the other models listed.
- In conclusion, the performance score for **KNN model with PCA = 30**:
  - When predicting 18 activities individually: Accuracy score = F1 score = **78%**
  - When predicting 3 groups of activities: Accuracy score = F1 score = **91%**