**Data import**

1. Get user ID’s from the filenames
2. Compare UserData files & Tappy files.
   * Use only those ID’s that have both a UserData file and a Tappy file

**User Data:**

1. Change ‘Gender’ column to OHE ‘Male’
2. Convert string features (Parkinsons, Tremors, Levadopa, etc) to binary columns
3. OHE ‘Sided’, ‘UPDRS’ and ‘Impact’

**Tappy Data:** Create new features.

1. Mean, std, skewness, kurtosis for:
   1. Hold Times per hand
   2. Latency Times per direction
2. Mean difference of
   1. Hold Time per hand
   2. Latency Time for different hand direction (LR – RL)
   3. Latency Time for same hand direction (LL – RR)

**Data cleaning**

**Missing Data**

1. User Data:
   1. DiagnosisYear 🡪 fillna(0)
   2. BirthYear 🡪 fillna(0)
2. Tappy Data:
   1. Mean imputation on NaN values.

**Outliers**

1. User Data:
   1. Nothing done
2. Tappy Data:
   1. Nothing done

**Factor Reduction / Feature Extraction:**

1. Extract relevant features for:
   1. Hold Times
   2. Latency Times
2. Based on:
   1. ~~# of factors~~
   2. Cumulative explained variance ratio >0.9 ?
   3. Others?

**Results PCA:**

1. Hold times: 3 factors with EVR > 0.9
   1. PC 1 & 2 correlate strongly
   2. PC 1 & 3 correlate strongly
   3. PC 2 & 3 is ok.
2. Lat times: 6 factors with EVR > 0.9
   1. PC 3 correlates strongly with every other factor.
   2. Other PC’s are ok.