Slide 2 Introduction

The data is collected from a wearable single accelerometer which is mounted on the chest. The sampling frequency of the accelerometer is 52hz. This is conducted on 15 people and each of them have different movements which the accelerometer records. The dataset which is available is in CSV format. The target column of the dataset is the actions performed by each of the 15 people.

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

Slide 3 – Data Prep

There are 15 datasets present, 1 dataset for each person and his actions. These datasets were loaded but did not have a column name. Column names were given, and sequence column was dropped because it was not necessary.

Person indication column was added to each of these datasets. This was done by adding a new column called ‘person’ and all the datasets were concatenated into one.

The parameters in each of them included: count, x\_acceleration, y\_acceleration, z\_acceleration, Label (Target feature). The labels are codified numbers which are –

1. Working at Computer
2. Standing up, Walking and Going up-down the stairs
3. Standing
4. Walking
5. Going Up-Down Stairs
6. Walking and Talking with someone
7. Talking while standing

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

Slide 4 - State the hypothesis

**Null Hypothesis** : Each of the accelerations in the data is interdependent with the label.

* For example:
* x axis, y axis acceleration is dependent on forward and backward movement like walking and climbing stairs
* z axis acceleration is dependent on upward movement like climbing up the stairs, standing up.

**Alternate Hypothesis** : The accelerations are not dependent on the label.

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

Slide 5 – Hypothesis continued

* + X-Acceleration vs Labels
    - Most of the values which were taken a reading was for Standing, walking and working on the computer
  + Y-Acceleration vs Labels
    - Working on the computer and Talking while standing triggered most of these values
  + Z-Acceleration vs Labels
    - It is seen that going up/down the stairs had triggered most of the values for z- acceleration

So we can conclude that alternate hypothesis can be rejected.

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

Slide – 6

## Data Preparation

* The target was label encoded. Although it was already between 1 and 7, it was more consistent to keep the data between 0 to 6. This was done using sklearn label encoder.
* Moving on to the other column, it is a numerical descriptive data. This needs to be standard scaled. This is much more efficient than min max scaler as the mean is 0 and the new standard deviation is 1
* Splitting the data into test, train with split of 30:70. We can now move on to fitting a classifier.

## Fitting a classifier

* K-Nearest Neighbour: The accuracy without any parameters generated was 0.718
* Decision Tree Classifier: When the data was fit in the DT the accuracy generated was 0.518

## Hyper Parameter Fine Tuning

* KNN classifier

Using repeated Stratified K folds, fine tuning for the parameters for Decision tree was done and the following results were obtained

* + Best Parameters: gini index, max depth of 12 and min sample split is 3
  + Best DT score was 0.713

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

Slide – 7

After looking at the accuracy scores of KNN and Decision Trees, we can conclude that KNN is doing better than DT. The prediction score for KNN was obtained as 0.756. Comparing the two classification matrices and confusion reports we can see how many True Positive and False Positives are being returned. We can also look at the F1 score for better understanding of each of the target feature. Looking at the F1-scores, we can see the model is not very good when it comes to label 1,4,5. While it is doing extremely well for the other labels.

Slide = 8

KNN has given a decent accuracy of predicted score of 75.6%. The accuracy increased drastically when parameter tuned. But it looks like KNN can predict accurately only for some labels like sitting and working on a computer, standing and Talking while standing. To get better results, we need to use different classifiers with appropriate feature selection. This shows that there probably needs to be more parameters for getting a better score for going up/down the stairs, walking. Right now, to avoid the conclusion we can confidently say the accuracy is valid only for some of the features. Maybe if this was in a controlled and monitored environment, we could get better accuracy of the results but then again, the whole point of an accelerometer is to be implemented in real world situations.