

Assigment -2 Python For Data Science

VARUN RAMESH – s3793675

June 10, 2020

RMIT University

+61 480142470

To Predict the activity of the user based on the movement of chest mounted Accelerometer

Contents

[Introduction 1](#_Toc42699222)

[Methodology 2](#_Toc42699223)

[Retrieving Data and Data Preparation 2](#_Toc42699224)

[Data Preparation 2](#_Toc42699225)

[Data Exploration 2](#_Toc42699226)

[Labels – Pie Chart 4](#_Toc42699227)

[Acceleration x vs Label Density Plot 5](#_Toc42699228)

[Acceleration y vs Label Density Plot 5](#_Toc42699229)

[Acceleration z and Label Density plot 6](#_Toc42699230)

[Evaluation between attributes 6](#_Toc42699231)

[Scatter Matrix: 6](#_Toc42699232)

[Correlation Matrix and Heat Map 8](#_Toc42699233)

[Data Modelling 9](#_Toc42699234)

[Data Preparation 9](#_Toc42699235)

[Fitting a classifier 10](#_Toc42699236)

[Hyper Parameter Fine Tuning 10](#_Toc42699237)

[Predictions and Results 11](#_Toc42699238)

[Conclusions 11](#_Toc42699239)

[References 12](#_Toc42699240)

I certify that this is all my own original work. If I took any parts from elsewhere, then they were non-essential parts of the assignment, and they are clearly attributed in my submission. I will show we I agree to this honor code by typing “Yes”:

# 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.

There are 15 datasets given based on each person. 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

The data is pre-processed and fit into a classifier to predict the action based on the accelerometer readings. This way it gives a better user interface capability if incorporated with watches and phones. The steps for doing this are –

1. Retrieving Data and Preparation
2. Exploration between attributes
3. Data Modelling
4. Fitting a classifier
5. Predicting

# Methodology

## Retrieving Data and Data Preparation

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.

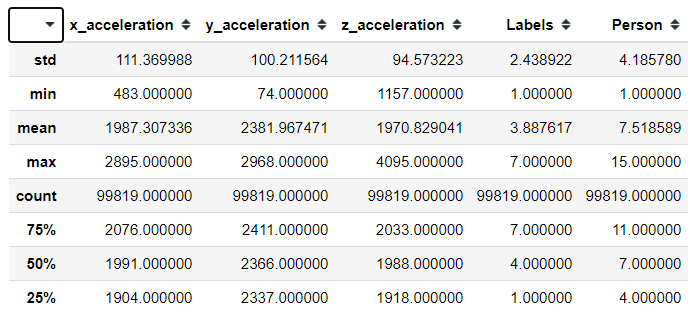
Person indication column was added to each of these datasets. This was done by adding a new column called ‘person’ and all the datasets were concatenated into one.

## Data Preparation

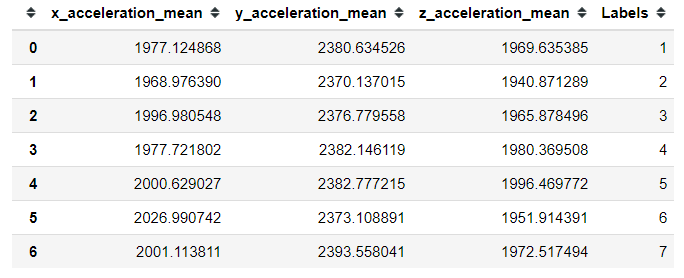
A random sample of 100,000 was taken from the dataset which was 1,926,896 rows. Data had no null values however it had a label called zero which needed to be treated as a missing value. This was because there was no label assigned to zero.

# Data Exploration

The descriptive statistics the data are:

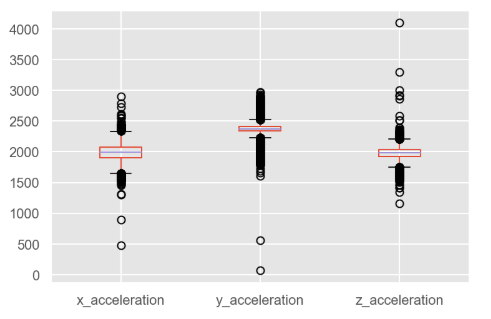


After having the descriptive statistics, the means of accelerations were found out for each of the label. The mean for each of the activities should be different, but let is find out:

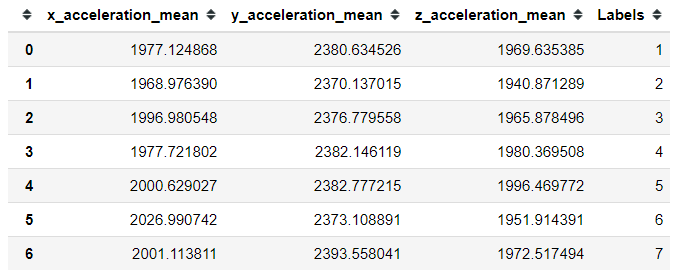


From the above means, it looks like y\_acceleration is always more irrespective of the activity done. Looks like activity number 2 (Standing up, Walking and Going up\down the stairs is the least for all parameters). The label looks very inconclusive considering it has all the activities in it. The highest value for z\_acceleration is when the person is going up/down the stairs (label 5). The hypothesis from y\_acceleration mean is very inconclusive. x\_acceleration is the highest when the person in walking and going up/down the stairs (label 5,6)

Moving on to see if there are any outliers, this is the plot which was obtained.

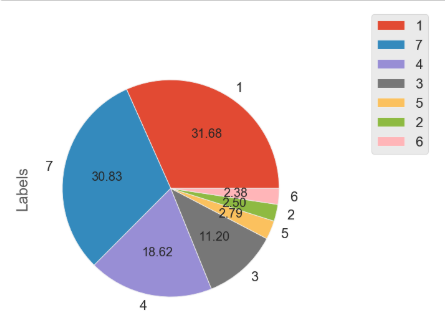


Removing the outliers, the descriptive statistics did not change this shows. So, it is best to keep the data with the probable outliers.



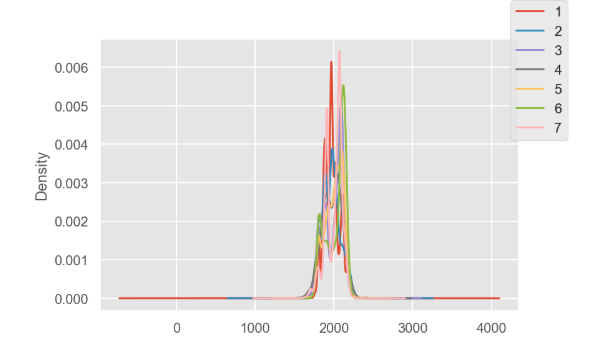
### Labels – Pie Chart

Moving on to exploring the column of only label, we found that most of the data was distributed among Label 1, 7 and 4. These are working on computer, talking while standing and walking respectively.



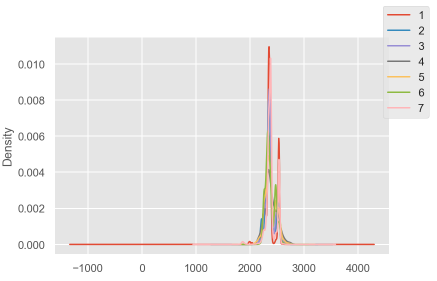
### Acceleration x vs Label Density Plot

Comparing both these graphs, we can see that x\_acceleration is the between 1600 - 2300. This is true for all the labels. However, as we can see the pink line represents label 7 which is talking while standing. This has the highest probability for higher values.



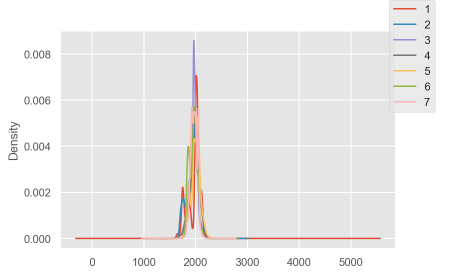
### Acceleration y vs Label Density Plot

We can see that most of the y\_acceleration values lie from 2300 - 2700. Label 1 and 7 have the highest spike in the graph. The average probability density of the data is much higher than x\_acceleration. This could mean that there is continuous movement in the y\_direction.



### Acceleration z and Label Density plot

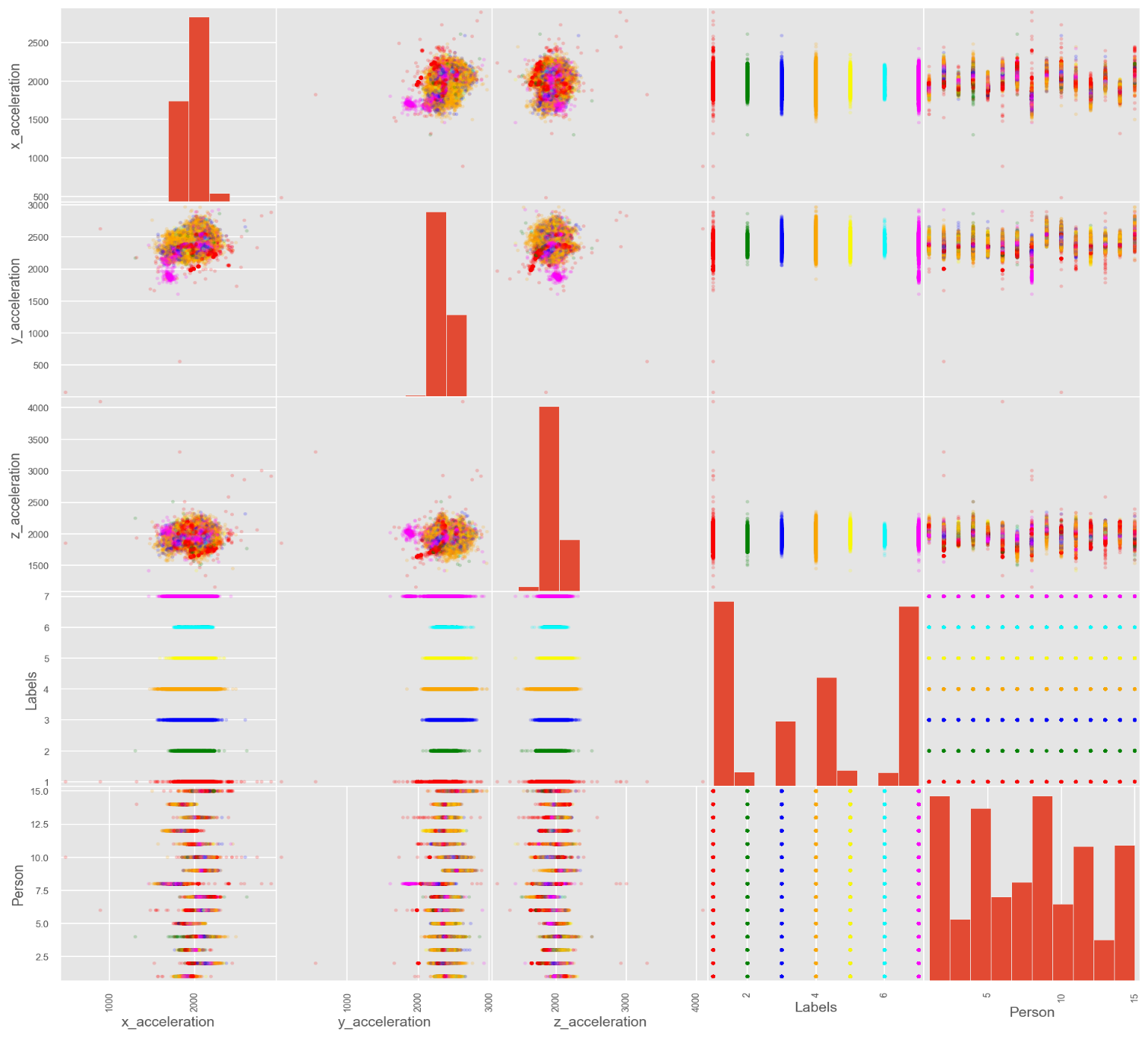
As we can see, the majority of the values are between 1800 to 2200. The highest spike of accelerometer is noticed when standing up. We can see that z\_acceleration has significantly lower acceleration rate compared to its counter part accelerations. So we can assume that any spike in the acceleration is due to an immediate activity in the upward direction.



## Evaluation between attributes

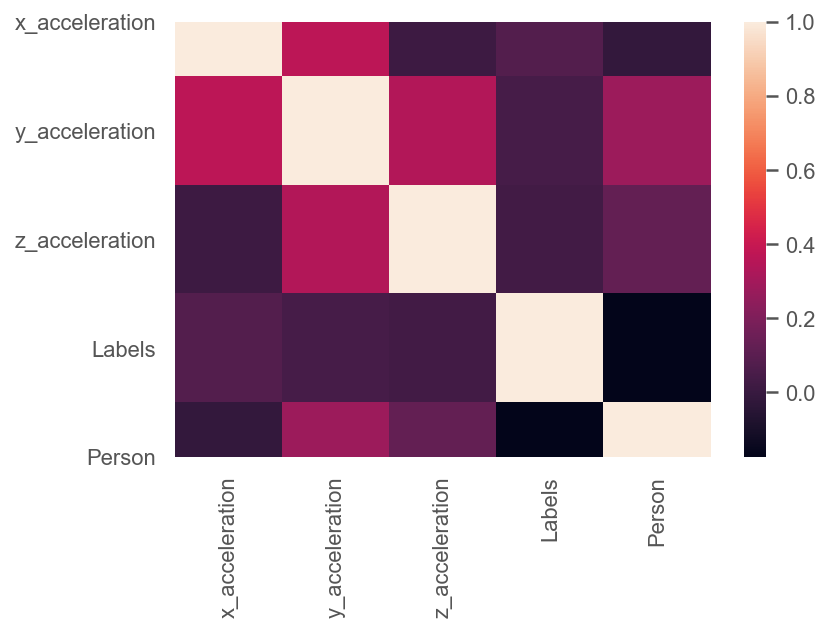
### Scatter Matrix:

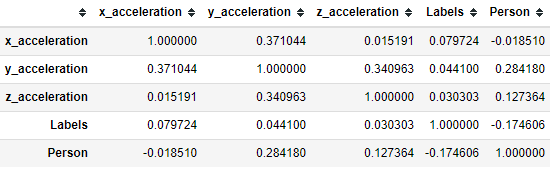
* We get an overview of all the attributes and the comparison between each other. The diagonal consists of histograms of each of the attributes which are compared with each other.
* The labels (actions) are colour coded as follows:
  + Red – Working at the computer
  + Green – Standing up, walking and Going up/down the stairs
  + Blue – Standing
  + Orange – Walking
  + Yellow – Going up/Down the stairs
  + Cyan – Walking and talking to someone
  + Magenta – Talking while standing
* X-Acceleration vs Y-Acceleration
  + The scatter plot has a lot of yellow, and a few red, Magenta. This means most of the acceleration is recorded when Going up/Down the stairs. However it is also noticed that working on the computer as well as talking while standing also had more x & y acceleration recorded than walking/ walking and talking to someone
* Y-Acceleration vs Z-Acceleration
  + The values for y & z acceleration are much lower compared to x & y. This is because z acceleration is not much while standing and walking and it is to a minimal. Although, we can see going up/down the stairs triggers z acceleration as well as y acceleration
* X-Acceleration vs Z-Acceleration
  + We can see in this the graph is redder and purpled. This means most of these values are for working at the computer and talking while standing
* Acceleration vs Labels
  + 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



### Correlation Matrix and Heat Map

Looking at the Pearson correlation we can see that x & y acceleration and y & z acceleration have close to 40% correlation while x & z acceleration have almost zero correlation. This maybe because it is very hard for an activity to be in both x & z axis. The only possible activities that can have both x & z activities are moving sideways and also moving upwards which is very hard to achieve in the given labels above.





# Data Modelling

## 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

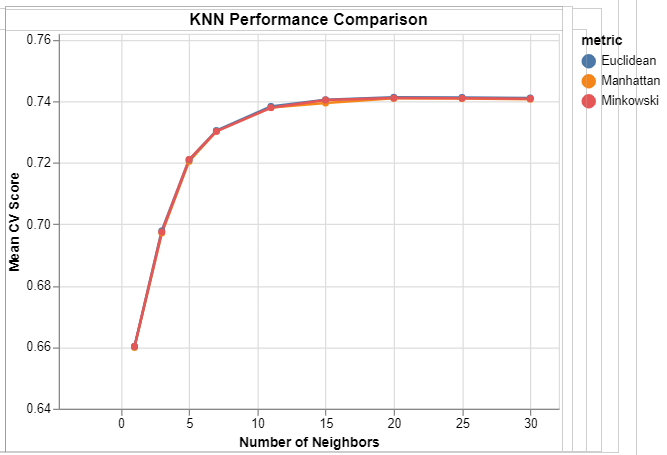
* K-Nearest Neighbour: The accuracy without any parameters generated was 0.718
* Decision Tree Classifier: When the data was fit in the DT the accuracy generated was 0.518

## 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

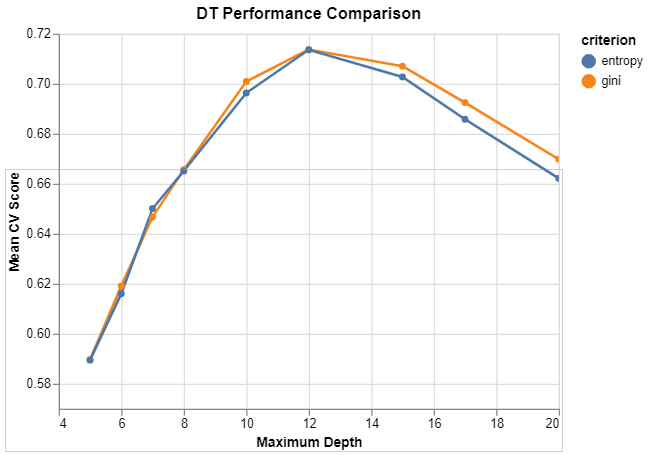
* + Best Parameters: gini index, max depth of 12 and min sample split is 3
  + Best DT score was 0.713
  + The graph shows how Maximum depth varies with mean CV score below.



* Decision Tree:

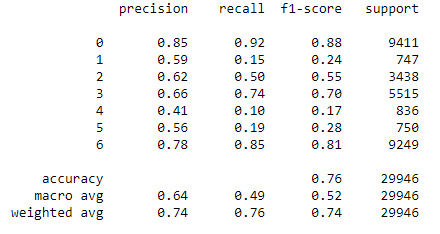
Using repeated Stratified K folds, fine tuning for the parameters for Decision tree was done and the following results were obtained

* + Best Parameters: gini index, max depth of 12 and min sample split is 3
  + Best DT score was 0.713
  + The graph shows how Maximum depth varies with mean CV score below.



# Predictions and Results

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.



# Conclusions

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.

# References

* Casale, Pierluigi & Pujol, Oriol & Radeva, Petia. (2012). Personalization and user verification in wearable systems using biometric walking patterns. Personal and Ubiquitous Computing - PUC. 16. 1-18. 10.1007/s00779-011-0415-z.
* Casale, Pierluigi & Pujol, Oriol & Radeva, Petia. (2012). Personalization and user verification in wearable systems using biometric walking patterns. Personal and Ubiquitous Computing - PUC. 16. 1-18. 10.1007/s00779-011-0415-z.
* Aksakali, V. (2019, July 16). *ml\_tutorials.* Retrieved from gitHub: https://github.com/vaksakalli/ml\_tutorials