College of Computer and Information Sciences
Department of Information Technology
IT 362 - Data Science Principles
First semester 1442 AH

جـــامـــعـــة
الملك سعود
King Saud University
1957

# IT 461 - Machine learning

## Course Project - Phase I

# The Effect of Exercises on People's Activeness

| Students Name | ID |
|---|---|
| Shaikha Bin Ateeq | 438201519 |
| Alanoud Alotaibi | 437200739 |

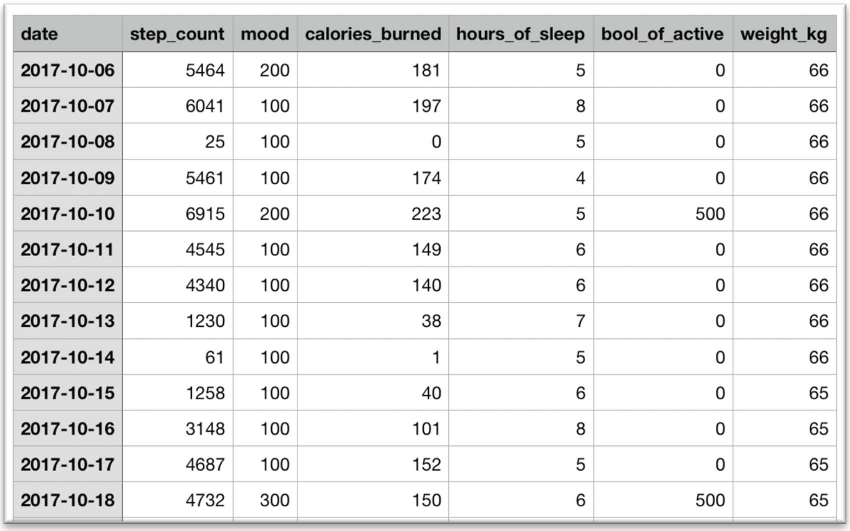
### Why I chose this dataset?

Since the exercise one of the most important thing in our lives it's can help to prevent excess weight gain or help maintain weight loss. When you engage in physical activity, you burn calories. The more intense the activity, the more calories you burn. Regular trips to the gym are great, but don't worry if you can't find a large chunk of time to exercise every day. Any amount of activity is better than none at all. To reap the benefits of exercise, just get more active throughout your day.

Our problem is about {"Does exercise/working-out improve a person's activeness?"}. The purpose of the project was to establish through two sets of data (control and experimental) if working-out/exercise promotes an increase in the daily step-count or not.

**We took this dataset from Kaggle site**: https://www.kaggle.com/aroojanwarkhan/fitness-data-trends

The dataset has **7** Attributes (columns) and **96** observation (rows).



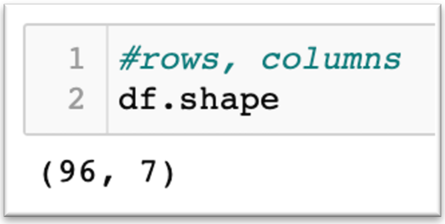
| date | step_count | mood | calories_burned | hours_of_sleep | bool_of_active | weight_kg |
|---|---|---|---|---|---|---|
| **2017-10-06** | 5464 | 200 | 181 | 5 | 0 | 66 |
| **2017-10-07** | 6041 | 100 | 197 | 8 | 0 | 66 |
| **2017-10-08** | 25 | 100 | 0 | 5 | 0 | 66 |
| **2017-10-09** | 5461 | 100 | 174 | 4 | 0 | 66 |
| **2017-10-10** | 6915 | 200 | 223 | 5 | 500 | 66 |
| **2017-10-11** | 4545 | 100 | 149 | 6 | 0 | 66 |
| **2017-10-12** | 4340 | 100 | 140 | 6 | 0 | 66 |
| **2017-10-13** | 1230 | 100 | 38 | 7 | 0 | 66 |
| **2017-10-14** | 61 | 100 | 1 | 5 | 0 | 66 |
| **2017-10-15** | 1258 | 100 | 40 | 6 | 0 | 65 |
| **2017-10-16** | 3148 | 100 | 101 | 8 | 0 | 65 |
| **2017-10-17** | 4687 | 100 | 152 | 5 | 0 | 65 |
| **2017-10-18** | 4732 | 300 | 150 | 6 | 500 | 65 |

- **Date:** the date which's doing the exercise ("Type: Interval").
- **Step count:** the number of steps that's take in a day ("Type: Discrete").
- **Mood:** either "**Happy**", "**Neutral**" or "**Sad**" which were given numeric values of **300**, **200** and **100** respectively (Type: Ordinal).
- **Calories:** The Burned calories in a day ("Type: Continuous').
- **hours of sleep:** number of hours per a day ("Type: Continuous").
- **Bool of active:** Feeling of activeness was measured in either "**Active**" or "**Inactive**" which were given numeric values of **500** or **0** respectively ("Type: Binary").
- **Weight:** weight in kg (Type: Continuous).

## Does it need pre-processsing, including normalization?

Yes, we need pre-processing for attribute “Date”, We don’t need it’s not added any benefit in analysis, so we delete it. Also, check if the dataset has a duplication row or missing values (No duplication, No missing values). No need for normalization all attributes integer.

**Find how many rows and columns in the dataset**

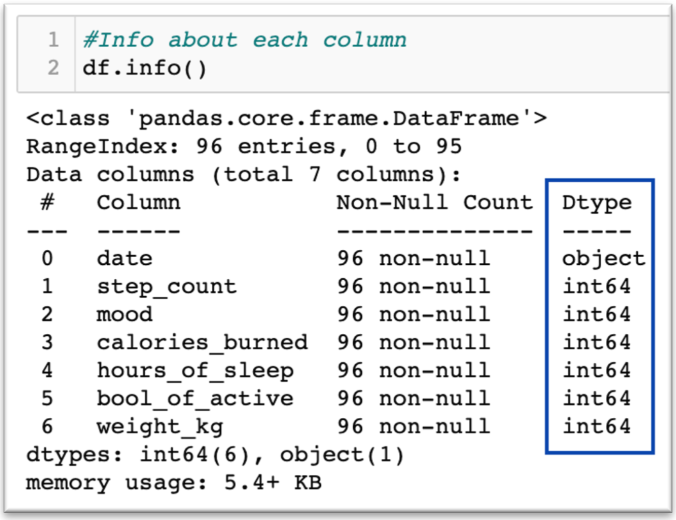


```
#rows, columns
df.shape
```

```
(96, 7)
```

Rows: 96
Columns: 7

**Find the columns data type**

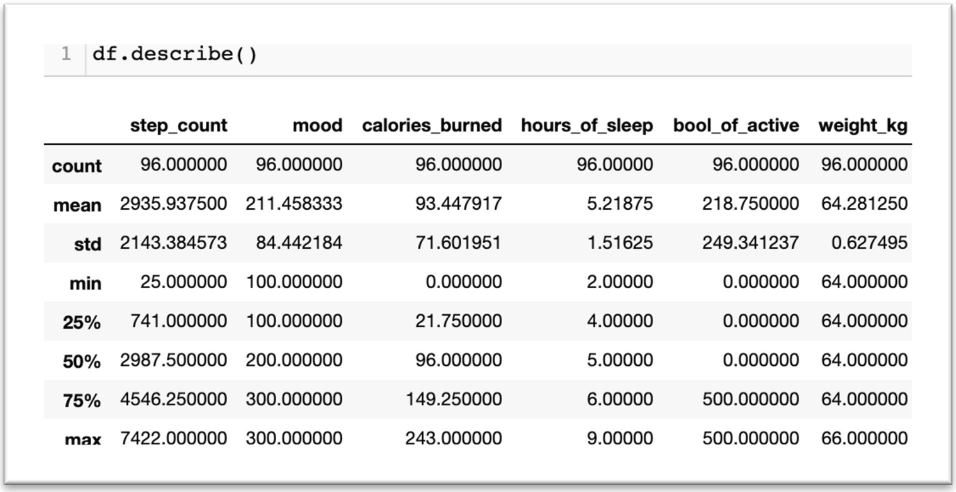


```
#Info about each column
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 96 entries, 0 to 95
Data columns (total 7 columns):
 #   Column           Non-Null Count  Dtype
---  ------           --------------  -----
 0   date             96 non-null     object
 1   step_count       96 non-null     int64
 2   mood             96 non-null     int64
 3   calories_burned  96 non-null     int64
 4   hours_of_sleep   96 non-null     int64
 5   bool_of_active   96 non-null     int64
 6   weight_kg        96 non-null     int64
dtypes: int64(6), object(1)
memory usage: 5.4+ KB
```

All the columns are integer except date is object.

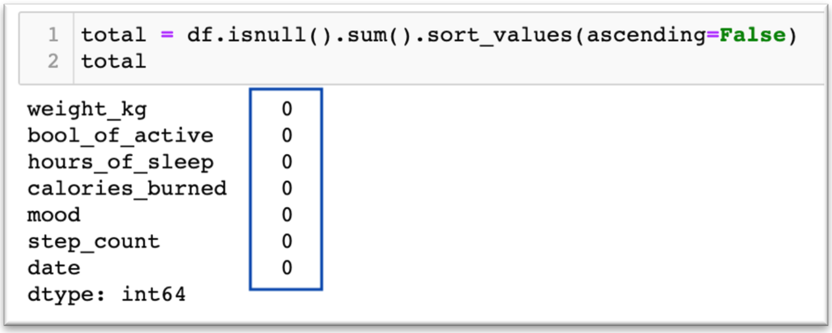
**Statistic description for numerical columns**



```
df.describe()
```

| | step_count | mood | calories_burned | hours_of_sleep | bool_of_active | weight_kg |
|---|---|---|---|---|---|---|
| **count** | 96.000000 | 96.000000 | 96.000000 | 96.00000 | 96.000000 | 96.000000 |
| **mean** | 2935.937500 | 211.458333 | 93.447917 | 5.21875 | 218.750000 | 64.281250 |
| **std** | 2143.384573 | 84.442184 | 71.601951 | 1.51625 | 249.341237 | 0.627495 |
| **min** | 25.000000 | 100.000000 | 0.000000 | 2.00000 | 0.000000 | 64.000000 |
| **25%** | 741.000000 | 100.000000 | 21.750000 | 4.00000 | 0.000000 | 64.000000 |
| **50%** | 2987.500000 | 200.000000 | 96.000000 | 5.00000 | 0.000000 | 64.000000 |
| **75%** | 4546.250000 | 300.000000 | 149.250000 | 6.00000 | 500.000000 | 64.000000 |
| **max** | 7422.000000 | 300.000000 | 243.000000 | 9.00000 | 500.000000 | 66.000000 |

**Checking if there are missing values**

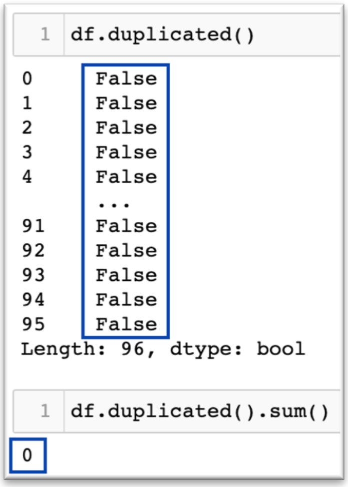


```
total = df.isnull().sum().sort_values(ascending=False)
total
```

```
weight_kg         0
bool_of_active    0
hours_of_sleep    0
calories_burned   0
mood              0
step_count        0
date              0
dtype: int64
```

There are no mussing values.

**Find if theres a duplication rows**



```
df.duplicated()
```

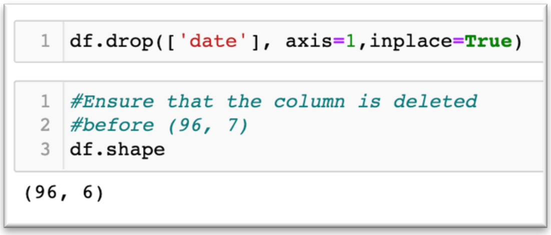
```
0     False
1     False
2     False
3     False
4     False
      ...
91    False
92    False
93    False
94    False
95    False
Length: 96, dtype: bool
```

```
df.duplicated().sum()
```

```
0
```

There are no duplication rows.

**Delete Date column**



```
df.drop(['date'], axis=1,inplace=True)
```

```
#Ensure that the column is deleted
#before (96, 7)
df.shape
```

```
(96, 6)
```

### Is it classification? clustering? dimension reduction?

The dataset is Classification, since the Feeling of activeness was measured in either "Active" or "Inactive" which were given numeric values of 500 and 0 respectively.

## Which ML methods work well with my dataset? Why?

- **Logistic Regression.**

  It used to **predict** the probability of a categorical dependent variable.
  We choose logistic regression Because we have a Categorical dependent variable that splits our data set into **active** or **unactive** based on **0** or **500.**

  The **dependent** variable "**response**" is "**Bool of active"** which is contain categorical value in our data set

  - **500** represent active.
  - **0** represent unactive.

  The **independent** variable that we choice to predict is "**mood":**

  - **300 - Happy**
  - **200 - Neutral**
  - **100 - Sad**

  Since there is a relation between them "if you are **happy**, you may be **active**".

- **Support Vector Machine (SVM).**

  Support Vector Machine (SVM) is a very popular Machine Learning algorithm that is used in both Regression and Classification. Support Vector Machine is similar to Linear Regression in that the equation of the line.

  We choose the algorithm support Victor machine because it's a linear model for classification and regression problem it can be linear or noun liner, which is could be helpful for us.
  The simple idea of support Victor machine is to create line or a hyper plan which is separate the data into classes either **active** or **inactive**.

  The **dependent** variable "**response**" is "**Bool of active**"
  The **independent** variable that we choice to predict is "**mood**":