

**IT 461 Practical Machine Learning**

**Fall 2020**

**Project Part II**

**Group Member Names:**

|  |  |
| --- | --- |
| **Student Name:** | **ID:** |
| **Shaikha Bin Ateeq** | **438201519** |
| **Alanoud Alotaibi** | **437200539** |

**Main Topic of Problem:**

**The Effect of Exercises on People’s Activeness**

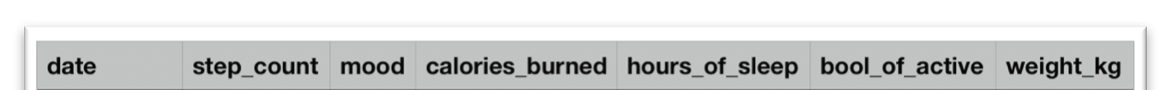
1. Introduction
2. Introducing the topic and why is it useful to use ML algorithms on it.

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. The main topic is about The Effect of Exercises on People’s Activeness.

which is mean when my number of steps that’s take in a day is high and Burned calories in a day also high, and my mood is happy also I sleep very well, does these features effect my Activeness? Shall say I am active person?

It useful to use ML algorithms on it because we want to know if this features effect my Activeness or not , the best solution in using ML algorithms on it receive and analyse input data to predict output values within an acceptable range. As new data is fed to these algorithms, they learn and optimise their operations to improve performance also since it can give a good accuracy if my assumptions is right or not

**We took this dataset from Kaggle site** The dataset has **7** Attributes (columns) and **96** observation (rows).



1. What is the problem you want to solve and what are you proposing to solve it with?

Our problem is about {"Does exercise/working-out improve a person’s activeness?”}.

The purpose of the project we want to measured the Feeling of activeness either "Active" or "Inactive" which were given numeric values of 500 and 0.

And That It's based on the features (predict ) in our data set to find )response( which is Bool\_of\_active

We proposing to solve by using Classification problem , since the Feeling of activeness was measured in either "Active" or "Inactive" therefore we chose two related algorithms to the Classification problem which is Logistic Regression And Support Vector Machine (SVM)

1. Data Preprocessing
2. Show what preprocessing techniques you will be using
3. Checking if there are missing values
4. Find if there’s a duplication row
5. Delete Date column
6. normalization all attributes integer
7. Why did you choose to apply these techniques?

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).since It’s important to understand these [different types of missing data](https://en.wikipedia.org/wiki/Missing_data) from a statistics point of view. The type of missing data will influence how you deal with filling in the missing values.

Normalization avoids these problems by creating new values that maintain the general distribution and ratios in the source data,

No need for normalization ,all attributes integer after delete “Date” since it object

1. Provide examples of preprocessing steps

|  |
| --- |
| **Find how many rows and columns in the dataset** |
| Graphical user interface, text, application  Description automatically generated |
| Rows: 96  Columns: 7 |
| **Find the columns data type** |
| Text  Description automatically generated |
| All the columns are integer except date is object. |
| **Statistic description for numerical columns** |
| Table  Description automatically generated |
| **Checking if there are missing values** |
| **A picture containing table  Description automatically generated** |
| There are no mussing values. |
| **Find if theres a duplication rows** |
| **Table  Description automatically generated** |
| There are no duplication rows. |
| **Delete Date column** |
| Graphical user interface, text, application  Description automatically generated |
| التحويل من ٥٠٠ ال ١ |
|  |

1. Methodology
2. Explain the ML algorithm used

* **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 }**

* **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”** 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 }**

1. Implementation (with screenshots)

**The code of splitting data:**

**A picture containing scatter chart

Description automatically generated**

**Logistic Regression code:**

**Text

Description automatically generated**

**Support Vector Machine (SVM) code:**

Text

Description automatically generated

|  |  |  |  |
| --- | --- | --- | --- |
| Algorithm/Split | Test=20 & train=80 | Test=45& train=55 | Test=80 & train=20 |
| * **Logistic Regression.** | Accuracy = 0.65 | Accuracy= 0.61 | Accuracy= 0.63 |
| Test Error=0.35  Train Error = 0.36 | Test Error=0.39  Train Error = 0.33 | Test Error=0.36  Train Error = 0.31 |
| * **Support Vector Machine (SVM).** | Accuracy=0.65 | Accuracy= 0.64 | Accuracy 0.66 |
| Test Error=0.35  Train Error = 0.34 | Test Error=0.36  Train Error = 0.33 | Test Error=0.33  Train Error = 0.39 |

We try multiple splitting for Test and train, the best one was when the Test=45& train= 55 since the accuracy of both models are high also there is no overfitting because of the test error is higher than training error

For other splitting, test error is less than the training error, and this means there is a underfitting {sampling bias }in our test

* **Logistic Regression** screenshots of accuracy and error Test=45& train= 55:

Text

Description automatically generated

* **Support Vector Machine (SVM)** screenshots of accuracy and error Test=45& train= 55:

Graphical user interface, text, application, email

Description automatically generated

1. What are the initial parameters of the model if any?

**Logistic Regression:**

Graphical user interface, text, application

Description automatically generated

The intercept (b0): (0.01036201)

Coefficient (b1) =(-2.11189379)

1. Evaluation and Results
2. Show your results (tables, figure, confusion matrices …etc)

|  |  |  |
| --- | --- | --- |
| Results/ Algorithm | **Logistic Regression.** | **Support Vector Machine (SVM).** |
| tables | Table  Description automatically generated | Table  Description automatically generated |
| figure | Chart, line chart  Description automatically generated | Chart, line chart  Description automatically generated |
| confusion matrices | Graphical user interface, chart, application  Description automatically generated | A picture containing chart  Description automatically generated |

1. Plot your output (visualization of error)

|  |  |  |
| --- | --- | --- |
|  | **Logistic Regression.** | **Support Vector Machine (SVM).** |
| **visualization of error)** |  |  |

1. Which evaluation techniques is better and why?

The techniques which is better is Support Vector Machine (SVM), Since we got the highest Accuracy= 0.64 without the vulnerability of overfitting or underfitting

to improve your models :

in phase 1 we didn’t select the independent value correctly since we say that we have only one independent variable which is “mood”, we thought that the mood can give us the felling of activeness for the person

But that assumption was not affected as much we think

To improve our result is by change the selection of the independent variable to {step\_count, mood, calories\_burned, hours\_of\_sleep}

Next time we will ensure that we have data set with enough row

Since our data set small, we have challenges working with it.

maybe when we choose step and calorie that will be better analysis because there are good correlation between them

Chart

Description automatically generated

1. Analysis and conclusion
2. Explain your results in analytical way instead of numbers. Meaning, why do you think one algorithm performed that way and the other didn’t?

The area covered by the curve is the area between the orange line (ROC) and the axis. This area covered is AUC. The bigger the area covered, the better the machine learning models is at distinguishing the given classes. The Support Vector Machine (SVM), have a bigger area under the curve more than Logistic Regression also svm auc= .72 / log auc= .59

Chart, line chart

Description automatically generated

precision it measures how likely the prediction of the positive class is correct.and since the support vector machine (svm) have a high precision which is for unactive =1 active =0,48 and that more than logistic regression. also recall gives how good the model is to recognize a positive class. Which is for unactive =0.45 active =1

Table

Description automatically generated

1. References (if Any)

Docs.microsoft.com. 2020. *ML Studio (Classic): Normalize Data - Azure*. [online] Available at: <https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/normalize-data> [Accessed 26 November 2020].

Sullivan, J., 2020. *Data Cleaning With Python And Pandas: Detecting Missing Values*. [online] Medium. Available at: <https://towardsdatascience.com/data-cleaning-with-python-and-pandas-detecting-missing-values-3e9c6ebcf78b> [Accessed 26 November 2020].