****

Subject:- Machine Learning

**Weekly Report Submission 4**

GROUP: **DATA DYNAMOS**

Submitted to faculty: Prof. Mehul Raval

Date of Submission: 18/03/2023

**Student Details**

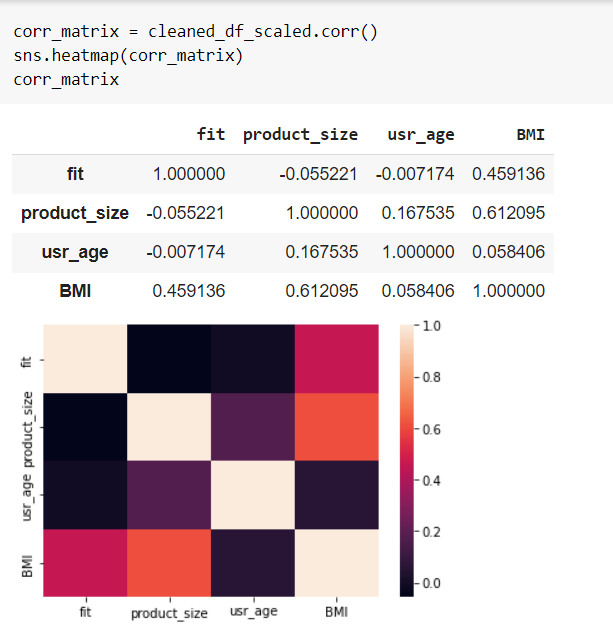
| Roll No. | Name of the Student | Name of the Program |
| --- | --- | --- |
| AU2040122 | Aditi Vasa | Btech CSE |
| AU2040002 | Shrey Somani | Btech CSE |
| AU2040196 | Vandan Shah | Btech CSE |
| AU2040048 | Ronit Shah | Btech CSE |

(Winter Semester)

**Report**

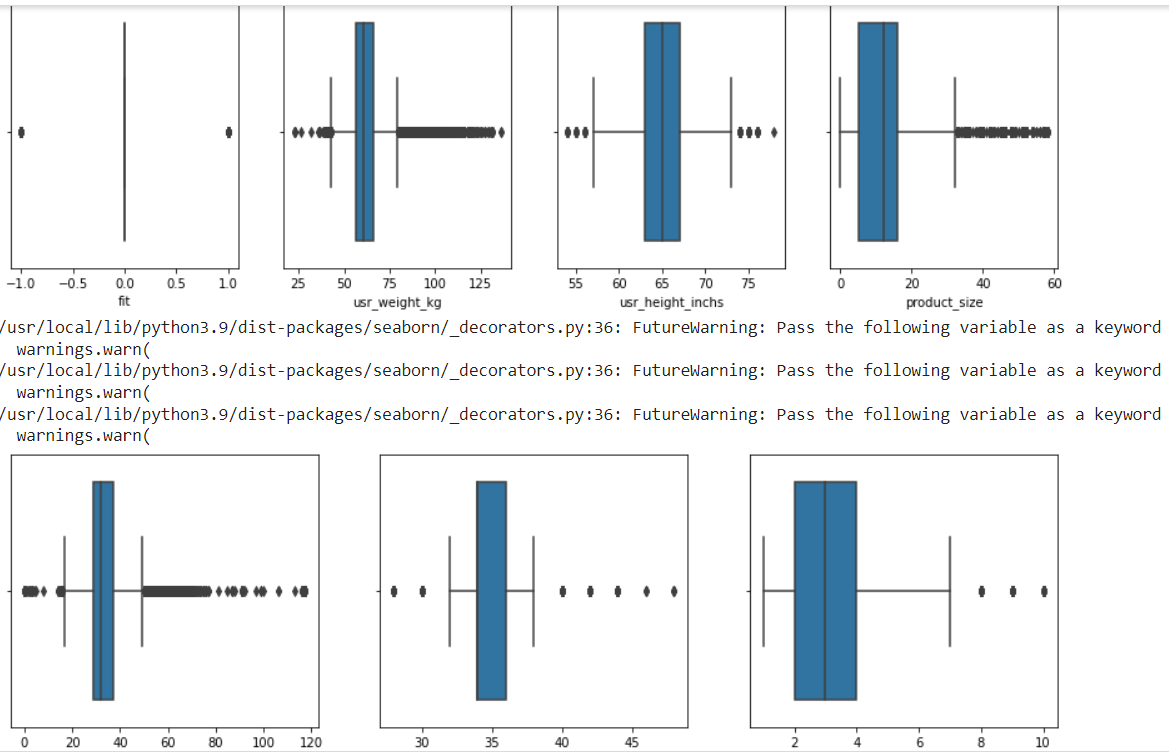
**Correlation matrix:**

Implemented correlation matrix to find correlation between various variables and we have done BMI as body weight, user height.

****

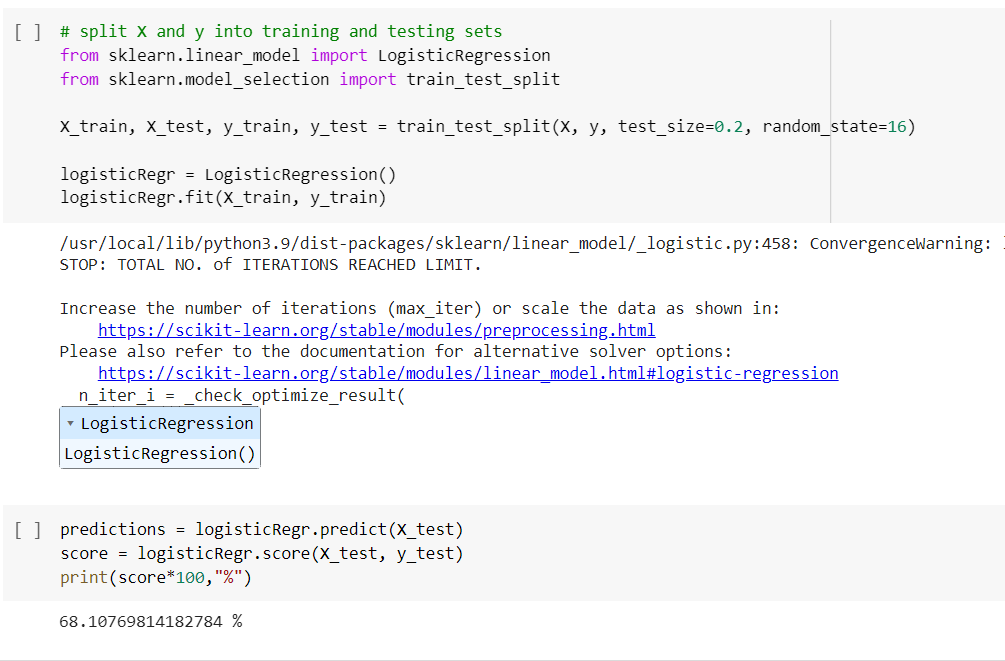
**Box and whisker plot:**

Implemented box and whisker plot to find out outliers and found that the variable with more outlier are product size, user weight, user age and try to reduce number of outliers.

****

**Started with Logistic Regression:**

Implemented logistic regression algorithm to find out accuracy of our model and train our model to get the accuracy. And we get an accuracy of around 68%.

****

**Summary:**

After performing data cleaning, data processing and encoding, we performed Exploratory Data Analysis. The next step was to plot the correlation matrix to understand which variables are strongly related and are affecting the other variables. It was observed that BMI and product size had a comparatively strong correlation. We also tried to plot Box and Whisker’s plots to detect the outliers in each variable. Lastly, using all the essential information we applied Variance Threshold to remove all the values below the threshold. Finally, we performed logistic regression on 9 columns and achieved an accuracy of 68%.

**Plan for Next Week:**

* We aim to normalize the values of outliers found after plotting the box and whisker’s plot.
* In the following week, we plan to perform further EDA to understand the interdependency of variables on each other using plots like line charts and histograms
* We also aim to understand which model will perform better on our data, increase our accuracy and gives desired output. Presently, we have just performed Logistic Regression on a trial and error basis to understand the accuracy. However, the plan now is to understand the data and use a justified model
* Other than that, since we have a multiclass output, we aim to use Quadratic Decision boundary.