

**Health plan predictor for people with Overweight**

**Introduction**

Based on one of the biggest health issues of our country that keeps growing that affects all the different sector of our society, overweight, and some of the most recent tendencies of society of trying to be on their ideal health parameter. The project is going to implement clusters and linear regression to create groups depending on their personal metrics and a real data sets and develop a new diet, to help them to rich a healthier lifestyle.

**K-Means**

It is part of unsupervised learning algorithms. The goal of K-means is clustering. It is part of hard clustering, that means that every data only belongs to one cluster only.

**Clusters**

It is used to group samples or events, based on their proximity to each other. Clusters have been used in a wide variety of applications. Clustering doesn’t consider human bias and finds its own patterns.

**Linear Regression**

It is used to model the relationship between to variables fitting a linear equation. Before fitting a model, they should determine the relationship between the variables of interest.

The most common form of a linear equation is the model of

In which Y is the dependent variable and “X” the explanatory one, the slope is given by “b” and “a” is the value of “Y” when its value is 0.

**Data set Used**

The patient data set was built and provided by Nutritionist Maria Lujambio, based on the different NOM’s stablished by the “Secretaría de Salud” like the NOM 043, NOM 015 and some organizations like the American Dietetic Association and the Mexican system of equivalents. Using this data set the objective is to create Cluster based on their age and BMI, this groups help the Nutritionist and the patient to see where they are standing on the groups and based on that implementing linear regression the diet plan of portions is going to be created and displayed in a graph.

**How the diet is calculated**

To develop a proper diet plan in the code we had to implement the Harris-Bennedict formula that allows us to know more precisely the number of calories that each metabolism requires in relation to your gender, weight, age, height and physical activity. First it calculates your minimum energy required for your body with the next equations depending on your gender.

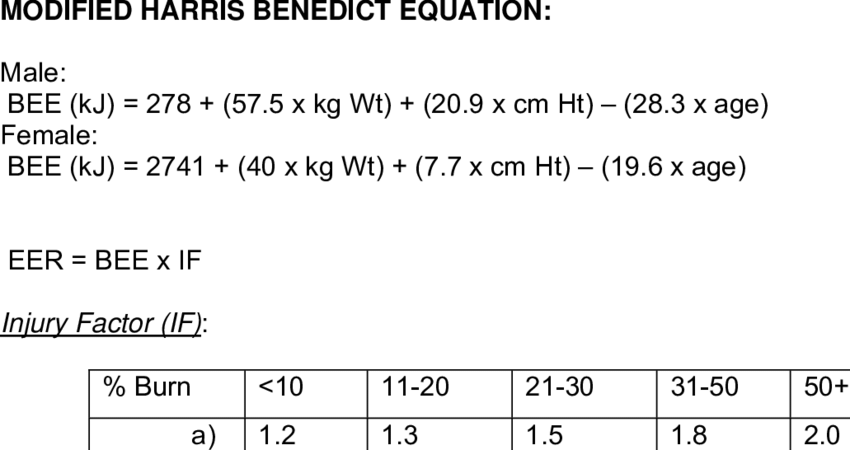


Figure 1. Harris-Benedict equation for Energy Consumption

Then we add the energy your body consumes depending on your physical activity and the energy your body consumes when you eat. With that we obtain your Total Energy Consumption. The next thing is depending of your BMI if we need to rest 500 kcals of your Total Energy Consumption or keep them the same. The reason of why diminishing 500 kcal from your Total Energy Consumption is to lose around 0.5 kg that is the amount recommended when losing weight that will not cause any health problem.

**How the Code works**

This code is meant to be used in conjunction with a Nutritionist for some details at the end.

At the begging it is going to ask you for your Age, weight in kg, height in meters, gender, and your physical activity, once that data is provided it calculates your BMI and your total energy consumption depending on your gender using an if. Then the cluster starts running, comparing BMI and age, it displays the data set without clusters and the one of clusters with your information displayed in a different color.

After that it will display in which group you belong and what does that mean. The next thing the code runs the linear regression and displays your new diet based on the total energy consumption that you should have depending on what cluster you belong to. This is going to be done by linear regression, for it to work we took 30% of our data to train the model and 70% for test.

**Results**

First, we tried to create the clusters comparing the weight and height like the BMI charts to see what the clusters showed and how it related the data. We saw that the clusters didn’t show anything that made sense apparently and everything was overlapped. So, the next thing that we tried was implementing DBSCAN but again we had a result that didn’t made sense. So, we changed the data that we were relating and now it made sense and it plot the graph that it’s shown in Figure 3. We chose K-means instead of DBSCAN because DBSCAN didn´t create any cluster and the silhouette shown by k-means was of 0.330, that even though it’s near to 0 it’s an acceptable range since the database has a lot of values that are really close between each other but it’s because of the parameters of their age, height and weight.

Gráfico, Gráfico de dispersión

Descripción generada automáticamente

Figure 2. Graph of dataset

Gráfico, Gráfico de dispersión

Descripción generada automáticamente

Figure 3. Graph of dataset K-means

Gráfico, Gráfico de dispersión

Descripción generada automáticamente

Figure 4. Graph of dataset DBSCAN

Then for the linear regression we obtained the next information for our regression. We can see that in al of the regressions we had good percentage of accuracy in our models. That can be consulted in the Table 1.

**Conclusion**

From the different things that were tried to test the code, we saw that if we tried to implement a dataset that doesn’t makes sense and has random values for our linear regression, it’s not able to calculate anything and shows an error. For the clusters it’s very important that when creating then and when choosing what data to take into account for them we need to have clear what’s the relation between them, but we don’t have to know how it’s going to be grouped.

**References**

Equation of Harris-Benedict. Published on:

<https://www.biobiochile.cl/noticias/2014/04/05/ecuacion-de-harris-benedict-un-facil-metodo-para-calcular-las-calorias-que-necesitas.shtml>

Linear Regression. Published on:

<http://www.stat.yale.edu/Courses/1997-98/101/linreg.htm>

A walkthrough on A.I Clustering — understanding user behavior across a website. Published on:

<https://becominghuman.ai/a-walkthrough-on-a-i-clustering-understanding-user-behaviour-across-a-website-ce2ac0f48699>

Clustering: What it is? When to use it? Published on:

<https://towardsai.net/p/machine-learning/clustering-what-it-is-when-to-use-it>

**Tables**

|  |  |  |  |
| --- | --- | --- | --- |
| Regression | Coefficients | Mean Squared error | Coefficient of determination |
| Verduras | [ 2.26e-04 -1.11e-03 7.69e-02 -1.22e+00 3.02e-03 -1.21e-03 1.06e-03 3.87e-03] | 0.08 | 0.96 |
| Frutas | [-4.26e-03 9.87e-03 7.52e-01 -9.08e-01  2.07e-01 -1.74e-03 5.05e-04 1.89e-03] | 0.09 | 0.87 |
| Cereales | [-9.19e-04 5.14e-03 4.51e-01 6.36e-01 1.99e-01 -7.96e-06 -3.16e-04 1.45e-03] | 0.10 | 0.74 |
| Proteinas | [-1.52e-03 1.57e-03 -1.57e-01 -3.60e-01 -6.34e-03 -3.52e-04 1.95e-04 2.53e-03] | 0.07 | 0.93 |
| Grasas | [-4.26e-03 9.87e-03 7.52e-01 -9.08e-01 2.07e-01 -1.74e-03 5.05e-04 1.89e-03] | 0.09 | 0.87 |

Table 1. Table of coefficients and accuracy of regressions