TDP OF ADULTS ARE ASSOCIATED WITH DIET QUALITY

DS201: FINAL PROJECT REPORT

Submitted By

Team2

Chirag (Student ID: 12140520), Rahul (Student ID: 12141300), Anant Yadav (Student ID: 12140190)

Course Instructor: Dr. Nitin Khanna

Department of Electrical Engineering and Computer Science

Indian Institute of Tecchnology Bhilai

December 7, 2022

Contents

Keywords Temporal Dietary Patterns · Healthy Eating Index · Dietary Intake · Diet Quantity · Dietary Patterns

1 Introduction

This paper describes the relation between Temporal Dietary Patterns and Diet Quality of clusters derived from US adults having similar eating habits. Temporal dietary patterns is an emerging topic for doing research in health of individuals. It incorporates multiple dietary characteristics, such as foods or nutrients, in analysis, and is often explored in relation to a particular health or dietary outcome. This project is based on applying this method on US non-pregnant adults of age 20-65 years and the data is taken from NHANES datasets. The data used here is from 2007 to 2018.

Clusters were derived from the cleaned data using various clustering algorithms. Diet Quality is measured in terms of HEI-2015 index derived from 5 quantities. The idea is to observe if there is any relation between the diet quality of clusters derived from people having similar dietary patterns.

Dietary pattern observation isn't new. It is based on various factors like quantity of nutrient intake and frequency of intake. Adding time to this quantity makes it Temporal Dietary Pattern. This means that here, the time of eating is also taken into consideration. Previous researches by other researchers also observed that TDP is better than normal Dietary Patterns and can be used to analyse various factors like obesity, diet quality etcetra.

2 Materials & Methods

NHANES dataset is a free open-source data provided by US government to carry out various studies on the US citizens. The data used here is from year 2007 to 2018. It had to be cleaned first and for that, various cleaning methods were used ranging from removing the rows having NULL values to dropping the unnecessary columns and combining two or more different datasets. After doing all this, 45,000 rows were left in the data and then the rest of the process was done.

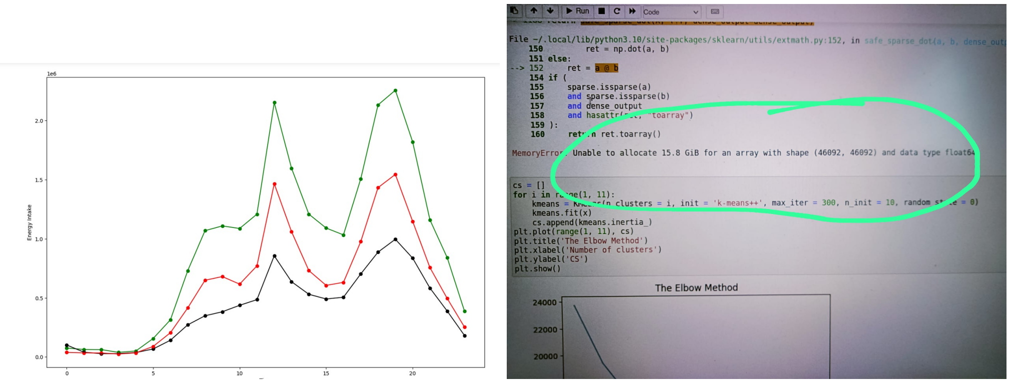
Clustering was done using K-means clustering with Euclidean Distance as the Distance Metric. Elbow shaped graph was derived and from there, number of clusters that were chosen, was 3. As for the diet quality analysis, HEI index was chosen. HEI is updated every 5 years and number of components that it depends upon, are changed. HEI-2015 index based on 5 different components was chosen and using multiple linear regression, it was derived. Sklearn library is used for K-Means clustering. Another library used is tslearn.

3 Results

KMeans clustering with Euclidean Distance as the distance metric is a potential method for making clusters. The problem with Euclidean distance is that if the vectors, that it is comparing, are of similar size but in different dimensions, the distance would be abruptly high. This is not what is required in this problem. Reason is that, here the dietary patterns are associated with time stamps and the requirement is to compare the ones that are of similar dietary pattern (not the exact ones). So, here, Euclidean distance fails and the clusters that are obtained are of no use.

Thus, what's needed as the distance metric is the Dynamic Time Wrapping as it takes the input in the form of a vector having its components in the form of a tuple having timestamp and the value of that component at that instance. This distance metric is potentially useful for comparing time-series data. Also, KMeans clustering needs to be changed into Kernal KMeans clustering as KMeans clustering couldn't make the desired clusters. What was intended, was that the individuals having similar Temporal Dietary Patterns will be grouped under same cluster. But KMeans couldn't give us that.

Kernal KMeans clustering is potentially a very having algorithm. The outcome of that was that it exceeded the computation capacity of the local computer and killed the kernal of Jupyter Notebook. Then it was tested on google colab but to our surprise, google colab denied to run it. It took a lot of time and didn't give any outcome then also. So, the only option we're left with was to use the KMeans clustering method. Here comes the next obstacle, the desired outcome needs to be using the frequency of eating of each individual and to plot it for highest and second-highest energy consumed during the 24 hours. But, to our surprise, what we achieved was that the clusters were made on the basis of energy consumed and not on the time of eating. This raised an issue that even if we would separate out the highest and second highest energy consumption for all the individuals of each clusters and plot them on a graph, what we would get is that all the graphs have similar time of eating. This is contradicting to what was achieved in the reference paper what we took as a reference.



```
File ~/.local/lib/python3.10/site-packages/sklearn/utils/extmath.py:152, in safe_sparse_dot(a, b, dense_outp
    150         ret = np.dot(a, b)
    151 else:
--> 152     ret = a @ b
    154 if (
    155     sparse.issparse(a)
    156     and sparse.issparse(b)
    157     and dense_output
    158     and hasattr(ret, "toarray")
    159 ):
    160     return ret.toarray()

MemoryError: Unable to allocate 15.8 GiB for an array with shape (46092, 46092) and data type float64
```

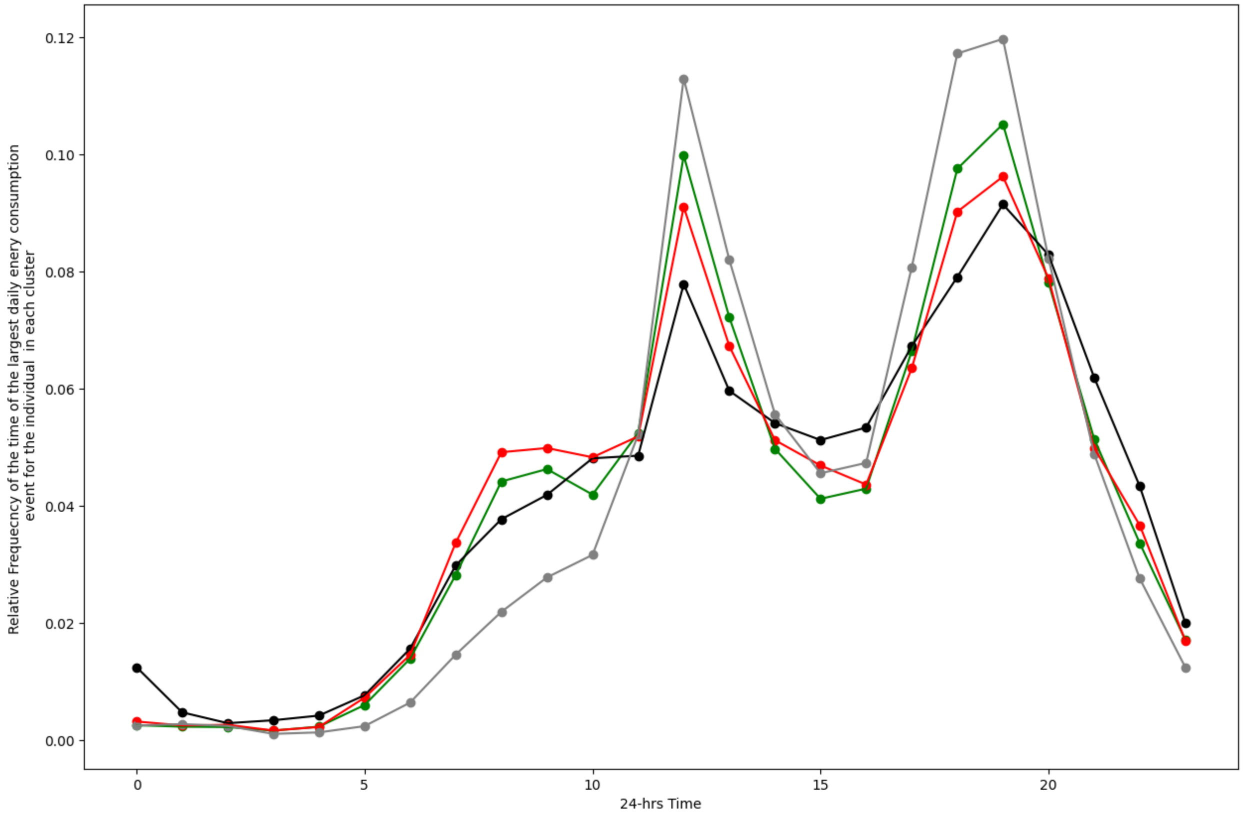
```
cs = []
for i in range(1, 11):
    kmeans = KMeans(n_clusters = i, init = 'k-means++', max_iter = 300, n_init = 10, random_state = 0)
    kmeans.fit(x)
    cs.append(kmeans.inertia_)
plt.plot(range(1, 11), cs)
plt.title('The Elbow Method')
plt.xlabel('Number of clusters')
plt.ylabel('CS')
plt.show()
```

(a) Clusters formed by KMeans Clustering

(b) Memory Error

4 Discussion

The Temporal Dietary Patterns that we were able to derive from the US Adults of 20-65 years of age, were exhibiting the similar dietary patterns. This led us to the problem of relating them with diet quality derived using HEI-2015 index as all of them were of similar dietary patterns which wasn't what was intended. The HEI-2015 index is based on 5 components that were combined together to quantify the diet quality of each individual. The values of HEI-2015 index for the population data can be found inside the code.



The given image is the Relative Frequency of the time of the largest daily energy consumption event for the individual in each cluster. As we can clearly see, the clusters are not different in terms of their shapes. This signifies that using KMeans clustering wasn't a good choice.

HEI-2015, according to international norms, has 12 components based on different vegetables etcetra. But the data we had is having dietary intake in the form of fats, vitamins etcetra. So, we had to convert the 12 component HEI-2015 to 5 component quantity.

Conclusions

What we can conclude from this project is that KMeans clustering method couldn't perform clustering in the desired way because of which, the relation between TDP and diet quality couldn't be figured out. Had we had powerful resources like servers or workstations, the Kernal KMeans clustering would be applied with Dynamic Time Wrapping as the distance metric and it is expected to give us the required results.

Acknowledgements

Dr. Nitin Khanna has given us this wonderful project because of which, we could explore so much about temporal dietary patterns, HEI indices, clustering algorithms etc. We express our great gratitute towards him for making us aware about this project work.