Toronto Metropolitan University

CPS803 Assignment 4 Report

Machine Learning

William Masoen | 500971193
Professor Elodie Lugez
November 27th, 2024

Background

For this assignment, I will be using the *Estimation of Obesity Levels Based On Eating Habits and Physical Condition* dataset from UC Irvine Machine Learning Repository. The dataset categorizes instances into 7 categories; normal weight, overweight 1, overweight 2, obese 1, obese 2, obese 3, and insufficient weight. This particular dataset is interesting to me because of its size and potential to provide insights into the factors affecting obesity levels.

The code I implemented in script_WilliamMasoen.py consists of three major sections. The first section is responsible for fetching the dataset from ucimlrepo and loading it into the variable "data" with pandas.

The second section is where data preprocessing happens. There are 4 manipulation methods that I used; data cleaning, encoding categorical values, feature scaling, and feature reduction with PCA.

The third and last section holds the clustering algorithm. I am using the DBSCAN algorithm for my dataset as it works best with it.

Methods

Data Manipulation

This part of the report will go in depth about the manipulation techniques used in the code. I used 4 manipulation methods; data cleaning, encoding categorical values, feature scaling, and feature reduction with PCA. These manipulation techniques are used to transform raw data to make it more suitable for analysis and modelling.

1. **Data cleaning**

   Upon verifying for missing and unusual values, the dataset seems to not have any. With that said, the data cleaning process is skipped as noise and outliers will be handled in the clustering algorithm itself.

   The verification process takes place by printing out the number of missing values and all distinct values in each column. The code shows 0 missing values for all columns. In addition, after carefully analyzing the distinct values, there seems to be no unusual values that are out of ordinary. So we can conclude that the dataset is clean and does not require a data cleaning process.

2. **Encoding categorical values**

   This part of the manipulation process happens to transform categorical values into numerical formats suitable for algorithms. This step is necessary because it makes raw data compatible with the clustering algorithm. One example of this is the Gender feature in the dataset. The raw data will have values such as Male or Female, but after encoding it to boolean values, the column name changed to Gender_Male and values would be True or False, where False indicates gender is female, and True for being a male. (Shown on the next page)

Gender
Female
Female
Male
Male
Male
Male

Encode categorical values →

Gender_Male
False
False
True
True
True
True

3. Feature scaling

For this part of data manipulation techniques, I used the standard scaler approach. This method standardized features to make sure that they are on a similar scale, making it more uniform.

Age	Height	Weight
21	1.62	64
21	1.52	56
23	1.8	77
27	1.8	87
22	1.78	89.8
29	1.62	53
23	1.5	55
22	1.64	53
24	1.78	64
22	1.72	68
26	1.85	105

Standard scaling →

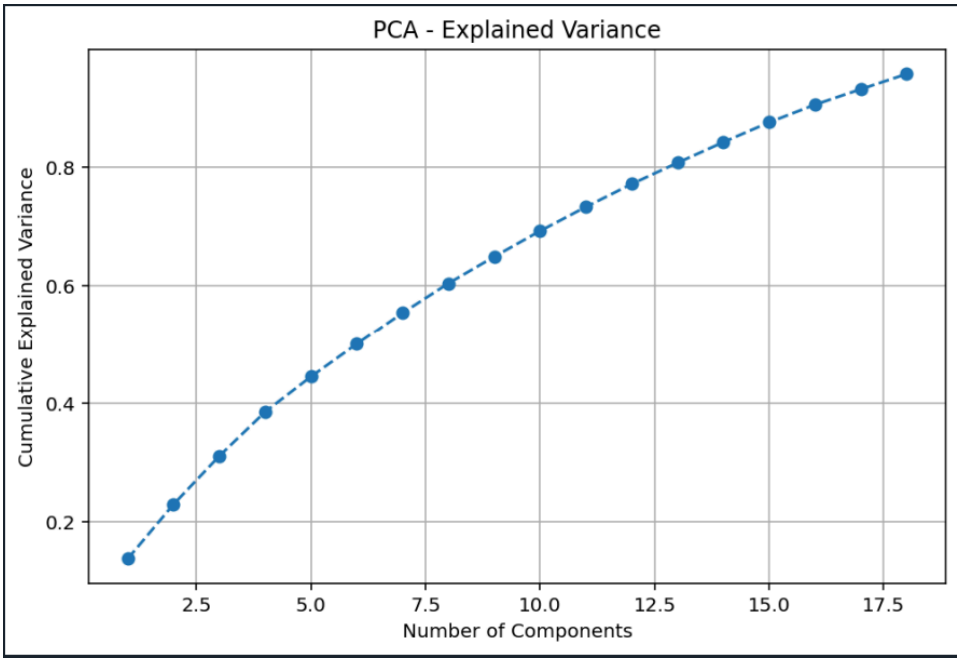
Age	Height	Weight
−0.522124	−0.875589	−0.862558
−0.522124	−1.9476	−1.16808
−0.206889	1.05403	−0.36609
0.423582	1.05403	0.0158084
−0.364507	0.839627	0.12274
0.738817	−0.875589	−1.28265
−0.206889	−2.162	−1.20627
−0.364507	−0.661187	−1.28265
−0.0492713	0.839627	−0.862558
−0.364507	0.196421	−0.709799
0.265964	1.59003	0.703226

4. Feature reduction with PCA

The feature reduction technique is used to reduce the dimensions of the data using PCA (Principal Component Analysis). The PCA technique transforms high dimensional data into lower dimensional data, while retaining as much variance as possible [3]. So the code will analyze the features and remove those that do not provide a significant variance. The dataset after encoded and scaled started off with a shape of (2111, 23), which means it has 2111 instances, and 23 columns. After the PCA technique was applied, the data has a shape of (2111, and 18), which means 5 columns have been successfully reduced while retaining the same number of instances.

```
Shape after encoding: (2111, 23)
Shape after scaling: (2111, 23)
Number of components that explain 95% of the variance: 18
Shape after PCA: (2111, 18)
```

In addition, I used Explained Variance to provide insights into how much information (variance) is retained by each component for reduction. In the code, I specified n_components = 0.95, which means I want to retain 95% of the variance from the dataset. The chart below shows which number of components meet that criteria, which is 18 components. That is why the PCA reduction only reduces to 18 components. (Shown on the next page)



Clustering Process and Results

The technique that I used to solve the practical clustering problem is DBSCAN as it suits my dataset the best. As my dataset contains some outliers and noise, DBSCAN can identify it and exclude it from belonging to any clusters. On top of that, my data contains more than 2000 instances which makes it hard for other methods such as K-Means and Hierarchical to solve. Before using DBSCAN, I have tested with K-Means and Hierarchical. However, upon analyzing the results, they seem to perform poorly. Both K-Means and Hierarchical were only producing an ARI and NMI score of 0.12 and 0.39 respectively.

The ARI is "statistical measure used in data clustering analysis. It quantifies the similarity between two partitions of a dataset by comparing the assignments of data points to clusters" [1]. In other words, this scoring metric will show how close your clustering results are to the true labels. With a score ranging from -1 to 1, where anything less than 0 means your model is performing worse than random clustering, and anything above 0 is considered better than random clustering. My clustering algorithm which used DBSCAN was able to output a score of 0.855 on the ARI scale.

```
Adjusted Rand Index (ARI, excluding noise): 0.8549664153957088
```

The NMI is a "metric for assessing the performance of community detection algorithms. The NMI facilitates comparisons between two clusters or communities, yielding a value that ranges from 0 to 1" [2]. In this case, if values are closer to 0, it means that the clustering is unrelated to the true labels. Whereas values closer to 1 means better agreement between the predicted clusters and the true label. My clustering algorithm was able to output a score of 0.752 on the NMI scale.

```
Normalized Mutual Information (NMI, excluding noise): 0.7521169008323662
```

In the code, I also specified the number of eps and min_samples of 0.7 and 6 respectively. The eps variable represents the radius of the cluster. For example, any instance within 0.7 units from the centroid will belong to that cluster. In regards to min_samples, this represents the minimum number of points within the eps radius to form a dense region. Which means that only a centroid with minimum 6 points within its radius is going to be considered a core point. After testing multiple values, I concluded that 0.7 and 6 are the values that produce the highest ARI and NMI score.

Conclusion

To conclude, this assignment explored clustering techniques to identify patterns within the Estimation of *Obesity Levels Based on Eating Habits and Physical Condition* dataset. After preprocessing the data with encoding, scaling, and reduction with PCA, I applied DBSCAN as the clustering method as it can handle noise and outliers. The results show that it performs significantly better than K-Means and Hierarchical methods, producing high ARI and NMI scores that show accurate clustering compared with the true labels.

Dataset

Estimation of Obesity Levels Based On Eating Habits and Physical Condition [Dataset]. (2019). UCI Machine Learning Repository. https://doi.org/10.24432/C5H31Z

Reference

[1] Yang, Y. (2017). *Adjusted rand index*. Adjusted Rand Index - an overview | ScienceDirect Topics. https://www.sciencedirect.com/topics/computer-science/adjusted-rand-index

[2] Mahmoudi, A., & Jemielniak, D. (2024). Proof of biased behavior of normalized mutual information. *Scientific Reports*, *14*(1). https://doi.org/10.1038/s41598-024-59073-9

[3] Jaadi, Z. (2024). *Principal Component Analysis (PCA) explained*. Principal Component Analysis (PCA): A Step-by-Step Explanation. https://builtin.com/data-science/step-step-explanation-principal-component-analysis