STEM Prasetiya Mulya

Clustering Analysis using K-Means, Gaussian, and BIRCH: Prasetiya Mulya Student Health and Wellness

Felicita | Felicia Austin | Jonathan Evan | Tri Yohana
Group 4

K-Means is the Best Model for Our Data Based on Three Evaluation Metrics We Used

Part I: Data Processing

Collected Data → Data Engineering → Exploratory Data Analysis → PCA Analysis

Part II: Modeling

Clustering Models

- K-Means
- Gaussian Mixture Model
- BIRCH

Part III: Evaluating Models

Method of Evaluation

- Silhouette Score
- Calinski-Harabasz Index
- Davies-Bouldin Index

Part IV: Conclusion

K-means clustering is the **best clustering method** for the variables based on **highest** silhouette value, **highest** calinski-harabasz index, and **lowest** davies-bouldin index

Four Categories of Data are Collected from 329 Respondent and Processed

Collected Data

Biodata	Food	Drinks	Lifestyle
Age	Meals Consumed	Water Consumed	Vitamin Consumed
Gender	Fast Food & Junk Food	Sugary Drinks	Alcohol Consumed
Height	Fruits and Vegetables	Caffeinated Drinks	Workout Time
Weight			Resting Time
			Cigarettes Smoked

Legends

*Numerical

*Categorical

Besides these 15 data, there are also class and major that will only be used for EDA purposes

Data Engineering

Some data engineering are done to pre-process these variables before we do any further analysis

Data Cleaning

- Remove unwanted columns such as "Timestamp"
- Renaming columns
- Changing "yes" and "no" categorical responses to 0 and 1
- Looking for duplicated data
- Removing outliers

Introducing New Metrics for EDA

- BMI: from weight and height
- Exercise to Sleep Ratio: from exercise and sleep

There are 6 of Interesting Findings based on EDA

Exploratory Data Analysis

1 Higher Healthier Food Consumption

PM students consumes 54% more fruit and vegetables compared to junk and fast food

2 Female is Healthier

Female students have healthier lifestyle: more veggies, less junk food, less sugar, less alcohol, less smoker

3 High Sweet & Caffeine Consumption

PM students consumes around 5.5 glasses of sweetened and caffeinated drinks per week

4 More Sleep, More Weight

PM students that sleeps more than 6 hours a day have higher potential for being overweight

5 Older Means Higher Caffeine

Strong positive correlation (0.73) exist between age and consumption of caffeinated drinks among PM students

6 Low Exercise to Sleep Ratio

Most students have ESR less than 0.1 (1 hour of exercise for 10 hours of sleep)

PCA Analysis

To simplify our datasets that has 15 variables, we will do **dimensional reduction** to only 2 using PCA from **sklearn.decomposition**

	principal component 1	principal component 2
0	0.064133	-0.753226
1	-1.374205	-0.808126
2	-2.913764	-1.504345
3	-2.209478	0.116412
4	-0.640636	-1.226615

Example of PCA results using .head() method

From **explained_variance_ratio_** property, we got that **principal component 1 holds 23.1% of the information** while the **principal component 2 holds only 15.3% of the information**. This means **61.6% information was lost** due to the dimensional reduction. This was probably due to low correlation between variables

There are Three Clustering Model We Wanted to Compare for This Data

Clustering Models

Models

K-Means	Partitions data into a specified number of clusters by minimizing the distance (error) between data points and cluster centroids.
Gaussian Mixture Model	Assumes data points are generated from a mixture of Gaussian distributions and estimates the parameters of these distributions to assign data points to clusters probabilistically
BIRCH	Constructs a hierarchical clustering structure by incrementally summarizing data points into smaller and more manageable subclusters, allowing for efficient and scalable clustering.

Evaluation Metrics

Silhouette Score

Measures cluster definition by averaging distances between points within and outside the cluster

Calinski-Harabasz Index

Compares inter-cluster and intra-cluster dispersion, with a higher score indicating distinct clusters.

Davies-Bouldin Index

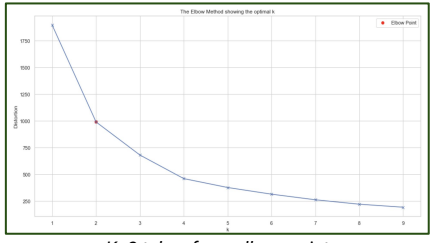
measures cluster similarity based on size and distance, with a lower score indicating well-defined clusters.

We will compare these three models using silhouette value, calinski-harabasz index, and davies-bouldin index to find the best clustering method for our data

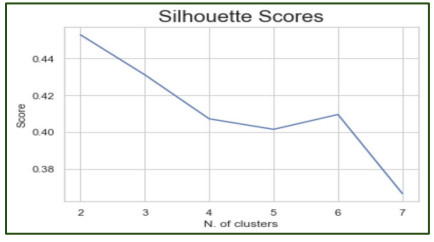
K-Means

1 Number of Cluster

Using elbow method and silhouette method, we got the optimum number of cluster is 2



K=2 taken from elbow point

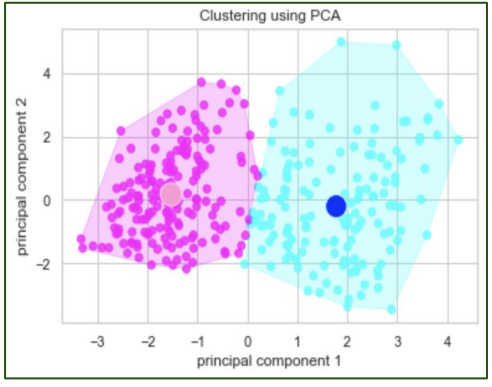


K=2 taken from highest silhouette score

2 Clustering Process

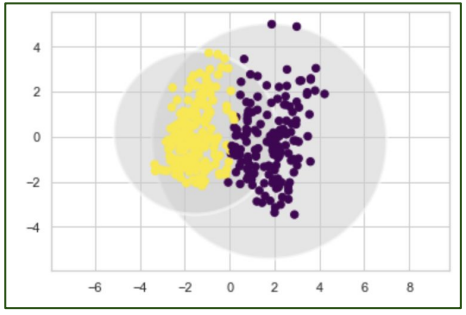
After 8 iterations of K, we reached SSE of 991.2 and converged to two centroids of:

- (1.78, -0.20)
- (-1.53, 0.17)



Cluster 0 contains 176 data points
Cluster 1 contains 153 data points

3 Evaluation Metrics



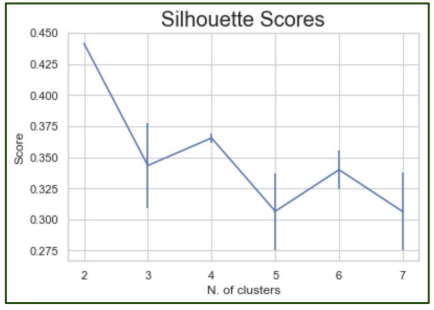
K-Means model circular representation

Silhouette Score	**0.4413**
Calinski-Harabasz Index	**298.9545**
Davies-Bouldin Index	**0.9278**

Gaussian Mixture Model

1 Number of Cluster

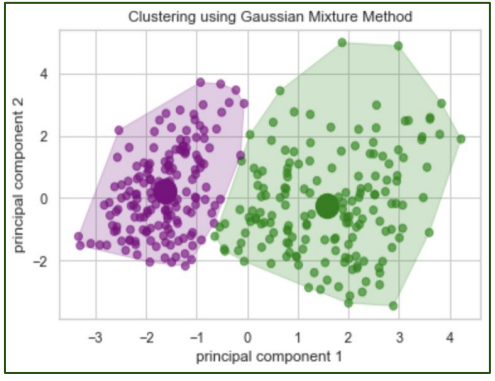
Using silhouette method, we got the optimum number of cluster is 2



K=2 taken from highest silhouette score

2 Clustering Process

Using GMM of 2 cluster, we have the following cluster

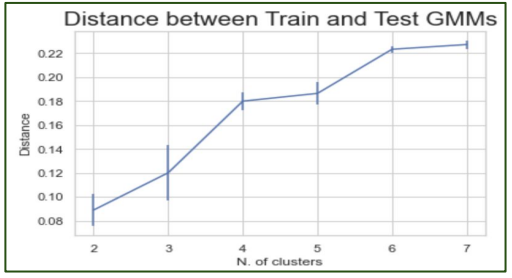


GMM Clustering result visualization

It can be seen that cluster 1 seems to be more sparse compared to cluster 0 who are more packed

3 Evaluation Metrics

Using Bayesian Information Criterion, we confirm 2 clusters has the lowest distance between train and test gmms

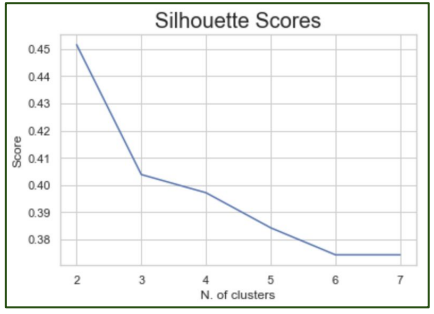


Silhouette Score	**0.4329**
Calinski-Harabasz Index	**289.3414**
Davies-Bouldin Index	**0.9355**

BIRCH

1 Number of Cluster

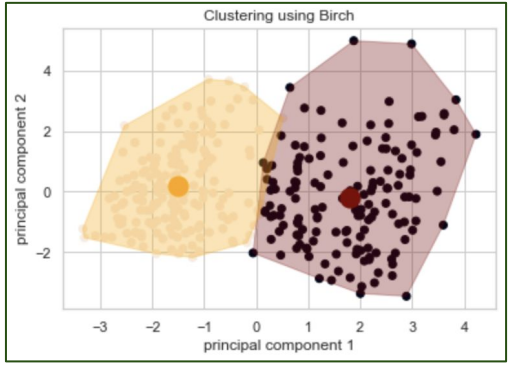
Using silhouette method, we got the optimum number of cluster is 2



K=2 taken from highest silhouette score

2 Clustering Process

Using GMM of 2 cluster, we have the following cluster



BIRCH Clustering result visualization

It can be seen that cluster 1 seems to be more sparse compared to cluster 0 who are more packed

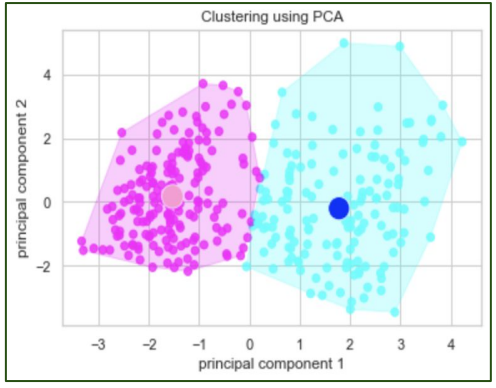
3 Evaluation Metrics

Silhouette Score	**0.4403**
Calinski-Harabasz Index	**297.8821**
Davies-Bouldin Index	**0.9294**

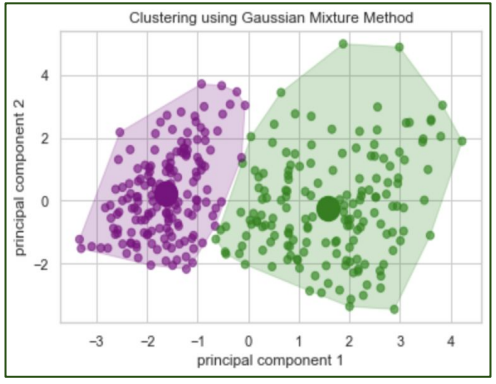
K-Means is the Best Model for Our Data Based on the Three Evaluation Metrics

Visualization

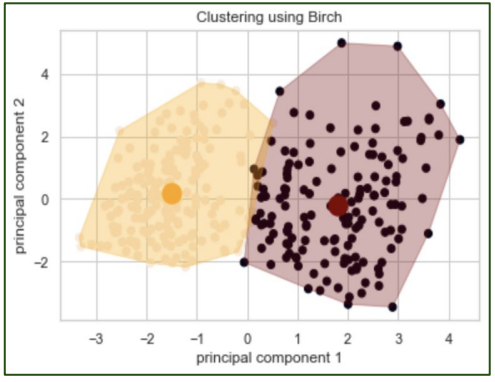
K-Means



Gaussian Mixture Model



Gaussian Mixture Model



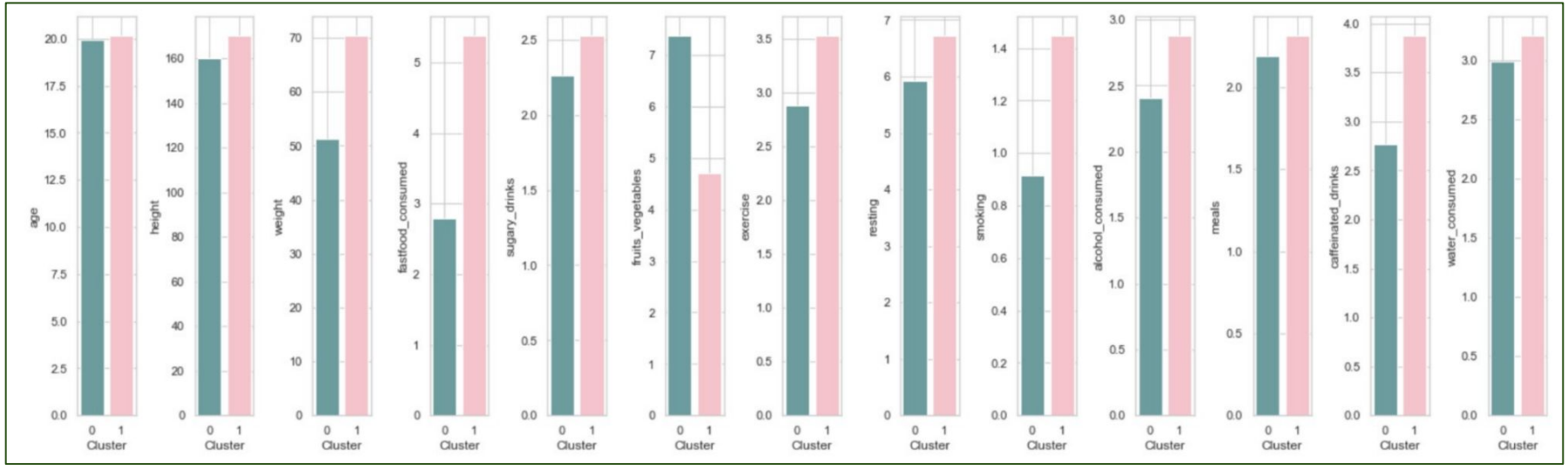
Eval. Metrics

	K-means	Gaussian Mixture	Birch
Silhouette	0.441397	0.432913	0.440333
Calinski	298.954544	289.341431	297.882086
Davies-Bouldin	0.927809	0.935487	0.929373

Based on highest silhouette value, highest calinski-harabasz index, and lowest davies-bouldin index, it can be deduced that the **best clustering method** for the variables i**s K-means** clustering.

Result

K-Means Clustering Result Showed Cluster 0 Tends to be More Healthy



Even though not generally applicable to every single variable, but we can conclude that overall, **cluster 0 is more healthy than cluster 1** since cluster 1 consume more fast food, less vegetables, more smoking, and drink alcohol & caffeine more compared to cluster 0

STEM Prasetiya Mulya

Thank You!

Any questions?

Appendix

EDA Graphs