37631926

KB seitei

Clustering fitness tracker data.

Introduction

Data clustering is a fundamental task in data mining and machine learning that involves partitioning a set of data points into groups, or clusters, where members of each group share greater similarity with each other than with those in other groups. This unsupervised learning technique is widely applied in areas such as data summarization, learning, segmentation, and target marketing.

In the absence of labelled data, clustering serves as a concise model of the dataset, functioning either as a summary or a generative model. By identifying inherent structures within the data, clustering enables the discovery of hidden patterns and relationships, facilitating more informed decision-making.

The primary objective of clustering is to divide data points into distinct groups such that intra-group similarities are maximized while inter-group similarities are minimized. This process aids in understanding the underlying distribution and characteristics of the data, which is crucial for various analytical tasks and strategic applications.

Data Generation

To stimulate data the is given data for three users we firstly import important libraries in Google colab for random data generation, plotting graphs and kmeans clustering. to ensure we get consistent result we set the random seed to a number, to get the same random numbers which will make the results consistent and reproducible. We the generated random user counts for each cluster and combine all the generated data into one single dataset to store in a 2D array.

<u>Observations</u>

<u>Visualising The data before clustering</u>

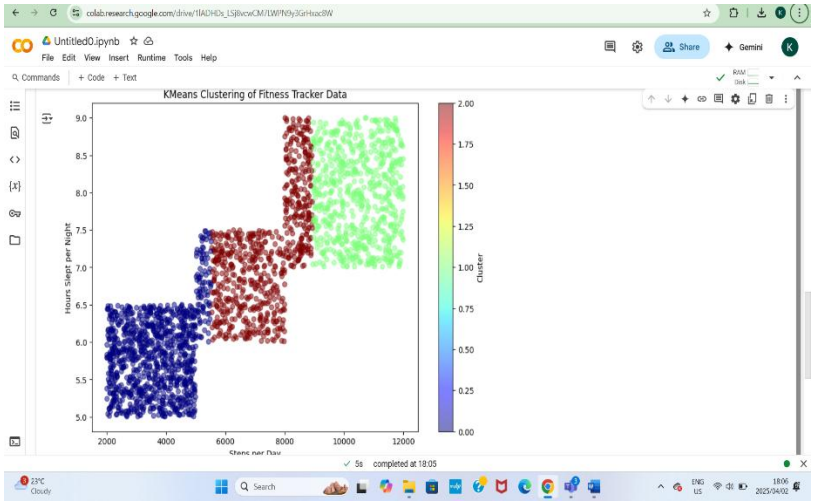
This cluster 3 falls between 0 to 4000 steps per day and 5.0 to 6.5 hours slept per night. These individuals tend to be less physically active and sleep fewer hours.

The second group is positioned between 4000 to 8000 steps per day and 6.5 to 7.5 hours slept per night. They maintain a balanced level of physical activity and get a moderate amount of sleep. The cluster 1 is spread between 8000 to 12000 steps per day and 7.5 to 9.0 hours slept per night. This suggests that increased daily movement might be correlated with longer sleep duration. This visualization hints at a potential relationship between physical activity and sleep patterns higher activity levels seem to be linked to longer sleep durations.

Clustering

Here we create a model, we set the random seed to 30 and use the fit method to compute the k means clustering on the provided dataset data. We effectively group similar data point together then retrieve the cluster for each data point .

Cluster Visualization



Some individuals in Cluster 2 (Moderate Activity) have sleep durations that are closer to Cluster 3 (High Activity).Likewise, some data points in Cluster 1 (Low Activity) overlap with Cluster 2, meaning that even less active individuals sometimes get the same amount of sleep as moderately active ones.

Conclusion

The data before clustering does not show any clear overlapping regions or outliers while the clustered data shows everything. The general trend remains consistent higher step counts tend to correspond with longer sleep durations, However, the spread of the data shows that this isn't a simple rule, some outliers may go against this pattern. individuals fit neatly into a group, while others fall between groups.

Click this link to view the CODE:https://colab.research.google.com/drive/1lADHDs_LSj8vcwCM7LWPN9y3GrHxac8W?usp=sharing or

```
-*- coding: utf-8 -*-
"""Untitled0.ipynb

Automatically generated by Colab.

Original file is located at
```

```
https://colab.research.google.com/drive/1lADHDs_LSj8vcwCM7LWPN9y3GrHxac
8W
"""

import numpy as np
import matplotlib.pyplot as plt
from sklearn.cluster import KMeans

#seitei 37631926
np.random.seed(30)

num_users = np.random.randint(100, 1000, 3)


steps_active = np.random.randint(8000, 12000, num_users[0])
sleep_active = np.random.uniform(7, 9, num_users[0])

steps_moderate = np.random.randint(5000, 8000, num_users[1])
sleep_moderate = np.random.uniform(6, 7.5, num_users[1])
#seitei 37631926
steps_least_active = np.random.randint(2000, 5000, num_users[2])
sleep_least_active = np.random.uniform(5, 6.5, num_users[2])

steps = np.concatenate([steps_active, steps_moderate,
steps_least_active])
sleep = np.concatenate([sleep_active, sleep_moderate,
sleep_least_active])


data = np.column_stack((steps, sleep))

print("Total user generated: {0}".format(len(data)))

#seitei 37631926
plt.figure(figsize=(10,7))
plt.scatter(steps, sleep, color='orange', alpha=0.5)
plt.xlabel("Steps per Day")
plt.ylabel("Hours Slept per Night")
plt.title("Fitness Tracker Data Visualization")
plt.show()

kmeans = KMeans(n_clusters=3, n_init=10)
kmeans.fit(data)
labels = kmeans.labels_
#seitei 37631926
plt.figure(figsize=(10,7))
plt.scatter(steps, sleep, c=labels, cmap='jet', alpha=0.5)
plt.xlabel("Steps per Day")
plt.ylabel("Hours Slept per Night")
plt.title("KMeans Clustering of Fitness Tracker Data")
plt.colorbar(label="Cluster")
plt.show()
#seitei 37631926
for i in range(3):
    avg_steps = steps[labels == i].mean()
    avg_sleep = sleep[labels == i].mean()
```

```
    print(f"Cluster {i+1}: Avg Steps = {avg_steps:.2f}, Avg Sleep =
{avg_sleep:.2f} hours")
```

References

Charu C, Chandan K(2014)*Data clustering algorithms and Applications,CRC PRESS*.