Лабораторная работа №7. Кластеризация данных методом K-Means

Освоить методы кластеризации данных и сегментации объектов.

Датасет для работы

Используется набор данных Mall Customers Dataset.

Источник: Mall Customers Dataset

Теоретические сведения

Кластеризация — метод машинного обучения без учителя, позволяющий разделить объекты на группы по схожести признаков.

Алгоритм K-Means:

1. Выбор количества кластеров.
2. Определение центров кластеров.
3. Распределение объектов.
4. Пересчет центров.

Задание

1. Загрузить набор данных.
2. Выбрать признаки для кластеризации.
3. Выполнить масштабирование данных.
4. Построить модель K-Means.
5. Визуализировать результаты кластеризации.

Пример выполнения

```
import pandas as pd
import matplotlib.pyplot as plt

from sklearn.cluster import KMeans
from sklearn.preprocessing import StandardScaler
```

```
# Загрузка данных
data = pd.read_csv("clean_data.csv")

# Выбор признаков
X = data[['Feature1', 'Feature2']]

# Масштабирование
scaler = StandardScaler() X_scaled
= scaler.fit_transform(X)

# K-Means
kmeans = KMeans(n_clusters=3, random_state=42) kmeans.fit(X_scaled)

# Метки кластеров data['Cluster']
= kmeans.labels_

# Визуализация
plt.scatter(X['Feature1'], X['Feature2'],
        c=data['Cluster'])

plt.xlabel("Feature1") plt.ylabel("Feature2")
plt.title("Кластеризация K-Means")
plt.show()
```

Контрольные вопросы

1. Что такое кластеризация?
2. В чем отличие обучения с учителем и без учителя?
3. Для чего выполняется масштабирование данных?
4. Как работает алгоритм K-Means?
   Что такое центр кластера?