Лабораторная работ 4.
Понижение размерности данных.

Цель: познакомиться с методами понижение размерности данных главных компонент и факторного анализа.
Задачи:

1. Стандартизировать данные;
2. Разделить данные на класс и признаки;
3. Провести понижение размерности;
4. Определить значение дисперсии;
5. Построить диаграмму рассеяния.

Пример выполнения лабораторной работы.

Выведем сводную информацию о подготовленном датасете.

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
from sklearn import preprocessing
from sklearn.preprocessing import StandardScaler
from sklearn.decomposition import PCA
df = pd.read_csv('lab4.csv',encoding='cp1251')
print(df.info())
```

Результат:

```
RangeIndex: 198 entries, 0 to 197
Data columns (total 5 columns):
 #   Column           Non-Null Count  Dtype
---  ------           --------------  -----
 0   id               198 non-null    int64
 1   gender           198 non-null    object
 2   age              198 non-null    int64
 3   income           198 non-null    int64
 4   spending_rating  198 non-null    int64
dtypes: int64(4), object(1)
memory usage: 7.9+ KB
```

Разделим данные на признаки и классы. В данном случае пол (мужской, женский) будет классом, а возраст, доход и рейтинг трат признаками. С учетом этого стандартизируем данные.

```
df = pd.read_csv('lab4.csv',encoding='cp1251')
variables = ['age', 'income', 'spending_rating']
x = df.loc[:, variables].values
y = df.loc[:,['gender']].values
x = StandardScaler().fit_transform(x)
```

```
x = pd.DataFrame(x)
print(x)
```

Результат:

```
            0         1         2
0   -1.424076 -1.798943 -0.422716
1   -1.279921 -1.798943  1.219183
2   -1.351999 -1.759745 -1.712779
3   -1.135767 -1.759745  1.062811
4   -0.559147 -1.720548 -0.383623
..        ...       ...       ...
193  0.594093  2.316830 -1.321851
194  0.449938  2.552017 -0.852737
195 -0.487069  2.552017  0.945533
196 -0.487069  2.983193 -1.243665
197 -0.631224  2.983193  1.297368

[198 rows x 3 columns]
```

Проводим понижение размерности методом главных компонент.

```
pca = PCA()
x_pca = pca.fit_transform(x)
x_pca = pd.DataFrame(x_pca)
print(x_pca)
```

Результат:

```
            0         1         2
0   -0.636056 -1.806655  1.332032
1   -1.693911 -1.867253  0.069542
2    0.324541 -1.716548  2.192340
3   -1.483144 -1.819480  0.077549
4   -0.055512 -1.714284  0.691553
..        ...       ...       ...
193  1.261044  2.376292  0.480561
194  0.818420  2.590621  0.247520
195 -1.113995  2.504567 -0.360859
196  0.415104  3.019606  1.180273
197 -1.481606  2.919221 -0.513846

[198 rows x 3 columns]
```

Выводим значения дисперсии и собственные числа, которые соответствуют компонентам.

```
print(pca.explained_variance_ratio_)
print(pca.singular_values_)
```

Результат:

```
[0.43283775 0.33319535 0.2339669 ]
[16.03451356 14.06833462 11.7888225 ]
```

Первый основной компонент составляет 43,28% дисперсии, второй 33,31%, третий 23,39%.

Добавим наш класс к новым данным.

```
x_pca['gender']=y
```

```
x_pca.columns = ['PC1','PC2','PC3','gender']
print(x_pca.head())
```

Результат:

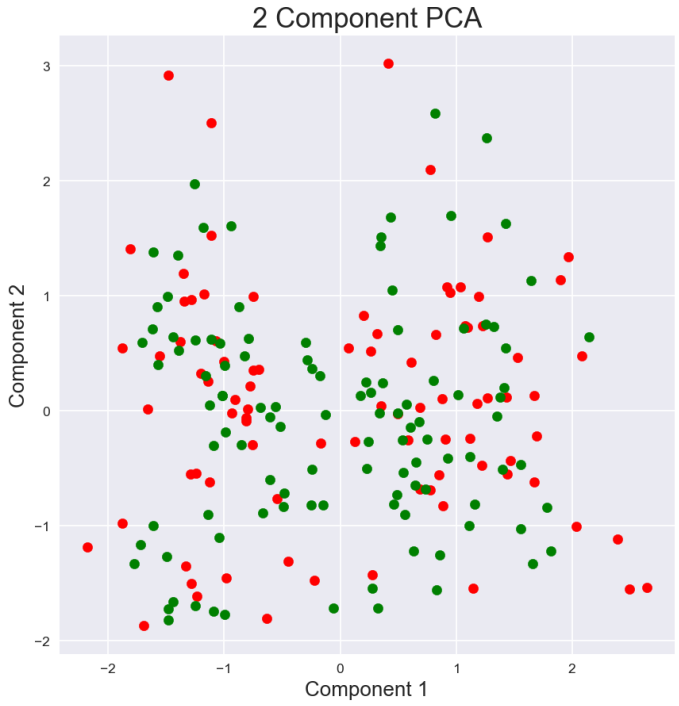
	PC1	PC2	PC3	gender
0	-0.636056	-1.806655	1.332032	Male
1	-1.693911	-1.867253	0.069542	Male
2	0.324541	-1.716548	2.192340	Female
3	-1.483144	-1.819480	0.077549	Female
4	-0.055512	-1.714284	0.691553	Female

Построим, диаграмму рассеяния используя первые два компонента.

```
fig = plt.figure(figsize = (8,10))
ax = fig.add_subplot(1,1,1)
ax.set_xlabel('Component 1', fontsize = 15)
ax.set_ylabel('Component 2', fontsize = 15)
ax.set_title('2 Component PCA', fontsize = 20)
gender = ['Male', 'Female']
colors = ['r', 'g']
for gender, color in zip(gender,colors):
    indicesToKeep= x_pca['gender'] == gender
    ax.scatter(x_pca.loc[indicesToKeep, 'PC1']
               , x_pca.loc[indicesToKeep, 'PC2']
               , c = color, s = 50)


plt.show()
```

Результат:



В ходе лабораторной работы мы провели понижение размерности данных методом главных компонент и методом факторного анализа, сравнили данные и получили …

Задание.

1. Стандартизировать данные методом MinMaxScaler;
2. Разделить данные на класс и признаки;
3. Провести понижение размерности методом главных компонент и методом факторного анализа;
4. Определить значение дисперсии;
5. Построить диаграмму рассеяния;
6. Сравнить результат и описать разницу в методах.

Формат отчета.

Протокол лабораторной работы в формате PDF, который должен содержать, поэтапное выполнение всех задач с текстовым описанием ваших действий и экранными формами, отображающими данные действия.