МИНОБРНАУКИ РОССИИ
САНКТ-ПЕТЕРБУРГСКИЙ ГОСУДАРСТВЕННЫЙ
ЭЛЕКТРОТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ
«ЛЭТИ» ИМ. В.И. УЛЬЯНОВА (ЛЕНИНА)
Кафедра МОЭВМ

ОТЧЕТ

по лабораторной работе №2

по дисциплине «Машинное обучение»

Тема: Понижение размерности пространства признаков

Студент гр. 6307 _____ Мишанов А. А.

Преподаватель _____ Жангиров Т. Р.

Санкт-Петербург

2020

Цель работы

Ознакомиться с методами понижения размерности данных из библиотеки Scikit Learn.

Ход работы

Загрузка данных

1. Загружен датасет по ссылке. Данные загружены в датафрейм с разделением на описательные признаки и признак, отображающий класс.

	RI	Na	Mg	Al	Si	K	Ca	Ba	Fe	Type
0	1.52101	13.64	4.49	1.10	71.78	0.06	8.75	0.0	0.0	1
1	1.51761	13.89	3.60	1.36	72.73	0.48	7.83	0.0	0.0	1
2	1.51618	13.53	3.55	1.54	72.99	0.39	7.78	0.0	0.0	1
3	1.51766	13.21	3.69	1.29	72.61	0.57	8.22	0.0	0.0	1
4	1.51742	13.27	3.62	1.24	73.08	0.55	8.07	0.0	0.0	1

Рисунок 1. Загруженный датасет

2. Построены диаграммы рассеяния для пар признаков.

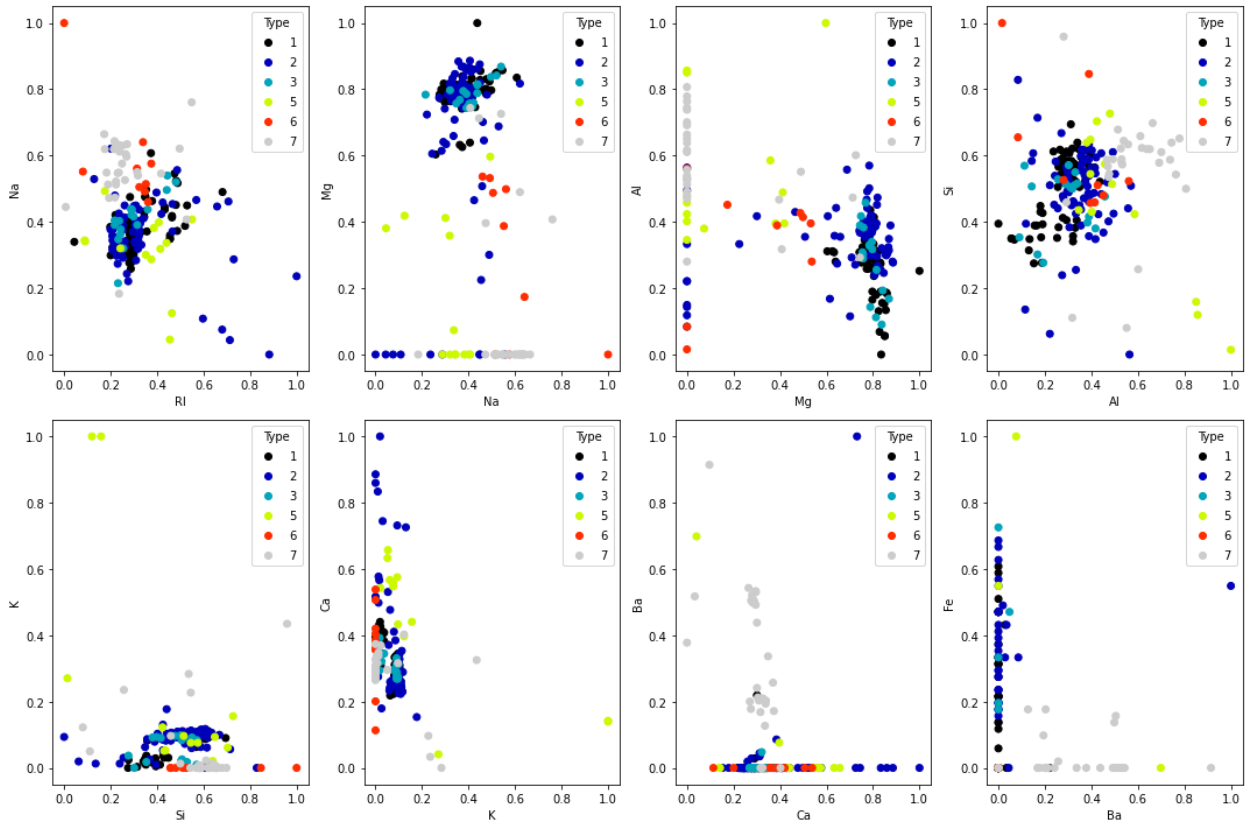


Рисунок 2. Диаграммы рассеяния признаков

Метод главных компонент

1. Выполнено понижение пространства признаков до размерности 2 с помощью метода главных компонент. Диаграмма рассеяния для новой пары признаков представлена на рисунке 3.

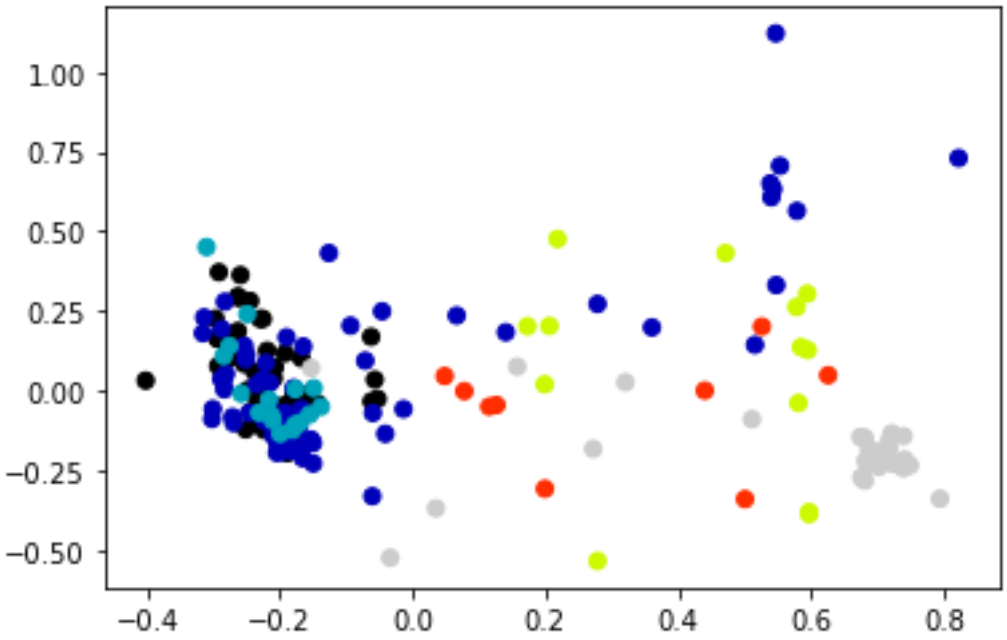


Рисунок 3. Диаграмма рассеяния для двух компонент

2. Были получены значения объясненной дисперсии и собственные числа.
   - Значения объясненной дисперсии: [0.45429569, 0.17990097]
   - Собственные числа: [5.1049308 , 3.21245688]
3. Изменяя количество компонент, определим количество, при котором компоненты объясняют не менее 85% дисперсии данных.

   [**0.454296, 0.179901, 0.126495, 0.097978**, 0.068624, 0.042141, 0.026098, 0.004328, 0.000139].

   Первые четыре компоненты объясняют не менее 85% дисперсии данных.
4. Восстановление данных с помощью inverse_transform при 4 компонентах.

Таблица 1. Восстановленные данные

	RI	Na	Mg	Al	Si	K	Ca	Ba	Fe
count	214.0000	214.0000	214.0000	214.0000	214.0000	214.0000	214.0000	214.0000	214.0000
mean	0.316744	0.402684	0.597891	0.359784	0.507310	0.080041	0.327785	0.055570	0.111783
std	0.130009	0.068790	0.320561	0.135910	0.129659	0.043794	0.128754	0.132283	0.188632

Таблица 2. Исходные данные

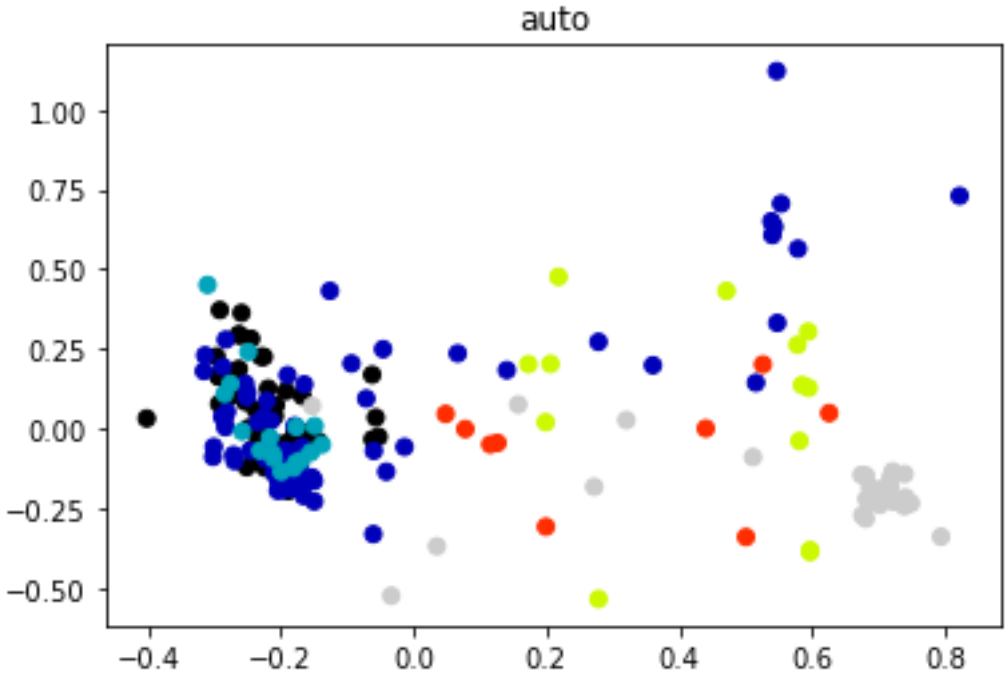
	RI	Na	Mg	Al	Si	K	Ca	Ba	Fe
count	214.0000	214.0000	214.0000	214.0000	214.0000	214.0000	214.0000	214.0000	214.0000
mean	0.316744	0.402684	0.597891	0.359784	0.507310	0.080041	0.327785	0.055570	0.111783
std	0.133313	0.122798	0.321249	0.155536	0.138312	0.105023	0.132263	0.157847	0.191056

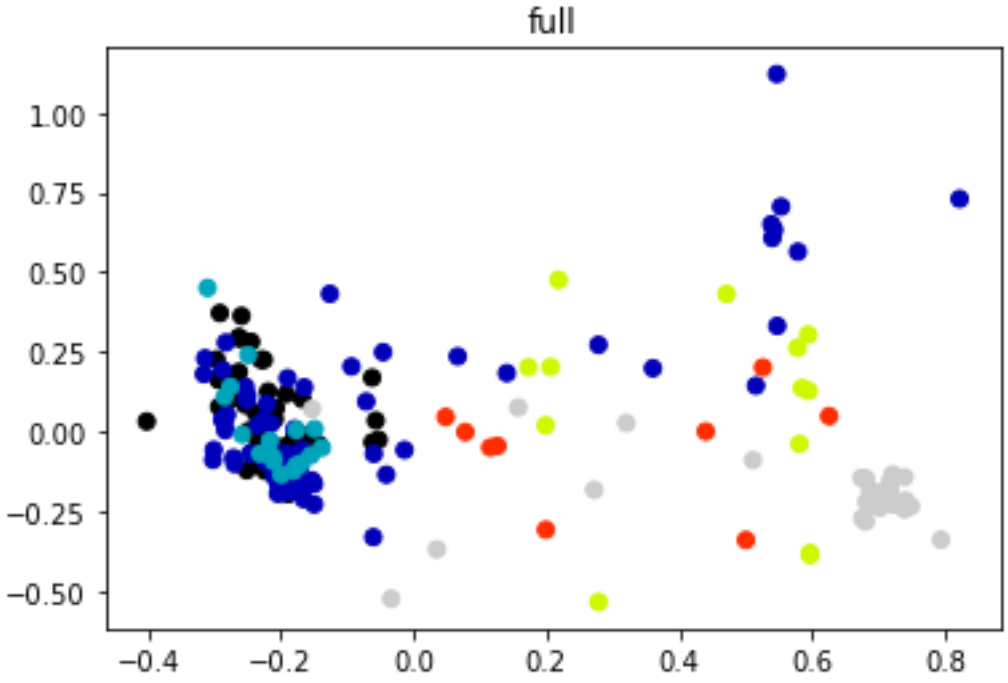
Видно, что при восстановление данных, некоторые признаки теряют часть СКО.

5. Исследование метода главных компонент при различных параметрах *svd_solver*.

   Параметр *svd_solver* может принимать значения: *auto, full, arpack, randomized.*

   Диаграмма рассеяния при различных параметрах представлена на рисунке 4.


auto
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.4
−0.2
0.0
0.2
0.4
0.6
0.8


full
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.4
−0.2
0.0
0.2
0.4
0.6
0.8

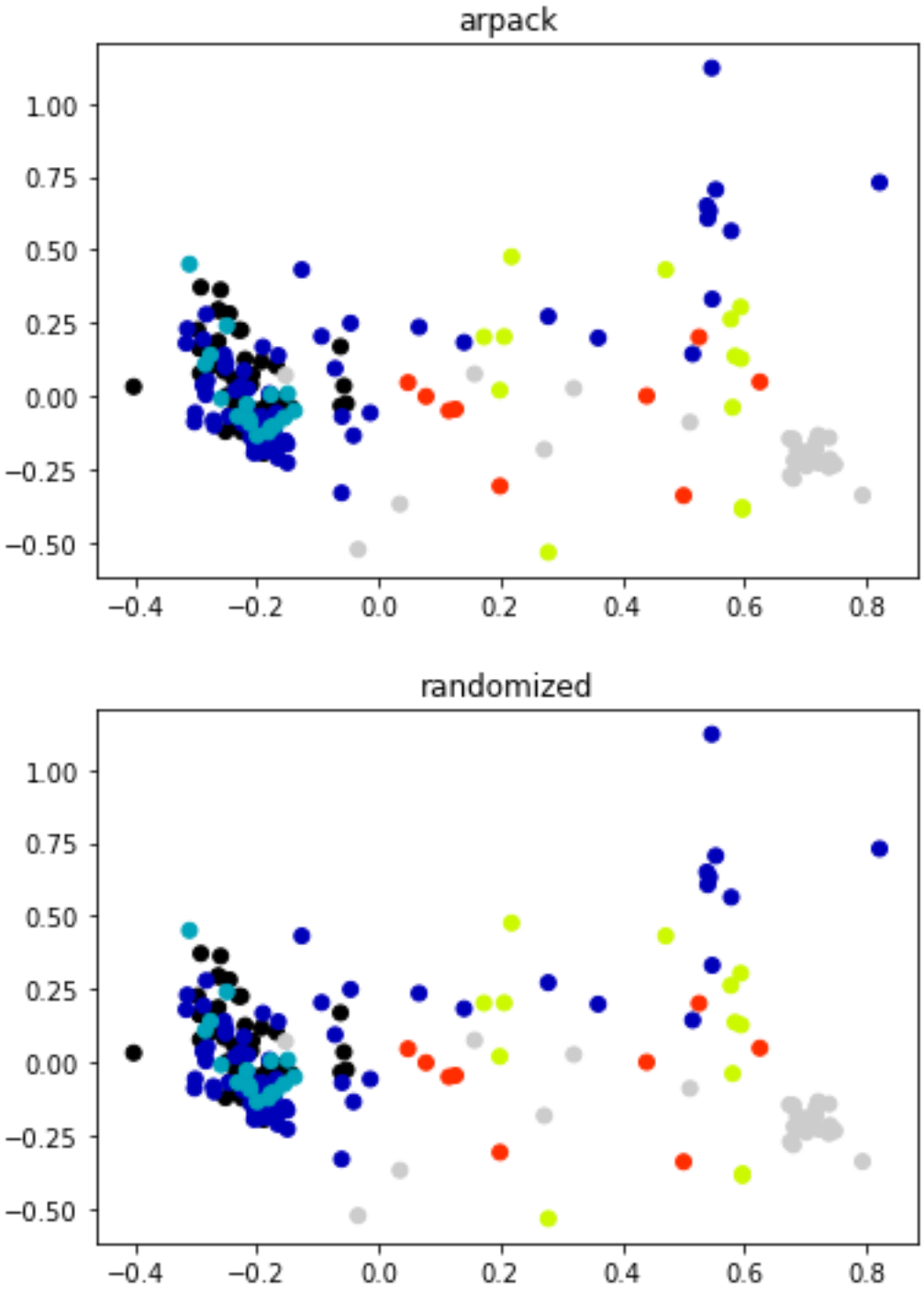
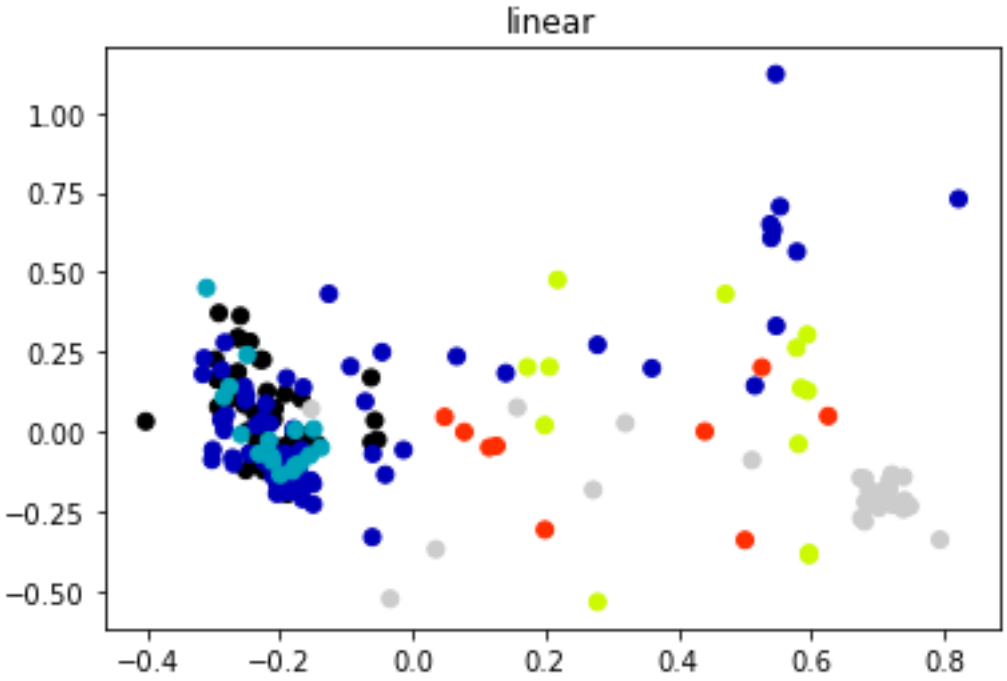
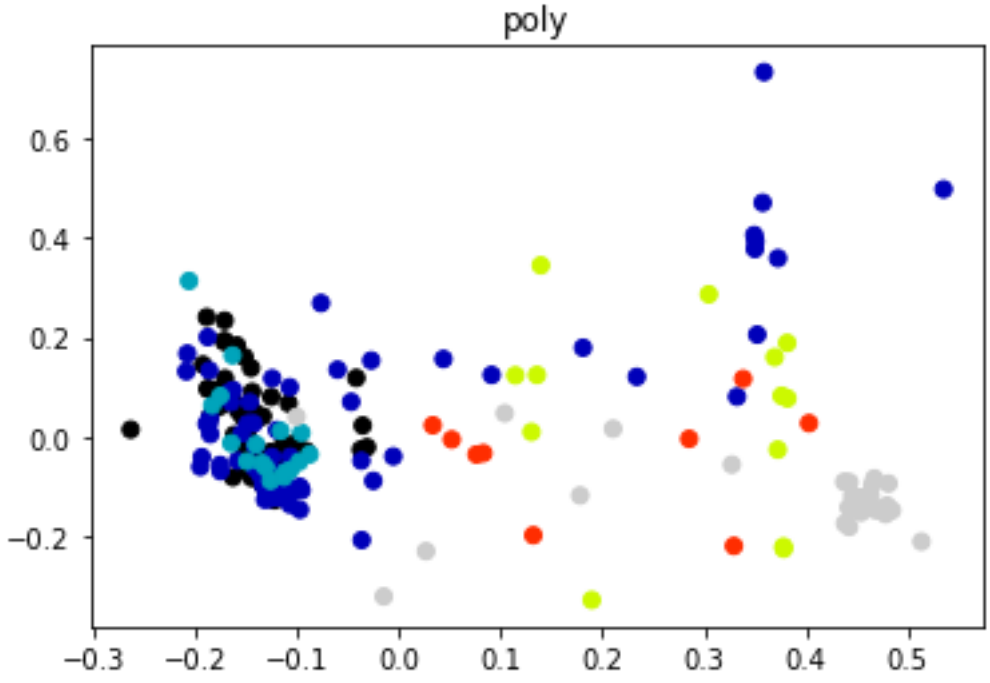


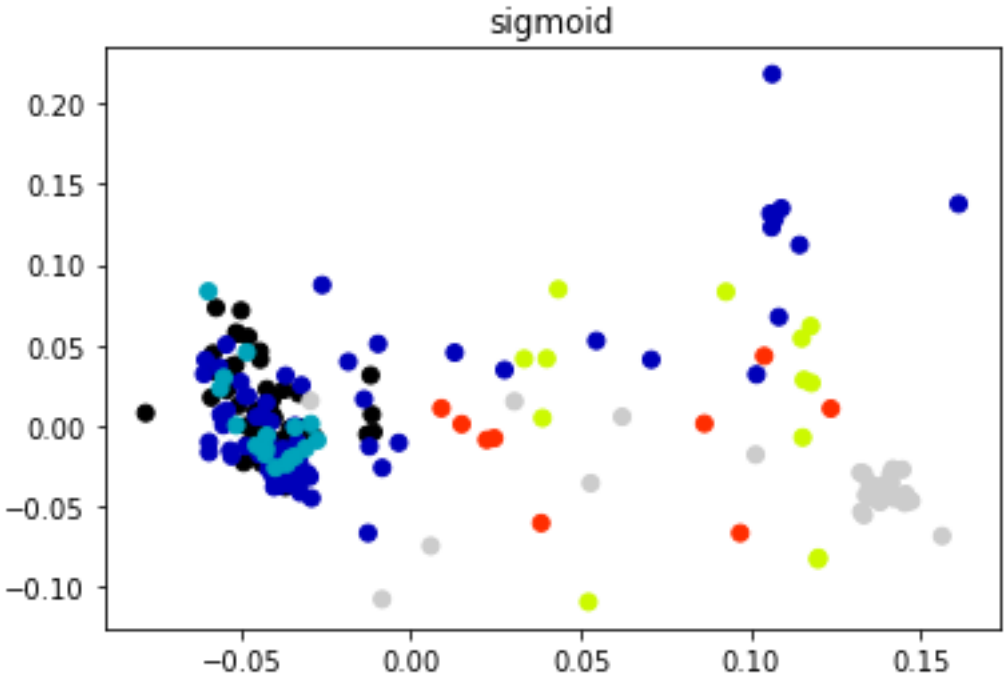
Рисунок 4. Диаграммы рассеяния при различных параметрах *svd_solver*

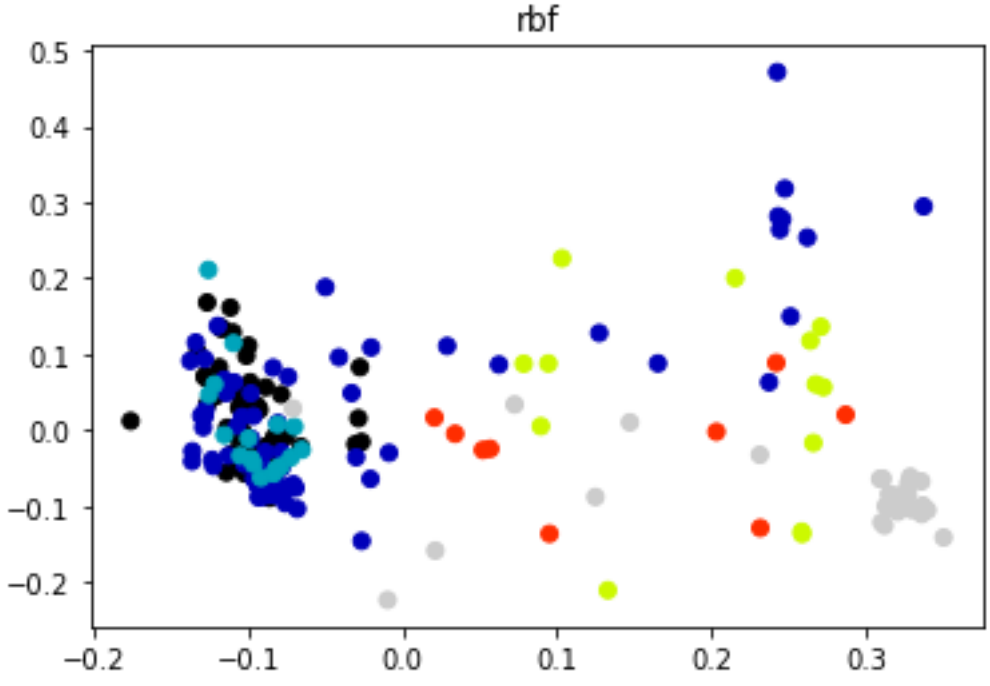
Модификация метода главных компонент

1. По аналогии с PCA исследуем KernelPCA для различных параметров kernel.


linear
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.4
−0.2
0.0
0.2
0.4
0.6
0.8


poly
0.6
0.4
0.2
0.0
−0.2
−0.3
−0.2
−0.1
0.0
0.1
0.2
0.3
0.4
0.5


sigmoid
0.20
0.15
0.10
0.05
0.00
−0.05
−0.10
−0.05
0.00
0.05
0.10
0.15


rbf
0.5
0.4
0.3
0.2
0.1
0.0
−0.1
−0.2
−0.2
−0.1
0.0
0.1
0.2
0.3

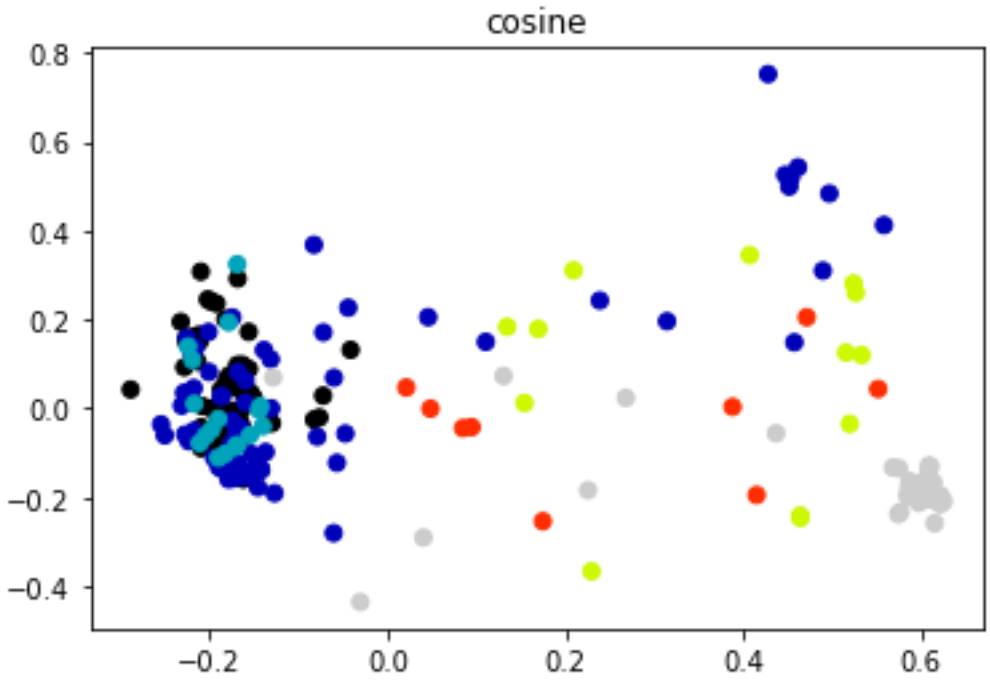


Рисунок 5. Диаграммы рассеяния при различных параметрах kernel

2. По диаграммам видно, что данные распределены одинаково, отличается только масштаб.
3. При *kernel='linear'* KernelPCA ведет себя также, как и PCA.
4. Аналогично исследуем SparcePCA.

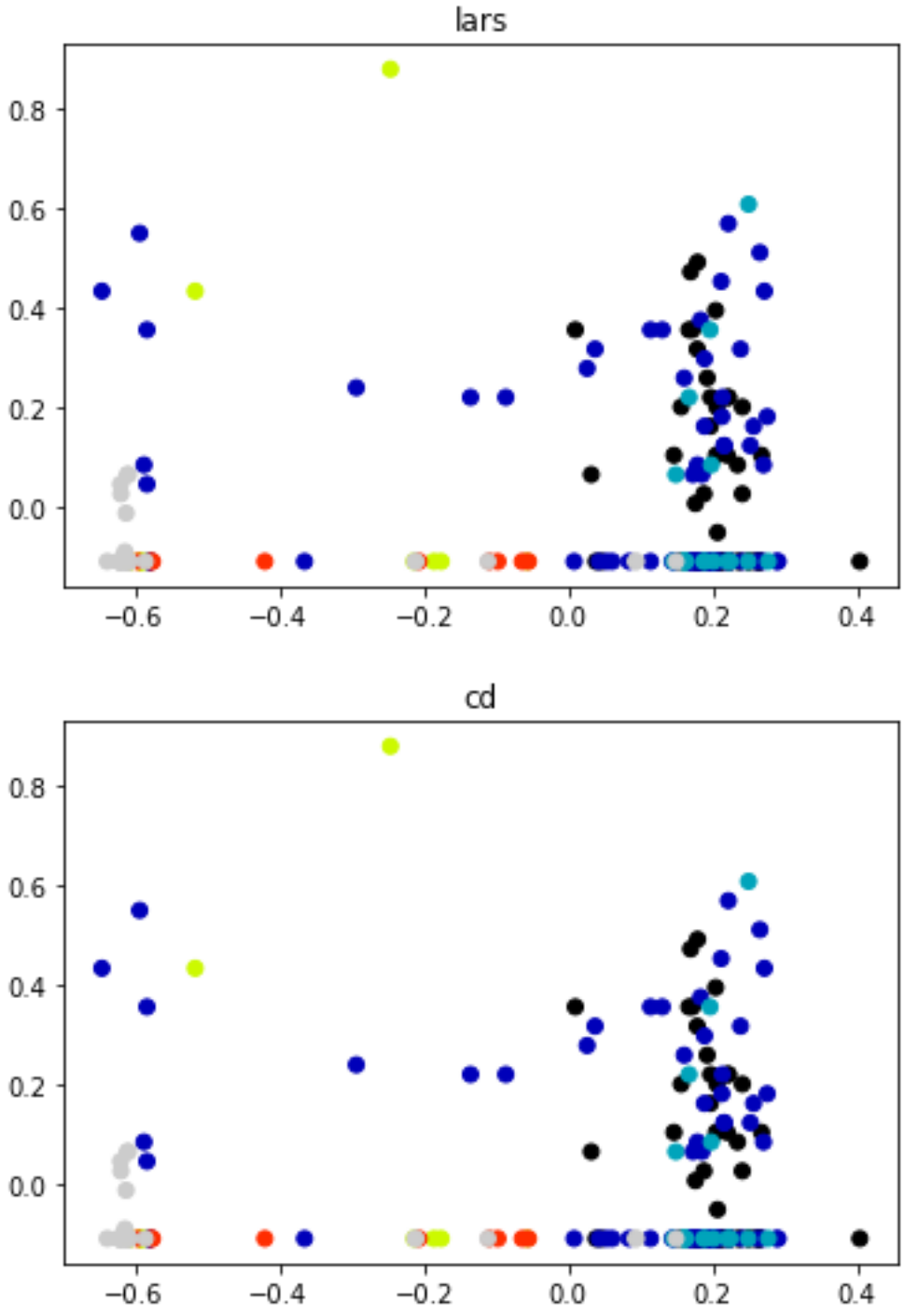


Рисунок 6. Диаграммы рассеяния при различных параметрах method

Факторный анализ

1. Выполнено понижение пространства признаков до размерности 2 с помощью факторного анализа. Диаграмма рассеяния для новой пары признаков представлена на рисунке 7.

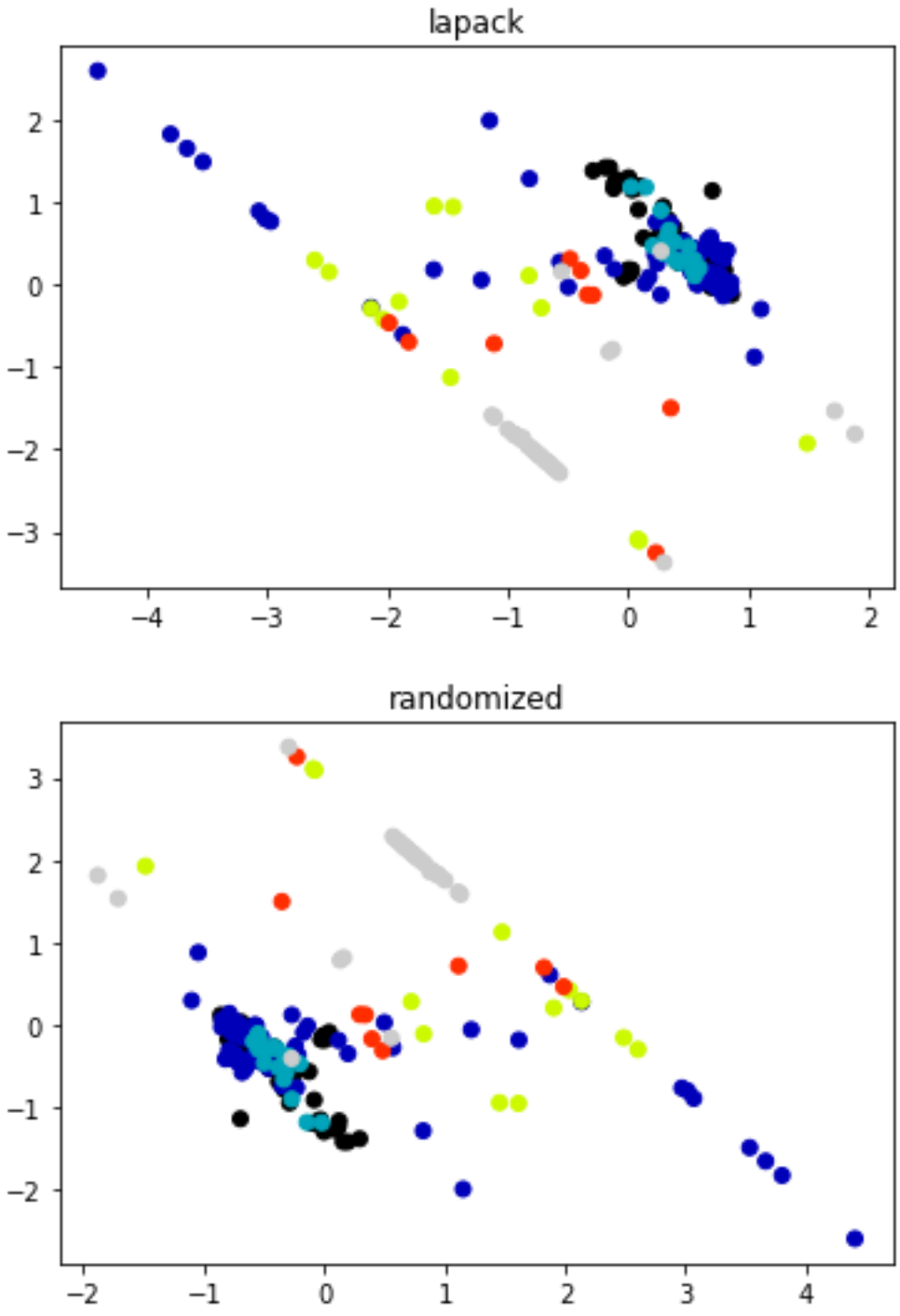


Рисунок 7. Диаграмма рассеяния после факторного анализа

2. Различия между методом главных компонент и факторным анализом.
   - Главные компоненты - это случайные величины максимальной дисперсии, построенные на основе линейных комбинаций входных характеристик. Они являются проекциями на оси главных компонентов, которые представляют собой линии, которые минимизируют средний квадрат расстояния до каждой точки в

наборе данных. Чтобы гарантировать уникальность, все оси главных компонентов должны быть ортогональными.

- Факторный анализ выполняет минимизацию корреляции между исходными переменными и факторами.
- PCA подходит для уменьшения размерности пространства признаков, а факторный анализ для поиска скрытых зависимостей.

Вывод

В результате работы были получены практические навыки по применению методов понижения пространства признаков библиотеки Scikit Learn.