МИНОБРНАУКИ РОССИИ

САНКТ-ПЕТЕРБУРГСКИЙ ГОСУДАРСТВЕННЫЙ

ЭЛЕКТРОТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ

«ЛЭТИ» ИМ. В.И. УЛЬЯНОВА (ЛЕНИНА)

Кафедра МО ЭВМ

ОТЧЕТ

по лабораторной работе №2

по дисциплине «Машинное обучение»

Тема: Понижение размерности пространства признаков

Студент гр. 6307 _____ Ходос А.А.

Преподаватель _____ Жангиров Т.Р.

Санкт-Петербург

2020

Цель работы

Ознакомиться с методами понижения размерности данных из библиотеки Scikit Learn.

Ход работы

1. Загрузка данных

Данные загруженного датасета разделены на описательные признаки и признаки отображающие класс. Данные приведены к интервалу [0; 1]. Построенные диаграммы рассеяния для пар признаков приведены на рисунке 1.

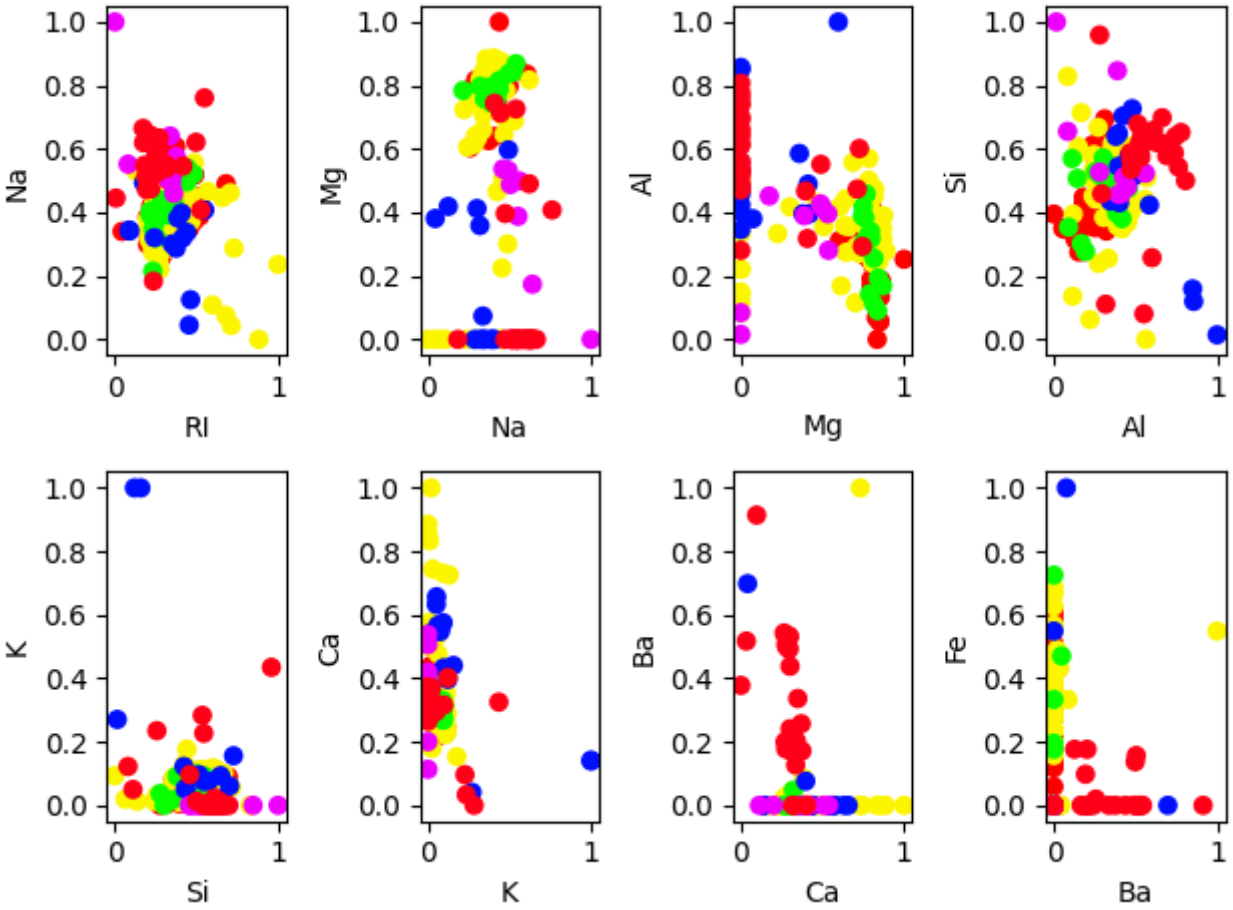

Рисунок 1

Для определения соответствия цвета на диаграмме и класса в датасете была построена диаграмма зависимости цвета от класса, приведенная на рисунке 2.

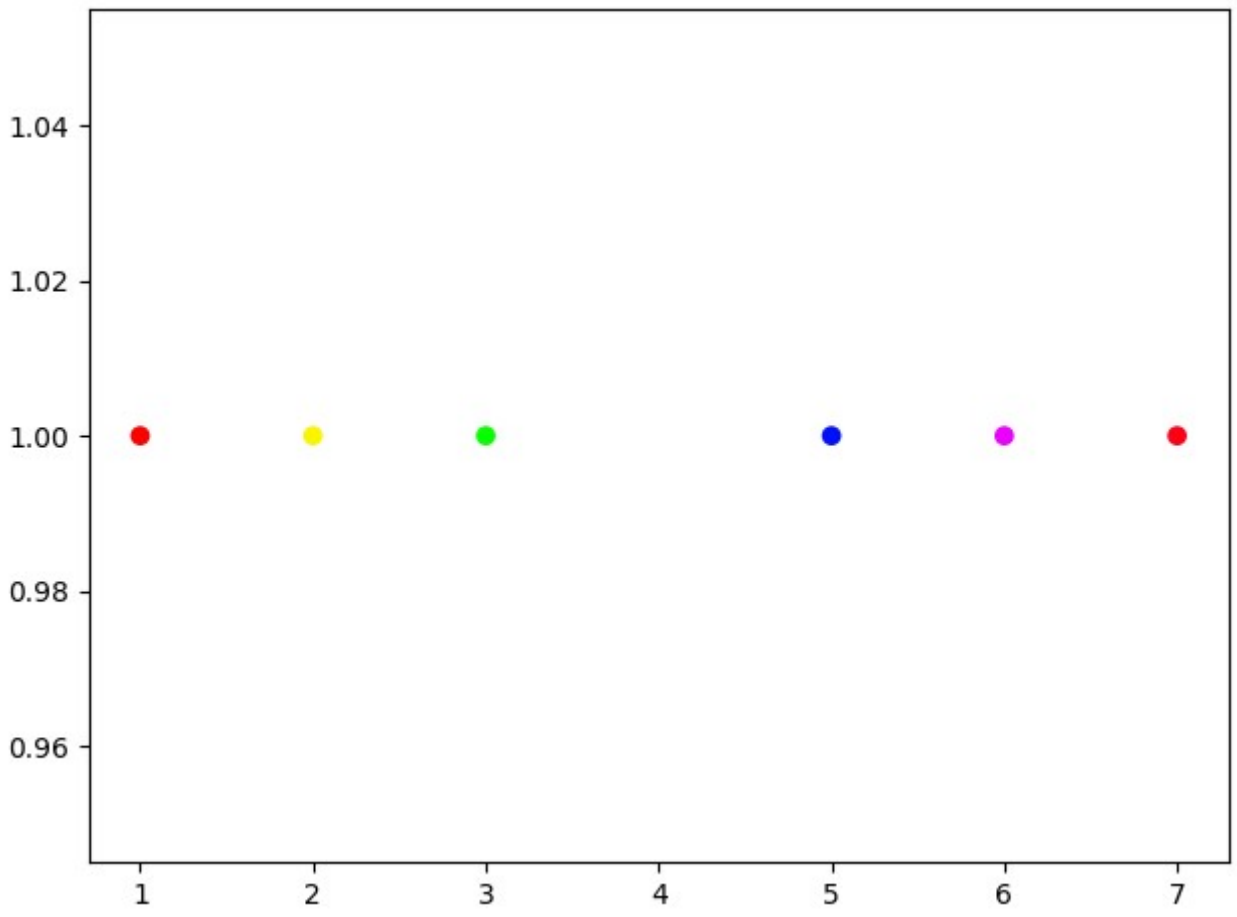

Рисунок 2

Таким образом цвет каждого класса:

1 : красный

2 : желтый

3 : зеленый

5 : синий

6 : фиолетовый

7: красный

2. Метод главных компонент

Используя метод главных компонент, было проведено понижение размерности пространства до размерности 2. Диаграмма рассеяния для новой пары компонент приведена на рисунке 3.

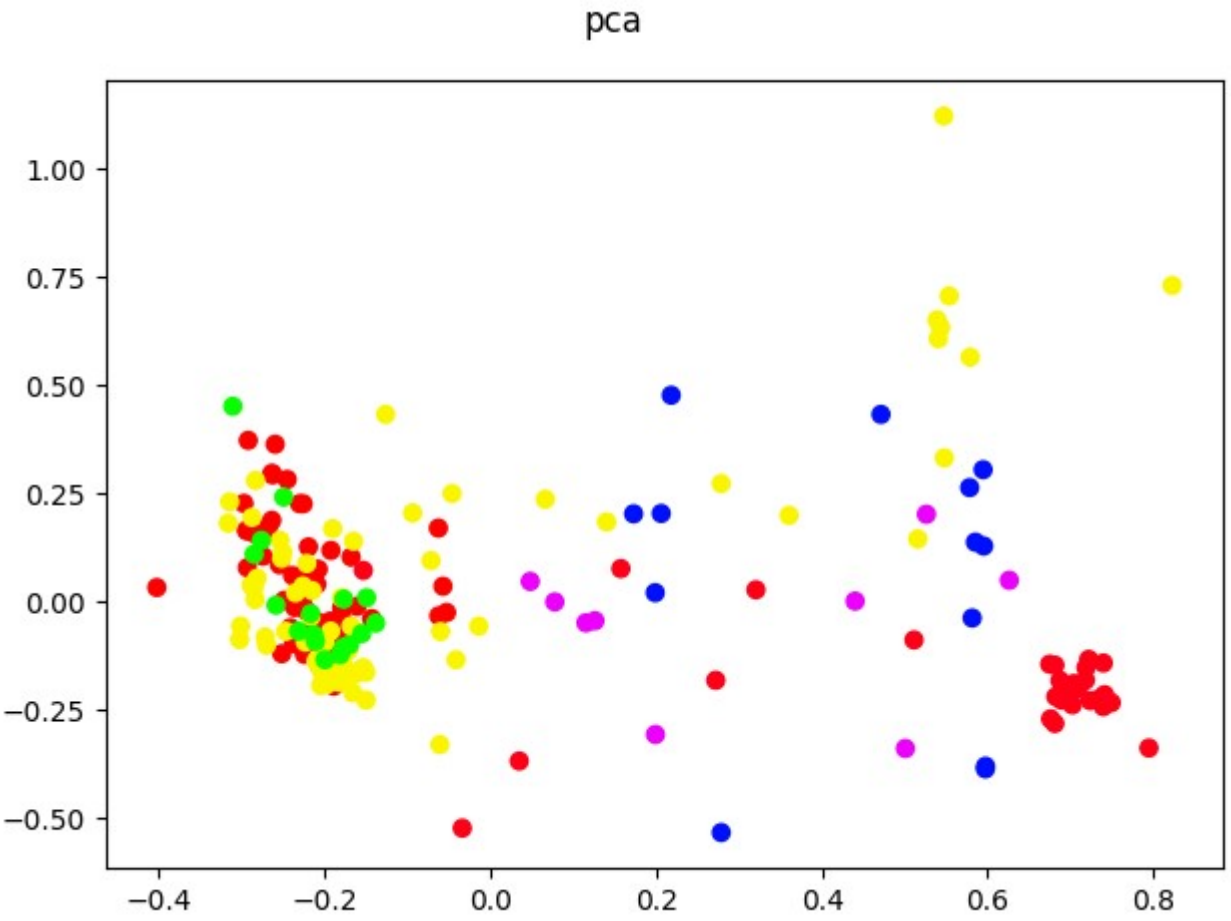

Рисунок 3

Значения объясненной дисперсии (в процентах): [45.42956891 17.9900973]

Собственные числа: [5.1049308 3.21245688]

Определим зависимость объясненной дисперсии от количества компонент:

1 : 0.45429568907468526

2 : 0.634196662104278

5 : 0.9272937149511479

6 : 0.9694347221994033

3 : 0.7606912558548664 7 : 0.9955326243472864

4 : 0.8586697305102718 8 : 0.9998605862637865

Применив метод обратной трансформации, восстановим исходные данные. Диаграмма рассеяния для пар восстановленных признаков продемонстрирована на рисунке 4.

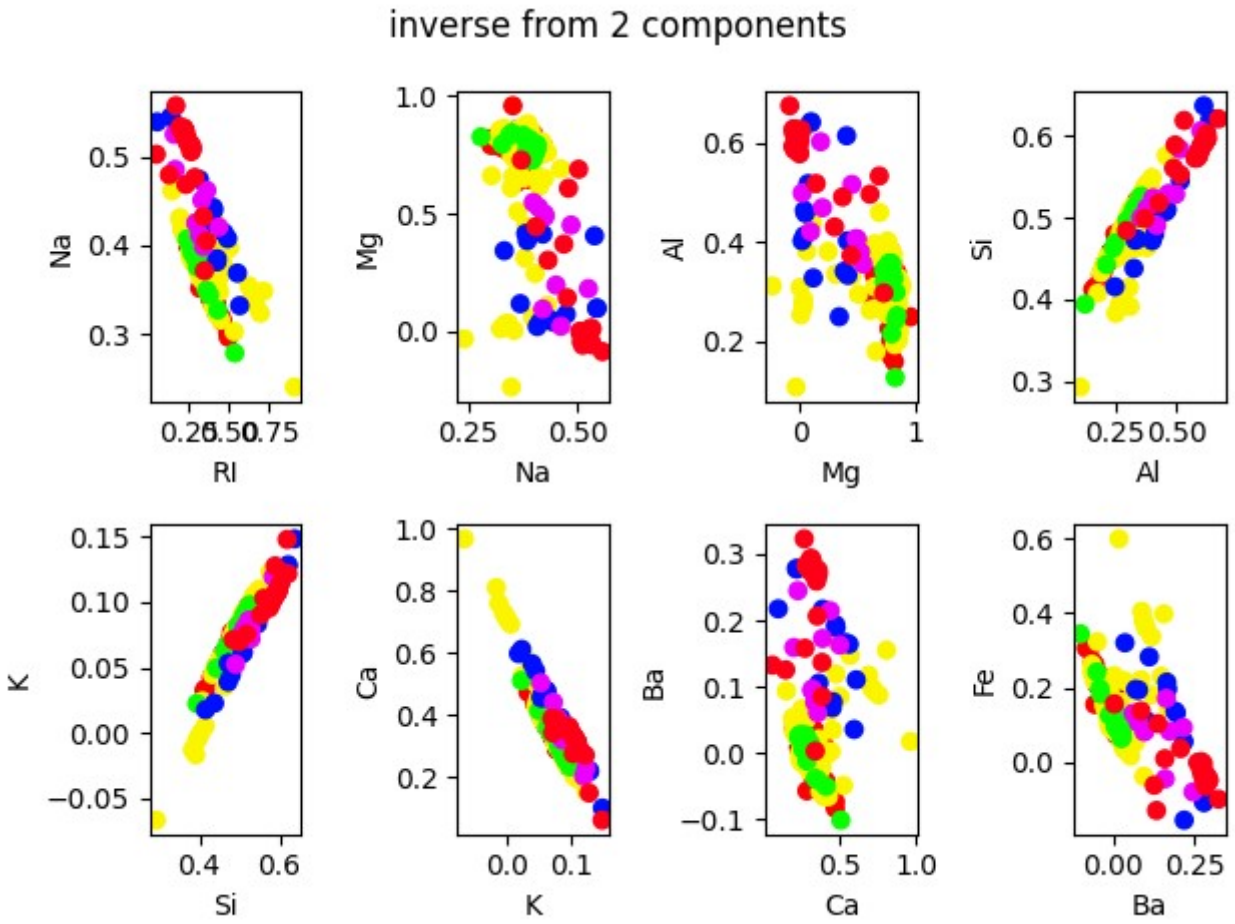

Рисунок 4

Результаты метода главных компонент при различных svd_solver, продемонстрированы на рисунках 5-8. Общая объясненная дисперсия и время обработки данных при разных количествах компонент:

2 auto : ratio 0.634196662104278 in 0.0069963932037353516 s

2 full : ratio 0.634196662104278 in 0.0018379688262939453 s

2 arpack : ratio 0.634196662104278 in 0.4258880615234375 s

2 randomized : ratio 0.6341966621042786 in 0.05556368827819824 s

4 auto : ratio 0.8586697305102718 in 0.0036165714263916016 s

4 full : ratio 0.8586697305102718 in 0.012143373489379883 s

4 arpack : ratio 0.8586697305102717 in 0.009063005447387695 s

4 randomized : ratio 0.8586697305102715 in 0.0039441585540771484 s

6 auto : ratio 0.9694347221994033 in 0.0028450489044189453 s

6 full : ratio 0.9694347221994033 in 0.0011112689971923828 s

6 arpack : ratio 0.9694347221994031 in 0.005422115325927734 s

6 randomized : ratio 0.969434722199403 in 0.0033032894134521484 s

8 auto : ratio 0.9998605862637865 in 0.0024139881134033203 s

8 full : ratio 0.9998605862637865 in 0.0011773109436035156 s

8 arpack : ratio 0.9998605862637866 in 0.005196332931518555 s

8 randomized : ratio 0.9998605862637872 in 0.036295175552368164 s

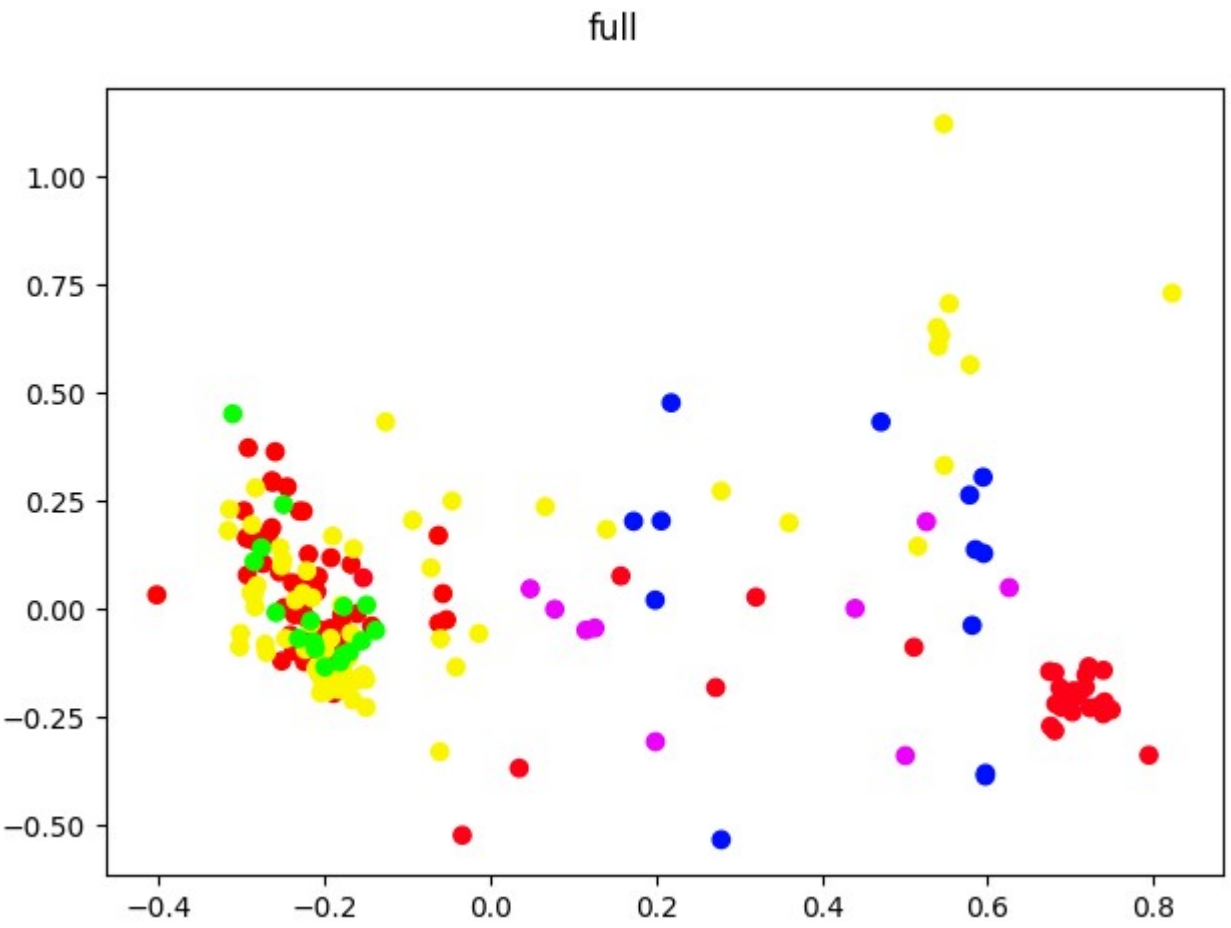

Рисунок 5

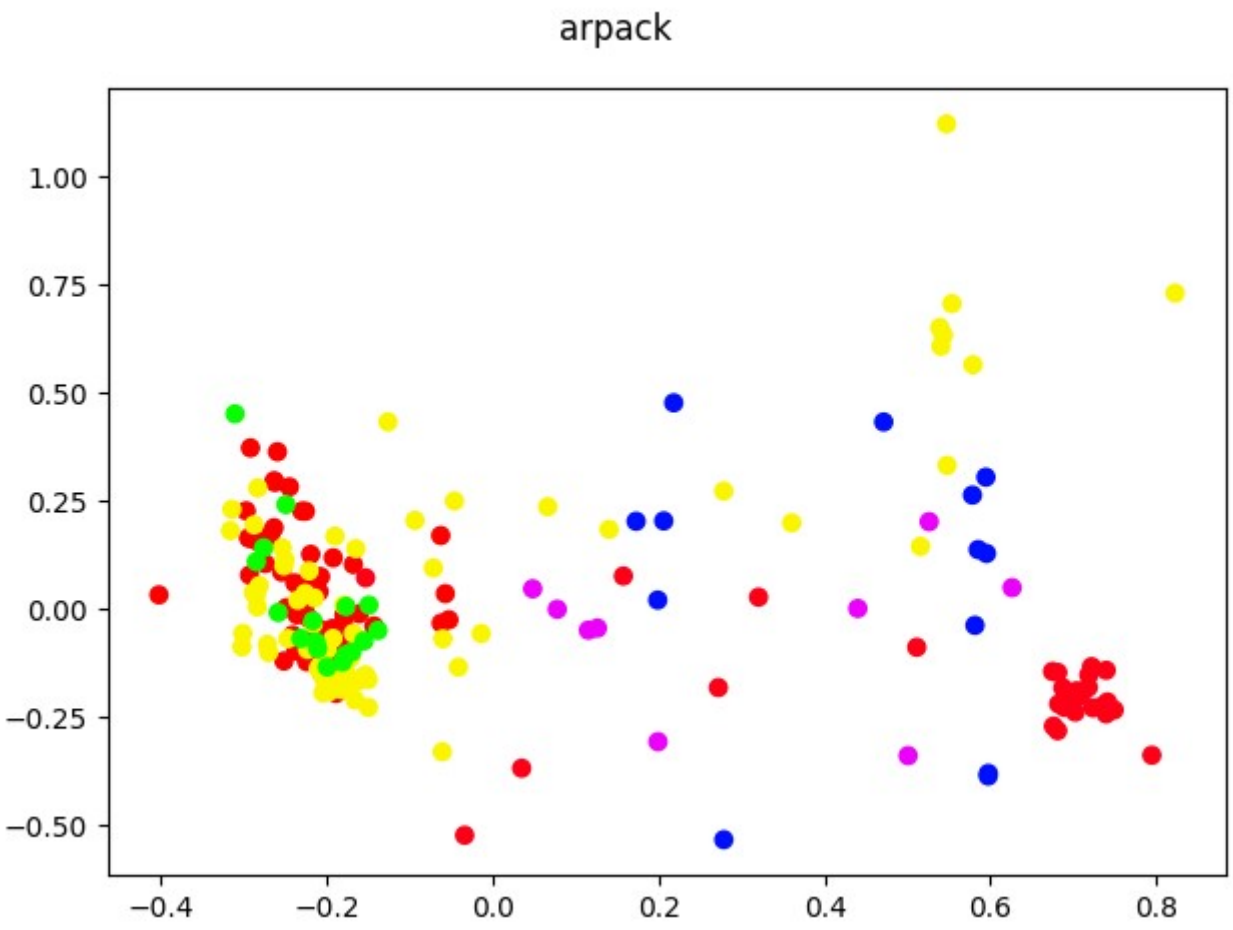

Рисунок 6

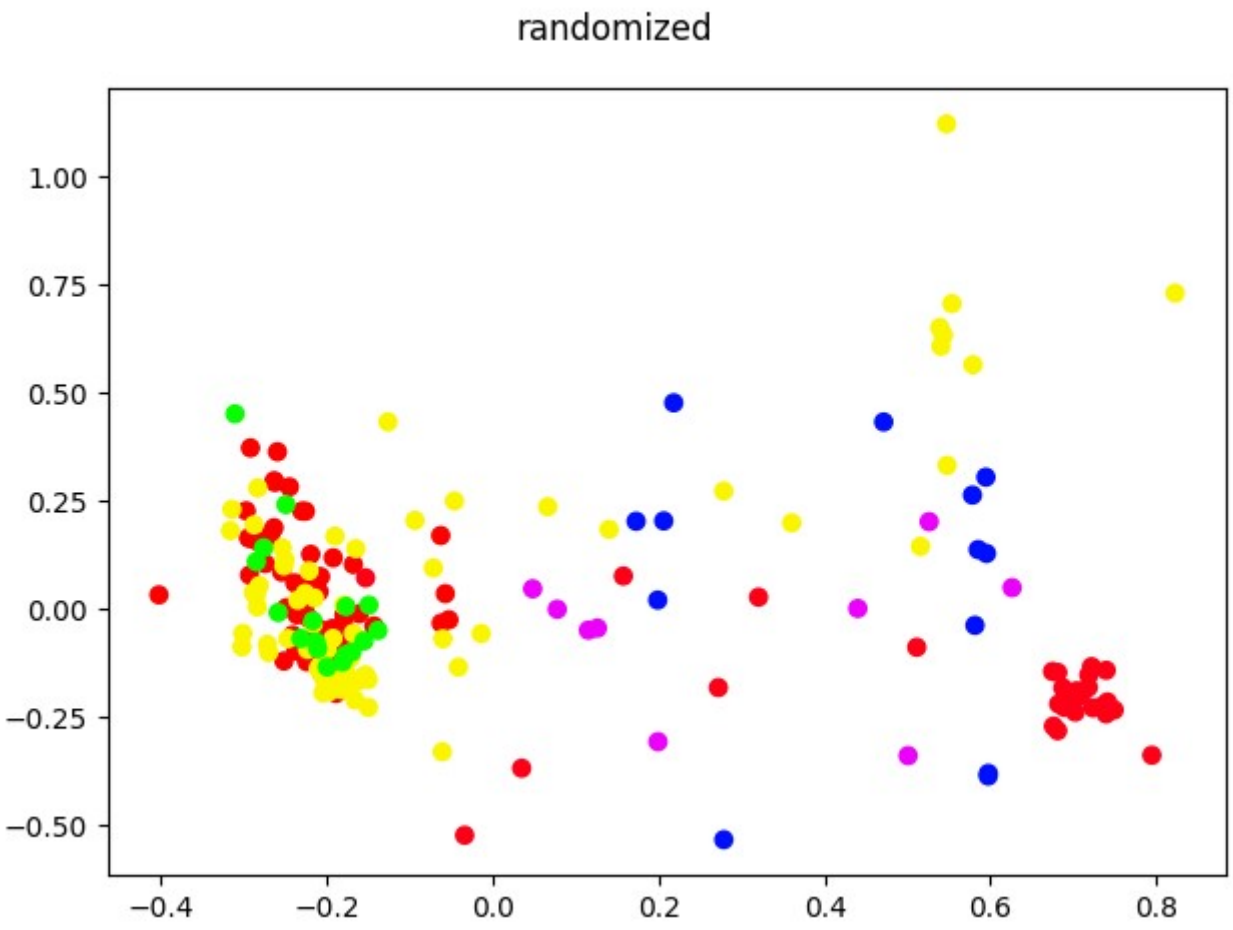

Рисунок 7

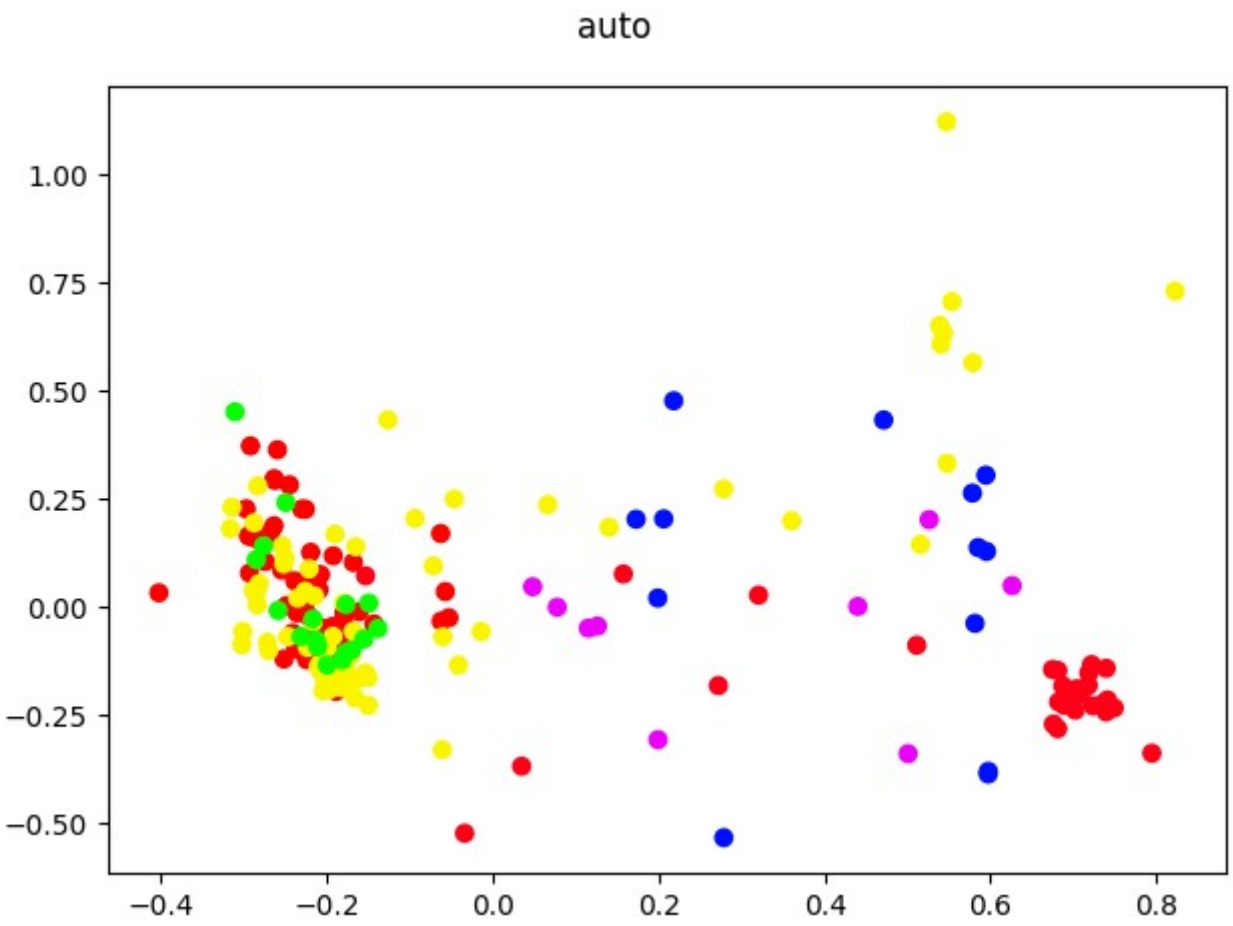

Рисунок 8

Видно, что результаты с разными svd_solver идентичны, отличается лишь время обработки данных.

3. Модификация метода главных компонент

По аналогии исследуем ядерный метода главных компонент для разных ядер. Результаты продемонстрированы на рисунках 9-13. Время работы для разных количеств компонент и ядер:

2 linear : 0.306396484375 s

2 poly : 0.01939535140991211 s

2 rbf : 0.5858860015869141 s

2 sigmoid : 0.03866863250732422 s

2 cosine : 0.011745691299438477 s

4 linear : 0.00761866569519043 s

4 poly : 0.01421666145324707 s

4 rbf : 0.013158321380615234 s

4 sigmoid : 0.015854835510253906 s

4 cosine : 0.009826898574829102 s

6 linear : 0.023562192916870117 s

6 poly : 0.018614530563354492 s

6 rbf : 0.018798828125 s

6 sigmoid : 0.022713184356689453 s

6 cosine : 0.05688667297363281 s

8 linear : 0.010929346084594727 s

8 poly : 0.0176699161529541 s

8 rbf : 0.013443470001220703 s

8 sigmoid : 0.020114421844482422 s

8 cosine : 0.024050474166870117 s

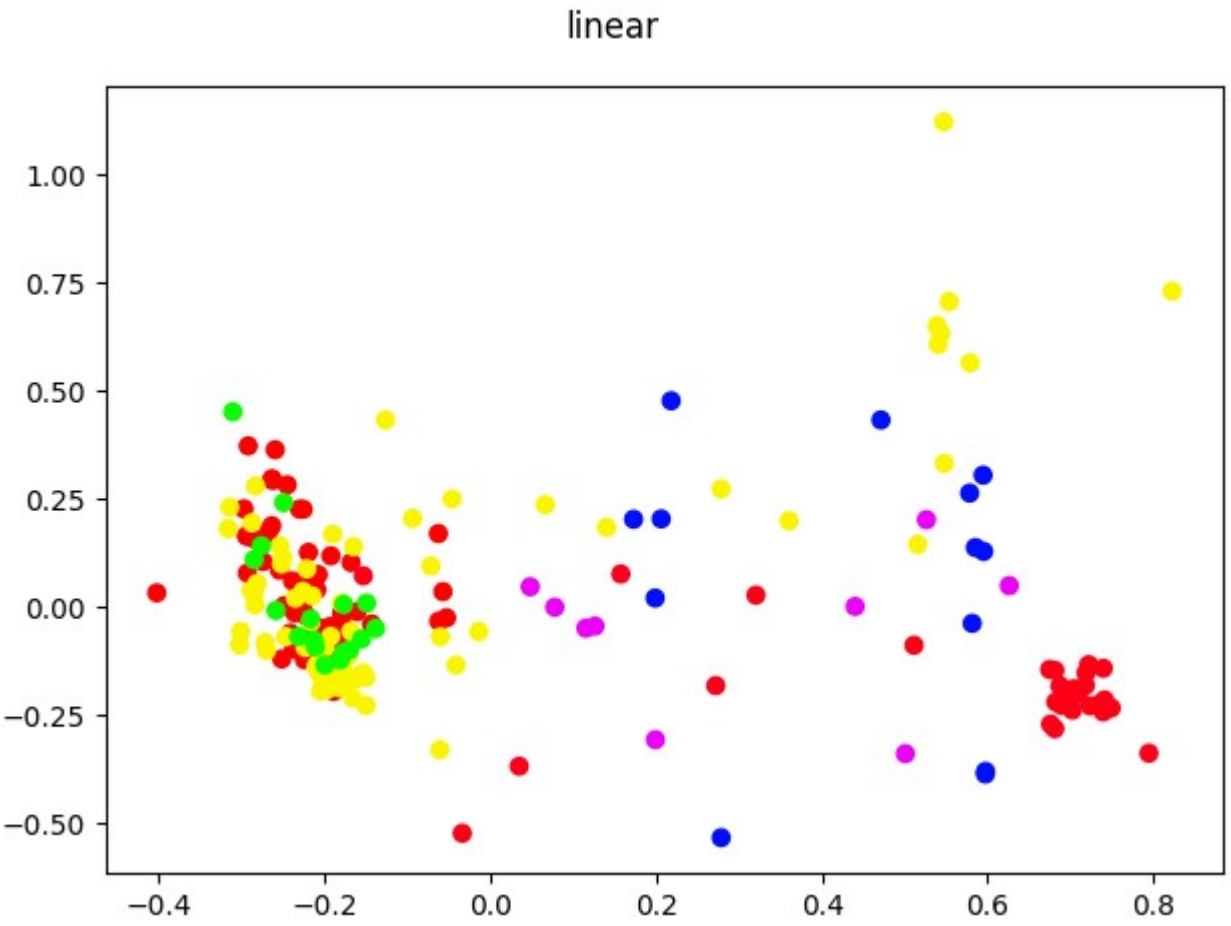

Рисунок 9

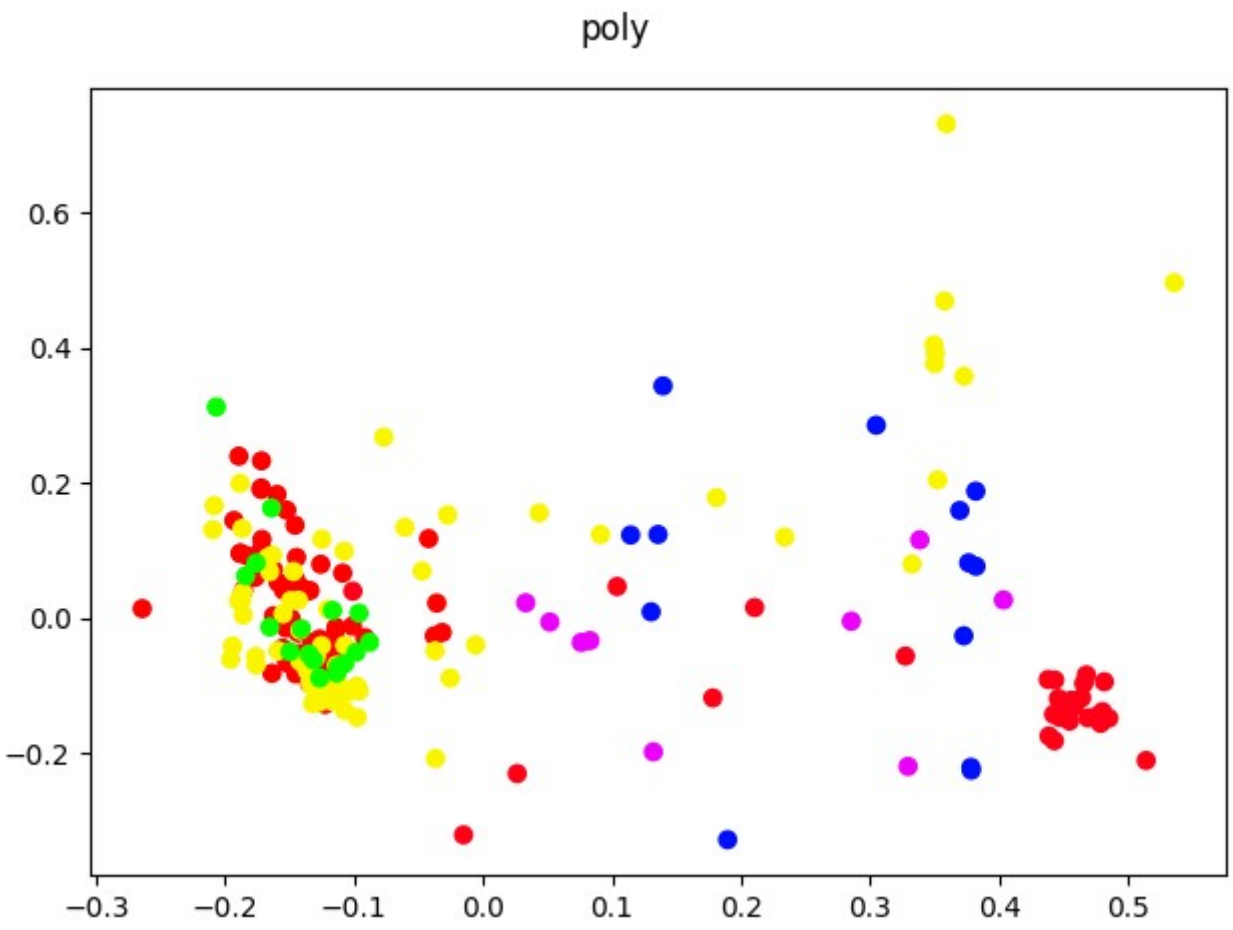

Рисунок 10

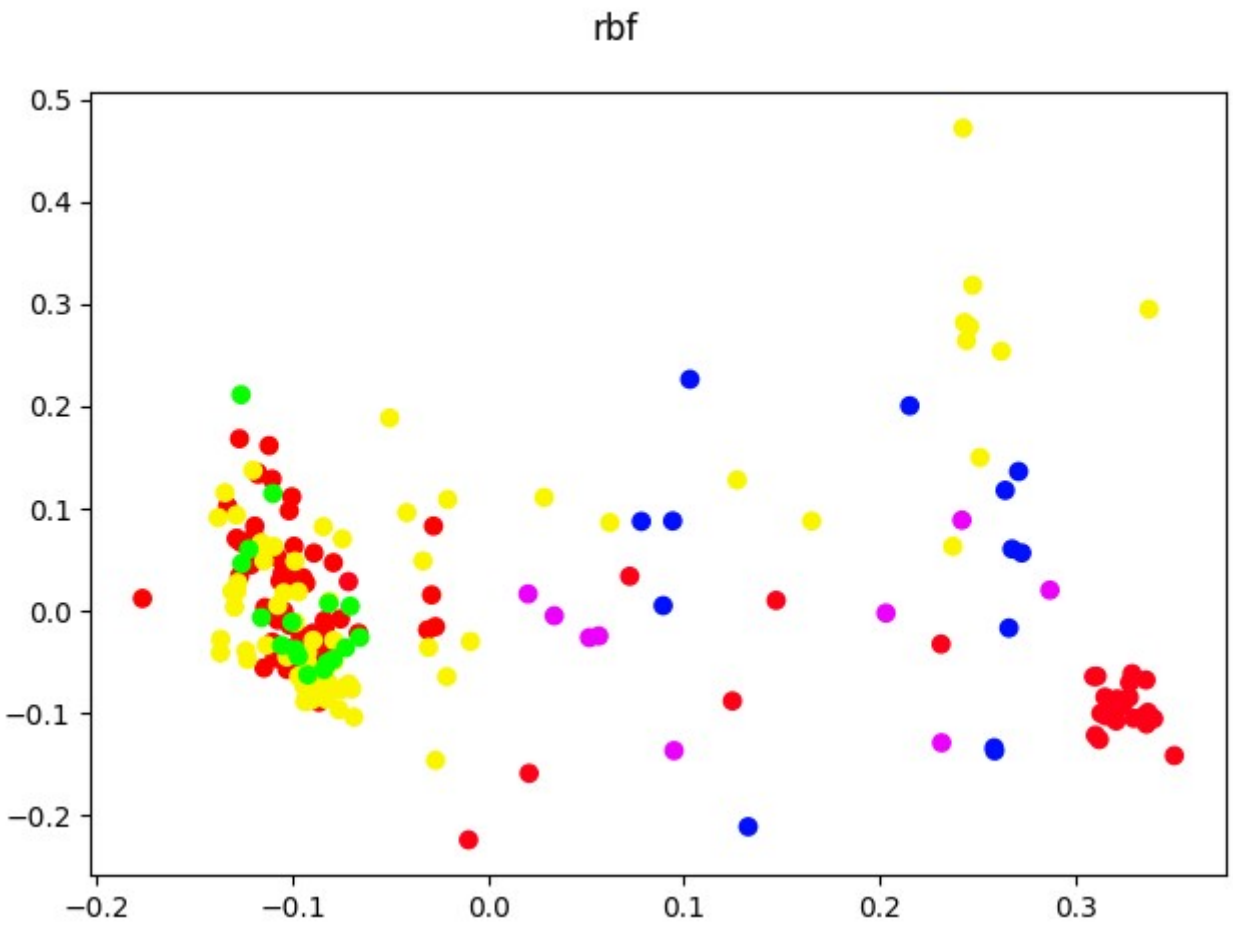

Рисунок 11

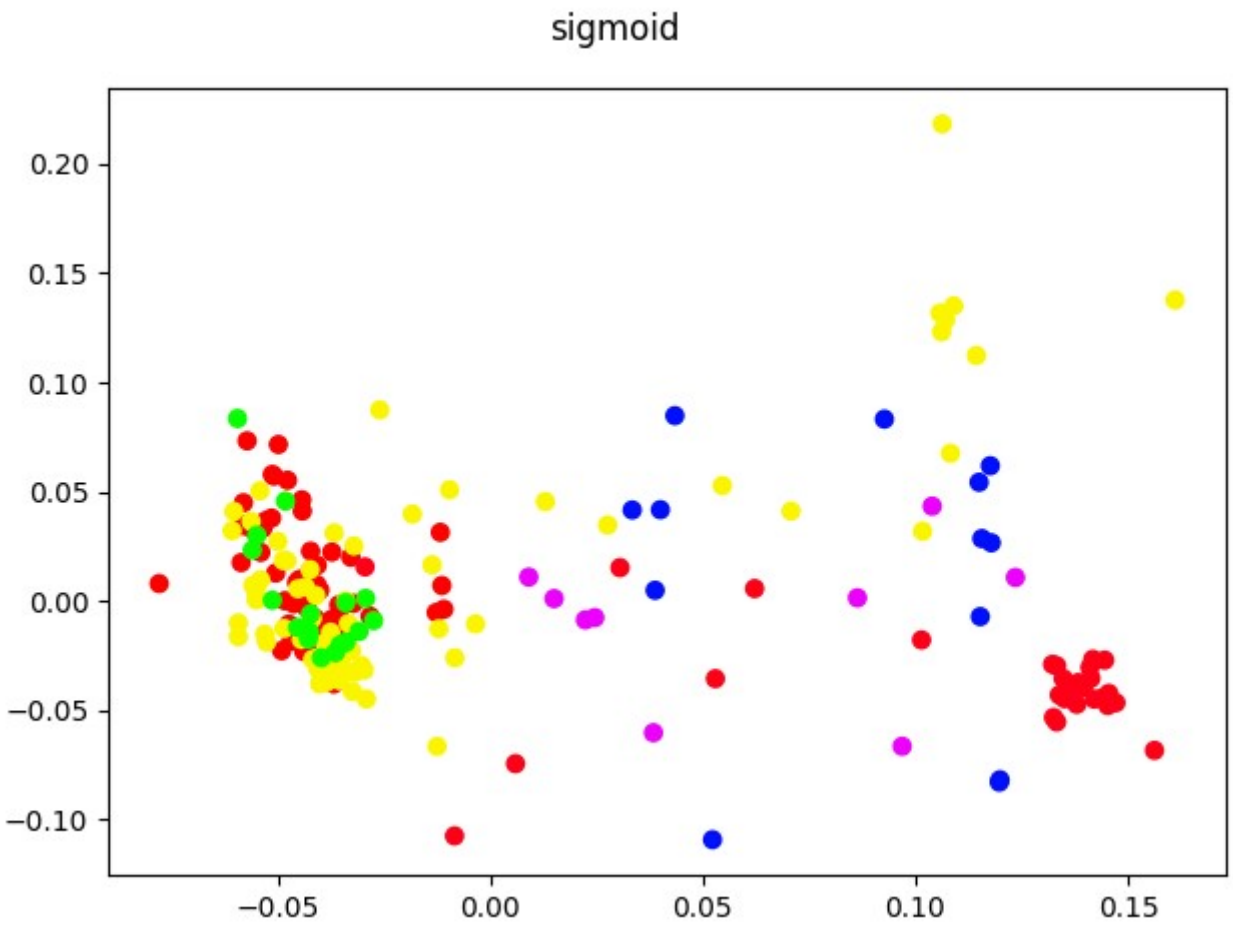

Рисунок 12

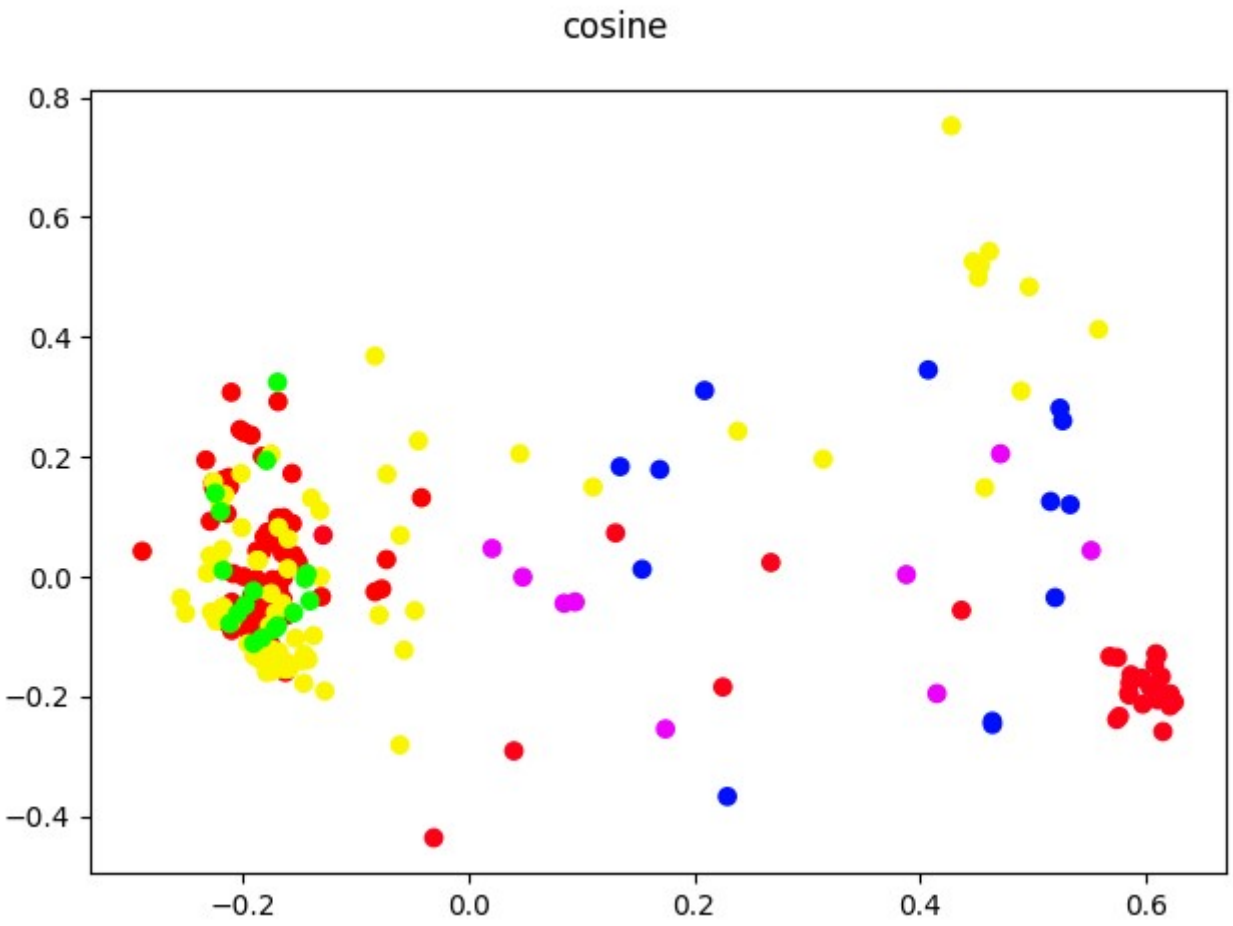

Рисунок 13

По рисунку видно, что ядерный метод главных компонент я линейным ядром работа так же как и метод главных компонент. Объясненная дисперсия примерно одинакова для всех ядер.

Аналогично исследуем разреженный метод главных компонент. Результаты работы продемонстрированы на рисунках 14-15. Время работы при разных количествах компонент и методов:

2 lars : 0.38828372955322266 s

2 cd : 0.12188553810119629 s

4 lars : 0.05005359649658203 s

4 cd : 0.01860809326171875 s

6 lars : 0.1327214241027832 s

6 cd : 0.10652923583984375 s

8 lars : 0.22406959533691406 s

8 cd : 0.09517073631286621 s

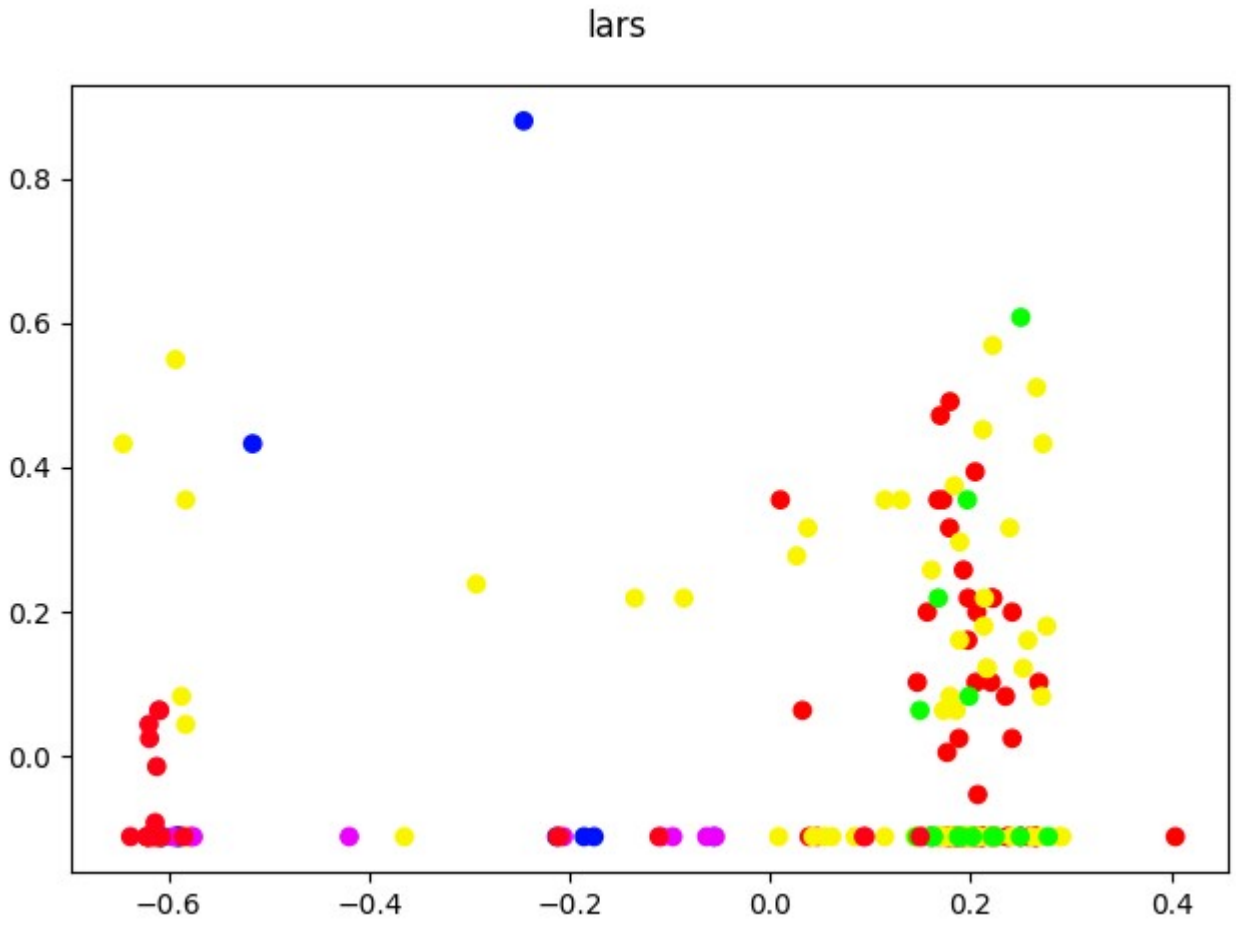

Рисунок 14

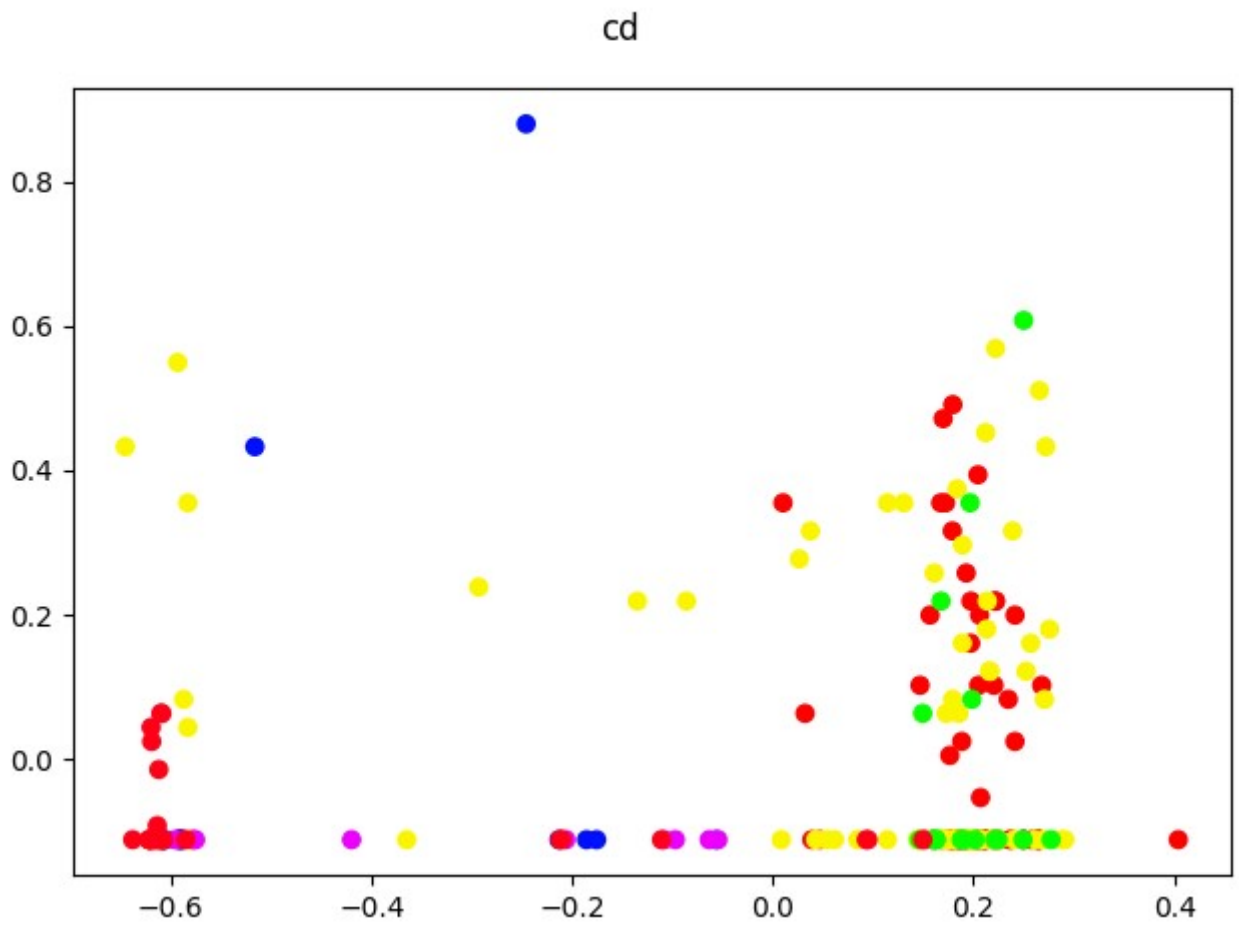

Рисунок 15

Разреженный метод главных компонент находит набор разреженных компонентов, которые могут оптимально восстановить данные. Можно сделать вывод, что этот метод лучше подходит для данных большой размерности. Для двух компонентов получили следующую матрицу компонентов:

array([[0., 0., 0.99804243, -0.03718353, 0., 0., 0., -0.0502861 , 0.],

[0., 0., 0., 0., 0., 0., 0., 0., 1.]])

4. Факторный анализ

Проведем понижение размерности используя факторный анализ. Результат работы представлен на рисунках 16-17. Время выполнения при разных количествах компонент и методах:

2 lapack : 0.32977724075317383 s

2 randomized : 0.4400608539581299 s

4 lapack : 0.10771560668945312 s

4 randomized : 0.49739980697631836 s

6 lapack : 0.07892274856567383 s

6 randomized : 0.3307662010192871 s

8 lapack : 0.03267312049865723 s

8 randomized : 0.12043333053588867 s

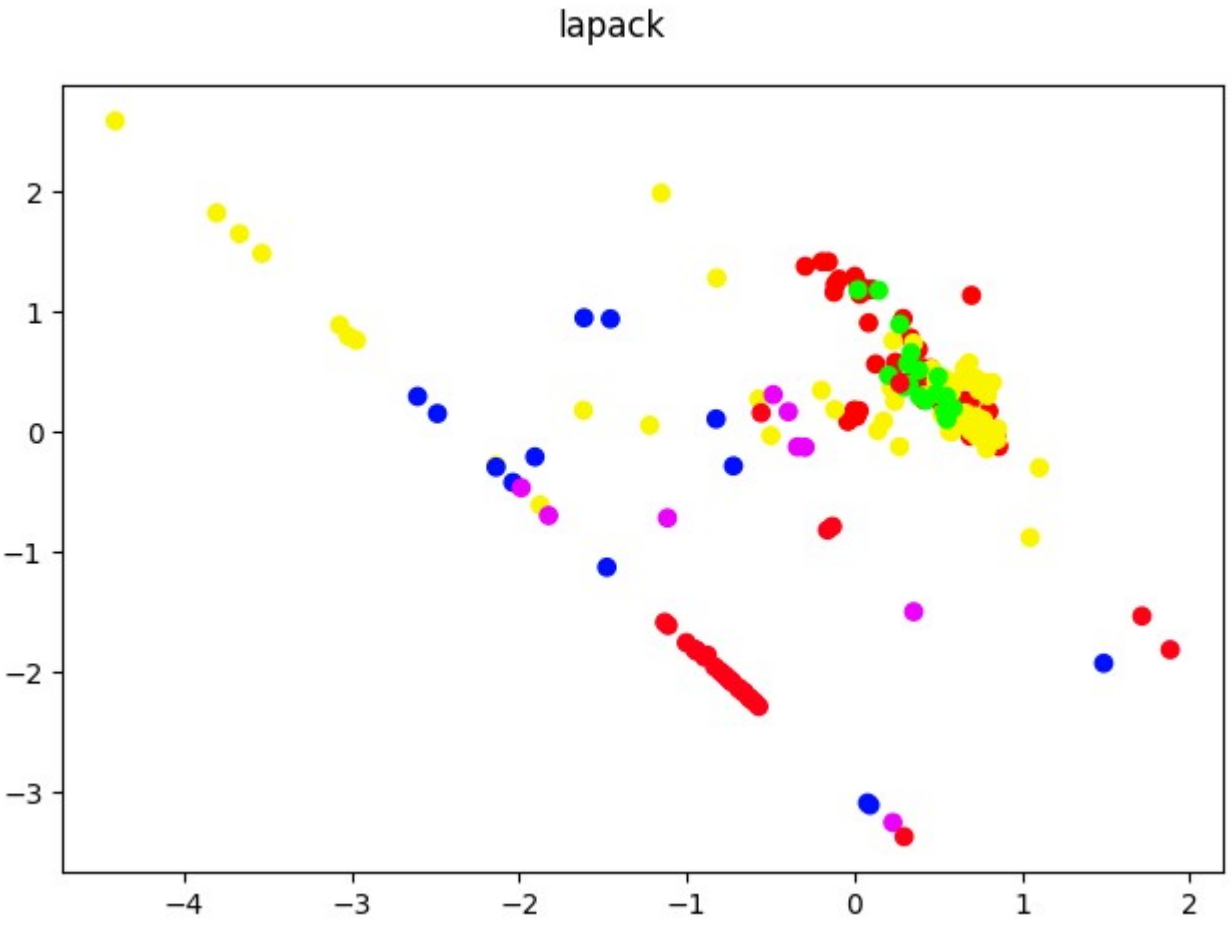

Рисунок 15

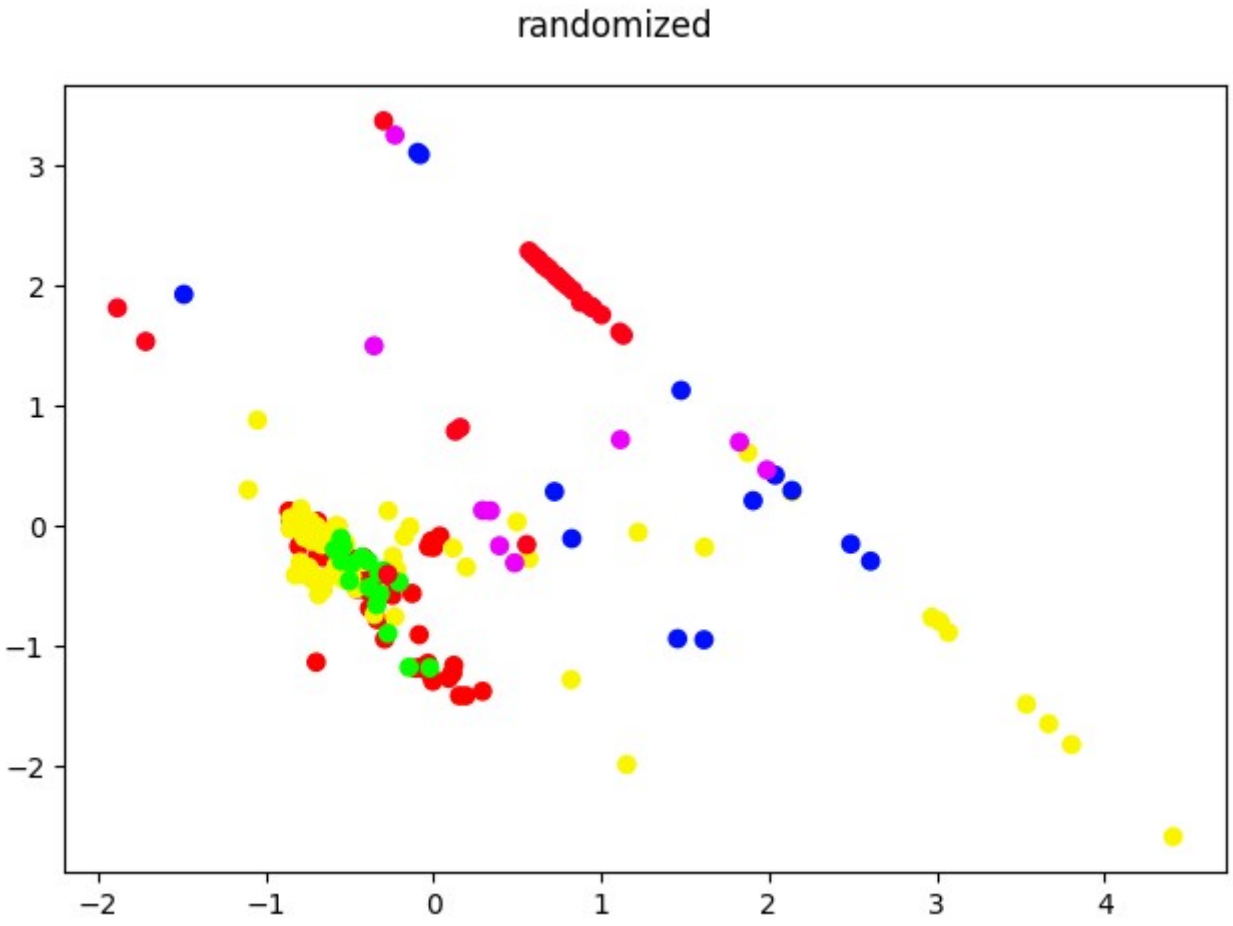

Рисунок 16

Факторный анализ предполагает, что наблюдаемые данные зависят от меньшего количества неизвестных переменных и случайной ошибки. Анализ выполняет минимизацию корреляции между исходными переменными и факторами.

Факторный анализ лучше подходит для поиска скрытых зависимостей, в то время как PCA — лучше для уменьшения размерности.

Вывод

Были получены навыки работы с методами понижения размерности данных из библиотеки Scikit Learn, такими как метод главных компонент (в том числе модифицированные) и факторный анализ.