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. 6302 _____ Барбарич И.Г.

Руководитель _____ Жангиров Т. Р.

Санкт-Петербург

2020

Цель работы

Ознакомиться с методами предобработки данных из библиотеки Scikit Learn

1. **Загрузка данных**

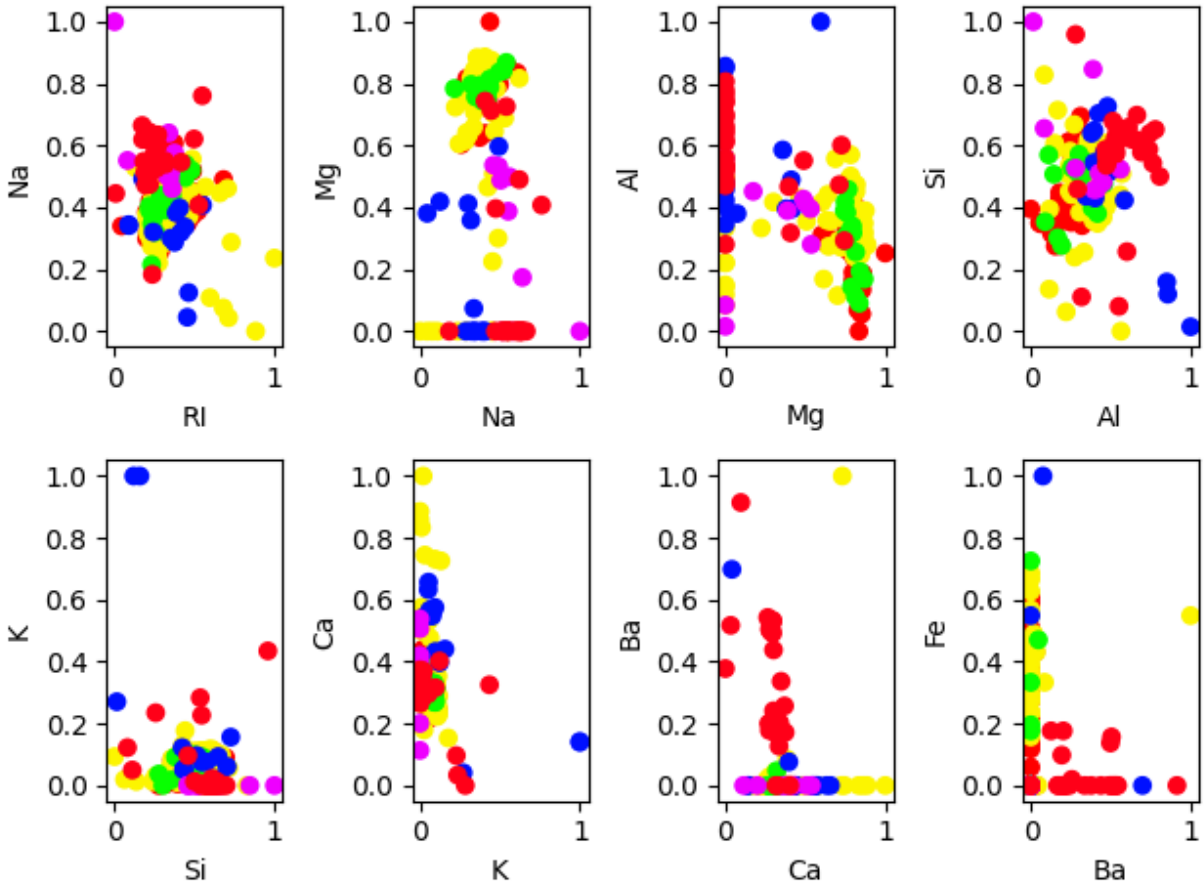


Рисунок 1. Диаграммы рассеивания для пар признаков.

Чтобы определить соответствие цвета на диаграмме и класса в датасете, необходимо воспользоваться функцией scatter, но использовать метки классов (label).

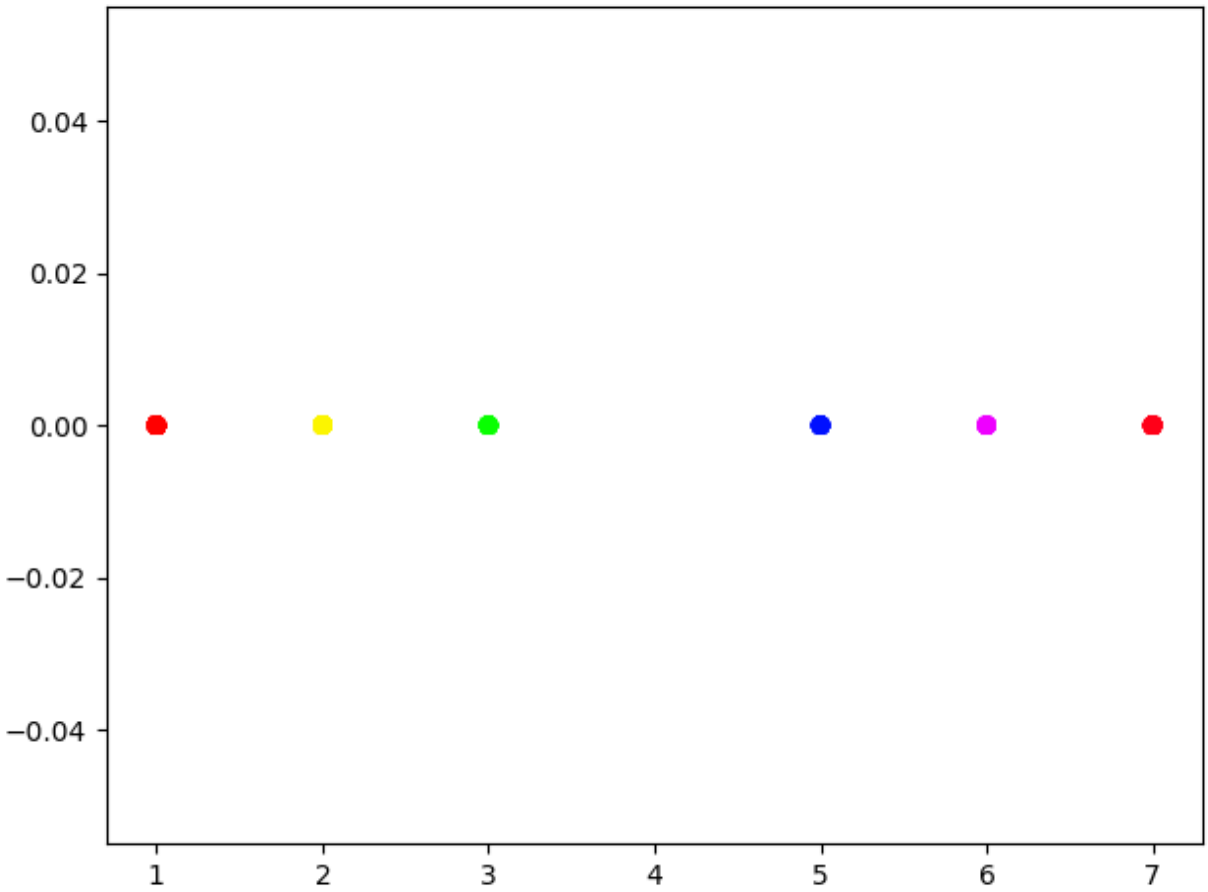


Рисунок 2. Соответствие цвета на диаграмме и класса в датасете.

2. Метод главных компонент

1. Используя метод главных компонент (PCA). Проведите понижение размерности пространства до размерности 2.

```
pca = PCA(n_components = 2)
pca_data = pca.fit(data).transform(data)
```

Рисунок 3.

2. Выведите значение объясненной дисперсии в процентах и собственные числа, соответствующие компонентам.

```
C:\Users\Lion\PycharmProjec
[0.45429569 0.17990097]
[5.1049308  3.21245688]
```

Рисунок 4

3. Постройте диаграмму рассеяния после метода главных компонент

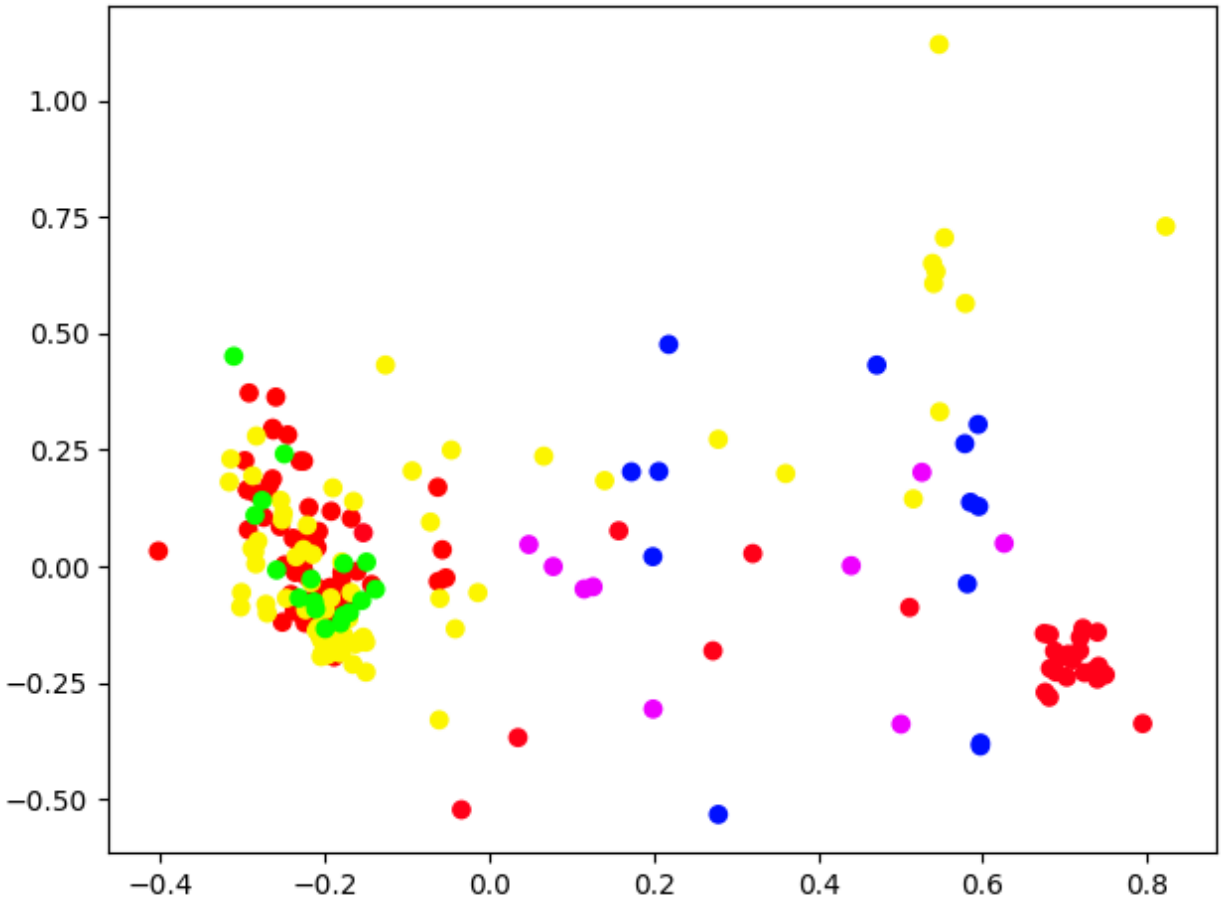


Рисунок 5.

4. Проанализируйте и обоснуйте полученные результаты

Обнаружено скопление данных 1,2,3, 7. Уменьшение линейной размерности с помощью сингулярного разложения данных, чтобы спроецировать их в пространство с более низкой размерностью. Входные параметры центрируются, но не масштабируются для каждого объекта перед применением SVD.

5. Изменяя количество компонент, определите количество, при котором компоненты объясняют не менее 85% дисперсии данных

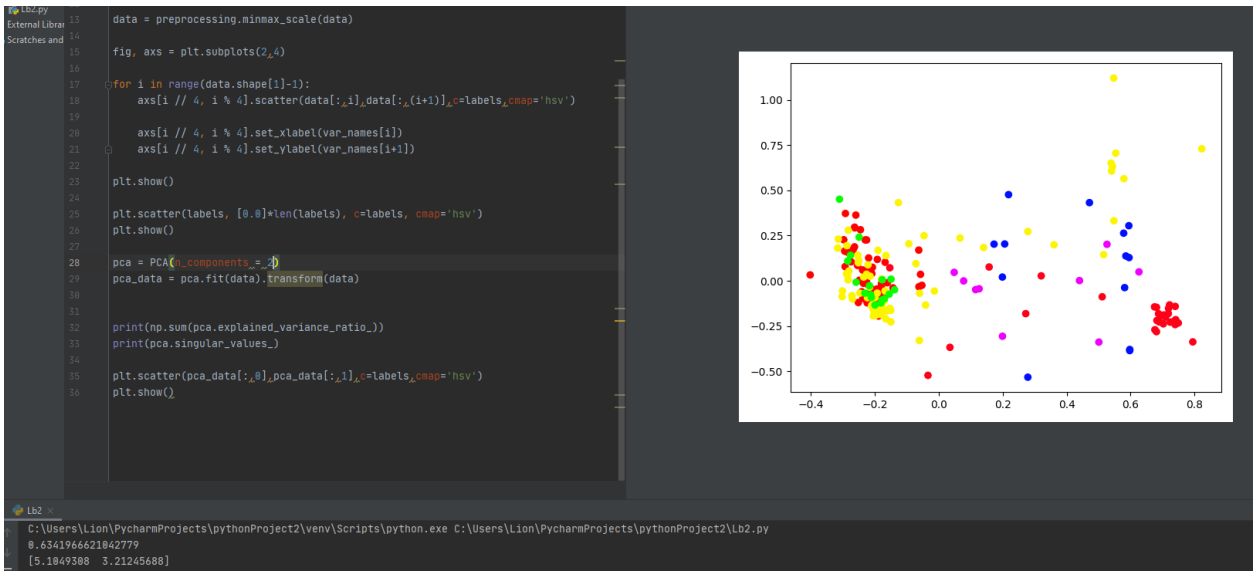


Рисунок 6. При 2 = 0.63

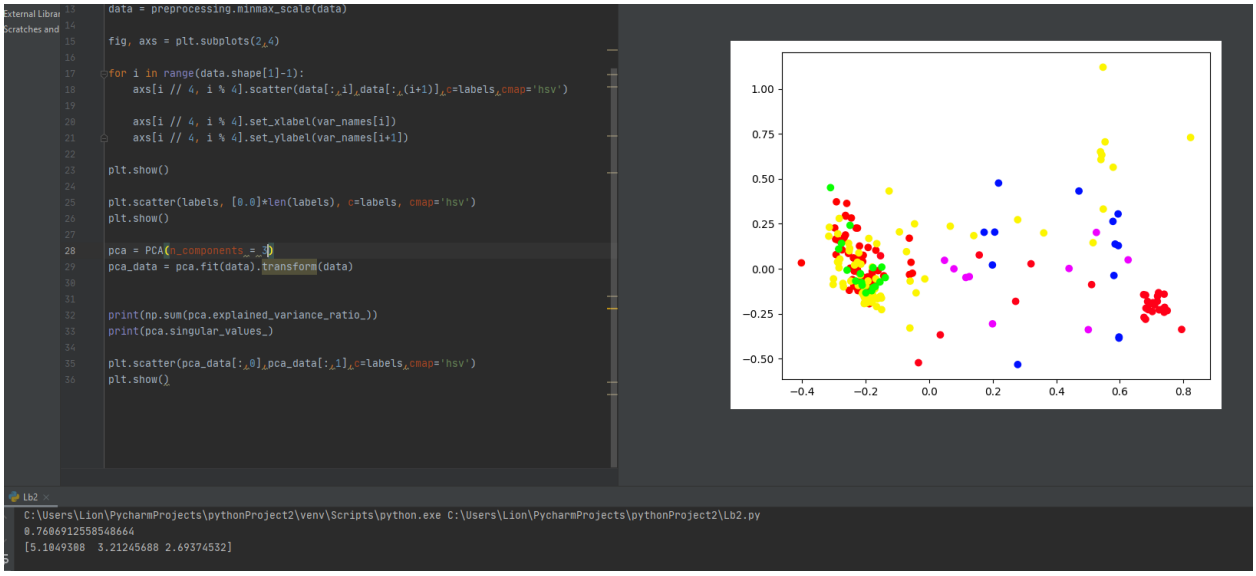


Рисунок 7. При 3 = 0.76

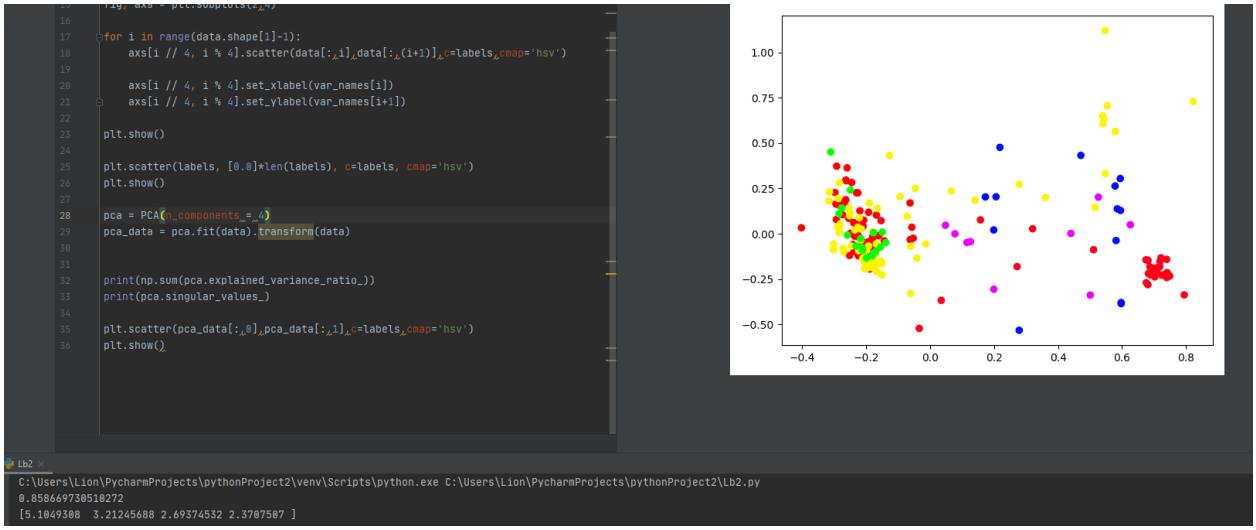


Рисунок 8. При 4 = 0,86

6. Используя метод inverse_transform восстановите данные, сравните с исходными.

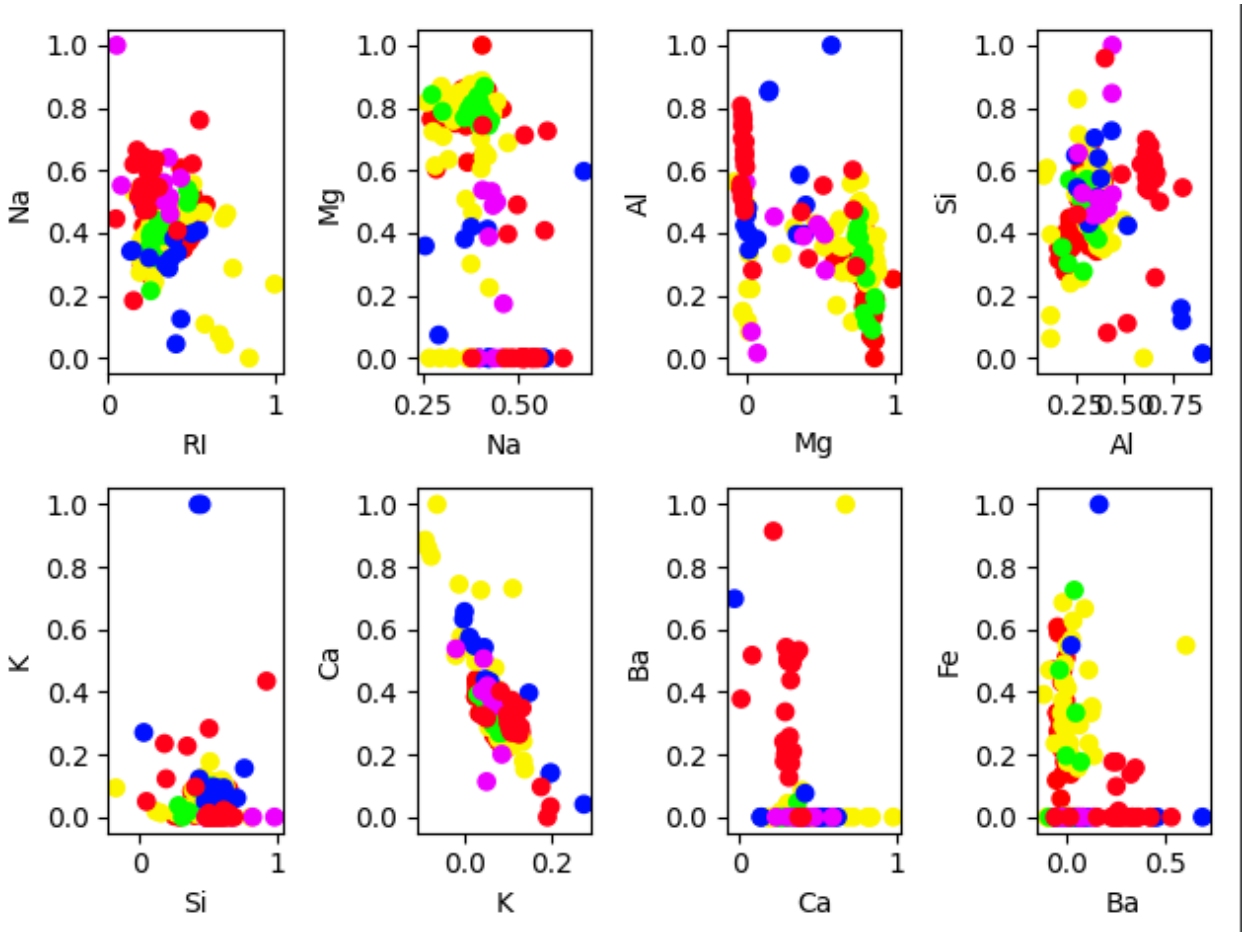


Рисунок 9.

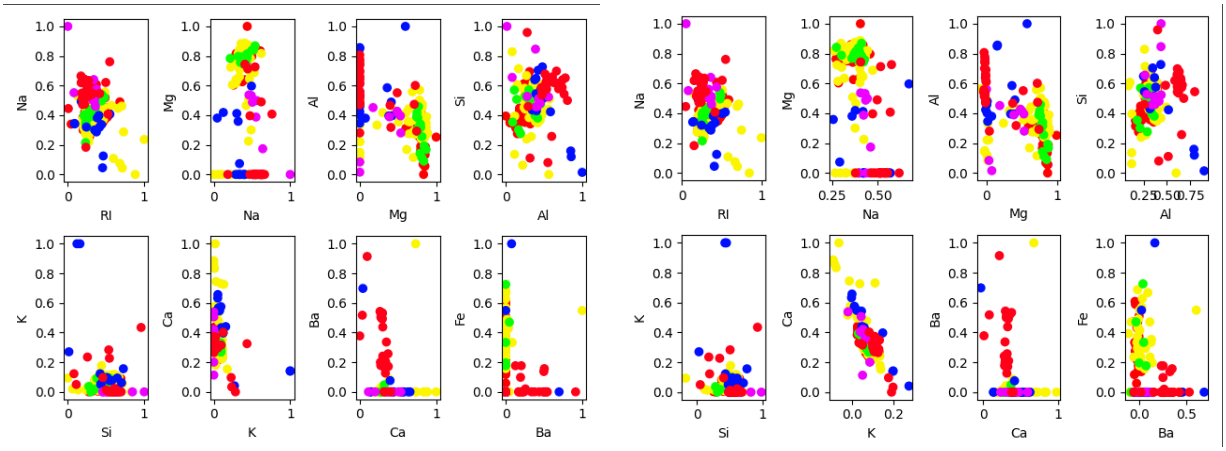


Рисунок 10. Сравнение с исходными. 1 – до, 2 – после.

Данные похожи, но есть отличия, т.к. не были учтены еще 14% дисперсии данных.

7. Исследуйте метод главных компонент при различных параметрах svd_solver

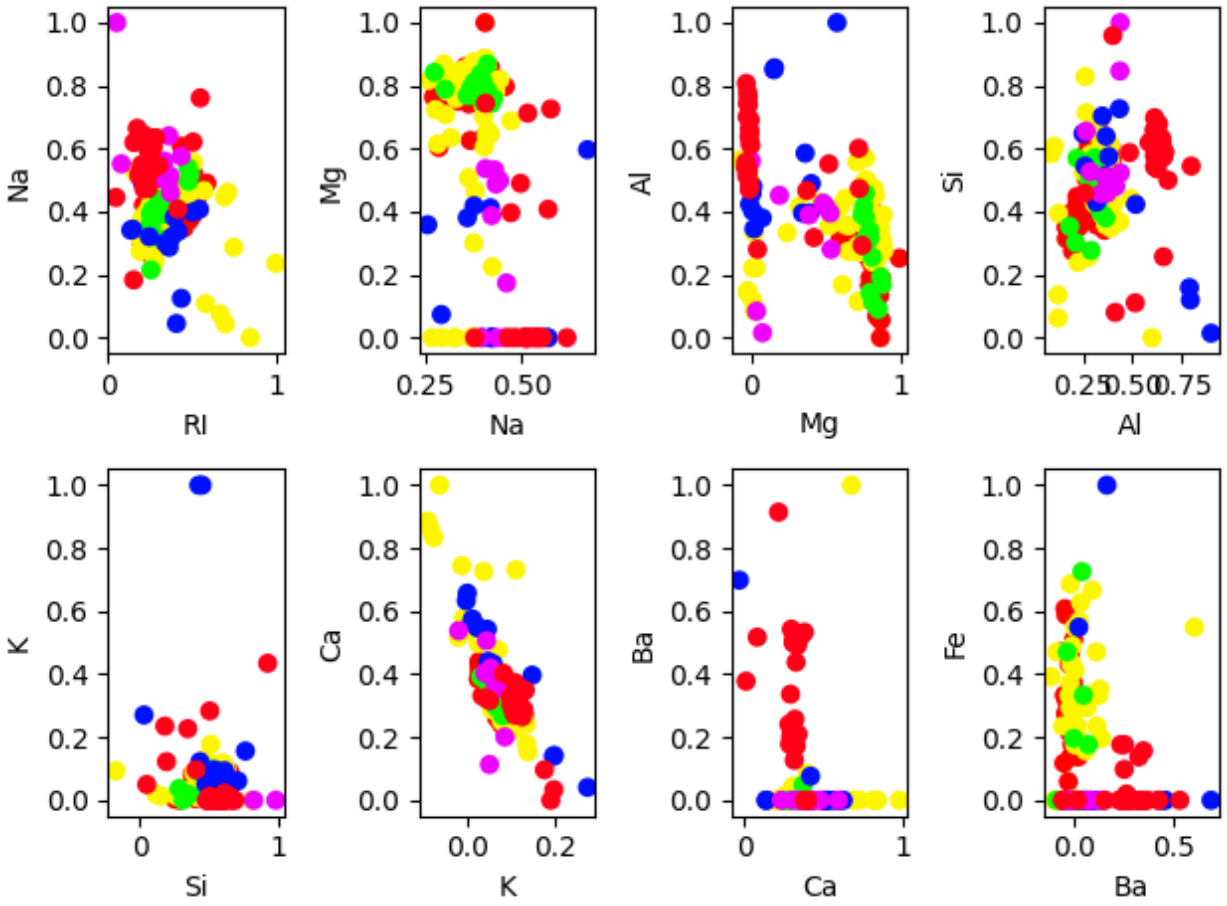
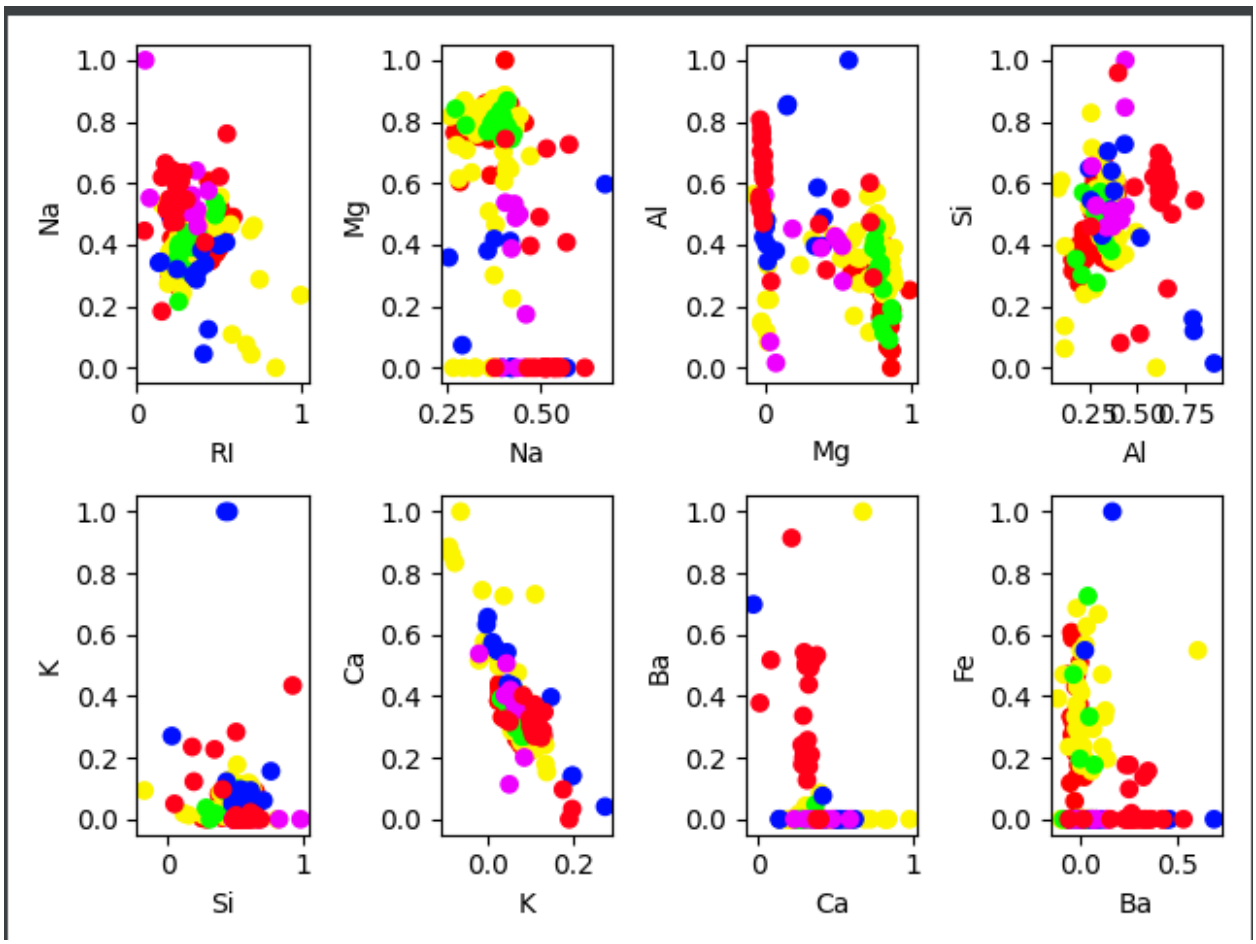
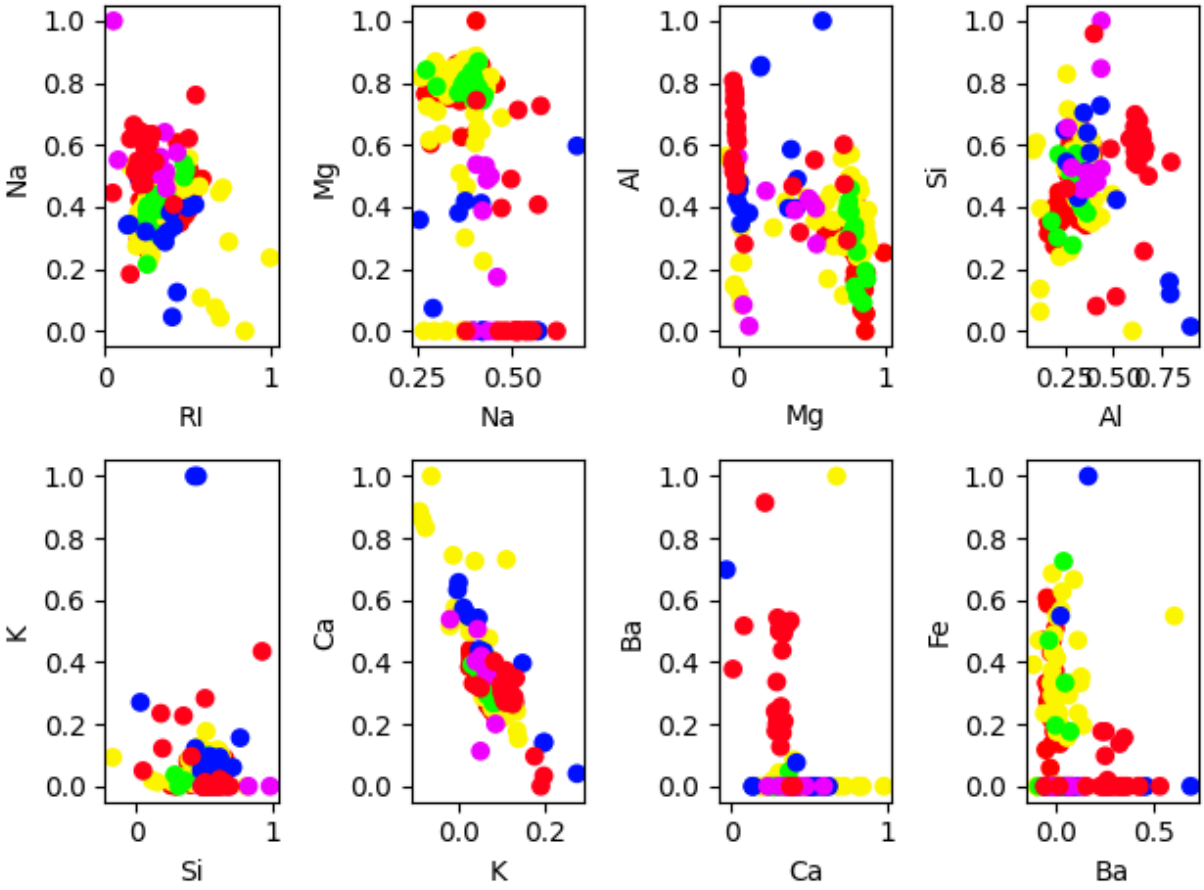


Рисунок 11. svd_solver = ‘auto’



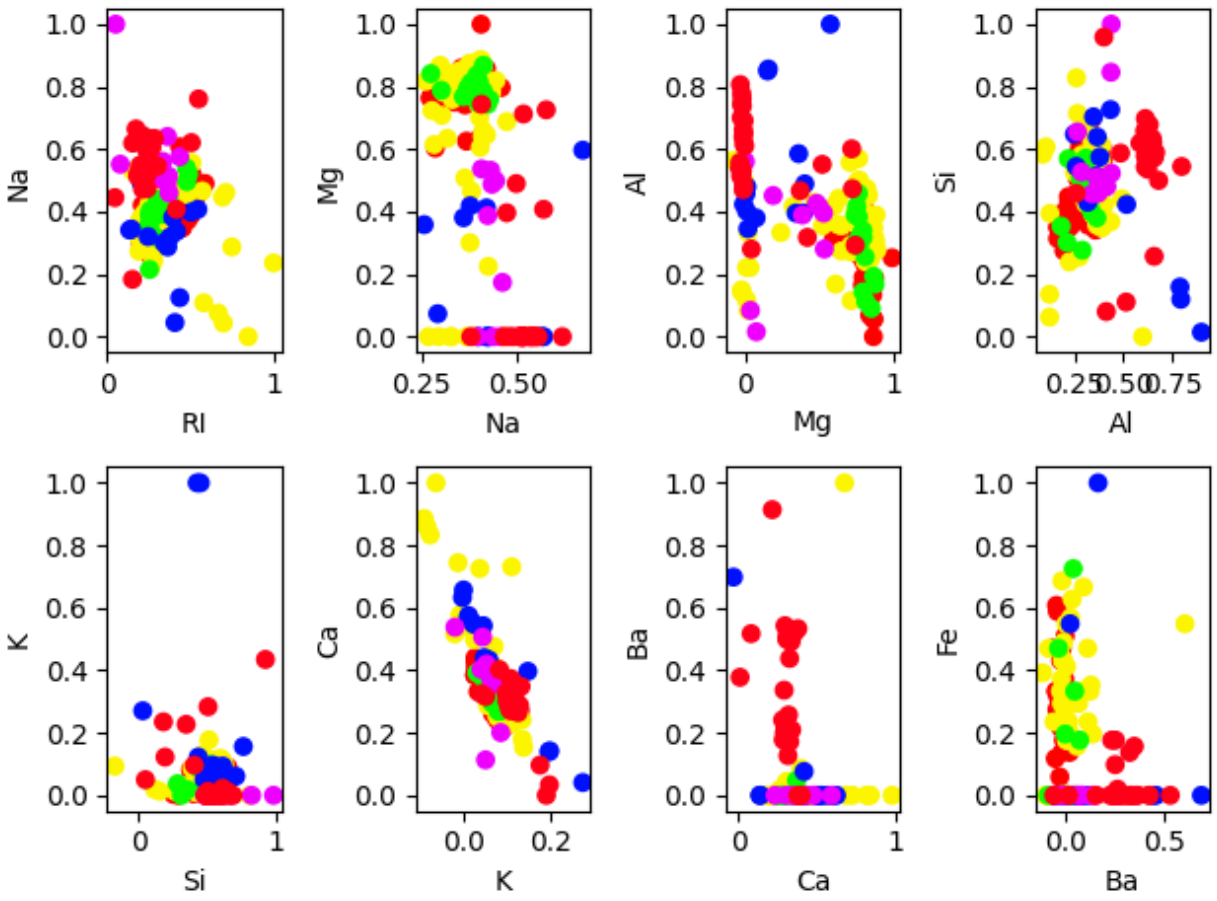

```
0.858669730510272
[5.1049308  3.21245688 2.69374532 2.3707507 ]
```

Рисунок 12. svd_solver = ‘full’




```
0.8586697305102717
[5.1049308  3.21245688 2.69374532 2.3707507 ]
```

Рисунок 13. svd_solver = ‘arpack’




```
0.8586697305102718
[5.1049308  3.21245688 2.69374532 2.3707507 ]
```

Рисунок 14. svd_solver = ‘randomized’

Разницы в результате не замечено. Но т.к. этот параметр отвечает за способ вычисления, есть вероятность, что различается скорость вычислений.

Модификации метода главных компонент

1. По аналогии с PCA исследуйте KernelPCA для различных параметров kernel и различных параметрах для ядра.

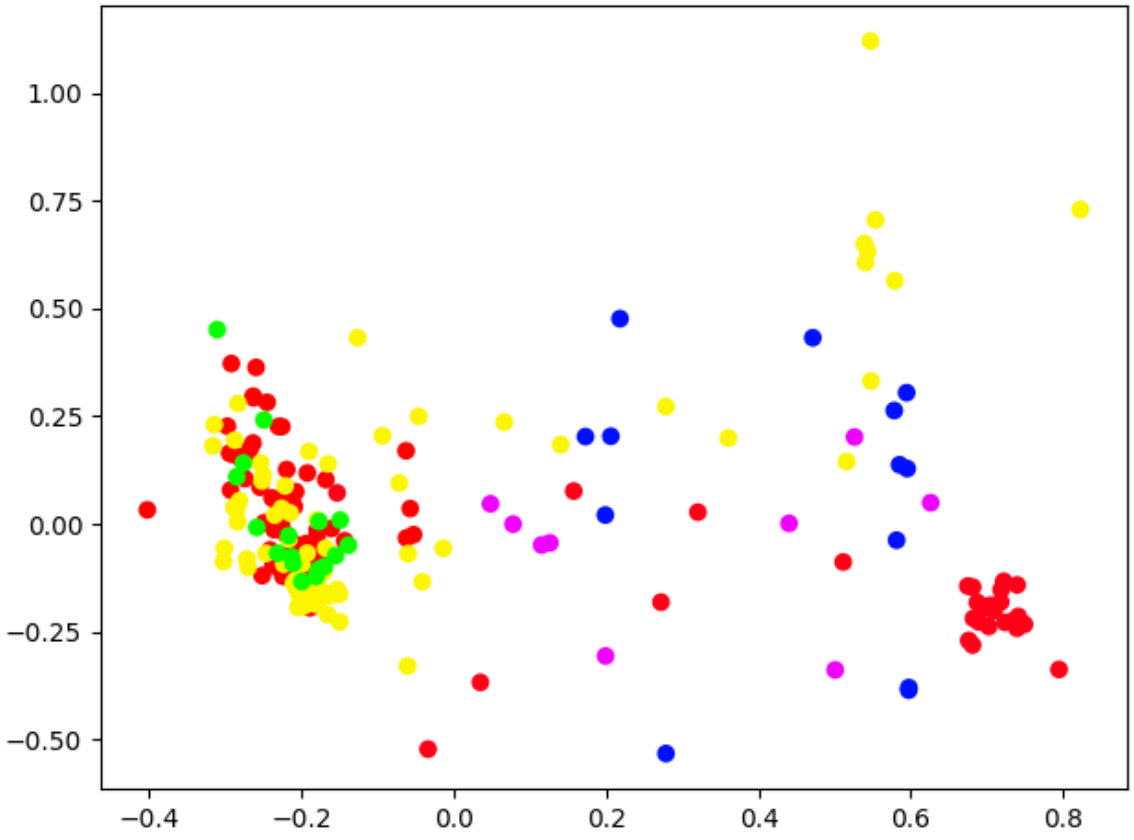


Рисунок 15. linear

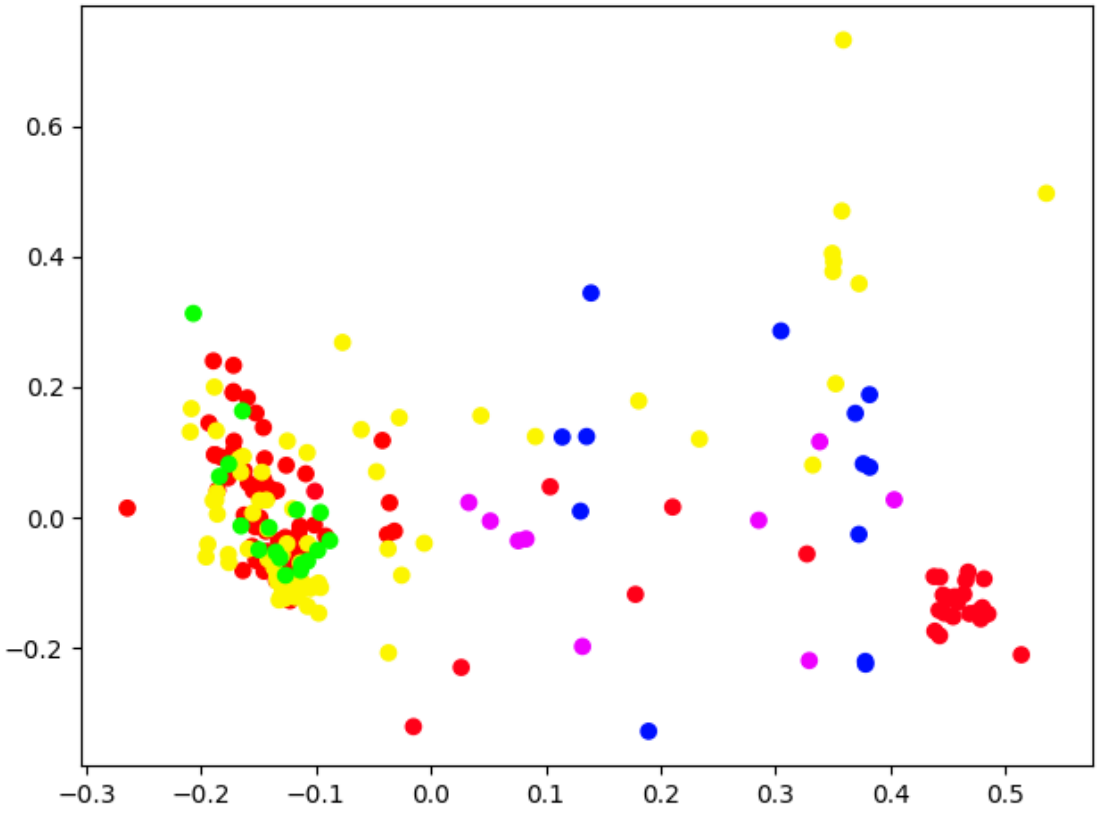


Рисунок 16. poly

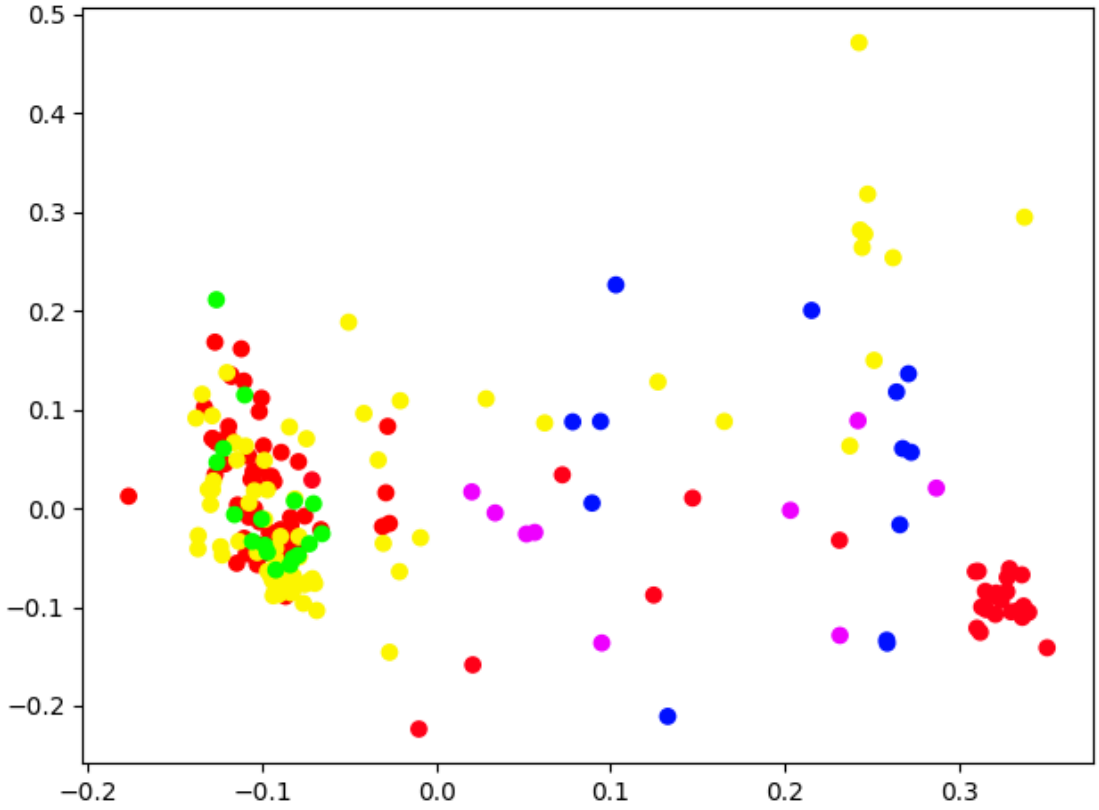


Рисунок 17. rbf

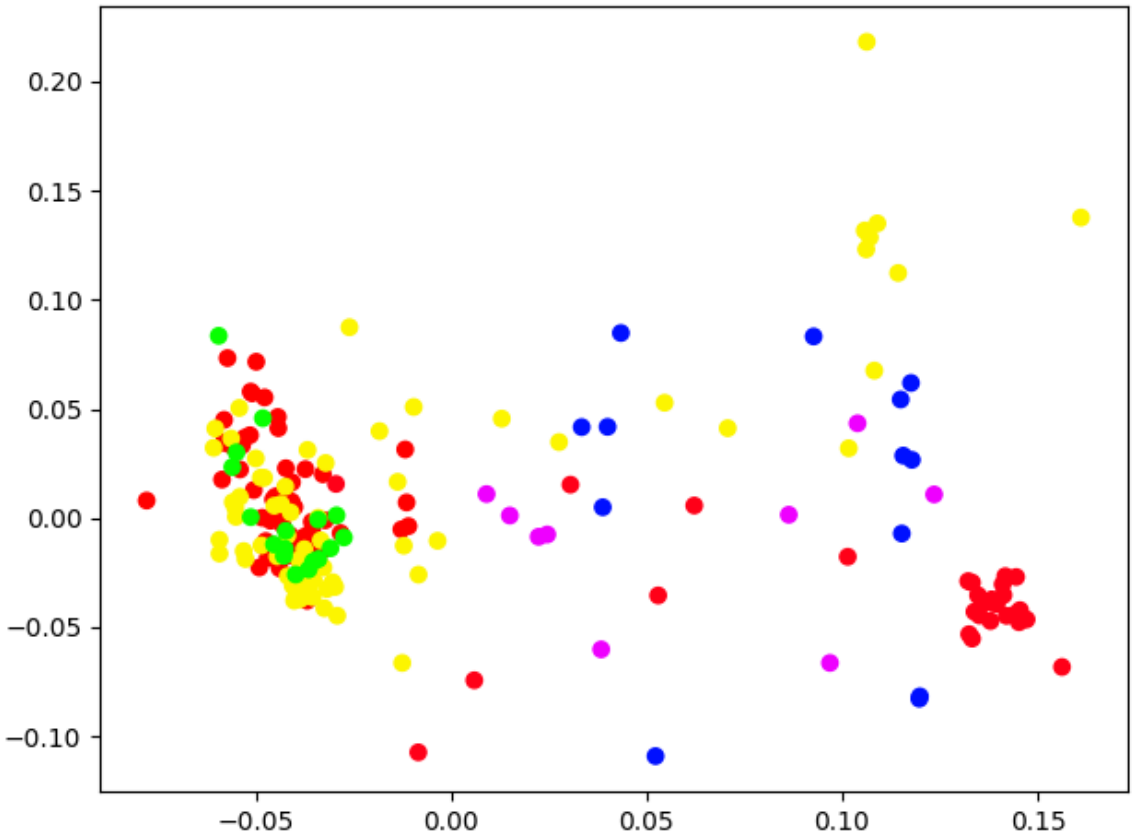


Рисунок 18. sigmoid

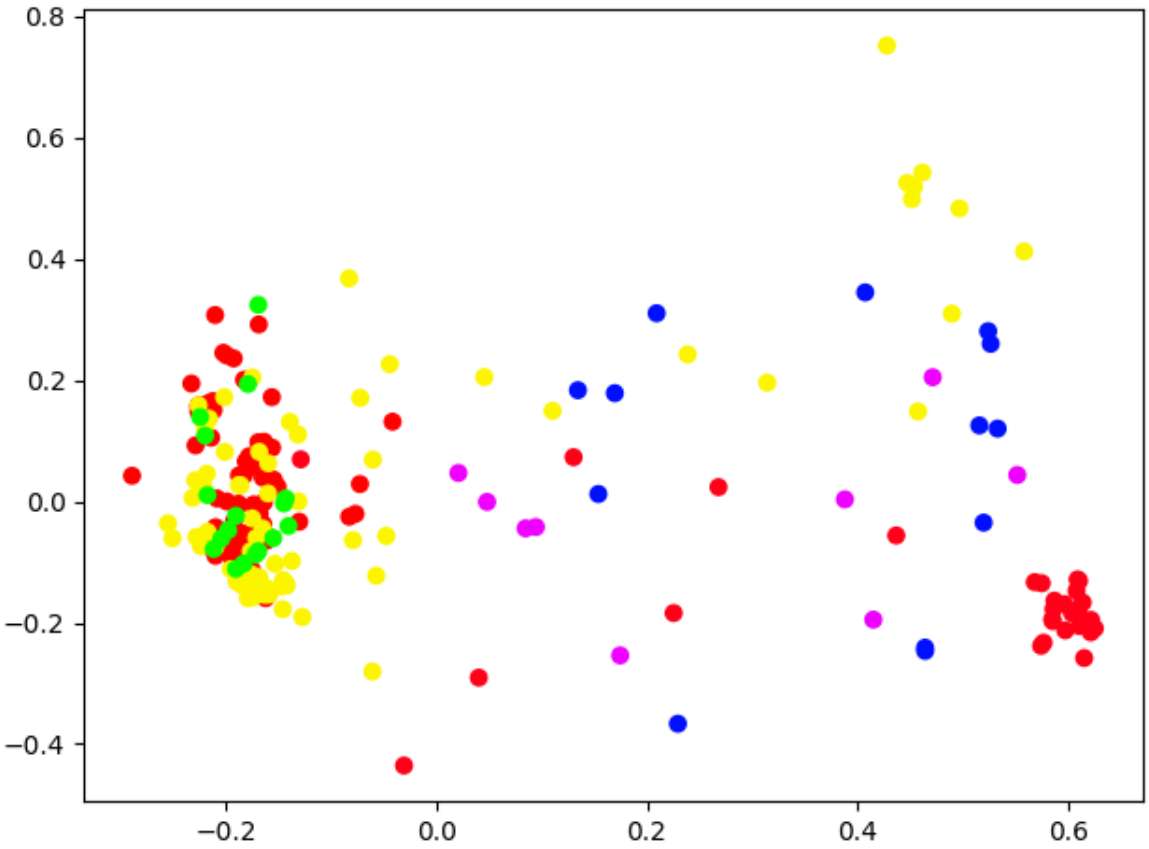


Рисунок 19. cosine

Отличие – масштаб и центрирования.

2. Определите, при каких параметрах KernelPCA работает также как PCA.

PCA – линейной преобразование, поэтому на графике при параметре linear будут одинаковы с KPCA.

3. Аналогично исследуйте SparcePCA.

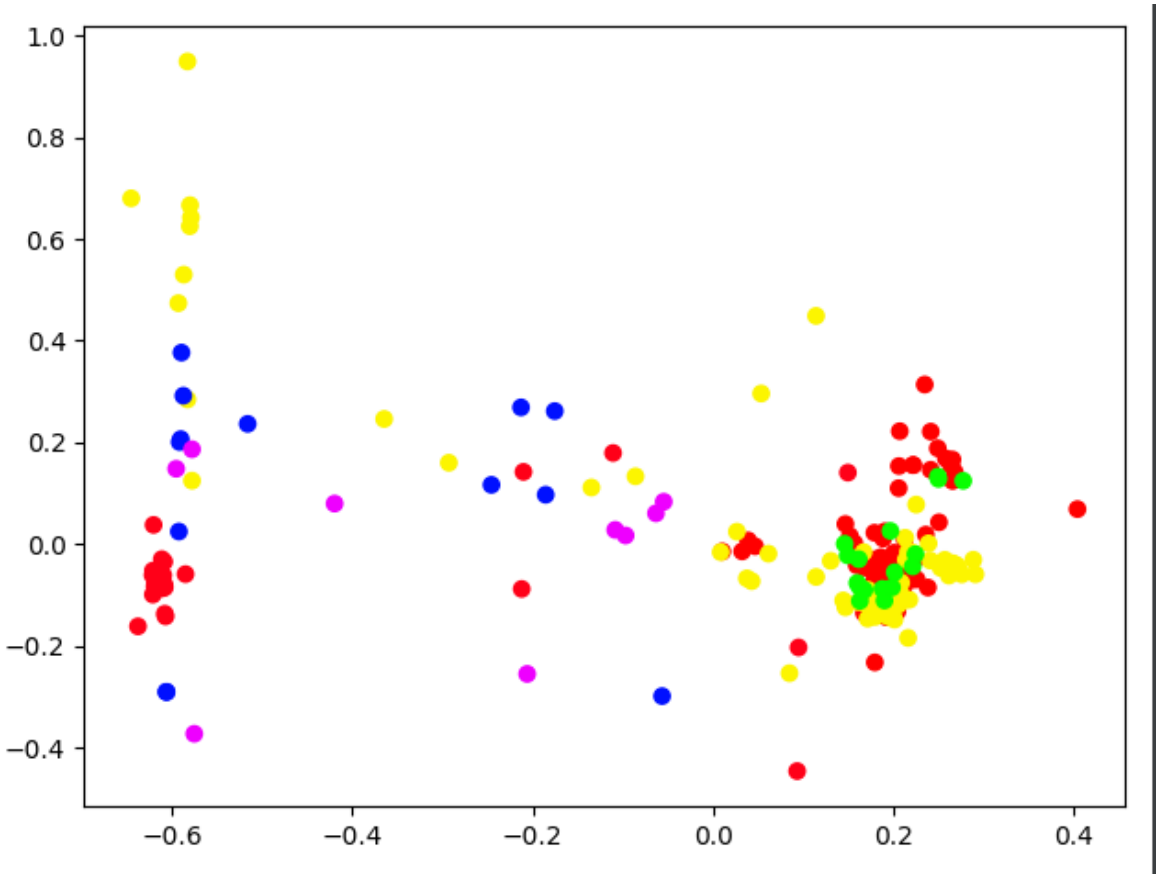


Рисунок 20. Alpha = 1

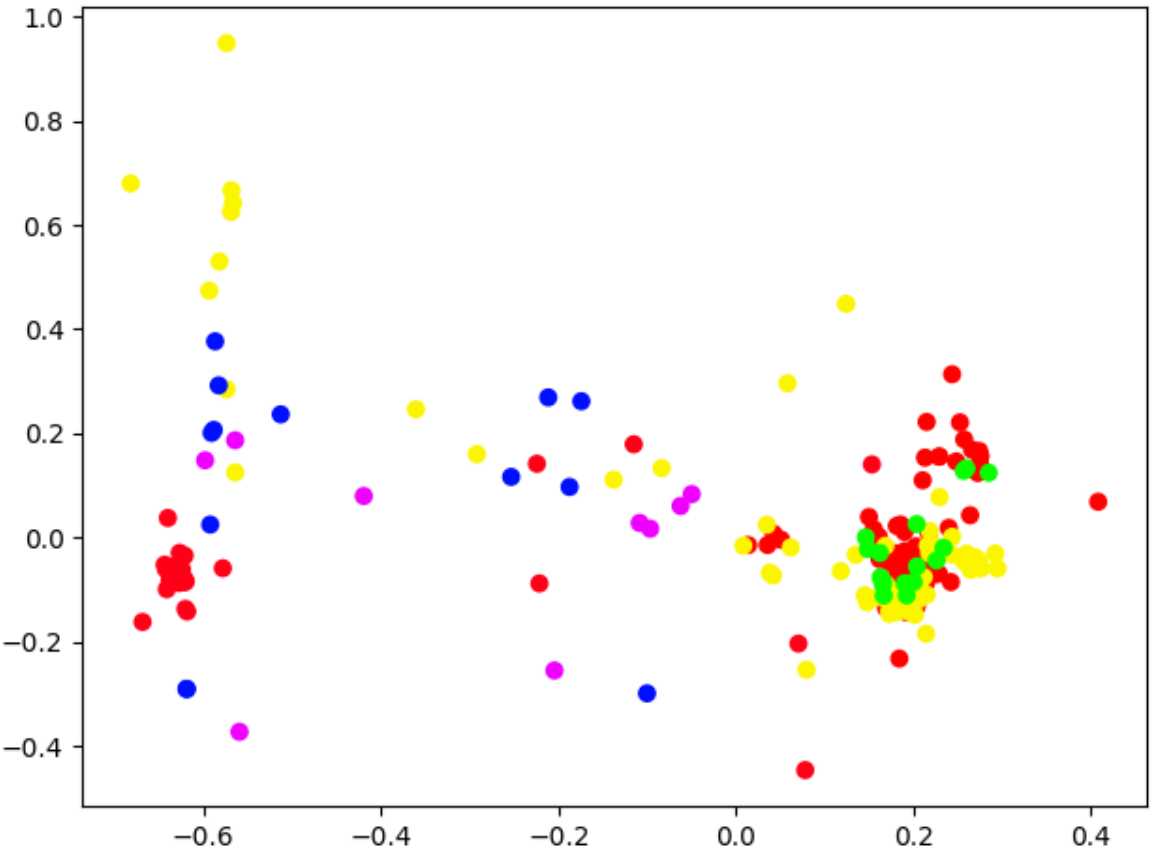


Рисунок 20. Alpha = 0.9

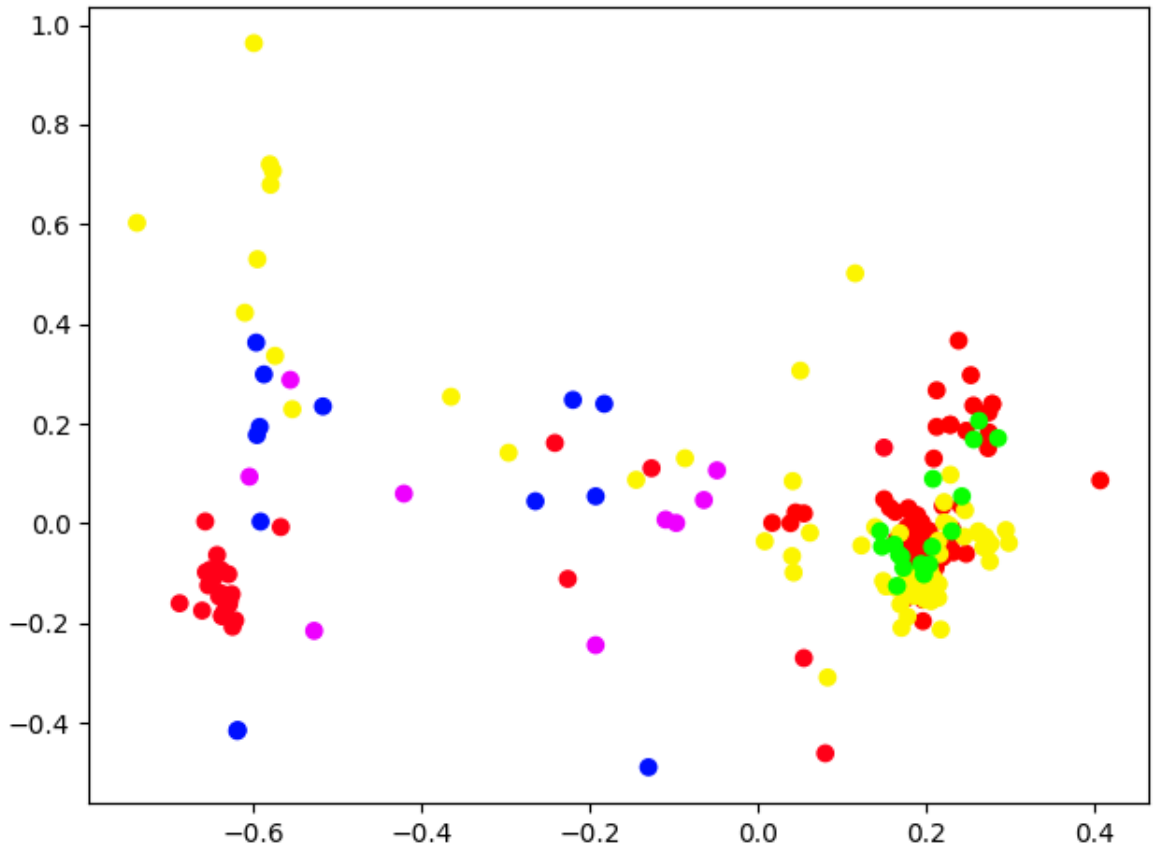


Рисунок 20. Alpha = 0.8

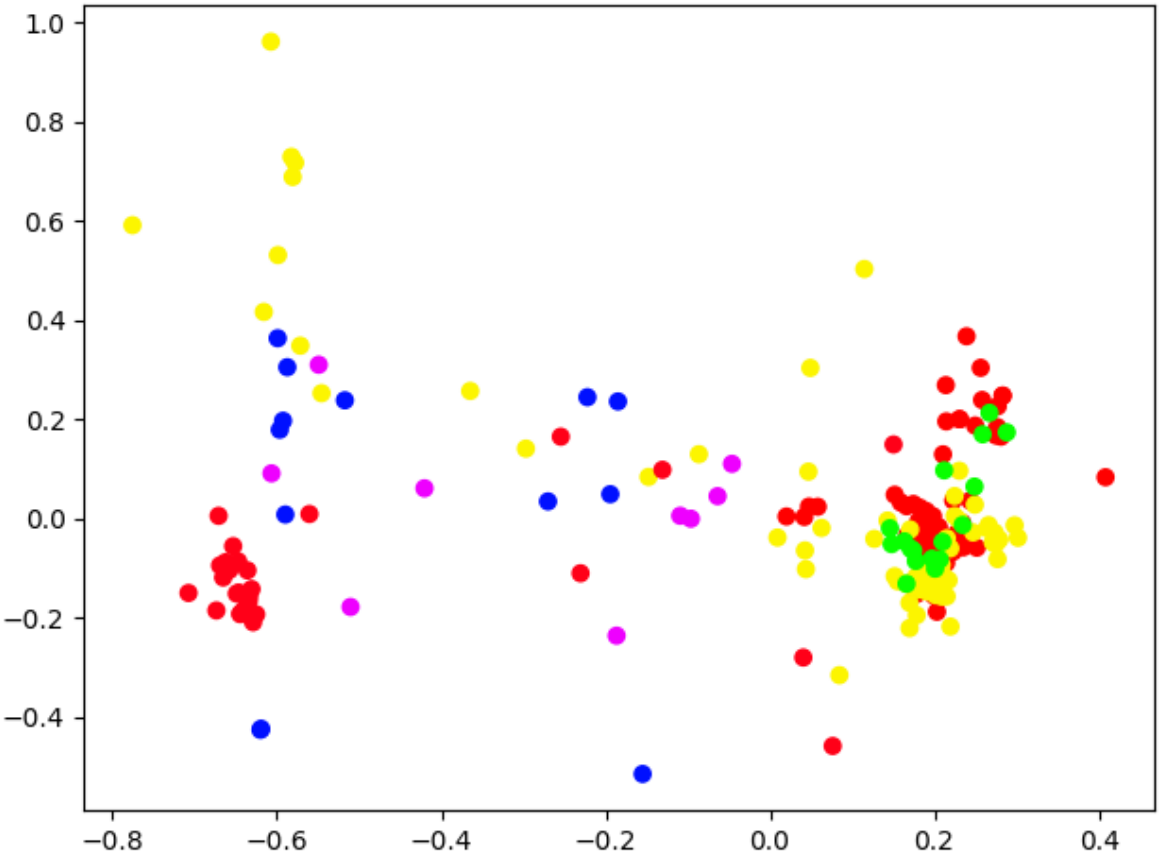


Рисунок 20. Alpha = 0.7

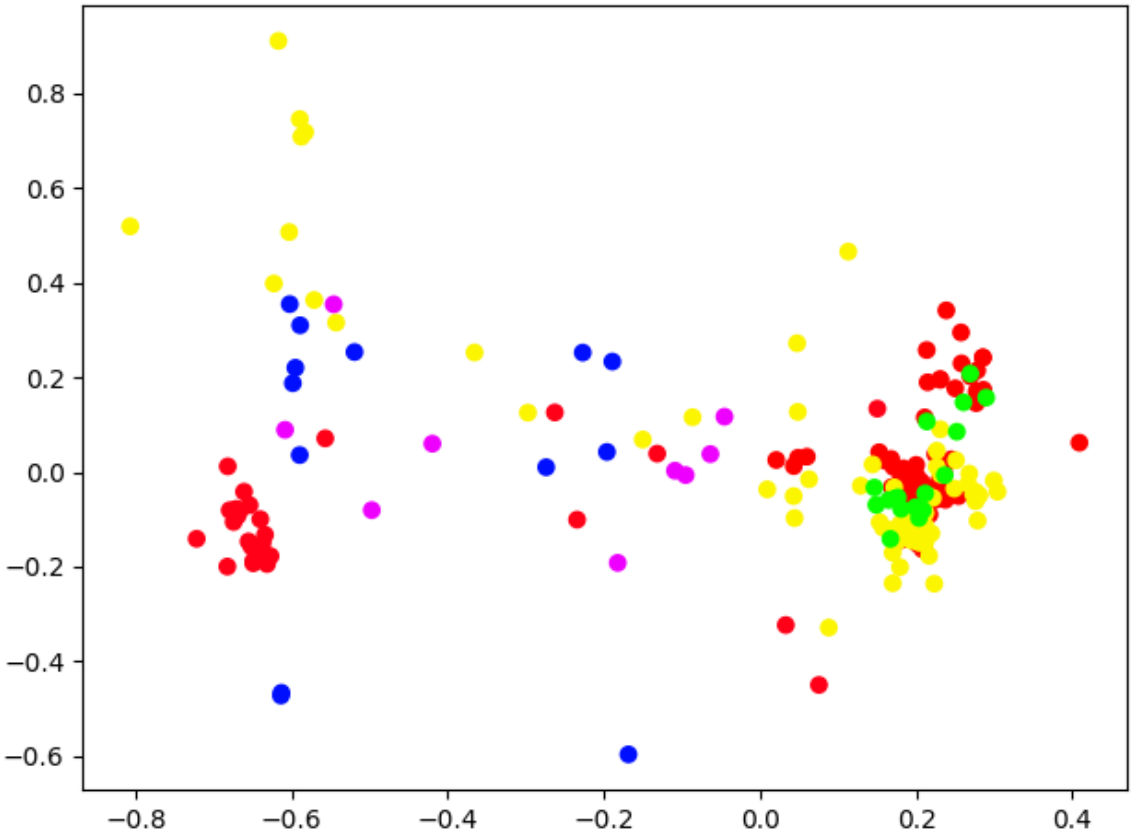


Рисунок 20. Alpha = 0.6

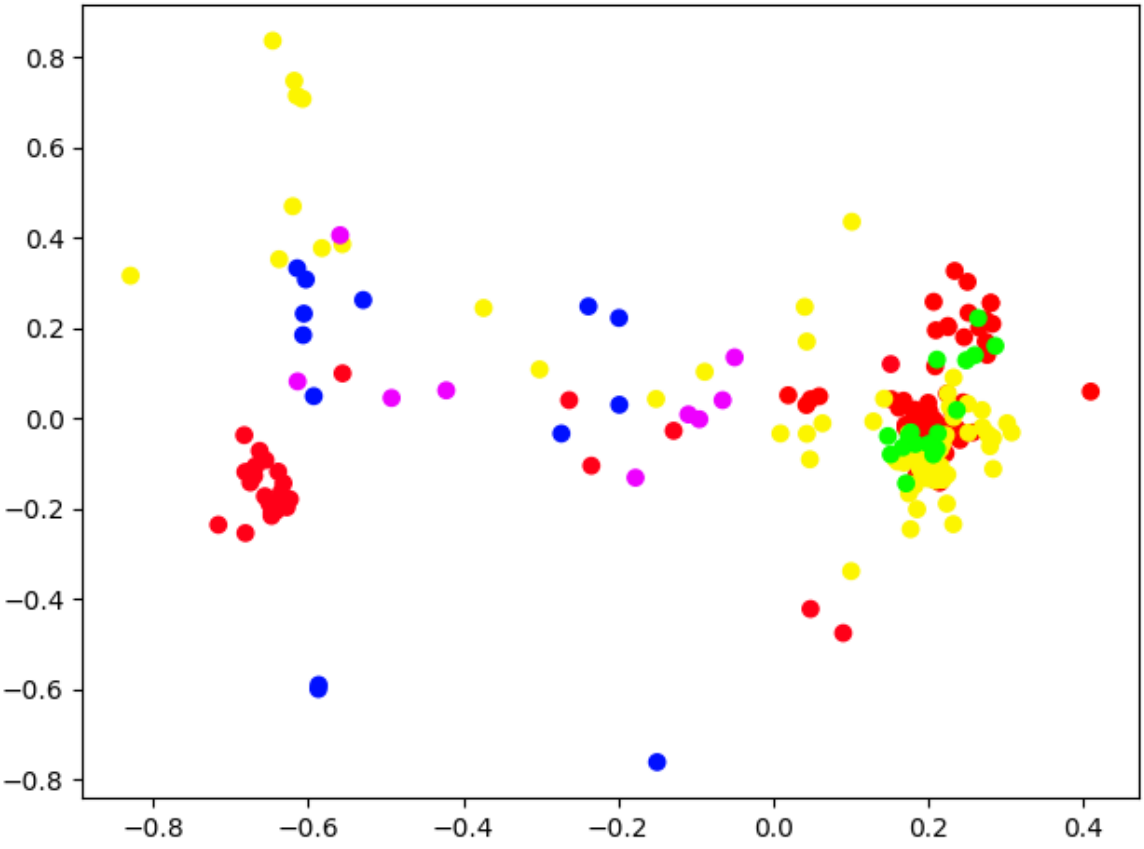


Рисунок 20. Alpha = 0.5

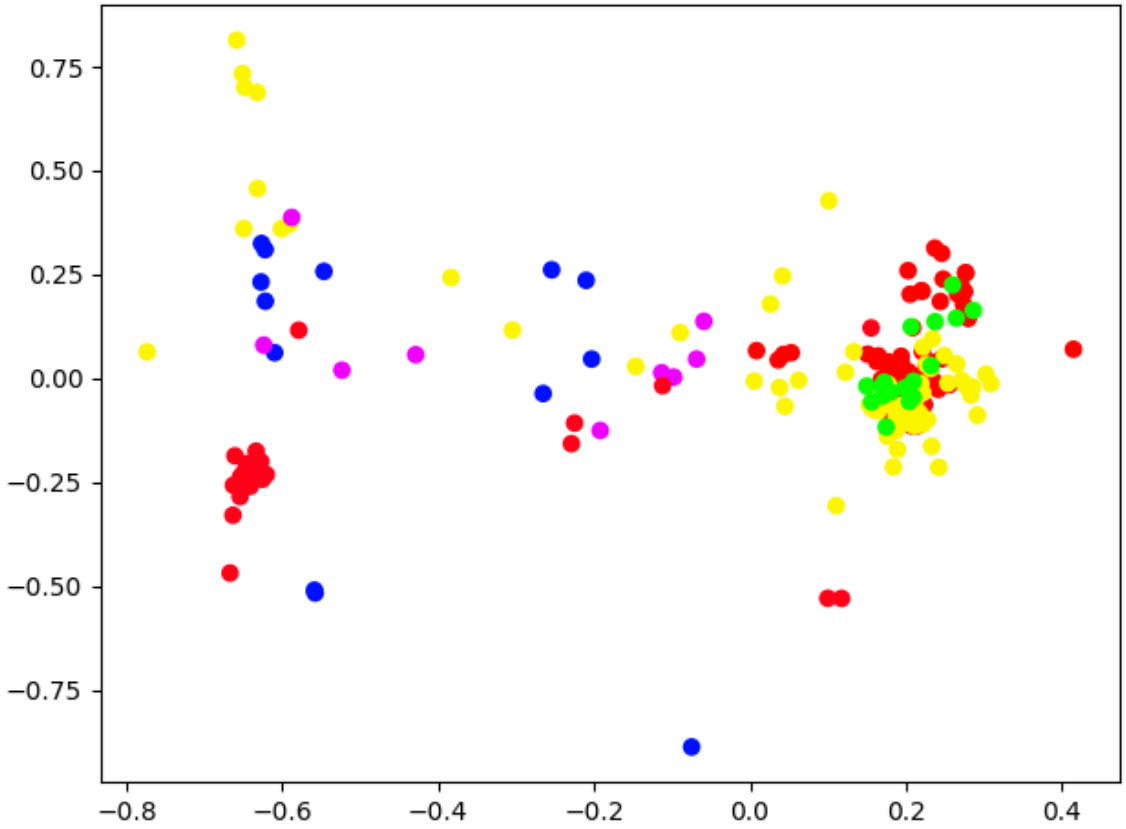


Рисунок 20. Alpha = 0.4

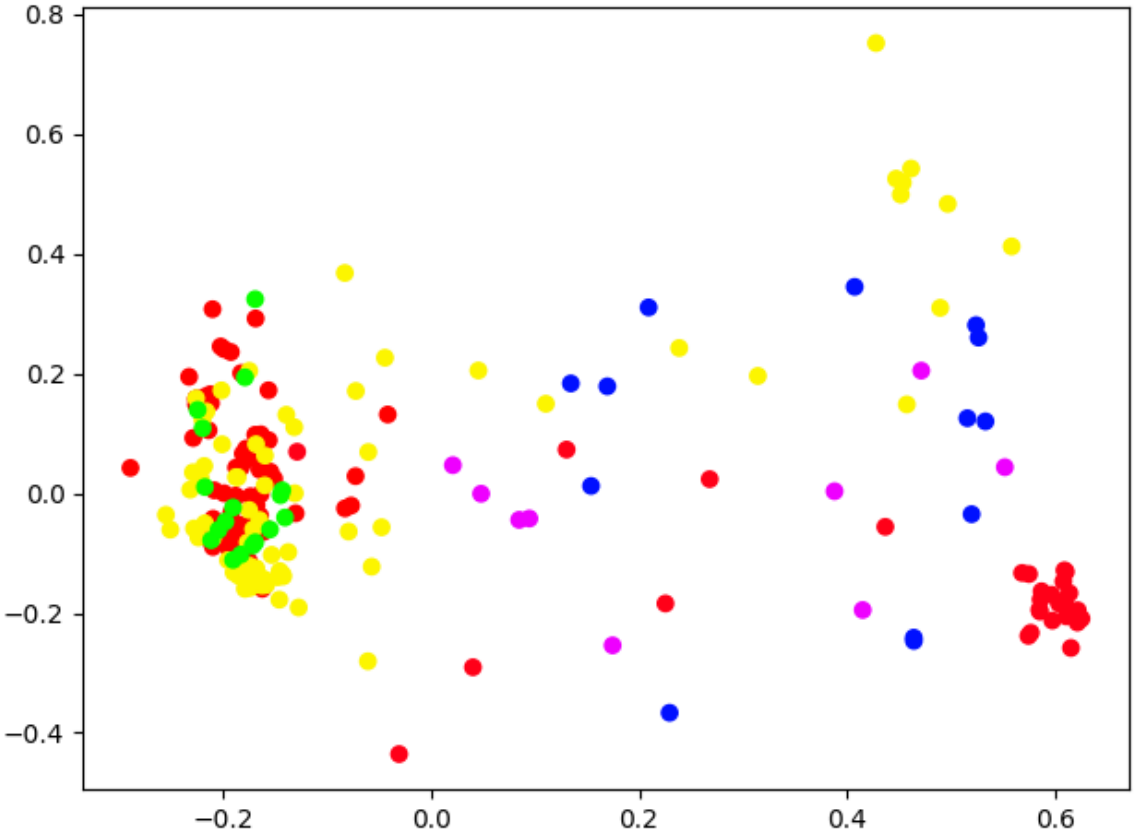


Рисунок 20. Alpha = 0.3

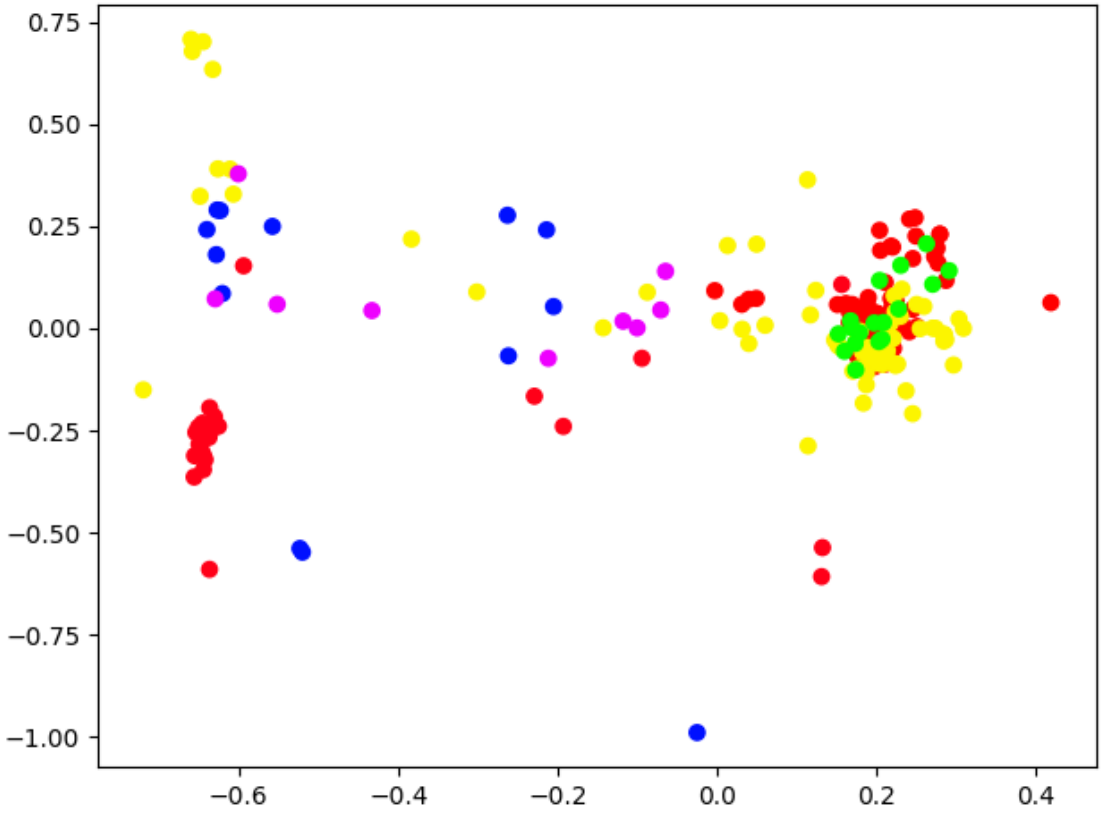


Рисунок 20. Alpha = 0.2

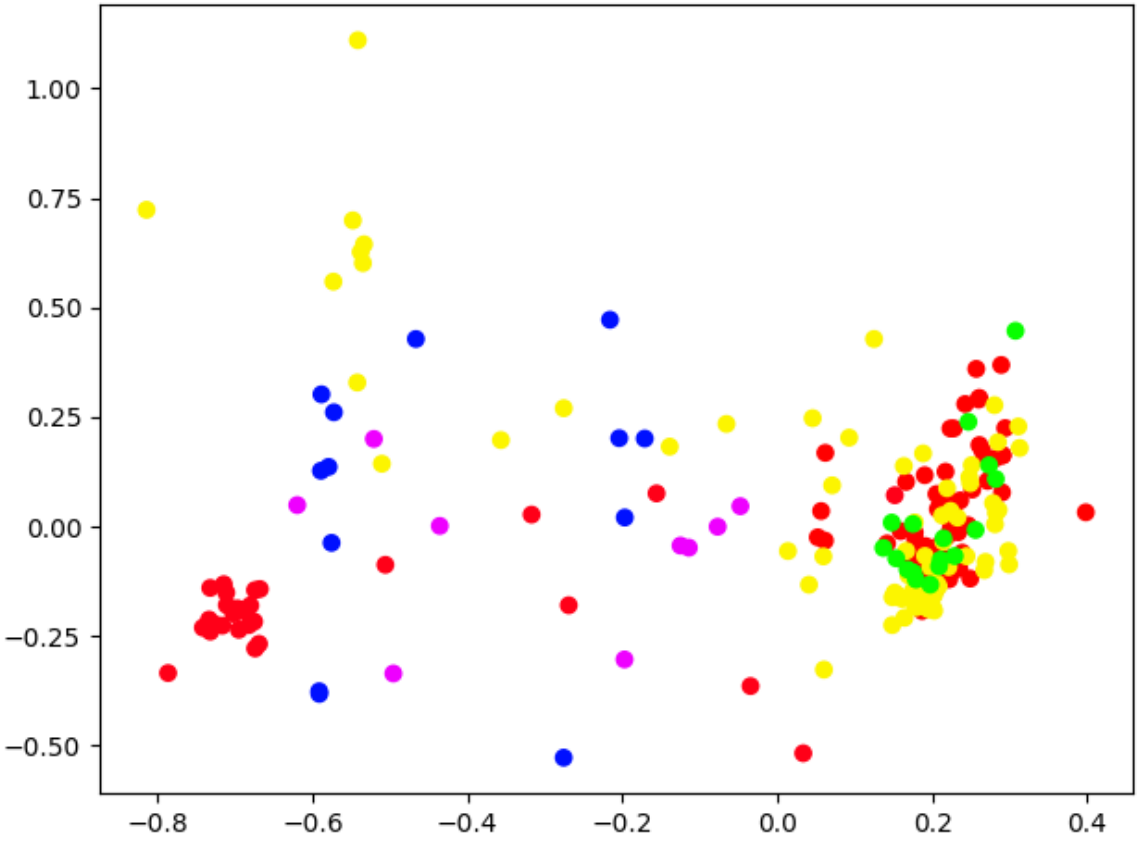


Рисунок 20. Alpha = 0.0

4. Проанализируйте и обоснуйте полученные результаты.

При понижении alpha до 0 становится немного похоже на результат у PCA только отраженным.

Факторный анализ

1. Проведите понижении размерности используя факторный анализ FactorAnalysis

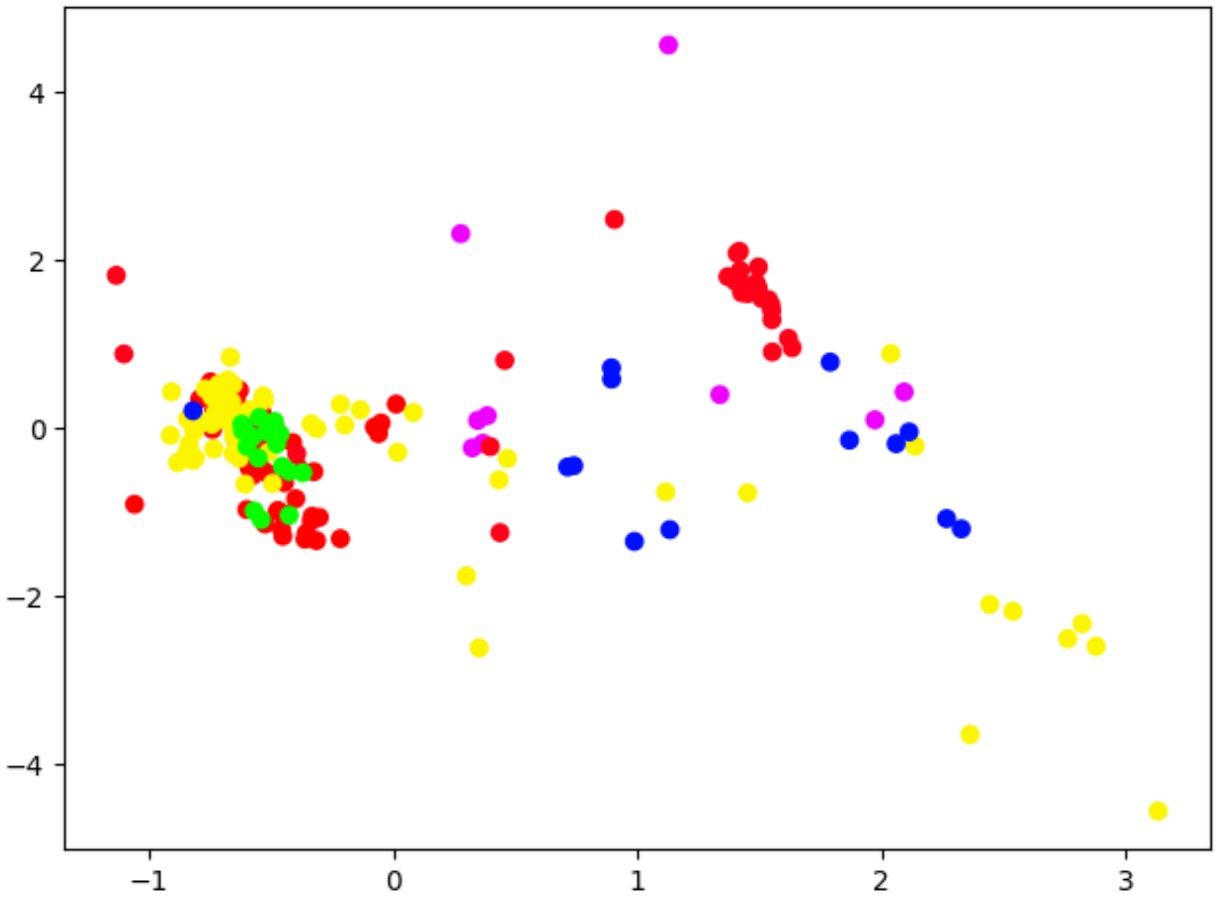


Рисунок 21. Для 4 компонентов.

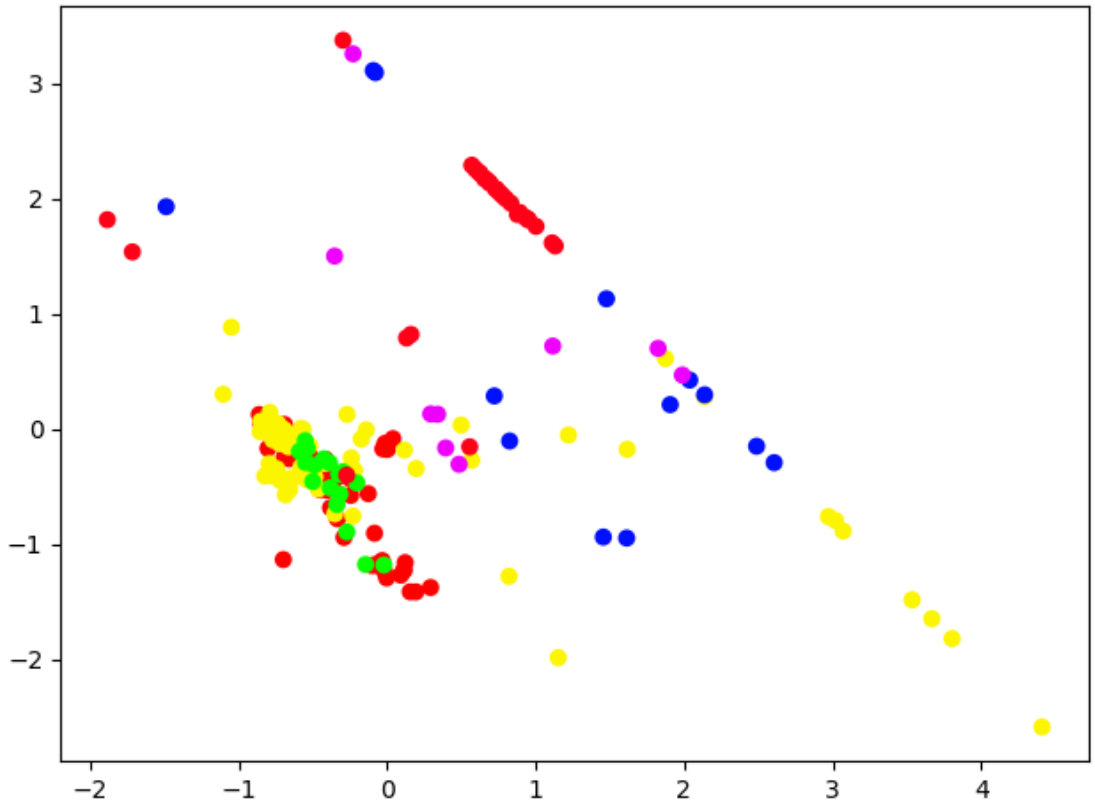


Рисунок 21. Для 2 компонентов.

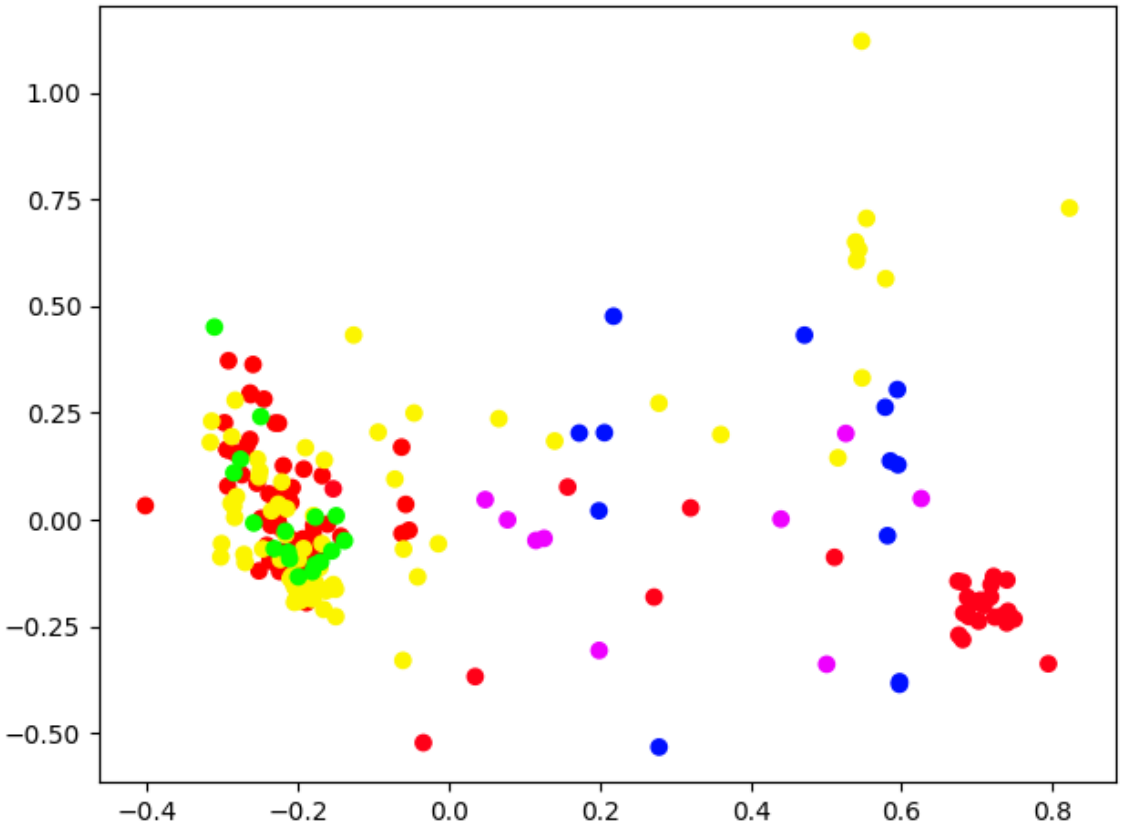


Рисунок 21. PCA для 2 компонентов.

2. Сравните полученные результаты с PCA

График очень отличается от PCA.

3. Объясните в чем разница между методом главных компонент и факторным анализом.

Факторный анализ используется как более широкий метод, предсказывающий наблюдаемые переменные из теоретически скрытых факторов.

Вывод

В результате работы ознакомился с методами предобработки данных из библиотеки Scikit Learn

.