МИНОБРНАУКИ РОССИИ

Санкт-Петербургский государственный электротехнический университет «ЛЭТИ» им. В. И. Ульянова (Ленина)

ОТЧЁТ
по лабораторной работе №2
по дисциплине «Машинное обучение»

Тема: «Понижение размерности пространства признаков»

Студент гр. 6307 Гарифуллин В.Ф.

Преподаватель Жангиров Т.Р.

Санкт-Петербург
2020

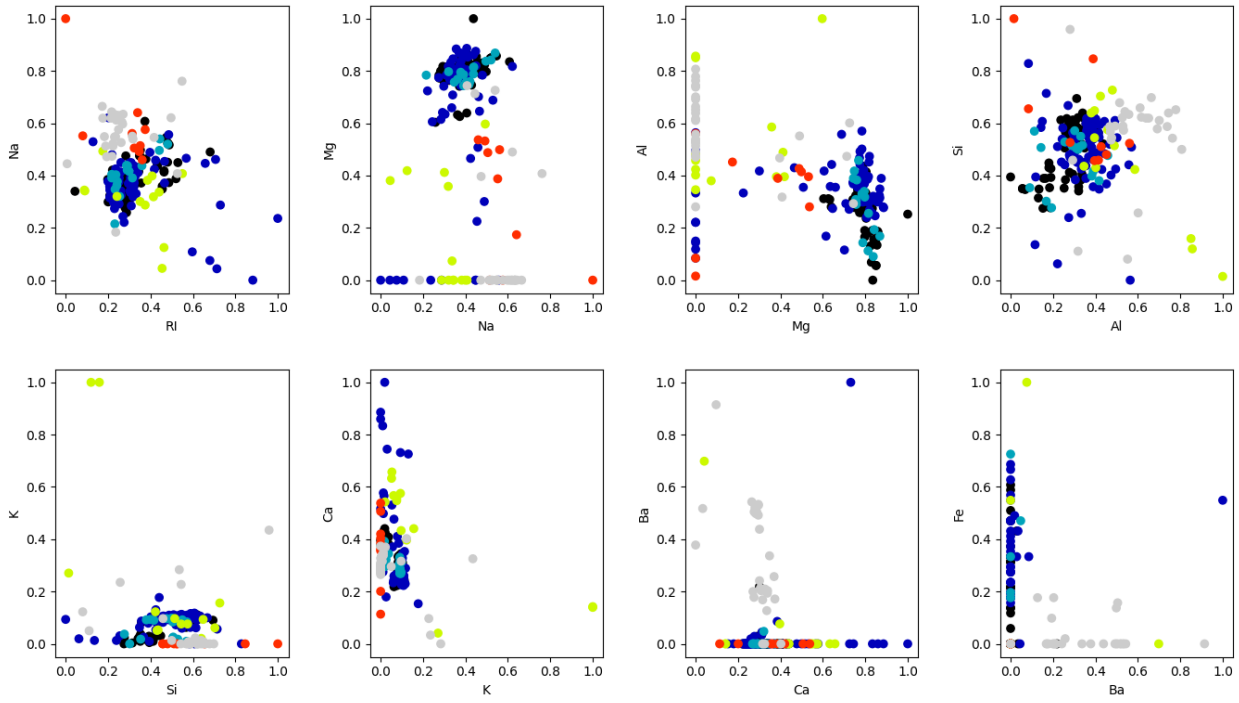
Цель работы

Ознакомиться с методами понижения размерности данных из библиотеки Scikit Learn

Ход работы

Загрузка данных

1. Провести нормировку данных к интервалу [0 1]
   Построить диаграммы рассеяния для пар признаков. Самостоятельно определите соответствие цвета на диаграмме и класса в датасете



1 - черный.
2 - синий.
3 - голубой.
5 - зеленый.
6 - оранжевый.
7 - серый.

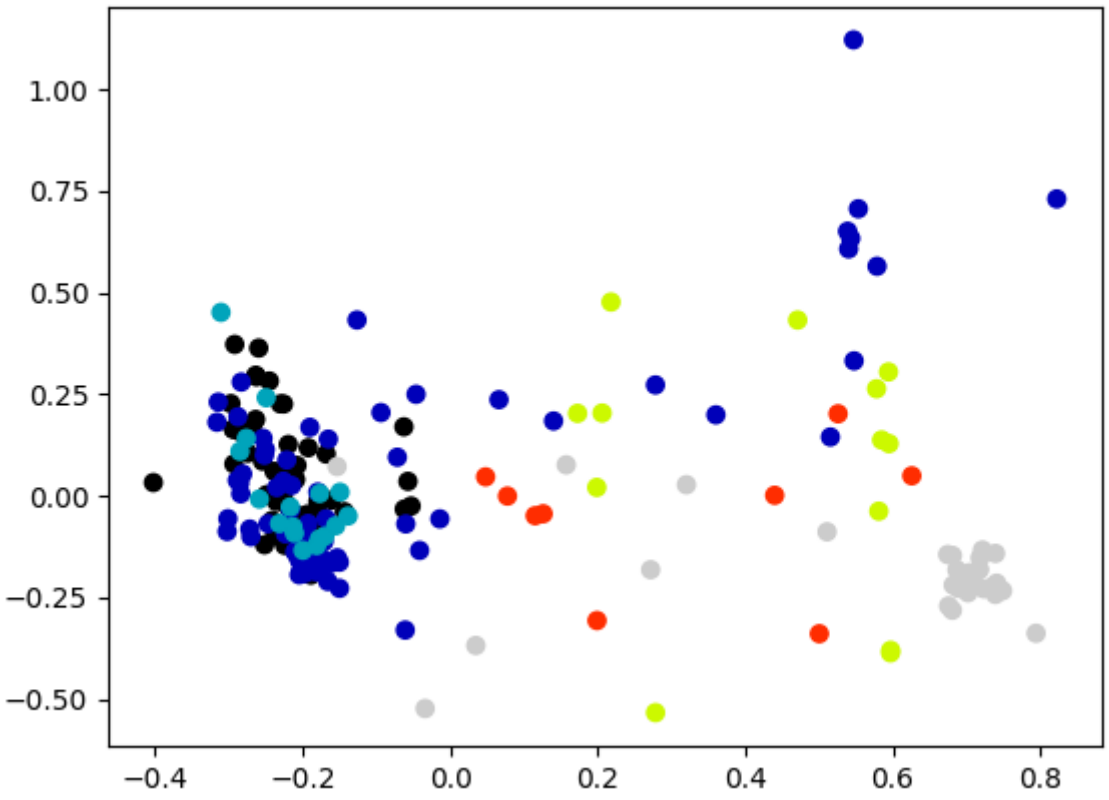
Метод главных компонент

1. Используя метод главных компонент (PCA). Проведите понижение размерности пространства до размерности 2. Выведите значение объясненной дисперсии в процентах и собственные числа соответствующие компонентам.

```
[0.45429569 0.17990097]
[5.1049308  3.21245688]
```

Значения объяснённой дисперсии процентах:
45% и 18%
Собственные числа:
5.1 и 3.2

2. Постройте диаграмму рассеяния после метода главных компонент



3. Проанализируйте и обоснуйте полученные результаты
   Ось X – первая главная компонента (объясненная дисперсия по ней максимальна). Ось Y – вторая главная компонента.

4. Изменяя количество компонент, определите количество при котором компоненты объясняют не менее 85% дисперсии данных
   Количество компонент, при котором компоненты объясняют не менее 85% дисперсии данных = 4

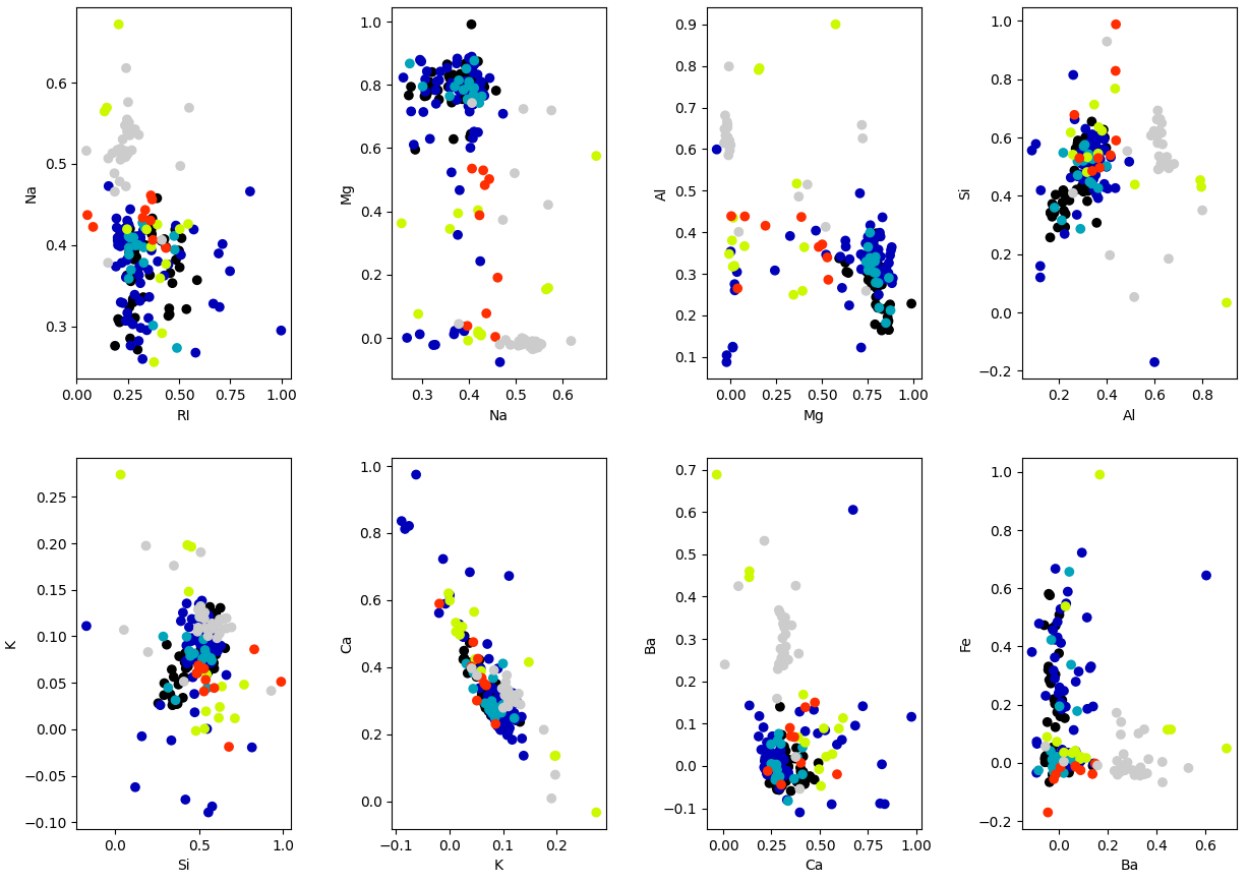
Расчёт суммы объясненной дисперсии:

```
print(np.sum(pca.explained_variance_ratio_))
```

Результат при 4 компонентах:

```
0.8586697305102717
```

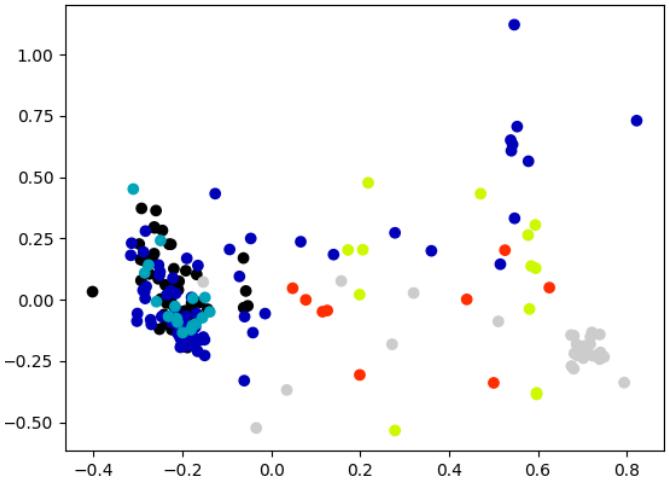
5. Используя метод inverse_transform восстановите данные, сравните с исходными



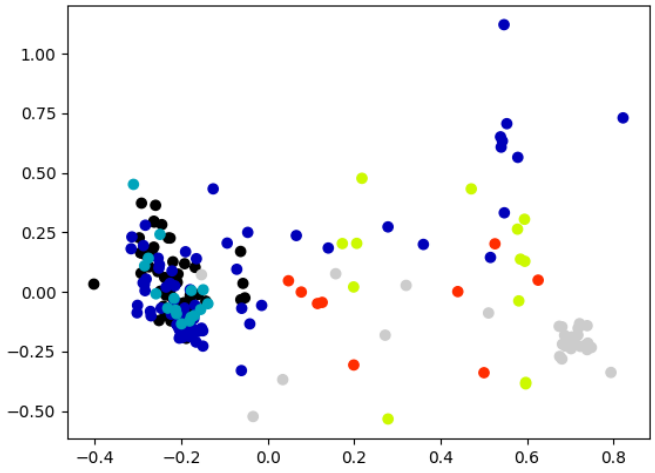
Часть информации о данных была потеряна, однако большинство информации сохранилось.

6. Исследуйте метод главных компонент при различных параметрах svd_solver

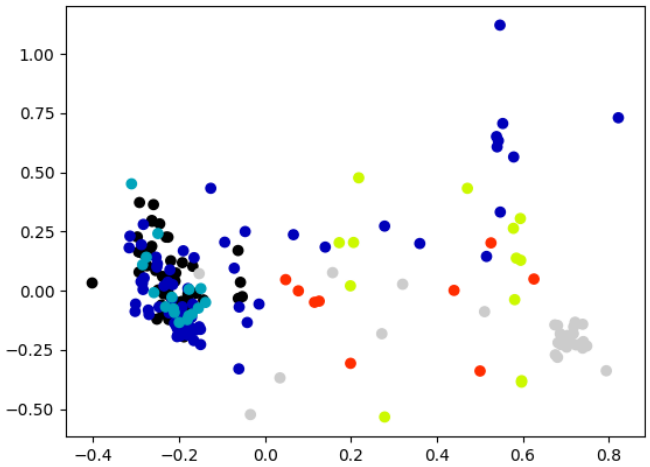
full:



arpack:



randomized:

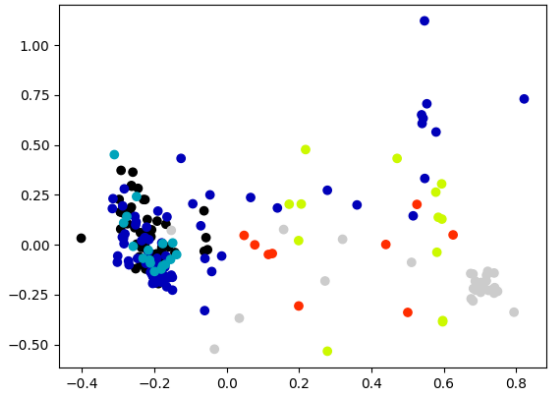


Для двух компонент результаты при всех параметрах идентичные. Параметр full может подобрать количество компонент в зависимости от необходимого уровня объяснённой дисперсии.

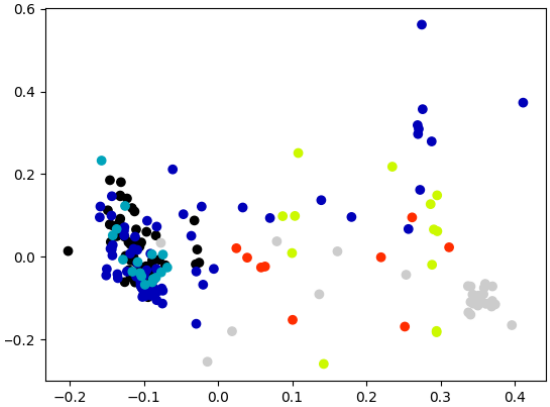
Модификации метода главных компонент

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

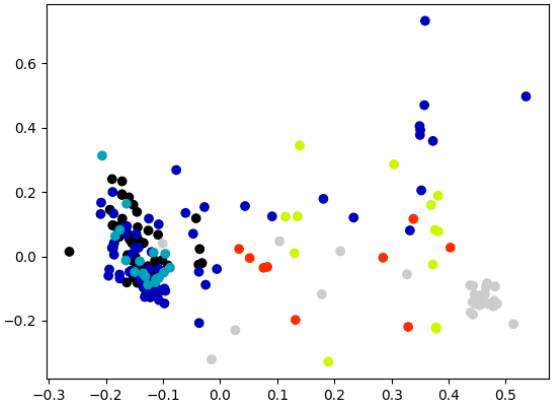
kernel = linear



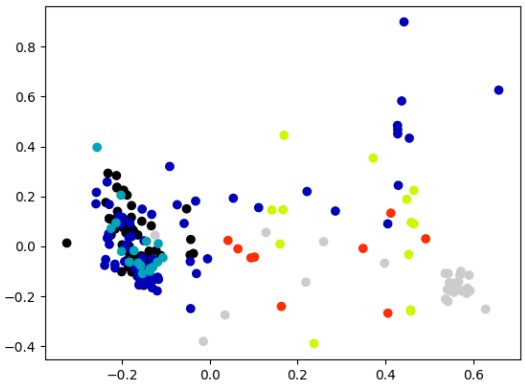
kernel = poly, degree = 2



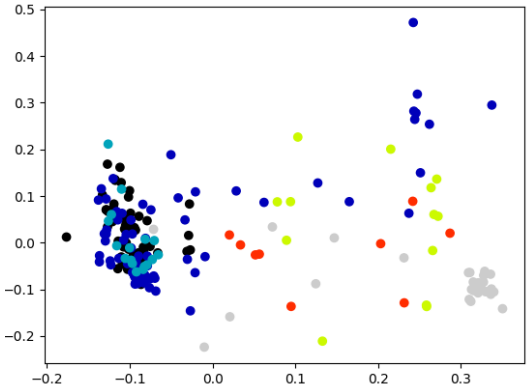
kernel = poly, degree = 3



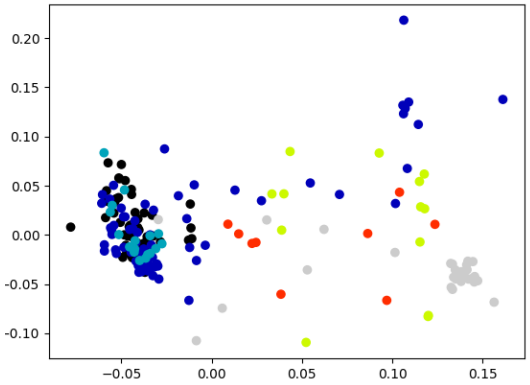
kernel = poly, degree = 4



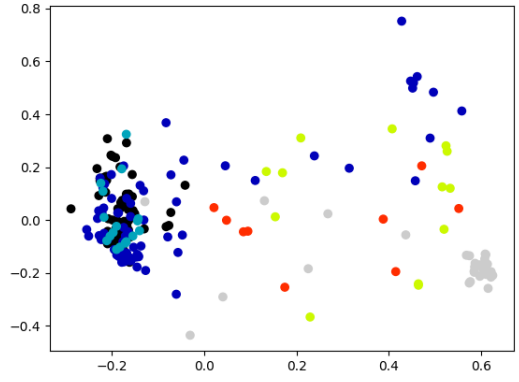
kernel = rbf



kernel = sigmoid



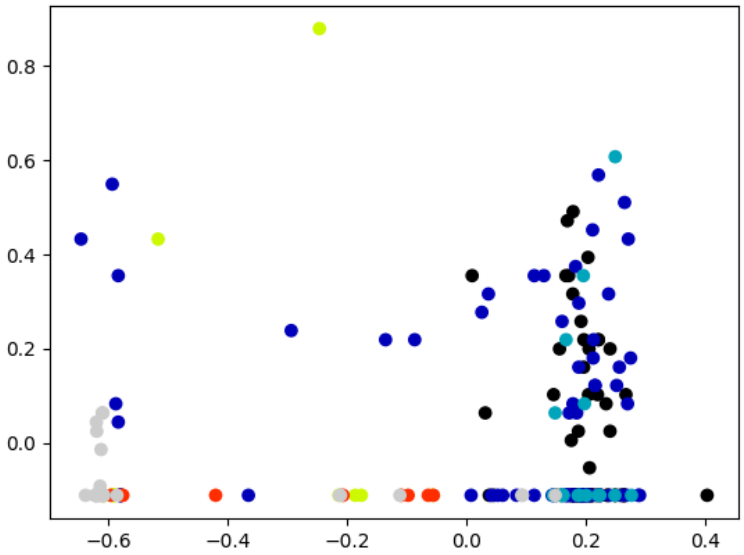
kernel = cosine



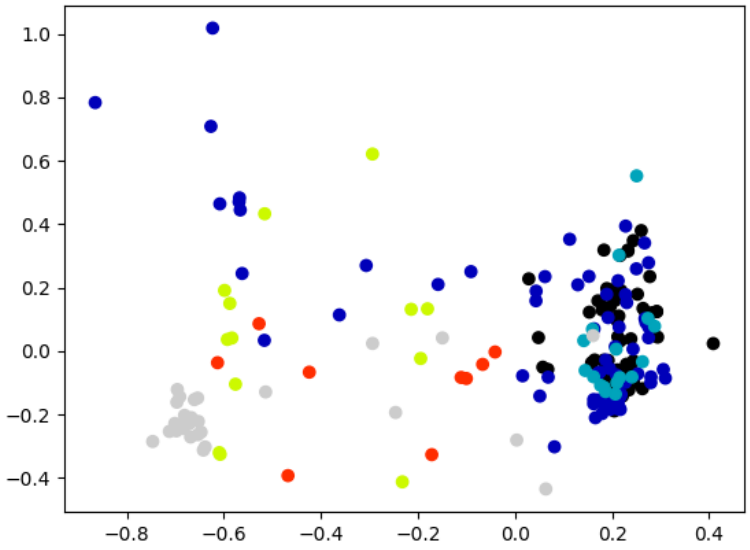
Почти все параметры дают примерно одинаковый результат. Различается только масштаб. Немного отличается результат при использовании ядра cosine.

2. Определите, при каких параметрах KernelPCA работает также как PCA
   Результаты совпадают при использовании ядра linear.

3. Аналогично исследуйте SparcePCA
   alpha = 1



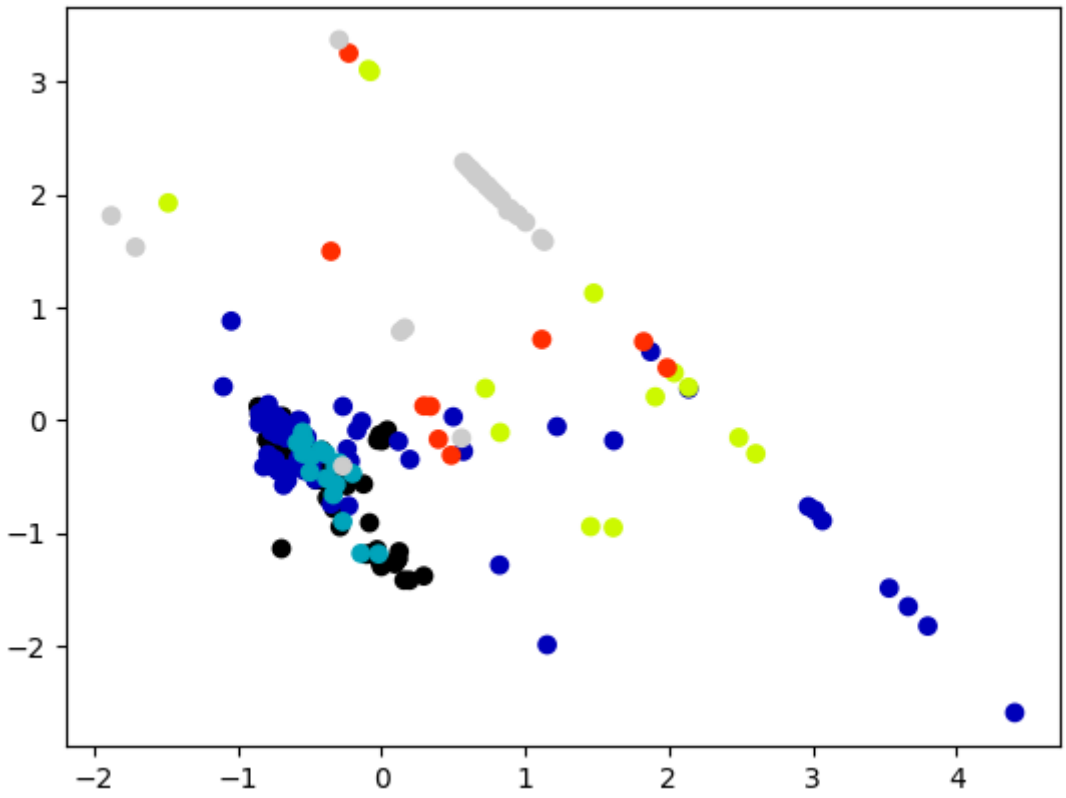
   alpha = 0.25



   При alpha = 1 результат отличается от всех остальных методов. При alpha = 0.25 результат такой же, как при PCA, но с инвертированной осью X

Факторный анализ

1. Проведите понижении размерности используя факторный анализ FactorAnalysis. Сравните полученные результаты с PCA. Объясните в чем разница между методом главных компонент и факторным анализом



При методе главных компонент первая главная компонента выбирается так, чтобы она имела максимальную дисперсию. Следующая выбирается также, при этом её корреляция с предыдущей должна быть минимальна.

Факторный анализ – выделение факторов из признаков, более емко отражающих свойства объекта. При анализе в один фактор объединяются сильно коррелирующие между собой переменные.

Вывод

В ходе работы было произведено понижение размерности данных с помощью метода главных компонент, модификаций метода главных компонент и факторного анализа.

Понижение размерности данных позволяет выделить наиболее важные признаки для дальнейшего анализа данных.