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»

Студентка гр. 6307 _____ Кичерова А. Д.
Преподаватель _____ Жангиров Т. Р.

Санкт-Петербург

2020

Цель работы

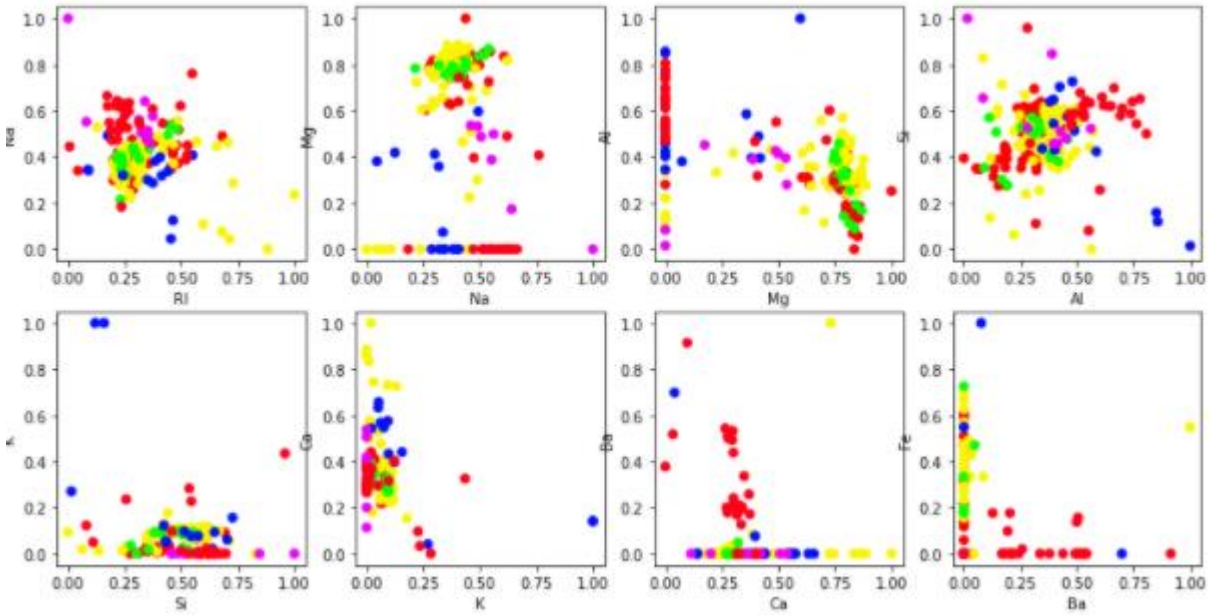
Ознакомиться с методами понижения размерности данных из библиотеки Scikit Learn.

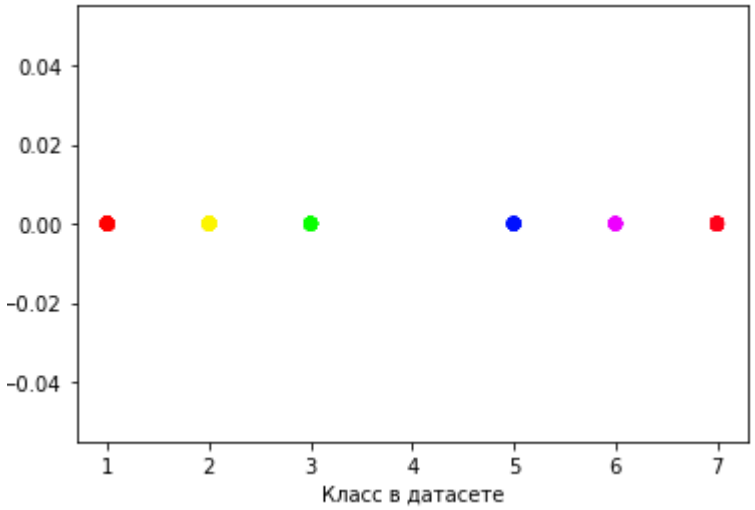
Выполнение

Загрузка данных

1. Датасет был загружен в датафрейм. Данные были приведены к интервалу [0 1].

2. Была построена диаграмма рассеяния для пар признаков, а так же были определены соответствия цвета на диаграмме и класса в датасете.



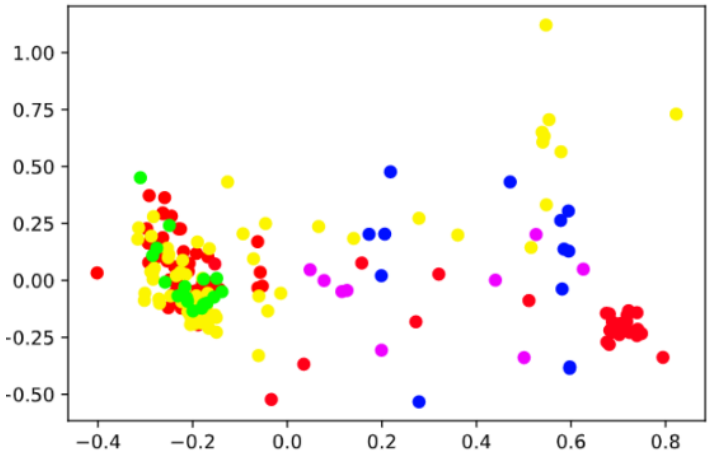


Метод главных компонент

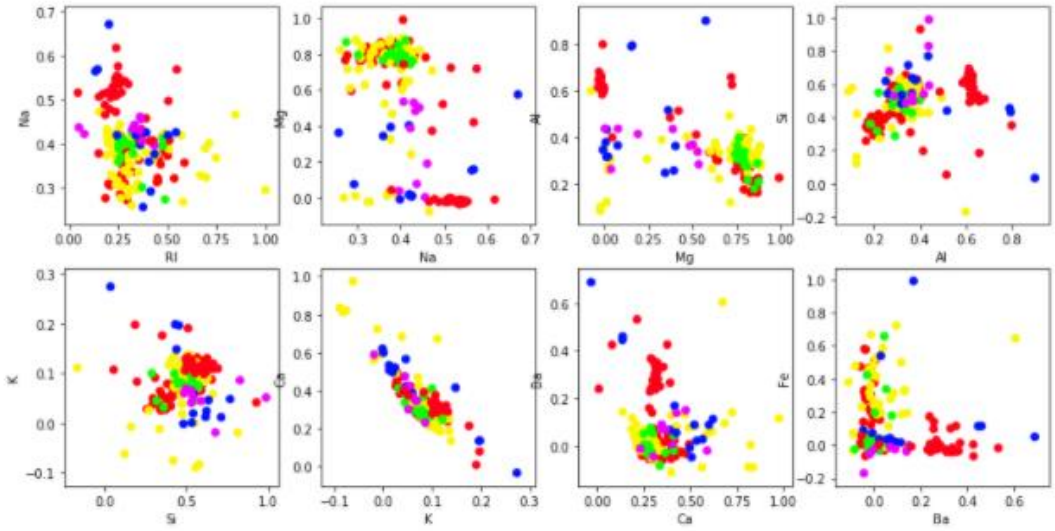
1. Используя метод главных компонент было приведено понижение размерности пространства до 2, после этого было выведено значение объясненной дисперсии и собственные числа.

```
Var:  [0.45429569 0.17990097]
Eigenvals:  [5.1049308  3.21245688]
```

Видно, что данные имеют около 63.5 % информации.



2. В ходе применения PCA для разного количества компонент было выяснено, что данные содержат не менее 85% информации при 4 компонентах. При этом из-за потери информации почти в 15% восстановленные данные отличаются от исходных.

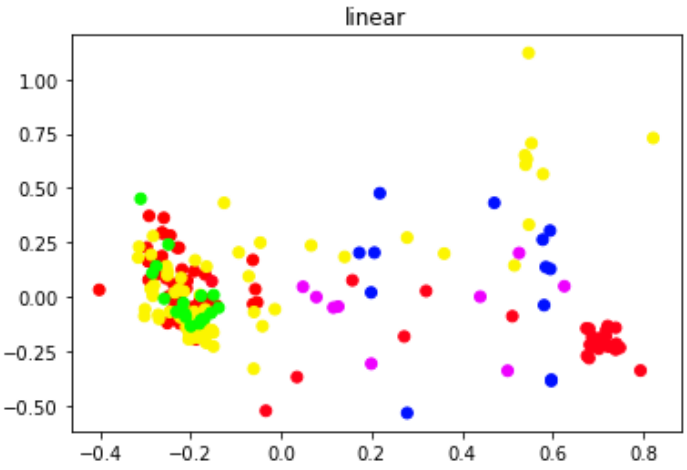


Модификации метода главных компонент

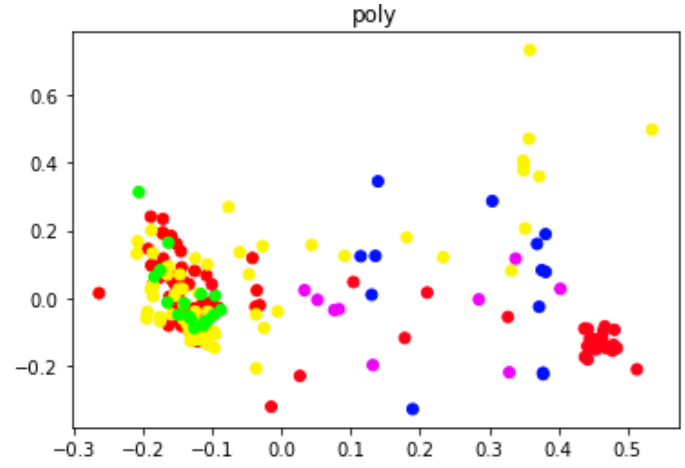
1. KernelPCA позволяет нелинейно уменьшить размерность с помощью использования ядерных функций. Ядерные функции выбираются параметром kernel.

Параметр	Ядерная функция
linear	$k(x,y) = x^T y$
poly	$k(x,y) = (\gamma x^T y + c_0)^d$
rbf	$k(x,y) = \exp(-\gamma \|\|x - y\|\|^2)$
cosine	$k(x,y) = \frac{xy^T}{\|\|x\|\| \, \|\|y\|\|}$
sigmoid	$k(x,y) = \tanh(\gamma x^T y + c_0)$

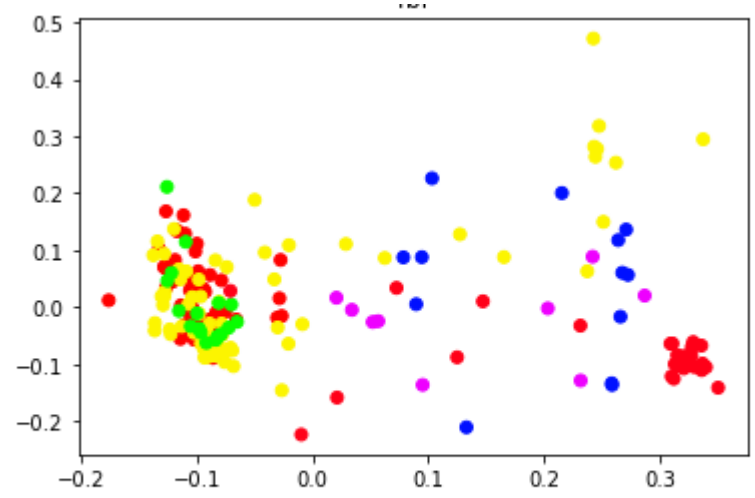
В ходе применения различных ядерных функций было выяснено, что обобщенная дисперсия ядерных матриц различаются незначительно, а собственные значения сильно разнятся.



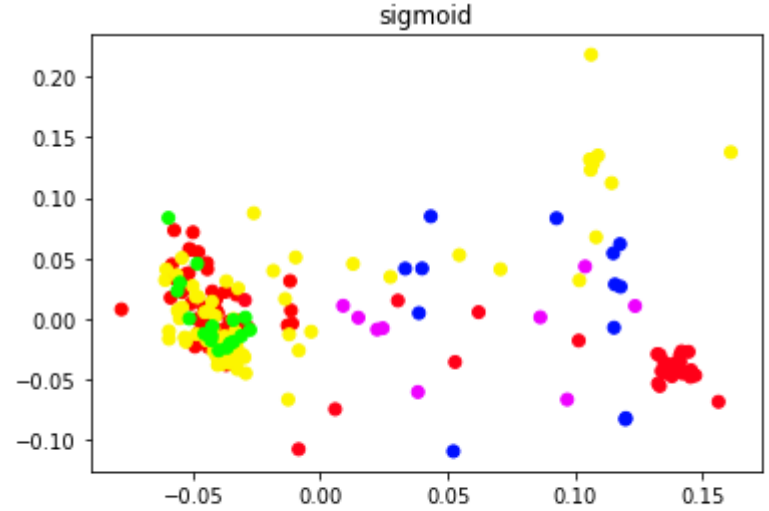

```
Eigenvals [26.06031845 10.31987923  7.25626387  5.6204589 ]
Var 0.8586697305102717
```



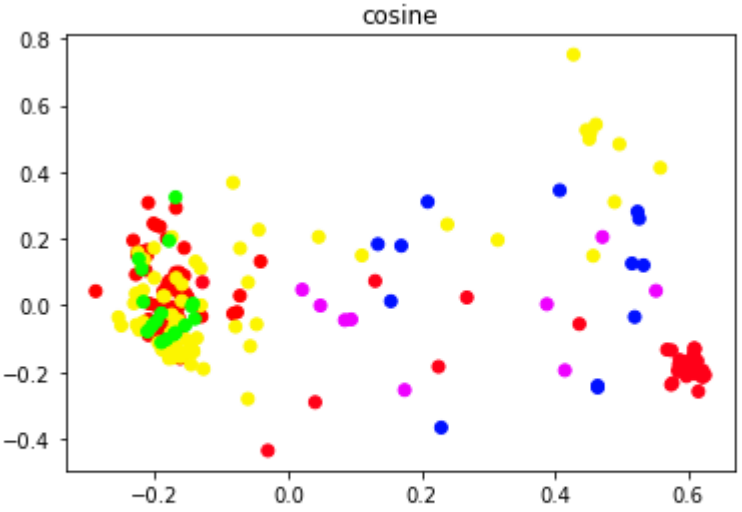

```
Eigenvals [10.9181964   4.31937695  3.1188508   2.36791674]
Var 0.8490317237345795
```




```
Eigenvals [5.35145251 2.0180542  1.4957381  1.11090453]
Var 0.8499381817557572
```




```
Eigenvals [1.00618101 0.39983752 0.27409853 0.2161195 ]
Var 0.8618649340638378
```

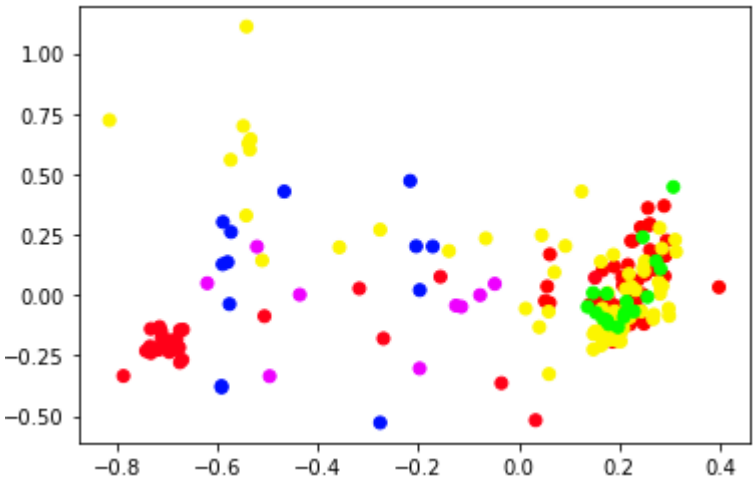


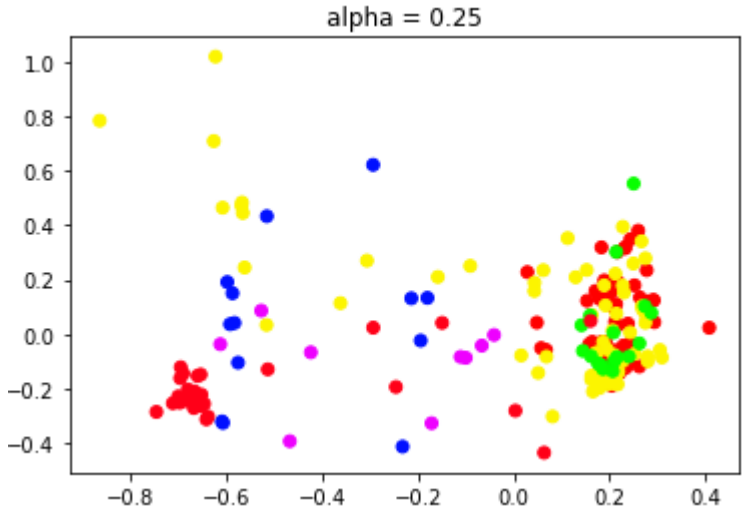

```
Eigenvals [18.31403041  6.47538495  4.6959991   3.57812492]
Var 0.8599435287117831
```

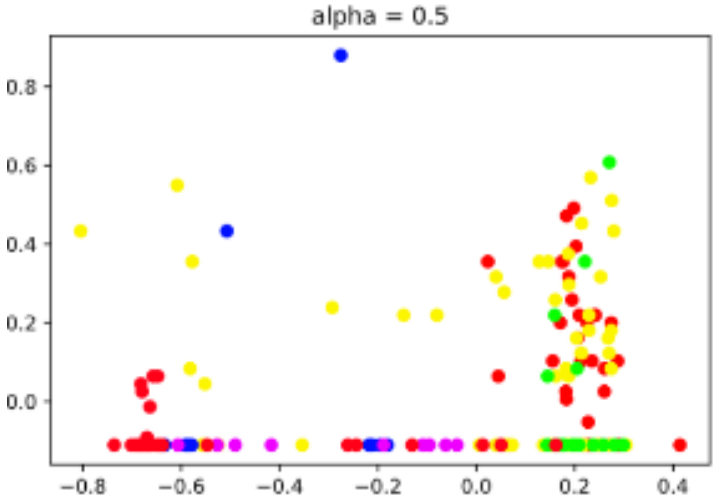
2. KernelPCA работает как PCA при линейной ядерной функции.

3. SparsePCA извлекает набор разряженных компонентов, которые лучше всего восстанавливают данные.

SparsePCA вычисляет различные результаты при изменении параметра, который контролирует разреженность. Чем выше значение, тем компоненты более разрежены.

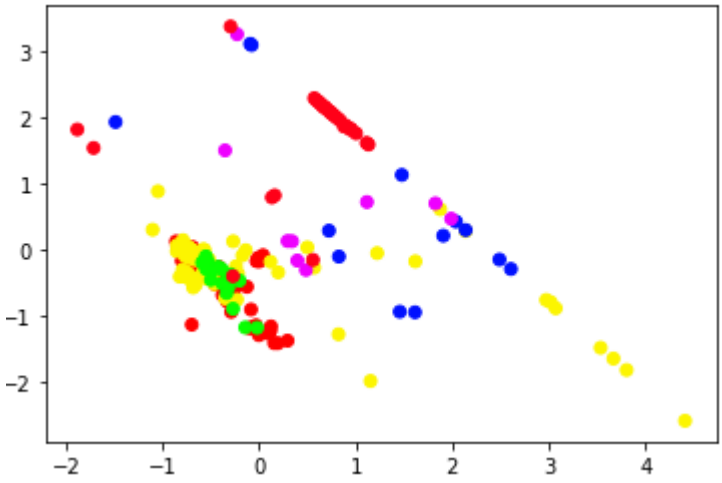






Факторные анализ

Сравнения результатов PCA и факторного анализа показывают различия в полученных данных.



Основные различия PCA и FA:

1. Компоненты PCA полностью независимы друг от друга, факторные анализ не накладывает таких ограничений.
2. В PCA каждый компонент (фактор) - линейная комбинация переменных, тогда как в FA это переменные, которые выражаются как линейные комбинации факторов (включая компоненты общности и уникальности, как вы сказали).

Выводы

В данной работе были изучены методы понижения размерности PCA, KernelPCA, SparsePCA и FA.

Разное количество компонент в PCA объясняют разное количетсво дисперсии данных.

KernelPCA используется для поиска нелинейных зависимостей данных. На предложенном наборе данных объясненная дисперсия было 85% вне зависимости от выбранного ядра.