Мартынов Д.А.

МПИ-20-4-2

Лабораторная работа 2

В работе производится обработка датасетов при помощи метода главных компонент с целью уменьшить размерность исходных датасетов. Метод главных компонент (PCA) взят из библиотеки sklearn, однако есть и ручная реализация, которая показывает такие же результаты на тестовом датасете.

В качестве исходных датасетов был взят Iris, Breast cancer и сгенерированный датасет с 60 признаками.

Критерии, которые использовались для оценки количества главных компонент:

1. Кайзера

Согласно данному критерию, стоит выбирать те главные компоненты, чьи собственные значения превышают единицу.

1. Каменистая осыпь

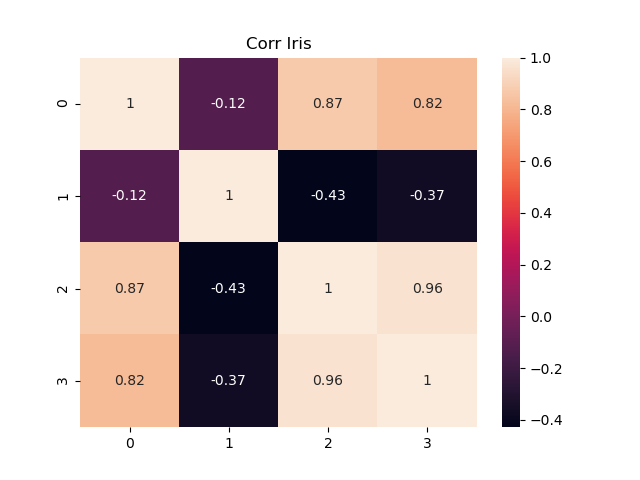
Строится так называемый Scree-plot – график, по горизонтали которого отложены главные компоненты, а по вертикали – дисперсию исходных данных, которую они описывают. Согласно критерию, стоит выбрать такое количество главных компонент, которое на графике предшествует сильному замедлению убывания.

1. Сломанная трость

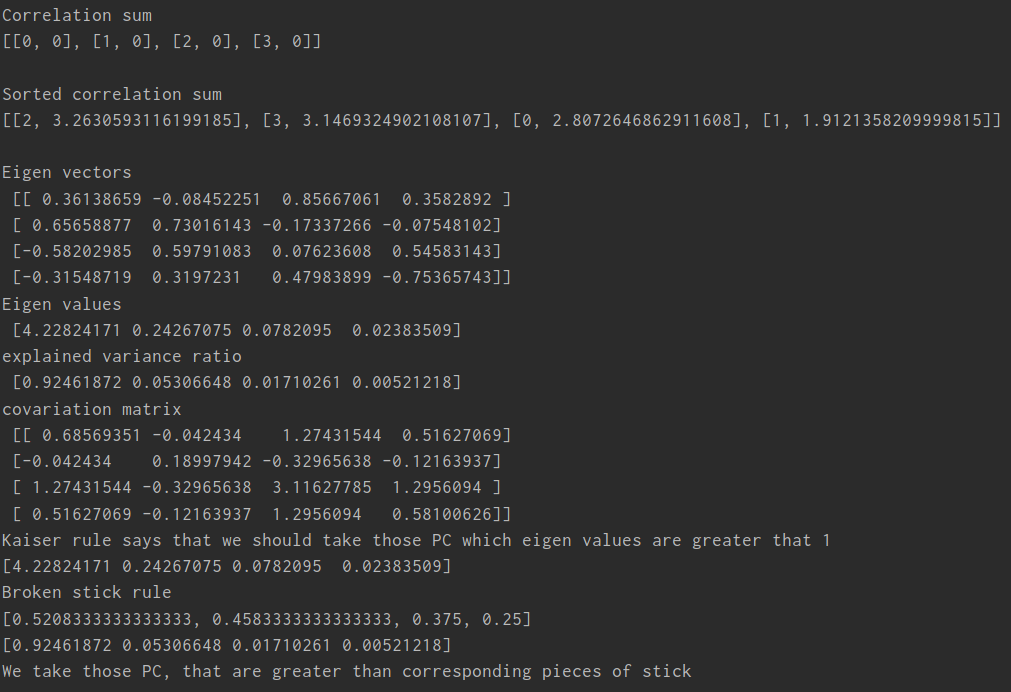
Сначала считается распределение длин кусков трости, сломанной в n-1 местах, где n – количество главных компонент. Далее эти длины сравниваются с нормированными собственными числами компонент. Согласно критерию, стоит выбрать те главные компоненты, нормированные собственные числа которых больше соответствующих длин кусков трости.

1. **Датасет Iris**

Матрица корреляций:

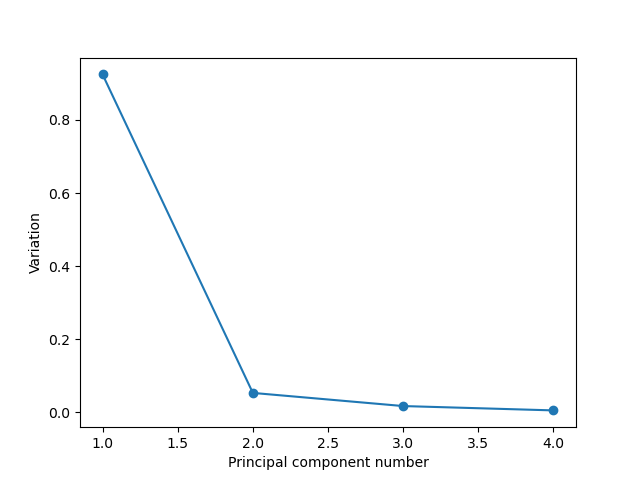


Выявим признаки, которые меньше других коррелирует со всеми

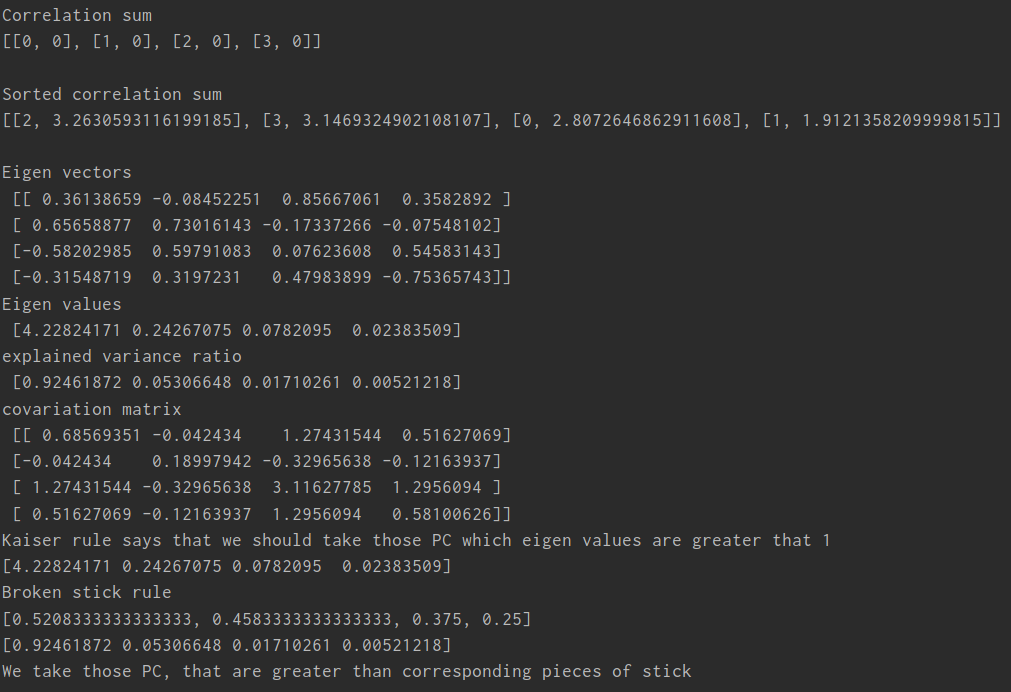


В списке выше признаки отсортированы по убыванию суммы корреляции.

Далее прогоним модель с датасетом и посчитаем критерии Кайзера, сломанной трости и каменистой осыпи:

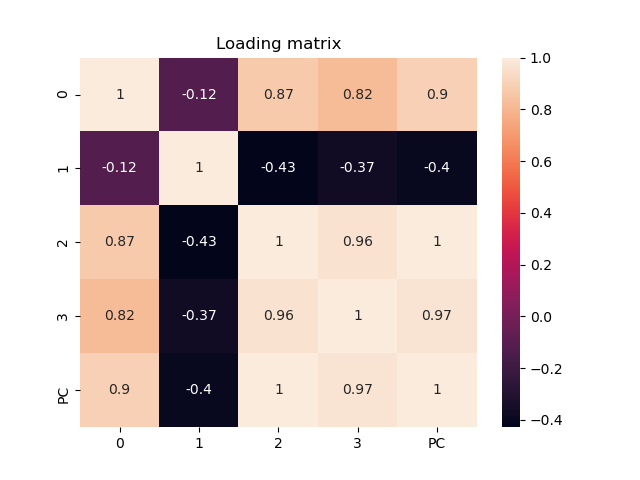


На графике – каменистая осыпь. Тут стоит выбрать 1 компоненту, так как сильнее всего замедляется убывание дисперсии после нее.



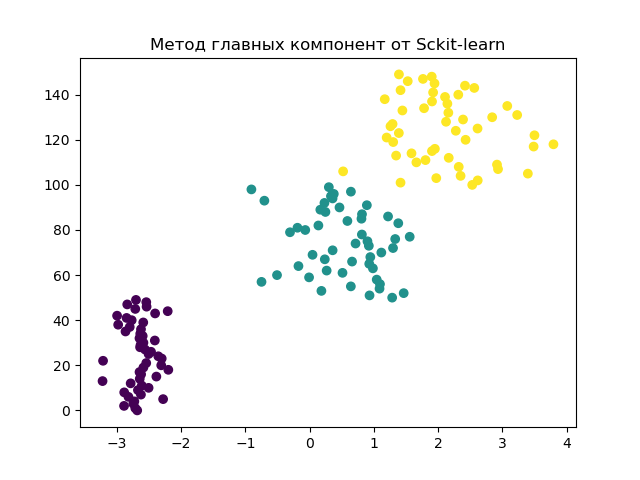
По критериям Кайзера и сломанной трости выбираем 1 главную компоненту.

Далее прогоним модель еще раз, указав количество главных компонент (одна), которое было выбрано согласно трем критериям, после чего посчитаем матрицу нагрузок:



Суммарная нагрузка для единственной главной компоненты составила ~ 3.26

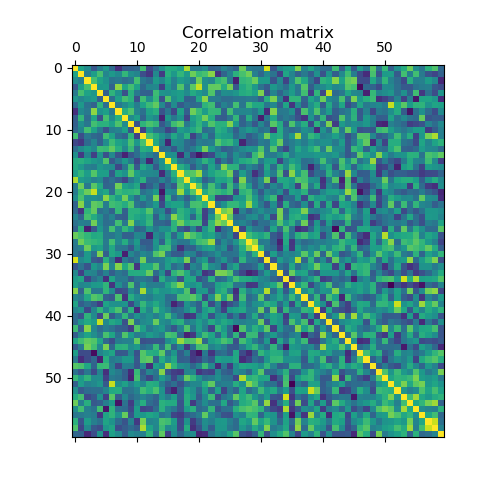
Результат работы метода главных компонент на графике:



Для данного датасета метод PCA отработал отлично. Получились линейно разделимые одномерные данные (по вертикальной оси отложен отрезок равный по длине количеству элементов датасета).

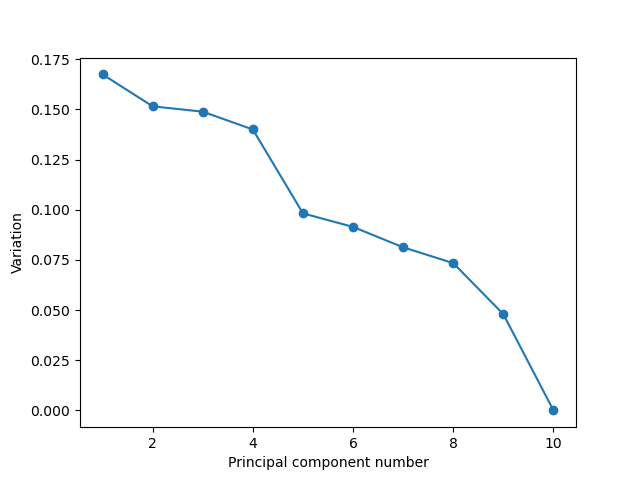
1. **Сгенерированный датасет с 60 признаками**

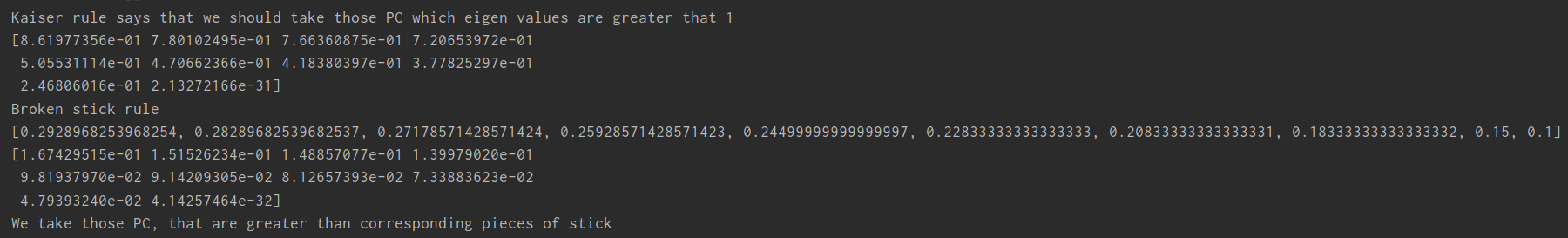
Матрица корреляций и список признаков, отсортированный по убыванию суммарной корреляции:



[44, 19.887824988288376], [17, 19.397768447023303], [50, 19.167291518404774], [54, 19.08307925116375], [55, 19.01108342318842], [20, 18.797846275023886], [33, 18.379143895652362], [19, 18.27298964084667], [46, 18.107610611168987], [47, 18.0885573334321], [59, 18.05008301710746], [27, 17.915829170321157], [35, 17.906818151477385], [22, 17.888150444140845], [37, 17.848562431077266], [9, 17.829653383844732], [58, 17.82202425323187], [34, 17.8086663991141], [5, 17.765409834766903], [42, 17.663352225263445], [28, 17.65219199985175], [53, 17.621048820210312], [1, 17.584430588829708], [26, 17.549199048237206], [56, 17.39757282648001], [38, 17.338256500910052], [43, 17.26284270700624], [25, 17.249219188843046], [8, 17.22969849384847], [32, 17.127815128471582], [3, 17.127477598113188], [52, 17.009504509153324], [18, 16.979802945402252], [49, 16.801788954455493], [31, 16.771340920749314], [40, 16.635594923708958], [45, 16.63535730819705], [21, 16.59981077828828], [7, 16.554258746407307], [11, 16.547878797585806], [0, 16.48889361538277], [2, 16.409872871034224], [36, 16.342744177930463], [23, 16.255259423009022], [24, 16.100513419086504], [29, 16.014314502923032], [39, 15.87181860016241], [14, 15.745216904346362], [16, 15.648459219522742], [12, 15.612377710056052], [51, 15.547983751099158], [48, 15.54363587890779], [4, 15.426699936147745], [57, 15.243301720634205], [6, 15.148164202075428], [10, 14.921968065449452], [30, 14.918781270911785], [41, 14.894451359345922], [15, 14.724980985803109], [13, 14.345663264722557]

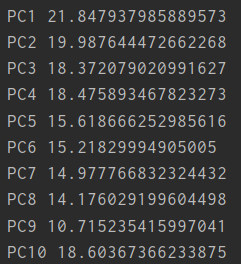
Критерии Кайзера, каменистой осыпи и сломанной трости:





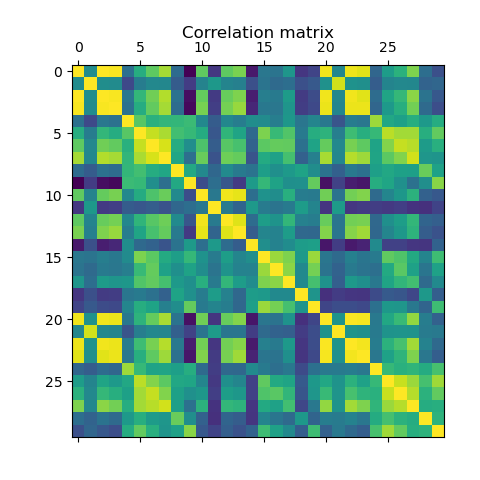
По всем трем критериям видно, что стоит оставить 10 компонент. Также всего десятью компонентами удалось описать всю дисперсию исходного датасета (то есть мы вообще не потеряли информацию).

Матрицу факторных нагрузок не превожу, так как на 1 графике уже становится неразборчиво, что и с чем коррелирует. Суммарная нагрузка каждой главной компоненты получилась следующая:



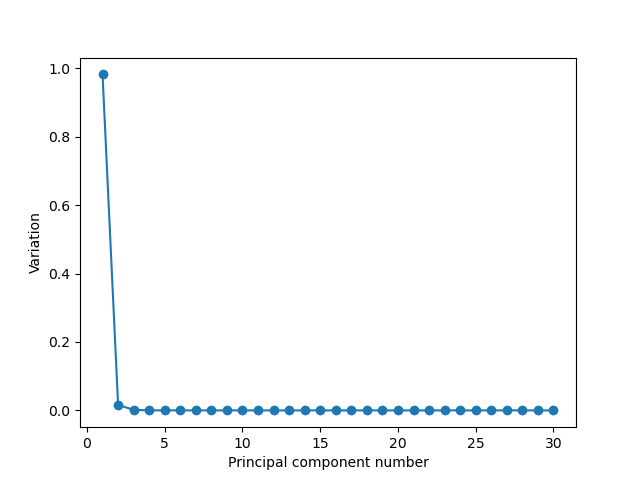
1. **Датасет Breast cancer**

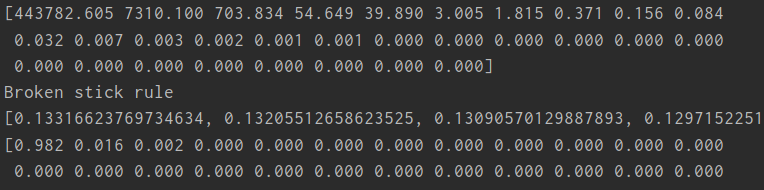
Матрица корреляций и список признаков, отсортированный по убыванию суммарной корреляции:



[6, 17.569165901064622], [7, 17.087954528979], [5, 16.8560501208885], [27, 16.774799372574897], [26, 15.741578587594848], [22, 15.545658163161336], [20, 15.116466380367932], [2, 14.848288377761403], [25, 14.848170681381578], [23, 14.752930725556851], [0, 14.424730626381303], [3, 14.355845037848889], [12, 14.06621015940526], [10, 13.650033534466328], [17, 13.315838171263344], [15, 13.134997986345255], [13, 13.12376413882894], [16, 11.646806633113362], [4, 10.937094773776703], [8, 10.826959084391445], [9, 10.472072444428338], [29, 10.461828286710162], [24, 9.888970172151947], [28, 9.595036842207444], [19, 9.379810621135018], [21, 8.141792512088076], [1, 7.893363343498187], [14, 6.7886114612534465], [18, 6.645739885421403], [11, 5.669495688298069]

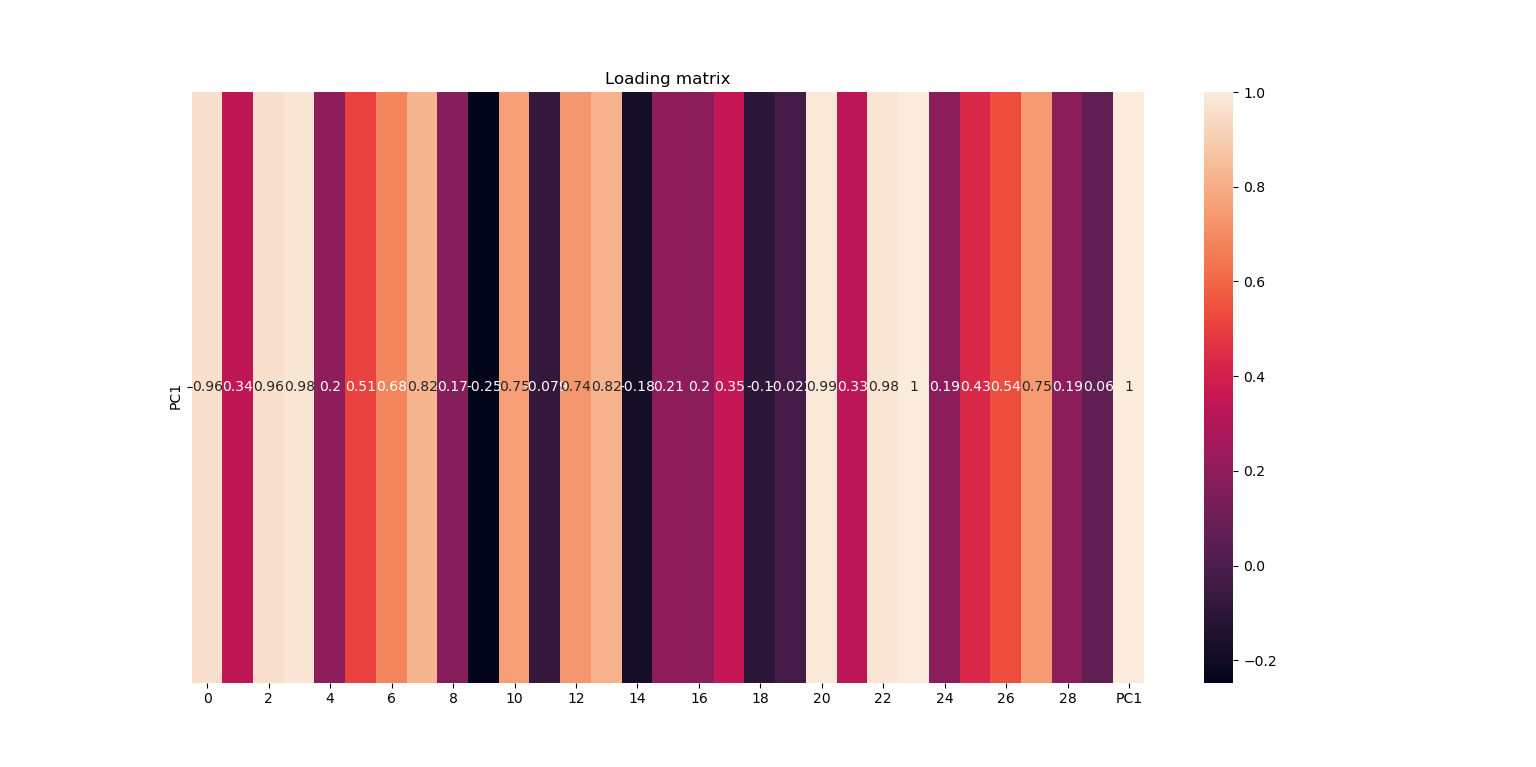
Критерии Кайзера, каменистой осыпи и сломанной трости:



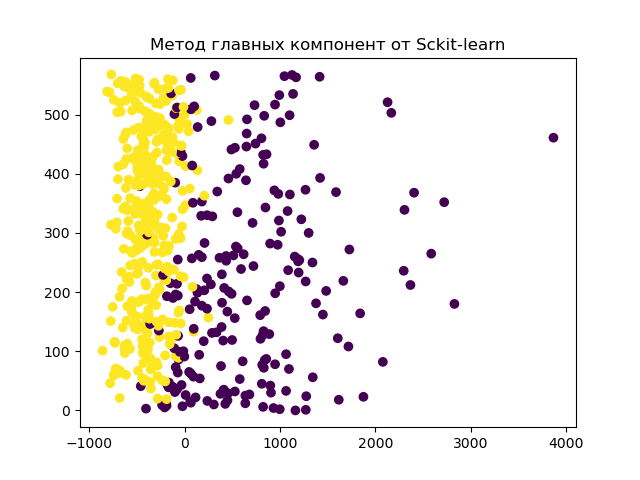


Судя по критериям, стоит оставить 1 главную компоненту

Суммарная нагрузка для главной компоненты получились равной ~14.77, а матрица нагрузок следующей:



Однако несмотря на то, что единственная главная компонента покрывает 98.2% дисперсии исходного датасета, хорошо разделить классы по ней не получится.



Если целью использования метода главных компонент является понижение размерности множества для упрощения дальнейшей классификации, то стоит попробовать выставить число главных компонент 2-3, а затем воспользоваться агломеративным методом кластеризации.