**Задание к лабораторной работе № 1.**

Используемые пакеты sklearn: datasets, decomposition, cross\_validation, neighbors, metrics, model\_selection, preprocessing

Для датасета breast\_cancer (входит в scikit):

I. Построить диаграмму рассеяния для двух произвольно взятых признаков.

Рассчитать матрицу ковариации исходного набора данных (X) и для исходного набора, спроецированного на главные компоненты (X\_reduced). Проверить, что главные компоненты ортогональны.

Сравнить собственные значения матрицы ковариации X со значениями дисперсии главных компонент.

Рассчитать total variation (след матрицы ковариации) для X и X\_reduced. Показать, что данный параметр не меняется при проецировании на главные компоненты.

Построить графики % объясненной дисперсии: а) для исходных признаков, б) для главных компонент.

II. Разделить выборку на тренировочную и тестовую (90% и 10%). Обучить классификатор на основе метода k ближайших соседей при фиксированном (произвольном) k. Оценить качество классификации на тренировочной/тестовой выборке. Сделать 10-fold кросс-валидацию при фиксированном k, оценить дисперсию. Построить графики зависимости точности на тренировочном/тестовом наборе от числа k (с дисперсией) для n-fold кросс-валидации для разных значений n (n = 2, 5, 8, 10). Сделать вывод о сравнительном качестве предсказаний при разных способах оценки точности. Определить наиболее приемлемый диапазон значений k (числа ближайших соседей).

III. Осуществить п. 2 не для исходной выборки, а для X\_reduced (число главных компонент варьировать, исследовать диапазон k, определенный в п. 2). Сравнить качество классификации при обучении классификатора по исходному набору/ по главным компонентам. Исследовать качество классификации при вариации числа главных компонент.