Задание 1.

[1] Выбрать один из датасетов из перечня:
- Iris
- Palmer Archipelago (Antarctica) penguin
- Wine Quality
- *Любой другой датасет, в котором есть три класса и четыре количественных (недискретных) признака.*

[2] Вывести в табличной форме статистику по датасету, включая
- Размерность всего датасета
- Количество признаков
- Количество целевых классов и объектов в каждом из классов
- Процент объектов с неопределенными признаками
- Иные ключевые характеристики датасета

Выбрать три класса и четыре количественных (недискретных) признака. Сформировать на их основе «отфильтрованный» датасет для дальнейшего анализа, удалив из датасета все объекты, для которых не определены значения хотя бы одного из выбранных четырех количественных признаков.

[3] Выполнить визуализацию датасета по всем парам выбранных количественных переменных, обозначая:
- в графиках с разными парами переменных объекты из разных классов различными по форме и цвету точками,
- в графиках с одной и той же парой переменных – гистограммы с достаточным числом разбиений (обычно – не менее 20), либо плотности распределения переменной по оси признака.

[4] В табличном варианте оценить степень сопряженности пар признаков-переменных на всем датасете, используя коэффициент корреляции Пирсона.
В табличном варианте оценить степень сопряженности пар признаков-переменных в каждом классе датасета, используя коэффициент корреляции Пирсона.

[5] Выбрать пару целевых классов и все количественные признаки.
Используя метод LDA (линейный дискриминантный анализ), построить решающую функцию алгоритма, разграниченные решающей функцией зоны и отдельные объекты классов на всех парах количественных признаков.

[6] Для одной из пар количественных признаков из пункта [5] на одном рисунке одновременно построить (а) решающую функцию LDA и (б) линейную регрессию одного количественного признака от другого.

[7] Выбрать два количественных признака и пару целевых классов
На отдельных рисунках с осями количественных признаков построить решающие функции, разграниченные решающей функцией зоны и отдельные объекты классов для методов (а) LDA, (б) SVM, (в) логистическая регрессия, (г) наивный байесовский классификатор

[8] Выбрать целевой класс и для каждого метода из пункта [7]:
- Вывести матрицу ошибок.
- Вывести значения sensitivity, specificity, precision, recall.
- Построить ROC кривую и рассчитать метрику AUC.

Задание 2.

[1] Используя make_blobs с любым random_state, сгенерировать датасет df1, в котором есть три класса с размером каждого класса 1000 и четыре количественных (недискретных) признака.

[2] Не забываем повторять шаги с задания 1
- ключевые характеристики датасета
- корреляции
- визуализация на всех парах переменных

[3] На основе созданного в пункте [1] датасета сгенерировать отдельные дополнительные датасеты (df2, df5, df10…), в которых объекты одного класса повторены 2 раза, 5 раз, 10 раз, 20 раз, 50 раз, 100 раз, 1000 раз, 10k раз, а количество объектов в остальных классов неизменно.

[4] Выбрать пару классов (включая класс с повторенными объектами) и пару количественных признаков.
Используя метод LDA (линейный дискриминантный анализ), для каждого из датасетов df1, df2, df5, df10, df20, df50, df100, df1000, df10k, построить решающую функцию алгоритма, разграниченные решающей функцией зоны и отдельные объекты классов.

[5] Повторить пункт **[4]** для алгоритма SVM.

[6] Для каждого из датасетов df1, df2, df5, df10, df20, df50, df100, df1000, df10k из пункта **[4]** восстановить в таблицу координаты следующих точек:
- центр отрезка, соединяющего центры масс выбранных классов
- общий центр масс выбранных классов
- точку пересечения решающей функции и отрезка, соединяющего центры масс выбранных классов.

В виде графиков визуализировать зависимости между количеством повторов в классе с повторенными объектами и координатами найденных точек.

[7] Выбрать целевой класс для решений из пункта **[4]**.
Для каждого из решений из пункта **[4]**:
- Построить ROC кривую и рассчитать метрику AUROC.
- Построить PR кривую и рассчитать метрику AUPRC.
- (*) Построить PRgain кривую и рассчитать метрику AUPRgainC.

[8] В пункте **[7]** выбрать другой целевой класс.
- Построить ROC кривую и рассчитать метрику AUROC.
- Построить PR кривую и рассчитать метрику AUPRC.
- (*) Построить PRgain кривую и рассчитать метрику AUPRgainC.

[9] Для датасета 10k на основе 3-fold, 5-fold, 10-fold, 20-fold, 50-fold, 100-fold кросс-валидации построить кривые AUROC и AUPRC с доверительными интервалами (CI95). Вместо CI95 можно взять CI90, CI80 или другой вариант доверительного интервала.

https://stackoverflow.com/questions/55541254/precision-recall-curve-with-n-fold-cross-validation-showing-standard-deviation
https://stackoverflow.com/questions/29656550/how-to-plot-pr-curve-over-10-folds-of-cross-validation-in-scikit-learn