Задание 1. Intro.

[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. LDA.

[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

Задание 3. LogReg.

[1] *(аналогично Заданию 2)* Используя make_blobs с любым random_state, сгенерировать датасет df, в котором есть ***три*** класса с размером каждого класса ***100***, ***четыре*** количественных (недискретных) признака ,а центры классов зафиксированы в следующих точках: Класс 0 – (+1, +1, +1, +1), Класс 1 – (-1,-1,-1,-1), Класс 2 – (-1,+1,-1,+1).

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

[3] На основе созданного в пункте [1] датасета сгенерировать отдельные дополнительные датасеты (**df_A_B**), в которых к классу 0 добавлено **A** одинаковых точек с координатами (+**B,-B,+B,-B**), где A = 1, 10, 100 и B = 5, 10, 20, при этом количество объектов в остальных классах неизменно.

[4] Выбрать пару классов (включая класс с повторенными объектами) и один количественный признак.
Для каждого из датасетов df и df_A_B в своем пространстве X-Y(количественный признак-класс) построить и визуализировать объекты, линию линейной регрессии и линию логистической регрессии. Регрессии строить на паре X-Y(количественный признак-класс).
Оценить качество работы полученных на основе логистической регрессии классификаторов, используя ROC кривые, восстановив на графике ROC кривых точку классификации Sensitivity-Specificity и доверительные интервалы CI95 бутстрепом (n=1000).
Все полученные графики возможно расположить в две колонки: левая колонка – визуализация пространства и регрессий, правая – графики с ROC кривой и точкой.

[5] Для выбранной пары классов (включая класс с повторенными объектами) на всех признаках вычислить уравнение множественной линейной регрессии, где Y – класс, Xi – признаки.
Используя полученные уравнения множественной линейной регрессии, на основе каждого из датасетов df и df_A_B сформировать новые датасеты logdf и logdf_A_B, в каждом из которых есть только один признак X, сформированный на основе соответствующего уравнения множественной линейной регрессии, а переменная Y – отнесение к классу.

[6] Для каждого из датасетов logdf и logdf_A_B в своем пространстве X-Y (количественный признак-класс) построить и визуализировать объекты, линию простой линейной регрессии и линию логистической регрессии. Регрессии строить на паре X-Y(количественный признак-класс).
Оценить качество работы полученных на основе логистической регрессии классификаторов, используя ROC кривые и восстановив на графике ROC кривых точку классификации Sensitivity-Specificity и доверительные интервалы CI95.
Все полученные графики возможно расположить в две колонки.

[7] Для выбранной пары классов (включая класс с повторенными объектами), визуализировать один из датасетов df_A_B на всех парах переменных, построив на графиках объекты, линии множественной линейной регрессии, разделение классов на основе логистической регрессии решающей функцией с contour_plot уровнями классификации.

anwarruff.com/exploring-logistic-regression/
www.alpha-quantum.com/blog/logistic-regression/logistic-regression-from-scratch/

Задание 4. PCA, tSNE, UMAP

[1] Сгенерировать двумерные датасеты df_a (размерности 4x4, 16x16, 256x256), в которых в каждой строке x начиная с колонки y(=x) значения следующих (a/2) ячеек равны «1» *(если достигнут конец строки датасета, сверх размера датасета не заполняем, на новую строку не переходим)*, а оставшиеся ячейки датасета равны «0». Выполнить визуализацию датасетов в пространстве X-Y.

[2] Используя метод главных компонент (основанный на корреляциях), для каждого датасета df_a отдельно визуализировать:
- Scree plot;
- Отображения отдельных объектов-строк в двумерных пространствах первых компонент от 1 до a (a на выбор от 4 до 8) – без отображения старых векторов-переменных;
- Отображения отдельных объектов-строк в двумерных пространствах первых компонент от 1 до a (a на выбор от 4 до 8) – с отображением старых векторов-переменных.
- Отображения отдельных объектов-строк в двумерных пространствах первых компонент от c до c+a (a на выбор от 4 до 8, c – любая со 129) – без отображения старых векторов-переменных;
- Отображения отдельных объектов-строк в двумерных пространствах первых компонент от c до c+a (a на выбор от 4 до 8, c – любая со 129)– с отображением старых векторов-переменных.

[3] Используя методы UMAP, PacMAP и tSNE, для каждого датасета df_a отдельно визуализировать проекции с нескольких рендом-стартов, а также после предвращения датасета методом PCA.

[4] Используя make_blobs, сгенерировать на основе номеров строк (параметр x), столбцов (параметр y) и значений («0», «1») ячеек датасета df_256 новый датасет df_256v, в котором в каждой точке с центром (x,y) и дисперсией d (d – на выбор, в диапазоне от 5 до 20) сформировано по 10 случайных точек класса «1» или «0».

[5] Используя метод главных компонент (основанный на корреляциях), для датасета df_256v выполнить визуализации аналогично первым трем подпунктам пункта [2].

[6] Для датасета df_256v повторить пункт [3].

[7] Взять датасет с Kaggle, в котором есть >=10 переменных, >=10000 объектов и несколько классов. Для каждого датасета не забываем повторять шаги с задания 1 - ключевые характеристики датасета, корреляции, визуализация на всех парах var.

[8] Используя метод главных компонент (основанный на корреляциях), для датасета из пункта [7] выполнить визуализации аналогично первым трем подпунктам пункта [2]. Определить различными способами (метод Кайзера, метод ломанной трости), сколько компонент необходимо оставлять для регрессионной модели.

[9] Для датасета из пункта [7] повторить пункт [3].

https://umap-learn.readthedocs.io/en/latest/
https://github.com/YingfanWang/PaCMAP