Домашнее задание №3

В данном домашнем задании вам необходимо будет научиться кластеризовать данные о пациентах с заболеваниями спины. В данных приведены измерения по 6 переменным:

- pelvic incidence
- pelvic tilt
- lumbar lordosis angle
- sacral slope
- pelvic radius
- degree_spondylolisthesis
- class

Для кластеризации в данном задании вам предлагается взять переменные *pelvic_radius* и *degree_spondylolisthesis*. Однако, вы можете поэкспериментировать и предложить другие переменные.

1. Нарисуйте Scatter-Plot для базовой классификации (данные лежат в переменной *class*). (1 балл)

NB! Для следующий операций создайте отдельный датасет, который не содержит переменной *class*.

2. Оцените, какое количество кластеров будет оптимальным для этих данных. Изобразите график. (2 балла)

3. Напишите классификатор Kmeans на основании данных об оптимальном числе кластеров, проведите кластеризацию на ваших данных. Визуализируйте полученные результаты и отразите центроиды на графике. (2 + 2 балла)

4. Напишите иерархический классификатор (агломеративный алгоритм) , попробуйте подобрать метод, который лучше всего будет кластеризовать. (2 + 2 балла)

5. Визуализируйте полученные результаты иерархической кластеризации. Предположите, какой из них работает лучше всего. (1 + 1 балл)

6. Сравните метрики качества модели (*'ARI', 'AMI','Homogenity' 'Completeness', 'V-measure' 'Silhouette'*). На основании полученных метрик сделайте вывод о том, какой алгоритм использовать лучше всего (1 + 2 балла).