**ЛР 3 – Кластеризация**

**Задание на ЛР №3**

В соответствии с индивидуальным заданием (вариантом), размещенным на странице «ЛР 3, Вариант хх» в записной книжке команды MS Teams, выполнить следующие работы:

1. Считайте из заданного набора данных репозитария UCI значения двух признаков и метки класса.
2. Если среди меток класса имеются пропущенные значения, то удалите записи с пропущенными метками класса. Если в признаках имеются пропущенные значения, то замените их на медианные значения того класса, к которому относится запись с пропущенным значением в признаке.
3. Если количество различных меток класса больше пяти, то объедините некоторые (наименее многочисленные) классы, чтобы общее количество классов не превышало пять.
4. Визуализируйте набор данных в виде точек плоскости с координатами, соответствующими двум признакам, отображая точки различных классов разными цветами. Подпишите оси и рисунок, создайте легенду набора данных.
5. Проведите кластеризацию набора данных из двух признаков с помощью алгоритмов, указанных в индивидуальном задании, для случая, когда количество кластеров равно количеству классов в исходном наборе (с учетом корректировки). В случае отсутствия сходимости алгоритма измените аргументы по умолчанию или используйте для кластеризации случайную выборку из набора данных.
6. Для каждого из алгоритмов кластеризации, указанных в индивидуальном задании, постройте матрицу сопряженности, используя функцию contingency\_matrix() из scikit-learn, и найдите значения мер качества кластеризации, указанные в индивидуальном задании, на основании данных в матрице сопряженности, не используя другие функции из scikit-learn или других фреймворков.
7. Определите алгоритм кластеризации, оптимальный с точки зрения меры качества кластеризации, указанной в индивидуальном задании.
8. Для оптимального алгоритма кластеризации из предыдущего пункта визуализируйте области принятия решения и набор данных в виде точек на плоскости с координатами, соответствующими двум признакам, отображая точки различных кластеров разными цветами. Подпишите оси и рисунок, создайте легенду набора данных.

Результат лабораторной работы оформить в виде отчета согласно шаблона (шаблон отчета находится в учебных материалах команды в формате .ipynb). Обязательно сопровождать представленный программный код комментариями и ссылками на выполненные пункты задания. Отчет представить в виде файлов .ipynb и .pdf (файл .pdf получить распечаткой файла .ipynb). Не архивировать файлы.

Отчет по лабораторной работе представить как результат выполнения задания MS Teams (представить файл с отчетом и файл с программой). **НЕ ЗАБЫТЬ НАЖАТЬ НА КНОПКУ СДАЧИ ЗАДАНИЯ**.

Меры качества:

* Чистота (purity)
* F-мера
* условная энтропия,
* парные меры TP, FN, FP, TN,
* индекс Фоулкса – Мэллоуса
* индекс Жаккара
* индекс Rand