***Лабораторная работа №2***

**ИЗУЧЕНИЕ МЕТОДОВ ВИЗУАЛИЗАЦИИ И КЛАСТЕРИЗАЦИИ ДАННЫХ**

**Описание датасетов**

**1.**Набор данных ***«Рост-Вес-Возраст-Позиция»*** (см. Л.Р.№1).

**2.** Набор данных ***«Iris Data Set»* –** ***«Ирисы Фишера»*** (см. Л.Р.№1)***.***

**Задание на лабораторную работу**

1. Провести визуализация данных в трехмерном пространстве (*3D Plot*). Для выборки «Iris» обосновать выбор трех измерений (сделать дополнительные исследования, например карта корреляций). Сравнить трехмерные диаграммы для выборок и сделать предварительные выводы (Исходное количество кластеров считаем неизвестным). В отчёт включить обоснование выбора измерений (результаты исследований и вывод по ним) и графики.
2. Провести иерархическую кластеризацию. Исследовать зависимость результатов иерархической классификации от выбора меры близости (евклидово расстояние, манхэттенское расстояние, расстояние Чебышева, косинусное) и правила объединения кластеров (одиночная связь, полная связь, невзвешенная средняя связь, невзвешенная центроидная связь, метод Уорда). Проанализировать диаграмму изменения расстояний при объединении кластеров. Оценить предположительное число кластеров, на которое разделяется исследуемая совокупность. В отчёт включить наилучшие результаты и обосновать почему они наилучшие (сравнить с плохим/средним результатом). Так же включить обоснование предположительного числа кластеров.
3. Построить проекцию исследуемой совокупности многомерных наблюдений на плоскость с помощью *метода главных* *компонент (PCA)*. Оценить предположительное число кластеров, на которое разделяется исследуемая совокупность. Проанализировать количество выделенных компонент и точность метода*.* Проанализировать взаимосвязь между полученными главными компонентами и исходными признаками*.* В отчёт включить результаты и обоснование предположительного числа кластеров.
4. Построить проекцию исследуемой совокупности многомерных наблюдений на плоскость с помощью *стохастического вложения соседей с t-распределением (t-sne)*. Оценить предположительное число кластеров, на которое разделяется исследуемая совокупность. Сравнить полученные результаты с результатами предыдущего пункта. В отчёт включить результаты и обоснование предположительного числа кластеров.
5. На основе проведенных исследований сформулировать и обосновать предположение о числе кластеров. (Сделать общий вывод с примерами)
6. Провести неиерархическую кластеризацию с помощью центроидного метода k-средних и проверить выдвинутое предположение о числе кластеров (для этого требуется посмотреть большее и меньшее количество кластеров).
7. Провести визуализация данных в трехмерном пространстве (*3D Plot*) (как в пп. 1). При отображении указать принадлежность наблюдения к кластеру (раскрасить точки на графике в соответствии с принадлежностью к кластеру):
   * Исходного распределения
   * Распределения полученного в пп. 6

Сравнить полученные графики. (В отчёт включить оба графики и вывод по полученным результатам)

1. Сравнить результаты, пункта 5 и 6, сделать окончательные выводы о структуре исследуемых выборок.