Лабораторная работа №3

Выполнил: Поздняков Иван

Ссылка на гитхаб:

1. Необходимо реализовать три метода кластерного анализа:
   * k-means
   * Неиерархический метод
   * Иерархический агломеративный метод

С евклидовой и манхеттенской метрикой

Ответ:

* Для k-means я использовал библиотеку sklearn.cluster с евклидовой метрикой
* Для иерархической я использовал linkage из sklearn
* Для Неиерархического я использовал k-means с манхетоннской метрикой

Реализация представлена на гитхаб

2. Реализовать вычисление не менее 4 различных показателей качества разделения

Ответ:

2.1 Для анализа результатов применяется силуэт

**Анализ силуэта баллов**

Счет имеет диапазон [-1, 1]. Ниже приводится анализ этой оценки —

* **Оценка +1** — Оценка около +1 означает, что выборка находится далеко от соседнего кластера.
* **Оценка 0** — Оценка 0 указывает на то, что выборка находится или очень близка к границе решения между двумя соседними кластерами.
* **Оценка -1** — Отрицательная оценка означает, что выборки были назначены в неправильные кластеры.

2.2) также я использовал точность, применяя проверочные БД

2.3) Также использовал: precision, recall, f1-score, support

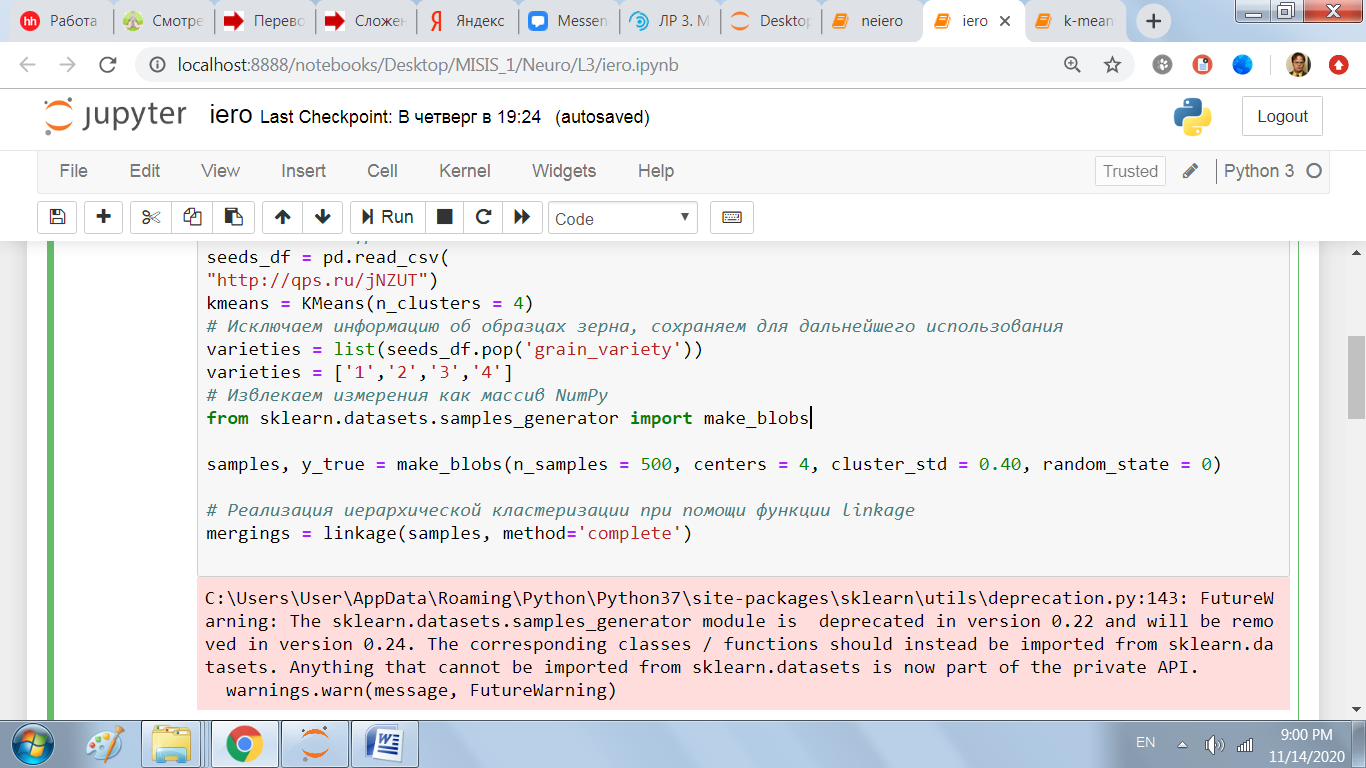
2.4) macro avg и weighted avg

3. Провести сравнительный анализ качества работы методов кластерного анализа на различных датасетах (сгенерированных самостоятельно и эталонных датасетах).

* линейно разделимые множества (с расстоянием между группами в 10^3 раз больше, чем диаметр группы)
* линейно разделимые множества (группы расположены близко или касаются друг друга)
* линейно неразделимое множество (средняя площадь пересечения классов 10-20%)
* линейно неразделимое множество (средняя площадь пересечения классов 50-70%)
* 2 эталонных датасета с различным количеством признаков, но не менее 7

Ответ:

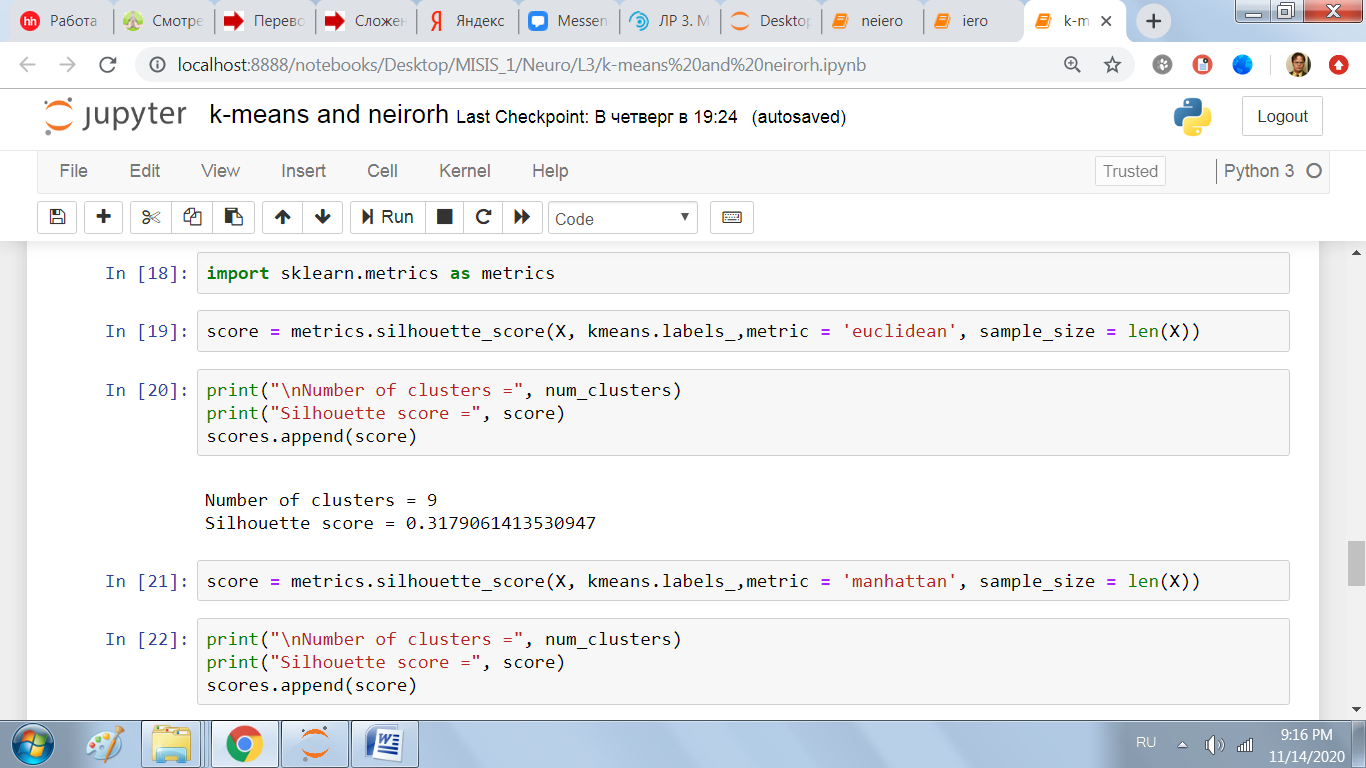
Я создавал собственные датасеты, используя make\_blobs из sklearn.datasets.samples\_generator

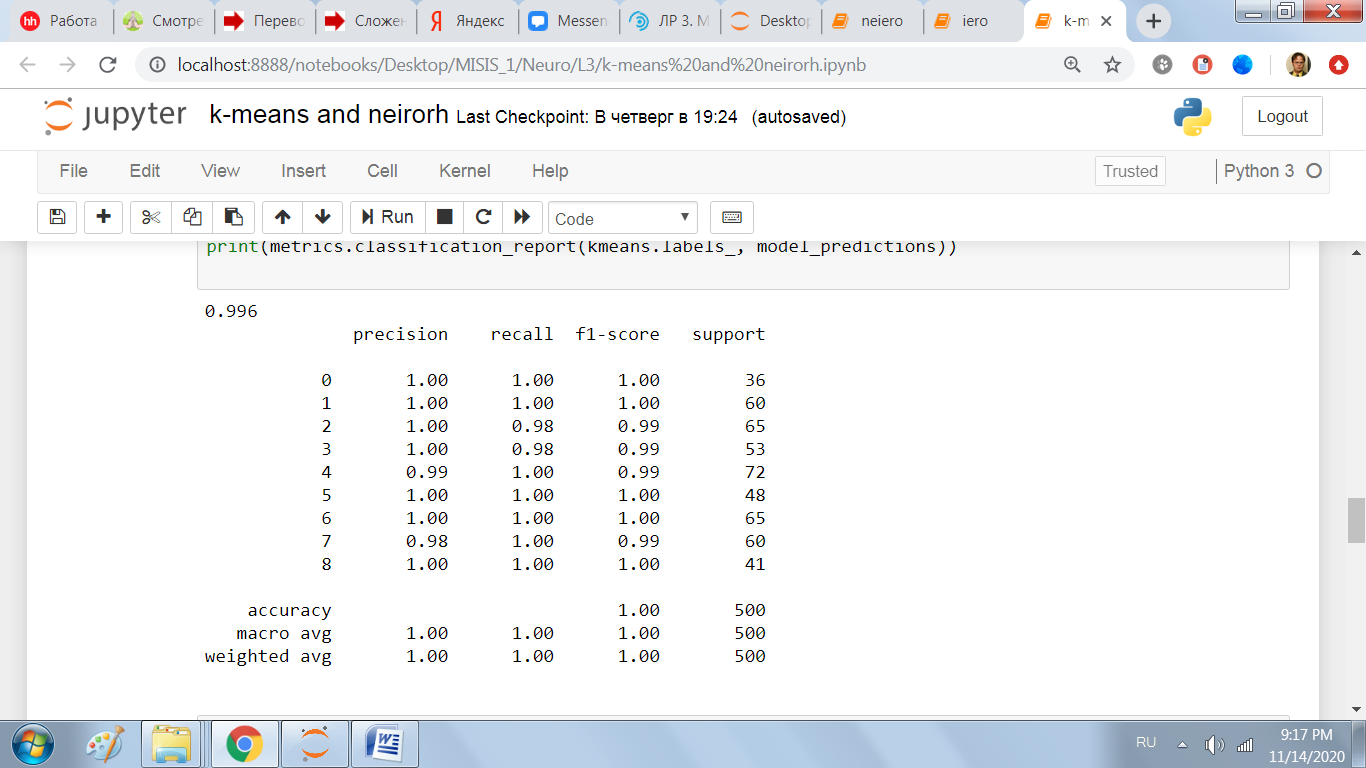


Задавал количество элементов, количество классов и уровень отдалённости классов.

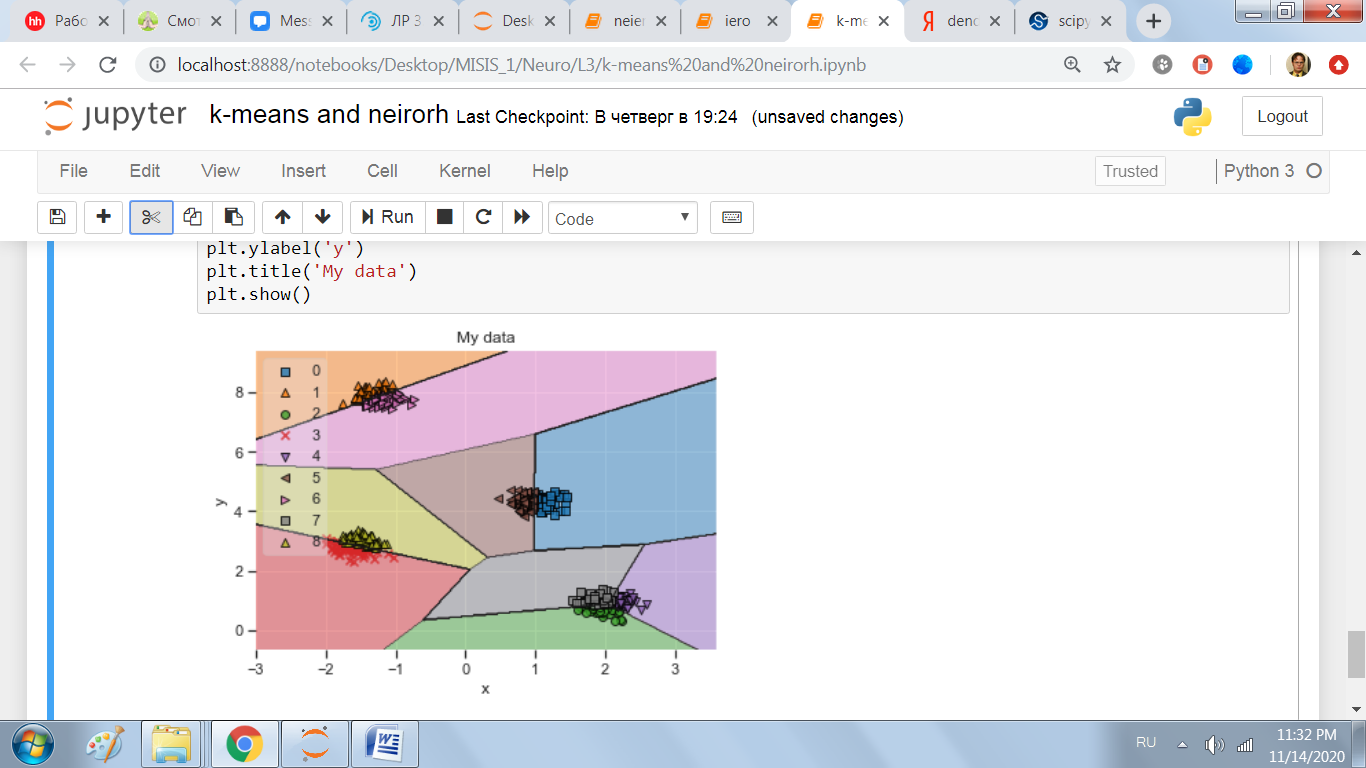
1. Линейно разделимые множества (с расстоянием между группами в 10^3 раз больше, чем диаметр группы)

Показатели K-means



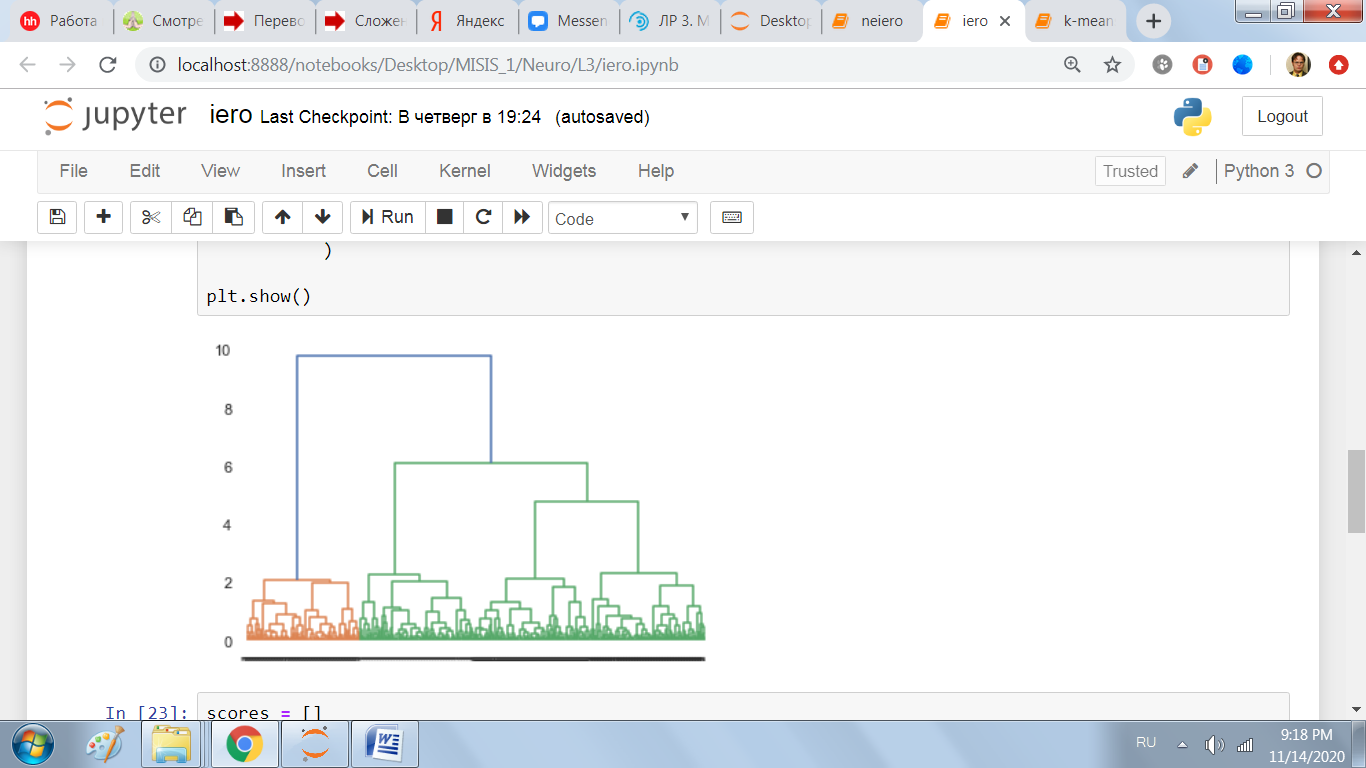


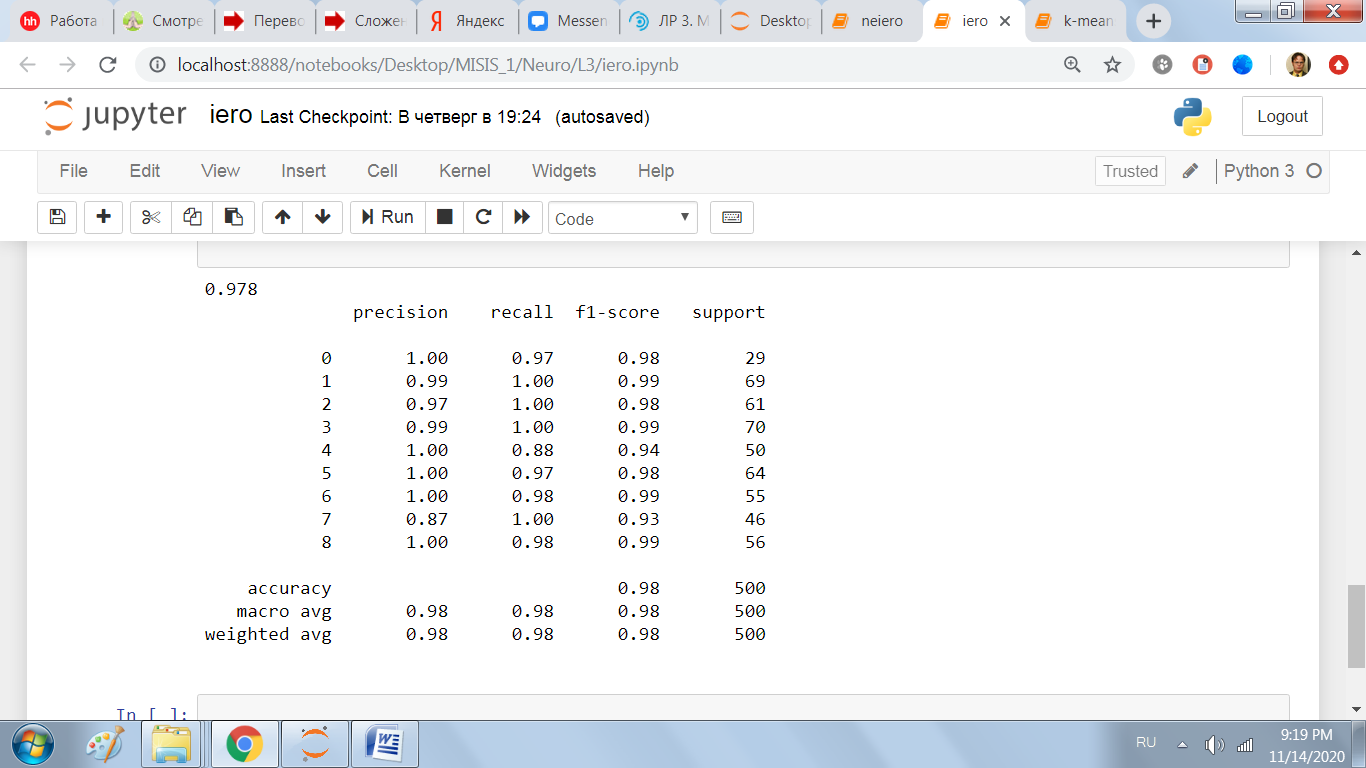
Для отображения передал свою модель и проверочную дату

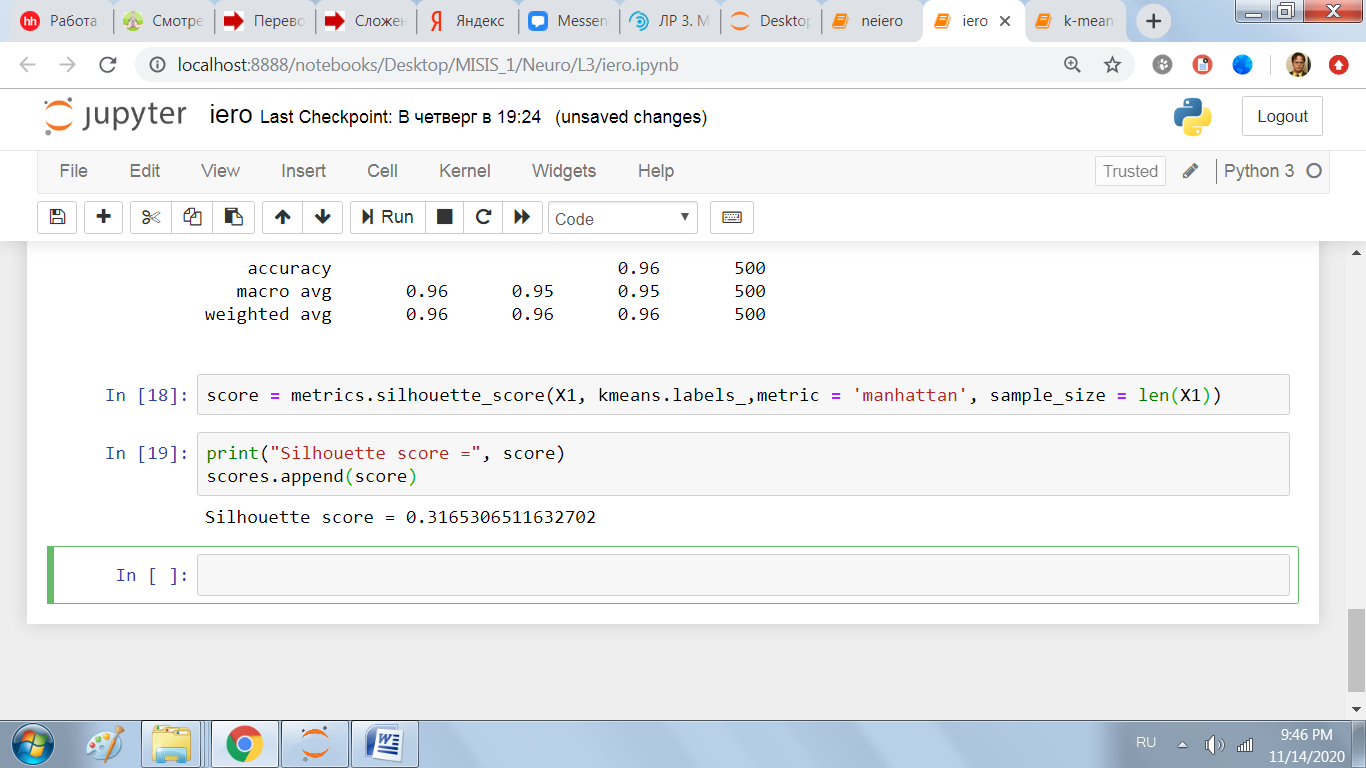


Иерархическая

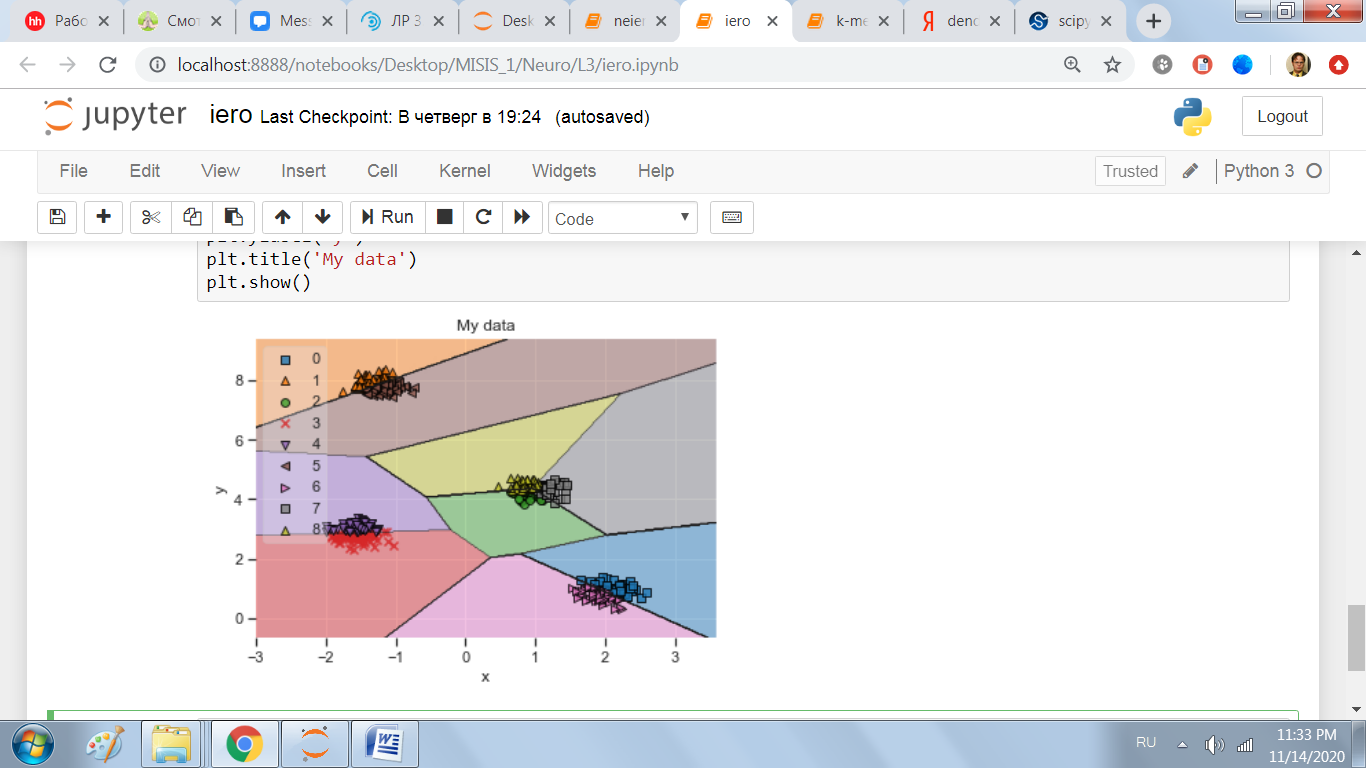
Применил обученную модель для тестового датасета, используя функцию dendrogram



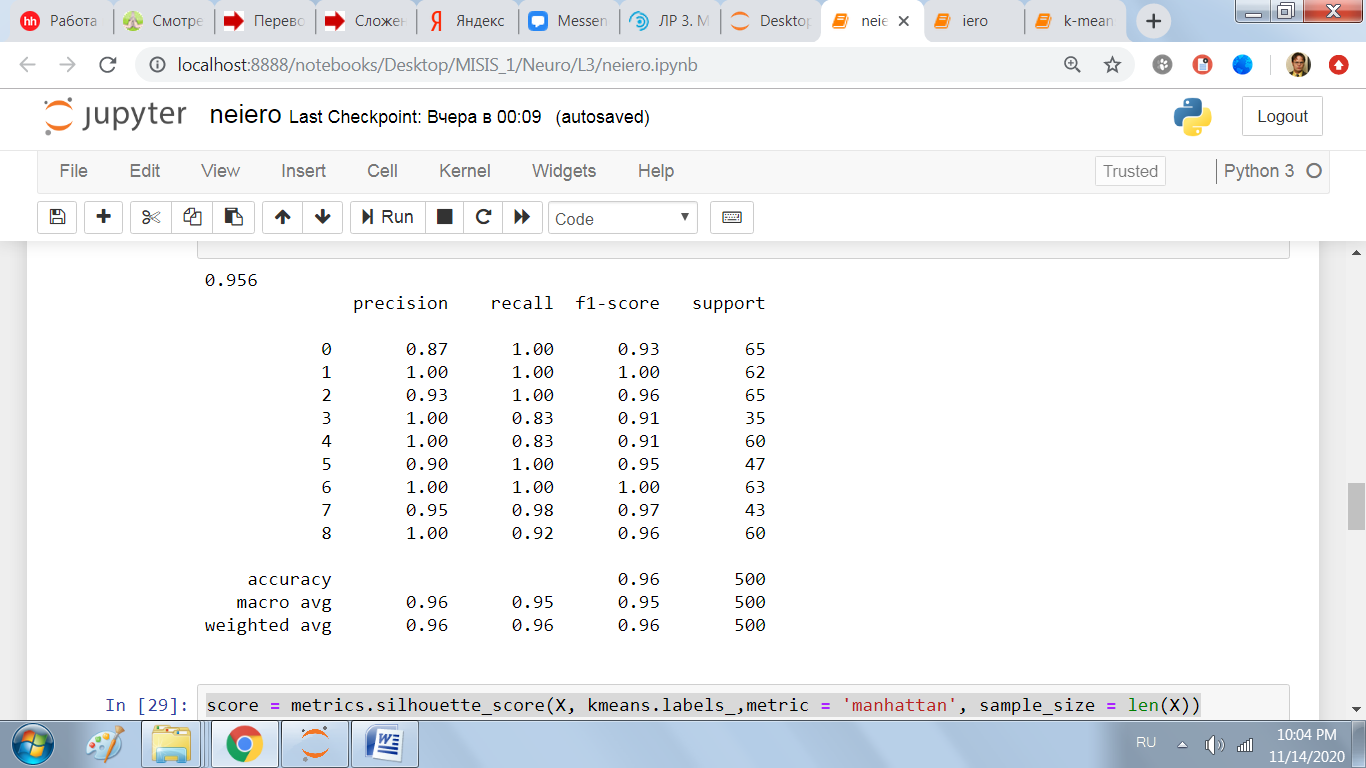


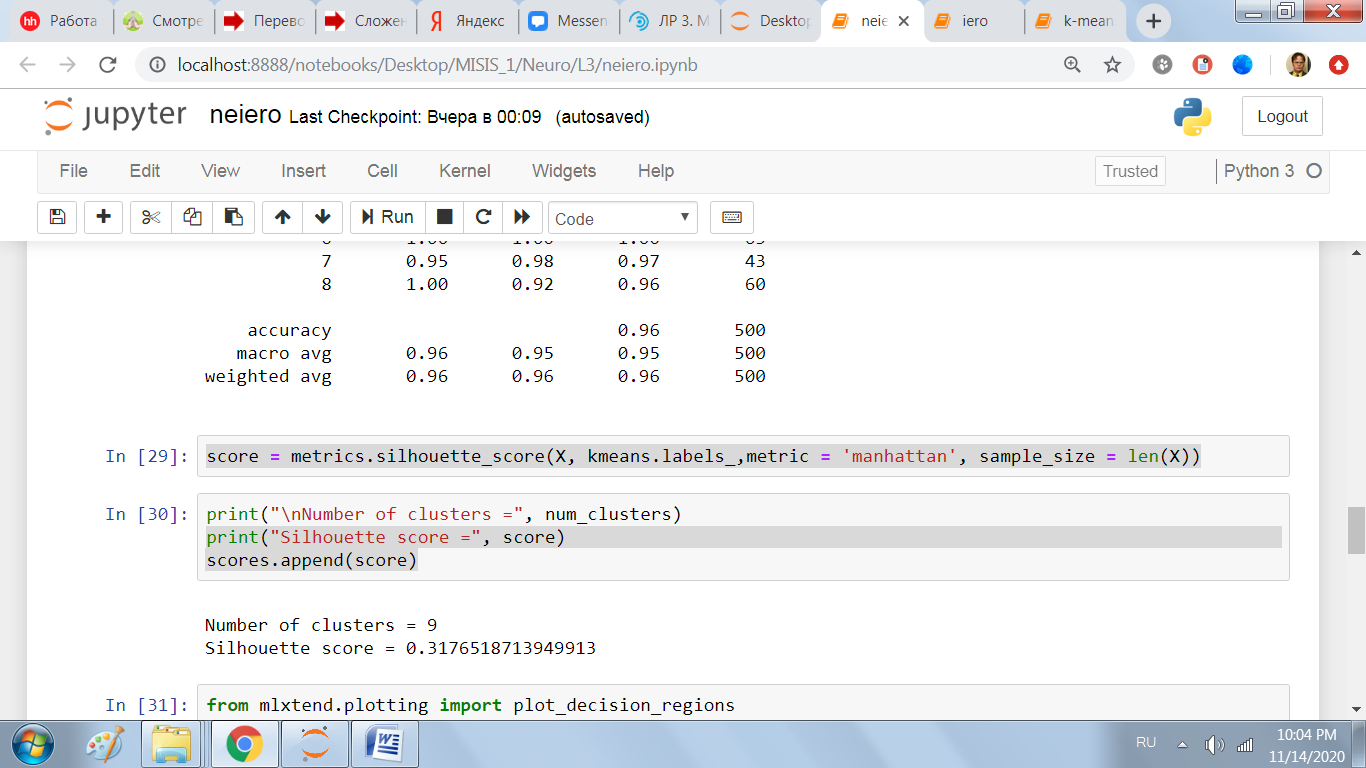


Для отображения передал свою модель и проверочную дату

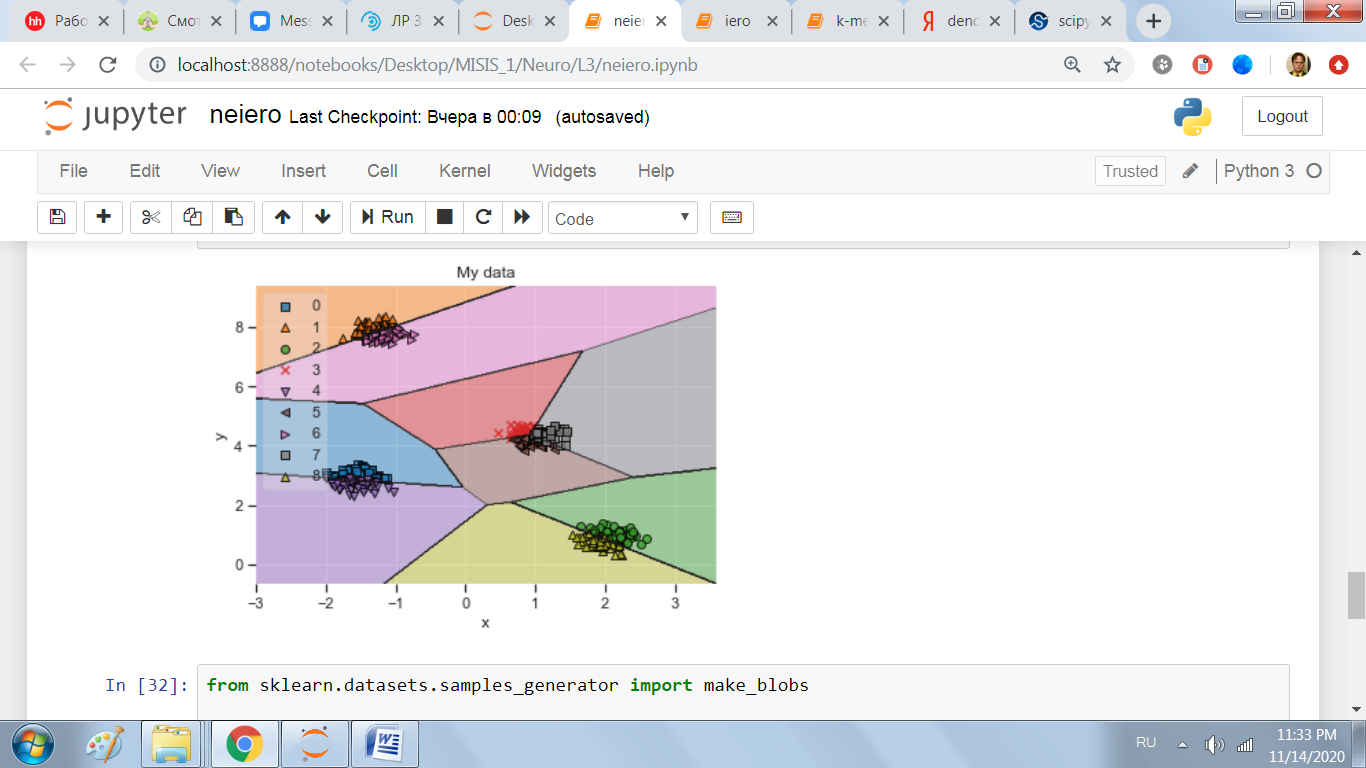


Неирархическая





Для отображения передал свою модель и проверочную дату

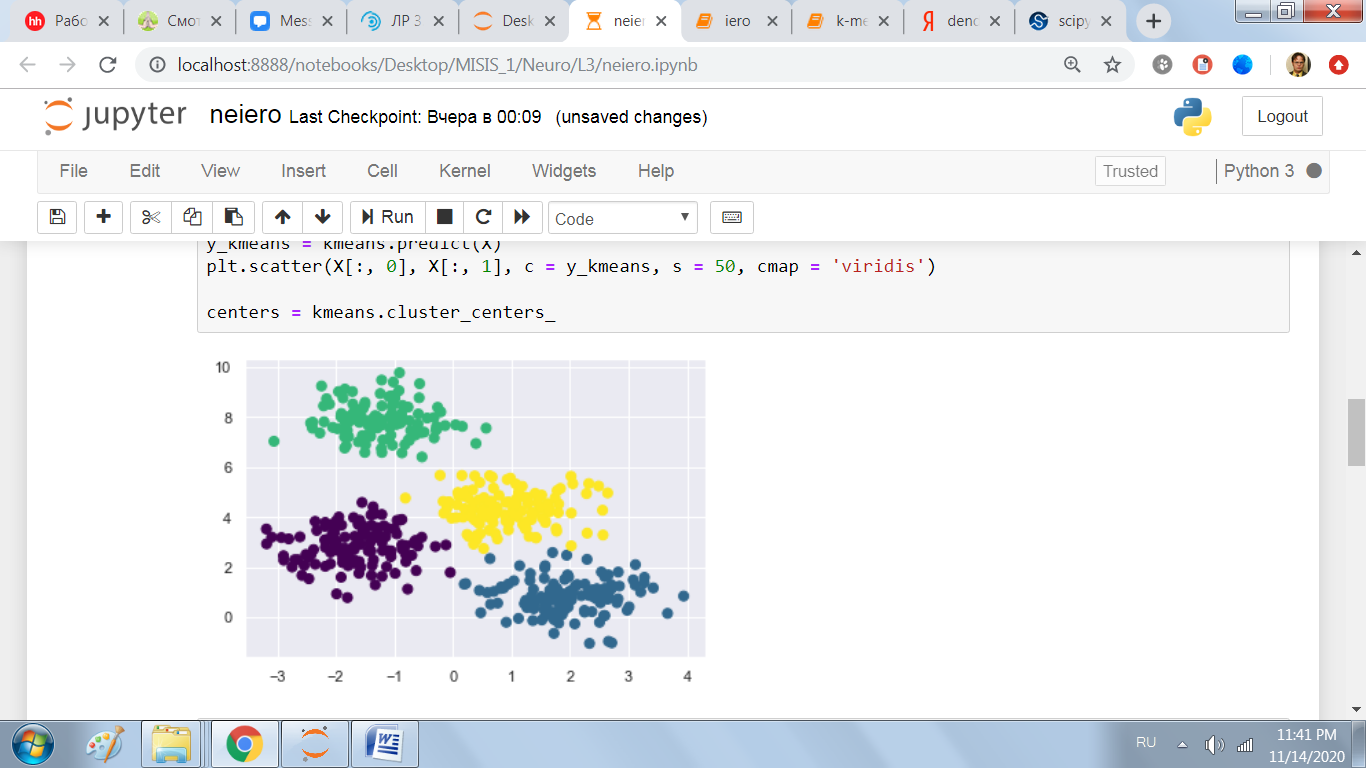


2) линейно разделимые множества (группы расположены близко или касаются друг друга)

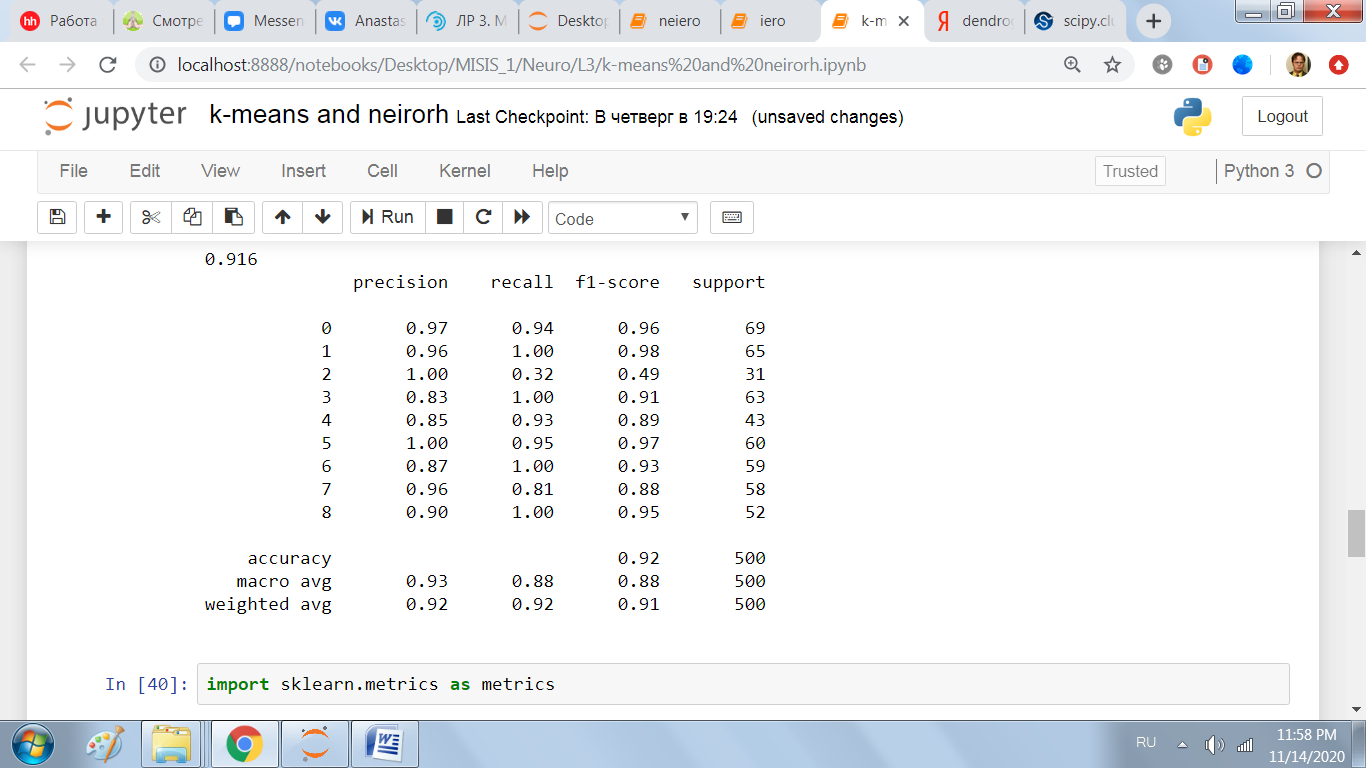
Ответ:

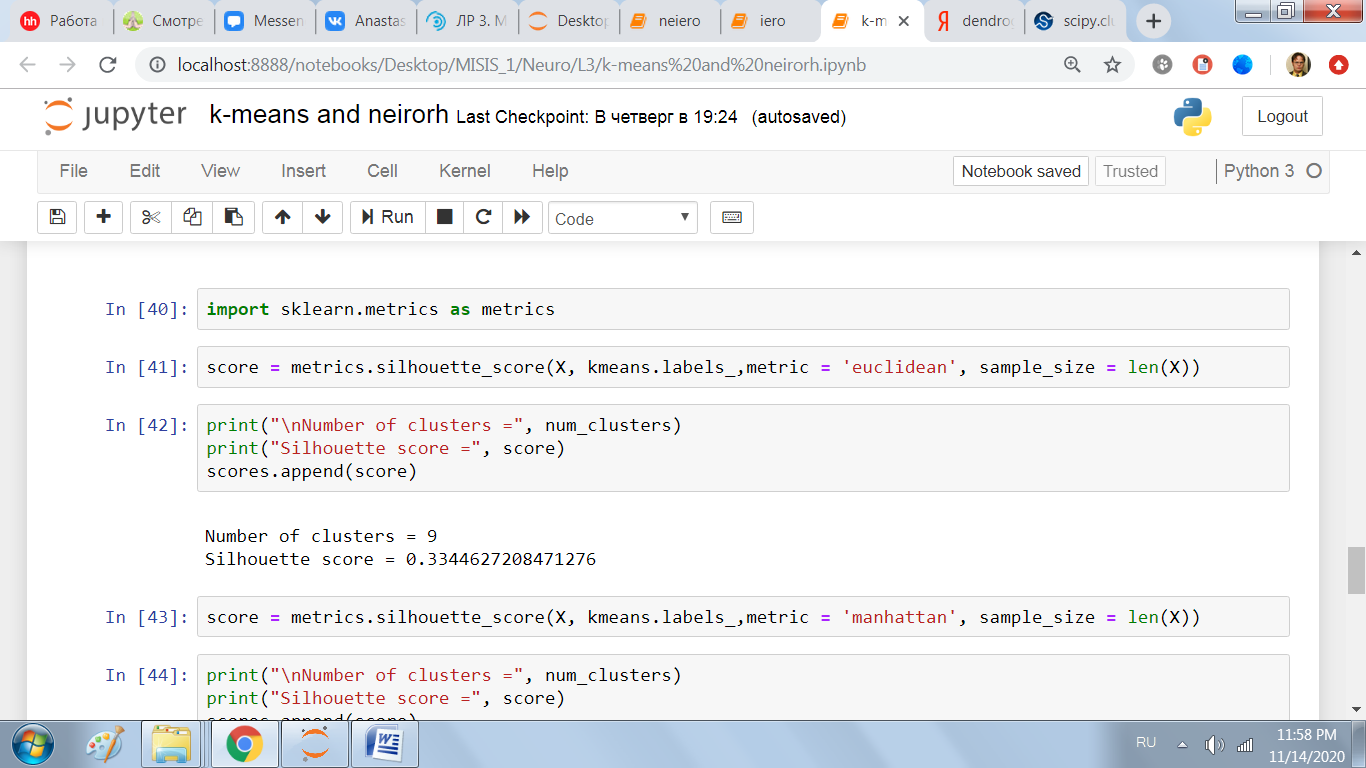
Регулирую даты через параметры

Получаю

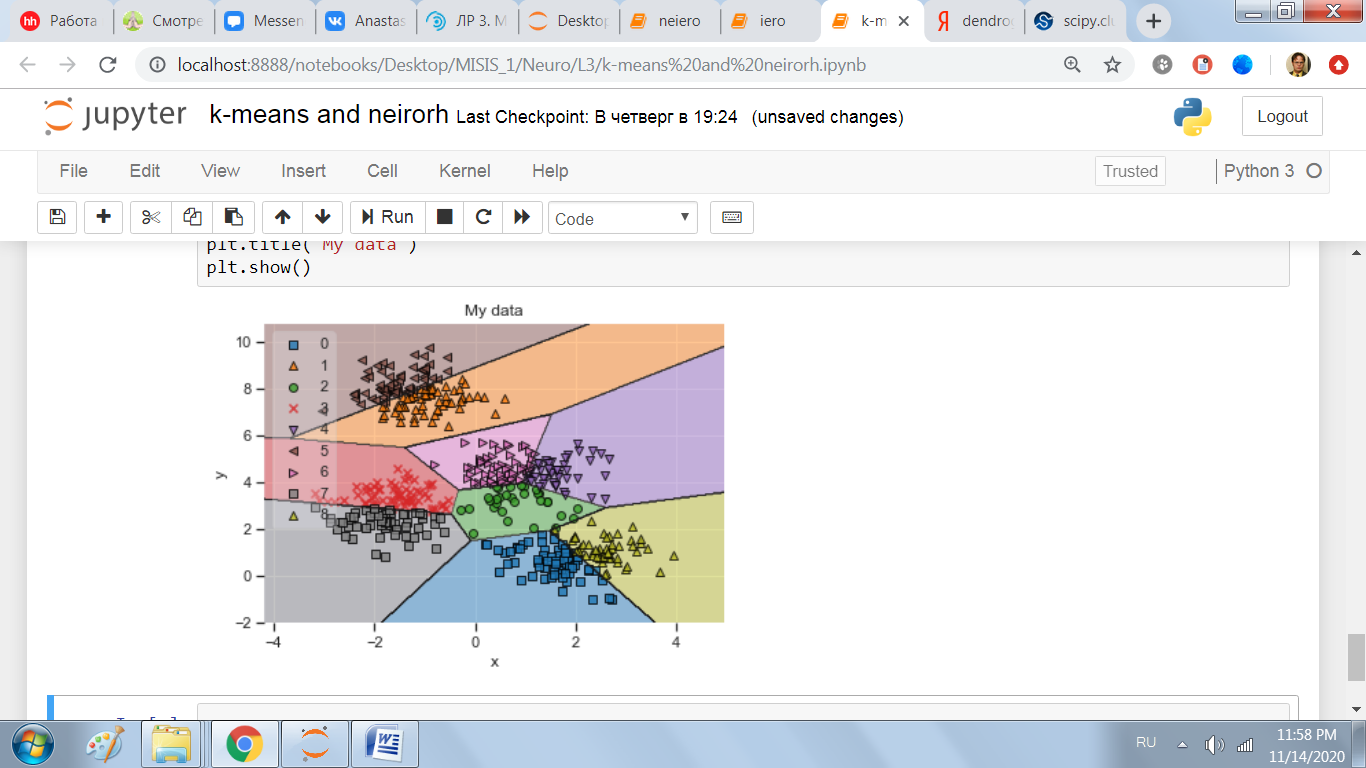


Показатели K-means



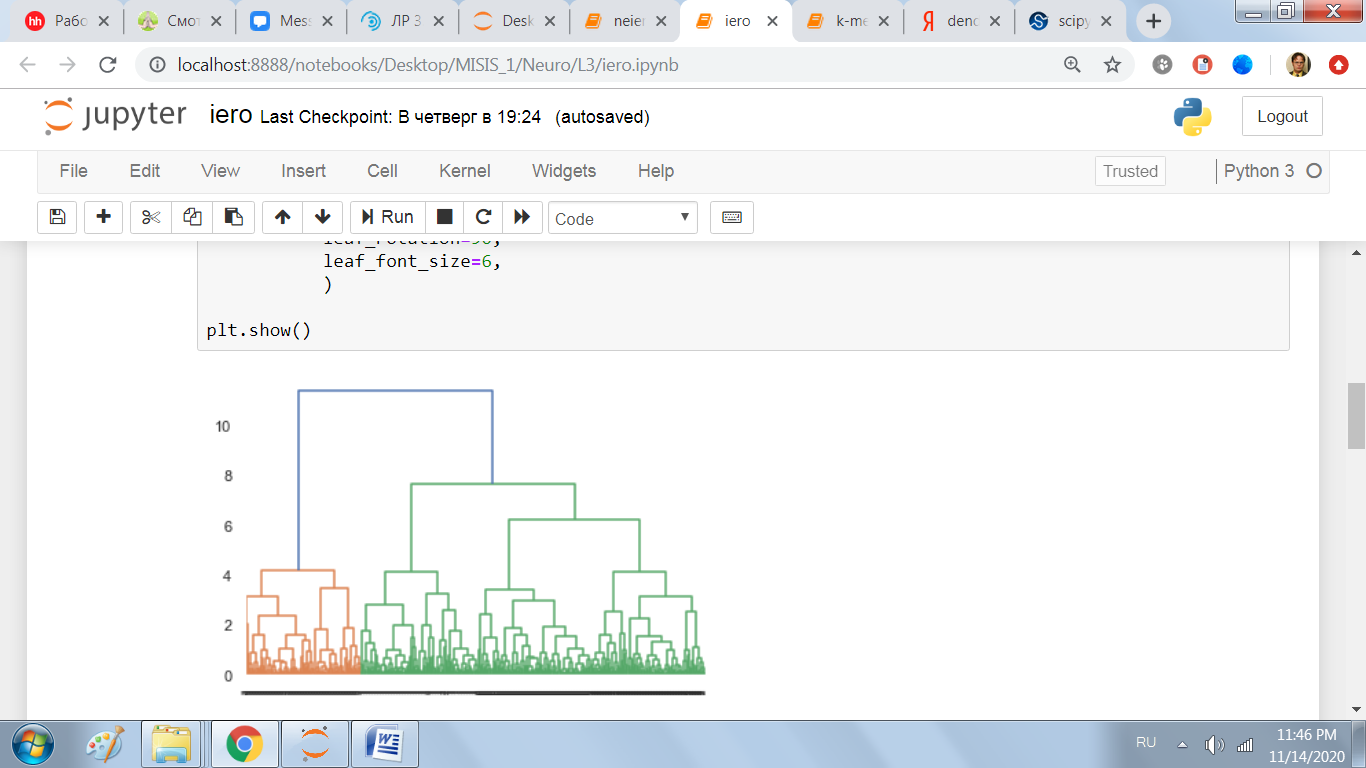


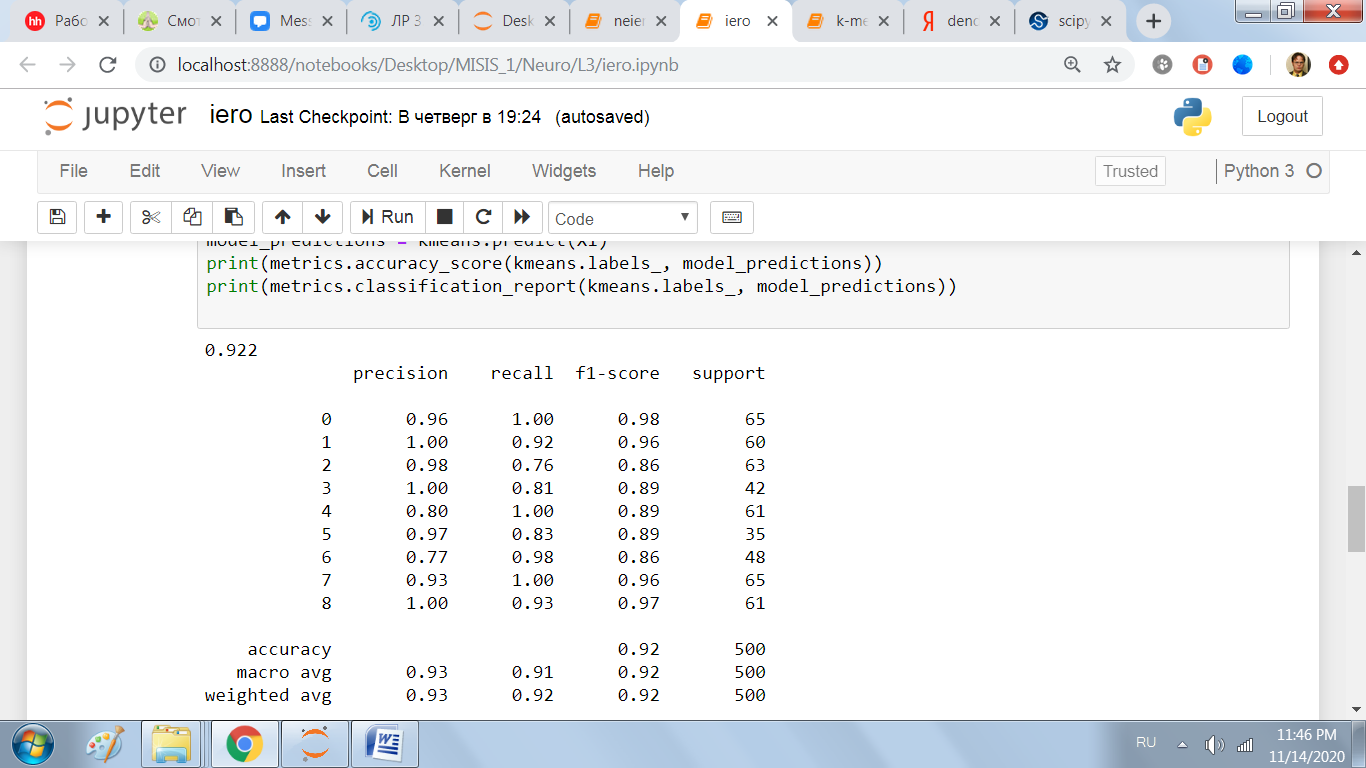
Для отображения передал свою модель и проверочную дату

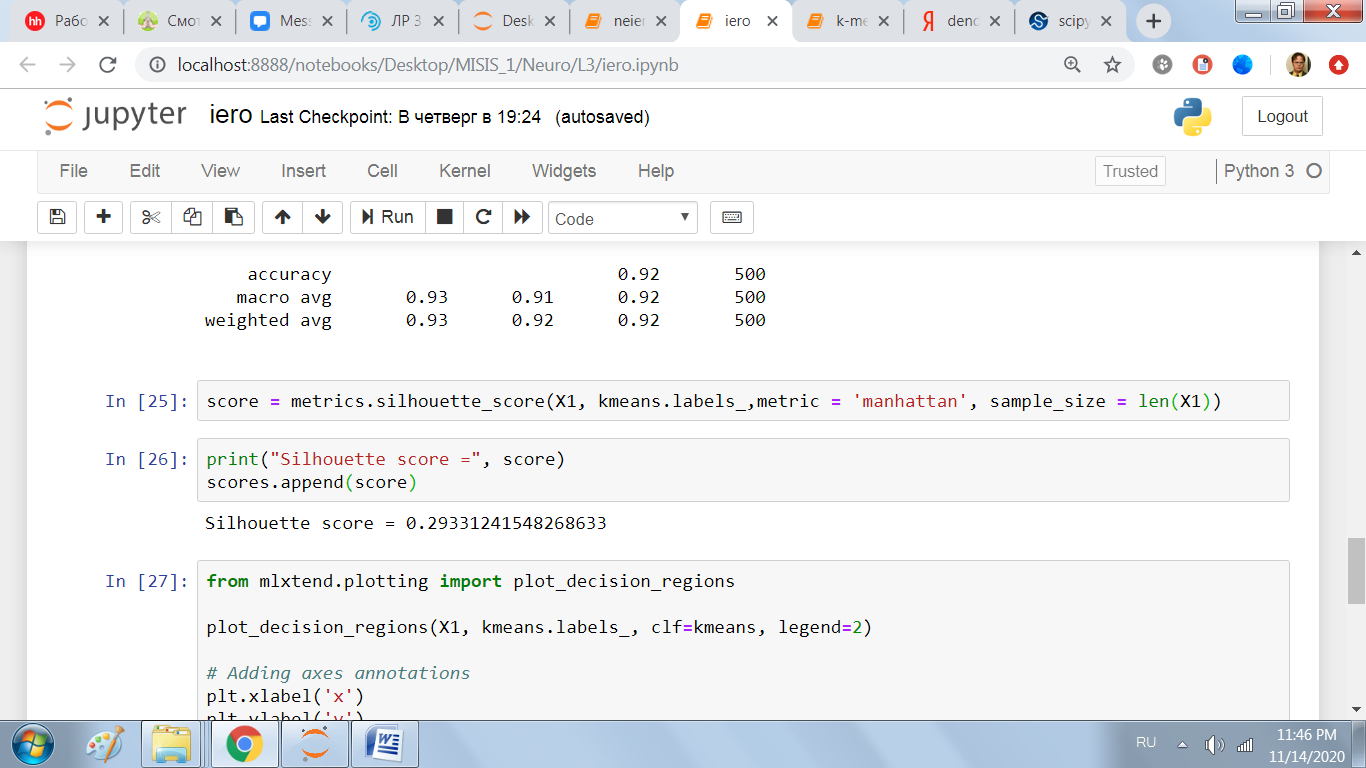


Иерархическая

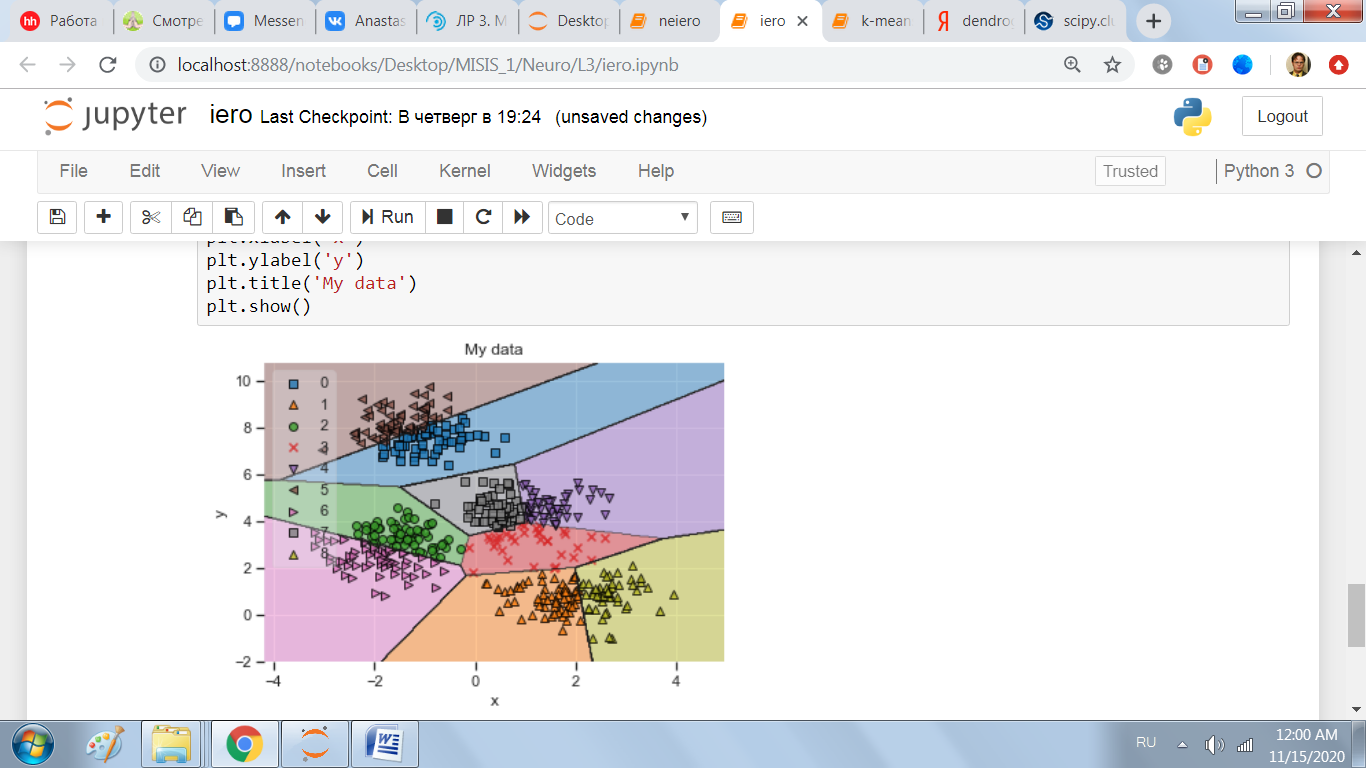
Применил обученную модель для тестового датасета, используя функцию dendrogram



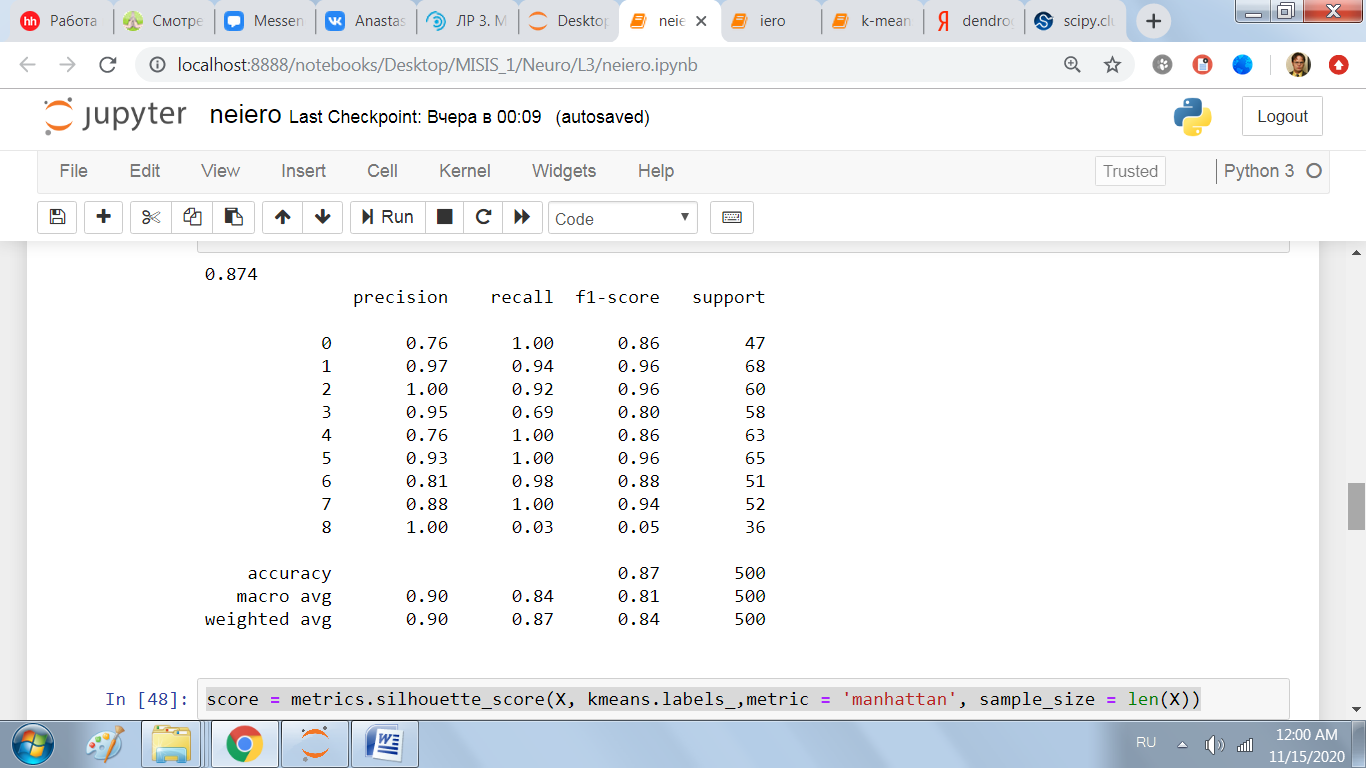


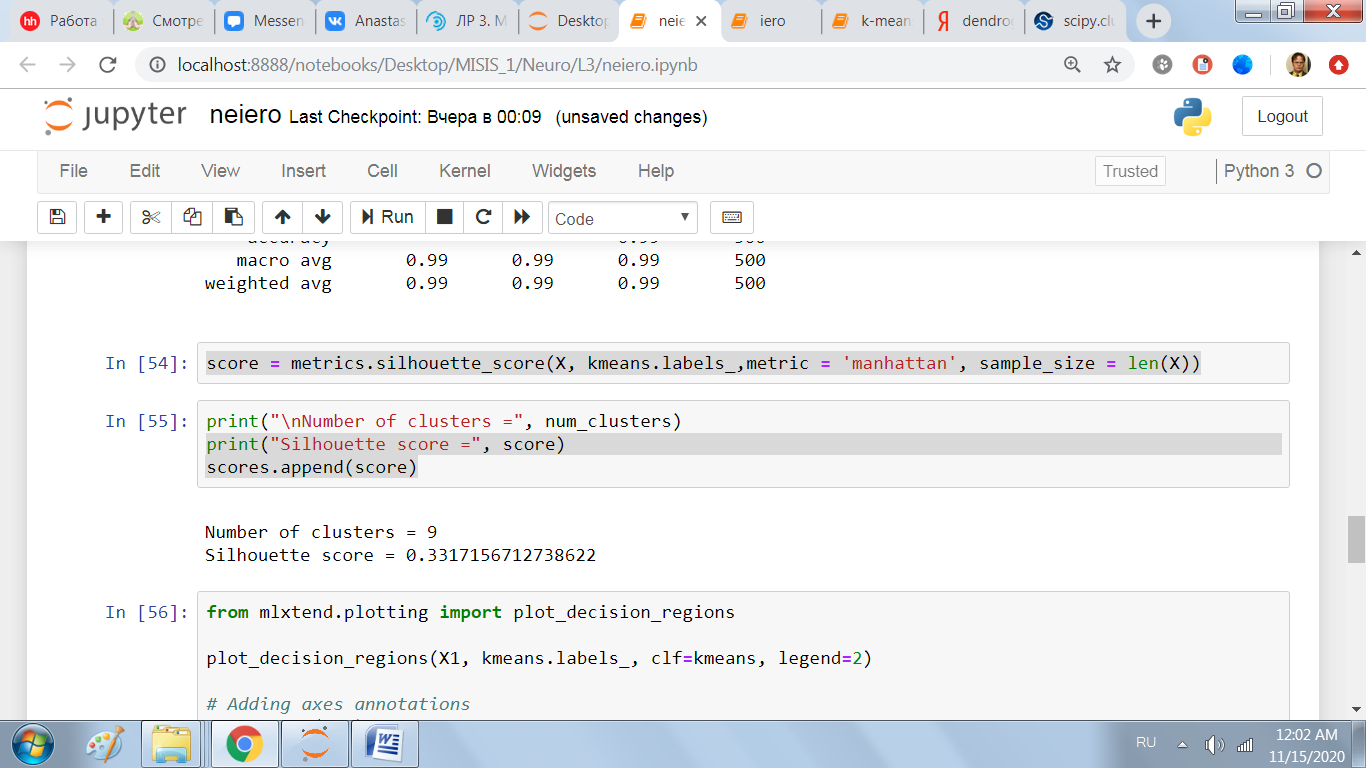


Для отображения передал свою модель и проверочную дату

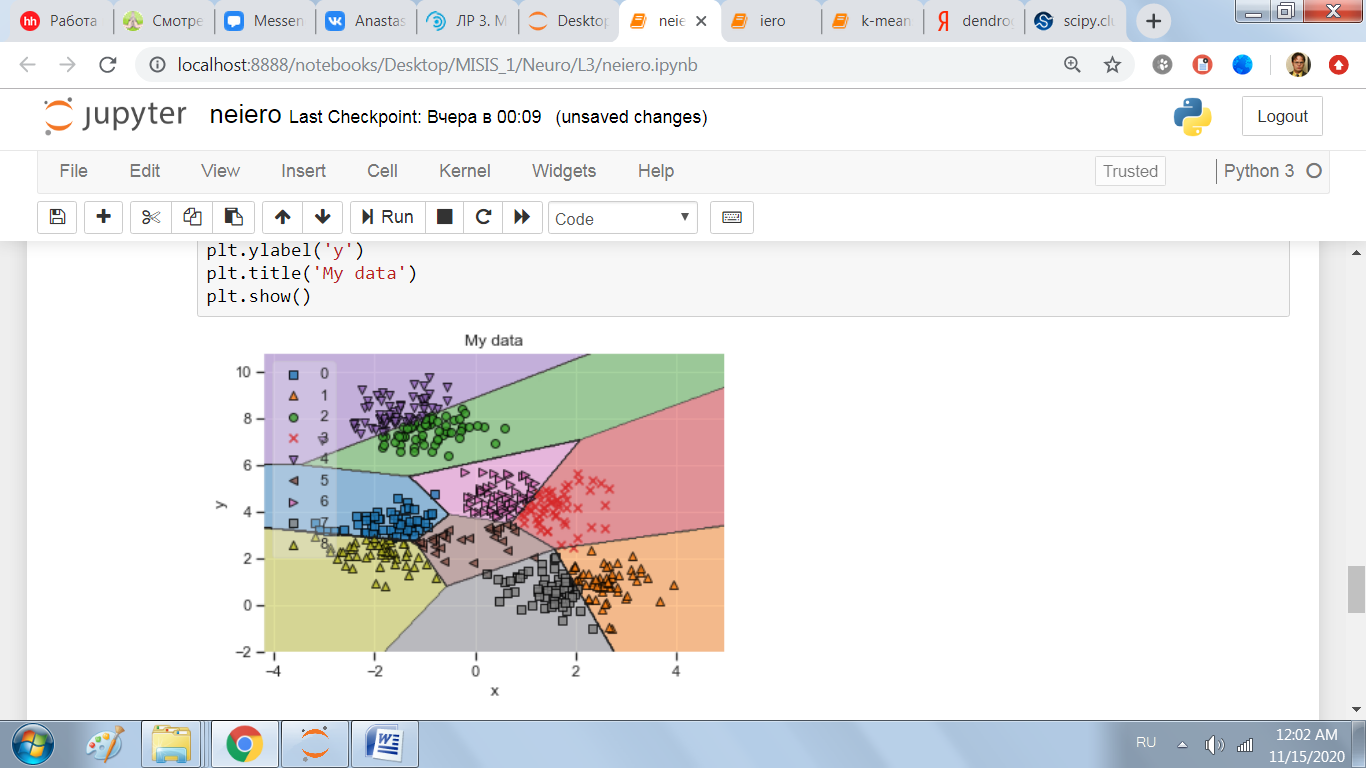


Неирархическая





Для отображения передал свою модель и проверочную дату



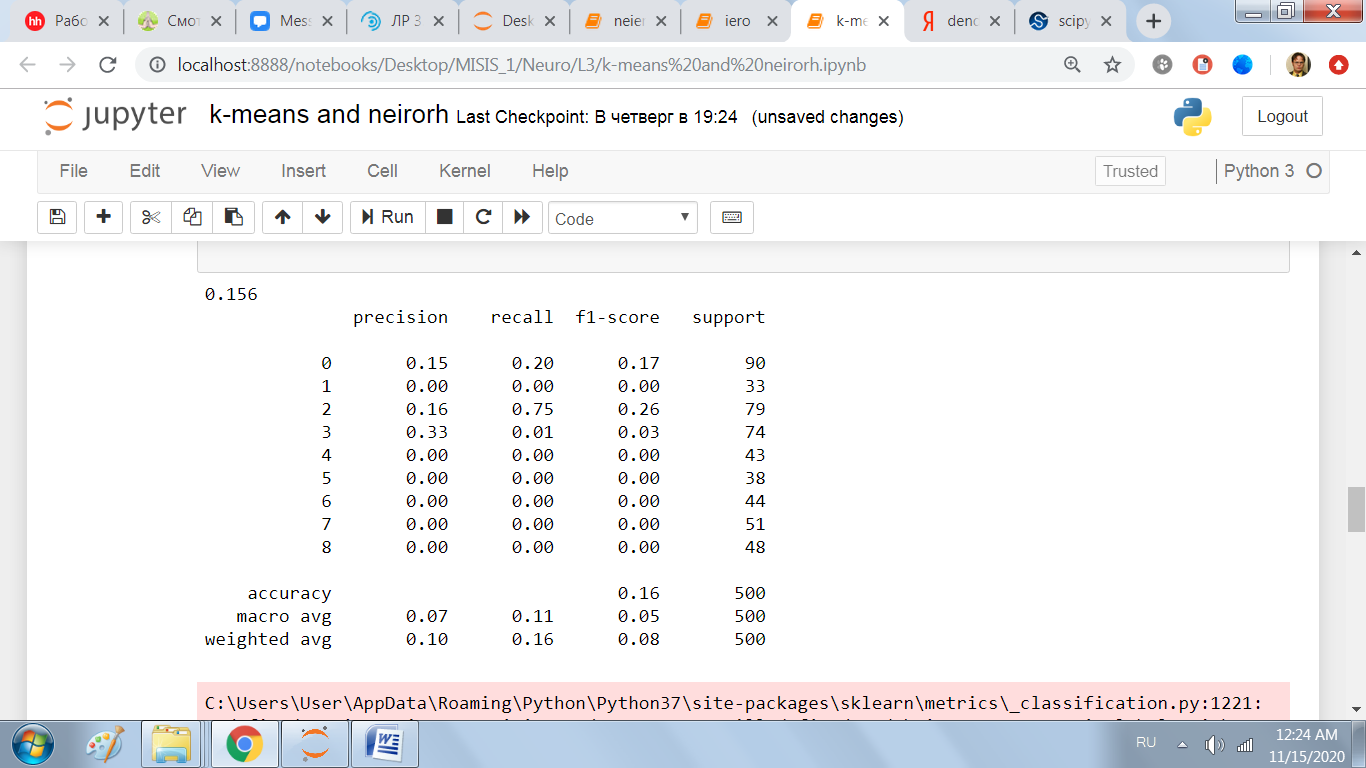
3) линейно неразделимое множество (средняя площадь пересечения классов 50-70%)

Ответ:

Регулирую даты через параметры

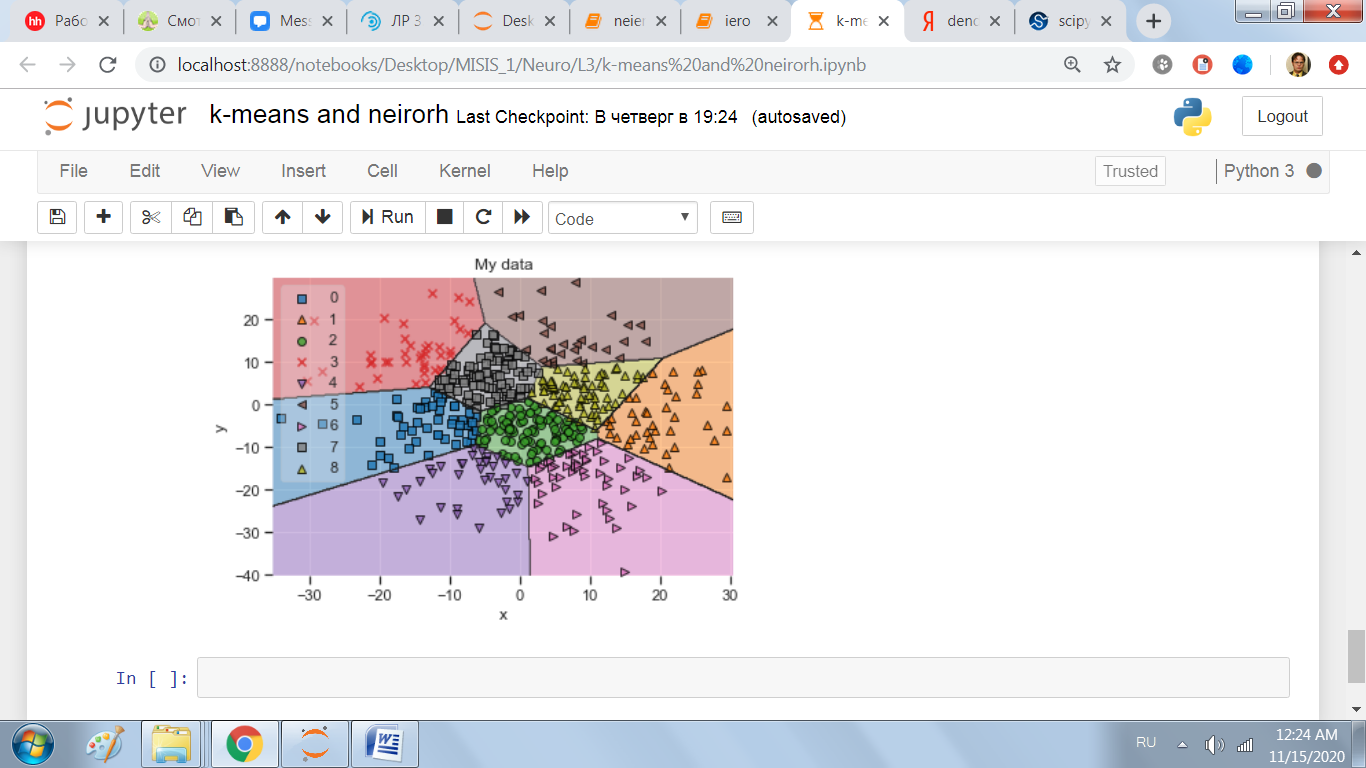
Получаю

Показатели K-means



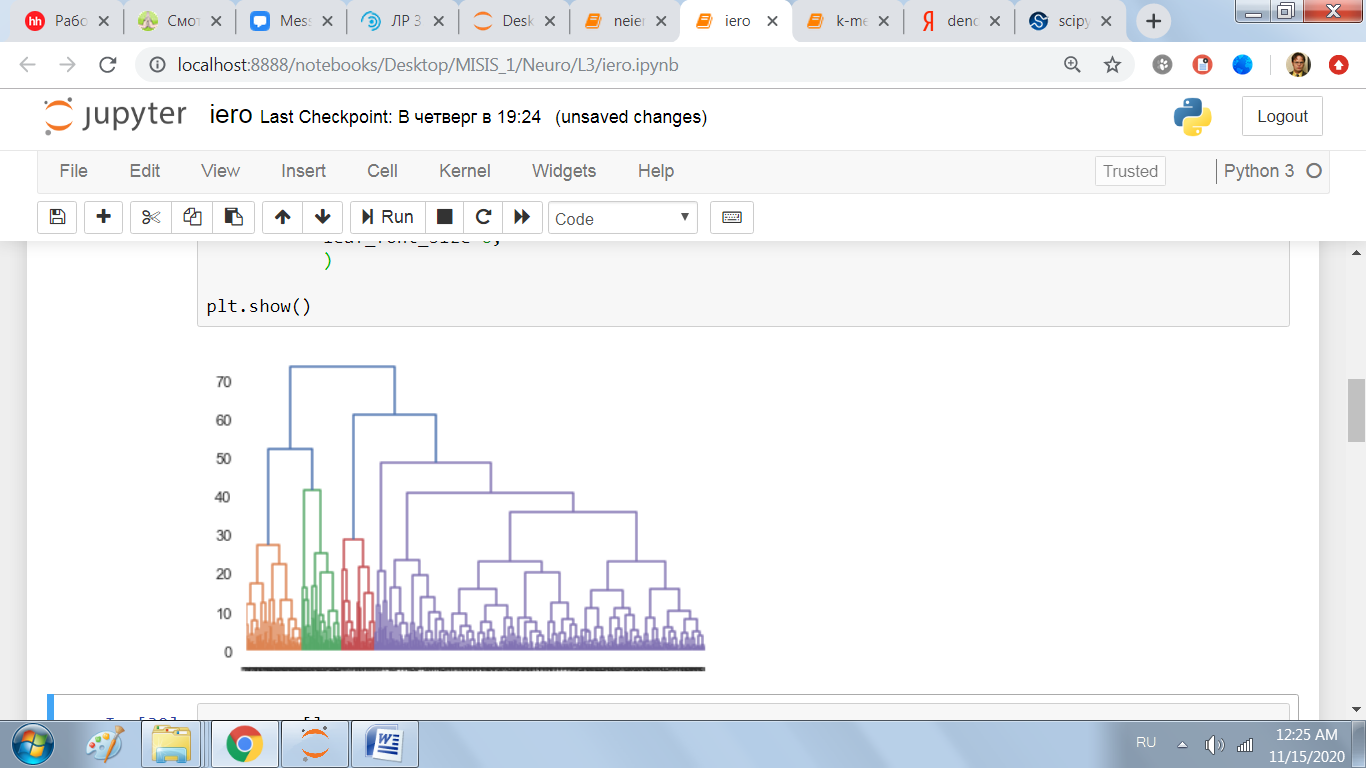


Для отображения передал свою модель и проверочную дату

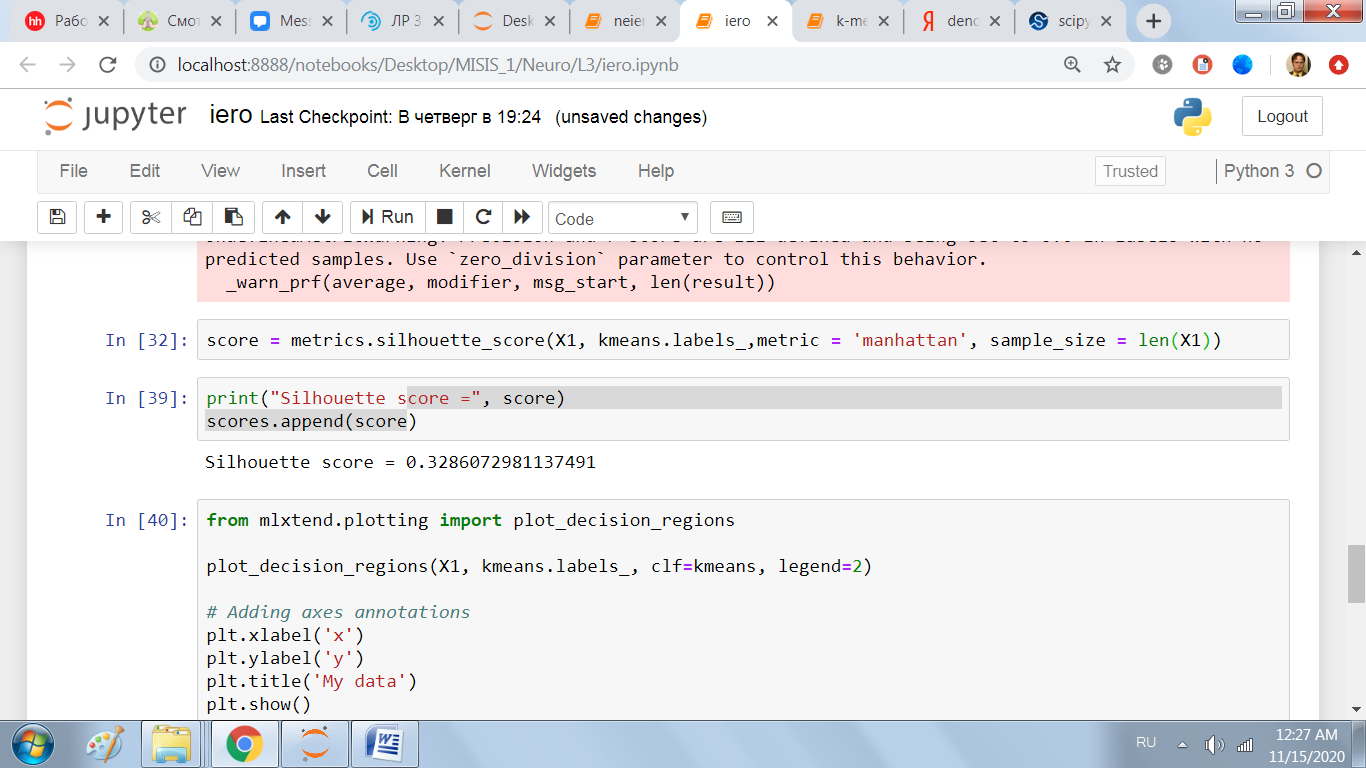


Иерархическая

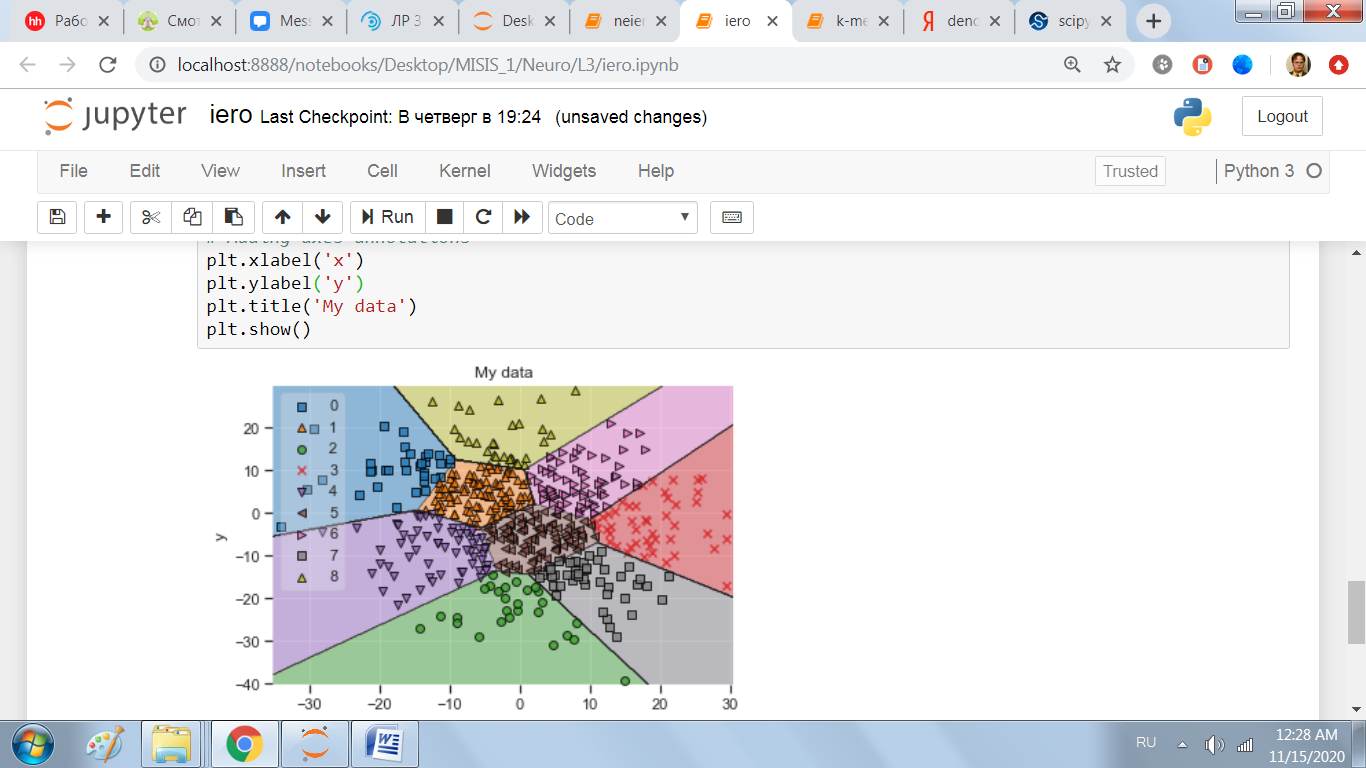
Применил обученную модель для тестового датасета, используя функцию dendrogram



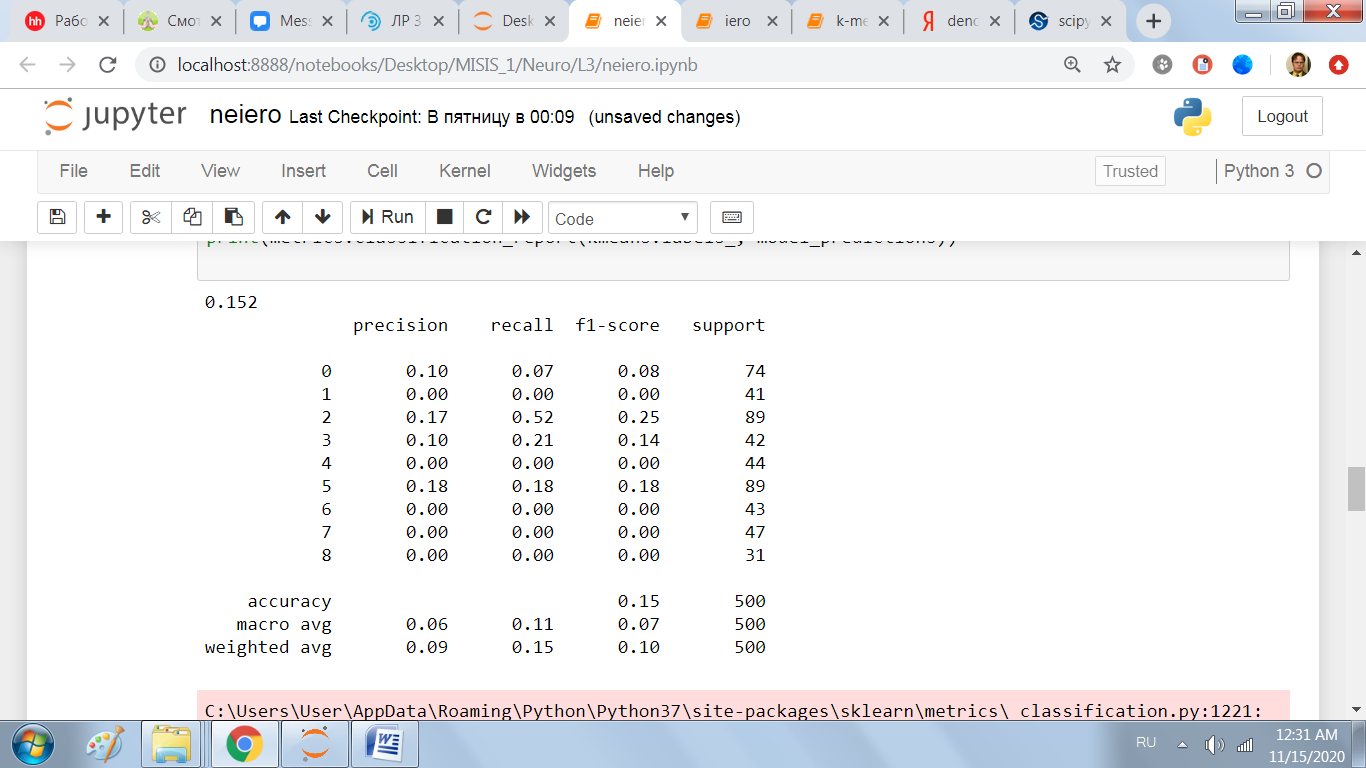


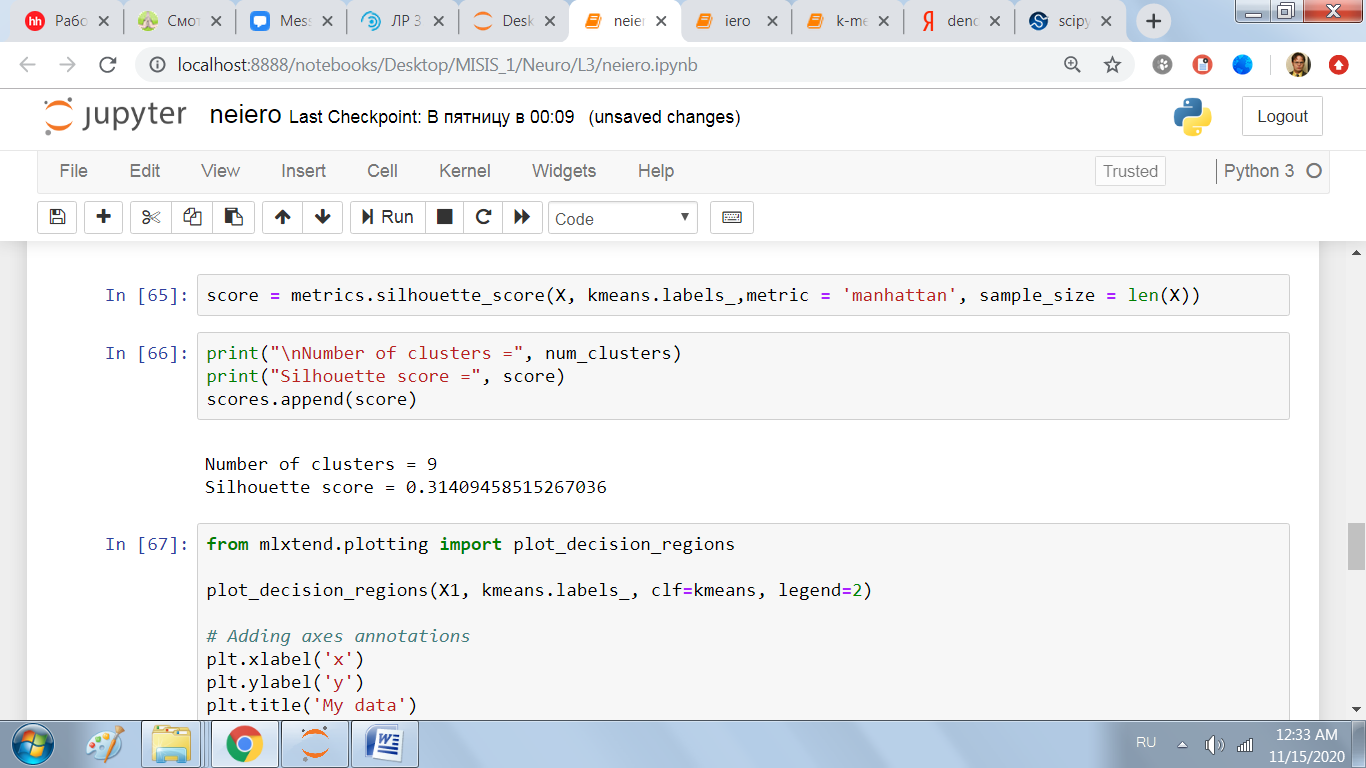


Для отображения передал свою модель и проверочную дату



Неирархическая





Для отображения передал свою модель и проверочную дату



Следовательно: чем ближе расположены середины выборок, тем хуже проходит кластеризация, что было продемонстрировано на датасетах.