МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ

федеральное государственное бюджетное образовательное учреждение

высшего образования

**«УЛЬЯНОВСКИЙ ГОСУДАРСТВЕННЫЙ ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ»**

Кафедра «Вычислительная техника»

Дисциплина «Интеллектуальные САПР»

**Лабораторная работа №5**

**«Алгоритмы кластеризации данных»**

Выполнил:

студент группы ИВТАПбд-41

Кондратьев П. С.

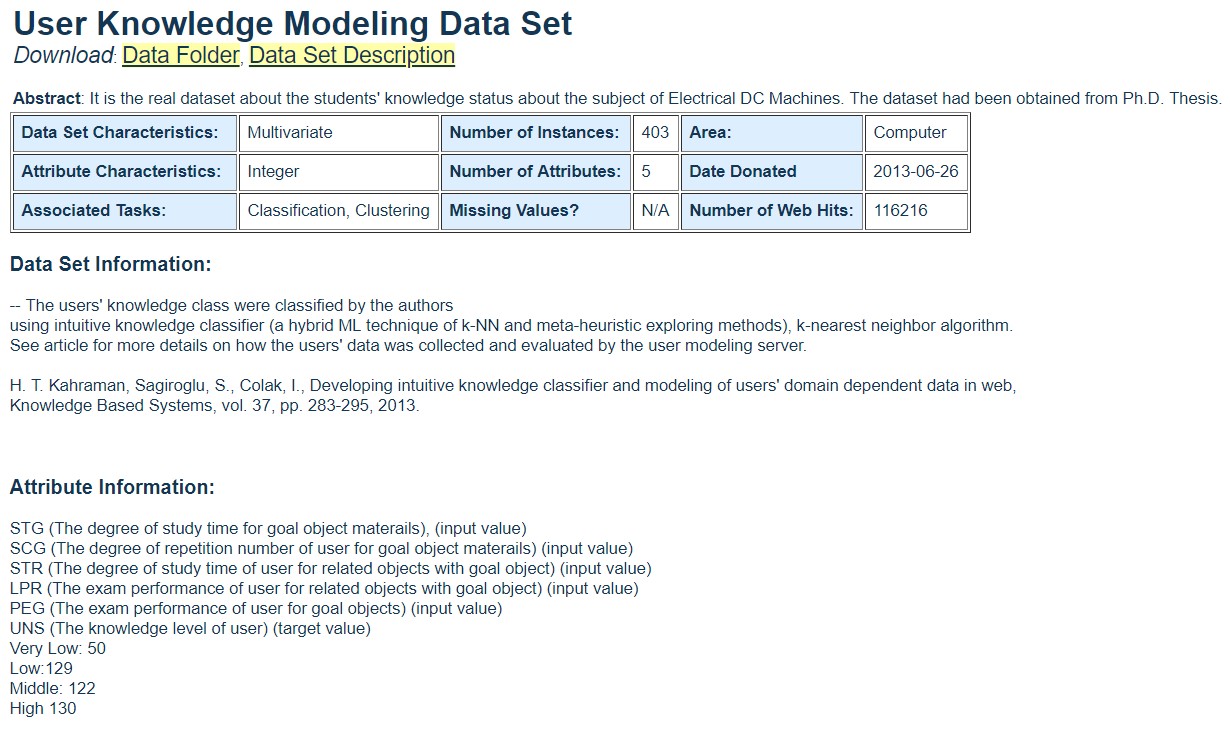
Проверил:

к.т.н. Святов К. В.

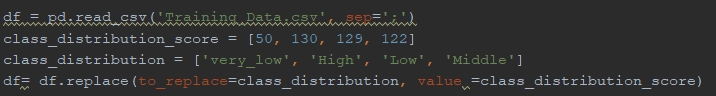
Ульяновск, 2019

Задание:

Data Set



1. Произвести масштабирование признаков (scaling).
2. С использованием библиотеки scikit-learn написать программу с использованием алгоритмов кластеризации данных, позволяющую разделить исходную выборку на классы, соответствующие предложенной вариантом задаче (<http://scikit-learn.org/stable/modules/clustering.html>).



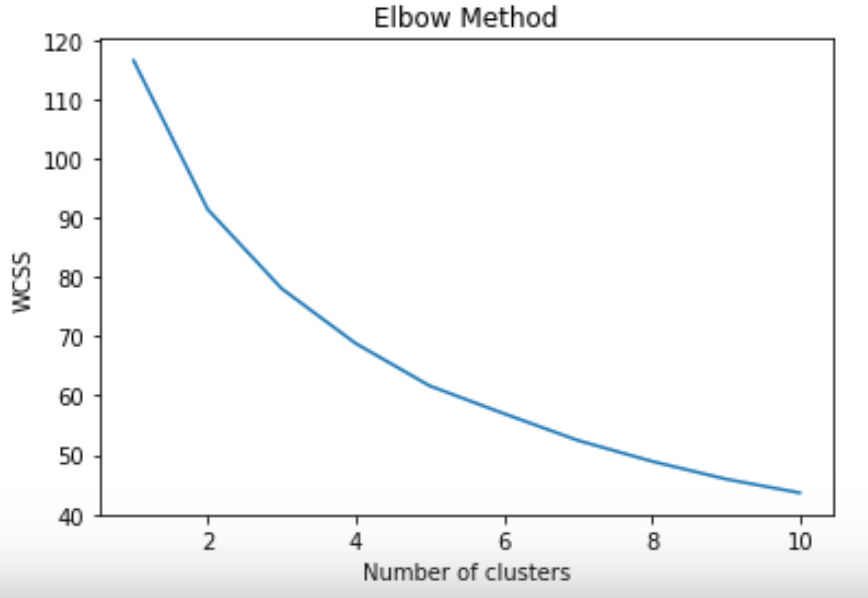
Массивы данных берутся из UCI Machine Learning Repository

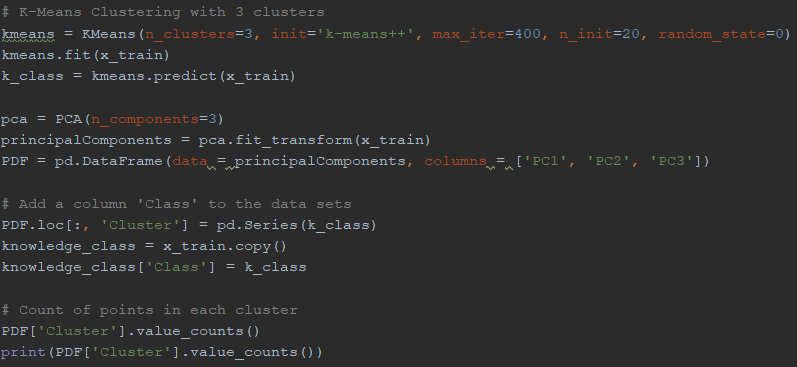
Датасет для 11 варианта - Synthetic Control Chart Time Series. В нем содердиться реальный набор данных о состоянии знаний студентов по предмету электрические машины постоянного тока.

Так как в Training\_Data содержит 259 строк и Test\_Data 146, было решено объединить файлы т.к. объём данных мал для определить наилучший алгоритм кластеризации.

1. Провести эксперименты и определить наилучший алгоритм кластеризации, параметры алгоритма. Необходимо использовать не менее 3-х алгоритмов. Данные экспериментов необходимо представить в отчете (графики, ход проведения эксперимента, выводы).

Число кластеров было выбрано с помощью метода локтя. В методе локтя оптимальное число кластеров выбирается как точка, за пределами которой скорость уменьшения суммы квадратов внутри кластеров начинает значительно падать. В некоторых случаях нам не нужно использовать метод локтя, если мы уверены в количестве требуемых кластеров. Например, в этом случае предположим, что мы хотим сформировать 3 кластера знаний учащихся, чтобы иметь возможность классифицировать их в три различные группы и потенциально использовать различные стратегии, чтобы помочь им улучшить свои знания.

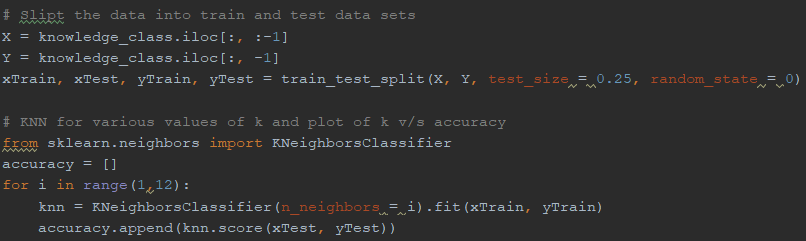


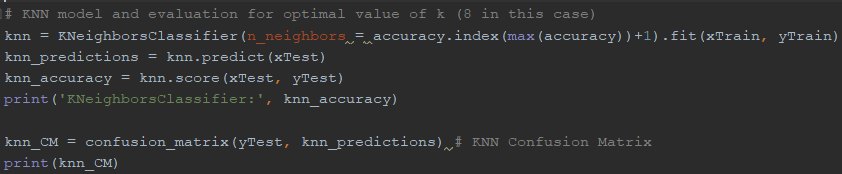


Далее мы хотим выполнить классификацию по невидимым данным и новым категориальным целевым значениям класса. Мы можем использовать методы многоклассовой классификации в машинном обучении на этих данных. Сначала мы разделим данные на обучающие и тестовые наборы. Затем мы обучим модели машинного обучения на данных trainnig и оценим их производительность на тестовых данных. Существует множество способов оценить производительность модели. Здесь мы будем использовать самую простую метрику-точность для оценки наших моделей.

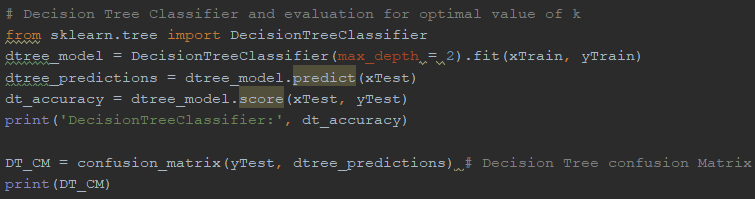
Алгоритмы, которые будут использоваться для этой задачи многоклассовой классификации, и причина, по которой они были выбраны из списка всех алгоритмов, указаны ниже:

* KNN (K-ближайшие соседи) - KNN использует расстояние в качестве метрики, и метки для набора данных также были получены с использованием расстояния в качестве метрики, когда мы применяли кластеризацию K-средних. Таким образом, KNN может хорошо работать с этим набором данных.
* Классификатор дерева решений-мы почти всегда хотим применить несколько методов машинного обучения к любому набору данных и сравнить их на основе подходящей оценочной метрики, а не выбирать одну окончательную модель, основанную только на intusion. Несмотря на то, что решение tess может не работать лучше всего на небольших данных, таких как этот, они хорошо поддаются интерпретации.
* Наивный Байес-основанный на предположении, что переменные независимы и делают вероятностную оценку с использованием гипотезы максимального правдоподобия, этот алгоритм является очень эффективным по сравнению с другими моделями машинного Лернинга.

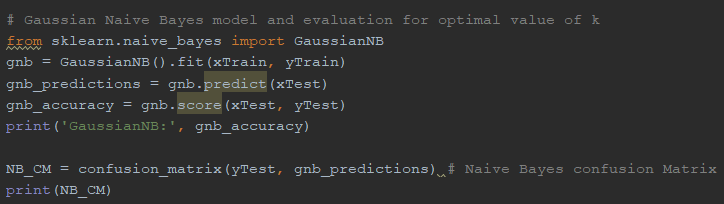




KNeighborsClassifier: 0.9306930693069307



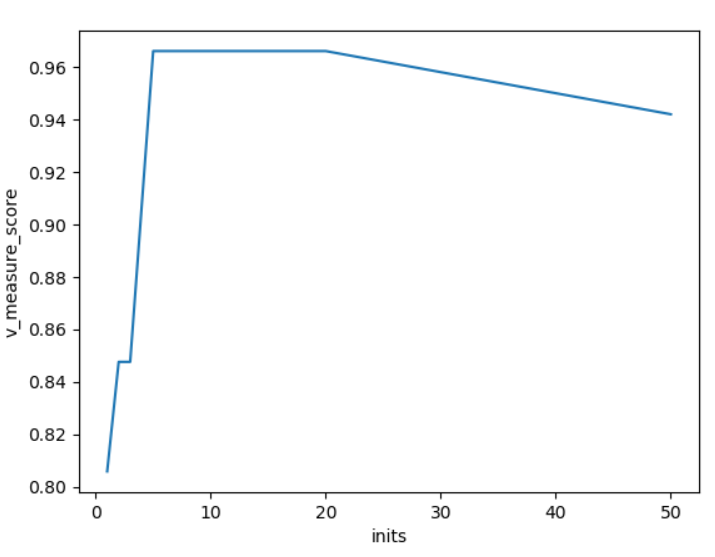
DecisionTreeClassifier: 0.9306930693069307

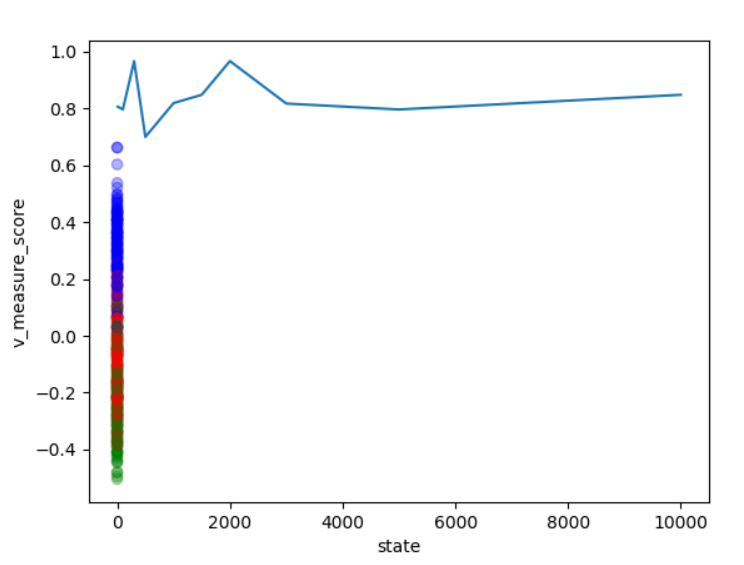


GaussianNB: 0.9504950495049505

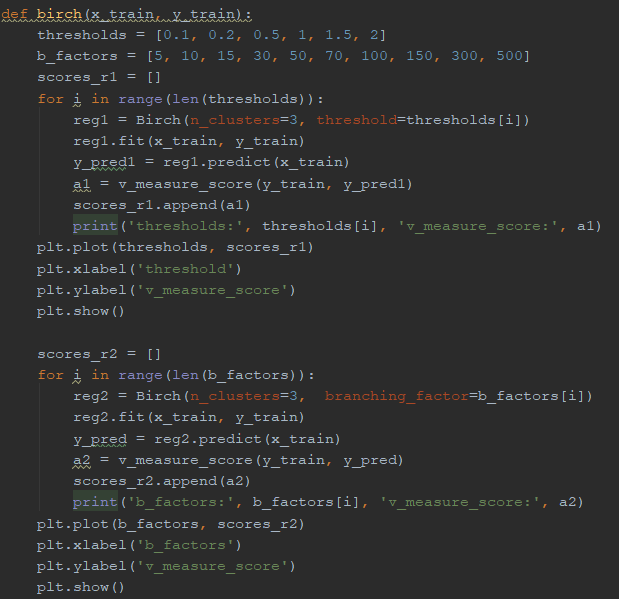
Реализован Kmeans с изменяющими параметрами random\_state и n\_init.

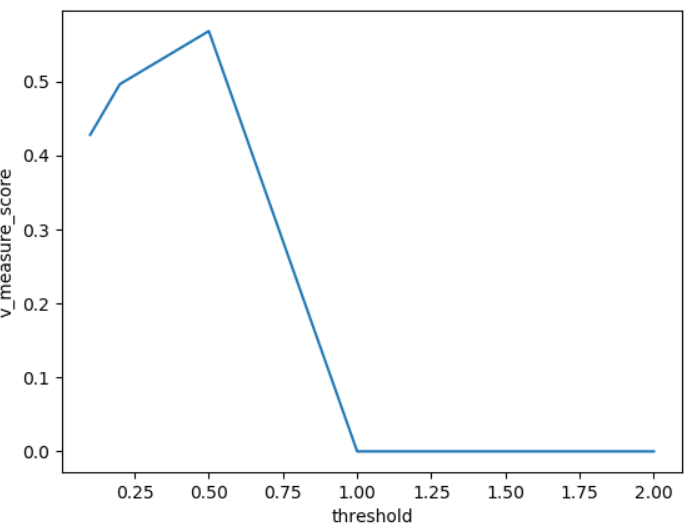


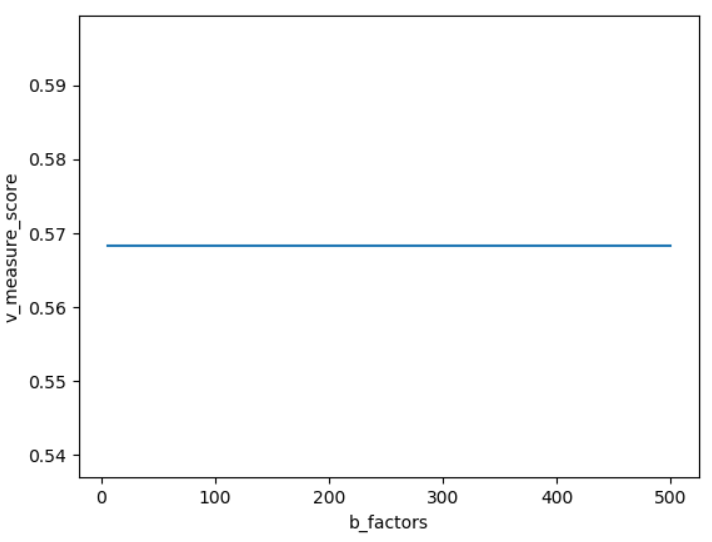




Реализован birch с изменяющими параметрами threshold и branching\_factor

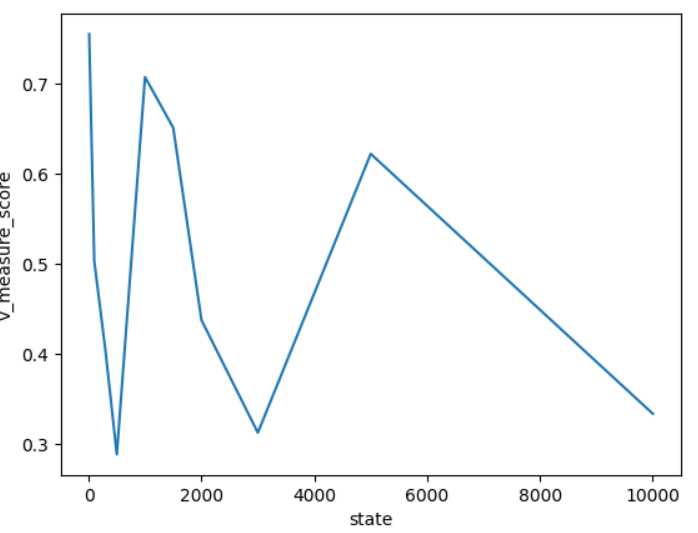


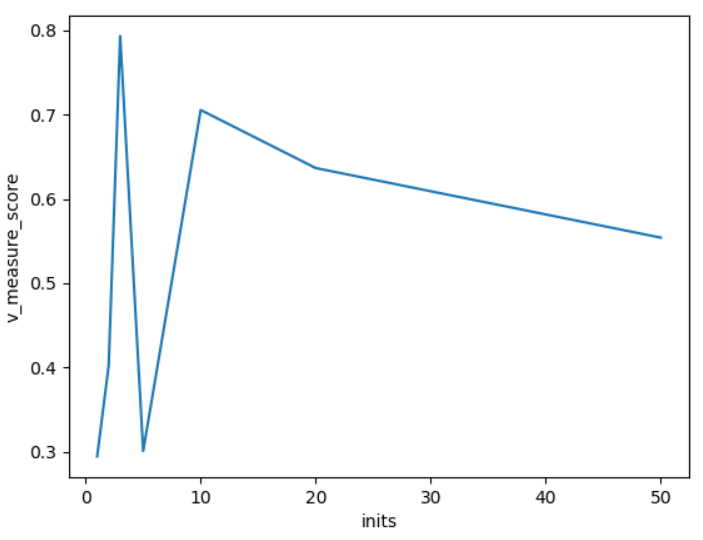




Реализован birch с изменяющими параметрами random\_state и n\_init.







Вывод: в ходе работы были изучены 3 модели алгоритмов кластеризации. Проверены экспериментально наилучшие алгоритмы кластеризации (параметры алгоритма).