|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| logoМИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ | | | | | | | | | | | | | | | | | | | | |
| Федеральное государственное автономное образовательное  учреждение высшего образования | | | | | | | | | | | | | | | | | | | | |
| **Дальневосточный федеральный университет** | | | | | | | | | | | | | | | | | | | | |
|  | | | | | | | | | | | | | | | | | | | | |
|  | | | | | | | | | | | | | | | | | | | | |
|  | | | | | | | | | | | | | | | | | | | | |
| **ШКОЛА ЕСТЕСТВЕННЫХ НАУК** | | | | | | | | | | | | | | | | | | | | |
|  | | | | | | | | | | | | | | | | | | | | |
|  | | | | | | | | | | | | | | | | | | | | |
| **Кафедра информационной безопасности** | | | | | | | | | | | | | | | | | | | | |
|  | | | | | | | | | | | | | | | | | | | | |
|  | | | | | | | | | | | | | | | | | | | | |
|  | | | | | | | | | | | | | | | | | | | | |
| **О Т Ч Е Т** | | | | | | | | | | | | | | | | | | | | |
| о прохождении учебной практики (учебно-лабораторного практикума) | | | | | | | | | | | | | | | | | | | | |
|  | | | | | | | | | | | | | | | | | | | | |
|  | | | | | | | | | | | | | | | | | | | | |
|  | | | | | | | | | | | | | | | | | | | | |
|  | | | | | | | | | |  |  | Выполнил студент  гр. С8117-10.05.01ммзи | | | | | | | | |
|  |  |  | | | | | | Вяргизова Ю.В. | | |
|  | | | | | | | | | |  |  | (подпись) | | | | | |  | | |
|  | | | | | | | | | | | | | | | | | | | | |
| Отчет защищен с оценкой | | | | | | | | | |  |  | Руководитель практики | | | | | | |  | |
|  | | | | | | | | | |  |  | Старший преподаватель кафедры информационной безопасности ШЕН | | | | | | | | |
|  | | | |  | С.С. Зотов | | | | |  |  |  | | | | |  | С.С. Зотов | | |
| (подпись) | | | |  | (И.О. Фамилия) | | | | |  |  | (подпись) | | | | |  | (И.О. Фамилия) | | |
| « | 26 | » | июня | | | | | | 2021 г. |  |  |  | | | | | | | | |
|  | | | | | | | | | | | | | | | | | | | | |
| Регистрационный № | | | | | | | |  | |  |  | Практика пройдена в срок | | | | | | | | |
| « | 26 | » | июня | | | | | | 2021 г. |  |  | с | « | 22 | » | февраля | | | | 2021 г. |
|  | | | | | | | | | |  |  | по | « | 26 | » | июня | | | | 2021 г. |
|  | | | | | |  |  | | |  |  | на предприятии | | | | | | | | |
| (подпись) | | | | | |  | (И.О. Фамилия) | | |  |  |
|  | | | | | | | | | |  |  | Кафедра информационной | | | | | | | | |
|  | | | | | | | | | |  |  | безопасности ШЕН ДВФУ | | | | | | | | |
|  | | | | | | | | | |  |  |  | | | | | | | | |
|  | | | | | | | | | | | | | | | | | | | | |
|  | | | | | | | | | | | | | | | | | | | | |
|  | | | | | | | | | | | | | | | | | | | | |
|  | | | | | | | | | | | | | | | | | | | | |
|  | | | | | | | | | | | | | | | | | | | | |
| г. Владивосток | | | | | | | | | | | | | | | | | | | | |
| 2021 | | | | | | | | | | | | | | | | | | | | |

# **Характеристика**

Выдана студенту 4 курса, специальности «Компьютерная безопасность», специализации «Математические методы защиты информации», Вяргизовой Юлии Владимировне.

Вяргизова Юлия Владимировна, в период с 22.02.2021 по 26.06.2021 года, проходил учебную практику (учебно-лабораторный практикум) на кафедре информационной безопасности ШЕН ДВФУ.

За время прохождения практики Юлия проявила усердие, тягу к знаниям, огромное желание и трудолюбие, а также неподдельный интерес к изучению материала, требуемого для написания отчета. Приходила на консультацию вовремя с перечнем вопросов, с подробным и исчерпывающим описанием о текущем состоянии практики, со списком отмеченных задач. Внимательно изучала предложенные материалы и литературу на интересующую тематику.

Вяргизова Ю.В. полностью выполнила предусмотренную программу практики, продемонстрировала умения самостоятельно решать практические вопросы, применяя теоретическую базу, полученную в учебный период, а также при самостоятельном обучении.

При выполнении поставленных задач Вяргизова Ю.В. характеризуется инициативностью, сообразительностью и ответственностью.

Старший преподаватель кафедры

информационной безопасности \_\_\_\_\_\_\_\_\_ Зотов С.С.

# **ДНЕВНИК СТУДЕНТА**

|  |  |  |  |
| --- | --- | --- | --- |
| Дата | Рабочее место | Краткое содержание выполняемых работ | Отметки руководителя |
| 22.02.21 – 27.04.21 | КИБ | Выбор темы практической работы |  |
| 28.04.21 – 30.04.21 | КИБ | Поиск материала |  |
| 01.05.21 – 05.05.21 | КИБ | Анализ найденного материала |  |
| 06.05.21 – 14.06.21 | КИБ | Реализация алгоритмов кластеризации |  |
| 15.06.21 – 20.06.21 | КИБ | Написание отчёта по проделанной работе |  |
| 21.06.21 – 26.06.21 | КИБ | Сдача готового отчета преподавателю |  |

Студент \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_Вяргизова Ю.В.

подпись Ф.И.О.

Руководитель практики от ДВФУ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_Зотов С.С.

подпись Ф.И.О.

Оглавление

[Характеристика 2](#_Toc75166477)

[ДНЕВНИК СТУДЕНТА 3](#_Toc75166478)

[Задание на практику 4](#_Toc75166479)

[Введение 5](#_Toc75166480)

[1 Кластеризация 6](#_Toc75166481)

[1.1 K-means кластеризация 9](#_Toc75166482)

[1.2 Агломеративная кластеризация 9](#_Toc75166483)

[2 Подготовка данных 10](#_Toc75166484)

[2.1 EDA 12](#_Toc75166485)

[2.2 Подготовка датасета 23](#_Toc75166486)

[3 Демонстрация работы алгоритма 24](#_Toc75166487)

[3.1 Алгоритм k-means 24](#_Toc75166488)

[3.2 Алгоритм агломеративной кластеризации 27](#_Toc75166489)

[Заключение 29](#_Toc75166490)

# **Задание на практику**

- Практическое изучение алгоритмов кластеризации данных.

- Написание отчета по практике о проделанной работе.

# 

# **Введение**

Учебная практика (учебно-лабораторный практикум) проходила на кафедре информационной безопасности ШЕН ДВФУ в период с 22 февраля 2021 года по 26 июня 2021 года.

Целью прохождения практики является приобретение практических и теоретических навыков по специальности, а также навыков оформления проведенного исследования в отчетной форме.

Задачи практики:

1. Изучить существующие алгоритмы кластеризации данных.
2. Выбрать датасет и выполнить на нем несколько алгоритмов кластеризации.
3. Сравнить выбранные алгоритмы между собой.
4. На основе полученных знаний написать отчет по практике о проделанной работе.

# **1 Кластеризация**

Кластерный анализ или кластеризация – это задача группировки набора объектов таким образом, чтобы объекты в одной группе (называемой кластером) были более похожи (в некотором смысле) друг на друга, чем на объекты в других группах (кластерах). Это основная задача исследовательского анализа данных и общий метод статистического анализа данных, используемый во многих областях, включая распознавание образов, анализ изображений, поиск информации, биоинформатику, сжатие данных, компьютерную графику и машинное обучение.

Применение кластерного анализа в общем виде сводится к следующим этапам:

1) Отбор выборки объектов для кластеризации.

2) Определение множества переменных, по которым будут оцениваться объекты в выборке. При необходимости – нормализация значений переменных.

3) Вычисление значений меры сходства между объектами.

4) Применение метода кластерного анализа для создания групп сходных объектов (кластеров).

5) Представление результатов анализа.

После получения и анализа результатов возможна корректировка выбранной метрики и метода кластеризации до получения оптимального результата.

Применение кластеризации несет в себе несколько целей:

- Понимание данных путём выявления кластерной структуры. Разбиение выборки на группы схожих объектов позволяет упростить дальнейшую обработку данных и принятия решений, применяя к каждому кластеру свой метод анализа.

- Сжатие данных. Если исходная выборка избыточно большая, то можно сократить её, оставив по одному наиболее типичному представителю от каждого кластера.

- Обнаружение новизны. Выделяются нетипичные объекты, которые не удаётся присоединить ни к одному из кластеров.

Существует также метод группировки набора объектов, называемый классификация. Классификация — один из разделов машинного обучения, посвященный решению следующей задачи. Имеется множество объектов (ситуаций), разделённых, некоторым образом, на классы. Задано конечное множество объектов, для которых известно, к каким классам они относятся. Это множество называется обучающей выборкой. Классовая принадлежность остальных объектов не известна. Требуется построить алгоритм, способный классифицировать произвольный объект из исходного множества.

Кластеризация отличается от классификации тем, что изначально не задано множество объектов, для которых известно, к каким классам они относятся, и даже могут быть неизвестны сами классы.

Решение задачи кластеризации принципиально неоднозначно, и тому есть несколько причин:

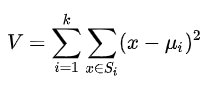
- Не существует однозначно наилучшего критерия качества кластеризации. Известен целый ряд эвристических критериев, а также ряд алгоритмов, не имеющих чётко выраженного критерия, но осуществляющих достаточно разумную кластеризацию «по построению». Все они могут давать разные результаты.

- Число кластеров, как правило, неизвестно заранее и устанавливается в соответствии с некоторым субъективным критерием.

- Результат кластеризации существенно зависит от метрики, выбор которой, как правило, также субъективен и определяется экспертом.

## **1.1 K-means кластеризация**

Метод k-means — наиболее популярный метод кластеризации. Действие алгоритма таково, что он стремится минимизировать суммарное квадратичное отклонение точек кластеров от центров этих кластеров:



 Данный алгоритм разбивает [множество](https://ru.wikipedia.org/wiki/%D0%9C%D0%BD%D0%BE%D0%B6%D0%B5%D1%81%D1%82%D0%B2%D0%BE) элементов [векторного пространства](https://ru.wikipedia.org/wiki/%D0%9B%D0%B8%D0%BD%D0%B5%D0%B9%D0%BD%D0%BE%D0%B5_%D0%BF%D1%80%D0%BE%D1%81%D1%82%D1%80%D0%B0%D0%BD%D1%81%D1%82%D0%B2%D0%BE) на заранее известное число кластеров *k*.

Основная идея заключается в том, что на каждой итерации перевычисляется [центр масс](https://ru.wikipedia.org/wiki/%D0%A6%D0%B5%D0%BD%D1%82%D1%80_%D0%BC%D0%B0%D1%81%D1%81) для каждого кластера, полученного на предыдущем шаге, затем векторы разбиваются на кластеры вновь в соответствии с тем, какой из новых центров оказался ближе по выбранной метрике.

Алгоритм завершается, когда на какой-то итерации не происходит изменения внутрикластерного расстояния. Это происходит за конечное число итераций, так как количество возможных разбиений конечного множества конечно, а на каждом шаге суммарное квадратичное отклонение *V* уменьшается, поэтому зацикливание невозможно.

Проблемы k-means:

* Не гарантируется достижение глобального минимума суммарного квадратичного отклонения *V*, а только одного из локальных минимумов.
* Результат зависит от выбора исходных центров кластеров, их оптимальный выбор неизвестен.
* Число кластеров надо знать заранее.

## **1.2 Агломеративная кластеризация**

Иерархическая кластеризация (также графовые алгоритмы кластеризации и иерархический кластерный анализ) — совокупность алгоритмов упорядочивания данных, направленных на создание иерархии (дерева) вложенных кластеров. Выделяют два класса методов иерархической кластеризации:

- Агломеративные методы: новые кластеры создаются путем объединения более мелких кластеров и, таким образом, дерево создается от листьев к стволу;

- Дивизивные или дивизионные методы: новые кластеры создаются путем деления более крупных кластеров на более мелкие и, таким образом, дерево создается от ствола к листьям. На практике такой подход нигде не применяется из-за больших вычислительных трудностей так как необходимо рассчитать большое количество всевозможных комбинаций деления.

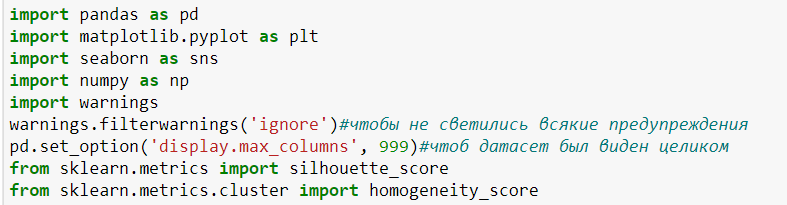
Алгоритмы иерархической кластеризации предполагают, что анализируемое множество объектов характеризуется определённой степенью связности. Как и большинство визуальных способов представления зависимостей графы быстро теряют наглядность при увеличении числа кластеров.

Под дендрограммой обычно понимается дерево, построенное по матрице мер близости. Дендрограмма позволяет изобразить взаимные связи между объектами из заданного множества. Для создания дендрограммы требуется матрица сходства (или различия), которая определяет уровень сходства между парами кластеров.

# **2 Подготовка данных**

Используется датасет [UNSW\_NB15 | Kaggle](https://www.kaggle.com/mrwellsdavid/unsw-nb15?select=UNSW_NB15_testing-set.csv) для классификации по обнаружению сетевых атак. Датасет содержит девять типов атак, а именно: Fuzzers, Analysis, Backdoors, DoS, Exploits, Generic, Reconnaissance, Shellcode и Worms, и 48 признаков.

Импортируем все нужные библиотеки и зададим параметры для лучшего отображения графиков:



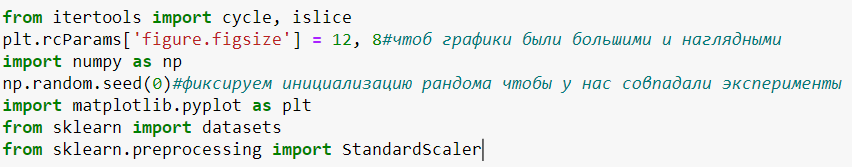


Рис. 1.Импорт всех нужных библиотек.

Читаем датасет и отображаем первые 5 строк:

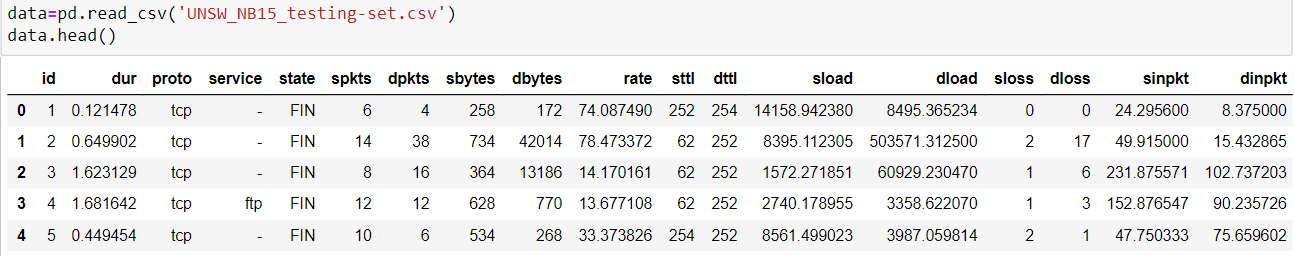
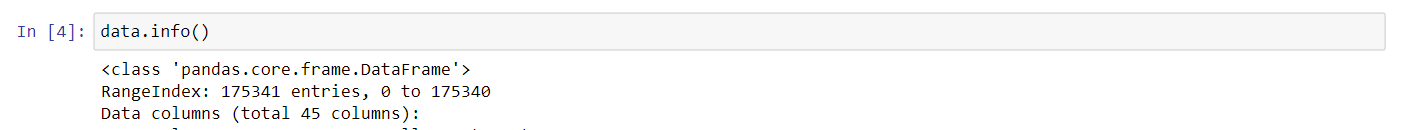
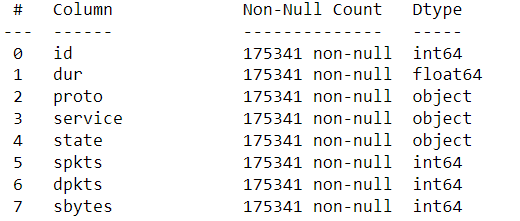
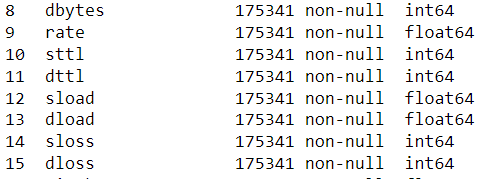


Рис. 2.Чтение датасета.

Получаем информацию о датафрейме: 





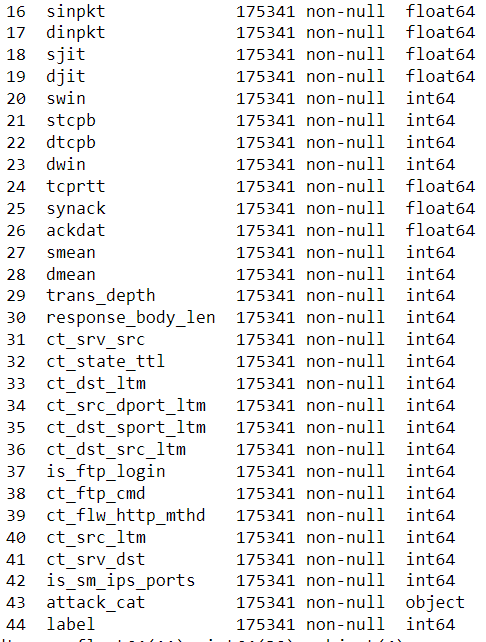




Рис. 3. Получение информации о датафрейме.

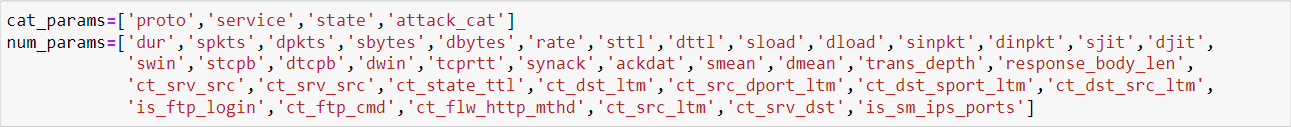
Категорийные и числовые параметры:

Рис. 4.Категорийные и числовые параметры.

## **2.1 EDA**

Вспомогательная функция, которая вычисляет все описательные все описательные статистические характеристики параметра такие как минимум, максимум, значения и строит график распределения:

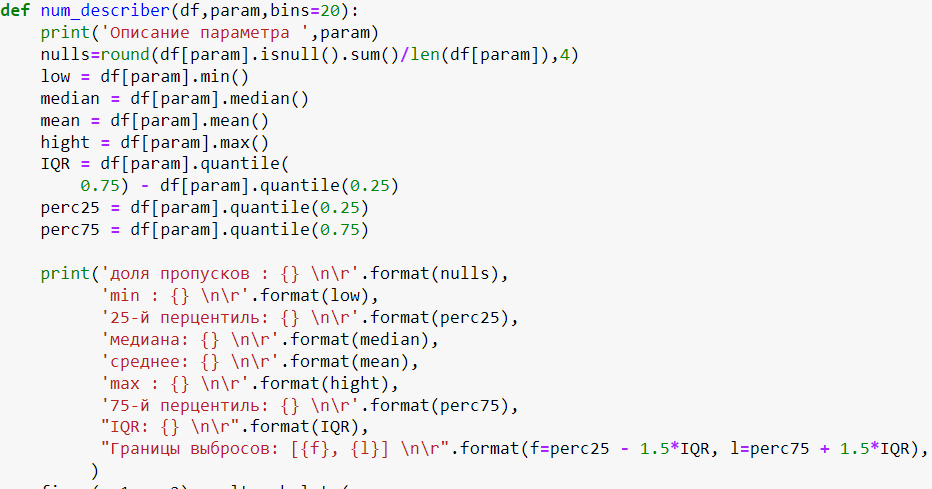
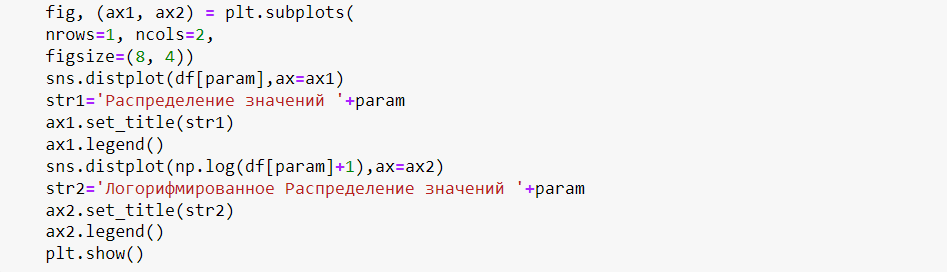


Рис. 4.Вспомогательная функция.

Сделаем разведку числовых параметров:

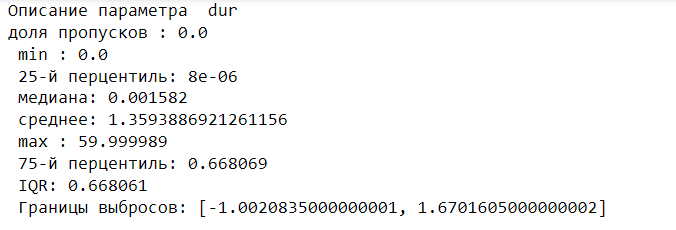
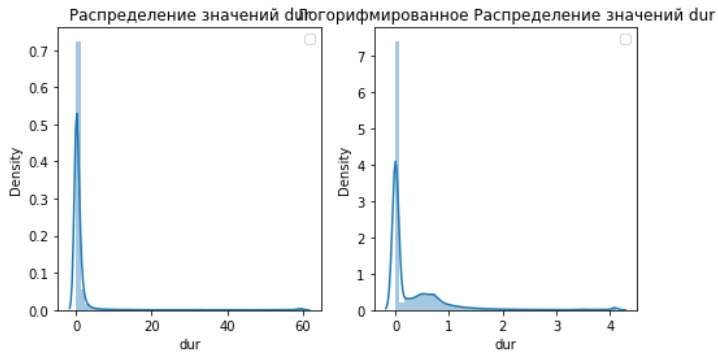
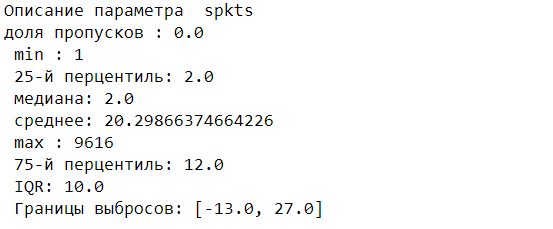
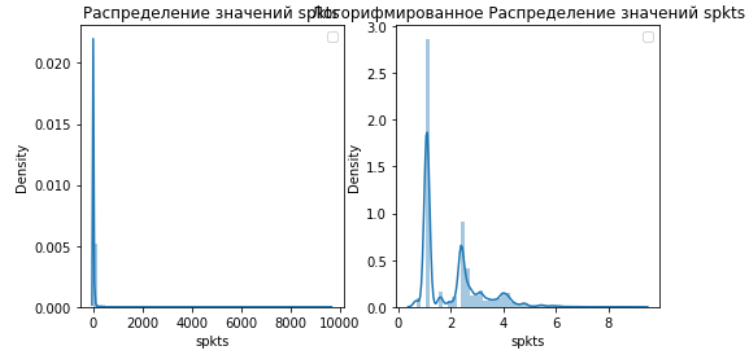
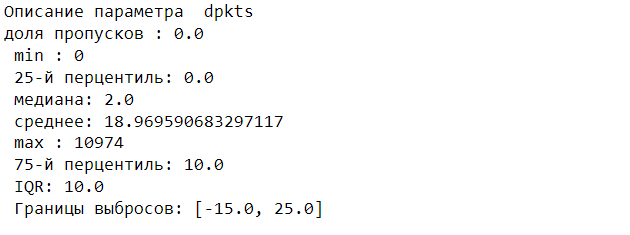
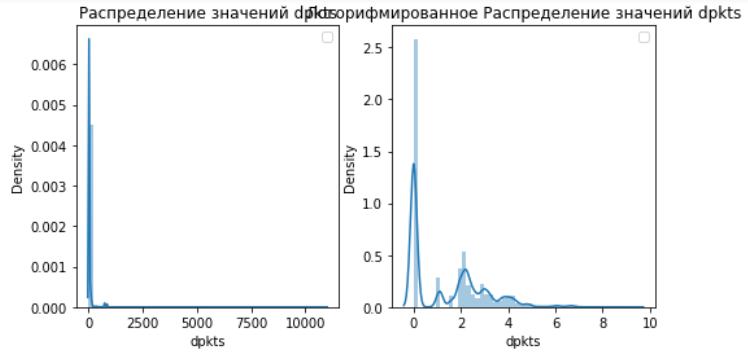


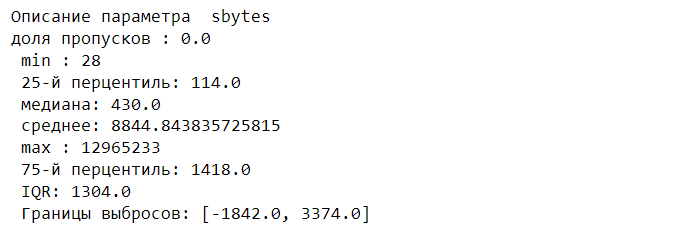
Рис. 5.Разведка числовых параметров.

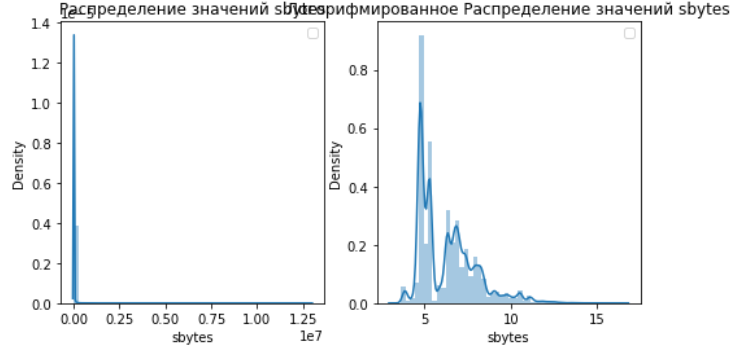


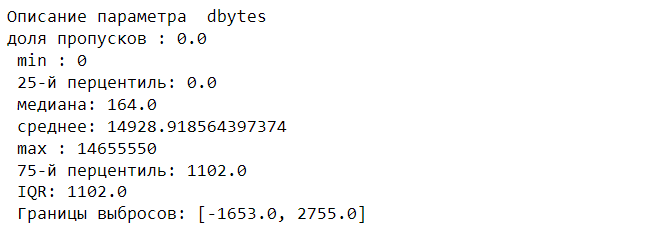


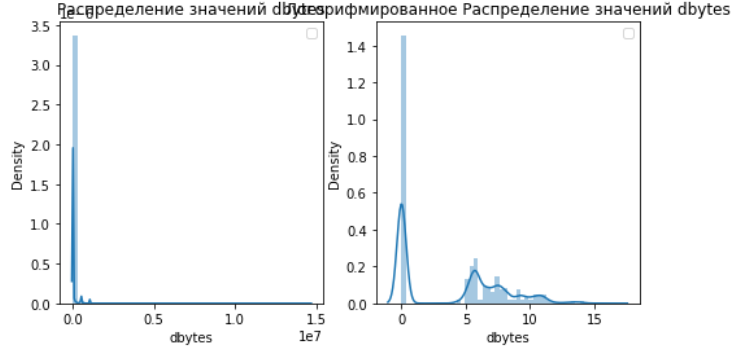


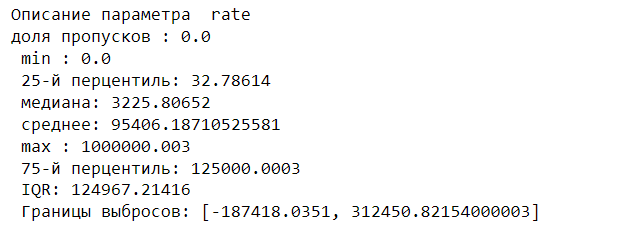


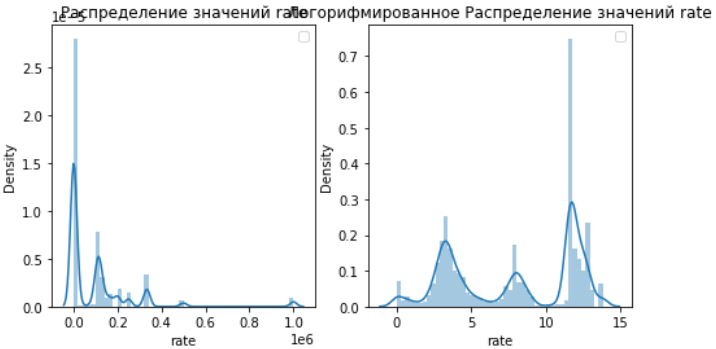


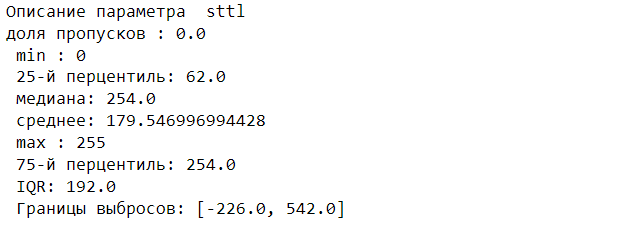


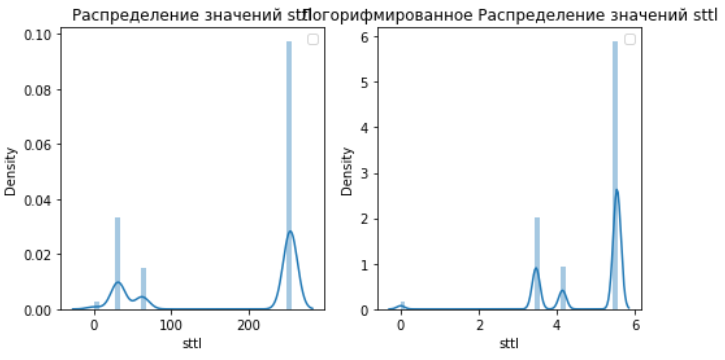


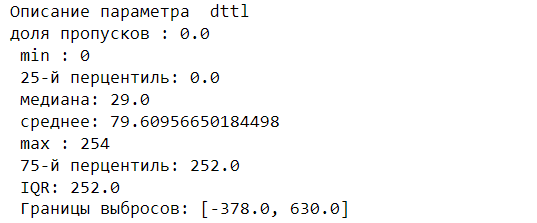


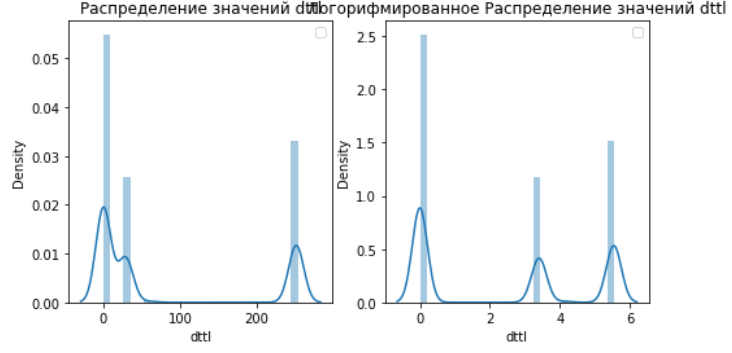


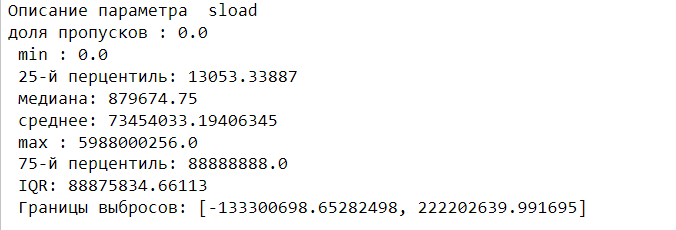


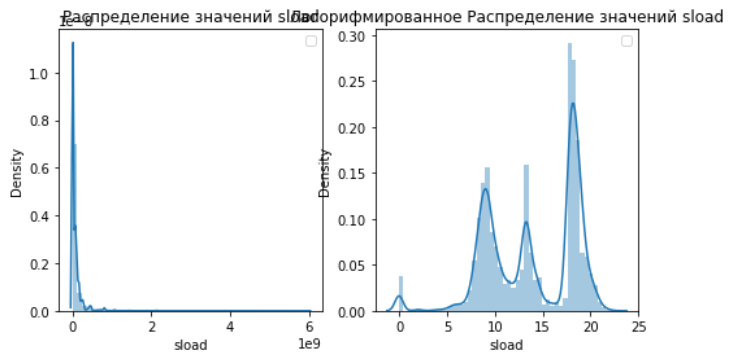


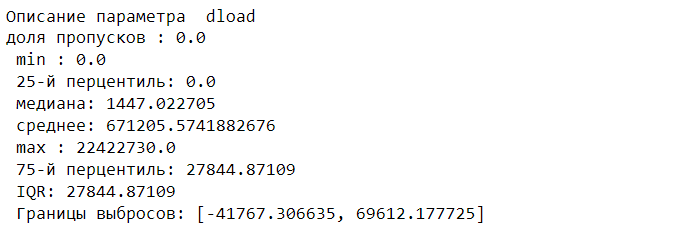






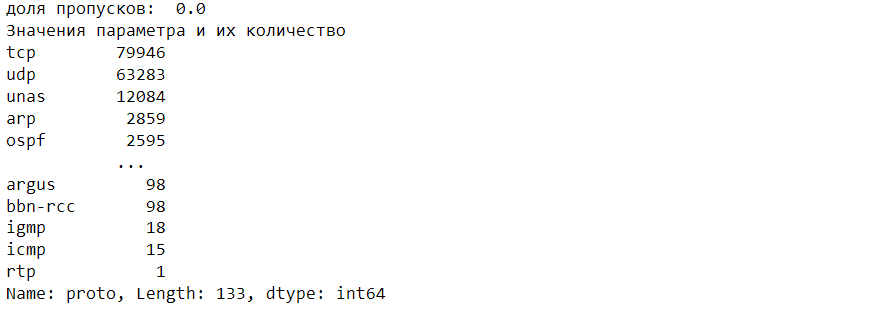


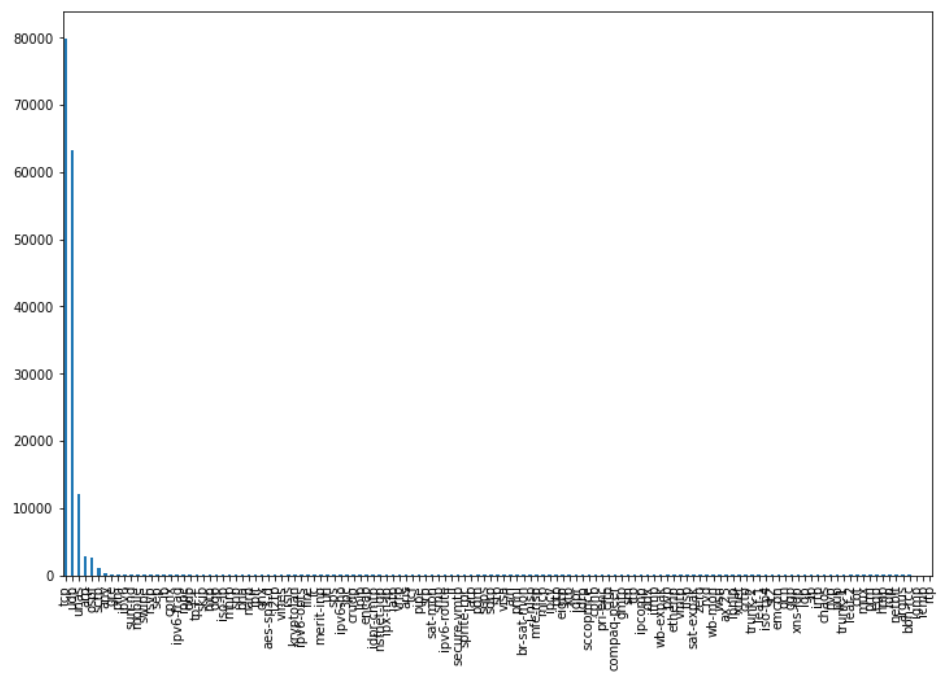




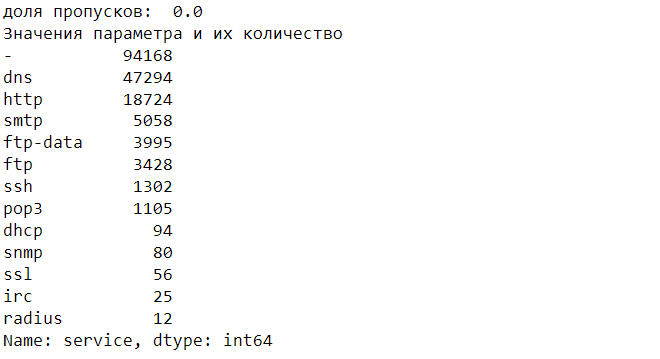
Сделаем осмотр категорийных параметров:

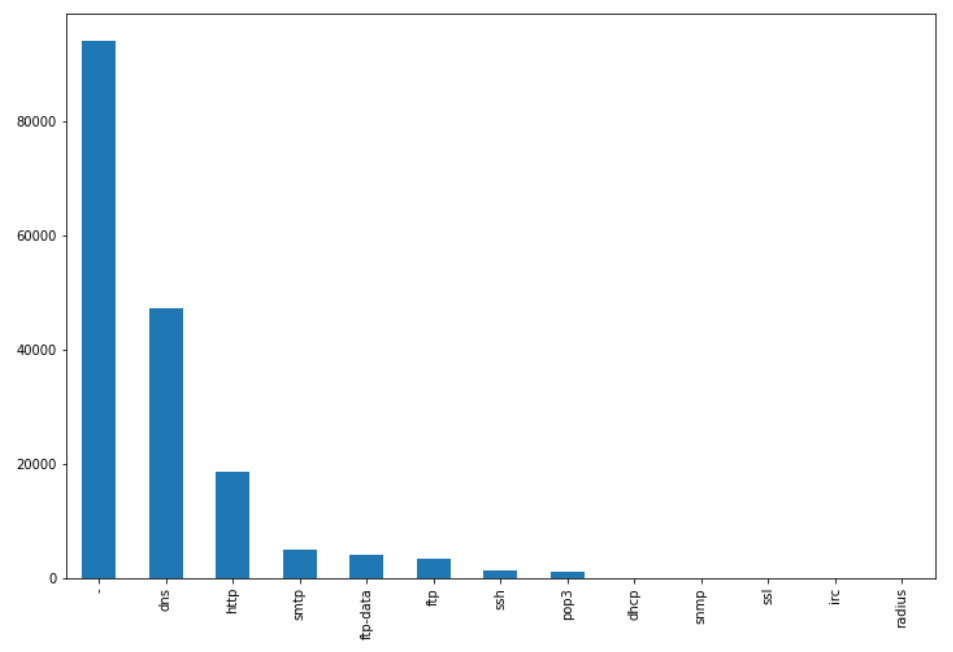




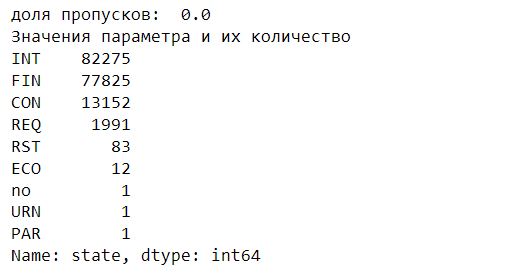


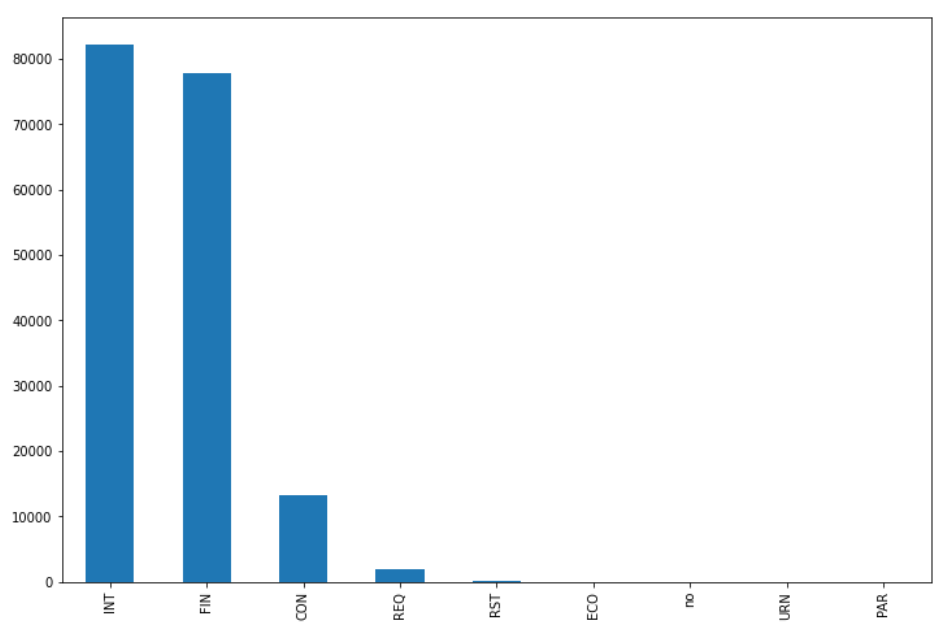






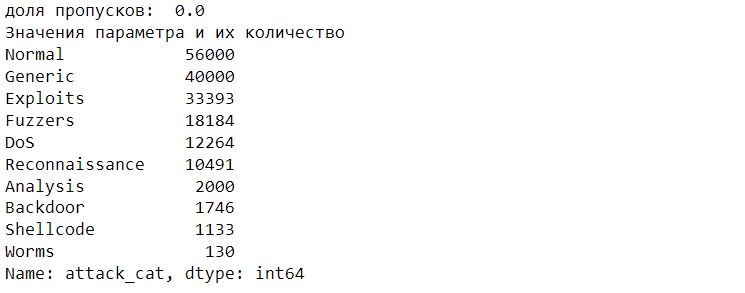






Типы атак, которые и будут нашими классами, на которые мы будем кластеризовать:





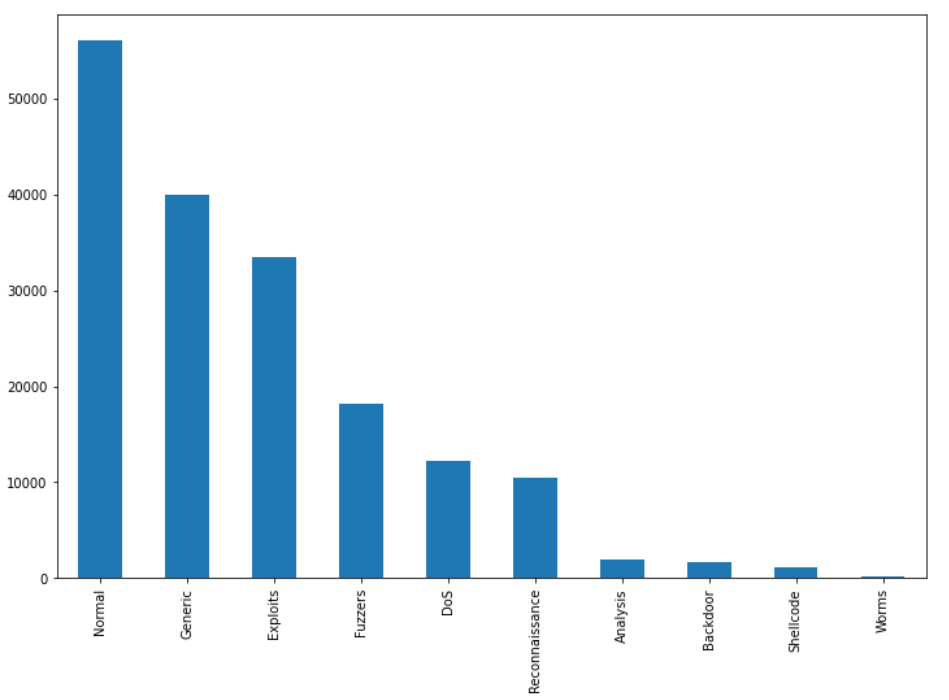


Рис. 5. Осмотр категорийных параметров.

## **2.2 Подготовка датасета**



Рис. 6.Удаление ненужных столбцов.

Неполный датафрейм:

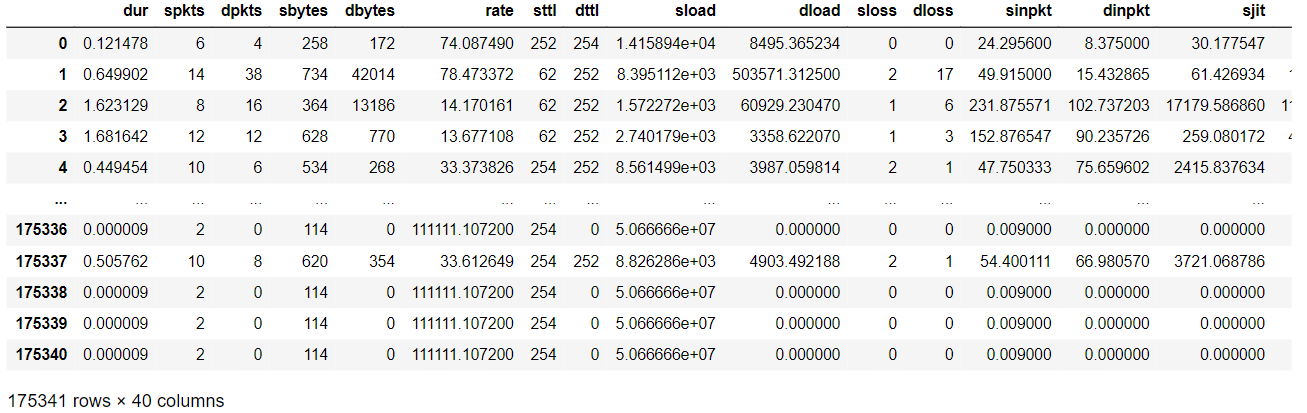
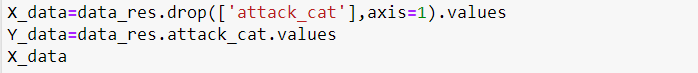
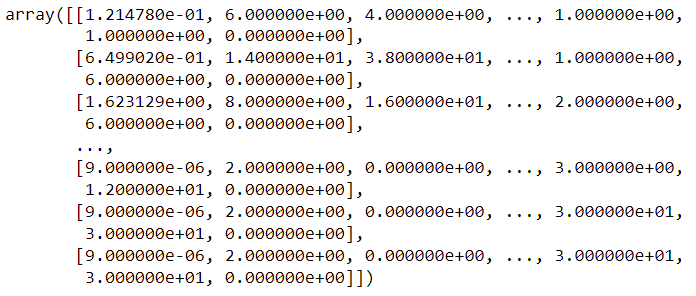
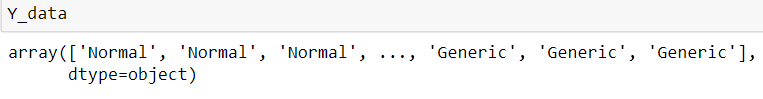


Рис. 7.Неполный датафрейм.

Алгоритмы работают с матрицами или векторами, поэтому подать датафрейм на вход не получится, берем только значения:



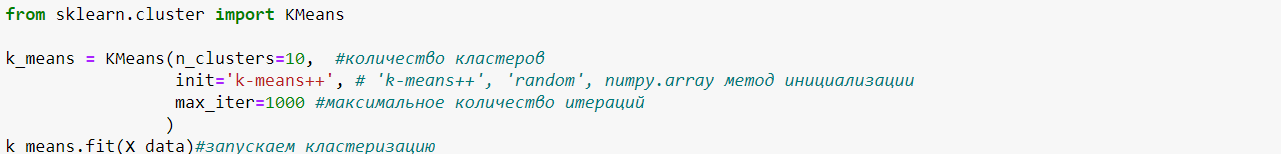


# **3 Демонстрация работы алгоритма**

## **3.1 Алгоритм k-means**

Выберем оптимальное кол-во кластеров для наших данных: в нашем случае это 10 кластеров(по виду атак),один из которых в центре.

Добавим случайным образом 10 центроидов. Центроиды – это предполагаемые центры будущих кластеров. Для каждой точки мы можем посчитать, к какому центроиду она ближе и окрасить в соответствующий цвет. Перенесем центроид в центр выборки, к которой мы его отнесли: то есть расположим его так, чтобы расстояния от объектов кластера до центроида были как можно меньше. Повторим эти шаги, пока алгоритм не сойдется.





Выведем лэйблы, или метки кластера, которые выдал наш алгоритм для обучающей выборки.

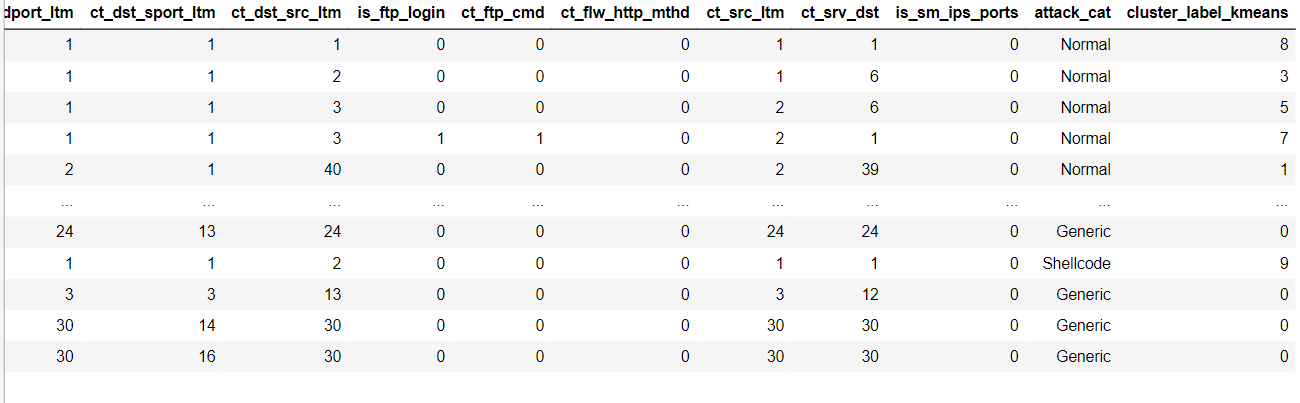


Предсказания в виде вектора не особо информативны. Приклеим наши предсказания к датасету data\_res. Если отлистать датасет в конец и посмотреть на столбцы attack\_cat, cluster\_label\_kmeans видно, что много ошибок даже по первым строкам где класс Normal. Основные причины:

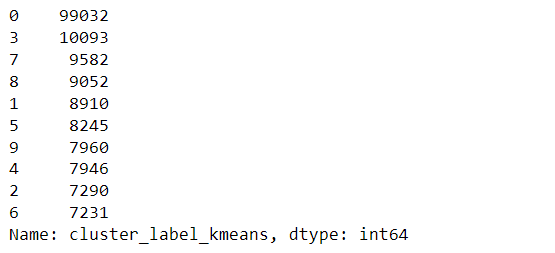
1.Большая размерность пространства.

2.Алгоритмы кластеризации не могут обрабатывать категорийные параметры, а если их включить то только усугубим проблему 1.

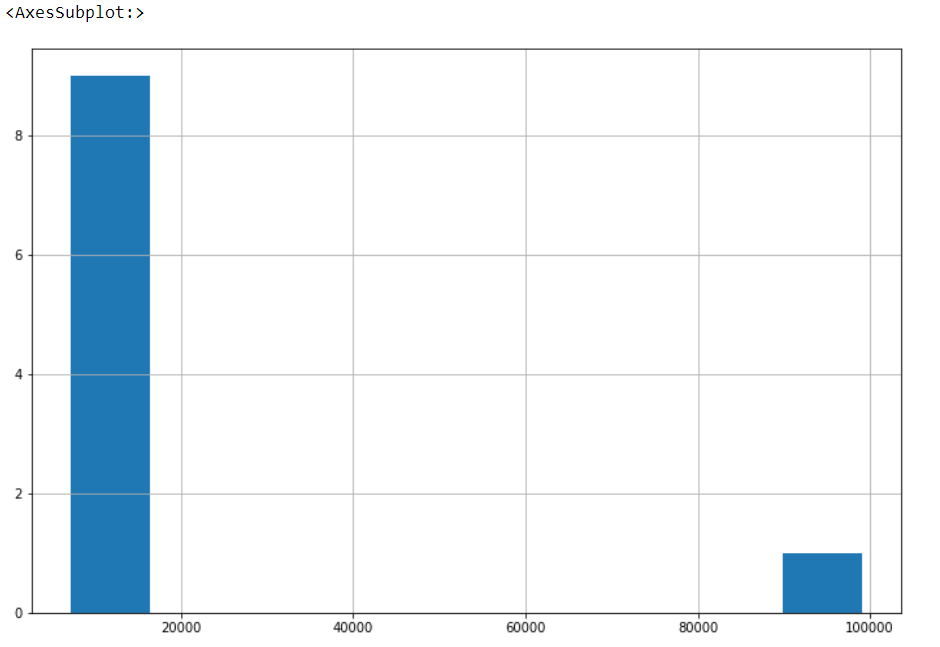




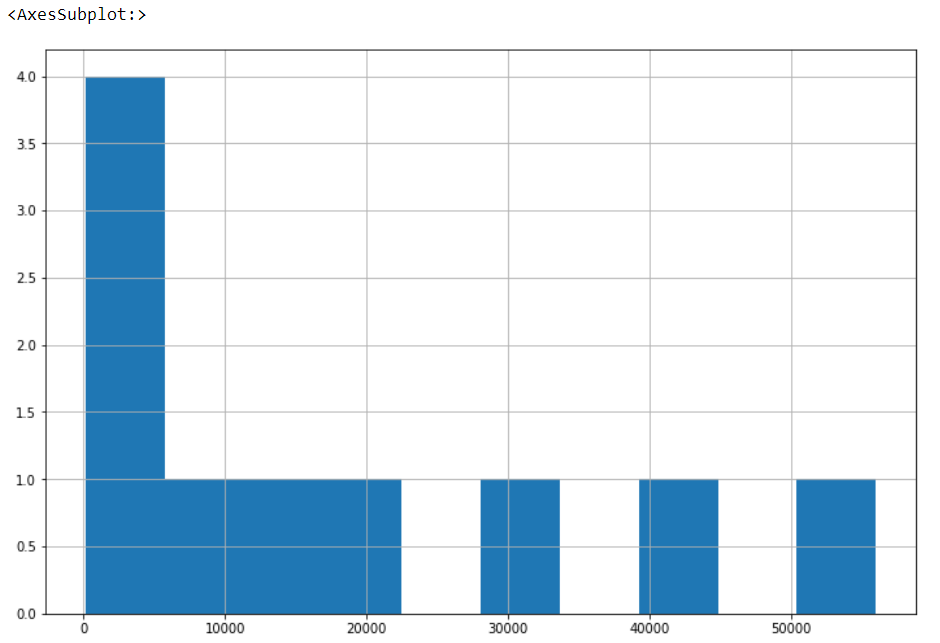








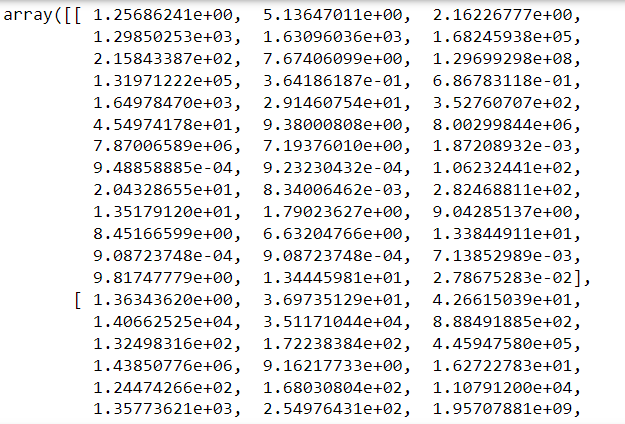




Если сравнить две гистограммы выше, видно, что пропорция классов/кластеров нарушилась.

Посмотрим на координаты центров кластеров для каждого кластера. Из-за большой размерности получается не особо информативно.





После того как провели кластеризацию и известны центры кластеров для новых объектов, можно сразу считать к какому кластеру он относится.

Пример на первой строке датасета:





## **3.2 Алгоритм агломеративной кластеризации**

Как было сказано выше, иерархическая кластеризация бывает двух видов: агломеративная и дивизивная. Будем использовать агломеративную кластеризацию, так как дивизивная обладает большой вычислительной сложностью.

Создадим объект агломеративной кластеризации со следующими свойствами: разбиение ровно на 10 кластеров, использование евклидовой метрики связи, как самой простой (можно также использовать метрики L1, L2, манхэттенскую метрику, косинусную или вычисленную заранее вместе с матрицей расстояний), использование критерия связи минимальной дисперсии Уорда (можно также использовать полный критерий, средний или одиночный). Код подготовки алгоритма представлен на рис. 1.

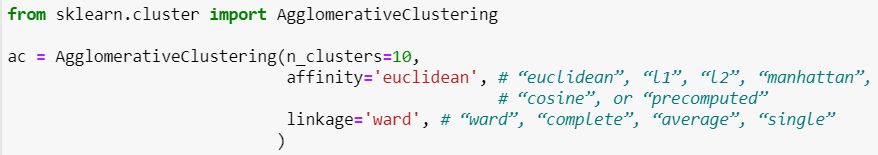


Рис. 8. Подготовка агломеративной кластеризации.

Затем кластеризуем матрицу и посмотрим номера классов элементов, как указано на рис. 2.

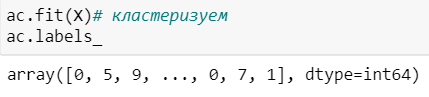


Рис. 9. Кластеризация.

Затем выведем кол-во элементов в кластерах, как указано на рис. 3. Таким образом алгоритм агломеративной кластеризации был выполним.

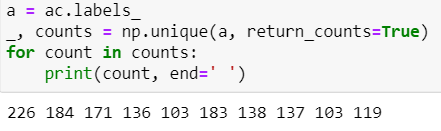


Рис. 10. Количество элементов в кластерах.

# **Заключение**

Для достижения данной цели, в процессе прохождения учебной практики (учебно-лабораторного практикума) изучила существующие методы кластеризации, а также сравнила их. На основе полученных знаний произвела кластеризацию выбранного датасета.

Также были изучены требования к написанию отчета по практике. В результате прохождения практики был составлен отчет по практике, соответствующий предъявленным требованиям.

В ходе прохождения практики все задачи были выполнены, а цель достигнута.