

0.1 Пункт 1

Для проведения эксперимента фиксировалась выборка размером 100, k равное 5 и d равное 0.2. α варьировалась от 0.5 до 10 с делением на 60 значений. Усреднение шло по 10 различным значениям, так как такое уже позволило понять форму большинства распределений.

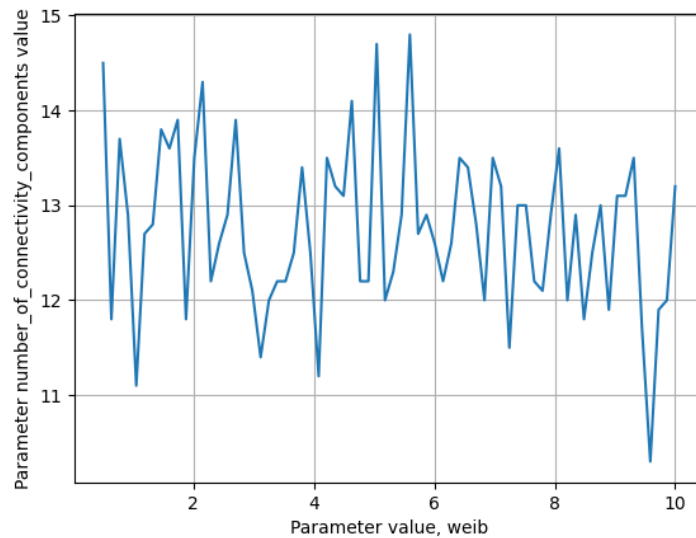


Рисунок 1 – Зависимость числа компонент связности от α при распределении *weibull*

В случае анализа числа компонент связности в knn при обоих распределениях (рис. 1 и рис. 2) их распределение судя по всему независимо от α и имеет при наших условиях среднее около 12.5 в случае *weibull* и 11 в случае *exp*.

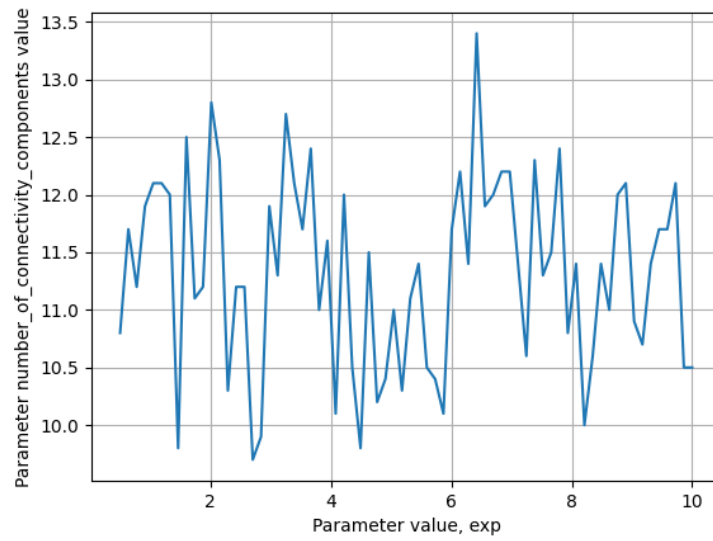


Рисунок 2 – Зависимость числа компонент связности от α при распределении *exp*

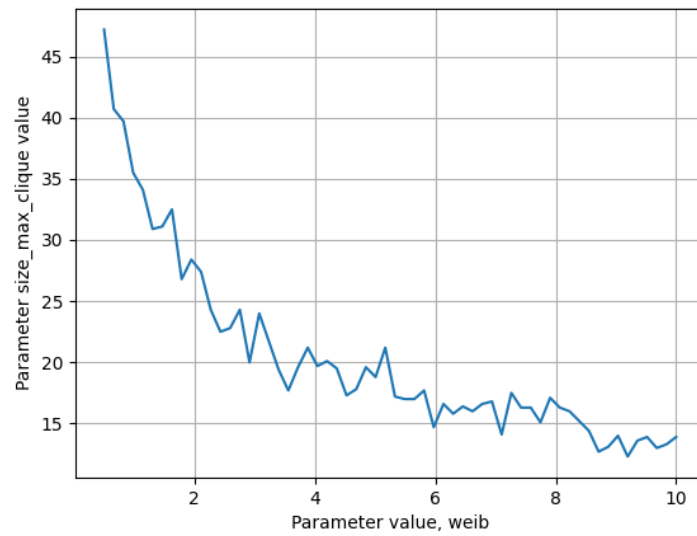


Рисунок 3 – Зависимость максимальной клики от α при распределении *weibull*

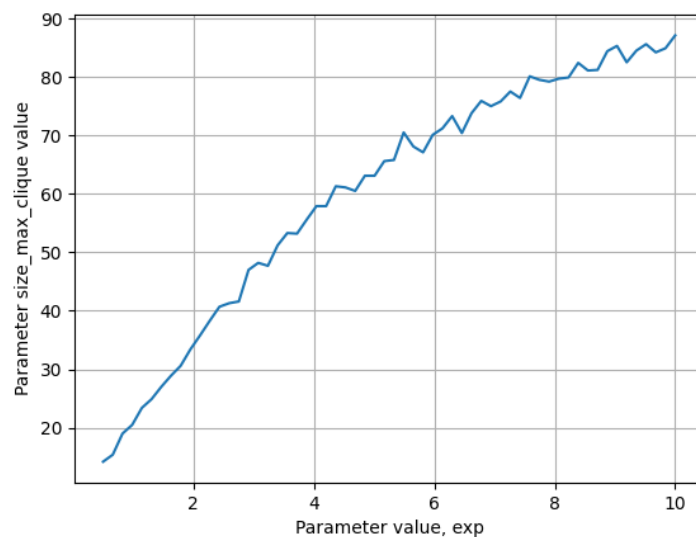


Рисунок 4 – Зависимость максимальной клики от α при распределении exp

В случае же с максимальной кликой видно (рис. 3 и рис. 4), что распределение напоминает степенную функцию, но с совершенно разными степенями. По моим расчетам при наших условиях степень составляет около в случае weibull и $2/5$ в случае с exp.

0.2 Пункт 2

Для проведения эксперимента фиксировалась данное в задании α и значения k проходили от 2 до 20 с шагом 1, значения d проходили от 0.05 до 10 с делением на 60 участков и значения n проходили от 50 до 100 с шагом 2. Усреднение шло по 10 различным значениям аналогично первому пункту.

Тут можно отметить, что от k зависимость степенная с отрицательным коэффициентом в обоих случаях, от d зависимость степенная с коэффициентом меньше 0, а от n зависимость линейная положительная во всех случаях, при том с меньшей дисперсией при подсчете кликового числа.

0.3 Пункт 3

После запуска функции мощность полученного A на выборке размером 300 и с количеством итераций 1000 составило `power = 0.9919999999999998`, `error = 0.9999999999999998` для `knn` и `power = 0.5780000000000001`, `error = 1.0` для `dist`. Это говорит о том, что в случае с числом компонент связности принимаемые значения похожи друг на друга и обеих плотностей, а вот кликовое число достаточно разнится, но все равно вероятность ошибиться можно оценить примерно как 50 на 50.

0.4 Анализ функций по их параметрам

Для проведения экспериментов фиксировалась выборка размера размера 100, k равное 5 и d равное 0.2:

1) `stud` распределение (рис 15)

Максимальная степень графа не влияет на числовую характеристику при изменении параметра

2) `lap` распределение (рис 16)

Размер максимального независимого множества к числовой характеристике стремиться к прямой зависимости, то есть чем больше параметр тем больше размер максимального независимого множества

0.5 Анализ функций по k и d

Для проведения экспериментов фиксировалась выборка размера 100:

1)stud распределение(рис 17)

Максимальная степень графа имеет линейную зависимость при изменении параметра k при создании gk

2)lap распределение(рис 18)

Размер максимального независимого множества k d проявляет примерно обратную зависимость

0.6 Анализ функций по выборке n

Для проведения экспериментов фиксировалась k равное 5 и d равное 0.2:

1)stud распределение(рис 19)

Максимальная степень графа не влияет на числовую характеристику при изменении выборки

2)lap распределение (рис 20)

Размер максимального независимого множества k размеру выборки n стремиться к прямой зависимости, то есть чем больше d тем больше размер максимального независимого множества

0.7 Пункт 3

Для Лапласа и Стьюдента:

После запуска функции при выборке 300 и количества итераций 1000 мощность A вышла 0.13, а ошибка 1.0. Вероятность неправильно принять H1 составляет не более 13 процентов.

0.8 Часть 2. Исследование важность характеристик

Важность признаков у stud и lap при постоянном выбранном нами n:

max_degree: 0.6129

size_max_independent_set: 0.3871

Важность признаков у exp и weib при постоянном выбранном нами n:

`number_of_connectivity_components: 0.4483`

`size_max_clique: 0.5517`

Посмотрим как выглядит график при различных n (рисунки 21 и 22):

Для `stud` и `lap` - синяя линия - `max_degree`, а желтая - `size_max_independent_set`. Из этого можно сделать вывод, что для определения большую роль играет максимальная степень

Для `exp` и `weib` - синяя линия - `number_of_connectivity_components`, а желтая - `size_max_clique`. Из этого можно сделать вывод, что для определения большую роль играет число компонент связности.

0.9 Часть 2. Исследование метрики

Для `stud` и `lap` - при минимальном n - лучшим алгоритмом будет K -ближайших соседей, при остальных n , чем больше n тем лучше результат, а при максимальном n точность будет равна 1 для всех алгоритмов.

Также для каждого n мы вывели Confusion matrix (рисунки 23-25)

Ошибка первого рода: 0.5

Мощность: 0.5

Точность: 1.0

Для `exp` и `weib` - при минимальном n - лучшим алгоритмом будет Логистическая регрессия и K -ближайших соседей, при остальных n , чем больше n тем лучше результат, а при максимальном n Дерево и Логистическая регрессия будут лучшими алгоритмами.

Также для каждого n мы вывели Confusion matrix (рисунки 26-28)

Ошибка первого рода: 0.02

Мощность: 0.23

Точность: 0.71

0.10 Реализация

В 1 части - каждый реализовывал свои функции

Создание `gd` - Лев, создание `gk` - Илья. Первый пункт - Лев и Илья, второй пункт - Лев и Илья, третий пункт - Илья.

Во 2 части - первый и второй пункт - Лев, третий - Илья.

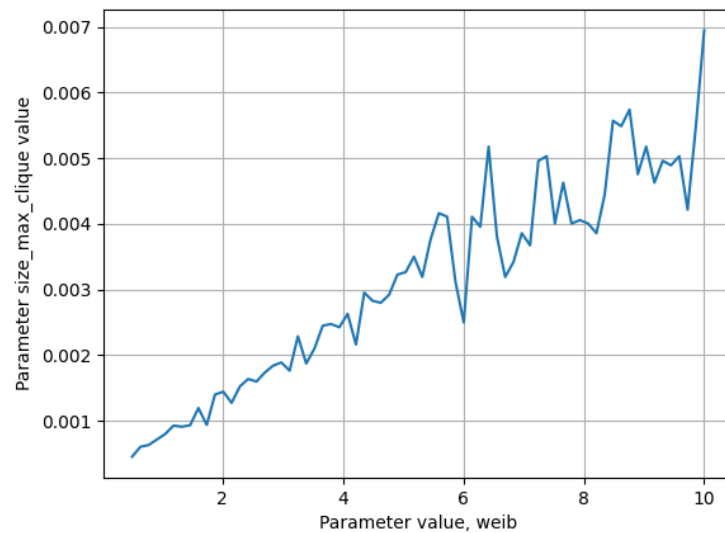


Рисунок 5 – Зависимость максимальной клики от α при распределении weibull после выравнивание возведением в степень

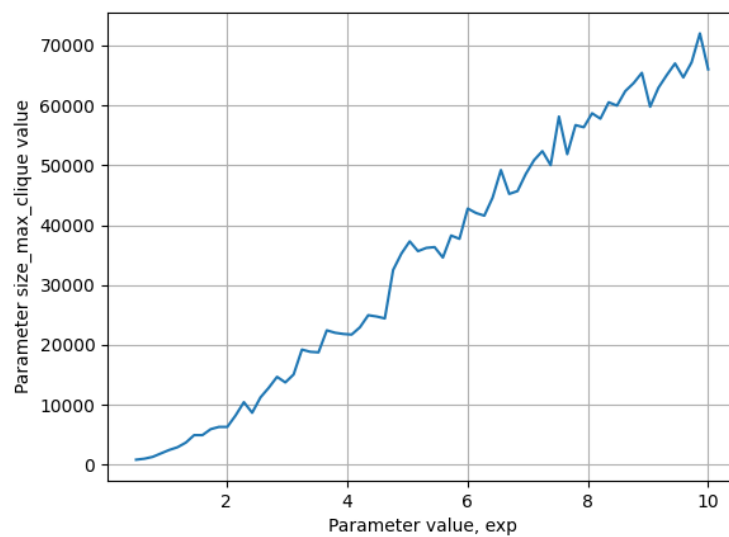


Рисунок 6 – Зависимость максимальной клики от α при распределении exp после выравнивание возведением в степень

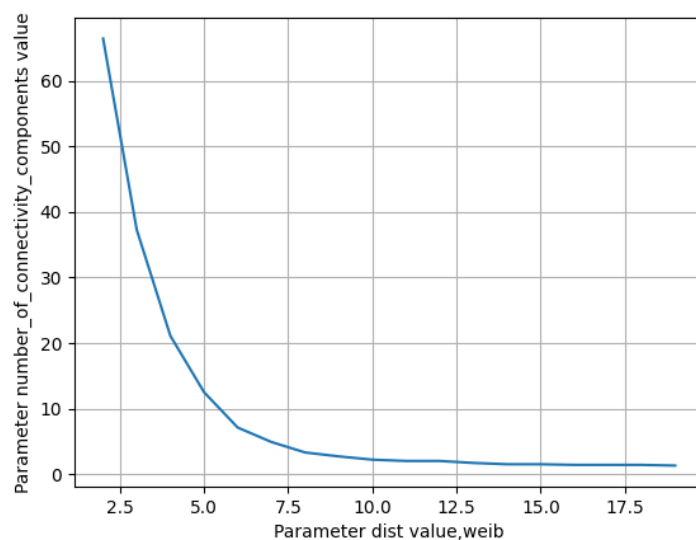


Рисунок 7 – Зависимость числа компонент связности от k при распределении weibull

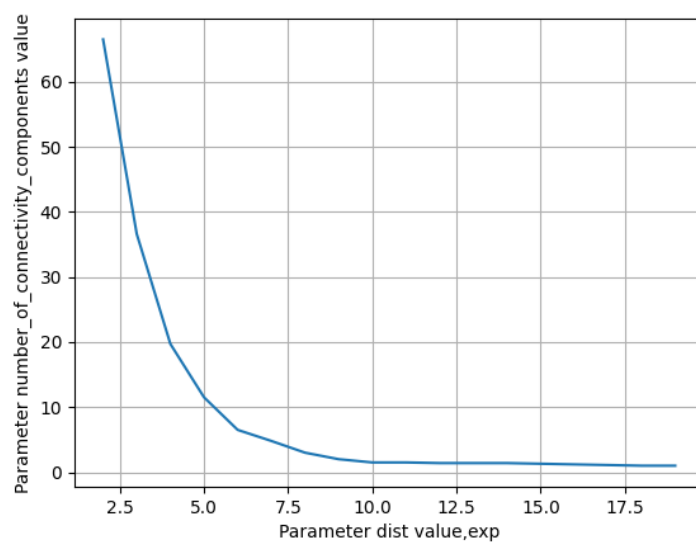


Рисунок 8 – Зависимость числа компонент связности от k при распределении exp

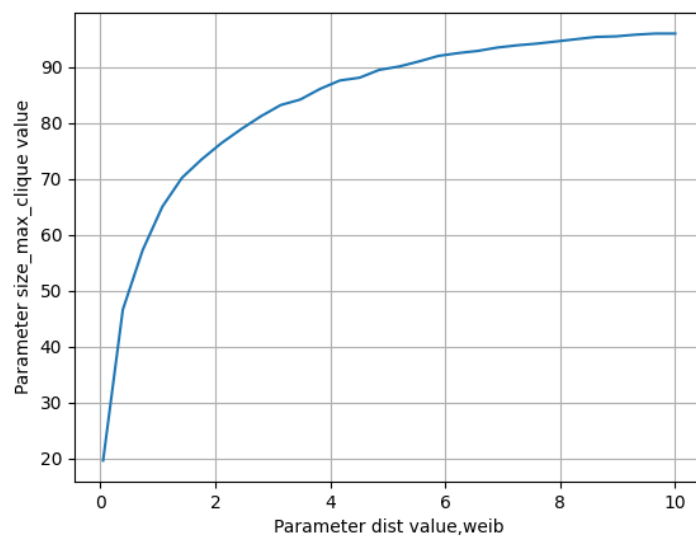


Рисунок 9 – Зависимость размера максимальной клики от d при распределении *weibull*

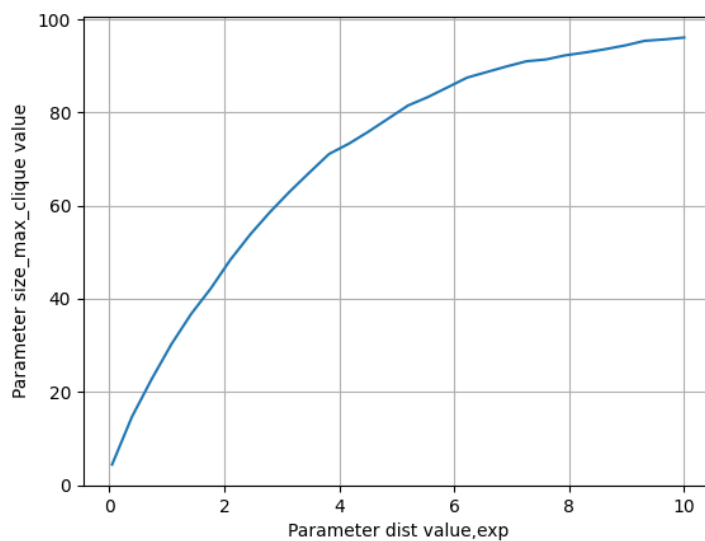


Рисунок 10 – Зависимость размера максимальной клики от d при распределении *exp*

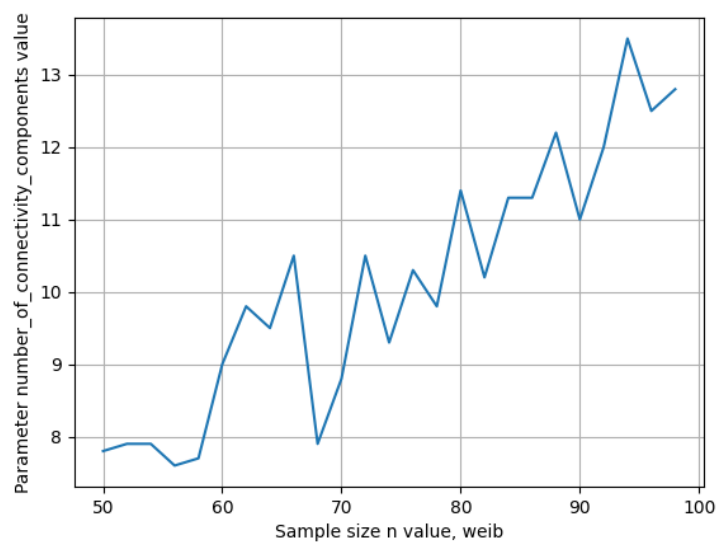


Рисунок 11 – Зависимость числа компонент связности n при распределении *weibull*

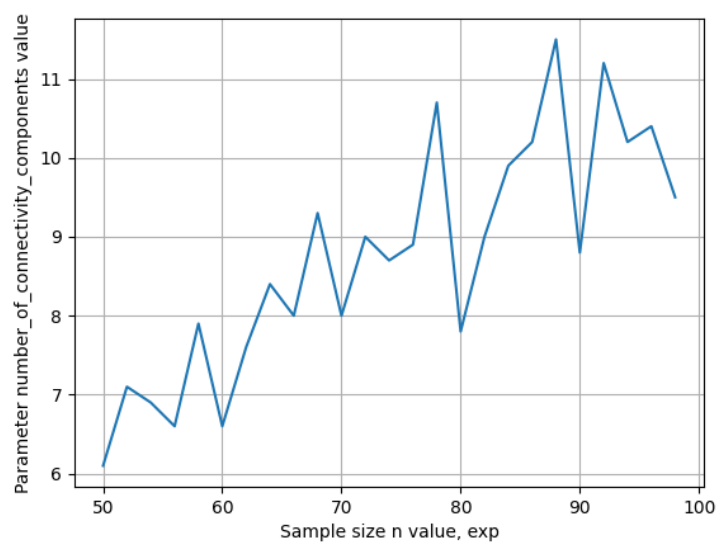


Рисунок 12 – Зависимость числа компонент связности n при распределении *exp*

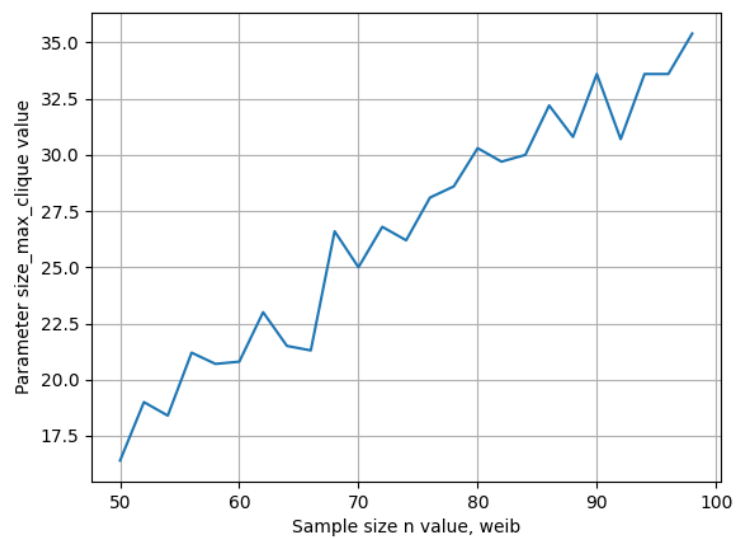


Рисунок 13 – Зависимость размера максимальной клики от n при распределении *weibull*

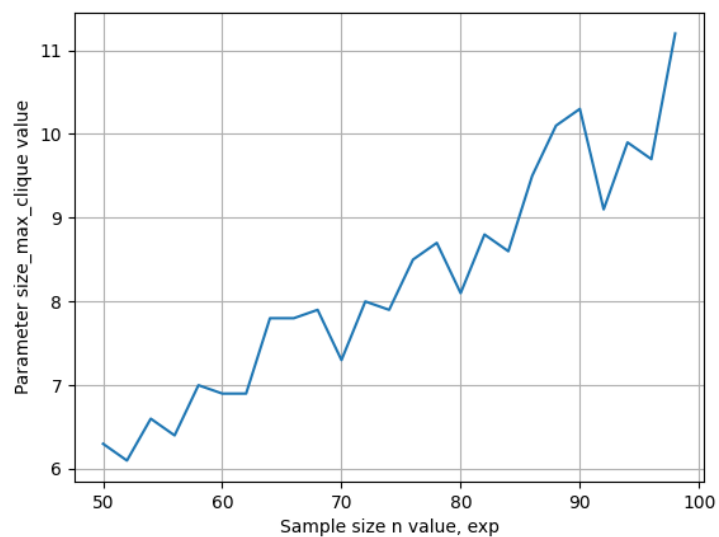


Рисунок 14 – Зависимость размера максимальной клики от n при распределении *exp*

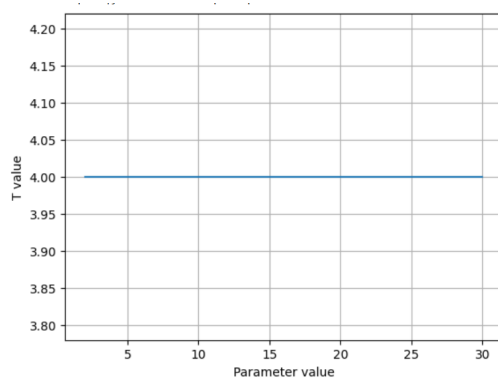


Рисунок 15 – Анализ по параметрам - *stud* распределение

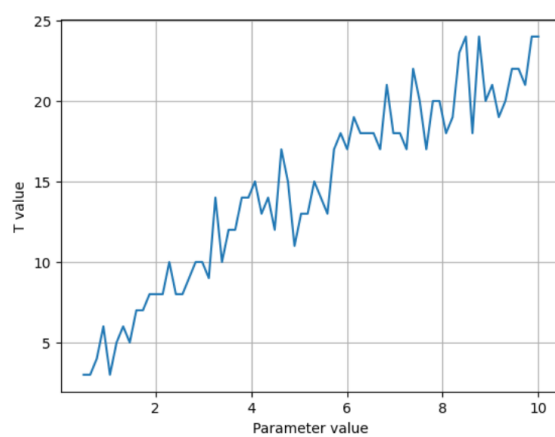


Рисунок 16 – Анализ по параметрам - *lap* распределение

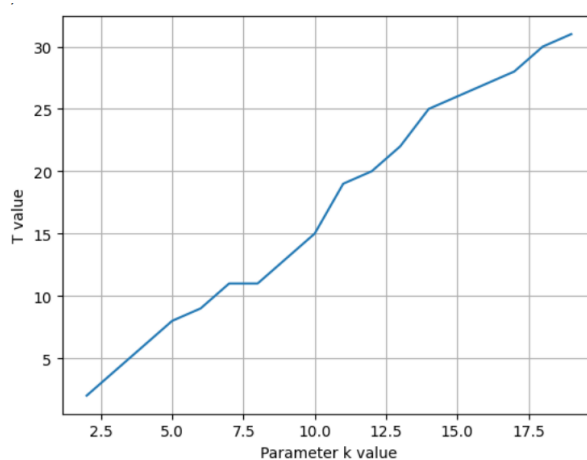


Рисунок 17 – Анализ по *k* - *stud* распределение

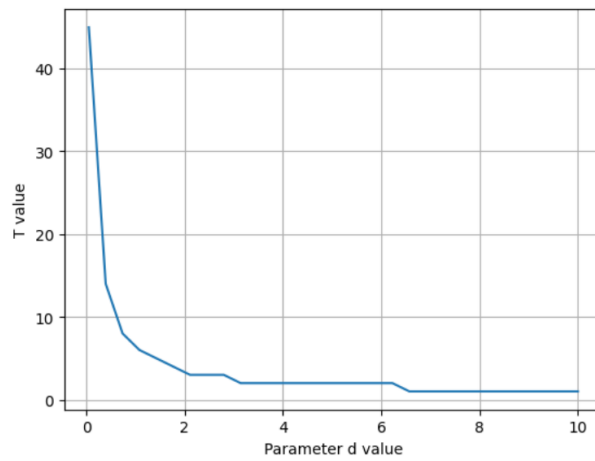


Рисунок 18 – Анализ по d - lap распределение

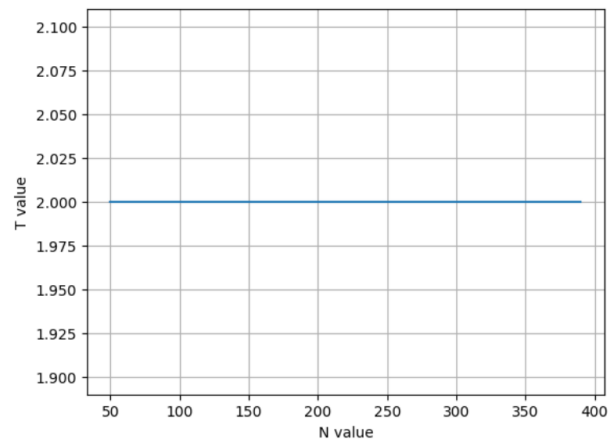


Рисунок 19 – Анализ по n - $stud$ распределение

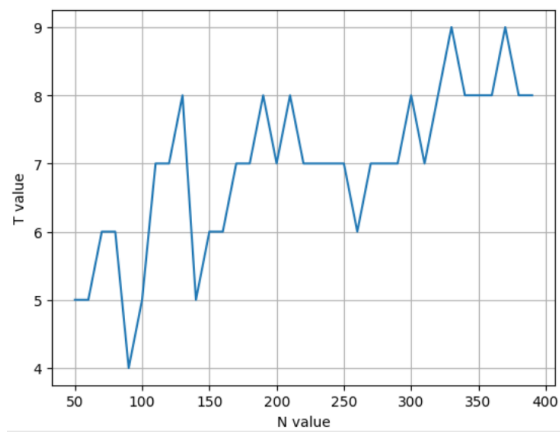


Рисунок 20 – Анализ по n - lap распределение

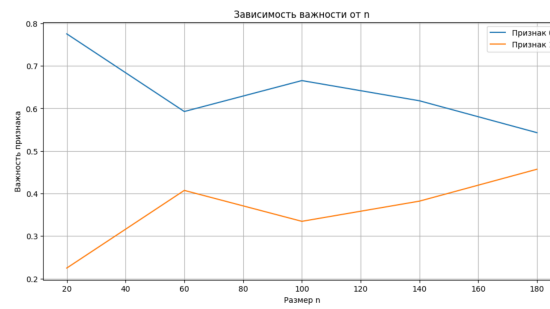


Рисунок 21 – Важность признаков у stud и lap

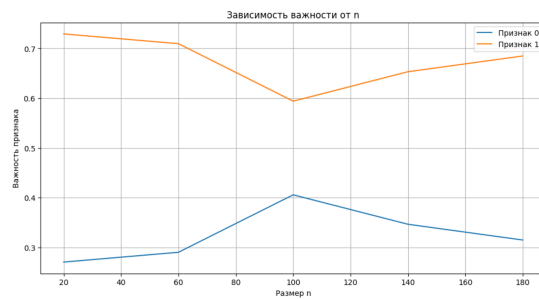


Рисунок 22 – Важность признаков у ехr и weib

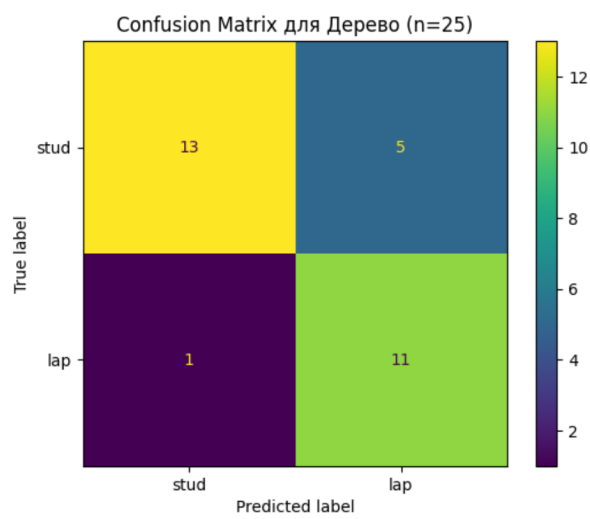


Рисунок 23 – Confusion matrix lap and stud n=25

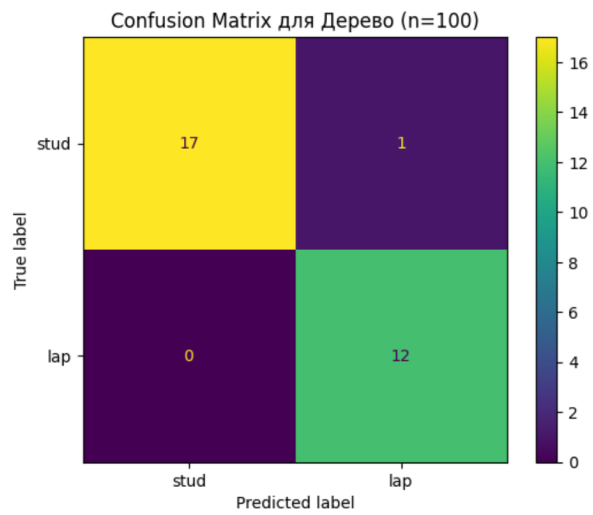


Рисунок 24 – Confusion matrix lap and stud n=100

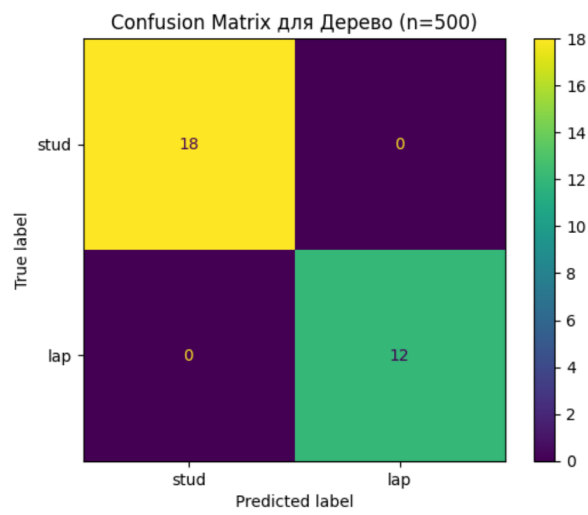


Рисунок 25 – Confusion matrix lap and stud n=500

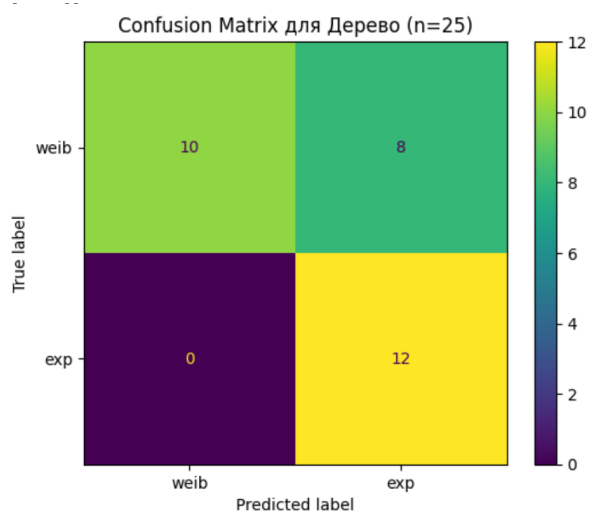


Рисунок 26 – Confusion matrix weib and exp n=25

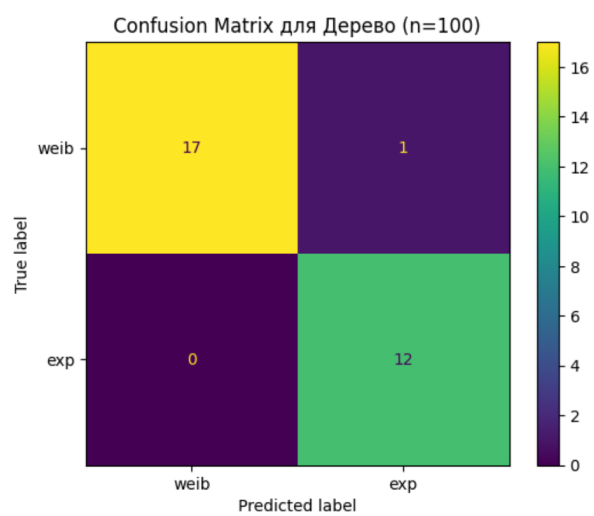


Рисунок 27 – Confusion matrix weib and exp n=100

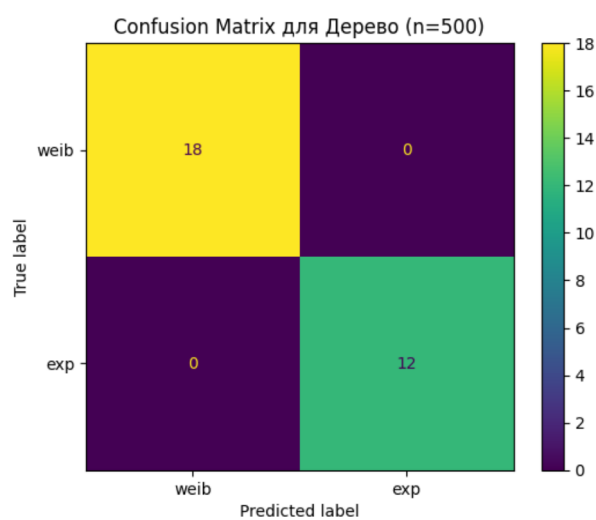


Рисунок 28 – Confusion matrix weib and exp n=500