

Отчет по проекту: Задачи по случайным графам

Бахурин Виктор и Стахова Екатерина

30 мая 2025 г.

Содержание

1	Введение	2
2	Описание кода	2
2.1	Используемые инструменты	2
2.2	UML-диаграмма	2
2.3	Реализованные алгоритмы	2
2.3.1	<i>fast_chromatic_number()</i>	2
2.3.2	<i>fast_max_independent_set_size()</i>	2
2.3.3	<i>greedy()</i>	2
3	Описание экспериментов	3
3.1	Эксперимент 1	3
3.1.1	Цель	3
3.1.2	Результаты	3
3.2	Эксперимент 2	5
3.2.1	Цель	5
3.2.2	Результаты	5
3.3	Эксперимент 3	7
3.3.1	Цель	7
3.3.2	Результаты	8
3.4	Эксперимент 4	8
3.4.1	Цель	8
3.4.2	Результаты	9
3.5	Промежуточный вывод	9
3.6	Эксперимент 5	10
3.6.1	Цель	10
3.6.2	Результаты	10
3.7	Эксперимент 6	10
3.7.1	Цель	10
3.7.2	Результаты	11
3.7.3	Результаты	11
3.8	Эксперимент 7	12
3.8.1	Цель	12
3.8.2	Результаты	12
3.8.3	Результаты	13
3.9	Оценка собственного классификатора	16

1 Введение

Часть I. Исследование свойств характеристики

2 Описание кода

2.1 Используемые инструменты

- Язык программирования: Python 3.10
- Основные библиотеки: numpy, networkx, matplotlib, scikit-learn
- Система контроля версий: Git (GitHub/GitLab)
- Дополнительные инструменты: Jupyter Notebook, PyCharm, Google Colab

2.2 UML-диаграмма

Мы не реализовывали свои классы.

2.3 Реализованные алгоритмы

2.3.1 *fast_chromatic_number()*

- **Назначение:** Вычисление хроматического числа для случайного графа построенного на данной выборке.
- **Входные данные:** list - выборка
- **Выходные данные:** int - хроматическое число
- **Сложность:** $O(n \log(n))$

2.3.2 *fast_max_independent_set_size()*

- **Назначение:** Вычисление размера максимального независимого множества для случайного графа построенного на данной выборке.
- **Входные данные:** graph - граф
- **Выходные данные:** int - размер независимого множества
- **Сложность:** $O(n+m)$

2.3.3 *greedy()*

- **Назначение:** Жадное построение множества A , максимизирующие мощность критерия, при заданной допустимой ошибке первого рода.
- **Входные данные:** T_{H_0} , T_{H_1} , α - два набора наблюдений и максимальная допустимая ошибка первого рода.
- **Выходные данные:** A , *current_error*, power - множество A , ошибка первого рода, мощность критерия.
- **Сложность:** $O(n \log(n))$

3 Описание экспериментов

3.1 Эксперимент 1

3.1.1 Цель

Исследовать, как ведет себя числовая характеристика T в зависимости от параметров распределений и , зафиксировав размер выборки и параметр процедуры построения графа KNN.

3.1.2 Результаты

Мы получили интересный результат. График для нормального распределения выглядит хаотичнее, чем график для Student-t(); в графике Student-t() прослеживается рост $E[in_\delta(G)]$ с ростом параметра . И еще одно интересное наблюдение: для интересующих нас параметров распределений v_0 и 0 график распределения Student-t() ниже графика нормального распределения.

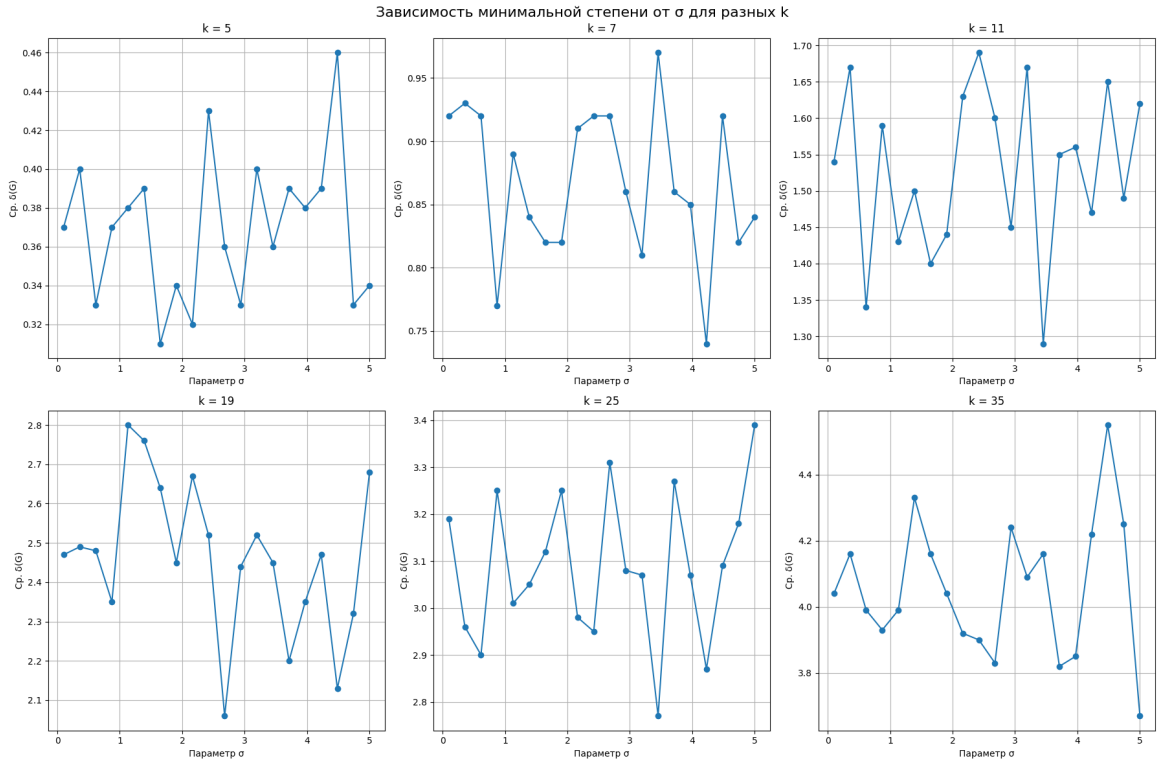
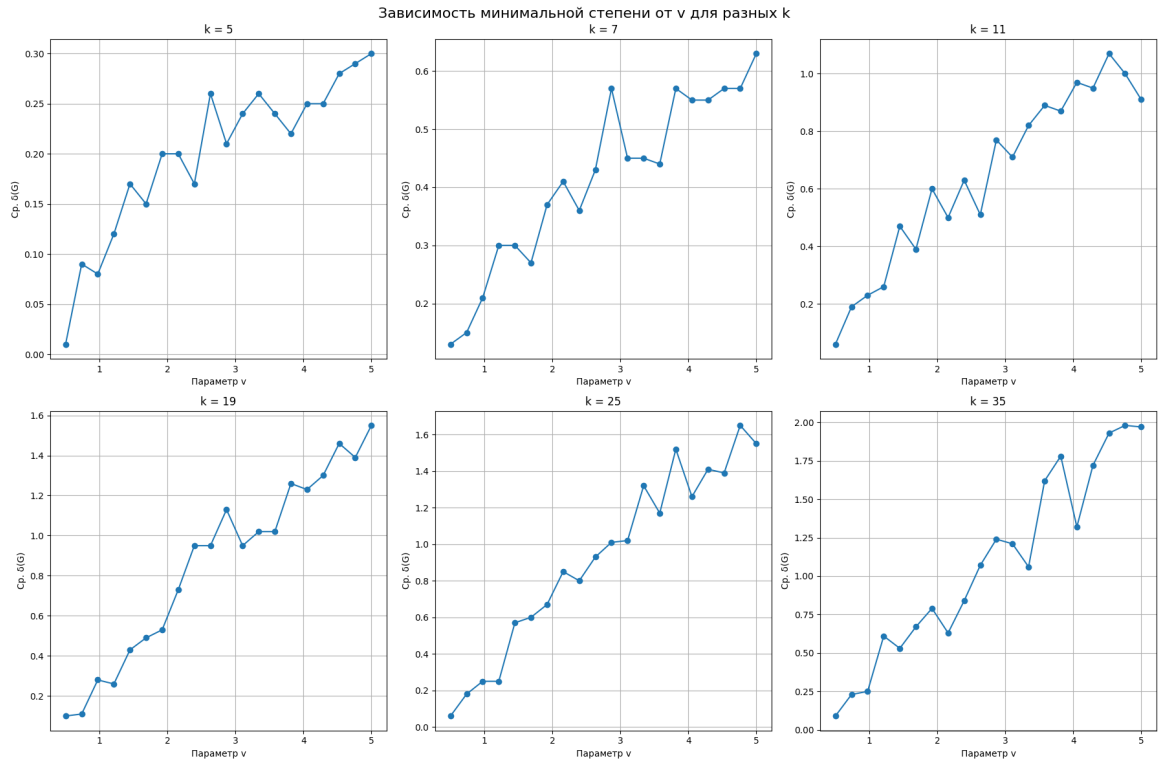
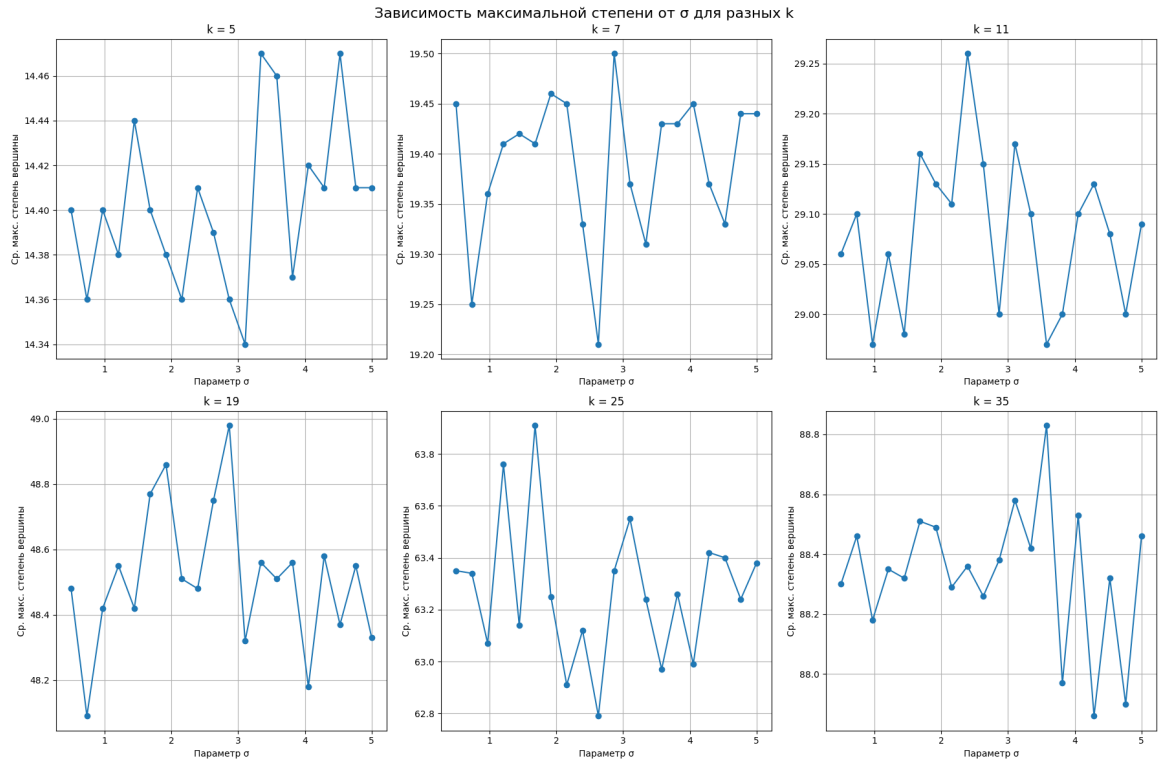


Рис. 1: $E[in_\delta(G)]$ для KNN графа построенного на $Normal(0, \sigma)$



Графики экспоненциального и гамма-распределения выглядят хаотично. Не прослеживается никакая зависимость от параметров.



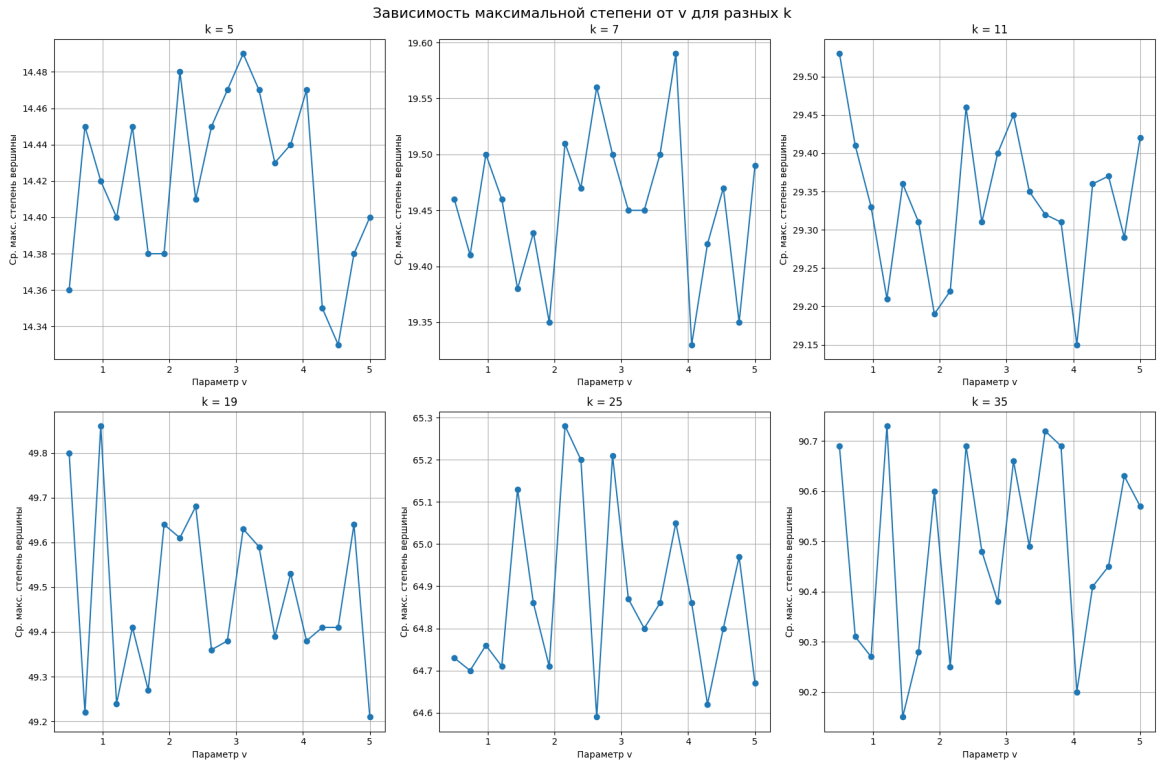


Рис. 4: Максимальная степень вершины для KNN графа построенного на $\Gamma_{\text{Gamma}}()$

3.2 Эксперимент 2

3.2.1 Цель

Исследовать, как ведет себя числовая характеристика T в зависимости от параметров распределений и , зафиксировав размер выборки и параметр процедуры построения графа dist .

3.2.2 Результаты

Характеристика (G) на дистанционном графе показывает разные результаты для разных выборок. Для нормального распределения с ростом параметра σ хроматическое число убывает, а для распределения $\text{Student-t}()$ с ростом параметра v $\chi(G)$ наоборот растет.

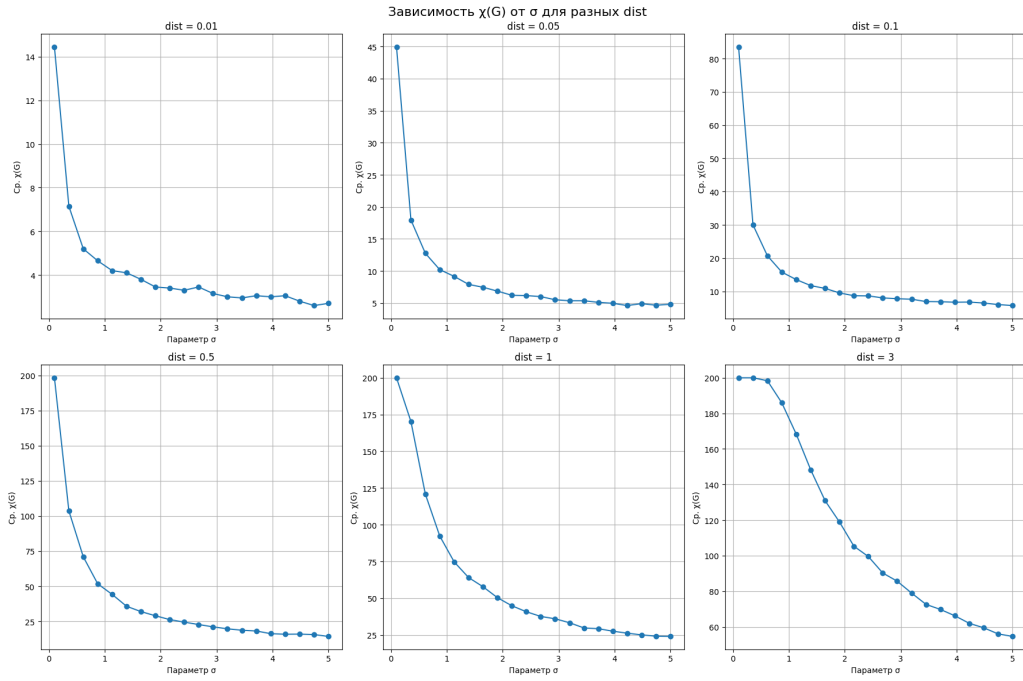


Рис. 5: $E[\chi(G)]$ для dist графа построенного на $Normal(0, \sigma)$

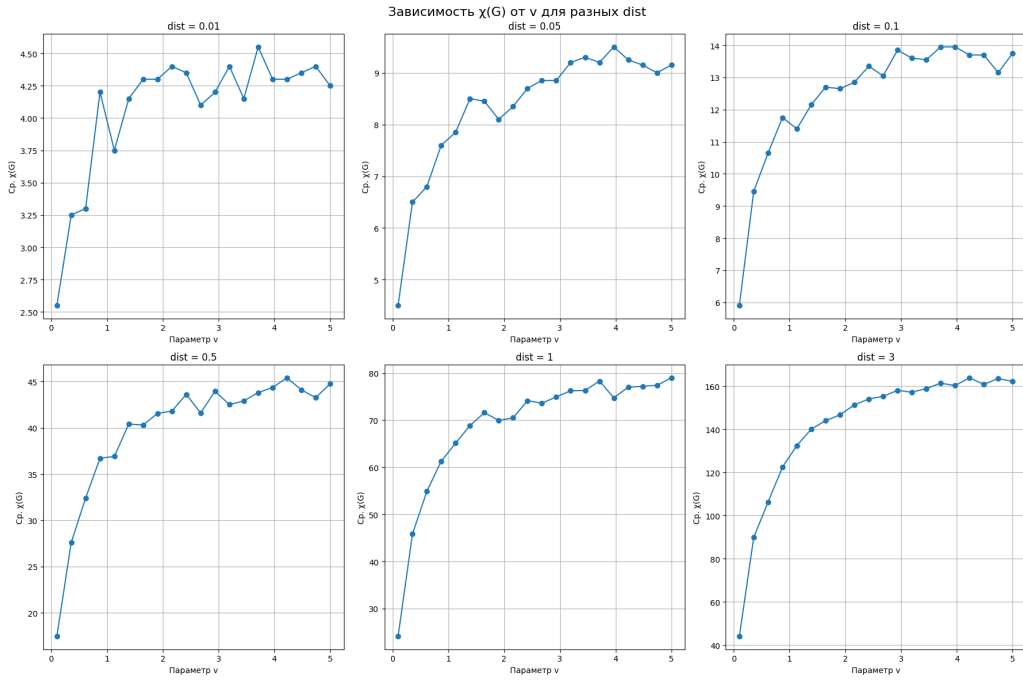


Рис. 6: $E[\chi(G)]$ для dist графа построенного на $Student - t()$

Размер максимального независимого множества убывает с увеличением параметра q и v . Однако для гамма-распределения зависимость несколько более хаотичная.

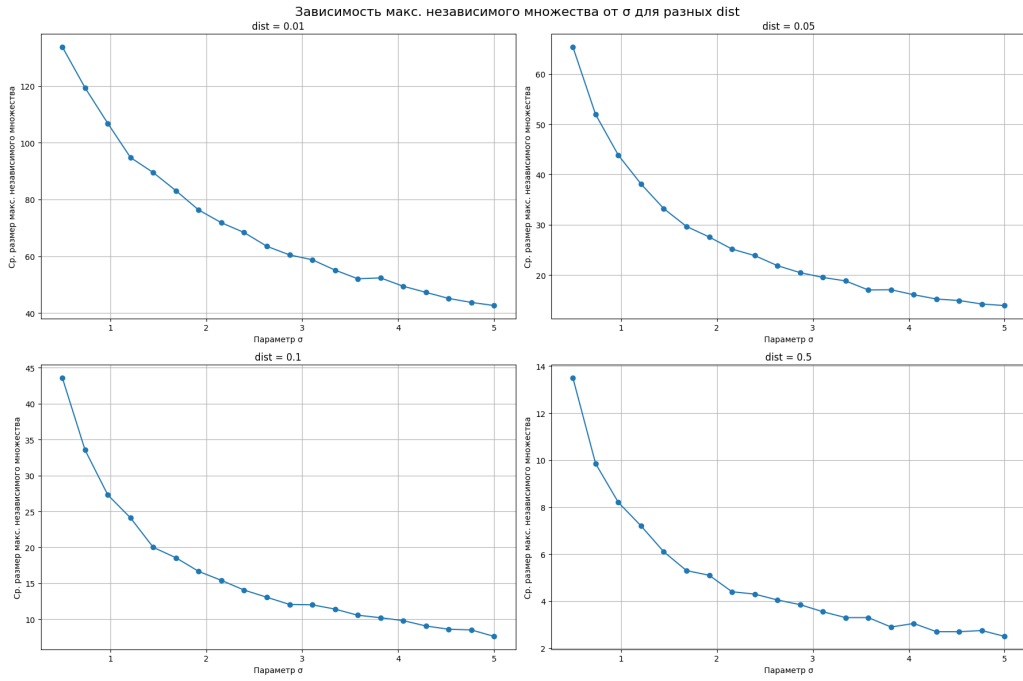


Рис. 7: Размер макс. независимого множества для dist графа построенного на $Exp(\sigma)$

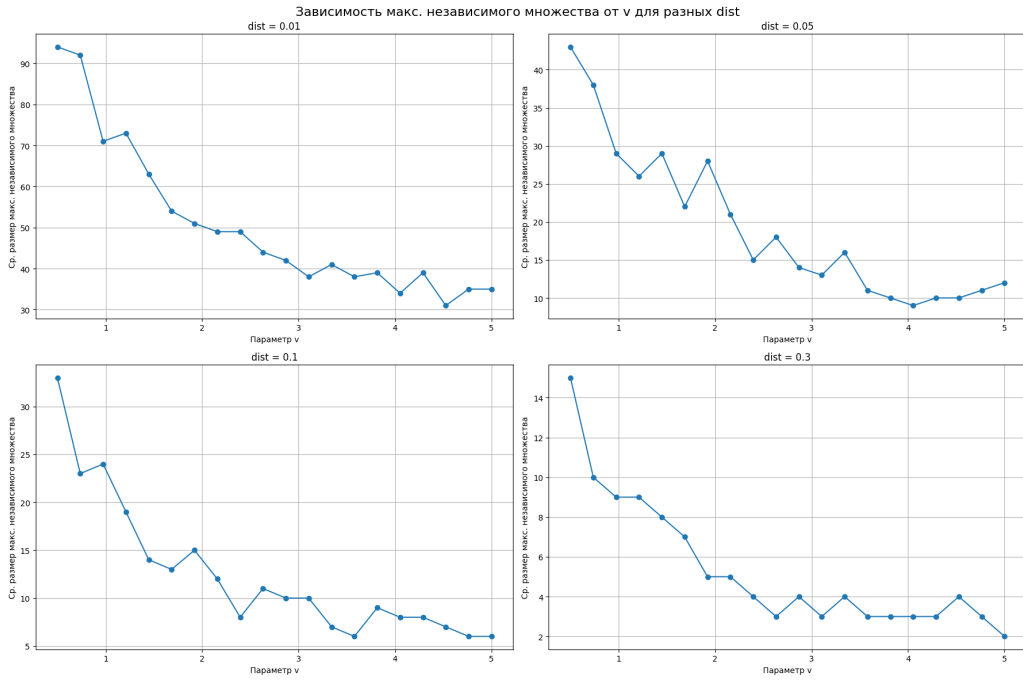


Рис. 8: Размер макс. независимого множества для dist графа построенного на $Gamma()$

3.3 Эксперимент 3

3.3.1 Цель

Исследовать, как ведет себя числовая характеристика T в зависимости от параметров процедуры построения графа KNN и размера выборки при фиксированных значениях $\theta = \theta_0$ и $v = v_0$.

3.3.2 Результаты

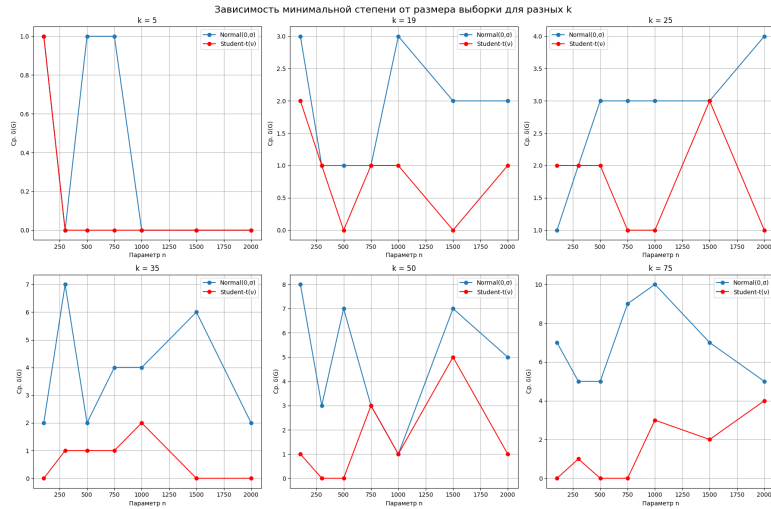


Рис. 9: $E[in_\delta(G)]$ для KNN графа

График для Normal выше, чем график для Student. Это может помочь в проверке истинности H_0 и H_1 .

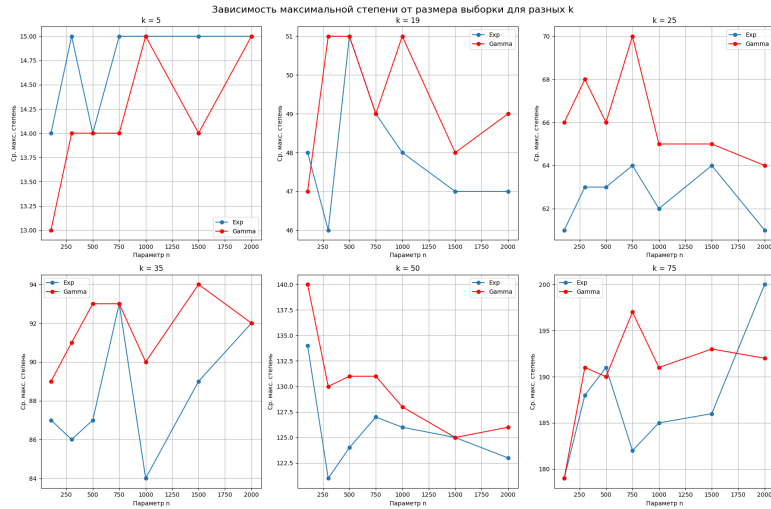


Рис. 10: Макс. степень вершины для KNN графа

При больших k график Гамма-распределения находится выше экспоненциального распределения.

3.4 Эксперимент 4

3.4.1 Цель

Исследовать, как ведет себя числовая характеристика T в зависимости от параметров процедуры построения дистанционного графа и размера выборки при фиксированных значениях $\theta = \theta_0$ и $v = v_0$.

3.4.2 Результаты

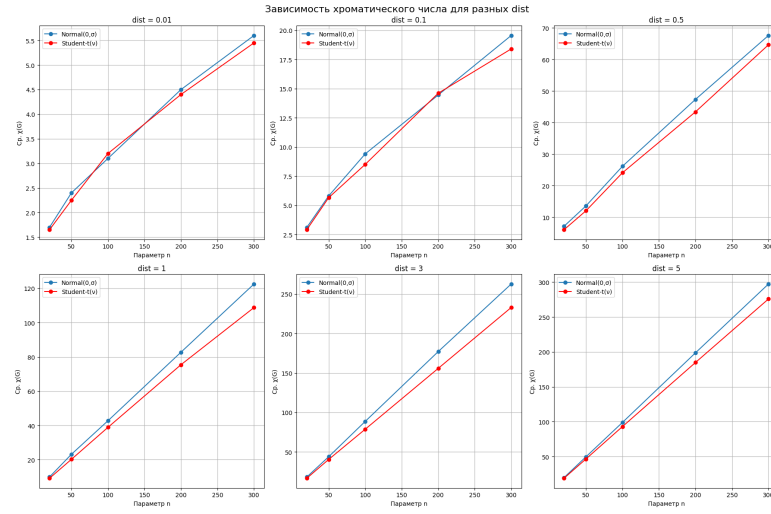


Рис. 11: $E[\chi(G)]$ для dist графа

К сожалению, данные графики не сильно отличаются, в среднем график для Student-t() ниже, чем график $Normal(0, \sigma)$.

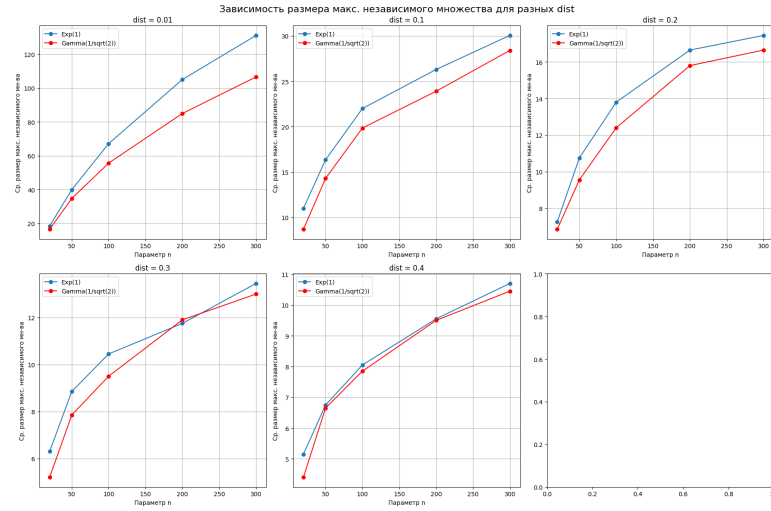


Рис. 12: Размер макс. независимого множества для dist графа

Экспонента в среднем выше, чем гамма-распределение.

3.5 Промежуточный вывод

Если обобщить результаты, полученные в предыдущих пунктах, то можно заметить, что каждая из характеристик показывает разные значения на случайных графах, построенных на распределениях $Student - t()$ и нормальном распределении $Normal(0, \sigma)$. Это означает, что существует возможность использовать их для проверки истинности гипотез H_0 и H_1 . Аналогичные рассуждения верны для экспоненциального и гамма распределений.

3.6 Эксперимент 5

3.6.1 Цель

Построить множество A в предположении $\theta = \theta_0$ и $v = v_0$ при максимальной допустимой вероятности ошибки первого рода $\alpha = 0.055$. Оценить мощность полученного критерия.

3.6.2 Результаты

Для каждой характеристики удалось построить множество A .

Используя характеристику $in_d(G)$ на графе KNN получен следующий результат:

Ошибка первого рода $\alpha = 0.035$.

Мощность полученного критерия 0.717.

Используя характеристику $\chi(G)$ на графе dist получен следующий результат:

Ошибка первого рода $\alpha = 0.045$.

Мощность полученного критерия 0.594.

В первом случае результат значительно лучше.

Используя характеристику макс. степень вершины в графе knn получен следующий результат:

Ошибка первого рода $\alpha = 0.039$.

Мощность полученного критерия 0.303.

Используя характеристику размер макс. независимого множества в графе dist получен следующий результат:

Ошибка первого рода $\alpha = 0.053$.

Мощность полученного критерия 0.314.

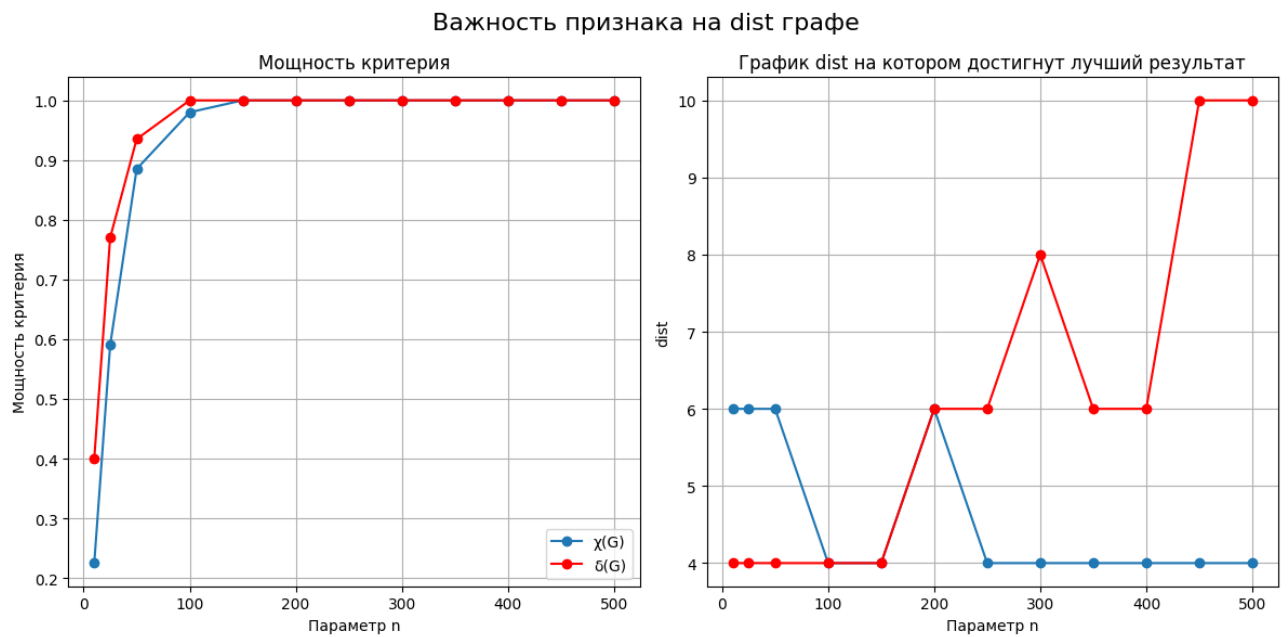
3.7 Эксперимент 6

3.7.1 Цель

Исследование важности характеристик, как признаков классификации. Узнать, меняется ли важность характеристик с ростом n .

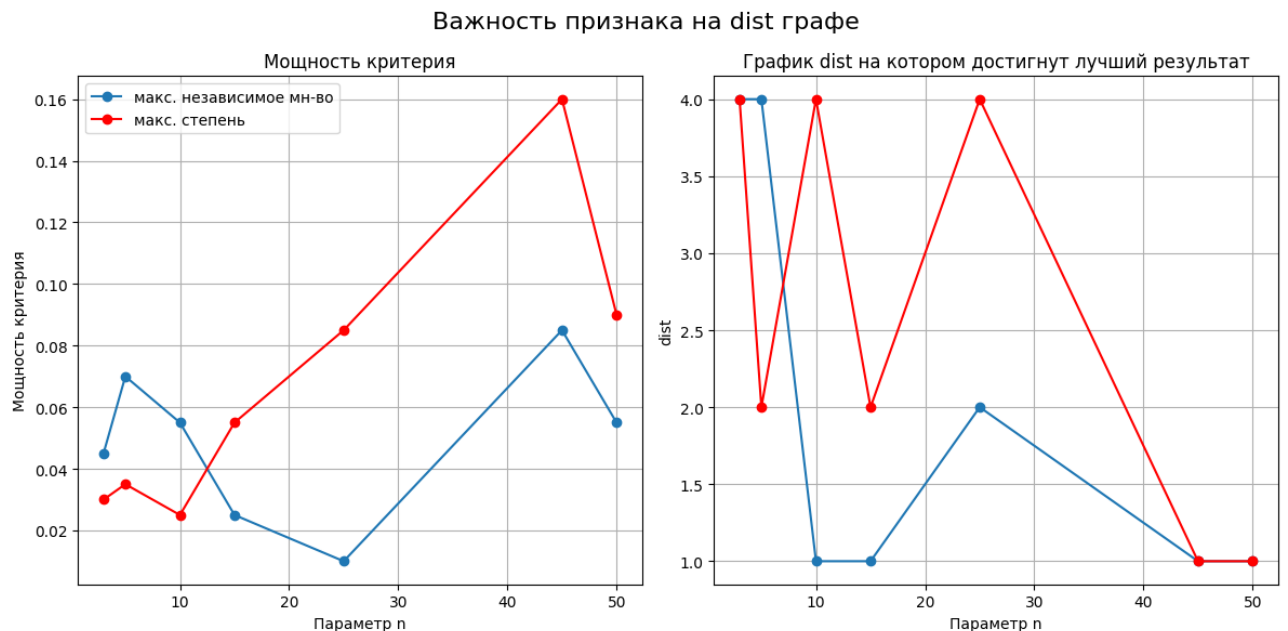
В качестве графа был выбран dist, так как он неориентированный.

3.7.2 Результаты



Мы видим, что каждый из признаков (хроматическое число и мин. степень) дает хорошую мощность критерия на значениях n от 50. При этом мы также видим значение $dist$, на котором было получено лучшее значение мощности критерия.

3.7.3 Результаты



В то же время признаки макс. степень вершины и макс. размер независимого множества дают очень маленькое значение на коротких выборках (до 50) и хорошие показатели, близкие к 1, на выборках (n от 150).

3.8 Эксперимент 7

3.8.1 Цель

Применить разные классификационные алгоритмы и оценить метрики качества. В качестве классификаторов были выбраны следующие: LogisticRegression, RandomForestClassifier и MyClassifier.

3.8.2 Результаты

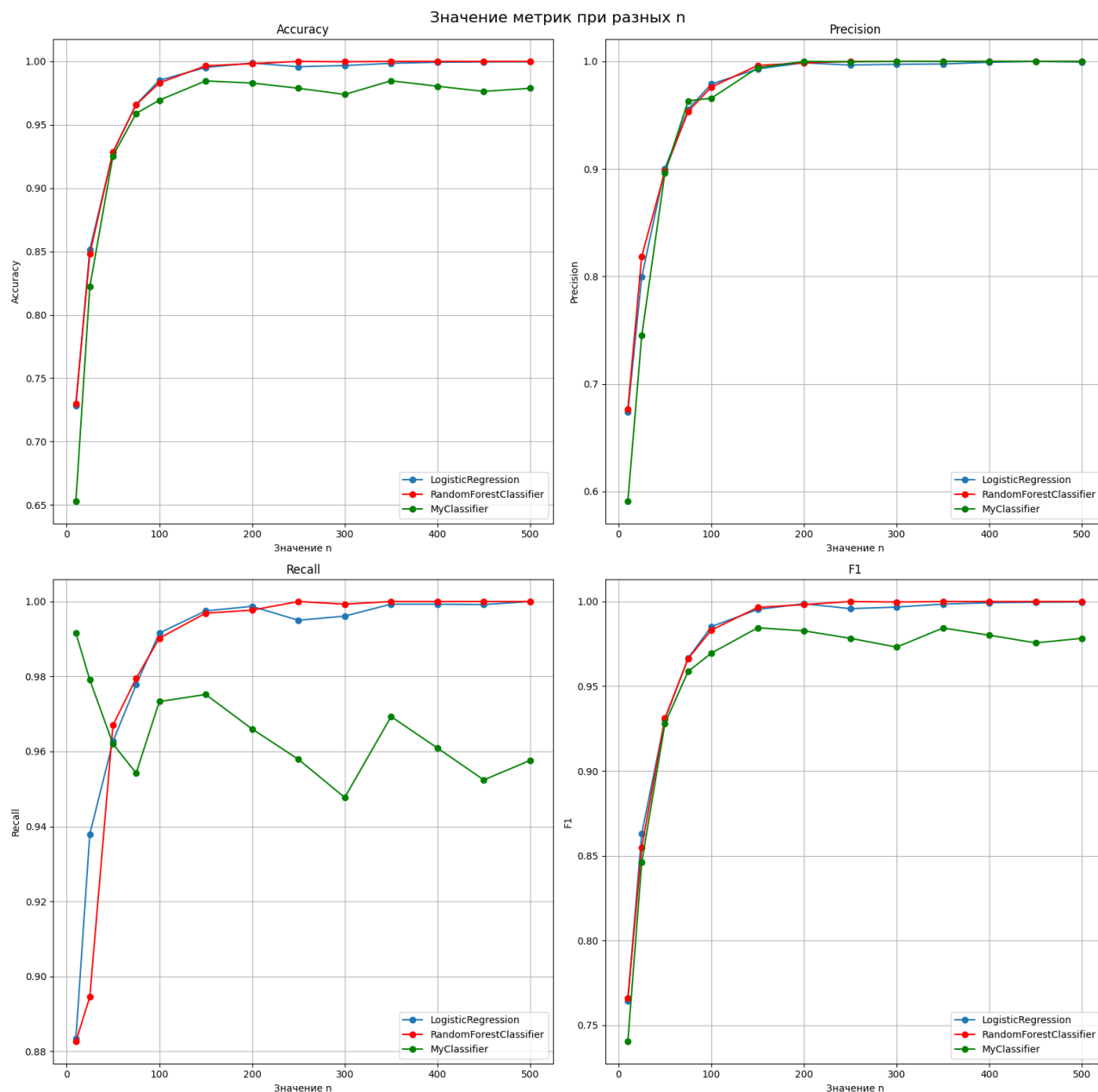


Рис. 15: Значения метрик

Мы видим, что используя характеристики хроматическое число и мин. степень вершины, мы получаем точность 95% уже при $n=75$. Это хороший результат. Стоит отметить, что собственный классификатор показывает не очень хорошую, относительно других классификаторов, метрику recall, это вызвано его реализацией: в случае, когда ха-

рактические характеристики указывают на разные гипотезы, ответ выбирается равновероятно, это и дает данную погрешность. В остальных метриках классификаторы ведут себя схоже: RandomForestClassifier чуть лучше, LogisticRegression чуть хуже, MyClassifier еще чуть хуже.

3.8.3 Результаты

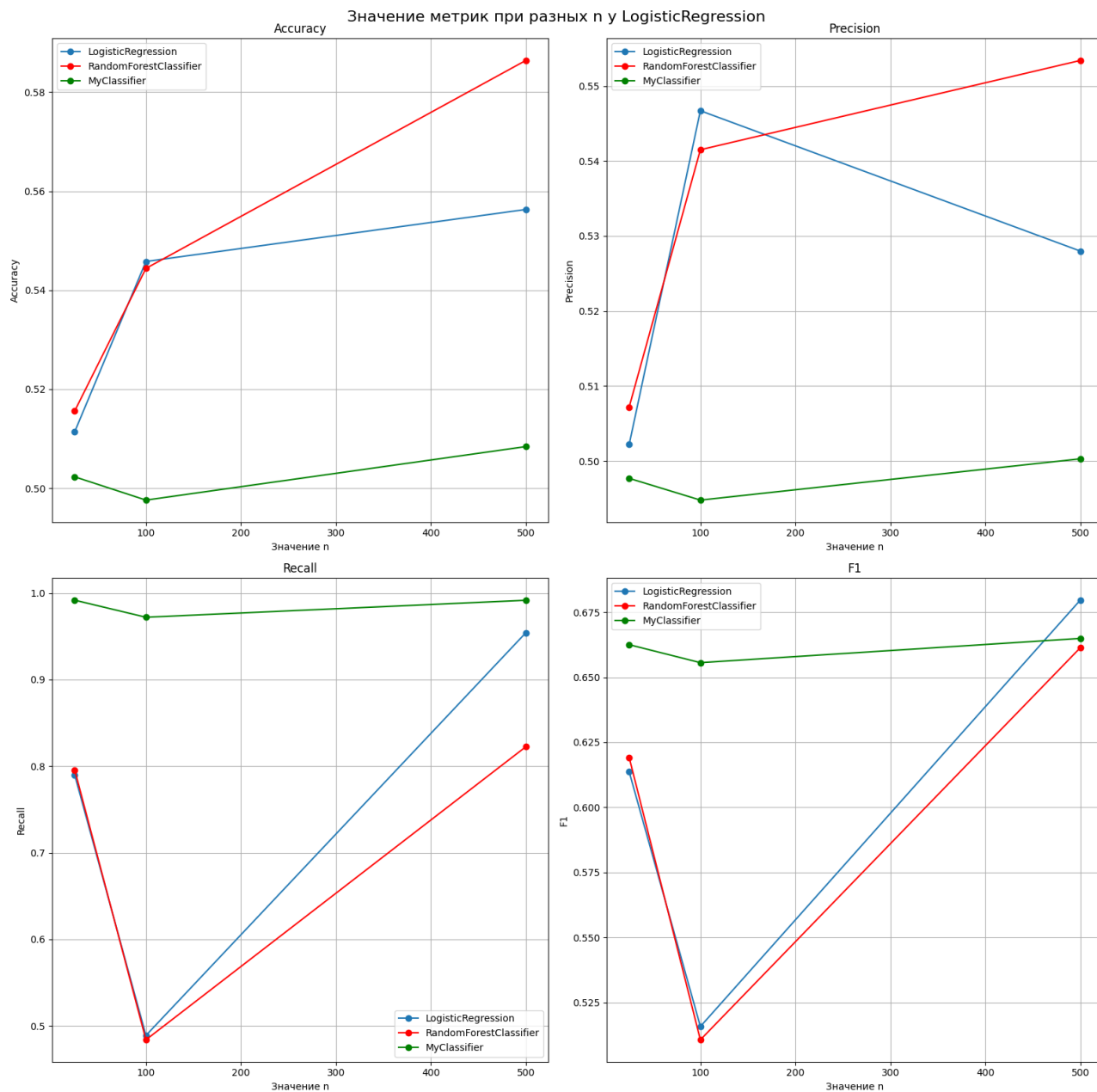


Рис. 16: Значения метрик

Для признаков макс. степень вершины и макс. размер независимого множества показатели метрик уже значительно ниже, не превосходят 0,65 %. Это может быть связано с применением жадного (но не точного) алгоритма для поиска макс. размера независимого множества, используемого для ускорения вычислений. Однако на метрике Recall MyClassifier выдает результат близкий к 1.

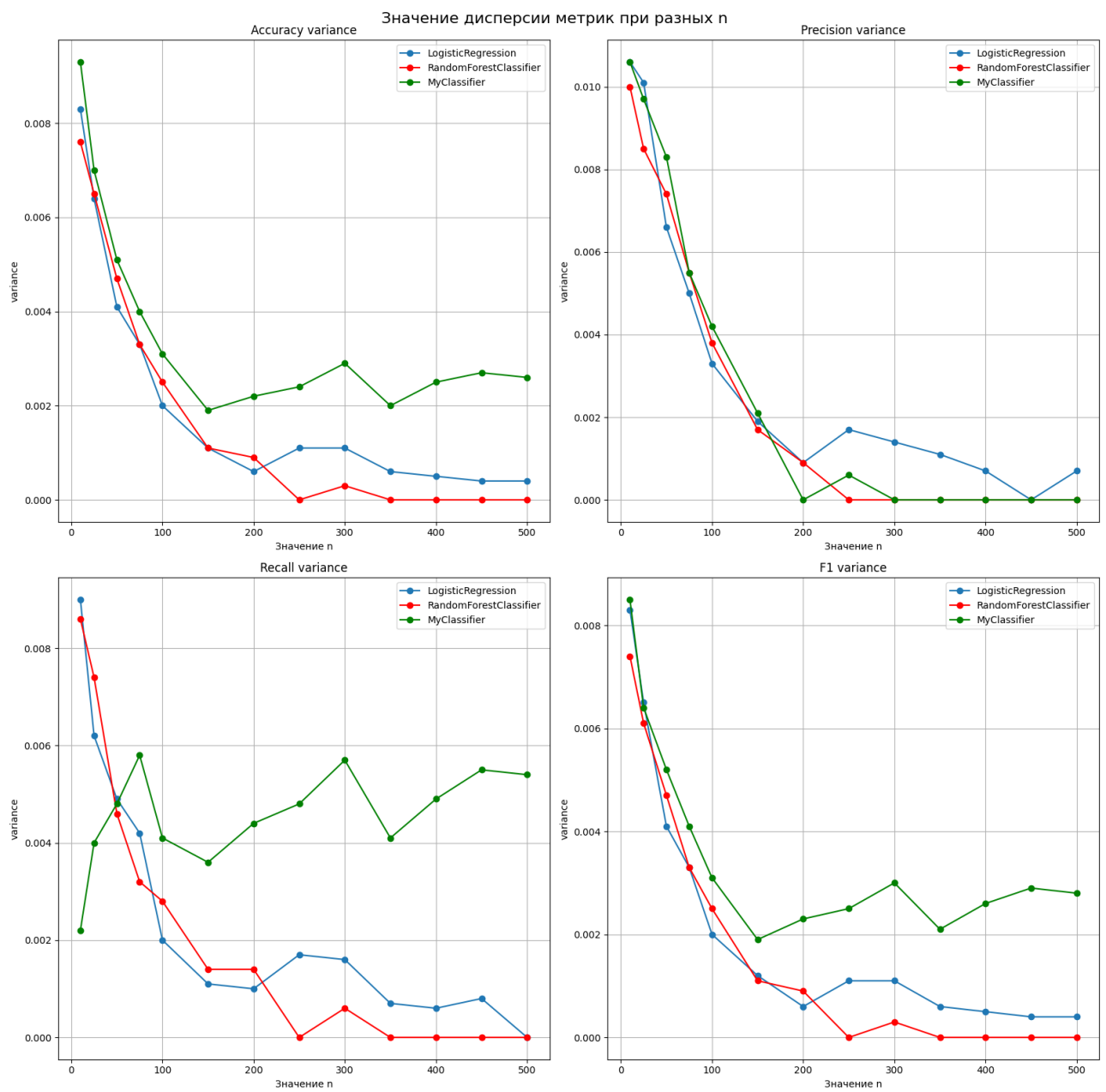


Рис. 17: Значения дисперсий метрик

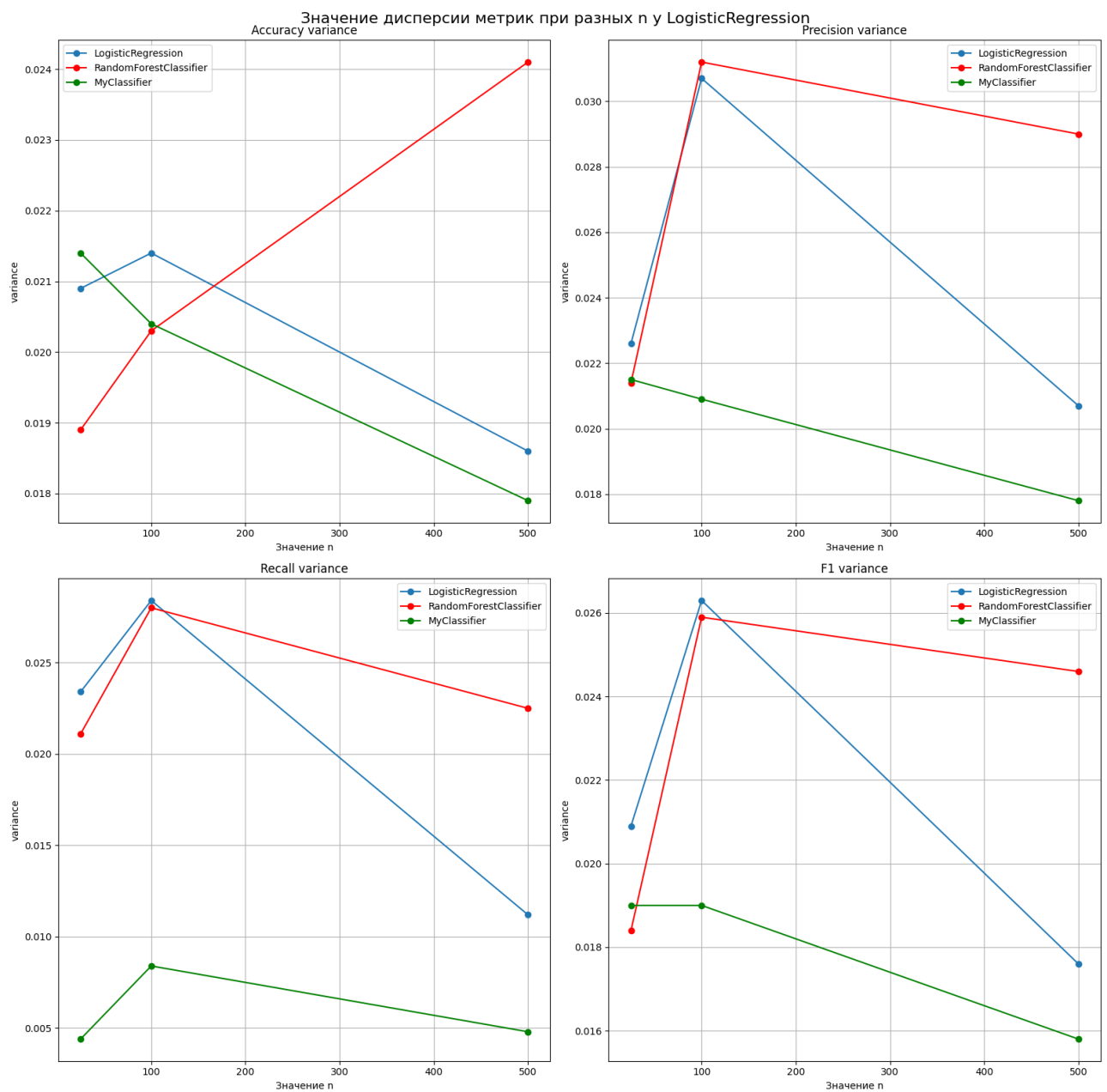


Рис. 18: Значения дисперсий метрик

Посмотрим на дисперсию метрик, для всех метрик и классификаторов верно, что коэффициент вариации находится в диапазоне от 10% до 0% и уменьшается с ростом n (исключение метрика recall и MyClassifier, но причину этого я уже описывал выше). Хороший ли это показатель, зависит от задачи.

3.9 Оценка собственного классификатора

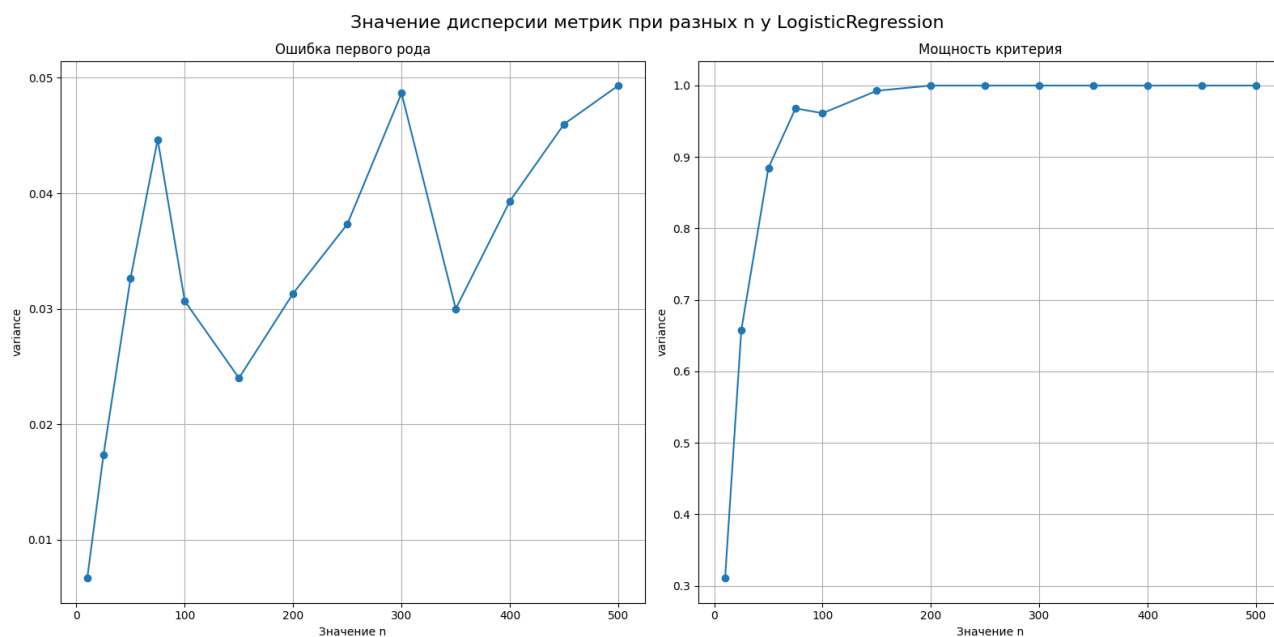


Рис. 19: Ошибка первого рода и мощность критерия

По графику заметно, что при значении ошибки первого рода в диапазоне 0.055 мы получили хорошие результаты мощности критерия при n от 50, мощность критерия в районе 90%. Я считаю, что для двух характеристик это неплохой результат. Возможно, если реализовывать классификатор другим способом, можно добиться лучшего (например, если руками реализовать RandomForestClassifier), но для столь наивной реализации результат неплох.