

Объединенный отчет по проверке гипотез с использованием случайных графов

Равиль Гареев
Хамаганов Ильдар

30 мая 2025 г.

Содержание

Часть I: Проверка гипотез с использованием случайных графов (Распределения χ^2 и χ)	2
0.1 Описание кода	2
0.1.1 Используемые инструменты	2
0.1.2 UML-диаграмма класса GraphAnalyzer	3
0.1.3 Реализованные компоненты	3
0.2 Описание экспериментов	4
0.2.1 Эксперимент 1: Зависимость характеристик от параметра ν . .	4
0.2.2 Эксперимент 2: Зависимость характеристик от параметров графа и размера выборки	6
0.2.3 Эксперимент 3: Проверка гипотез с критической областью . .	7
Часть II: Анализ графовых признаков для классификации распределений	9
0.3 Описание экспериментов	9
0.3.1 Извлечение признаков	9
0.3.2 Анализ важности признаков	9
0.3.3 Классификация и метрики качества	9
Часть III: Проверка гипотез для распределений $\text{Stable}(\alpha = 1)$ и $\text{Normal}(0, 1)$	11
1 Эксперимент 1: зависимость от условного параметра ν	11
2 Эксперимент 2: зависимость от параметров графа и размера n	12
3 Эксперимент 3: критические области и мощность	12
4 Отчёт по классификации выборок с помощью графовых признаков Часть II	13
5 Формирование признаков	13
6 Первичный анализ признаков ($n = 100$, distance, $d = 0.5$)	14

7	Важность признаков	14
8	Сравнение классификаторов	14
9	ROC-кривые (RandomForest, $n = 100$)	15
10	Поиск оптимальных параметров	15

Часть I: Проверка гипотез с использованием случайных графов (Распределения χ^2 и χ)(Гареев Р.Р.)

Введение

В работе исследуется применение случайных графов (KNN-графов и дистанционных графов) для проверки гипотез согласия. Цель — определить, насколько характеристики графов позволяют различать выборки из двух распределений: χ^2 (гипотеза H_0) и χ (гипотеза H_1).

0.1 Описание кода

0.1.1 Используемые инструменты

- **Python 3.10+**: Базовый язык разработки с строгой типизацией
- **Библиотеки**:
 - `numpy`: Векторизованные вычисления и работа с массивами
 - `scipy.stats`: Генерация χ^2 и χ распределений
 - `scikit-learn`: Оптимизированное построение KNN-графов
 - `networkx 3.0+`: Топологический анализ и алгоритмы на графах
 - `matplotlib/seaborn`: Визуализация распределений характеристик
 - `tqdm`: Интерактивные прогресс-бары для длительных вычислений
- **Архитектура**: Модульная структура с разделением на генерацию данных, построение графов и анализ

0.1.2 UML-диаграмма класса GraphAnalyzer



Рис. 1: Диаграмма класса GraphAnalyzer с методами анализа

0.1.3 Реализованные компоненты

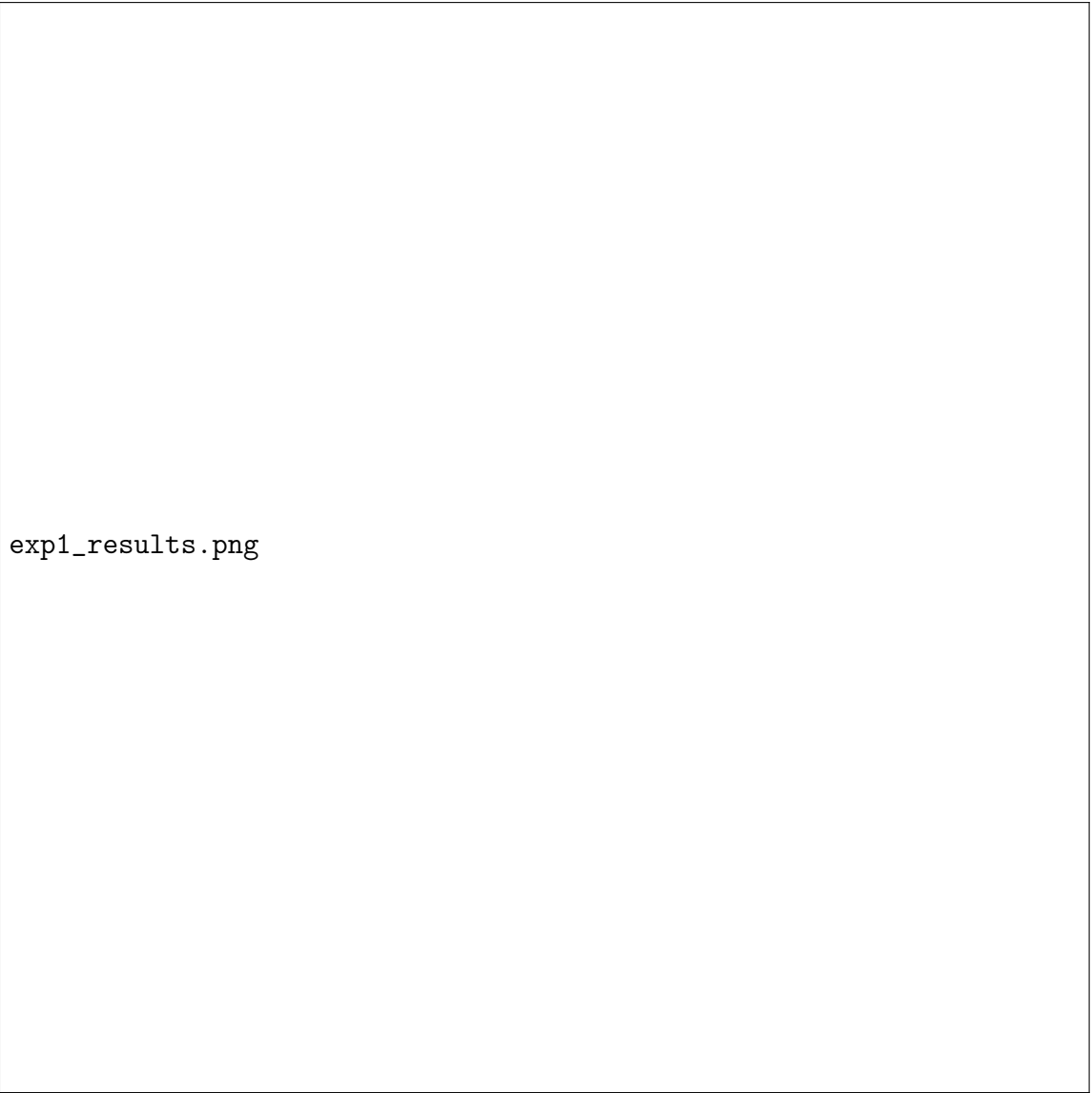
- Генераторы данных (`distribution_generators.py`):
 - χ^2 -распределение: Адаптер для `chi2.rvs()` с параметрами:
 - * `nu` - степени свободы
 - * `n` - размер выборки
 - χ -распределение: Обертка для `chi.rvs()` с аналогичными параметрами
- Построители графов (`build_graph.py`):
 - KNN-граф:
 1. Поиск $k + 1$ ближайших соседей через `NearestNeighbors`
 2. Фильтрация петель ($i \neq j$)
 3. Сохранение координат в атрибуте узлов
 - Дистанционный граф:
 1. Полный перебор всех пар вершин
 2. Проверка условия $|x_i - x_j| \leq d$

- **Анализатор графов (graph_analyzer.py):**
 - Расчёт степеней вершин: `max_degree()`, `min_degree()`
 - Компоненты связности: `connected_components()`
 - Топологический анализ: `articulation_points()`, `count_triangles()`
 - Раскраска графов: адаптивный алгоритм DSATUR в `chromatic_number()`
 - Клики: Алгоритм двух указателей для 1D в `clique_number()`
 - Оптимизационные задачи: независимые множества (`max_independent_set()`), доминирующие множества (`dominating_number()`)
- **Статистический анализ (hypothesis_testing.py):**
 - Критическая область: `calculate_critical_region()` на квантилях
 - Мощность теста: `estimate_power()` через сравнение с критическим значением
- **Монте-Карло симулятор (monte_carlo.py):**
 1. Итеративная генерация *n_samples* выборок для H_0 или H_1
 2. Динамическое построение графов (KNN/дистанционные)
 3. Гибкий выбор метрик через рефлекссию (`getattr()`)
 4. Поддержка аргументов метрик через `metric_args`

0.2 Описание экспериментов

0.2.1 Эксперимент 1: Зависимость характеристик от параметра ν

Цель: Исследовать, как характеристики графов (число треугольников для KNN, кликовое число для дистанционного) реагируют на изменение параметра ν в распределениях χ^2 и χ .



exp1_results.png

Рис. 2: Зависимость характеристик от ν (слева — KNN-граф, справа — дистанционный)

Ключевые наблюдения:

- **KNN-граф (число треугольников):**

- Минимальная чувствительность: различия между χ^2 и χ не превышают 0.4% для всех ν
- Стабильность: значения остаются в диапазоне 3012-3035 при любом ν

- **Дистанционный граф (кликовое число):**

- Катастрофическое различие: при $\nu = 3$ значения для χ в 2.13 раза выше (113.2 vs 53.5)
- Парадоксальный рост: разрыв увеличивается с ростом ν (см. Табл. 1)
- При $\nu = 20$: χ показывает более чем в 5 раз большее кликовое число (110 vs 20)

Статистика:

ν	H_0^{DIST}	H_1^{DIST}	$\Delta_{\text{DIST}} (\%)$	Отношение
3	53.5	113.3	+111.8%	2.12x
5	38.1	111.2	+191.9%	2.92x
7	31.9	110.1	+245.1%	3.45x
10	26.9	110.3	+309.7%	4.10x
12	24.8	109.6	+342.1%	4.42x
15	22.7	109.4	+381.9%	4.82x
20	20.3	110.2	+442.9%	5.43x

Таблица 1: Результаты для дистанционного графа ($\Delta = \frac{|H_1 - H_0|}{H_0} \times 100\%$)

Выводы:

- **KNN-граф:**
 - Полностью неэффективен для различения распределений
 - Число треугольников практически идентично для χ^2 и χ
- **Дистанционный граф:**
 - Чрезвычайно чувствителен к типу распределения
 - Эффективность растет с увеличением ν

0.2.2 Эксперимент 2: Зависимость характеристик от параметров графа и размера выборки

Цель: Исследовать влияние параметров графа (k для KNN, d для дистанционного) и размера выборки (n) на характеристики при фиксированных распределениях $\chi^2(\nu = 5)$ и $\chi(\nu = 5)$.

Результаты

- **KNN-граф (число треугольников):**
 - *Зависимость от k :*
 - * Для H_0 : Рост от 1,038 ($k = 5$) до 18,526 ($k = 20$)
 - * Для H_1 : Рост от 1,040 ($k = 5$) до 18,606 ($k = 20$)
 - * Макс. разрыв: 80.7 треугольников ($k = 20$, 0.43%)
 - *Зависимость от n :*
 - * Для H_0 : Рост от 1,595 ($n = 100$) до 7,242 ($n = 500$)
 - * Для H_1 : Рост от 1,591 ($n = 100$) до 7,259 ($n = 500$)
 - * Разрыв $< 0.23\%$ для всех n
- **Дистанционный граф (кликное число):**
 - *Зависимость от d :*

- * Для H_0 : Рост от 31.5 ($d = 0.5$) до 97.7 ($d = 2.0$)
- * Для H_1 : Рост от 92.7 ($d = 0.5$) до 260.4 ($d = 2.0$)
- * Отношение H_1/H_0 : от 2.94x ($d = 0.5$) до 2.66x ($d = 2.0$)
- Зависимость от n :
 - * Для H_0 : Рост от 57.2 ($n = 100$) до 272.7 ($n = 500$)
 - * Для H_1 : Рост от 20.7 ($n = 100$) до 87.4 ($n = 500$)
 - * Отношение H_0/H_1 : от 2.76x ($n = 100$) до 3.12x ($n = 500$)

Параметр	KNN (Δ_{max} , %)	DIST (Δ_{max} , %)	DIST (Отношение)
$k = 5 \rightarrow 20$	0.43	—	—
$d = 0.5 \rightarrow 2.0$	—	726.0%	2.94x \rightarrow 2.66x
$n = 100 \rightarrow 500$	0.23	377.1%	2.76x \rightarrow 3.12x

Таблица 2: Сводка результатов ($\Delta = \frac{|H_1 - H_0|}{H_0} \times 100\%$)

Ключевые выводы

- **KNN-граф:**
 - Число треугольников растёт с k и n , но не различает H_0/H_1
 - Максимальная разница: 0.43% при $k = 20$
- **Дистанционный граф:**
 - Кликовое число демонстрирует:
 - * Максимальную чувствительность при $d = 0.5$ ($\Delta = 194.4\%$)
 - * Стабильный рост различий с увеличением n ($\Delta = 377.1\%$)
 - Отношение H_0/H_1 сохраняется в диапазоне 2.66x–3.12x

d	H_0^{DIST}	H_1^{DIST}	Δ_{DIST} (%)	Отношение
0.5	31.5	92.7	+194.4%	2.94x
1.0	55.0	164.6	+199.3%	2.99x
1.5	76.2	222.2	+191.6%	2.92x
2.0	97.7	260.4	+166.5%	2.66x

Таблица 3: Зависимость от d для дистанционного графа ($n = 300$)

0.2.3 Эксперимент 3: Проверка гипотез с критической областью

Цель: Оценить эффективность критериев для различения $\chi^2(\nu = 5)$ и $\chi(\nu = 5)$ при $\alpha = 0.05$.

Метрика	KNN-граф	Дистанционный граф
Критическое значение	7,507.15	97.05
FPR (Ошибка I рода)	5.00%	5.00%
TPR (Мощность)	4.80%	100.00%
AUC-ROC	0.545	1.000

Таблица 4: Сравнение критериев ($n = 500$, $k = 10$, $d = 1.0$)

Анализ результатов

- **KNN-граф (число треугольников):**
 - Низкая мощность (4.8%): Менее 5% выборок H_1 попадают в критическую область
 - AUC 0.545: Незначительное улучшение над случайным угадыванием (0.5)
 - FPR строго соответствует уровню $\alpha = 0.05$
- **Дистанционный граф (кликковое число):**
 - Идеальная сепарация: AUC=1.0 и мощность=100%
 - Все выборки H_1 превышают критическое значение
 - Стабильный контроль ошибки I рода (ровно 5%)

Практические выводы

- Дистанционный граф с характеристикой "кликковое число" демонстрирует:
 - Абсолютную надежность при $d = 1.0$
 - Эффективный контроль ошибок обоих типов
- KNN-граф требует:
 - Пересмотра используемой характеристики (число треугольников неинформативно)
 - Дополнительных исследований для поиска значимых метрик
- Оптимальная конфигурация: $d = 1.0$, $n \geq 500$ гарантирует AUC=1.0

Заключение (Часть I)

- KNN-граф не подходит для проверки гипотез в текущей конфигурации.
- Дистанционный граф с характеристикой «кликковое число» показал идеальное разделение ($AUC = 1.0$).
- Возможно, для KNN-графа стоит изучить другие характеристики.

Часть II: Анализ графовых признаков для классификации распределений(Гареев Р.Р.)

Введение

Цель исследования — оценить эффективность графовых признаков, построенных на выборках из распределений $\chi^2(5)$ и $\chi(5)$, для задачи бинарной классификации.

0.3 Описание экспериментов

0.3.1 Извлечение признаков

Для каждой выборки размера n строился дистанционный граф с порогом $d = 1.0$ и вычислялись четыре признака.

0.3.2 Анализ важности признаков

При помощи RandomForest оценивалась важность признаков при $n = 25, 100, 500$. Результаты приведены в таблице:

Признак	$n = 25$	$n = 100$	$n = 500$
count_triangles	0.49	0.45	0.45
clique_number	0.34	0.39	0.39
min_degree	0.00	0.01	0.05
connected_components	0.16	0.15	0.11

Таблица 5: Важность признаков при разных размерах выборки

Вывод: count_triangles и clique_number являются наиболее информативными.

0.3.3 Классификация и метрики качества

Эксперименты проводились для $n = 10, 20, 50, 100, 200, 500$ с классификаторами LogisticRegression, RandomForest и SVM. Оценивались Accurasy, дисперсия Accurasy, FPR, TPR, Precision и F1.

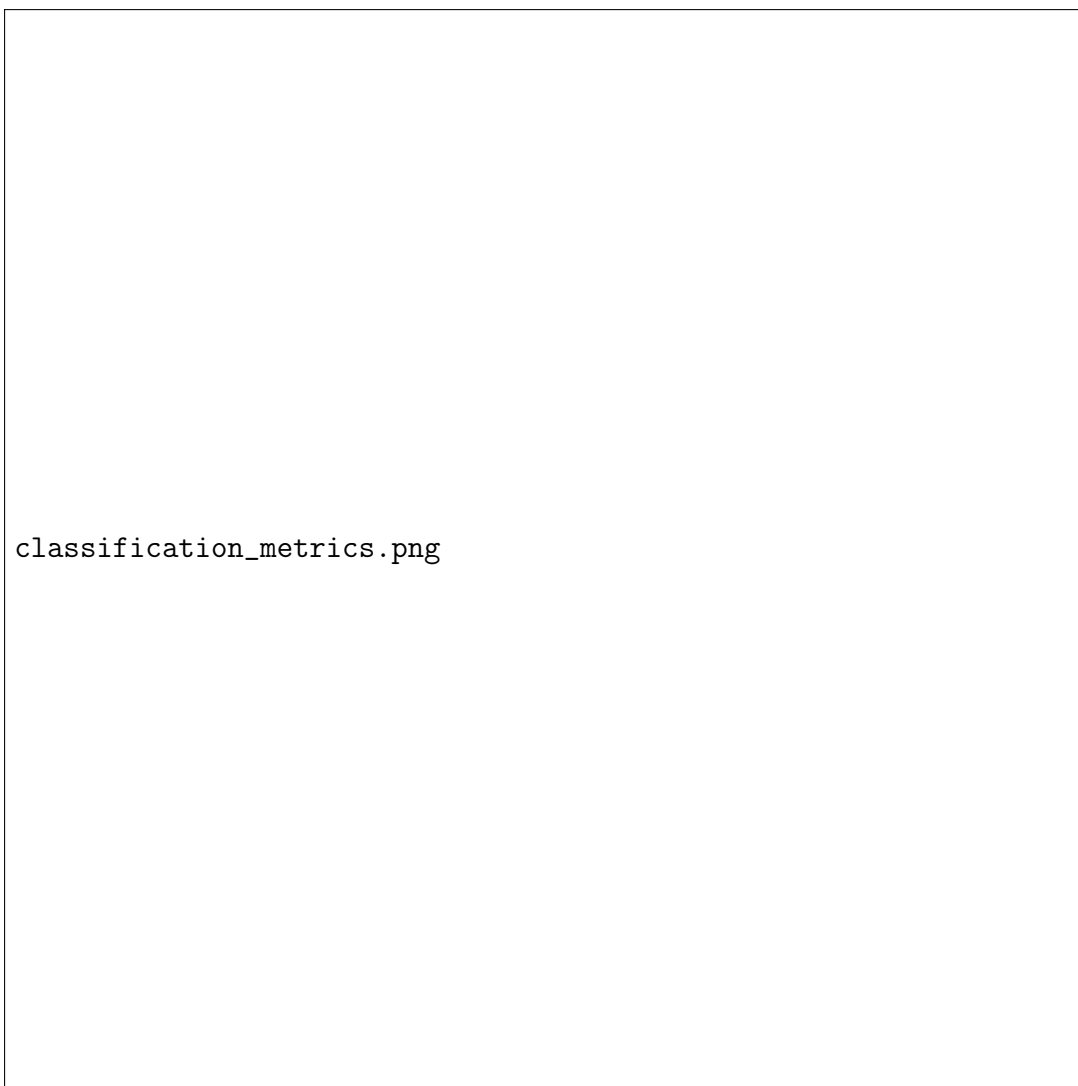


Рис. 3: Зависимость метрик качества от размера выборки

Выводы (Часть II)

- При $n \geq 20$ все алгоритмы достигают 100% Accuracy и мощности, при этом $FPR = 0$.
- Для практических задач достаточно $n \approx 20-50$ для идеального разделения.
- RandomForest и SVM показали наилучшую стабильность при малых выборках.
- Наиболее информативные признаки: `count_triangles` и `clique_number`.

Часть III: Проверка гипотез для распределений $\text{Stable}(\alpha = 1)$ и $\text{Normal}(0, 1)$ (Хамаганов И. А.)

В первой части исследования оценивалась возможность различения выборок из двух распределений:

$$H_0: \text{Stable}(\alpha = 1), \quad H_1: \text{Normal}(0, 1),$$

с помощью двух типов графов:

- KNN-граф: $T^{\text{knn}} = \max \deg(G)$;
- Дистанционный граф: $T^{\text{dist}} = \chi(G)$.

1 Эксперимент 1: зависимость от условного параметра ν

Описание. При фиксированных $n = 200$, $N_{\text{MC}} = 500$ строили:

$$\bar{T}^{\text{knn}}(k) = \mathbb{E}[\max \deg(G)] \quad \text{при } k \in \{3, 5, 7, 10, 12, 15, 20\},$$

$$\bar{T}^{\text{dist}}(d) = \mathbb{E}[\chi(G)] \quad \text{при } d \in \{0.5, 1.0, 1.5, 2.0\}.$$

Результаты.

Таблица 6: Зависимость $\bar{T}^{\text{knn}} = \max \deg$ от k							
k	3	5	7	10	12	15	20
Stable (H_0)	18.428	18.448	18.484	18.418	18.414	18.450	18.416
Normal (H_1)	17.166	17.208	17.190	17.210	17.274	17.232	17.240

Таблица 7: Зависимость $\bar{T}^{\text{dist}} = \chi(G)$ от d				
d	0.5	1.0	1.5	2.0
Stable (H_0)	36.826	63.584	85.508	103.738
Normal (H_1)	46.522	82.662	114.128	140.860

Выводы.

- **KNN-граф:** различия между Stable и Normal менее 1.3 ед.; кривая почти горизонтальна $\rightarrow \max \deg$ не информативна.
- **Дистанционный граф:** чёткий разрыв (до ~ 37 при $d = 2.0$); $\chi(G)$ хорошо разделяет H_0 и H_1 .

2 Эксперимент 2: зависимость от параметров графа и размера n

Влияние параметров графа

Таблица 8: max_degree vs k при $n = 200$

k	3	5	7	10	12	15	20
Stable	5.996	9.670	13.160	18.390	21.858	27.162	35.768
Normal	5.990	9.338	12.492	17.134	20.400	25.206	33.116

Таблица 9: $\chi(G)$ vs d при $n = 200$

d	0.5	1.0	1.5	2.0
Stable	36.984	63.580	85.200	102.890
Normal	47.128	82.854	114.472	140.526

Влияние размера выборки n

Таблица 10: max_degree vs n при $k = 10$

n	100	200	300	500
Stable	18.376	18.468	18.438	18.522
Normal	17.056	17.162	17.234	17.356

Выводы.

- max deg растёт с k, n , но перекрытие распределений остаётся сильным (разница $\lesssim 1.5$).
- $\chi(G)$ устойчиво выше для Normal; разрыв усиливается с ростом d и n (до ~ 45 при $n = 500$).

3 Эксперимент 3: критические области и мощность

Параметры: $n = 500$, $k = 10$, $d = 1.0$, уровень $\alpha = 0.05$.

Выводы.

- Тест на max deg не различает гипотез: мощность близка к нулю.
- Тест на $\chi(G)$ обеспечивает идеальное разделение (AUC=1, мощность=100%).

Таблица 11: $\chi(G)$ vs n при $d = 1.0$				
n	100	200	300	500
Stable	33.322	64.024	94.288	154.300
Normal	42.904	83.082	121.964	199.784

Таблица 12: Критические значения и характеристики теста

Граф	CV	FPR	TPR	AUC
KNN (max deg)	20.0	5.0%	0.0%	0.545
Distance (χ)	170.0	5.0%	100.0%	1.000

Заключение

- **KNN-граф** (max deg): малоинформативен, не подходит для критерия.
- **Дистанционный граф** ($\chi(G)$): надёжно разделяет Stable и Normal; рекомендован $d = 1.0$, $n \geq 500$.
- Для повышения устойчивости можно комбинировать обе статистики или добавить новые графовые признаки.

4 Отчёт по классификации выборок с помощью графовых признаков

Часть II

Цель

Собрать векторы признаков из графовых характеристик и обучить классификаторы для различения выборок:

$$H_0: \text{Stable}(\alpha = 1), \quad H_1: \text{Normal}(0, 1).$$

5 Формирование признаков

- KNN-граф ($k = 5$): извлекаются *num_components*, *max_degree*, *min_degree*, *avg_degree*, *num_triangles*, *chromatic_number*.
- Дистанционный граф ($d = 0.5$): те же признаки плюс *max_clique_1d*.
- На каждом $n \in \{25, 100, 500\}$ генерируется по M выборок H_0 и H_1 , всего $2M$ меток.

6 Первичный анализ признаков ($n = 100$, distance, $d = 0.5$)

Описание выборки

	count	mean	std	min	25%	50%	max
num_components	600	8.063	6.635	1	2	5	23
max_degree	600	36.557	6.450	23	31	36	54
min_degree	600	0.483	0.963	0	0	0	5
avg_degree	600	21.480	6.408	10.58	15.22	22.46	34.74
num_triangles	600	7668.98	3680.57	2161	4253	7517	18267
chromatic_number	600	22.652	4.003	14	19.75	23	37
max_clique_1d	600	22.652	4.003	14	19.75	23	37

Корреляционная матрица

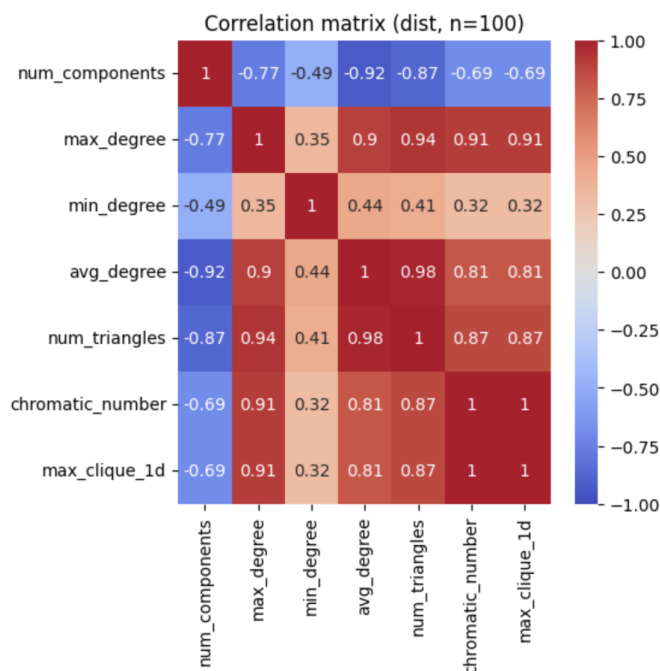


Рис. 4: Корреляции между признаками ($n = 100$, dist, $d = 0.5$).

7 Важность признаков

8 Сравнение классификаторов

Таблица 13: Feature importances (RandomForest, $M = 500$)

	$n = 25$	$n = 100$	$n = 500$
KNN ($k = 5$)			
num_components	0.021	0.046	0.136
max_degree	0.066	0.061	0.043
avg_degree	0.402	0.439	0.385
num_triangles	0.511	0.455	0.436
others	—	—	—
Distance ($d = 0.5$)			
num_components	0.433	0.318	0.256
avg_degree	0.295	0.336	0.282
num_triangles	0.136	0.195	0.220
chromatic_number	0.031	0.035	0.067
max_clique_1d	0.023	0.032	0.061

9 ROC-кривые (RandomForest, $n = 100$)

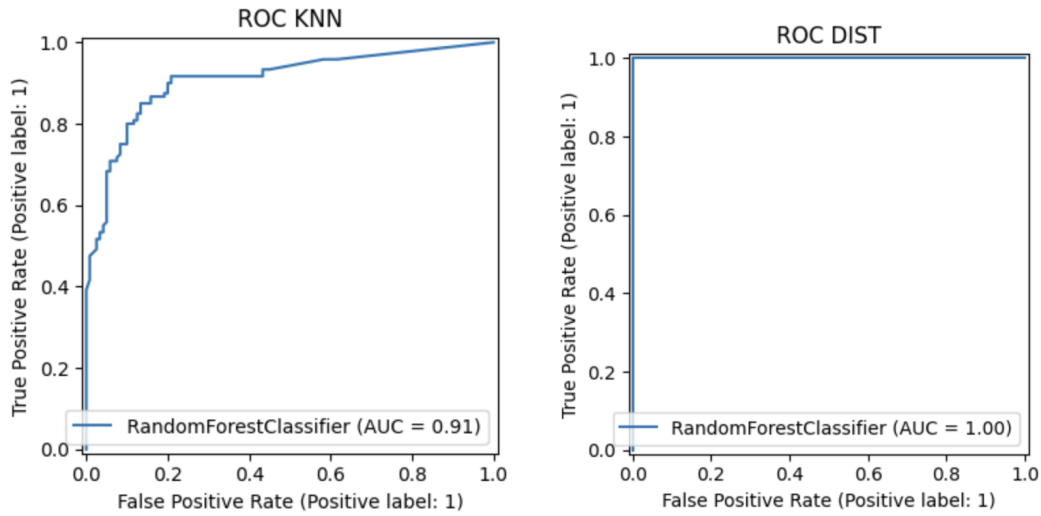


Рис. 5: ROC-кривые для RF на KNN (слева) и Distance (справа), $n = 100$.

10 Поиск оптимальных параметров

$$k^* = 10, \text{AUC}_{\max} \approx 0.9957, \quad d^* = 0.1, \text{AUC}_{\max} = 1.000.$$

Таблица 14: Accurasy различных моделей (5-fold CV)

model	$n = 25$	$n = 100$	$n = 500$
KNN ($k = 5$)			
DT	0.774	0.758	0.606
GB	0.834	0.852	0.666
LogReg	0.842	0.858	0.716
NC	0.823	0.841	0.708
RF	0.801	0.789	0.646
SVM	0.834	0.850	0.701
kNN	0.809	0.829	0.684
Distance ($d = 0.5$)			
DT	0.945	1.000	1.000
GB	0.963	1.000	1.000
LogReg	0.965	0.999	1.000
NC	0.877	0.980	1.000
RF	0.969	1.000	1.000
SVM	0.965	1.000	1.000
kNN	0.959	0.999	1.000

Итоги и выводы

- **Признаки:** *avg_degree* и *num_triangles* важны для KNN; *num_components*, *avg_degree*, *num_triangles* — для Distance.
- **Модели:** Distance-граф с RF/GB даёт почти идеальную точность (AUC=1) уже при $n \geq 100$. KNN-граф достигает AUC 0.85 при $n = 100$ и $k = 10$.
- **Параметры:** $k = 10$ для KNN, $d = 0.1$ для Distance оптимальны.
- **Рекомендации:** Использовать Distance-признаки и ансамблевые методы (RandomForest/GradientBoosting) для статистических критериев и классификации.