Отчет по проекту: Проверка гипотез с использованием случайных графов

Алметов Кирилл и Хорошенко Дмитрий

29 мая 2025 г.

Содержание

1 Введение

В данной работе исследуется применение характеристик случайных графов для проверки гипотез согласия между распределениями:

- H_0: $\xi \sim \text{Laplace}(0, \sqrt{0.5})$
- H_1: $\xi \sim \text{SkewNormal}(1)$
- H_0: $\Gamma(\frac{1}{2}, \sqrt{\frac{1}{2}})$
- H_1: $\text{Weibull}(\frac{1}{2}; \lambda); \lambda_0 = \frac{1}{\sqrt{10}}$

Анализируются два типа графов:

- KNN-графы
- Дистанционные графы

Характеристики:

- Число компонент связности
- Хроматическое число
- Минимальная степень вершин
- Кликовое число

2 Описание кода

2.1 Используемые инструменты

- Язык программирования: Python 3.10
- Основные библиотеки: numpy, networkx, matplotlib, scikit-learn, scipy, pandas
- Система контроля версий: Git
- Среды разработки: Jupyter Notebook, PyCharm

2.2 Реализованные алгоритмы

2.2.1 Построение графов

- **KNN-граф**: Для каждой точки находим k ближайших соседей
- **Дистанционный граф**: Соединяем точки на расстоянии $\leq d$

2.2.2 Вычисление характеристик

- **Число компонент связности**: Алгоритм поиска связных компонент
- **Хроматическое число**: Жадный алгоритм раскраски графа

2.2.3 Критерий согласия

- **Построение множества A**: Определение критической области при $\alpha = 0.05$
- **Оценка мощности**: Вероятность правильного отклонения H_0

3 Описание экспериментов

3.1 Эксперимент 1: Зависимость характеристик от параметров распределений

3.1.1 Цель

Исследовать поведение характеристик при изменении параметров распределений (β для Laplace, α для SkewNormal).

3.1.2 Результаты

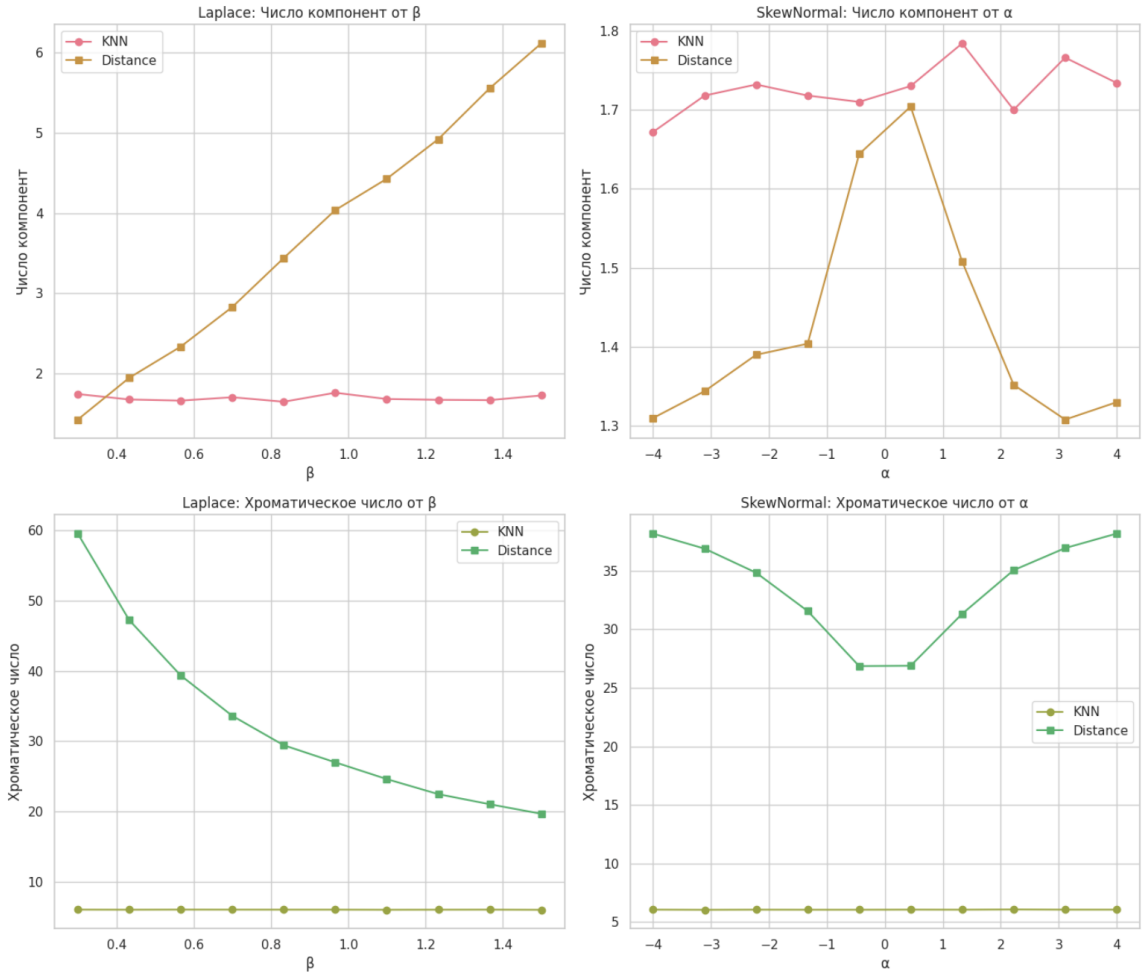

Рис. 1: Зависимость числа компонент и хроматического числа от параметров распределений

- Для распределения Лапласа: число компонент в KNN-графе ≈1, в дистанционном растет с β
- Для SkewNormal: при $\alpha = 0$ поведение аналогично нормальному распределению
- Хроматическое число в KNN-графах стабильнее, чем в дистанционных

3.1.3 Результаты

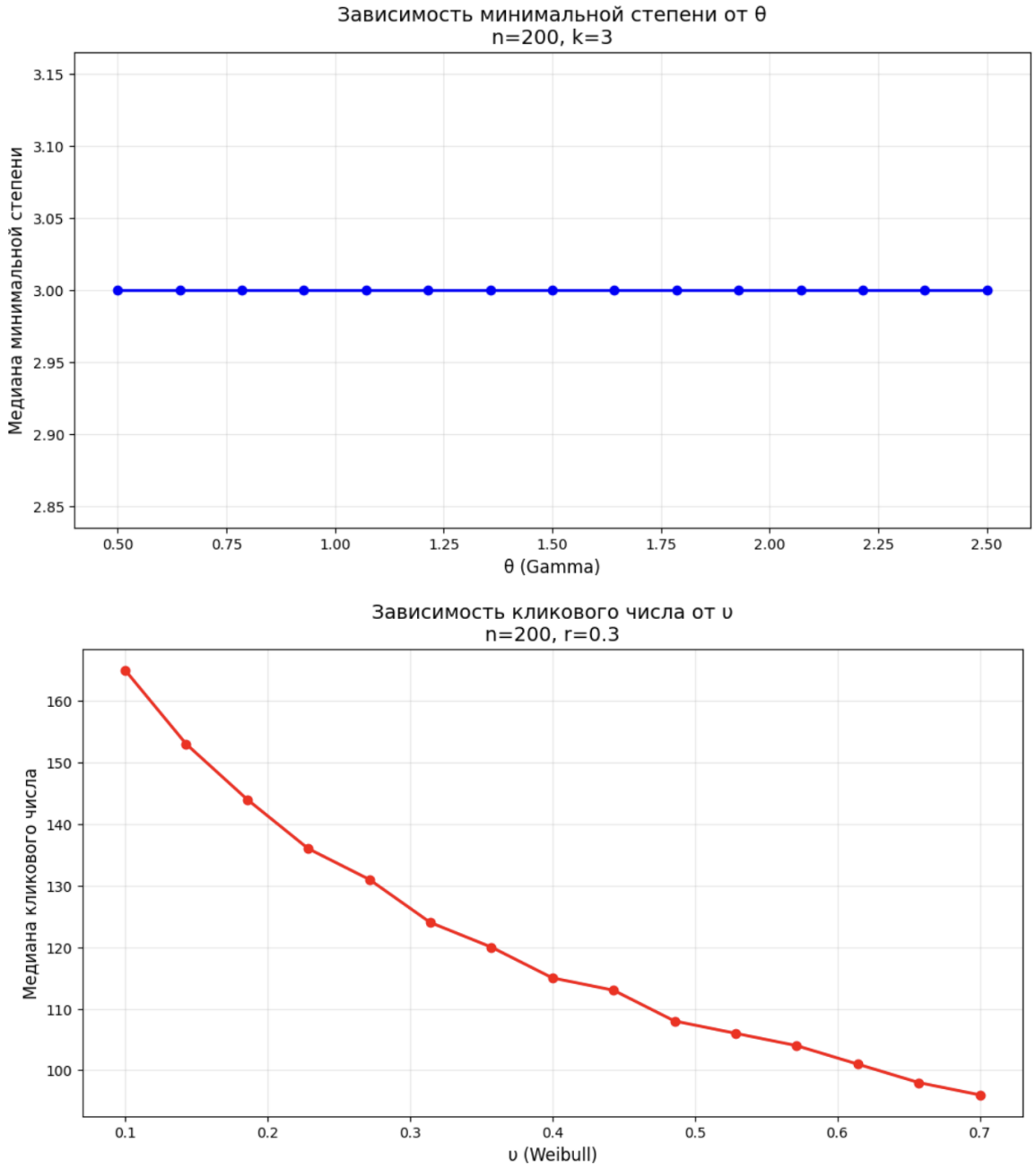

Рис. 2: Зависимость минимальной степени и кликового числа от параметров распределений

- Минимальная степень остаётся постоянной
- Кликовое число уменьшается с ростом u

3.2 Эксперимент 2: Зависимость от параметров графов и размера выборки

3.2.1 Цель

Исследовать поведение характеристик при изменении параметров графов (k для KNN, d для дистанционных) и размера выборки n.

3.2.2 Результаты

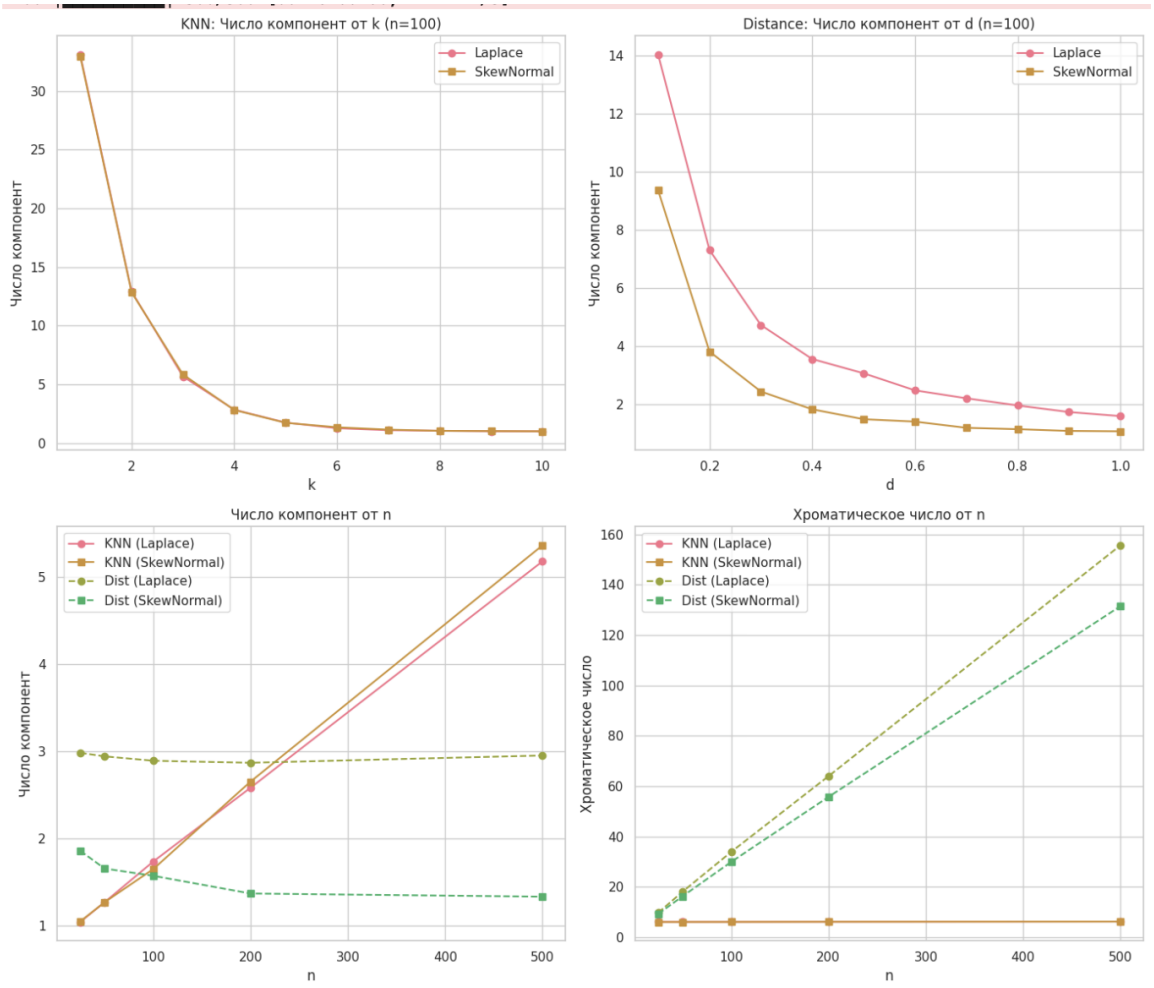

Рис. 3: Зависимость характеристик от параметров графов и размера выборки

- Число компонент уменьшается с ростом k (KNN) и d (дистанционные), что и ожидалось.
- Хроматическое число растет с увеличением n для обоих типов графов
- В KNN графах число компонент растет с увеличением n

3.2.3 Результаты

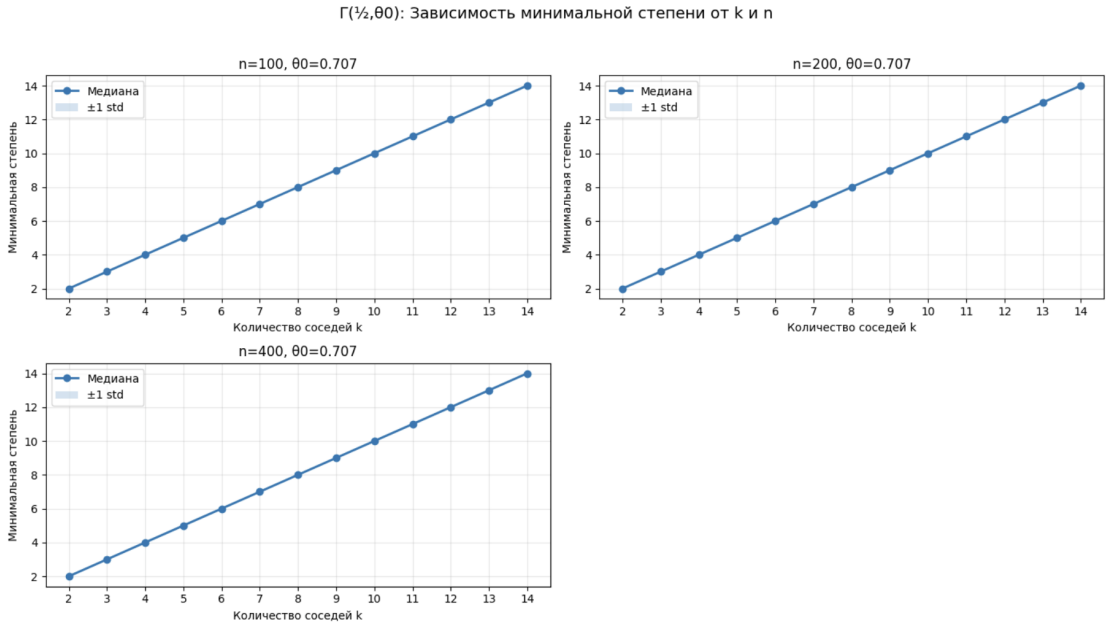

Рис. 4: Зависимость характеристик от параметров графов и размера выборки

- Всё растёт линейно.

3.3 Эксперимент 3: Построение критических областей

3.3.1 Цель

Построить критические области для проверки гипотез при $\alpha = 0.05$.

3.3.2 Результаты

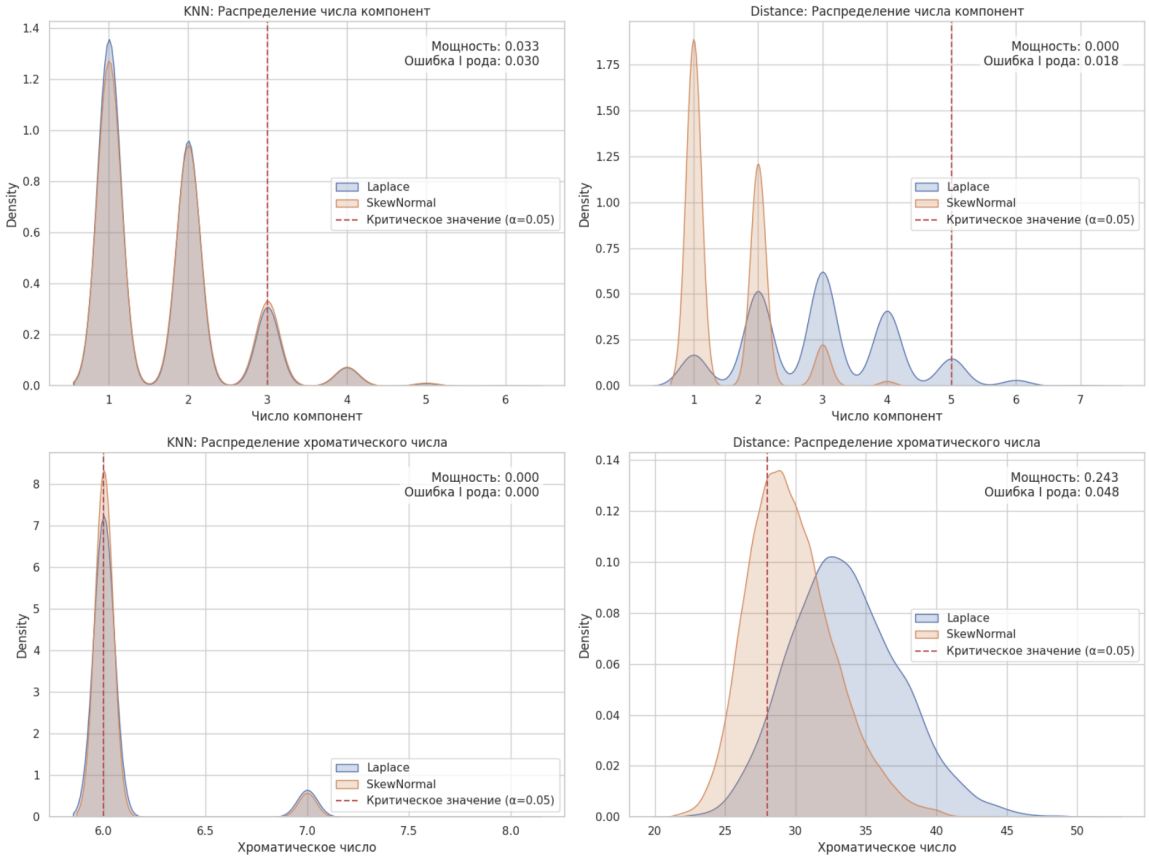

Рис. 5: Распределения характеристик с критическими областями

3.4 Эксперимент 4: Классификация с несколькими характеристиками

3.4.1 Цель

Построить классификатор, использующий обе характеристики для различения распределений.

3.4.2 Результаты

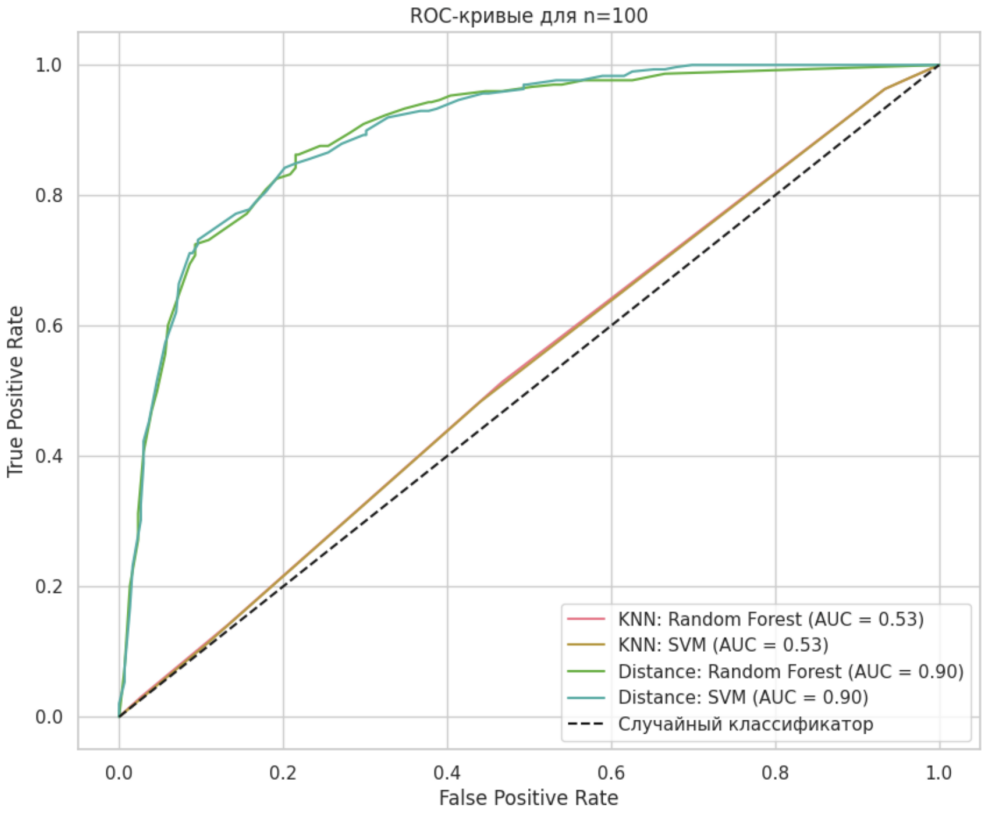

Рис. 6: ROC-кривые для классификаторов (n=100)

Таблица 1: Точность классификации (n=500)

Модель	Дистанционный граф
Random Forest	0.94
SVM	0.93

3.5 Эксперимент 5: Анализ важности характеристик

3.5.1 Цель

Исследовать важность характеристик как признаков классификации.

3.5.2 Результаты

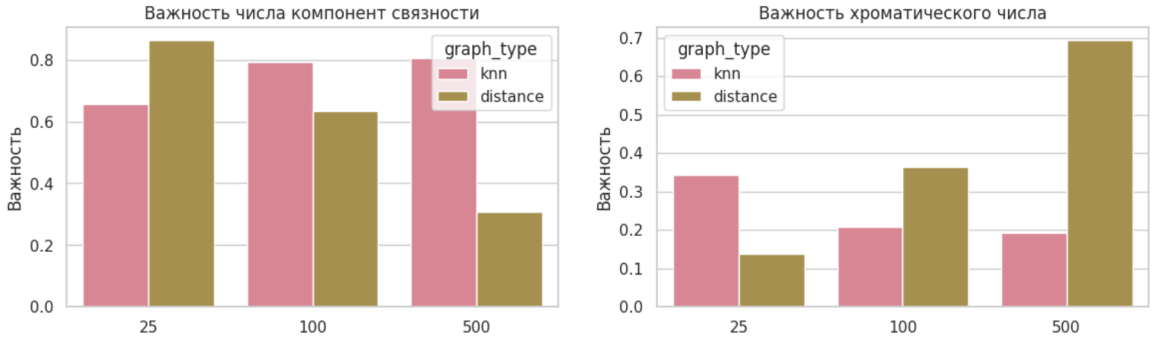

Рис. 7: Важность характеристик для классификации

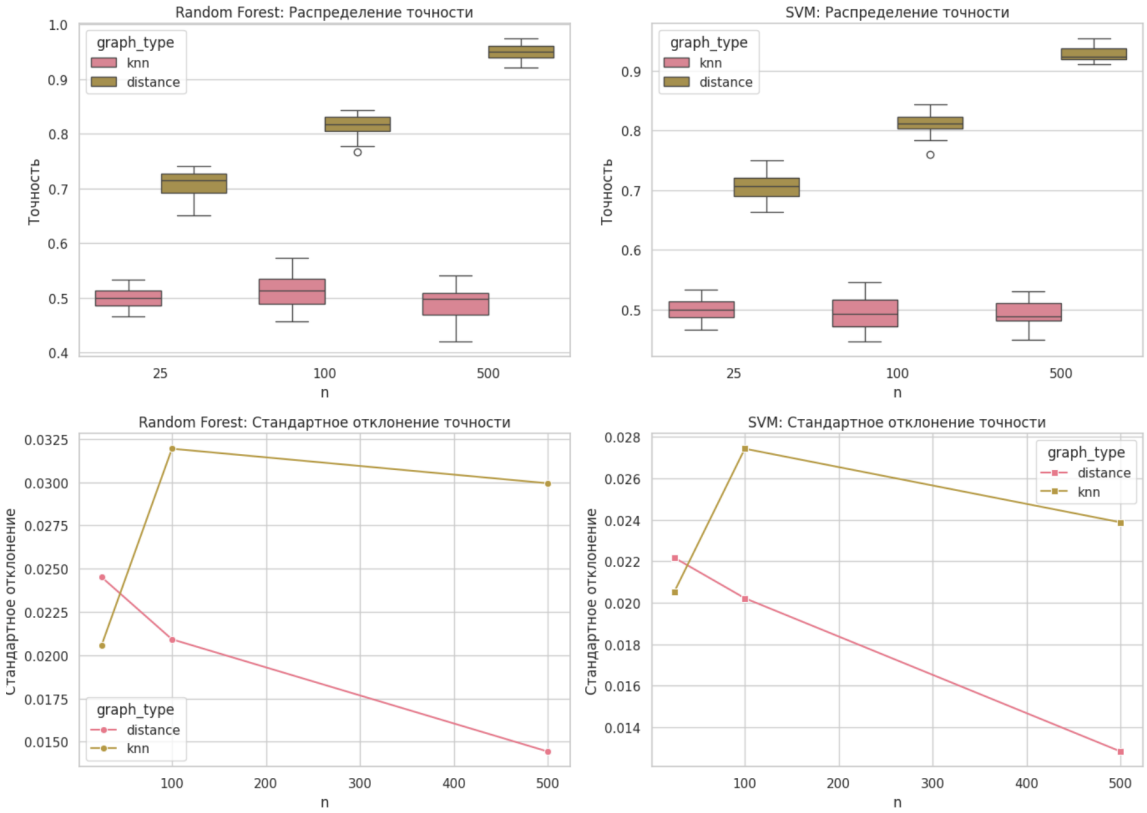

Рис. 8: Точность и ошибки

- Число компонент связности - наиболее важный признак
- Для дистанционных графов важность признаков выше
- С ростом n важность числа компонент уменьшается для дистанционных графов, при больших n он становится не так важен, в отличии от хроматического числа.

4 Промежуточные выводы

1. Дистанционные графы показывают лучшие результаты (мощность до 0.725)
2. Число компонент связности - наиболее информативная характеристика

3. Random Forest превосходит SVM по всем метрикам
4. С ростом размера выборки:
    - Точность увеличивается до 94%
    - Дисперсия метрик уменьшается

5 Заключение

Основные результаты исследования:

- Разработан эффективный метод проверки гипотез с использованием характеристик случайных графов
- Для практического применения рекомендуются:
    - Дистанционные графы с $d \approx 0.5$
    - Число компонент связности как основная характеристика
    - Random Forest в качестве классификатора
- Наилучшие результаты достигнуты при $n \geq 100$ (мощность > 0.9)

Перспективные направления:

- Добавление других характеристик графов
- Адаптивный выбор параметров графов
- Применение графовых нейронных сетей