**РГР**

**Анализ данных в среде R**

**Категория 1**

**Темы:**

1. Многомерный регрессионный анализ
2. Нелинейный регрессионный анализ
3. Пуассоновская регрессия
4. Полиномиальная регрессия
5. Кусочно-полиномиальная регрессия
6. Гребневая регрессия, метод лассо
7. Пошаговый регрессионный анализ
8. Анализ мощности
9. Перестановочные тесты
10. Бутстреп-анализ
11. Анализ главных компонент
12. Факторный анализ
13. Анализ пропущенных значений
14. Визуализация данных с помощью анимированных графиков
15. Построение ансамбля моделей классификации (случайные леса)
16. Методы бустинга регрессионных деревьев
17. Методы бэггинга регрессионных деревьев
18. Построение ансамбля моделей кластеризации
19. Построение ансамбля моделей прогнозирования временных рядов
20. Методы сравнения точности моделей
21. Нейросетевые методы для решения задач прогнозирования
22. Нейросетевые методы для решения задач классификации
23. Нейросетевые методы для решения задач кластеризации
24. Обработка пропущенных наблюдений во временных рядах
25. Выявление структурных изменений во временных рядах
26. Байесовские структурные модели временных рядов
27. Кластеризация временных рядов
28. Сравнение моделей. Оценка эффективности моделей. Lift и Profit кривые.
29. Сравнение моделей. Оценка эффективности моделей. ROC-анализ.
30. Деревья регрессии с многомерным откликом
31. Бинарные деревья решений (алгоритм ID3)
32. Анализ последовательностей событий
33. Критерии согласия, исследование законов распределения данных
34. Анализ сетей (визуализация)
35. Анализ сетей (модели случайных сетей)
36. Анализ сетей (статистические модели сетей)
37. Анализ сетей (динамические модели сетей)
38. Исследование сетевого трафика
39. Методы анализа статистики событий информационной безопасности
40. Методы обезличивания данных
41. Методы оценки риска раскрытия информации после обезличивания данных
42. Методы оценки информационных потерь после обезличивания данных
43. Методы обогащения данных
44. Исследование методов предобработки данных для улучшения точности моделей машинного обучения.
45. Применение неоднородных ансамблей для повышения точности решения задачи классификации (стекинг моделей).
46. Применение неоднородных ансамблей для повышения точности решения задачи регрессии (стекинг моделей).
47. Применение неоднородных ансамблей для повышения точности решения задачи кластеризации (стекинг моделей).
48. Исследование методов обработки несбалансированных классов в задачах классификации.
49. Оптимизация гиперпараметров модели машинного обучения с использованием алгоритмов оптимизации.
50. Анализ и сравнение алгоритмов детекции аномалий для выявления нестандартных паттернов в данных.
51. Исследование методов сжатия данных для уменьшения размера наборов данных при сохранении важной информации (методы понижения размерности, компрессия данных, семплирование, агрегация, …).
52. Оценка и сравнение методов декомпозиции данных с целью выделения факторов и компонент (PCA, факторный анализ, ICA, спектральное разложение, …).
53. Применение ансамблевых методов (бэггинг и/или бустинг, необходимо рассмотреть минимум 2 алгоритма) для улучшения качества классификации.
54. Применение ансамблевых методов (бэггинг и/или бустинг, необходимо рассмотреть минимум 2 алгоритма) для улучшения качества регрессии.
55. Исследование методов генерации синтетических данных для обучения моделей.
56. Исследование методов анализа и визуализации больших объемов данных с использованием графических процессоров (GPU).
57. Оценка и сравнение методов обнаружения дубликатов в больших наборах данных.
58. Исследование методов обработки данных для улучшения устойчивости моделей к шуму и выбросам.
59. Сравнение и анализ различных методов ансамблирования в задачах классификации (необходимо рассмотреть минимум 2 алгоритма).
60. Оценка и сравнение методов стекинга и блендинга моделей в ансамблях. Решаться может задача классификации или регрессии.
61. Исследование и применение методов интеллектуального управления данными.
62. Исследование влияния методов регуляризации (Dropout,   
    L1/L2-регуляризация, Elastic Net) на качество моделей.
63. Применение методов интерпретации моделей (SHAP, LIME) для объяснения результатов машинного обучения.
64. Применение методов transfer learning (трансферное обучение) для задач классификации или регрессии.
65. Визуализация многомерных данных методами t-SNE и UMAP.
66. Визуализация сетей и графовых структур (igraph, ggraph).
67. Методы кластеризации (k-means, DBSCAN, агломеративная кластеризация) и их сравнение.
68. Исследование методов регуляризованной регрессии для высокоразмерных данных.
69. Сравнение линейных и нелинейных моделей регрессии для анализа сложных зависимостей.

**В РГР должны быть представлены следующие разделы:**

1. Теоретическая часть. Описание метода: (задача, для решения которой разработан метод, основные формулы, алгоритм применения метода, требования к структуре и типу исходных данных, ограничения (условия использования) метода, достоинства, недостатки, …).
2. Практическая часть.

- постановка задачи, для решения которой будет использоваться рассматриваемый метод;

- исходные данные (данные должны быть реальными, взятыми из репозиториев данных по машинному обучению, например, ресурс Kaggle, репозиторий UCI (UCI Machine Learning Repository), либо собственные данные)

- описание хода решения задачи в R (привести код на R с комментариями и краткими пояснениями);

- выводы по результатам решения задачи в терминах предметной области (интерпретация результатов).

3. Выводы по работе в произвольной форме (сравнение R с другими средствами анализа данных, оценка рассматриваемого метода анализа данных, анализ результатов и т.п.).

*Допускается выбор собственной темы по согласованию с преподавателем.*

Категория 2

*Тематика*: «Разработка интерактивного приложения с использованием пакета Shiny на языке R». Набор реализованных методов студент определяет самостоятельно и согласовывает с преподавателем до начала работы.

Разрабатываемое приложение должно включать в себя следующие функциональные компоненты:

* *Загрузка и отображение исходных данных*. Приложение должно обеспечивать возможность загрузки и предварительного просмотра реальных наборов данных (минимум двух–трех).
* *Визуализация данных.* Необходимо реализовать средства интерактивной визуализации, включающие динамические графики и интерактивные таблицы, обеспечивающие наглядное представление распределений, зависимостей и результатов вычислений.
* *Выбор переменных для анализа.* Пользователь должен иметь возможность через графический интерфейс задавать целевую переменную, а также выбирать подмножество признаков, которые будут использоваться при анализе.
* *Выбор методов анализа и настройка параметров.* В приложении требуется реализовать не менее двух методов машинного обучения. Для каждого метода необходимо предусмотреть возможность настройки основных гиперпараметров непосредственно через элементы управления графического интерфейса.
* *Оценка качества моделей.* Приложение должно поддерживать вычисление и отображение не менее четырех метрик качества (например, Accuracy, Precision, Recall, F1-мера, AUC, RMSE — в зависимости от задачи). Результаты оценки следует представлять в табличной и графической форме.

**В РГР должны быть представлены следующие разделы:**

1. Теоретическая часть:

* Краткое описание фреймворка Shiny (или какого-либо другого пакета для создания интерактивных приложений для языка R): основные инструменты и возможности (реактивное программирование, структура приложения Shiny, алгоритм создания графического интерфейса пользователя (GUI) в Shiny), используемые (в РГР) элементы управления и т.д.
* Описание метода: задача, для решения которой разработан метод, основные формулы, алгоритм применения метода, требования к структуре и типу исходных данных, ограничения (условия использования) метода, сравнение методов, достоинства, недостатки и т.д.

1. Практическая часть:

* постановка задачи, для решения которой будет использоваться рассматриваемый метод;
* исходные данные (данные должны быть реальными, взятыми из репозиториев данных по машинному обучению, например, ресурс Kaggle, репозиторий UCI (UCI Machine Learning Repository), либо собственные данные);
* описание хода решения задачи в R с использованием фреймворка Shiny (привести код на R с комментариями и краткими пояснениями);
* выводы по результатам решения задачи в терминах предметной области (интерпретация результатов).

1. Выводы по работе в произвольной форме (сравнение технологии R Shiny с другими средствами и технологиями создания приложений анализа данных, оценка рассматриваемого метода анализа данных, анализ результатов и т.п.).

**Литература**

1. Кабаков Р.И. R в действии. Анализ и визуализация данных в программе R. – М.: ДМК Пресс, 2014. – 588 с.
2. Кабаков Р. И. R в действии / И. Р. Кабаков ; пер. с англ. А. Н. Киселева. – Москва : ДМК Пресс, 2023. – 768 с.
3. Шитиков В.К., Мастицкий С.Э. (2017) Классификация, регрессия и другие алгоритмы Data Mining с использованием R. 351 с. − Электронная книга, адрес доступа: https://github.com/ranalytics/data-mining.
4. Мастицкий С. Э. Статистический анализ и визуализация данных с помощью R / С. Э. Мастицкий, В. К. Шитиков. – Москва : ДМК Пресс, 2023. – 497 с.
5. Шитиков В.К., Мастицкий С.Э. (2017) Классификация, регрессия и другие алгоритмы Data Mining с использованием R. 351 с. − Электронная книга, адрес доступа: <https://github.com/ranalytics/data-mining>.
6. Анализ данных: учебник для вузов / под редакцией В. С. Мхитаряна. — Москва : Издательство Юрайт, 2025. — 448 с. — (Высшее образование). — ISBN 978-5-534-19964-2. — Текст : электронный // Образовательная платформа Юрайт [сайт]. — URL: <https://urait.ru/bcode/560311> (дата обращения: 22.06.2025).
7. Миркин, Б. Г. Базовые методы анализа данных : учебник и практикум для вузов / Б. Г. Миркин. — 3-е изд., перераб. и доп. — Москва : Издательство Юрайт, 2025. — 297 с. — (Высшее образование). — ISBN 978-5-534-19709-9. — Текст : электронный // Образовательная платформа Юрайт [сайт]. — URL: <https://urait.ru/bcode/560414> (дата обращения: 22.06.2025).
8. Джеймс Г., Уиттон Д., Хасти Т., Тибширани Р. Введение в статистическое обучение с примерами на языке R. – М.: ДМК Пресс, 2016. – 450 с.
9. Руниони Справочник по непараметрической статистике
10. Аптон. Анализ таблиц сопряженности
11. Барсегян А.А., Куприянов М.С., Степаненко В.В., Холод И.И. Методы и модели анализа данных: *OLAP* и *Data Mining*. – Спб.: БХВ-Петербург
12. Айвазян С.А., Мхитарян В.С. Прикладная статистика и основы эконометрики. Учебник для вузов. – М.: ЮНИТИ, 1998. – 1022 с.
13. Колемаев В.А., Староверов О.В., Турундаевский В.Б. Теория вероятностей и математическая статистика: учеб. пособие для экон. спец. вузов. – М.: высш. шк., 1991. – 400 с.
14. Факторный, дискриминантный и кластерный анализ/Пер. с англ. А.М. Хотинского. Под ред. И.С. Енюкова. -М.: Финансы и статистика, 1989.
15. Паклин Н.Б., Орешков В.И. Бизнес-аналитика: от данных к знаниям. – Спб.: Питер, 2013. – 704 с.
16. <https://otexts.com/fpp2/accuracy.html> - Прогнозирование временных рядов в R.
17. Айвазян С.А., Енюков И.С., Мешалкин Л.Д. Прикладная статистика: Основы моделирования и первичная обработка данных. – М.: «Финансы и статистика», 1983. – 471 с.
18. Бокс Дж., Дженкинс Г. Анализ временных рядов. Прогноз и управление. М.: Мир, вып. 1, 1974. – 406 с.; вып. 2 – 197 с.
19. Дж. Д. Лонг и Пол Титор Книга рецептов: Проверенные рецепты для статистики, анализа и визуализации данных / пер. с анг. Д. А. Беликова. – М.: ДМК Пресс, 2020. – 510 с.
20. Кендэл М. Временные ряды. – М.: Финансы и статистика, 1981. – 199 с.
21. Кильдишев Г.С., Френкель А.А. Анализ временных рядов и прогнозирование. – М.: Статистика, 1973. – 103 с.
22. **Мастицкий С. Э. (2020) Анализ временных рядов с помощью R. — Электронная книга, адрес доступа:** <https://ranalytics.github.io/tsa-with-r>
23. Дуглас Люк: Анализ сетей (графов) в среде R. Руководство пользователя. Цветное издание. (перевод с английского, ISBN: 978-5-97060-428-1 )

<https://github.com/polyakovyevgeniy/Network_Analysis_in_R_russian_translation/blob/master/%D0%94%D1%83%D0%B3%D0%BB%D0%B0%D1%81%20%D0%9B%D1%8E%D0%BA_%D0%90%D0%BD%D0%B0%D0%BB%D0%B8%D0%B7%20%D1%81%D0%B5%D1%82%D0%B5%D0%B9%20(%D0%B3%D1%80%D0%B0%D1%84%D0%BE%D0%B2)%20%D0%B2%20%D1%81%D1%80%D0%B5%D0%B4%D0%B5%20R.pdf>