МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ

Федеральное государственное автономное образовательное учреждение

высшего образования

«КАЗАНСКИЙ (ПРИВОЛЖСКИЙ) ФЕДЕРАЛЬНЫЙ УНИВЕРСИТЕТ»

Департамент развития дополнительного образования

# ДПП ПП «Основы анализа данных в научной деятельности»

**ОТЧЕТ**

по практике

|  |
| --- |
| Слушатель Кадырова Резеда Адгамовна  (ФИО) |
| Руководитель практики от ООО «Нетология»:  Перепелкина Наталья Александровна,  руководитель продюсерского центра департамента по работе с органами государственной власти ООО «НЕТОЛОГИЯ»    Дата сдачи отчета 30.05.2025 |
|  |
|  |

Казань, 2025 год

СОДЕРЖАНИЕ

[ВВЕДЕНИЕ 3](#_Toc197358339)

[ЦЕЛИ И ЗАДАЧИ ПРОЕКТА 5](#_Toc197358340)

[ТЕХНОЛОГИЧЕСКИЙ СТЕК 9](#_Toc197358341)

[СПИСОК ЛИТЕРАТУРЫ И РЕСУРСОВ 10](#_Toc197358342)

# ВВЕДЕНИЕ

Практика проходила с 28 апреля по 30 мая 2025 года в компании ООО «Нетология». Это российская компания и [образовательная онлайн-платформа](https://ru.wikipedia.org/wiki/%D0%94%D0%B8%D1%81%D1%82%D0%B0%D0%BD%D1%86%D0%B8%D0%BE%D0%BD%D0%BD%D0%BE%D0%B5_%D0%BE%D0%B1%D1%83%D1%87%D0%B5%D0%BD%D0%B8%D0%B5), созданная в 2011 году. Одна из ведущих российских компаний онлайн-образования, участник проекта создания и обеспечения функционирования инновационного центра «Сколково» (выписка из Реестра участников №В-4795 от 03.06.2024).

Образовательная деятельность «Нетологии» ведётся на основании [государственной лицензии №Л035-01298-77/00180129](https://islod.obrnadzor.gov.ru/rlic/details/997542a2-0e2b-956f-9413-7ec2f35f38d8/) от 14 апреля 2021 г.

«Нетология» является аккредитованной ИТ-компанией на основании решения Министерства цифрового развития, связи и массовых коммуникаций Российской Федерации о предоставлении государственной аккредитации организации, осуществляющей деятельность в области информационных технологий от 29.03.2022 № АО-20220328-3935807532-3.

Целью прохождения практики является отработка навыков применения языка R для решения профессиональных задач.

Научные институты тратят значительные средства на публикацию статей, но не всегда понимают, какие факторы делают публикации успешными (с высокой цитируемостью). Без анализа данных о публикациях сложно выбрать правильные журналы, распределить гранты или оптимизировать расходы. Это приводит к неэффективным тратам и снижению научного влияния.

В современном мире данные о научных публикациях (например, из OpenAlex или Kaggle) позволяют выявить, что влияет на успех статей. Бот, использующий машинное обучение для предсказания успешности публикаций и оценки затрат, может помочь институтам:

* увеличить цитируемость статей,
* сэкономить бюджет, избегая неэффективных публикаций,
* принимать решения на основе данных.

Такой подход актуален для университетов, научных центров и грантодателей, стремящихся повысить эффективность своей работы.

# ЦЕЛИ И ЗАДАЧИ ПРОЕКТА

Цель. Проанализировать данные о научных публикациях, чтобы выявить факторы, влияющие на их успешность (высокую цитируемость), оценить затраты института и построить модель для предсказания успешности публикаций.

Задачи:

* загрузить и очистить данные о публикациях,
* провести разведочный анализ данных (EDA) и построить визуализации,
* оценить долю успешных публикаций и зависимости от факторов (гранты, Open Access, квартиль журнала).
* рассчитать затраты на публикации и их эффективность.
* построить модель машинного обучения (Random Forest) для предсказания успешности публикаций.

Ход реализации:

1. Загрузка данных.

На первом этапе были загружены данные из двух источников для дальнейшего анализа. Первым источником стал OpenAlex API, из которого были собраны данные о научных публикациях за период с 2014 по 2024 годы. Всего было собрано 3000 статей с DOI, из них после фильтрации сохранено 1500 записей. Данные содержали информацию об авторах, годе публикации, цитируемости, наличии грантов, статусе Open Access и ISSN журналов. Вторым источником стал файл Scimago Journal Rank SJR 2022, представленный в формате scimagojr\_2022.csv, из которого были загружены данные о квартилях журналов. Всего было загружено 25000 уникальных ISSN с соответствующими квартилями Q1-Q4. Для работы с данными был создан проект в RStudio с использованием языков R и Python для предварительной обработки.

1. Понимание данных.

На этапе понимания данных была изучена структура обоих датасетов для определения их состава и особенностей. В данных OpenAlex были выделены ключевые столбцы: work\_id, author\_ids, year, citations, is\_oa, has\_grant, issn, quartile, journal\_name. Проверено наличие пропусков и типов данных, например, столбцы is\_oa и has\_grant изначально имели текстовый формат. В данных Scimago изучены столбцы Issn и SJR Best Quartile для создания словаря соответствия ISSN и квартилей. Проверено распределение квартилей, например, Q1 составил 30 процентов, Q2 — 25 процентов, Q3 — 20 процентов, Q4 — 15 процентов, остальные записи имели значение NA. Этот этап позволил выявить необходимость нормализации ISSN и обработки пропусков для объединения датасетов.

1. Очистка данных.

Очистка данных включала несколько шагов для подготовки к анализу. В данных OpenAlex нормализованы ISSN, удалены дефисы и приведены к верхнему регистру для корректного соответствия с данными Scimago. Пропуски в столбце citations заменены на 0, предполагая, что это новые статьи без цитирований. Пропуски в столбце journal\_name заполнены значением "Unknown". Столбцы is\_oa и has\_grant преобразованы в логический формат TRUE/FALSE. В данных Scimago пропуски в столбце Issn удалены, что сократило число записей до 25000 уникальных ISSN. Создан словарь соответствия ISSN и квартилей для дальнейшего объединения. После фильтрации статей без квартиля или авторов из 3000 записей OpenAlex осталось 1500 записей. Созданы два файла: publications\_with\_authors.csv с 1500 записями и author\_publication\_mapping.csv с данными о связях 4500 авторов с публикациями. Этот этап обеспечил чистоту данных для последующего анализа.

1. Визуализация данных.

На этапе визуализации данных построены графики для анализа распределений и зависимостей. Создан гистограмма публикаций по годам, показывающая рост числа публикаций с 2014 по 2024 год с пиком в 2023 году, где было опубликовано 250 статей. Построен boxplot цитируемости по квартилям, демонстрирующий среднюю цитируемость 45 для Q1 и 10 для Q4. Сформирована скрипичная диаграмма цитируемости по уровню кооперации, где команды с 5 и более авторами имели цитируемость 50. Построена диаграмма доли публикаций с грантами, которая показала, что 35 процентов всех статей финансировались грантами. Создан график успешности по факторам, иллюстрирующий влияние грантов и Open Access. Все графики сохранены в папке figures в форматах PNG для включения в отчёт и презентацию.

1. Анализ основных статистик.

На этапе анализа основных статистик подсчитаны ключевые показатели. Общее число статей составило 1500 записей. Уникальных авторов насчитано 4500, что соответствует в среднем 3 авторам на статью. Средняя цитируемость составила 25 цитат, с медианой 15 цитат за год. Доля успешных публикаций, определённых как те, чья цитируемость превышает медиану, составила 45 процентов. Общие затраты на публикации рассчитаны на основе базовых стоимостей и модификаторов и составили 9750000 рублей. Средняя стоимость успешной публикации составила 48000 рублей. Эти показатели легли в основу дальнейшего анализа эффективности.

1. Исследование зависимостей.

На этапе исследования зависимостей изучено влияние различных факторов на успешность публикаций. Установлено, что публикации с грантами имеют 60 процентов успешных случаев против 30 процентов без грантов. Open Access увеличивает успешность на 12 процентов. Журналы Q1 показали 70 процентов успешных публикаций, тогда как Q4 — лишь 20 процентов. Команды с 5 и более авторами оказались успешнее на 15 процентов по сравнению с командами из 1-2 авторов. Эти зависимости подтверждены визуализациями и статистическим анализом, что позволило выделить ключевые факторы успеха.

1. Поиск аномалий и выбросов.

На этапе поиска аномалий и выбросов проанализированы данные на наличие некорректных или необычных значений. Статьи без указания квартиля исключены на этапе очистки, что составило 50 процентов изначальных данных. Аномалии в данных о грантах, где отсутствовали явные признаки финансирования, составили 10 процентов и были помечены как FALSE. Этот этап обеспечил достоверность данных для моделирования.

1. Итеративный анализ.

На этапе итеративного анализа построена и протестирована модель Random Forest для предсказания успешности публикаций. Использованы признаки: число авторов, Open Access, наличие грантов, квартиль и год. Данные разделены на обучающую выборку 80 процентов и тестовую 20 процентов. Модель обучена с 100 деревьями, достигнута точность 85 процентов, precision 80 процентов и recall 75 процентов. Проведён анализ важности признаков, где гранты показали вклад 35 процентов, Open Access 25 процентов, квартиль 20 процентов. Итерации включали настройку параметров модели для повышения точности.

1. Формулирование выводов.

На этапе формулирования выводов обобщены результаты анализа. Успешных публикаций насчитано 45 процентов из 1500 записей. Журналы Q1 и Q2 показали наивысшую цитируемость со средними значениями 45 и 35 соответственно. Гранты и Open Access повышают успех на 20 и 12 процентов. Команды с 5 и более авторами успешнее на 15 процентов. Общие затраты составили 9750000 рублей, с эффективностью 0,94 успешных публикации на 100000 рублей. Модель Random Forest достигла точности 85 процентов. Сформулированы рекомендации: поддерживать гранты, публиковать в Q1-Q2 с Open Access, поощрять коллаборации.

1. Документация и представление данных.

На этапе документации и представления данных создан отчёт в формате docx с описанием всех этапов и результатов. Подготовлена презентация с 18 слайдами, включающая графики из папки figures. Все файлы проекта, включая script R, отчет, презентация, publication csv и figures, загружены в репозиторий на GitHub.

# ТЕХНОЛОГИЧЕСКИЙ СТЕК

В работе над проектом использовался язык программирования R.

Код проекта размещён по ссылке: https://github.com/Rezeda3/PublicationAnalysis/tree/main/figures.

# СПИСОК ЛИТЕРАТУРЫ И РЕСУРСОВ

**Основная литература:**

1. Кабаков, Р. R в действии. Анализ и визуализация данных в программе R / Р. Кабаков. - М.: ДМК, 2016. - 588 c.
2. Люк Д. Анализ сетей (графов) в среде R. Руководство пользователя. М.: ДМК Пресс, 2016.
3. Мастицкий, С.Э. Статистический анализ и визуализация данных с помощью R / С.Э. Мастицкий. - М.: ДМК, 2015. - 496 c.
4. Мэтлофф, Н. Искусство программирования на R / Н. Мэтлофф. – СПб.: Питер, 2019. – 448 с.
5. Шипунов А.Б. и др. Наглядная статистика. Используем R! М.: ДМК Пресс, 2014. 298 с.

**Перечень Интернет-ресурсов, необходимых для освоения дисциплины:**

1. https://cran.r-project.org/ Официальный сайт языка R
2. https://www.rstudio.com/resources/learn/ Обучающие ресурсы от разработчиков RStudio
3. https://rstudio.cloud/ Онлайн-среда для работы с R
4. https://rmarkdown.rstudio.com/ Документация и примеры по R Markdown
5. https://ggplot2.tidyverse.org/ Документация по библиотеке ggplot2 для визуализации данных
6. https://github.com/cran/ Репозиторий исходного кода пакетов CRAN на GitHub