

МИНОБРНАУКИ РОССИИ  
Федеральное государственное бюджетное образовательное учреждение  
высшего образования  
«Костромской государственный университет»  
(КГУ)

## РАБОЧАЯ ПРОГРАММА ДИСЦИПЛИНЫ

Технологии анализа больших данных

Направление подготовки: 01.03.02 “Прикладная математика и информатика”  
Профиль “Анализ данных”

Квалификация выпускника: бакалавр

**Кострома  
2025**

Рабочая программа дисциплины «Технологии анализа больших данных» разработана в соответствии с Федеральным государственным образовательным стандартом высшего образования по направлению подготовки Технологии анализа больших данных (уровень подготовки бакалавриат), утверждённым приказом №9 от 10.01.2018 г.

Разработал: Ивков Владимир Анатольевич, к.э.н., доцент

Рецензент: Козырев Сергей Борисович, доцент КГУ

**ПРОГРАММА УТВЕРЖДЕНА:**

На заседании кафедры Прикладной математики и информационных технологий:

Протокол заседания кафедры №4 от 18.02.2025 г.

Заведующий кафедрой Прикладной математики и информационных технологий

Ивков Владимир Анатольевич, доц., к.э.н.

## 1. Цели и задачи освоения дисциплины

**Цель дисциплины:** получение студентом целостного представления о возможностях и ограничениях современных статистических методов и машинного обучения, а также об особенностях анализа данных для решения задач, возникающих в анализе данных.

### Задачи дисциплины:

- дать базовые знания о современных методах анализа данных и инструментах работы с ними;
- научить выполнять сбор, очистку и предварительную обработку данных;
- сформировать навыки статистического анализа и визуализации данных;
- познакомить с основами машинного обучения и построением прогнозных моделей;
- развить умение интерпретировать результаты анализа и представлять их в наглядной форме.

## 2. Перечень планируемых результатов обучения по дисциплине

В результате освоения дисциплины обучающийся должен:

### освоить компетенцию:

ПК-1. Способность выполнять проверку работоспособности и рефакторинг кода программного обеспечения, интеграцию программных модулей и компонент.

ПК-2. Способность участвовать в практическом освоении систем управления качеством.

### Знать:

- Основные понятия и методы анализа данных;
- Источники открытых данных и способы их обработки;
- Основные статистические характеристики (среднее, медиана, дисперсия, корреляция);
- Базовые алгоритмы машинного обучения (регрессия, классификация, кластеризация);
- Принципы визуализации данных и типы графиков;
- Современные технологии анализа данных (нейронные сети, NLP).

### Уметь:

- Обработать и очистить данные с помощью Python (Pandas, NumPy);
- Проводить статистический анализ и проверку гипотез;
- Строить модели машинного обучения (Scikit-learn);
- Визуализировать данные (Matplotlib, Seaborn, Plotly);
- Применять нейросетевые модели для решения задач (Keras, TensorFlow).

### Владеть:

- Навыками работы с Jupyter Notebook и средой разработки;
- Методами предобработки и анализа данных;
- Техниками визуализации для представления результатов;
- Базовыми инструментами машинного обучения и глубокого обучения

## 3. Место дисциплины в структуре ОП ВО

Дисциплина «Технологии анализа больших данных» относится к части учебного плана, формируемой участниками образовательных отношений. Изучается в 7 семестре обучения. Она способствует углублению аналитической подготовки бакалавров, предваряя последующее изучение дисциплины «Интеллектуальный анализ данных». Также она даёт начальные практические навыки разработки решений в условиях ограниченного времени.

Изучение дисциплины основывается на ранее освоенных темах дисциплины «математика анализа данных», «Анализ данных на python».

**4. Объём дисциплины «Технологии анализа больших данных»**  
**4.1. Объём дисциплины в зачётных единицах с указанием академических (астрономических) часов и виды учебной работы**

Виды учебной работы,	Очная форма
Общая трудоёмкость в зачётных единицах	4
Общая трудоёмкость в часах	144
Аудиторные занятия в часах, в том числе:	46
Лекции	16
Лабораторные занятия	30
Практическая подготовка	–
Самостоятельная работа в часах	98
Форма промежуточной аттестации	Экзамен

**4.2. Объём контактной работы на 1 обучающегося**

Виды учебных занятий	Очная форма
Лекции	16
Лабораторные занятия	30
Консультации	2
Зачёт/зачёты	0,35
Экзамен/экзамены	-
Курсовые работы	-
Курсовые проекты	-
Практическая подготовка	-
Всего	48,35

**5. Содержание дисциплины «Технологии анализа больших данных»,  
структурированное по темам (разделам), с указанием количества часов и  
видов занятий**

**5.1 Тематический план учебной дисциплины**

№	Название раздела, темы	Всего з.е/час	Аудиторные занятия		Самостоятельная работа
			Лекц.	Лаб.	
1	Введение в анализ данных в	0,67/24	2	5	17
2	Поиск и обработка данных	0,67/24	2	5	17
3	Основы статистического анализа данных	0,67/24	2	5	17
4	Машинное обучение	0,67/24	2	5	17
5	Визуализация данных	0,67/24	4	5	15
6	Продвинутые методы анализа данных	0,67/24	4	5	15
	Итого:	4/144	16	30	98

**5.2. Содержание:**

**Тема 1. Введение в анализ данных.** Основные понятия анализа данных. Практические кейсы. Применение точных наук. Цифровые технологии. Данные для принятия решений.

**Тема 2. Поиск и обработка данных.** Источники данных. Открытые данные. Базы данных. Подбор данных под задачи, портал данных. Оценка качества данных.

**Тема 3. Основы статистического анализа.** Генеральная совокупность и выборка. Репрезентативность. Распределения. Типы переменных: количественные, порядковые, категориальные. Меры центральной тенденции: медиана, среднее, мода. Меры разброса: дисперсия, стандартное отклонение. Квантили и перцентили. Корреляция, p-value. Критерий Стьюдента. Ложные корреляции. Причинно-следственные связи.

**Тема 4. Машинное обучение.** Обучение с учителем и без учителя. Регрессия. Классификация. Кластеризация. Основная терминология. Практические примеры.

**Тема 5. Визуализация данных.** Графики плотности распределения. Диаграммы рассеяния. Гистограммы. Столбиковые, линейные, коробчатые, круговые диаграммы. Нестандартные типы визуализаций. Критерии выбора графиков.

**Тема 6. Продвинутое методы анализа.** Нейронные сети. Архитектура и принципы работы. Биологические аналоги. Области применения. Современные технологии в анализе данных.

## 6. Методические материалы для обучающихся по освоению дисциплины «Технологии анализа больших данных»

### 6.1. Самостоятельная работа обучающихся по дисциплине (модулю)

№ п/п	Раздел (тема) дисциплины	Задание	Часы	Методические рекомендации по выполнению задания	Форма контроля
1	Введение в анализ данных	Знакомство с сайтами. Регистрация на них	17	Используйте литературу	Индивидуальная проверка
2	Поиск и обработка данных	Изучение литературы, решение задач	17	Используйте литературу	Проверка домашних заданий
3	Основы статистического анализа данных	Изучение литературы, решение задач	17	Используйте литературу	Проверка домашних заданий
4	Машинное обучение	Изучение литературы, решение задач	17	Используйте литературу	Проверка домашних заданий
5	Визуализация данных	Изучение литературы, решение задач	15	Используйте литературу	Проверка домашних заданий
6	Продвинутые методы анализа данных	Изучение литературы, решение задач	15	Используйте литературу	Проверка домашних заданий

### 6.2. Тематика и задания для практических занятий

*Не предусмотрено.*

### 6.3. Тематика и задания для лабораторных занятий

Тема 1. Введение в анализ данных

- Разбор реальных кейсов – анализ готовых примеров применения data science.
- Знакомство с инструментами – базовые операции в Python (Pandas, NumPy) или R.
- Работа с датасетами – загрузка, первичный осмотр и очистка данных.

Тема 2. Поиск и обработка данных

- Поиск открытых данных – работа с Kaggle, государственными и научными порталами.

- Предобработка данных – обработка пропусков, аномалий, дубликатов.
- Формирование выборки – фильтрация и агрегация данных под задачу.

### Тема 3. Основы статистического анализа

- Описательная статистика – расчет среднего, медианы, дисперсии, визуализация распределений.
- Анализ взаимосвязей – корреляционный анализ, тепловые карты.
- Проверка гипотез – t-тест, ANOVA, интерпретация p-value.

### Тема 4. Машинное обучение

- Линейная регрессия – предсказание числовых значений.
- Классификация – логистическая регрессия, дерево решений.
- Кластеризация – метод k-средних, визуализация кластеров.

### Тема 5. Визуализация данных

- Построение базовых графиков – гистограммы, box-plot, scatter plot.
- Интерактивная визуализация – Plotly, Seaborn, Tableau (по выбору).
- Оптимизация графиков – подбор цветов, подписей, улучшение читаемости.

### Тема 6. Продвинутое методы анализа

- Нейросети на практике – реализация простой модели на Keras/TensorFlow.
- Работа с временными рядами – предсказание с использованием LSTM.
- Анализ текстовых данных – базовые методы NLP (токенизация, TF-IDF).

## **7. Перечень основной и дополнительной литературы, необходимой для освоения дисциплины «Технологии анализа больших данных»**

### *а) основная:*

1. Bluman A. Elementary statistics. McGraw-Hill, 2008
2. Bluman A. Probability demystified. McGraw-Hill, 2005
3. Tabachnick B.G., Fidell L.S Using Multivariate Statistics. Pearson Education, 2007

### *б) дополнительная:*

- Никитин А. Data Science для начинающих. – М.: Издательство, 2020. – 250 с.
- Хейдт М. Python для визуализации данных. – СПб.: Питер, 2022. – 320 с.
- Вандер Плас Дж. Python для сложных задач: наука о данных и машинное обучение. – М.: ДМК Пресс, 2018. – 470 с.

## **8. Перечень ресурсов информационно-телекоммуникационной сети «Интернет», необходимых для освоения дисциплины**

### *Электронные библиотечные системы:*

1. ЭБС Университетская библиотека онлайн - <http://biblioclub.ru>
2. ЭБС «Лань» <https://e.lanbook.com>
3. ЭБС «ZNANIUM.COM» <http://znanium.com>

## **9. Описание материально-технической базы, необходимой для**

## **осуществления образовательного процесса по дисциплине**

Лекционные и практические занятия проводятся в аудиториях с требуемым числом посадочных мест, оборудованные мультимедиа.

Лабораторные работы проводятся в компьютерных классах.

Лицензионное программное обеспечение:

LibreOffice 5.0, лицензия GNU LGPL.

Свободно распространяемое программное обеспечение:

- MS Excel