**БЕЛОРУССКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ**

УТВЕРЖДАЮ
Ректор Белорусского
государственного университета
_____А.Д.Король
27 июня 2025 г.
Регистрационный № 3744/м.

# МЕТОДЫ ОБРАБОТКИ И АНАЛИЗА РАЗНОРОДНЫХ ДАННЫХ

Учебная программа учреждения образования по учебной дисциплине для специальности:

**7-06-0533-05 Прикладная математика и информатика**

Профилизация: Интеллектуальные системы

2025 г.

Учебная программа составлена на основе ОСВО 7-06-0533-05-2023 и учебного плана № М53-5.3-54/уч. от 23.05.2025.

**СОСТАВИТЕЛИ:**

***М.М.Лукашевич***, доцент кафедры информационных систем управления факультета прикладной математики и информатики Белорусского государственного университета

**РЕЦЕНЗЕНТЫ:**

***А.С.Сидорович***, ведущий инженер-программист ИПУП «ИССОФТ СОЛЮШЕНЗ»;

***В.В.Старовойтов***, главный научный сотрудник государственного научного учреждения «Объединенный институт проблем информатики Национальной академии наук Беларуси», д.т.н., профессор

**РЕКОМЕНДОВАНА К УТВЕРЖДЕНИЮ:**

Кафедрой информационных систем управления БГУ
(протокол № 15 от 19.06.2025);

Научно-методическим советом БГУ
(протокол № 11 от 26.06.2025)

Заведующий кафедрой _____ А.М.Недзьведь

## ПОЯСНИТЕЛЬНАЯ ЗАПИСКА

**Цели и задачи учебной дисциплины**

Учебная дисциплина «Методы обработки и анализа разнородных данных» знакомит магистрантов с теоретическими основами и современными методами анализа, интеграции и машинного обучения на данных разнородной природы - включая текстовые, графовые, временные ряды, изображения и мультимодальные наборы. Дисциплина направлена на развитие ключевых профессиональных компетенций, необходимых для проектирования и реализации интеллектуальных систем в условиях сложных, неоднородных данных.

Цель учебной дисциплины – формирование у магистрантов компетенций, обеспечивающих способность: применять интеллектуальные методы и алгоритмы для решения задач поиска, распознавания и обработки данных; использовать методики проектирования технических процессов и систем при работе с гетерогенными данными; интегрировать и обрабатывать разнородные данные в рамках комплексных задач.

**Задачи учебной дисциплины**:

1. освоить методы и алгоритмы обработки, представления и интеграции разнородных данных, в том числе с использованием интеллектуальных подходов;
2. научиться проектировать конвейеры обработки данных и модели машинного обучения, адаптированные под специфику гетерогенных источников;
3. развить практические навыки интеграции данных различных типов (текст, изображение, граф, таблица, временной ряд) в единую модель или систему;
4. ознакомиться с современными инструментами и фреймворками, позволяющими эффективно реализовывать методы обработки разнородных данных.

**Место учебной дисциплины** в системе подготовки специалиста с высшим образованием (магистра).

Учебная дисциплина относится к модулю «Методы интеллектуального анализа данных» компонента учреждения образования.

Программа составлена с учетом межпредметных связей с учебными дисциплинами. Основой для изучения учебной дисциплины являются следующие учебные дисциплины первой ступени высшего образования: «Технологии анализа и визуализации данных», «Основы и методологии программирования» и дисциплина второй ступени высшего образования «Модели и методы искусственного интеллекта».

**Требования к компетенциям**

Освоение учебной дисциплины «Методы обработки и анализа разнородных данных» должно обеспечить формирование следующих компетенций:

***Специализированные компетенции:***

Использовать методы и алгоритмы (в том числе интеллектуальные) для решения задач поиска, распознавания и обработки данных.

Применять методики проектирования технических процессов и систем.

Применять навыки интеграции и использования разнородных данных.

В результате изучения дисциплины студент должен:

**знать:**

- особенности структуры, источников и типов разнородных (гетерогенных) данных;
- математические основы и ключевые методы машинного обучения, адаптированные для работы с разнородными данными;
- принципы проектирования и интеграции компонентов интеллектуальных систем, использующих данные разной природы;

**уметь:**

- проектировать и реализовывать конвейеры предварительной обработки, векторного представления и интеграции разнородных данных;
- выбирать и адаптировать алгоритмы машинного обучения под специфику гетерогенных наборов данных и решаемые прикладные задачи;
- строить и оценивать модели интеллектуальных систем на основе разнородных данных с учётом требований к интерпретируемости, надёжности и масштабируемости;

**иметь навык:**

- современными программными средствами и фреймворками для обработки и анализа разнородных данных;
- навыками интеграции данных различных типов и применения методов машинного обучения при решении практических задач в области интеллектуальных систем.

**Структура учебной дисциплины**

Дисциплина изучается в 1-ом и 2-ом семестрах. В соответствии с учебным планом всего на изучение учебной дисциплины «Методы обработки и анализа разнородных данных» отведено для очной формы получения высшего образования – 192 часа, в том числе 68 аудиторных часов, из них: лекции – 32 часа, лабораторные занятия – 36 часов.

***1 семестр:***

Лекции – 16 часов, лабораторные занятия – 18 часов.

Трудоемкость учебной дисциплины составляет 3 зачетные единицы.

Форма промежуточной аттестации – зачет.

***2 семестр:***

Лекции – 16 часов, лабораторные занятия – 18 часов.

Трудоемкость учебной дисциплины составляет 3 зачетные единицы.

Форма промежуточной аттестации – экзамен.

## СОДЕРЖАНИЕ УЧЕБНОГО МАТЕРИАЛА

### Раздел 1. Основы и базовые методы обработки разнородных данных

***Тема 1.1. Введение в разнородные данные***

Понятие разнородных (гетерогенных) данных. Классификация данных: структурированные, полуструктурированные, неструктурированные. Источники разнородных данных. Основные проблемы: масштаб, шум, отсутствие меток, семантическая несогласованность. Типовые задачи машинного обучения на разнородных данных: классификация, кластеризация, прогнозирование.

***Тема 1.2. Предварительная обработка и нормализация разнородных данных***

Очистка данных: обнаружение и обработка выбросов, интерполяция пропущенных значений. Нормализация и стандартизация признаков разной природы. Кодирование категориальных признаков. Балансировка несбалансированных классов.

***Тема 1.3. Векторизация и эмбеддинги разнородных данных***

Векторизация текстов. Представление изображений. Эмбеддинги временных рядов. Векторизация графов. Понятие мультимодальных эмбеддингов.

***Тема 1.4. Интеграция разнородных источников данных***

Стратегии интеграции. Проблема согласования шкал, единиц измерения и временных меток. Примеры интеграции: текст + табличные данные, изображение + временной ряд, сенсорные данные + метаданные.

***Тема 1.5. Снижение размерности для гетерогенных данных***

Классические методы: PCA, t-SNE, UMAP. Особенности применения к смешанным типам данных. Совместное снижение размерности: канонический корреляционный анализ (CCA), Multi-view PCA. Визуализация гетерогенных данных.

***Тема 1.6. Обучение без учителя на разнородных данных***

Кластеризация смешанных данных: k-prototypes, метрика Гауэра (Gower distance), DBSCAN. Мультимодальная кластеризация. Обнаружение аномалий: Isolation Forest, автоэнкодеры, One-Class SVM. Оценка качества кластеризации и детекции аномалий.

***Тема 1.7. Обучение с учителем на разнородных признаках***

Модели для гетерогенных данных: деревья решений, ансамбли (Random Forest, XGBoost), нейронные сети. Проблема доминирования одной модальности. Анализ важности признаков (feature importance) в моделях с разнородными входами.

***Тема 1.8. Основы мультимодального машинного обучения***

Понятие мультимодальности: текст + изображение, аудио + текст, видео + сенсорные данные. Архитектуры мультимодальных моделей. Обзор современных моделей. Области применения.

## Раздел 2. Продвинутые методы и приложения в интеллектуальных системах

***Тема 2.1. Обработка временных и пространственных разнородных данных***

Особенности многоканальных временных рядов и геопространственных данных. Методы представления: оконное преобразование, спектральный анализ, геохэширование. Интеграция с внешними источниками.

***Тема 2.2. Графовые и реляционные данные в машинном обучении***

Представление знаний в виде графов. Модели для гетерогенных графов. Интеграция графовых и табличных данных.

***Тема 2.3. Работа с неполными и несбалансированными разнородными данными***

Проблема отсутствующих модальностей и пропущенных признаков. Методы импутации: MICE, KNN-импутация, глубокие модели. Обучение с частичной информацией. Балансировка классов в мультимодальных задачах. Оценка устойчивости моделей к неполноте данных.

***Тема 2.4. Онлайн-обработка и потоковые разнородные данные***

Особенности потоковых данных: высокая скорость, изменчивость распределений. Методы адаптации моделей. Онлайн-интеграция источников.

***Тема 2.5. Интерпретируемость и объяснимость моделей на разнородных данных***

Задача объяснимого ИИ (XAI) в контексте гетерогенных моделей. Методы интерпретации: LIME, SHAP, attention visualization. Анализ вклада отдельных модальностей в предсказание.

***Тема 2.6. Этические и правовые аспекты обработки разнородных данных***

Источники и последствия смещений в разнородных данных. Проблемы приватности и конфиденциальности. Регуляторные требования. Принципы справедливого и ответственного ИИ.

***Тема 2.7. Современные инструменты и фреймворки***

Библиотеки для обработки разнородных данных. Фреймворки глубокого обучения. Специализированные инструменты. Средства управления экспериментами.

***Тема 2.8. Прикладные задачи и перспективы в интеллектуальных системах***

Реальные применения: здравоохранение, умные города, финансы, промышленность.

## УЧЕБНО-МЕТОДИЧЕСКАЯ КАРТА УЧЕБНОЙ ДИСЦИПЛИНЫ

Очная (дневная) форма получения высшего образования с применением дистанционных образовательных технологий (ДОТ)

| Номер раздела, темы | Название раздела, темы | Количество аудиторных часов | | | | | Форма контроля |
|---|---|---|---|---|---|---|---|
| | | Лекции | Практические занятия | Семинарские занятия | Лабораторные занятия | Иное | |
| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
| | **Методы обработки и анализа разнородных данных** | **32** | | | **36** | | |
| **1.** | **Основы и базовые методы обработки разнородных данных** | **16** | | | **18** | | |
| 1.1. | Введение в разнородные данные | 2 | | | | | Экспресс-опрос |
| 1.2. | Предварительная обработка и нормализация разнородных данных | 2 | | | 4 | | Контрольная работа |
| 1.3. | Векторизация и эмбеддинги разнородных данных | 2 | | | 4 | | Контрольная работа |
| 1.4. | Интеграция разнородных источников данных | 2 | | | | | Экспресс-опрос |
| 1.5. | Снижение размерности для гетерогенных данных | 2 | | | 4 | | Экспресс-опрос |
| 1.6. | Обучение без учителя на разнородных данных | 2 | | | 2 | | Контрольная работа |
| 1.7. | Обучение с учителем на разнородных признаках | 2 | | | 4 | | Контрольная работа |
| 1.8. | Основы мультимодального машинного обучения | 2 | | | | | Экспресс-опрос |
| **2.** | **Продвинутые методы и приложения в интеллектуальных системах** | **16** | | | **18** | | Экспресс-опрос |
| 2.1. | Обработка временных и пространственных разнородных данных | 2 | | | 4 | | Контрольная работа |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| | | | | | | | Экспресс-опрос |
| 2.2. | Графовые и реляционные данные в машинном обучении | 2 | | | | | |
| 2.3. | Работа с неполными и несбалансированными разнородными данными | 2 | | | 4 | | Контрольная работа |
| 2.4. | Онлайн-обработка и потоковые разнородные данные | 2 | | | 4 | | Экспресс-опрос |
| 2.5. | Интерпретируемость и объяснимость моделей на разнородных данных | 2 | | | 2 | | Контрольная работа |
| 2.6. | Этические и правовые аспекты обработки разнородных данных | 2 | | | | | Экспресс-опрос |
| 2.7. | Современные инструменты и фреймворки | 2 | | | 2 | | Контрольная работа |
| 2.8 | Прикладные задачи и перспективы в интеллектуальных системах | 2 | | | 2 | | Экспресс-опрос |

## ИНФОРМАЦИОННО-МЕТОДИЧЕСКАЯ ЧАСТЬ

### Основная литература

1. Серрано, Л. Грокаем машинное обучение / Луис Серрано ; [пер. с англ. Р. Чикин]. - Санкт-Петербург ; Москва ; Минск : Питер, 2024. - 511 с.
2. Вандер Плас, Дж. Python для сложных задач: наука о данных и машинное обучение / Дж. Вандер Плас ; [пер. с англ. И. Пальти]. - Санкт-Петербург ; Москва ; Минск : Питер, 2023. - 573 с.

### Дополнительная литература

1. Практический анализ временных рядов: прогнозирование со статистикой и машинное обучение. : Пер. с англ. - СПб. : ООО “Диалектика”, 2021. - 544 с.

### Перечень рекомендуемых средств диагностики и методика формирования итоговой отметки

Для диагностики компетенции в рамках учебной дисциплины рекомендуется использовать следующие формы:

1. Устная форма: экспресс-опрос.
2. Письменная форма: контрольные работы.

Формой промежуточной аттестации по дисциплине «Методы обработки и анализа разнородных данных» учебным планом предусмотрены зачет в 1-ом семестре и экзамен во 2-ом семестре.

Для формирования итоговой отметки по учебной дисциплине используется модульно-рейтинговая система оценки знаний студента, дающая возможность проследить и оценить динамику процесса достижения целей обучения. Рейтинговая система предусматривает использование весовых коэффициентов для текущей и промежуточной аттестации студентов по учебной дисциплине.

Формирование итоговой отметки в ходе проведения контрольных мероприятий текущей аттестации (примерные весовые коэффициенты, определяющие вклад текущей аттестации в отметку при прохождении промежуточной аттестации):

- контрольные работы – 50 %;
- экспресс-опрос – 50%.

Итоговая отметка по дисциплине рассчитывается на основе итоговой отметки текущей аттестации (модульно-рейтинговой системы оценки знаний) 50 % и экзаменационной отметки 50 %.

### Примерная тематика лабораторных занятий

1. Предварительная обработка и интеграция разнородных данных
2. Векторизация и построение эмбеддингов для данных разных типов

3. Снижение размерности и визуализация гетерогенных наборов данных
4. Кластеризация и обнаружение аномалий в разнородных данных
5. Обучение с учителем на гетерогенных признаках
6. Мультимодальное обучение: интеграция текста и изображений
7. Работа с неполными данными и вставка пропущенных модальностей
8. Проектирование и оценка интеллектуальной системы на разнородных данных

## Описание инновационных подходов и методов к преподаванию учебной дисциплины

При организации образовательного процесса большинства практических занятий используется ***практико-ориентированный подход***, который предполагает освоение содержания учебного материала через решение практических задач, а также приобретение навыков эффективного выполнения разных видов профессиональной деятельности.

Кроме этого, при организации образовательного процесса используется комбинация ***методов группового обучения***, ***проектного обучения*** и ***учебной дискуссии***. Комбинация методов предполагает: ориентацию на генерирование идей, приобретение навыков для решения исследовательских, творческих и коммуникационных задач, появление нового уровня понимания изучаемой темы, применение знаний (теорий, концепций) при решении проблем, определение способов их решения.

## Методические рекомендации по организации самостоятельной работы

Для организации самостоятельной работы студентов магистратуры по учебной дисциплине следует использовать информационно коммуникационные технологии:

Образовательный портал БГУ https://edufpmi.bsu.by,

систему AnyTask https://anytask.org/school/bsu,

разместить в сетевом доступе комплекс учебных и учебно-методических материалов (учебно-программные материалы, учебное издание для теоретического изучения дисциплины, презентации лекций, методические указания к практическим занятиям, электронные версии домашних заданий, материалы текущего контроля и текущей аттестации, позволяющие определить соответствие учебной деятельности обучающихся требованиям образовательных стандартов высшего образования и учебно-программной документации, в том числе вопросы для подготовки к экзамену, задания, вопросы для самоконтроля, список рекомендуемой литературы, информационных ресурсов и др.).

## Примерный перечень вопросов к зачету

1. Что такое разнородные (гетерогенные) данные?
2. Какие основные вызовы возникают при работе с разнородными данными?
3. Опишите этапы предварительной обработки разнородных данных.
4. В чём отличие нормализации от стандартизации?
5. Как кодируются категориальные признаки?
6. Что такое эмбеддинги? Приведите примеры эмбеддингов для текста, изображений и графов.
7. В чём суть метода TF-IDF? Как он применяется к текстовым данным?
8. Как работают Word2Vec и BERT? В чём их принципиальное отличие?
9. Что такое стратегии интеграции данных? Приведите примеры.
10. Как решается проблема семантической несогласованности при объединении источников?
11. Опишите метод главных компонент (PCA). Можно ли его применять к смешанным типам данных?
12. Что такое канонический корреляционный анализ (CCA)? Для чего он используется в контексте разнородных данных?
13. Какие метрики применяются для кластеризации смешанных данных?
14. В чём особенность алгоритма k-prototypes по сравнению с k-means?
15. Как обнаруживаются аномалии в разнородных данных? Опишите метод Isolation Forest.
16. Почему деревья решений и ансамбли (XGBoost, Random Forest) хорошо работают с гетерогенными признаками?
17. Как оценивается важность признаков в моделях с разнородными входами?
18. Приведите пример прикладной задачи, где используются данные разных типов (например, медицинская диагностика).
19. Что такое мультимодальность? Приведите три примера мультимодальных пар (модальность + модальность).
20. В чём суть архитектуры CLIP?

## Примерный перечень вопросов к экзамену

1. Какие особенности имеет анализ временных и пространственных данных в условиях их разнородности?
2. В чём заключается специфика работы с графовыми структурами в задачах машинного обучения?
3. Как машинное обучение адаптируется к данным с пропущенными или частично доступными модальностями?
4. Какие подходы используются для обработки потоковых данных, и чем они отличаются от анализа статических наборов?

5. Почему интерпретируемость моделей особенно важна при работе с разнородными данными?
6. Какие методы применяются для объяснения решений моделей, построенных на данных разных типов?
7. Какие этические риски возникают при интеграции и анализе разнородных данных?
8. Как обеспечивается защита персональных данных при построении интеллектуальных систем на гетерогенных источниках?
9. Какие программные инструменты и фреймворки наиболее эффективны для обработки разнородных данных?
10. Как организуется вычислительный процесс при работе с большими объёмами разнородной информации?
11. Что такое мультимодальное обучение и в чём его преимущества по сравнению с традиционными подходами?
12. Как современные архитектуры нейронных сетей поддерживают интеграцию различных типов данных?
13. Какие вызовы возникают при проектировании интеллектуальных систем на основе разнородных источников?
14. Как обеспечивается согласованность семантики при объединении данных из разных доменов?
15. Какие стратегии используются для балансировки вклада разных модальностей в итоговое решение модели?
16. В чём состоит роль предобученных моделей (foundation models) в обработке разнородных данных?
17. Как меняется подход к оценке качества моделей при переходе от однородных к гетерогенным данным?
18. Какие особенности проектирования систем машинного обучения учитываются при работе в реальном времени?
19. Какие тенденции определяют развитие методов анализа разнородных данных в ближайшие годы?
20. Приведите примеры применения интеллектуальных систем на основе разнородных данных в реальных отраслях (медицина, финансы, транспорт и др.).

## Рекомендуемая тематика контрольных работ

1. Методы подготовки и трансформации разнородных данных.
2. Построение числовых представлений для данных разной природы.
3. Анализ структуры смешанных данных с помощью методов кластеризации.
4. Разработка моделей предсказания на основе гетерогенных признаков.
5. Применение графовых подходов к анализу связных разнородных данных.

6. Стратегии обработки неполноты и дисбаланса в гетерогенных наборах.
7. Методы объяснения решений моделей на разнородных данных.
8. Использование современных библиотек для реализации ML-решений.

## ПРОТОКОЛ СОГЛАСОВАНИЯ УЧЕБНОЙ ПРОГРАММЫ УО

| Название учебной дисциплины, с которой требуется согласование | Название кафедры | Предложения об изменениях в содержании учебной программы учреждения образования по учебной дисциплине | Решение, принятое кафедрой, разработавшей учебную программу (с указанием даты и номера протокола) |
|---|---|---|---|
| Учебная дисциплина не требует согласования | | | |

Заведующий кафедрой
информационных систем управления,
д.т.н., доцент _____ А.М.Недзьведь

19.06.2025

**ДОПОЛНЕНИЯ И ИЗМЕНЕНИЯ К УЧЕБНОЙ ПРОГРАММЕ УО**

на _____/_____ учебный год

| № п/п | Дополнения и изменения | Основание |
|---|---|---|
| | | |

Учебная программа пересмотрена и одобрена на заседании кафедры
_____ (протокол № ____ от _____ 202_ г.)

Заведующий кафедрой
_____ _____

УТВЕРЖДАЮ
Декан факультета
_____ _____