

## Аннотация рабочей программы дисциплины

### **Б1.В.ДВ.01.01 «Подготовка данных машинного обучения»**

Направление подготовки/специальность 02.03.02 Фундаментальная информатика и информационные технологии

*(код и наименование направления подготовки/специальности)*

Курс 3 Семестр 6 Количество з.е. 2

**Объем трудоемкости:** 2 зачетных единиц (72 ч., из них – 34,2 час. аудиторной нагрузки: лекционных 16 ч., лабораторных работ - 16 ч., 37,8 часов самостоятельной работы, 2 часов КСР, 0,2 часа ИКР.), форма контроля – зачет.

**Цель освоения дисциплины «Подготовка данных машинного обучения»** является формирование у студентов систематизированных знаний, практических умений и навыков применения современных методов искусственного интеллекта, машинного обучения для решения задач подготовки данных для дальнейшего применения моделях различных предметных областей.

Дисциплина направлена на развитие способности собирать, размечать, преобразовывать данные и оценивать качество подготовленных данных.

#### **Задачи дисциплины**

1. Кроме методов решения типовых задач подготовки данных: обработка пропущенных значений, кодирование категориальных признаков, масштабирование и нормализация числовых данных и методов обработки выбросов (аномалий) в данных, изученных ранее в дисциплине «Многомерный статистический анализ и машинное обучение», усвоить принципы и методы feature engineering (создания и преобразования признаков) для повышения эффективности моделей машинного обучения.

2. Применять на практике методы работы с несбалансированными данными и подходы к разметке данных (labeling).

3. Применять на практике критерии и метрики для оценки качества подготовленных данных и их пригодности для решения конкретной задачи.

4. Приобрести практический навык применения методов предобработки данных: очистка от шума, обработка пропусков, кодирование категориальных переменных, масштабирование признаков. Сформировать умение создавать новые признаки (Feature Engineering) на основе существующих для улучшения предсказательной способности моделей. Освоить методы селекции признаков (Feature Selection) для отбора наиболее информативных переменных и уменьшения размерности данных.

5. Приобрести умение оценивать качество подготовленного набора данных с помощью визуализации и статистических метрик перед передачей его на этап моделирования.

#### **Место дисциплины (модуля) в структуре образовательной программы**

Дисциплина «Подготовка данных машинного обучения» относится к части, формируемой участниками образовательных отношений Блока "Дисциплины (модули) по выбору" учебного плана (Б1.В.ДВ).

Дисциплина изучается в 6-м семестре. Для успешного освоения необходимы знания, полученные в дисциплинах: «Алгебра и введение в тензорный анализ», «Теория вероятностей

и математическая статистика», «Многомерный статистический анализ», и «Машинное обучение», «Программирование».

Преподавание ведется в виде лекций и лабораторных занятий с использованием интерактивных методов. Лабораторные работы направлены на практическое освоение методов и инструментов классификации на реальных данных.

Дисциплина формирует компетенции, необходимые для выполнения выпускной квалификационной работы и профессиональной деятельности в области вычислительных технологий.

**Результаты обучения (знания, умения, опыт, компетенции):**

|              |                                                                                                                                                                                                                       |
|--------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <b>BD-1</b>  | <b>Способен осуществлять поиск, сбор, очистку и предварительный анализ данных</b>                                                                                                                                     |
| BD-1.1       | Обосновывает способы и варианты применения методов предварительного анализа данных в задачах ИИ, включая их математическое (алгоритмическое) преобразование и адаптацию к специфике задачи                            |
| BD-1.2       | Применяет методы анализа данных для проверки разведочных гипотез и подготовки данных к применению современных методов ИИ                                                                                              |
| BD-1.3       | Применяет методы понижения размерности для первичной интерпретации и визуализации многомерных данных                                                                                                                  |
| BD-1.4       | Отбирает признаки данных, значимые для исследования                                                                                                                                                                   |
| <b>BD-2</b>  | <b>Способен определять требования к наборам данных для решения задач машинного обучения, проводить разметку и анализ наборов данных, оценивать качество данных, обеспечивать непрерывную интеграцию данных</b>        |
| BD-2.1       | Определяет требования к наборам и качеству данных для решения задач машинного обучения                                                                                                                                |
| BD-2.2       | Работает с данными, в том числе собирает данные из разрозненных источников, проверяет данные на корректность                                                                                                          |
| <b>LLM-2</b> | <b>Способен дообучать, адаптировать и оптимизировать генеративные модели под специфические задачи и условия применения</b>                                                                                            |
| LLM-2.1      | Понимает принципы fine-tune                                                                                                                                                                                           |
| LLM-2.2      | Создаёт обучающие наборы данных                                                                                                                                                                                       |
| <b>MF-4</b>  | <b>Способен применять статистические методы для анализа данных, валидации моделей машинного обучения и проведения экспериментов в области ИИ</b>                                                                      |
| MF-4.1       | Применяет статистические методы анализа и машинного обучения для решения задач анализа данных и проведения экспериментов на данных.                                                                                   |
| MF-4.2       | Способен применять статистические методы для построения предсказательных моделей, включая методы для анализа и прогнозирования временных рядов, а также моделирования нестационарных случайных процессов.             |
| MF-4.3       | Способен применять статистические методы для оценки качества моделей ИИ, включая метрики и критерии для регрессии, классификации и кластеризации, а также для проведения статистических тестов для сравнения моделей. |
| <b>ML-2</b>  | <b>Способен применять фундаментальные принципы и методы машинного обучения включая подготовку данных оценку качества моделей и работу с признаками</b>                                                                |

|        |                                                                                              |
|--------|----------------------------------------------------------------------------------------------|
| ML-2.1 | Различает основные типы задач машинного обучения и применяет на практике принципы их решения |
|--------|----------------------------------------------------------------------------------------------|

### Содержание и структура дисциплины:

Распределение видов учебной работы и их трудоемкости по разделам дисциплины.

Разделы дисциплины, изучаемые в 6 семестре (очная форма)

| №                                   | Наименование разделов (тем)                                                                                                                                                                                                                                        | Количество часов |                   |    |           |                      |
|-------------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|------------------|-------------------|----|-----------|----------------------|
|                                     |                                                                                                                                                                                                                                                                    | Всего            | Аудиторная работа |    |           | Внеаудиторная работа |
|                                     |                                                                                                                                                                                                                                                                    |                  | Л                 | ПЗ | ЛР        |                      |
| 1.                                  | Введение в подготовку данных для МО. Планирование датасета под задачу.                                                                                                                                                                                             | 8                | 2                 |    | 2         | 4                    |
| 2.                                  | Первичный сбор данных и их оценка. Предобработка данных: очистка, обработка пропусков, выбросов.                                                                                                                                                                   | 8                | 2                 |    | 2         | 4                    |
| 3.                                  | Разметка данных и аугментация.                                                                                                                                                                                                                                     | 8                | 2                 |    | 2         | 4                    |
| 4.                                  | Предобработка изображений                                                                                                                                                                                                                                          | 8                | 2                 |    | 2         | 4                    |
| 5.                                  | Предобработка аудиоданных                                                                                                                                                                                                                                          | 8                | 2                 |    | 2         | 4                    |
| 6.                                  | Предобработка текстовых данных                                                                                                                                                                                                                                     | 9                | 2                 |    | 2         | 5                    |
| 7.                                  | Предобработка данных для временных рядов                                                                                                                                                                                                                           | 8,8              | 2                 |    | 2         | 4,8                  |
| 8.                                  | Статистический анализ данных (EDA, визуализация, анализ распределений, корреляции, анализ признаков, feature importance). Стратегии по улучшению метрик через данные (feature engineering, отбор признаков, балансировка). Дизайн эксперимента по улучшению данных | 11               | 2                 |    | 2         | 7                    |
| <b>ИТОГО по разделам дисциплины</b> |                                                                                                                                                                                                                                                                    | <b>69,8</b>      | <b>16</b>         |    | <b>16</b> | <b>37,8</b>          |
|                                     | Контроль самостоятельной работы (КСР)                                                                                                                                                                                                                              | 2                |                   |    |           |                      |
|                                     | Промежуточная аттестация (ИКР)                                                                                                                                                                                                                                     | 0,2              |                   |    |           |                      |
|                                     | Подготовка к текущему контролю                                                                                                                                                                                                                                     | -                |                   |    |           |                      |
|                                     | Общая трудоемкость по дисциплине                                                                                                                                                                                                                                   | 72               |                   |    |           |                      |

Примечание: Л – лекции, КСР – контрольные и самостоятельные работы, ЛР – лабораторные занятия, СРС – самостоятельная работа студента

### Курсовые проекты или работы.

Не предусмотрены учебным планом

**Вид аттестации:** ЛР, Комплексная итоговая работа, зачет.

Автор Приходько Т.А. – кандидат технических наук, доцент кафедры вычислительных технологий;