

УТВЕРЖДАЮ
Первый проректор, проректор
по УР

_____ А.Е. Рудин

Рабочая программа дисциплины

Б1.В.05

Модели и методы обработки данных

Учебный план: 2026-2027 09.03.03 ИИТА ПИИ ОО №1-1-181.plx

Кафедра: **33** Цифровых и аддитивных технологий

Направление подготовки:
(специальность) 09.03.03 Прикладная информатика

Профиль подготовки: Прикладной искусственный интеллект
(специализация)

Уровень образования: бакалавриат

Форма обучения: очная

План учебного процесса

Семестр (курс для ЗАО)		Контактная работа обучающихся		Сам. работа	Контроль, час.	Трудоёмкость, ЗЕТ	Форма промежуточной аттестации
		Лекции	Практ. занятия				
5	УП	16	64	135,75	0,25	6	Зачет
	РПД	16	64	135,75	0,25	6	
6	УП	17	34	30	27	3	Экзамен
	РПД	17	34	30	27	3	
7	УП		32	49	27	3	Экзамен
	РПД		32	49	27	3	
Итого	УП	33	130	214,75	54,25	12	
	РПД	33	130	214,75	54,25	12	

Рабочая программа дисциплины составлена в соответствии с федеральным государственным образовательным стандартом высшего образования по направлению подготовки 09.03.03 Прикладная информатика, утверждённым приказом Минобрнауки России от 19.09.2017 г. № 922

Составитель (и):

старший преподаватель

Кокорин Евгений Сергеевич

От кафедры составителя:

Заведующий кафедрой цифровых и аддитивных технологий

Сошников Антон
Владимирович

От выпускающей кафедры:

Заведующий кафедрой

Сошников Антон
Владимирович

Методический отдел:

1 ВВЕДЕНИЕ К РАБОЧЕЙ ПРОГРАММЕ ДИСЦИПЛИНЫ

1.1 Цель дисциплины: Сформировать у обучающегося базовые и прикладные компетенции в области обработки, анализа и представления данных, включая навыки работы с современными инструментами и технологиями подготовки и хранения данных, основами генерации, визуализации и распределённой обработки больших объёмов информации.

1.2 Задачи дисциплины:

Изучить этапы обработки данных: получение, очистку, подготовку и первичный анализ;
Ознакомиться с основными типами данных и форматами их хранения (CSV, Excel, JSON и др.);
Научиться использовать современные библиотеки Python (NumPy, Pandas) для обработки и анализа данных;
Освоить приёмы визуализации данных с помощью специализированных инструментов (matplotlib, pandas);
Сформировать навыки генерации и обогащения данных, в том числе для работы с временными рядами;
Освоить основы оценки качества данных и применения основных метрик для анализа полноты и корректности информации;
Получить базовые знания о технологиях распределённой обработки данных, включая MapReduce и Hadoop;
Развить умения самостоятельной работы с инструментами для анализа и визуализации данных в средах Jupyter Notebook и Google Colab;
Выработать навыки командной работы и презентации результатов анализа данных.

1.3 Требования к предварительной подготовке обучающегося:

Предварительная подготовка предполагает создание основы для формирования компетенций, указанных в п. 2, при изучении дисциплин:

Алгоритмизация и программирование

Дизайн-проектирование

Учебная практика (научно-исследовательская работа (получение первичных навыков научно-исследовательской работы))

2 КОМПЕТЕНЦИИ ОБУЧАЮЩЕГОСЯ, ФОРМИРУЕМЫЕ В РЕЗУЛЬТАТЕ ОСВОЕНИЯ ДИСЦИПЛИНЫ

ПК-2: Способен анализировать возможности реализации требований к компьютерному программному обеспечению					
Знать: Методологические подходы к разработке программного обеспечения для обработки данных.					
Уметь: Вырабатывать варианты реализации требований к компьютерному программному обеспечению в зависимости от типа данных и поставленных целей.					
Владеть: Навыками проведения сравнительного анализа вариантов реализации требований к компьютерному программному обеспечению.					
ПК-6: Способен подготавливать данные для проведения аналитических работ по исследованию больших данных					
Знать: Типы больших данных: метаданные; виды источников данных; особенности и возможности применения облачных технологий, облачных сервисов.					
Уметь: Определять этапы работы с данными из широкого круга источников, в том числе в режиме реального времени.					
Владеть: Навыками обеспечения структуры для хранения данных, с последующей обработкой.					
ПК-7: Способен проводить аналитическое исследование с применением технологий больших данных в соответствии с требованиями заказчика					
Знать: Типы больших данных: метаданные; виды источников данных; особенности и возможности применения облачных технологий, облачных сервисов.					
Уметь: Определять этапы работы с данными из широкого круга источников, в том числе в режиме реального времени.					
Владеть: Навыками обеспечения структуры для хранения данных, с последующей обработкой.					

3 РЕЗУЛЬТАТЫ ОБУЧЕНИЯ ПО ДИСЦИПЛИНЕ

Наименование и содержание разделов, тем и учебных занятий	Семестр (курс для ЗАО)	Контактная работа		СР (часы)	Инновац. формы занятий	Форма текущего контроля
		Лек. (часы)	Пр. (часы)			
Раздел 1. Введение в обработку данных	5					О
Тема 1. Источники и категории данных. Понятие пайплайна обработки данных. Практические занятия: Анализ источников, схема пайплайна.		2	8	15,25		
Тема 2. Получение, очистка и подготовка данных. Введение в NumPy и Pandas. Практические занятия: загрузка, очистка, преобразование данных.		2	7	15,25		

Тема 3. Основные типы данных и форматы хранения. Практические занятия: разбор CSV, Excel, JSON		2	7	15,25		Пр
Тема 4. Первичный анализ данных и выявление проблем. Практические занятия: поиск пропусков, базовые статистики		2	7	15		
Тема 5. Введение в инструменты Jupyter Notebook и Google Colab. Практические: работа с блокнотами, импорт данных		2	7	15		
Раздел 2. Основы работы с данными в Python						
Тема 6. Массивы и таблицы в NumPy и Pandas. Практические занятия: базовые операции с массивами, фильтрация		2	7	15		
Тема 7. Агрегация и группировка данных. Практические занятия: группировка, сводные таблицы		2	7	15		
Тема 8. Визуализация и графики (matplotlib, pandas). Практические занятия: построение простых графиков, анализ		1	7	15		
Тема 9. Экспорт и сохранение данных. Практические занятия: сохранение результатов обработки		1	7	15		
Итого в семестре (на курсе для ЗАО)		16	64	135,75		
Консультации и промежуточная аттестация (Зачет)		0,25				
Раздел 3. Генерация и обогащение данных	6					Пр
Тема 10. Генерация синтетических данных и аугментация. Практические занятия: генерация выборок, добавление шума		3	5	5		
Тема 11. Обогащение данных: новые признаки, объединение таблиц. Практические занятия: merge, join, feature engineering		3	5	5		
Тема 12. Основы работы с временными рядами. Практические занятия: разметка временных данных, выделение трендов		3	5	4		
Тема 13. Визуализация временных рядов. Практические занятия: построение графиков временных рядов		2	5	4		
Раздел 4. Анализ качества данных и метрики						Пр
Тема 14. Проверка данных на корректность и полноту. Практические занятия: валидация наборов		2	5	4		
Тема 15. Метрики качества данных: полнота, точность, дубликаты. Практические занятия: оценка метрик на учебных примерах		2	5	4		
Тема 16. Оценка распределений, визуальный контроль качества. Практические занятия: построение гистограмм, анализ распределения		2	4	4		
Итого в семестре (на курсе для ЗАО)		17	34	30		
Консультации и промежуточная аттестация (Экзамен)		2,5		24,5		
Раздел 5. Технологии распределённой обработки данных	7					О

Тема 17. Основы MapReduce: концепция, примеры. Практические занятия: составление схем MapReduce			8	10		
Тема 18. Введение в Hadoop и распределённые файловые системы. Практические занятия: моделирование простого кластера			8	13		
Тема 19. Моделирование MapReduce на практике. Практические занятия: решение учебной задачи			8	13		
Тема 20. Кейсы работы с большими данными. Практические занятия: обсуждение примеров, разбор типовых ошибок			8	13		
Итого в семестре (на курсе для ЗАО)			32	49		
Консультации и промежуточная аттестация (Экзамен)		2,5		24,5		
Всего контактная работа и СР по дисциплине		168,25		263,75		

4 КУРСОВОЕ ПРОЕКТИРОВАНИЕ

Курсовое проектирование учебным планом не предусмотрено

5. ФОНД ОЦЕНОЧНЫХ СРЕДСТВ ДЛЯ ПРОВЕДЕНИЯ ПРОМЕЖУТОЧНОЙ АТТЕСТАЦИИ

5.1 Описание показателей, критериев и системы оценивания результатов обучения

5.1.1 Показатели оценивания

Код компетенции	Показатели оценивания результатов обучения	Наименование оценочного средства
ПК-6	Формулирует понятие большие данные и раскрывает классификацию типов больших данных. Раскрывает основные этапы работы с данными из широкого круга источников. Демонстрирует результат построения структуры для хранения данных.	
ПК-2	Раскрывает список методологий для обработки данных. Формулирует требования к программному обеспечению для обработки данных для достижения поставленных целей. Представляет аналитическое заключение по анализу требований к программному обеспечению.	
ПК-7	Раскрывает сущность основных методов оценки моделей данных. Строит алгоритм для обработки больших данных с использованием современного программного обеспечения. Демонстрирует аналитическое заключение на основании обработки больших объемов данных.	

5.1.2 Система и критерии оценивания

Шкала оценивания	Критерии оценивания сформированности компетенций	
	Устное собеседование	Письменная работа
5 (отлично)	Полный, исчерпывающий ответ, явно демонстрирующий глубокое понимание предмета и широкую эрудицию в оцениваемой области. Критический, оригинальный подход к материалу. Учитываются баллы, накопленные в течение семестра.	Качество исполнения всех элементов задания полностью соответствует всем требованиям. Учитываются баллы, накопленные в течение семестра.
4 (хорошо)	Ответ стандартный, в целом качественный, основан на всех обязательных источниках информации. Присутствуют небольшие пробелы в знаниях или несущественные ошибки. Учитываются баллы, накопленные в течение семестра.	Работа выполнена в соответствии с заданием. Имеются отдельные несущественные ошибки или отступления от правил оформления работы. Учитываются баллы, накопленные в течение семестра.

3 (удовлетворительно)	Ответ неполный, основанный только на лекционных материалах. При понимании сущности предмета в целом – существенные ошибки или пробелы в знаниях сразу по нескольким темам, незнание (путаница) важных терминов. Учитываются баллы, накопленные в течение семестра.	Задание выполнено полностью, но с многочисленными существенными ошибками. При этом нарушены правила оформления или сроки представления работы. Учитываются баллы, накопленные в течение семестра.
2 (неудовлетворительно)	Неспособность ответить на вопрос без помощи экзаменатора. Незнание значительной части принципиально важных элементов дисциплины. Многочисленные грубые ошибки. Не учитываются баллы, накопленные в течение семестра.	Отсутствие одного или нескольких обязательных элементов задания, либо многочисленные грубые ошибки в работе, либо грубое нарушение правил оформления или сроков представления работы. Не учитываются баллы, накопленные в течение семестра.
Зачтено	Отвечает на теоретический вопрос по материалам лекций, возможно допуская несущественные ошибки. Учитываются баллы, накопленные в течение семестра.	Обучающийся своевременно выполнил практические задания. Учитываются баллы, накопленные в течение семестра.
Не зачтено	При ответе на вопрос допускает существенные ошибки. Не учитываются баллы, накопленные в течение семестра.	Своевременно не выполняет (выполнил частично) практические задания. Не учитываются баллы, накопленные в течение семестра.

5.2 Типовые контрольные задания или иные материалы, необходимые для оценки знаний, умений, навыков и (или) опыта деятельности

5.2.1 Перечень контрольных вопросов

№ п/п	Формулировки вопросов
Семестр 5	
1	Какие этапы включает в себя процесс очистки данных?
2	Перечислите основные источники и категории данных.
3	Что такое пайплайн обработки данных? Приведите пример.
4	Назовите распространённые форматы хранения данных.
5	Как загрузить данные из CSV-файла в Pandas DataFrame?
6	В чём особенности формата JSON по сравнению с Excel?
7	Назовите распространённые форматы хранения данных.
8	Как загрузить данные из CSV-файла в Pandas DataFrame?
9	В чём особенности формата JSON по сравнению с Excel?
10	Как осуществляется фильтрация данных в Pandas?
11	Какие функции Pandas позволяют группировать данные?
12	Как построить простой график по данным с помощью matplotlib?
13	В чём разница между Series и DataFrame в Pandas?
14	Как сохранить обработанные данные обратно в файл?
15	Какую функцию в Pandas использовать для объединения таблиц?
16	Какие инструменты используются для работы с ноутбуками?
17	Какие шаги включает первичный анализ данных?
18	Для чего используется Jupyter Notebook?
19	Как импортировать библиотеку Pandas в Python?
20	Каковы основные преимущества использования Google Colab?
Семестр 6	
21	Для чего используется генерация синтетических данных?
22	В каких случаях применяется аугментация данных?
23	Как создать новый признак (feature) в Pandas DataFrame?
24	В чём заключается объединение данных с помощью merge?
25	Какие задачи решаются при анализе временных рядов?
26	Как разметить временные данные для анализа?
27	Какие методы позволяют выявить тренды во временных рядах?
28	Как построить график временного ряда?

29	Почему важно проверять корректность и полноту данных?
30	Какие бывают метрики качества данных?
31	В чём отличие полноты и точности как метрик?
32	Как определить наличие дубликатов в наборе данных?
33	Как проверить распределение данных?
34	Какие графики используются для визуального контроля качества?
35	Чем полезна гистограмма для анализа данных?
36	Какие шаги включает в себя процесс валидации данных?
37	Что такое feature engineering?
38	Как осуществляется объединение нескольких DataFrame?
39	Почему важно анализировать пропуски во временных рядах?
40	Какие ошибки могут возникать при работе с временными рядами?
Семестр 7	
41	Объясните принцип работы MapReduce.
42	В чём преимущества распределённой обработки данных?
43	Какова основная задача этапа Map в MapReduce?
44	Для чего используется этап Reduce?
45	Что такое кластер Hadoop?
46	Как устроена распределённая файловая система HDFS?
47	Какие задачи решаются с помощью Hadoop?
48	Какие форматы данных поддерживаются в Hadoop?
49	Как смоделировать работу MapReduce на простом примере?
50	Какие существуют типовые ошибки при работе с большими данными?
51	Как организовать хранение больших массивов данных?
52	Для чего нужны распределённые вычисления?
53	Какие существуют типовые ошибки при работе с большими данными?
54	Как организовать хранение больших массивов данных?
55	Для чего нужны распределённые вычисления?
56	Как анализировать результат выполнения MapReduce?
57	Какие существуют способы оптимизации производительности при работе с большими данными?
58	В чём разница между централизованной и распределённой обработкой?
59	Какова роль мастер-узла в кластере Hadoop?
60	Какие программные инструменты используются для распределённого анализа данных?
61	Как осуществляется мониторинг и отладка процессов в распределённой системе?

5.2.2 Типовые тестовые задания

Не предусмотрено

5.2.3 Типовые практико-ориентированные задания (задачи, кейсы)

Собрать и загрузить набор данных из открытого источника, выполнить его очистку и первичный анализ (выявить пропуски, удалить дубликаты).

Провести агрегацию и группировку данных с помощью библиотеки Pandas, представить результаты в виде таблицы и диаграммы.

Выполнить визуализацию набора данных с помощью matplotlib и Pandas: построить графики, гистограммы и круговые диаграммы.

Сгенерировать синтетические данные для учебной задачи, обогатить их новыми признаками и сохранить результат в файл.

Провести анализ и визуализацию временного ряда: выделить тренды, построить соответствующие графики.

Разработать пайплайн обработки данных в Jupyter Notebook или Google Colab для решения прикладной задачи (например, анализ результатов опроса).

Оценить качество данных, рассчитать основные метрики (полнота, точность, количество дубликатов), оформить результаты анализа в виде отчёта.

Смоделировать работу алгоритма MapReduce для обработки большого объёма данных на примере учебного датасета.

Выполнить интеграцию и объединение данных из двух разных источников, выявить и устранить несоответствия.

Подготовить презентацию (или мини-отчёт) с визуализацией результатов анализа данных и выводами.

5.3 Методические материалы, определяющие процедуры оценивания знаний, умений, владений (навыков и (или) практического опыта деятельности)

5.3.1 Условия допуска обучающегося к промежуточной аттестации и порядок ликвидации академической задолженности

Проведение промежуточной аттестации регламентировано локальным нормативным актом СПбГУПТД «Положение о проведении текущего контроля успеваемости и промежуточной аттестации обучающихся»

5.3.2 Форма проведения промежуточной аттестации по дисциплине

Устная ☐ + Письменная ☐ Компьютерное тестирование ☐ Иная ☐

5.3.3 Особенности проведения промежуточной аттестации по дисциплине

зачета

- время на подготовку к устному собеседованию составляет 15 минут;
- выполнение кейс-задания осуществляется на компьютере за 60 минут.

экзамена

- время на подготовку к устному собеседованию составляет 30 минут;
- выполнение кейс-задания осуществляется на компьютере за 60 минут.

6. УЧЕБНО-МЕТОДИЧЕСКОЕ И ИНФОРМАЦИОННОЕ ОБЕСПЕЧЕНИЕ ДИСЦИПЛИНЫ

6.1 Учебная литература

Автор	Заглавие	Издательство	Год издания	Ссылка
6.1.1 Основная учебная литература				
Катунин, Г. П.	Основы мультимедийных технологий	Москва: Ай Пи Ар Медиа	2020	https://www.iprbooks.hop.ru/93614.html
Поляков, Е. А.	Web-дизайн	Саратов: Вузовское образование	2019	http://www.iprbookshop.ru/81868.html
6.1.2 Дополнительная учебная литература				
Поляков, Е. А.	Web-дизайн	Саратов: Вузовское образование	2019	http://www.iprbookshop.ru/81869.html
Немцова Т.И., Казанкова Т.В., Шнякин А.В. и др.	Компьютерная графика и web-дизайн	Москва: Форум	2022	https://ibooks.ru/reading.php?short=1&productid=361314

6.2 Перечень профессиональных баз данных и информационно-справочных систем

Информационная система «Единое окно доступа к образовательным ресурсам. Раздел. Информатика и информационные технологии» [Электронный ресурс]. URL: http://window.edu.ru/catalog/?p_rubr=2.2.75.6
Электронно-библиотечная система IPRbooks [Электронный ресурс]. URL: <http://www.iprbookshop.ru/>

6.3 Перечень лицензионного и свободно распространяемого программного обеспечения

MicrosoftOfficeProfessional

Microsoft Windows

Adobe Audition CC ALL Multiple Platforms Multi European Languages Team LicSub Level 4 (100+) Education

Device license

Figma

6.4 Описание материально-технической базы, необходимой для осуществления образовательного процесса по дисциплине

Аудитория	Оснащение
Компьютерный класс	Мультимедийное оборудование, компьютерная техника с возможностью подключения к сети «Интернет» и обеспечением доступа в электронную информационно-образовательную среду
Лекционная аудитория	Мультимедийное оборудование, специализированная мебель, доска