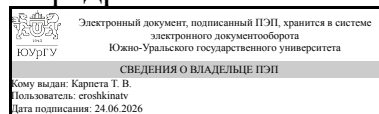


ЮЖНО-УРАЛЬСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ

УТВЕРЖДАЮ:
Заведующий выпускающей
кафедрой



Т. В. Карпета

РАБОЧАЯ ПРОГРАММА

дисциплины 1.Ф.М0.01 Предобработка и анализ данных
для направления 02.04.02 Фундаментальная информатика и информационные технологии

уровень Магистратура

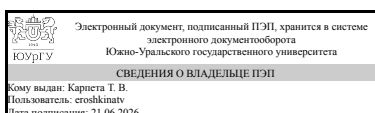
магистерская программа Инновационные методы и технологии машинного обучения и анализа данных

форма обучения очная

кафедра-разработчик Прикладная математика и программирование

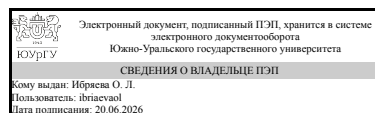
Рабочая программа составлена в соответствии с ФГОС ВО по направлению подготовки 02.04.02 Фундаментальная информатика и информационные технологии, утверждённым приказом Минобрнауки от 23.08.2017 № 811

Зав.кафедрой разработчика,
к.физ.-мат.н.



Т. В. Карпета

Разработчик программы,
д.техн.н., доцент



О. Л. Ибряева

1. Цели и задачи дисциплины

Целью освоения дисциплины является формирование системных знаний и практических навыков предобработки, очистки и анализа данных различной природы для последующего применения в задачах машинного обучения и анализа данных. В ходе изучения дисциплины студенты осваивают методы разведочного анализа, обработки пропусков и выбросов, кодирования и масштабирования признаков, работы с временными рядами и текстовыми данными, а также техники балансировки несбалансированных выборок. Особое внимание уделяется пониманию математических основ алгоритмов предобработки и их влияния на качество моделей машинного обучения. Дисциплина сочетает теоретическую подготовку с выполнением лабораторных работ на реальных наборах данных, что обеспечивает формирование компетенций, востребованных в промышленной аналитике и научных исследованиях. По итогам освоения курса студент способен самостоятельно проводить полный цикл подготовки данных от первичного исследования до формирования признакового пространства, готового к использованию в моделях искусственного интеллекта.

Краткое содержание дисциплины

Дисциплина охватывает полный цикл предобработки данных, начиная с разведочного анализа и статистического описания признаков, включая визуальный поиск аномалий и анализ распределений. Изучаются методы обработки пропусков различных типов (MCAR, MAR, MNAR), обнаружения и обработки выбросов. Рассматриваются техники кодирования категориальных признаков, масштабирования числовых переменных и трансформации распределений для устранения асимметрии. Значительная часть курса посвящена анализу временных рядов: исследованию стационарности, автокорреляционной структуры, построению моделей ARIMA и SARIMA, а также альтернативным представлениям рядов в виде рекуррентных изображений, угловых полей Грамиана и марковских переходных полей. Отдельные темы посвящены конструированию и автоматизированному извлечению признаков, включая полиномиальное расширение и использование библиотек TSFEL и tsfresh. Изучаются методы предобработки текстовых данных: токенизация, лемматизация, построение обратного индекса и векторизация с использованием TF-IDF. Завершается курс рассмотрением проблем дисбаланса классов и методов балансировки выборок, включая алгоритмы SMOTE и ADASYN.

2. Компетенции обучающегося, формируемые в результате освоения дисциплины

Планируемые результаты освоения ОП ВО (компетенции)	Планируемые результаты обучения по дисциплине
ПК-1 Способен разрабатывать и применять алгоритмы анализа данных, машинного обучения и нейросетевых технологий при создании интеллектуальных информационных систем	Знает: современные архитектуры глубоких нейронных сетей, их внутреннюю структуру, особенности обучения и регуляризации; основные методы машинного обучения, подходы к построению адаптивных моделей и их оценки в прикладных задачах Умеет: подбирать и настраивать архитектуры глубоких сетей для решения задач различного

	типа, интерпретировать полученные результаты; строить адаптивные модели машинного обучения, комбинировать методы машинного обучения при решении прикладных задач Имеет практический опыт: применения современных архитектур глубокого обучения для решения прикладных задач и оценки качества обученных моделей; использования современных инструментальных средств и систем программирования для разработки новых методов и моделей машинного обучения
--	---

3. Место дисциплины в структуре ОП ВО

Перечень предшествующих дисциплин, видов работ учебного плана	Перечень последующих дисциплин, видов работ
Нет	Основы глубокого обучения, Инжиниринг архитектур глубокого обучения, Методы и технологии машинного обучения

Требования к «входным» знаниям, умениям, навыкам студента, необходимым при освоении данной дисциплины и приобретенным в результате освоения предшествующих дисциплин:

Нет

4. Объём и виды учебной работы

Общая трудоемкость дисциплины составляет 3 з.е., 108 ч., 38,25 ч. контактной работы

Вид учебной работы	Всего часов	Распределение по семестрам в часах
		Номер семестра
		1
Общая трудоёмкость дисциплины	108	108
<i>Аудиторные занятия:</i>	32	32
Лекции (Л)	16	16
Практические занятия, семинары и (или) другие виды аудиторных занятий (ПЗ)	0	0
Лабораторные работы (ЛР)	16	16
<i>Самостоятельная работа (СРС)</i>	69,75	69,75
Решение индивидуальных заданий по теме Инженерия признаков, анализ сложных структур данных и балансировка	15	15
Решение индивидуальных заданий по теме Статистические основы предобработки табличных данных	15	15
Изучить встроенные структуры данных, функции и файлы в Python	10	10
Подготовка к зачету	19,75	19.75
Изучить пример выполнения разведочного анализа данных на	10	10

языке Python для набора данных Ирисы Фишера		
Консультации и промежуточная аттестация	6,25	6,25
Вид контроля (зачет, диф.зачет, экзамен)	-	зачет

5. Содержание дисциплины

№ раздела	Наименование разделов дисциплины	Объем аудиторных занятий по видам в часах			
		Всего	Л	ПЗ	ЛР
1	Статистические основы предобработки табличных данных	12	6	0	6
2	Анализ временных рядов и инженерия признаков	12	6	0	6
3	Специализированные методы и балансировка выборок	8	4	0	4

5.1. Лекции

№ лекции	№ раздела	Наименование или краткое содержание лекционного занятия	Кол-во часов
1	1	Разведочный анализ данных Этапы жизненного цикла анализа данных (CRISP-DM). Типы данных и методы первичного исследования набора данных. Статистические характеристики признаков, визуальный поиск аномалий, анализ распределений и взаимосвязей между переменными.	2
2	1	Очистка данных, обработка пропусков и выбросов Типы пропусков MCAR, MAR и MNAR. Методы удаления и восстановления пропусков, включая статистические методы и алгоритмы многомерной импутации. Статистические и алгоритмические методы обнаружения выбросов и аномалий. Особенности многомерного поиска аномалий, расстояние Махаланобиса.	2
3	1	Кодирование, масштабирование и преобразование признаков Методы кодирования категориальных признаков и статистические подходы для работы с категориями высокой кардинальности. Масштабирование признаков и методы трансформации распределений для устранения асимметрии. Риск утечки данных.	2
4	2	Временные ряды: анализ, предобработка и прогнозирование Особенности временных рядов, тренд, сезонность, циклические компоненты и автокорреляции. Методы сглаживания, детрендирования и выделения сезонных компонент. Лаговые признаки, скользящие статистики, методы проверки стационарности временных рядов. Основы классических моделей прогнозирования ARIMA и SARIMA.	2
5	2	Извлечение признаков и альтернативные представления временных рядов Методы извлечения статистических и спектральных характеристик временных рядов, спектрограммы, мел-спектрограммы и скалограммы. Методы преобразования временных рядов в двумерные представления (Recurrence Plot, Gramian Angular Field, Markov Transition Field).	2
6	2	Конструирование и отбор признаков Методы создания новых признаков на основе исходных данных. Автоматизированное извлечение признаков из временных рядов с использованием специализированных библиотек (TSFEL, tsfresh). Методы оценки информативности и отбора признаков. Проблема мультиколлинеарности, методы её обнаружения и устранения.	2
7	3	Предобработка текстовых данных и извлечение текстовых признаков Методы очистки и нормализации текстов, токенизация, удаление стоп-слов, стемминг	2

		и лемматизация. Классические методы представления текстов в виде признаков: Bag-of-Words, N-граммы и TF-IDF. Особенности подготовки текстовых данных для современных языковых моделей и систем искусственного интеллекта.	
8	3	Дисбаланс данных и балансировка выборок Проблемы несбалансированных наборов данных и их влияние на анализ данных и обучение моделей. Методы оценки дисбаланса классов. Подходы к балансировке выборок, генерация синтетических объектов (SMOTE, ADASYN). Особенности применения методов балансировки и риск искажения структуры данных.	2

5.2. Практические занятия, семинары

Не предусмотрены

5.3. Лабораторные работы

№ занятия	№ раздела	Наименование или краткое содержание лабораторной работы	Кол-во часов
1	1	Разведочный анализ данных (EDA) — описательная статистика, визуализация распределений, поиск аномалий	2
2	1	Обработка пропусков и выбросов — методы импутации, сравнение правил Тьюки и 3σ	2
3	1	Статистические основы предобработки табличных данных	2
4	2	Анализ временных рядов — стационарность, ACF/PACF, дифференцирование, ARIMA	2
5	2	Альтернативные представления временных рядов — GAF, RP, MTF, спектрограммы	2
6	2	Конструирование и отбор признаков — трансформации, полиномиальное расширение, методы отбора	2
7	3	Предобработка текстовых данных — токенизация, TF-IDF, обратный индекс	2
8	3	Инженерия признаков, анализ сложных структур данных и балансировка	2

5.4. Самостоятельная работа студента

Выполнение СРС			
Подвид СРС	Список литературы (с указанием разделов, глав, страниц) / ссылка на ресурс	Семестр	Кол-во часов
Решение индивидуальных заданий по теме Инженерия признаков, анализ сложных структур данных и балансировка	Приложенное пособие "Инженерия признаков, анализ сложных структур данных и балансировка"	1	15
Решение индивидуальных заданий по теме Статистические основы предобработки табличных данных	Приложенное пособие "Сборник индивидуальных заданий по теме Статистические основы предобработки табличных данных"	1	15
Изучить встроенные структуры данных, функции и файлы в Python	Маккинни, У. Python и анализ данных. Первичная обработка данных с применением pandas, NumPy и Jupiter : справочник / У. Маккинни ; перевод с английского А. А. Слинкина. — 3-е изд. — Москва : ДМК Пресс, 2023. — 536 с.	1	10

	— ISBN 978-5-93700-174-0. — Текст : электронный // Лань : электронно-библиотечная система. — URL: https://e.lanbook.com/book/348086 (дата обращения: 19.06.2026). — Режим доступа: для авториз. пользователей. С. 71-104		
Подготовка к зачету	Приложенные материалы для подготовки к зачету (список теоретических вопросов, примеры задач)	1	19,75
Изучить пример выполнения разведочного анализа данных на языке Python для набора данных Ирисы Фишера	Демидова, Л. А. Разведочный анализ данных. Python : учебно-методическое пособие / Л. А. Демидова. — Москва : РТУ МИРЭА, 2022 — Часть 1 — 2022. — 107 с. — Текст : электронный // Лань : электронно-библиотечная система. — URL: https://e.lanbook.com/book/310970 (дата обращения: 19.06.2026). — Режим доступа: для авториз. пользователей. С. 69 - 88	1	10

6. Фонд оценочных средств для проведения текущего контроля успеваемости, промежуточной аттестации

Контроль качества освоения образовательной программы осуществляется в соответствии с Положением о балльно-рейтинговой системе оценивания результатов учебной деятельности обучающихся.

6.1. Контрольные мероприятия (КМ)

№ КМ	Се-местр	Вид контроля	Название контрольного мероприятия	Вес	Макс. балл	Порядок начисления баллов	Учитывается в ПА
1	1	Текущий контроль	Лабораторная работа №1. Разведочный анализ данных	1	10	Корректность расчёта описательных статистик и квартилей — 2 балла. Качество, разнообразие и подписанность визуализаций — 2 балла. Правильность применения правил Тьюки и трёх сигм — 2 балла. Сравнение методов обнаружения выбросов с обоснованием расхождений — 2 балла. Качество выводов, интерпретация корреляций и формулировка следующих шагов — 2 балла.	зачет
2	1	Текущий контроль	Лабораторная работа №2. Обработка пропусков	1	10	Качество анализа пропусков, классификация по типам MCAR/MAR/MNAR и визуализация паттернов — 3 балла. Корректность применения минимум трёх стратегий импутации для числовых и двух для категориальных признаков — 3 балла.	зачет

						Сравнение влияния разных стратегий на распределения признаков с визуализацией — 2 балла. Оценка влияния импутации на взаимосвязи между признаками и качество итоговых рекомендаций — 2 балла.	
3	1	Текущий контроль	Лабораторная работа №3. Анализ временных рядов	1	10	Качество визуализации и интерпретация структуры временного ряда — 2 балла. Корректность расчёта ACF/PACF и применения теста стационарности — 2 балла. Правильность выполнения преобразований (дифференцирование, Бокс-Кокс) и оценка их эффекта — 2 балла. Обоснованность выбора параметров модели ARIMA/SARIMA — 2 балла. Качество прогноза, анализ остатков и итоговые выводы — 2 балла.	зачет
4	1	Текущий контроль	Лабораторная работа №4. Альтернативные представления временных рядов	1	10	Корректность построения минимум двумерных представлений — 3 балла. Качество построения и интерпретация спектрограмм и мел-спектрограмм — 2 балла. Интерпретация всех полученных изображений — 2 балла. Сравнительный анализ методов с обоснованием их применимости — 2 балла. Качество выводов о частотной структуре сигнала — 1 балл.	зачет
5	1	Текущий контроль	Лабораторная работа №5. Конструирование и отбор признаков	1	10	Корректность применения математических трансформаций и полиномиального расширения — 2 балла. Качество автоматизированного извлечения признаков из временных рядов — 2 балла. Применение и сравнение методов отбора признаков — 3 балла. Анализ мультиколлинеарности и итоговый набор признаков — 2 балла. Качество выводов и обоснованность рекомендаций — 1 балл.	зачет
6	1	Текущий контроль	Лабораторная работа №6. Предобработка текстовых данных	1	10	Корректность выполнения нормализации текста (токенизация, удаление стоп-слов, лемматизация) — 3 балла. Правильность расчёта весов TF-IDF для корпуса и запроса — 4 балла. Вычисление косинусного сходства и определение наиболее релевантного документа — 2 балла. Качество выводов и интерпретация весов терминов — 1 балл.	зачет
7	1	Текущий контроль	Индивидуальные задачи 1	1	10	Максимальный балл за работу составляет 10. Один балл начисляется за корректное определение квартилей и межквартильного размаха, один балл — за вычисление среднего значения и дисперсии, один балл — за вычисление	зачет

					стандартного отклонения и анализ полученных характеристик. Один балл начисляется за применение правила Тьюки, один балл — за применение правила 3σ и сравнение результатов двух методов. Один балл начисляется за обработку пропуска в числовом признаке, один балл — за выполнение нормализации и обработку пропуска в категориальном признаке. Один балл начисляется за корректное выполнение Label Encoding и One-Hot Encoding. Один балл начисляется за вычисление коэффициента корреляции Пирсона и анализ влияния выброса на результат, ещё один балл — за содержательную интерпретацию полученных результатов и ответ на индивидуальный вопрос варианта. При наличии арифметических ошибок балл за соответствующий элемент может быть снижен частично. Использование неверного метода решения или некорректной формулы приводит к неначислению балла за соответствующий элемент.		
8	1	Текущий контроль	Индивидуальные задачи 2	1	10	Максимальный балл за работу составляет 10. Один балл начисляется за корректный расчёт коэффициентов автокорреляции и выполнение сезонного дифференцирования. Один балл — за обоснованный выбор параметров p, q, P, Q по графикам ACF/PACF. Один балл — за применение преобразования Бокса–Кокса и обратное восстановление прогноза в исходную шкалу. Один балл — за корректное построение матрицы альтернативного представления (GAF, MTF или RP) с указанием всех промежуточных этапов. Один балл — за применение математической трансформации признака и сравнение гистограмм до и после преобразования. Один балл — за корректное выполнение полиномиального расширения и объяснение влияния новых признаков на модель. Один балл — за расчёт весов TF-IDF и косинусной близости. Один балл — за определение наиболее релевантного документа с кратким обоснованием. Один балл — за расчёт коэффициента дисбаланса IR и генерацию синтетического объекта SMOTE. Один балл — за содержательные пояснения в рамках	зачет

						расчётных задач (почему лаг 3 специфичен, как изменилась автокорреляция, почему выбранный документ релевантен и т.п.). При наличии арифметических ошибок балл за соответствующий элемент может быть снижен частично. Использование неверного метода решения или некорректной формулы приводит к неначислению балла за соответствующий элемент.	
9	1	Текущий контроль	Контрольная работа 1	1	10	Каждый вариант оценивается максимум в 10 баллов. Распределение баллов указано в каждом задании.	зачет
10	1	Текущий контроль	Контрольная работа 2	1	10	Каждый вариант — максимум 10 баллов. Задание 1 (4 балла): глубокое понимание одной из ключевых тем (временные ряды, тексты, признаки, дисбаланс, полиномы или интерпретация). Задание 2 (3 балла): корректное применение метода представления ряда с выбором метода на усмотрение студента. Задание 3 (3 балла): точное выполнение технической задачи (Бокс–Кокс, SMOTE, полиномы).	зачет
11	1	Промежуточная аттестация	Зачет	-	100	Распределение баллов по заданиям указано непосредственно в тексте билета: теоретический вопрос оценивается в 40 баллов, каждая из двух практических задач — в 30 баллов. При наличии арифметических ошибок балл за соответствующий элемент снижается частично. Использование неверной формулы или метода приводит к неначислению балла за соответствующий этап. Отсутствие пояснений и обоснований в теоретическом вопросе снижает балл на 30–50%. Отсутствие пошаговых выкладок в практических задачах снижает балл на 50% за соответствующий элемент.	зачет

6.2. Процедура проведения, критерии оценивания

Вид промежуточной аттестации	Процедура проведения	Критерии оценивания
зачет	Зачет проводится в билетной форме, каждый билет формируется путем случайной комбинации одного теоретического вопроса из перечня и двух практических задач из подборки, охватывающих различные темы курса. Студент получает билет и выполняет задания в письменной форме в течение 90 минут, при этом теоретический вопрос требует	В соответствии с пп. 2.5, 2.6 Положения

	развернутого ответа с обоснованиями, а практические задачи — пошаговых выкладок и интерпретации результатов. Итоговая оценка выставляется на основе набранных баллов по шкале: 90–100 баллов — отлично, 75–89 — хорошо, 60–74 — удовлетворительно, менее 60 — неудовлетворительно. При наличии арифметических ошибок или отсутствии пояснений балл за соответствующий элемент снижается частично согласно критериям оценивания.	
--	---	--

6.3. Паспорт фонда оценочных средств

Компетенции	Результаты обучения	№ КМ										
		1	2	3	4	5	6	7	8	9	10	11
ПК-1	Знает: современные архитектуры глубоких нейронных сетей, их внутреннюю структуру, особенности обучения и регуляризации; основные методы машинного обучения, подходы к построению адаптивных моделей и их оценки в прикладных задачах	+			+			+		+		+
ПК-1	Умеет: подбирать и настраивать архитектуры глубоких сетей для решения задач различного типа, интерпретировать полученные результаты; строить адаптивные модели машинного обучения, комбинировать методы машинного обучения при решении прикладных задач		+		+			+	+	+	+	+
ПК-1	Имеет практический опыт: применения современных архитектур глубокого обучения для решения прикладных задач и оценки качества обученных моделей; использования современных инструментальных средств и систем программирования для разработки новых методов и моделей машинного обучения			+			+		+		+	+

Типовые контрольные задания по каждому мероприятию находятся в приложениях.

7. Учебно-методическое и информационное обеспечение дисциплины

Печатная учебно-методическая документация

а) основная литература:

- Бендат Д. С. Применение корреляционного и спектрального анализа : пер. с англ. / Д. С. Бендат, А. Д. Пирсол. - М. : Мир, 1983. - 312 с.

б) дополнительная литература:

- Бендат Д. С. Прикладной анализ случайных данных / Пер. с англ. В. Е. Привольского, А. И. Кочубинского; Под ред. И. Н. Коваленко. - М. : Мир, 1989. - 540 с. : ил.
- Седжвик Р. Программирование на языке Python : учеб. курс ; пер. с англ. / Р. Седжвик, К. Уэйн, Р. Дондеро. - М.; СПб. : Диалектика, 2019. - 736 с. : ил.

в) отечественные и зарубежные журналы по дисциплине, имеющиеся в библиотеке:

Не предусмотрены

г) методические указания для студентов по освоению дисциплины:

- Сборник индивидуальных заданий по теме Инженерия признаков, анализ сложных структур данных и балансировка

2. Сборник индивидуальных заданий по теме Статистические основы предобработки табличных данных

из них: учебно-методическое обеспечение самостоятельной работы студента:

1. Сборник индивидуальных заданий по теме Инженерия признаков, анализ сложных структур данных и балансировка
2. Сборник индивидуальных заданий по теме Статистические основы предобработки табличных данных

Электронная учебно-методическая документация

№	Вид литературы	Наименование ресурса в электронной форме	Библиографическое описание
1	Основная литература	ЭБС издательства Лань	Демидова, Л. А. Разведочный анализ данных. Python : учебно-методическое пособие / Л. А. Демидова. — Москва : РТУ МИРЭА, 2022 — Часть 1 — 2022. — 107 с. — Текст : электронный // Лань : электронно-библиотечная система. — URL: https://e.lanbook.com/book/310970 (дата обращения: 19.06.2026). — Режим доступа: для авториз. пользователей.
2	Основная литература	ЭБС издательства Лань	Демидова, Л. А. Разведочный анализ данных. Python : учебно-методическое пособие / Л. А. Демидова. — Москва : РТУ МИРЭА, 2023 — Часть 2 — 2023. — 92 с. — ISBN 978-5-7339-1933-1. — Текст : электронный // Лань : электронно-библиотечная система. — URL: https://e.lanbook.com/book/382691 (дата обращения: 19.06.2026). — Режим доступа: для авториз. пользователей.
3	Основная литература	ЭБС издательства Лань	Кизбикенов, К. О. Прогнозирование и временные ряды : учебное пособие / К. О. Кизбикенов. — Барнаул : АлтГПУ, 2017. — 115 с. — ISBN 978-5-88210-869-3. — Текст : электронный // Лань : электронно-библиотечная система. — URL: https://e.lanbook.com/book/112174 (дата обращения: 19.06.2026). — Режим доступа: для авториз. пользователей.
4	Основная литература	ЭБС издательства Лань	Пшехотская, Е. А. Информационный поиск : учебное пособие / Е. А. Пшехотская, Ю. Н. Филиппович. — Москва : Московский Политех, 2023. — 108 с. — ISBN 978-5-2760-2830-9. — Текст : электронный // Лань : электронно-библиотечная система. — URL: https://e.lanbook.com/book/482786 (дата обращения: 19.06.2026). — Режим доступа: для авториз. пользователей.
5	Основная литература	ЭБС издательства Лань	Маккинни, У. Python и анализ данных. Первичная обработка данных с применением pandas, NumPy и Jupiter : справочник / У. Маккинни ; перевод с английского А. А. Слинкина. — 3-е изд. — Москва : ДМК Пресс, 2023. — 536 с. — ISBN 978-5-93700-174-0. — Текст : электронный // Лань : электронно-библиотечная система. — URL: https://e.lanbook.com/book/348086 (дата обращения: 19.06.2026). — Режим доступа: для авториз. пользователей.

Перечень используемого программного обеспечения:

Нет

Перечень используемых профессиональных баз данных и информационных справочных систем:

Нет

8. Материально-техническое обеспечение дисциплины

Не предусмотрено