МИНИСТЕРСТВО ТРАНСПОРТА РОССИЙСКОЙ ФЕДЕРАЦИИ

ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ

«РОССИЙСКИЙ УНИВЕРСИТЕТ ТРАНСПОРТА»
(РУТ (МИИТ)

Рабочая программа дисциплины (модуля),
как компонент образовательной программы
базового высшего образования
по направлению подготовки
09.03.02 Информационные системы и технологии,
утвержденной первым проректором РУТ (МИИТ)
Тимониным В.С.

РАБОЧАЯ ПРОГРАММА ДИСЦИПЛИНЫ (МОДУЛЯ)

Анализ данных

Направление подготовки: 09.03.02 Информационные системы и технологии

Направленность (профиль): Технологии искусственного интеллекта в транспортных системах

Форма обучения: Очная

1. Общие сведения о дисциплине (модуле).

Дисциплина «Анализ данных» формирует критически важный инженерный компетенс в конвейере машинного обучения – способность превращать сырые массивы телеметрической и инфраструктурной информации в качественные математические объекты. В условиях жесткого импортозамещения и цифровизации транспортной отрасли РФ рынок испытывает острый дефицит специалистов, способных проводить глубокий разведочный анализ, устранять аномалии и конструировать информативные признаки без использования проприетарных облачных AutoML-платформ. Студент погружается в реальный производственный процесс предиктивной аналитики, работая с зашумленными временными рядами и полуструктурированными логами. На практике осваивается полный цикл прикладного анализа данных – от первичного профилирования и статистической проверки гипотез до генерации признаков и оформления инженерного паспорта данных. Выпускник получает подтвержденный портфель воспроизводимых решений, что делает его конкурентоспособным кандидатом на позиции аналитика данных в ведущих транспортных и ИТ-компаниях.

Целью освоения дисциплины является формирование у обучающихся системных теоретических знаний и прикладных инженерных умений в области разведочного анализа данных и конструирования признаков для обеспечения высокого качества входной информации, необходимой для обучения моделей искусственного интеллекта в транспортных системах.

Для достижения поставленной цели в рамках дисциплины решается комплекс задач, направленных на формирование у обучающихся способности: осуществлять комплексное профилирование и аудит качества многомерных массивов данных, применять статистически обоснованные стратегии импутации пропусков и фильтрации аномалий, конструировать многомерные визуализации для выявления скрытых закономерностей, выполнять корреляционный анализ и проверку статистических гипотез, генерировать производные признаки с жестким соблюдением требования отсутствия утечек данных, а также оформлять результаты анализа в виде воспроизводимого интерактивного отчета и технического паспорта данных.

2. Планируемые результаты обучения по дисциплине (модулю).

Перечень формируемых результатов освоения образовательной программы (компетенций) в результате обучения по дисциплине (модулю):

ПК-7 - Способен осуществлять сбор, подготовку, разметку и анализ данных для обучения моделей искусственного интеллекта.

Обучение по дисциплине (модулю) предполагает, что по его результатам обучающийся будет:

Знать:

- концептуальные основы разведочного анализа данных и его критическую роль в конвейере машинного обучения;
- математический аппарат и программные методы автоматизированного профилирования многомерных массивов данных;
- классификацию механизмов возникновения пропущенных значений и методы их статистической идентификации;
- теоретические основы обнаружения выбросов и аномалий в одномерных и многомерных распределениях;
- алгоритмы и стратегии импутации пропущенных значений с учетом временной и физической природы транспортных данных;
- методы фильтрации шума и математического сглаживания временных рядов телеметрии;
- принципы когнитивной визуализации и грамматику графиков для исследования многомерных пространств;
- типологию статистических визуализаций в контексте прикладных транспортных задач;
- математический аппарат оценки корреляционных связей и границы их применимости;
- концепцию проверки статистических гипотез и критерии значимости при анализе эмпирических выборок;
- феномен ложной корреляции и методы контроля смешивающих переменных;
- теоретические основы конструирования признаков и его влияние на обобщающую способность моделей;
- математические методы генерации лаговых переменных, скользящих окон и временных агрегаций;
- принципы извлечения признаков из полуструктурированных данных;
- концепцию утечки данных и архитектурные паттерны ее предотвращения при трансформации датасетов;
- методы оценки важности признаков и алгоритмы устранения мультиколлинеарности;
- принципы снижения размерности данных и сохранения информативности матрицы объектов;

- стандарты оформления инженерной документации и принципы абсолютной воспроизводимости аналитического кода.

Уметь:

- уметь проводить комплексное профилирование и аудит качества многомерных массивов данных при помощи инструментов Pandas и специализированных библиотек EDA в условиях работы с зашумленными телеметрическими потоками;

- уметь применять статистически обоснованные стратегии импутации пропусков и фильтрации аномалий при помощи методов NumPy и SciPy с учетом физической природы датчиков и временных меток;

- уметь конструировать многомерные визуализации для выявления скрытых закономерностей при помощи Matplotlib и Seaborn в рамках этапа разведочного анализа транспортных данных;

- уметь выполнять корреляционный анализ и проверку статистических гипотез при помощи параметрических и непараметрических тестов для отсева ложных зависимостей в многомерных выборках;

- уметь генерировать производные признаки при помощи векторных операций Pandas с жестким соблюдением требования отсутствия утечек данных;

- уметь осуществлять отбор информативных признаков и устранение мультиколлинеарности при помощи методов оценки важности и матриц корреляций для подготовки финальной матрицы объектов;

- уметь оформлять результаты анализа в виде воспроизводимого интерактивного отчета и технического паспорта данных при помощи средств Markdown и офисных пакетов в соответствии со стандартами инженерной документации.

Владеть:

- навыками работы в интерактивных средах разработки для ведения воспроизводимых журналов аналитического кода;

- приемами векторных вычислений и оконных функций для обработки временных рядов телеметрии;

- методами настройки параметров многомерной графики для когнитивного представления результатов исследования;

- навыками оформления инженерной документации и паспортов данных в отечественных офисных пакетах.

3. Объем дисциплины (модуля).

3.1. Общая трудоемкость дисциплины (модуля).

Общая трудоемкость дисциплины (модуля) составляет 4 з.е. (144 академических часа(ов).

3.2. Объем дисциплины (модуля) в форме контактной работы обучающихся с педагогическими работниками и (или) лицами, привлекаемыми к реализации образовательной программы на иных условиях, при проведении учебных занятий:

Тип учебных занятий	Количество часов	
	Всего	Семестр №5
Контактная работа при проведении учебных занятий (всего):	64	64
В том числе:		
Занятия лекционного типа	32	32
Занятия семинарского типа	32	32

3.3. Объем дисциплины (модуля) в форме самостоятельной работы обучающихся, а также в форме контактной работы обучающихся с педагогическими работниками и (или) лицами, привлекаемыми к реализации образовательной программы на иных условиях, при проведении промежуточной аттестации составляет 80 академических часа (ов).

3.4. При обучении по индивидуальному учебному плану, в том числе при ускоренном обучении, объем дисциплины (модуля) может быть реализован полностью в форме самостоятельной работы обучающихся, а также в форме контактной работы обучающихся с педагогическими работниками и (или) лицами, привлекаемыми к реализации образовательной программы на иных условиях, при проведении промежуточной аттестации.

4. Содержание дисциплины (модуля).

4.1. Занятия лекционного типа.

№ п/п	Тематика лекционных занятий / краткое содержание
1	Введение в разведочный анализ данных и профилирование Рассматриваемые вопросы: - концептуальные основы EDA и его критическая роль в конвейере машинного обучения; - математический аппарат и программные методы автоматизированного профилирования многомерных массивов.
2	Природа пропущенных значений Рассматриваемые вопросы: - классификация механизмов возникновения пропусков; - методы статистической идентификации типа пропусков в транспортных датасетах.

№ п/п	Тематика лекционных занятий / краткое содержание
3	Обнаружение аномалий и выбросов Рассматриваемые вопросы: - теоретические основы детекции выбросов в одномерных и многомерных распределениях; - специфика поиска аномалий в потоках телеметрии и данных с датчиков инфраструктуры.
4	Стратегии импутации данных Рассматриваемые вопросы: - алгоритмы заполнения пропусков и оценка их влияния на статистические свойства выборки; - методы восстановления целостности временных рядов и пространственных данных.
5	Обработка шума и сглаживание сигналов Рассматриваемые вопросы: - методы математического сглаживания и фильтрации высокочастотного шума; - баланс между сохранением полезного сигнала и удалением артефактов измерения.
6	Грамматика графиков и когнитивная визуализация Рассматриваемые вопросы: - принципы когнитивного восприятия графической информации и этика визуализации; - выбор оптимальных типов графиков для решения аналитических задач.
7	Многомерный визуальный анализ Рассматриваемые вопросы: - типология визуализаций для исследования распределений, взаимосвязей и композиций; - инструменты интерактивного исследования данных в среде Jupyter.
8	Математический аппарат корреляционного анализа Рассматриваемые вопросы: - параметрические и непараметрические меры связи; - интерпретация коэффициентов корреляции и границы их применимости.
9	Проверка статистических гипотез Рассматриваемые вопросы: - концепция p-value, доверительные интервалы и ошибки I и II рода; - применение критериев значимости для сравнения выборок и оценки эффектов.
10	Ложные корреляции и смешивающие переменные Рассматриваемые вопросы: - феномен ложной корреляции и методы контроля смешивающих факторов; - различие между корреляционной зависимостью и причинно-следственной связью.
11	Теоретические основы конструирования признаков Рассматриваемые вопросы: - влияние качества признаков на обобщающую способность моделей машинного обучения; - роль предметной области и экспертных знаний в генерации гипотез для Feature Engineering.
12	Генерация признаков из временных рядов Рассматриваемые вопросы: - математические методы создания лаговых переменных и скользящих оконных статистик; - кодирование циклических паттернов и сезонности в транспортных процессах.
13	Извлечение признаков из полуструктурированных данных Рассматриваемые вопросы: - методы парсинга и векторизации временных меток, логов и текстовых атрибутов; - работа с геопространственными координатами и расчет дистанционных метрик.
14	Архитектура предотвращения утечек данных Рассматриваемые вопросы: - классификация утечек данных и их влияние на валидность моделей; - паттерны безопасного разбиения данных и трансформации признаков во времени.

№ п/п	Тематика лекционных занятий / краткое содержание
15	Отбор признаков и мультиколлинеарность Рассматриваемые вопросы: - методы оценки важности признаков и алгоритмы отбора; - диагностика мультиколлинеарности и методы ее устранения.
16	Финализация матрицы признаков и инженерная отчетность Рассматриваемые вопросы: - принципы снижения размерности данных и сохранения информативности; - стандарты оформления паспорта данных и обеспечения воспроизводимости кода.

4.2. Занятия семинарского типа.

Лабораторные работы

№ п/п	Наименование лабораторных работ / краткое содержание
1	Профилирование и аудит качества данных Студент анализирует сырой массив телеметрических данных – выявляет структурные дефекты и классифицирует типы пропусков. На основе предметной области формируется стратегия первичного аудита качества датасета. Итогом работы является аналитическая записка в офисном пакете – Р7-Офис или МойОфис – с описанием выявленных аномалий.
2	Автоматизированное профилирование датасета Студент загружает подготовленный на практическом занятии датасет в среду Jupyter Notebook и применяет библиотеки автоматизированного профилирования. В процессе работы генерируются сводные таблицы и базовые распределения для верификации ручного аудита. Результатом является интерактивный журнал с первичным срезом качества данных.
3	Проектирование критериев поиска аномалий Студент исследует физические границы показателей датчиков и проектирует математические критерии для поиска грубых выбросов. Разрабатывается алгоритм фильтрации аномалий с учетом инерционности транспортных процессов. Итогом становится техническое задание на очистку данных с заданными пороговыми значениями.
4	Программная фильтрация выбросов Студент реализует спроектированные алгоритмы фильтрации при помощи векторных операций NumPy и Pandas. В коде применяются различные методы – метод межквартильного размаха и изолирующий лес – для удаления артефактов измерения. Результатом работы является очищенный датафрейм с зафиксированными метками удаленных строк.
5	Разработка стратегий импутации пропусков Студент анализирует природу пропущенных значений и выбирает оптимальные стратегии импутации для различных групп признаков. Проектируется схема восстановления временных рядов с учетом сезонности и внешних факторов – таких как температура окружающей среды и интенсивность движения. Итогом является матрица соответствия типов пропусков и методов их заполнения.
6	Реализация конвейера импутации Студент программирует выбранные стратегии импутации, применяя интерполяцию и метод k-ближайших соседей. Реализуется конвейер обработки пропусков, сохраняющий статистические свойства исходной выборки. Результатом является полностью заполненный датасет без пропущенных значений.
7	Макетирование многомерного визуального анализа Студент формулирует гипотезы о взаимосвязи целевых метрик и факторов окружающей среды. Разрабатывается макет альбома визуализаций для проверки этих гипотез с точки зрения

№ п/п	Наименование лабораторных работ / краткое содержание
	когнитивного восприятия. Итогом служит детальная схема расположения графиков и выбора цветовых палитр.
8	Построение многомерных визуализаций Студент строит многомерные визуализации при помощи библиотек Seaborn и Matplotlib согласно утвержденному макету. В коде настраиваются параметры читаемости и интерактивности для детального исследования распределений. Результатом является набор графических артефактов, подтверждающих или опровергающих исходные гипотезы.
9	Планирование статистических тестов и проверки гипотез Студент планирует проведение статистических тестов для оценки значимости выявленных визуальных закономерностей. Подбираются параметрические и непараметрические критерии – такие как критерий Стьюдента или Манна-Уитни – с учетом нормальности распределений признаков. Итогом является протокол исследования с заданными уровнями значимости.
10	Корреляционный анализ и статистические тесты Студент вычисляет матрицы корреляций и применяет статистические тесты из библиотеки SciPy. В коде реализуется автоматическая фильтрация ложных корреляций и визуализация тепловых карт взаимосвязей. Результатом является таблица статистически значимых факторов, влияющих на целевую переменную.
11	Проектирование конструирования признаков Студент проектирует новые признаки на основе скользящих окон и лаговых переменных для предиктивной модели. Разрабатывается схема агрегации телеметрических данных за различные временные интервалы. Итогом является спецификация генерируемых признаков с физическим обоснованием их полезности.
12	Генерация признаков из временных рядов Студент программирует конвейер генерации признаков с использованием группировок и оконных функций Pandas. В коде строго контролируется отсутствие утечек данных из будущего – это критично при расчете скользящих статистик. Результатом является расширенная матрица объектов с новыми информативными столбцами.
13	Извлечение признаков из полуструктурированных данных Студент анализирует полуструктурированные атрибуты и проектирует методы извлечения из них числовых признаков. Разрабатываются алгоритмы парсинга временных меток и текстовых логов ошибок. Итогом становится карта преобразования неструктурированных полей в числовые векторы.
14	Парсинг и векторизация полуструктурированных полей Студент реализует функции извлечения признаков, применяя регулярные выражения и методы циклического кодирования времени. В коде формируется отдельный блок трансформации, интегрируемый в общий конвейер подготовки данных. Результатом является датасет, обогащенный признаками из логических и временных меток.
15	Отбор признаков и финализация паспорта данных Студент оценивает избыточность сконструированных признаков и проектирует стратегию их финального отбора. Разрабатывается формат инженерного паспорта данных для передачи матрицы признаков ML-инженерам. Итогом является структура итогового отчета и критерии мультиколлинеарности.
16	Оценка важности признаков и экспорт матрицы Студент применяет алгоритмы оценки важности признаков и вычисляет фактор инфляции дисперсии для удаления дублей. В коде формируется финальный датасет и автоматически генерируется черновик паспорта данных в формате Markdown. Результатом является готовое портфолио с воспроизводимым кодом и итоговой матрицей для обучения модели.

4.3. Самостоятельная работа обучающихся.

№ п/п	Вид самостоятельной работы
1	Изучение рекомендованной литературы.
2	Подготовка к лабораторным работам.
3	Подготовка к промежуточной аттестации.
4	Подготовка к текущему контролю.

5. Перечень изданий, которые рекомендуется использовать при освоении дисциплины (модуля).

№ п/п	Библиографическое описание	Место доступа
1	Маккинни, У. Python и анализ данных / У. Маккинни ; перевод с английского А. А. Слинкина. — 2-ое изд., испр. и доп. — Москва : ДМК Пресс, 2020. — 540 с. — ISBN 978-5-97060-590-5. — Текст : электронный	Лань : электронно-библиотечная система. — URL: https://e.lanbook.com/book/131721 (дата обращения: 19.06.2026)
2	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)
3	Панов, М. А. Анализ данных с использованием языка программирования Python : учебное пособие / М. А. Панов. — Екатеринбург : УрГЭУ, 2024. — 329 с. — Текст : электронный	Лань : электронно-библиотечная система. — URL: https://e.lanbook.com/book/481577 (дата обращения: 19.06.2026)
4	Демидова, Л. А. Интеллектуальный анализ данных на языке Python : учебно-методическое пособие / Л. А. Демидова. — Москва : РТУ МИРЭА, 2021. — 92 с. — Текст : электронный	Лань : электронно-библиотечная система. — URL: https://e.lanbook.com/book/218693 (дата обращения: 19.06.2026)
5	Груздев, А. В. Предварительная подготовка данных в Python / А. В. Груздев. — Москва : ДМК Пресс, 2023 — Том 1 : Инструменты и валидация — 2023. — 816 с. — ISBN 978-5-93700-156-6. — Текст : электронный	Лань : электронно-библиотечная система. — URL: https://e.lanbook.com/book/314945 (дата обращения: 19.06.2026)
6	Хуснуллин, И. Х. Методы обработки данных : учебное пособие / И. Х. Хуснуллин. — Уфа : БГПУ имени М. Акмуллы, 2024. — 92 с. — ISBN 978-5-00251-045-0. — Текст : электронный	Лань : электронно-библиотечная система. — URL: https://e.lanbook.com/book/498257 (дата обращения: 19.06.2026)
7	ернышев, С. А. Основы программирования на Python : учебное пособие для вузов / С. А. Чернышев. — Москва : Издательство Юрайт, 2022. — 286 с. — (Высшее образование). — ISBN 978-5-534-14350-8. — Текст : электронный	Образовательная платформа Юрайт [сайт]. — URL: https://urait.ru/bcode/496893 (дата обращения: 19.06.2026)

6. Перечень современных профессиональных баз данных и информационных справочных систем, которые могут использоваться при освоении дисциплины (модуля).

ЭБС «Лань» – https://e.lanbook.com/

Образовательная платформа «Юрайт» – https://urait.ru/

Pandas (обработка и анализ табличных данных) – Официальная документация. URL: https://pandas.pydata.org/docs/

NumPy (научные вычисления и векторные операции) – Официальная документация. URL: https://numpy.org/doc/

Matplotlib (базовая визуализация данных) – Официальная документация. URL: https://matplotlib.org/stable/contents.html

Seaborn (статистическая визуализация) – Официальная документация. URL: https://seaborn.pydata.org/

SciPy (статистические тесты и математические функции) – Официальная документация. URL: https://docs.scipy.org/doc/scipy/

Scikit-learn (алгоритмы отбора признаков и оценки важности) – API Reference. URL: https://scikit-learn.org/stable/api/index.html

7. Перечень лицензионного и свободно распространяемого программного обеспечения, в том числе отечественного производства, необходимого для освоения дисциплины (модуля).

Операционные системы – Astra Linux Special Edition / ALT Linux / РЕД ОС.

Офисные пакеты – Р7-Офис / МойОфис Стандартный (для подготовки отчетов и презентаций по ГОСТ).

Среда разработки – Anaconda Distribution, Jupyter Notebook / JupyterLab, VS Code Community Edition / VSCodium.

Технологический стек ИИ и Data Science – Python 3.10+, Pandas, NumPy, SciPy, Matplotlib, Seaborn, Plotly, Scikit-learn, CatBoost (разработка Яндекса), ydata-profiling.

Работа с API и сетями – Postman / Hoppscotch, curl.

СУБД – PostgreSQL / Postgres Pro (в реестре ПО РФ), расширение PostGIS, SQLite.

8. Описание материально-технической базы, необходимой для осуществления образовательного процесса по дисциплине (модулю).

Учебные аудитории для проведения учебных занятий, оснащенные компьютерной техникой и наборами демонстрационного оборудования.

Для лабораторных занятий – наличие персональных компьютеров вычислительного класса.

9. Форма промежуточной аттестации:

Экзамен в 5 семестре.

10. Оценочные материалы.

Оценочные материалы, применяемые при проведении промежуточной аттестации, разрабатываются в соответствии с локальным нормативным актом РУТ (МИИТ).

Авторы:

старший преподаватель кафедры
«Цифровые технологии управления
транспортными процессами» А.Ю. Кремнев

Согласовано:

Заведующий кафедрой ЦТУТП В.Е. Нутович

Председатель учебно-методической
комиссии Н.А. Андриянова