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
09.03.01 Информатика и вычислительная техника,
утвержденной первым проректором ФГАОУ ВО
РУТ (МИИТ) Тимониным В.С.

РАБОЧАЯ ПРОГРАММА ДИСЦИПЛИНЫ (МОДУЛЯ)

Предобработка и очистка данных

Направление подготовки: 09.03.01 Информатика и вычислительная техника

Направленность (профиль): Цифровая инженерия транспортных процессов

Форма обучения: Очная

Рабочая программа дисциплины (модуля) в виде электронного документа выгружена из единой корпоративной информационной системы управления университетом и соответствует оригиналу

Простая электронная подпись, выданная ФГАОУ ВО РУТ (МИИТ)
ID подписи: 937226
Подписал: руководитель образовательной программы
Проневич Ольга Борисовна
Дата: 08.09.2026

1. Общие сведения о дисциплине (модуле).

Целями освоения учебной дисциплины являются формирование у обучающихся системных знаний и практических навыков в области предобработки и очистки данных для их последующего использования в аналитических системах и системах искусственного интеллекта. В рамках курса рассматриваются основные этапы подготовки данных: от выявления и анализа «грязных» данных до преобразования и обогащения наборов данных для решения прикладных задач транспортной отрасли. Особое внимание уделяется работе с пропущенными значениями, выбросами, нормализации и стандартизации признаков, кодированию категориальных переменных, а также применению методов машинного обучения для сложных сценариев очистки данных.

Задачами освоения дисциплины являются:

Формирование теоретических знаний о типах данных (структурированные, полуструктурированные, неструктурированные), источниках возникновения «грязных» данных и влиянии качества данных на результаты моделирования.

Освоение методов обработки пропущенных значений: анализ природы пропусков (MCAR, MAR, MNAR), удаление пропусков, заполнение средними, медианными, модальными значениями, использование констант и продвинутых методов импутации.

Изучение подходов к обнаружению и обработке аномалий и выбросов с использованием статистических методов (IQR, Z-score), методов машинного обучения (Isolation Forest, One-Class SVM, LOF, DBSCAN) и стратегий обработки выбросов (удаление, трансформация, винзоризация, замена на пропуски).

Освоение алгоритмов нормализации и стандартизации данных: Min-Max Scaling, Z-score Standardization, Robust Scaling, а также понимание критериев выбора метода в зависимости от типа данных и решаемой задачи.

Формирование практических навыков кодирования категориальных признаков: Label Encoding, One-Hot Encoding, Count/Frequency Encoding, Target Encoding, а также анализ рисков и ограничений каждого метода.

Освоение методов работы с текстовыми данными: очистка текста, токенизация, стемминг и лемматизация, удаление стоп-слов, векторизация текста с использованием Bag of Words и TF-IDF.

Приобретение навыков работы с датами и временем: парсинг, извлечение признаков, вычисление временных интервалов, циклическое кодирование временных признаков.

Освоение методов генерации новых признаков (Feature Engineering), включая биннинг непрерывных признаков, создание полиномиальных признаков, методов отбора признаков и понижения размерности.

Формирование навыков применения методов очистки и предобработки данных с использованием SQL и Python (библиотеки Pandas, NumPy, Scikit-learn), включая построение полных пайплайнов обработки данных.

Знакомство с современными подходами к автоматизации предобработки с использованием искусственного интеллекта: автоматическая генерация кода, автоматизированный инжиниринг признаков, синтетическая генерация данных и сложные методы импутации пропусков с использованием ML-моделей.

Развитие компетенций по оценке качества данных и выбору оптимальной стратегии очистки с учетом специфики предметной области и требований к аналитическим задачам.

Дисциплина предназначена для формирования компетенций, необходимых для эффективной подготовки и очистки данных перед проведением аналитических исследований и построением моделей машинного обучения в транспортных организациях.

2. Планируемые результаты обучения по дисциплине (модулю).

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

ППК-ДА2 - Способен подготавливать и обрабатывать данные для аналитических исследований.

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

Знать:

Принципы подготовки данных для задач искусственного интеллекта и аналитических исследований.

Методы предобработки структурированных, полуструктурированных и неструктурированных данных.

Алгоритмы нормализации и стандартизации данных, критерии их выбора.

Техники обработки пропущенных значений и выбросов, включая статистические и ML-методы.

Критерии качества данных для различных аналитических задач.

Методы кодирования категориальных признаков и их ограничения.

Подходы к работе с текстовыми данными и временными рядами на этапе предобработки.

Уметь:

Применять методы очистки данных в зависимости от их типа и структуры.

Преобразовывать данные в требуемые форматы для дальнейшего анализа.

Выявлять и обрабатывать пропущенные значения и выбросы с использованием различных методов.

Выполнять нормализацию и стандартизацию признаков для подготовки данных к машинному обучению.

Кодировать категориальные переменные и создавать новые признаки.

Использовать библиотеки Python (Pandas, NumPy, Scikit-learn) и SQL для очистки и трансформации данных.

Владеть:

Навыками работы с «грязными» данными: выявление, анализ и устранение проблем.

Навыками предобработки и очистки данных на микроуровне (работа с отдельными признаками, строками, значениями).

Навыками построения воспроизводимых пайплайнов очистки и трансформации данных.

Навыками выбора и обоснования стратегии очистки данных для конкретной аналитической задачи.

Навыками работы с инструментами профилирования и контроля качества данных.

3. Объем дисциплины (модуля).

3.1. Общая трудоемкость дисциплины (модуля).

Общая трудоемкость дисциплины (модуля) составляет 2 з.е. (72 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 №1
Контактная работа при проведении учебных занятий (всего):	32	32

В том числе:		
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 40 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.

Не предусмотрено учебным планом

4.2. Занятия семинарского типа.

Практические занятия

№ п/п	Тематика практических занятий/краткое содержание
1	Тема 1. Введение в предобработку данных. Основные этапы Рассматриваются вопросы: Что такое «грязные данные» и их источники. Цели и задачи предобработки данных. Обзор пайплайна (ETL/ELT): извлечение, трансформация, загрузка. Влияние качества данных на результаты моделирования (Garbage In, Garbage Out). Типы данных: структурированные, неструктурированные, полуструктурированные.
2	Тема 2. Работа с пропущенными значениями Рассматриваются вопросы: Природа пропусков: MCAR, MAR, MNAR. Методы обработки: общий обзор стратегий. Удаление пропущенных значений (listwise/pairwise deletion). Заполнение пропущенных значений (Imputation): среднее, медиана, мода, использование констант.
3	Тема 3. Работы с аномалиями. Часть 1. Рассматриваются вопросы: Понятие выброса и аномалии. Разница между шумом, выбросом и аномалией. Контекстуальные и коллективные выбросы. Статистические методы обнаружения: Метод межквартильного размаха (IQR).

№ п/п	Тематика практических занятий/краткое содержание
	Метод Z-отклонения (Z-score) и его модификации для малых выборок. Визуальные методы: «Ящик с усами» (Boxplot), гистограммы и диаграммы рассеяния (Scatter plot).
4	Тема 4. Работы с аномалиями. Часть 2. Рассматриваются вопросы: Методы машинного обучения для обнаружения выбросов (обзор). Isolation Forest (Лес изоляции). Метод опорных векторов (One-Class SVM). Алгоритмы кластеризации для поиска аномалий (DBSCAN, K-means (расстояние до центроида)). Эллиптическая оболочка (Elliptic Envelope) и фактор локальных выбросов (LOF).
5	Тема 5. Работы с аномалиями. Часть 3. Рассматриваются вопросы: Стратегии обработки выбросов: когда удалять, а когда оставлять. Удаление строк с выбросами: риски и последствия. Трансформация данных для подавления выбросов (логарифмирование, извлечение корня). Винзоризация (Winsorization): ограничение экстремальных значений процентилями. Замена выбросов на пропуски и последующая импутация (Missing Value Imputation).
6	Тема 6. Нормализация и стандартизация Рассматриваются вопросы: Почему признаки нужно приводить к одному масштабу? (Градиентный спуск, метрики расстояния). Нормализация данных: Min-Max Scaling (приведение к диапазону [0, 1]). Стандартизация данных: Z-score Standardization (StandardScaler). Robust Scaling (устойчивое масштабирование на основе медианы и IQR). Сравнение методов: когда применять каждый из них на практике.
7	Тема 7. Кодирование категориальных признаков Рассматриваются вопросы: Разница между номинальными и порядковыми (ordinal) признаками. Label Encoding (меточное кодирование) и Ordinal Encoding: риски внесения ложного порядка. One-Hot Encoding (унитарное кодирование), создание бинарных векторов. Проблема «проклятия размерности» при большом количестве категорий. Счётчики (Count/Frequency Encoding) и Target Encoding (кодирование средним значением целевой переменной).
8	Тема 8. Работа с текстовыми данными Рассматриваются вопросы: Базовые методы очистки текста: приведение к нижнему регистру, удаление пунктуации, цифр и спецсимволов. Удаление стоп-слов и стемминг/лемматизация. Регулярные выражения (Regex): поиск email, телефонов, дат в неструктурированном тексте. Токенизация (разбиение на слова/предложения). Векторизация текста: Мешок слов (Bag of Words) и TF-IDF (Term Frequency – Inverse Document Frequency).
9	Тема 9. Работа с датами и временем. Рассматриваются вопросы: Парсинг дат и приведение к единому формату (datetime). Извлечение признаков из дат: день недели, месяц, квартал, день года. Вычисление временных интервалов (разницы между датами). Циклическая природа времени (перевод часов/дней в синусы и косинусы для моделей). Работа с временными зонами и летним временем.
10	Тема 10. Преобразование форматов и генерация признаков Рассматриваются вопросы: Приведение типов данных (строка в число, object в category для оптимизации памяти).

№ п/п	Тематика практических занятий/краткое содержание
	Понижение размерности данных (PCA, t-SNE) как метод предобработки. Биннинг (дискретизация) непрерывных признаков. Полиномиальные признаки и создание взаимодействий. Методы отбора признаков (Feature Selection): фильтры, обёртки, встроенные методы.
11	Тема 11. Области ИИ в предобработке и очистки данных. Часть 1 Рассматриваются вопросы: Автоматическая генерация кода для очистки (помощь LLM). Автоматизированный инжиниринг признаков (Deep Feature Synthesis, библиотека Featuretools). Умное обнаружение дубликатов: нечёткое сравнение строк (Fuzzy Matching, расстояние Левенштейна).
12	Тема 12. Области ИИ в предобработке и очистки данных. Часть 2 Рассматриваются вопросы: Синтетическая генерация данных для балансировки классов (SMOTE, CTGAN). Использование ML-моделей для сложной импутации пропусков (MICE, KNN-Imputer, MissForest).
13	Тема 13. Практические примеры предобработки и очистки данных на SQL и python. Часть 1 Рассматриваются вопросы: Очистка дублирующихся строк (DISTINCT, ROW_NUMBER OVER PARTITION). Обработка NULL и заполнение значений (COALESCE, CASE WHEN, оконные функции). Поиск выбросов и агрегация данных на SQL. Нормализация и кодирование прямо в запросе (CASE для One-Hot).
14	Тема 14. Практические примеры предобработки и очистки данных на SQL и python. Часть 2 Рассматриваются вопросы: Работа с DataFrame: фильтрация, сортировка, слияние (merge). Цепочки методов (Method chaining) и функция pipe(). Применение функций к группам (groupby + transform) для заполнения пропусков медианой группы. Оптимизация работы с памятью (изменение dtype).
15	Тема 15. Практические примеры предобработки и очистки данных на SQL и python. Часть 3 Рассматриваются вопросы: Написание пользовательских функций (User Defined Functions) для очистки. Библиотека Scikit-learn Pipelines. Класс ColumnTransformer: применение разных стратегий к разным колонкам одновременно. Пример полного пайплайна: от сырых данных до готового обучающего набора.
16	Тема 16. Самостоятельная практическая работа по предобработке и очистке реального набора данных Рассматриваются вопросы: Разбор реального датасета (транспорт). Поиск проблем. Составление стратегии очистки: какие методы выбрать и в какой последовательности. Защита решения: объяснение выбора методов обработки выбросов и пропусков.

4.3. Самостоятельная работа обучающихся.

№ п/п	Вид самостоятельной работы
1	Изучение литературы по очистке и предобработке данных.
2	Работа с документацией изученных библиотек.

№ п/п	Вид самостоятельной работы
3	Выполнение заданий по обработке пропусков и выбросов.
4	Подготовка к текущему контролю.
5	Подготовка к промежуточной аттестации.
6	Подготовка к текущему контролю.

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

№ п/п	Библиографическое описание	Место доступа
1	Алетдинова, А. А. Интеллектуальный анализ больших данных : учебное пособие / А. А. Алетдинова, М. Ш. Муртазина. — Новосибирск : НГТУ, 2023. — 66 с. — ISBN 978-5-7782-4899-1.	https://e.lanbook.com/book/404567
2	Газанова, Н. Ш. Методы искусственного интеллекта : учебно-методическое пособие / Н. Ш. Газанова, С. Н. Никольский. — Москва : РТУ МИРЭА, 2023. — 102 с. — ISBN 978-5-7339-1805-1.	https://e.lanbook.com/book/368756

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

https://habr.com/ru - база знаний в виде статей, обзоров

http://library.miit.ru/ - электронно-библиотечная система научно-технической библиотеки МИИТ

https://superset.apache.org/user-docs/intro/#documentation – документация Apache Superset

https://www.postgresql.org/ - официальная страница PostgreSQL

https://yandex.cloud/ru/docs/datalens/ - официальная документация Yandex Datalens

https://docs.docker.com/ - официальная документация Docker

https://www.metabase.com/docs/latest/ - официальная документация Metabase

https://tableau.pro/ - официальный сайт Tableau с лучшими практиками проектирования дашбордов

https://revealthedata.com/blog/ - Блог о визуализации данных и BI-системах

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

MS Office (Excel, Word, PowerPoint)
Apache Superset
Yandex Datalens
Metabase
PostgreSQL
Docker
Apache Airflow
Браузер Chrome / Яндекс браузер
Draw.io

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

Компьютер преподавателя с доступом к сети Интернет
Компьютеры студентов с установленным Python-окружением и табличным процессором
Доступ к учебной базе событийных данных и экспериментальным выборкам
Доступ к PostgreSQL и BI-системе для визуализации результатов
Доступ к облачной виртуальной машине или Docker-контейнерам для воспроизводимого анализа
Проектор и экран для проектора
Маркерная доска

9. Форма промежуточной аттестации:

Зачет в 1 семестре.

10. Оценочные материалы.

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

Авторы:

руководитель образовательной программы О.Б. Проневич

Согласовано:

Директор Д.В. Паринов

Руководитель образовательной программы О.Б. Проневич

Председатель учебно-методической комиссии Д.В. Паринов