МИНИСТЕРСТВО ТРАНСПОРТА РОССИЙСКОЙ ФЕДЕРАЦИИ

ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ

«РОССИЙСКИЙ УНИВЕРСИТЕТ ТРАНСПОРТА»
(РУТ (МИИТ)

АННОТАЦИЯ К РАБОЧЕЙ ПРОГРАММЕ ДИСЦИПЛИНЫ (МОДУЛЯ)

Предобработка и очистка данных

Направление подготовки:	09.03.01 – Информатика и вычислительная техника
Направленность (профиль):	Цифровая инженерия транспортных процессов
Форма обучения:	Очная

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.

Общая трудоемкость дисциплины (модуля) составляет 2 з.е. (72 академических часа(ов).