

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)

Машинное обучение и анализ данных

Направление подготовки: 09.03.01 Информатика и вычислительная техника

Направленность (профиль): IT-сервисы и технологии обработки данных на транспорте

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: директор центра Проневич Ольга Борисовна

Дата: 21.10.2026

1. Общие сведения о дисциплине (модуле).

Целью освоения учебной дисциплины является формирование у обучающихся системных знаний и практических навыков в области машинного обучения и анализа данных, включая понимание методов предобработки данных, принципов формирования выборок, техник балансировки и аугментации данных, а также применение классических алгоритмов машинного обучения для решения прикладных задач.

Задачами освоения дисциплины являются:

Формирование теоретических знаний о принципах формирования обучающих и тестовых выборок для различных задач машинного обучения.

Освоение практических навыков предобработки, очистки и преобразования данных для аналитических исследований.

Приобретение умений применять методы балансировки данных для работы с несбалансированными выборками.

Освоение техник аугментации данных для увеличения объёма и разнообразия обучающих выборок.

Формирование компетенций по применению классических алгоритмов машинного обучения для задач регрессии, классификации и кластеризации.

Приобретение навыков оценки качества моделей и интерпретации результатов анализа данных.

Дисциплина предназначена для формирования компетенций, необходимых для решения профессиональных задач, связанных с подготовкой и обработкой данных для аналитических исследований в транспортной отрасли и смежных областях.

2. Планируемые результаты обучения по дисциплине (модулю).

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

МО-1 - Способен проектировать, разрабатывает и улучшать алгоритмы и модели машинного обучения;

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

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

Знать:

Принципы формирования обучающих и тестовых выборок для различных задач машинного обучения, включая методы разделения данных и стратификацию.

Методы балансировки данных для работы с несбалансированными выборками, включая техники ресемплинга и синтетической генерации данных.

Техники аугментации данных для увеличения объёма и разнообразия обучающих выборок, а также методы их применения к различным типам данных.

Уметь:

Формировать обучающие, валидационные и тестовые выборки с учётом особенностей данных и требований решаемой задачи.

Применять методы балансировки данных для устранения дисбаланса классов и повышения качества моделей.

Использовать техники аугментации данных для увеличения объёма обучающей выборки и улучшения обобщающей способности моделей.

Владеть:

Навыками предобработки и подготовки данных для аналитических исследований с использованием стандартных библиотек.

Навыками применения методов балансировки и аугментации данных в рамках пайплайнов машинного обучения.

Навыками оценки качества моделей и интерпретации результатов анализа 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 №2
Контактная работа при проведении учебных занятий (всего):	48	48
В том числе:		
Занятия лекционного типа	16	16

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 24 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	Тема 1. Введение в машинное обучение и анализ данных. Рассматриваемые вопросы: - Определение машинного обучения, связь с анализом данных и искусственным интеллектом. - Классификация задач машинного обучения: обучение с учителем, без учителя, с подкреплением. - Жизненный цикл проекта по машинному обучению (CRISP-DM). - Роль подготовки данных в успехе аналитических проектов (Data-Centric AI). - Применение машинного обучения в транспортной отрасли: прогнозирование трафика, диагностика, оптимизация маршрутов.
2	Тема 2. Принципы формирования обучающих и тестовых выборок. Рассматриваемые вопросы: - Основные понятия: обучающая, валидационная и тестовая выборки. - Методы разделения данных: hold-out, кросс-валидация (k-fold, stratified). - Стратификация при разделении выборок для сохранения распределения классов. - Временное разделение для задач прогнозирования (без перемешивания). - Принципы формирования выборок для различных типов задач (регрессия, классификация, кластеризация).
3	Тема 3. Продвинутая предобработка и преобразование данных. Рассматриваемые вопросы: - Основные этапы предобработки данных в инженерном цикле. - Обработка пропущенных значений: удаление, заполнение (среднее, медиана, KNN-заполнение), интерполяция. - Выявление и обработка выбросов (IQR, Z-score, изолирующий лес). - Нормализация и стандартизация данных, выбор метода в зависимости от алгоритма. - Кодирование категориальных признаков (One-Hot, Target Encoding, Frequency Encoding).

№ п/п	Тематика лекционных занятий / краткое содержание
	<ul style="list-style-type: none"> - Feature Engineering: создание новых признаков на основе существующих. - Критерии качества данных для аналитических задач.
4	<p>Тема 4. Методы балансировки данных.</p> <p>Рассматриваемые вопросы:</p> <ul style="list-style-type: none"> - Проблема несбалансированных данных и её влияние на качество моделей (смещение в сторону большинства). - Методы недообучения (undersampling): случайное удаление примеров большинства, Cluster Centroids. - Методы переобучения (oversampling): случайное дублирование примеров меньшинства. - Синтетическая генерация данных: SMOTE и его модификации (ADASYN, Borderline-SMOTE). - Комбинированные подходы к балансировке (SMOTE + Tomek Links). - Оценка эффективности методов балансировки на реальных данных.
5	<p>Тема 5. Техники аугментации данных.</p> <p>Рассматриваемые вопросы:</p> <ul style="list-style-type: none"> - Понятие аугментации данных и её роль в машинном обучении (улучшение обобщения). - Аугментация для табличных данных: добавление гауссовского шума, трансформации, синтетические примеры. - Аугментация для текстовых данных: синонимическая замена, бэк-транслейшн, замена слов по маске. - Применение аугментации для улучшения обобщающей способности моделей. - Ограничения и риски при использовании аугментации (искажение распределения, выбросы).
6	<p>Тема 6. Сравнительный анализ алгоритмов классификации.</p> <p>Рассматриваемые вопросы:</p> <ul style="list-style-type: none"> - Постановка задачи классификации и выбор метрик для несбалансированных данных. - Логистическая регрессия: регуляризация (L1, L2) и интерпретация коэффициентов. - Метод опорных векторов: выбор ядра и настройка гиперпараметров. - Деревья решений: управление переобучением через ограничения. - Ансамблевые методы: случайный лес (бэггинг) и градиентный бустинг (XGBoost, LightGBM). - Критерии выбора алгоритма в зависимости от размера данных, интерпретируемости и требований к скорости.
7	<p>Тема 7. Алгоритмы регрессии и прогнозирования временных рядов.</p> <p>Рассматриваемые вопросы:</p> <ul style="list-style-type: none"> - Постановка задачи регрессии и выбор метрик (MSE, MAE, MAPE). - Линейная регрессия и её модификации: гребневая регрессия (Ridge) и LASSO (L1-регуляризация). - Сравнение методов регуляризации и их влияние на отбор признаков. - Основы прогнозирования временных рядов: тренд, сезонность, шум. - Модели прогнозирования: скользящее среднее, экспоненциальное сглаживание, ARIMA (концептуально). - Сравнительный анализ методов регрессии для различных типов данных.
8	<p>Тема 8. Кластеризация и комплексная оценка качества моделей.</p> <p>Рассматриваемые вопросы:</p> <ul style="list-style-type: none"> - Постановка задачи кластеризации и области применения. - Метод K-means: алгоритм, выбор числа кластеров (метод локтя, силуэт). - Иерархическая кластеризация: агломеративная и дивизивная. - Метрики качества классификации для сбалансированных и несбалансированных данных. - Метрики качества регрессии и их интерпретация. - Перекрестная проверка, кривые обучения и борьба с переобучением. - Визуализация и интерпретация результатов анализа для бизнес-задач.

4.2. Занятия семинарского типа.

Практические занятия

№ п/п	Тематика практических занятий/краткое содержание
1	<p>Тема 1. Установка и настройка среды для анализа данных.</p> <p>Рассматриваемые вопросы:</p> <ul style="list-style-type: none"> - Установка и настройка сред разработки (Anaconda, VS Code, Jupyter Lab). - Установка библиотек для анализа данных и машинного обучения (numpy, pandas, matplotlib, seaborn, scikit-learn, imbalanced-learn, xgboost). - Знакомство с интерфейсом и функционалом. - Проверка работоспособности инструментов и создание первого ноутбука.
2	<p>Тема 2. Разведочный анализ данных (EDA) и визуализация.</p> <p>Рассматриваемые вопросы:</p> <ul style="list-style-type: none"> - Загрузка и первичный просмотр данных (head, info, describe). - Описательная статистика и анализ распределений. - Визуализация распределений (гистограммы, boxplot, violinplot). - Выявление закономерностей и аномалий с помощью парных графиков. - Анализ корреляций между признаками.
3	<p>Тема 3. Продвинутое удаление пропущенных значений и выбросов.</p> <p>Рассматриваемые вопросы:</p> <ul style="list-style-type: none"> - Выявление пропущенных значений и их визуализация (тепловая карта). - Применение методов заполнения пропусков (среднее, медиана, KNN-заполнение, интерполяция). - Выявление и удаление выбросов с использованием IQR и изолирующего леса. - Оценка влияния обработки на распределение данных.
4	<p>Тема 4. Нормализация, стандартизация и кодирование признаков.</p> <p>Рассматриваемые вопросы:</p> <ul style="list-style-type: none"> - Применение нормализации (MinMaxScaler) и стандартизации (StandardScaler, RobustScaler). - Кодирование категориальных признаков (One-Hot Encoding, Target Encoding). - Сравнение различных методов кодирования на качестве модели. - Создание пайплайна предобработки с использованием ColumnTransformer.
5	<p>Тема 5. Формирование обучающей и тестовой выборки.</p> <p>Рассматриваемые вопросы:</p> <ul style="list-style-type: none"> - Разделение данных на обучающую, валидационную и тестовую выборки. - Применение стратификации для сохранения распределения классов. - Кросс-валидация на практике (cross_val_score). - Временное разделение для задач прогнозирования (TimeSeriesSplit). - Сравнение различных подходов к разделению.
6	<p>Тема 6. Методы балансировки данных на практике.</p> <p>Рассматриваемые вопросы:</p> <ul style="list-style-type: none"> - Выявление дисбаланса классов (визуализация, расчёт соотношения). - Применение методов недообучения (RandomUnderSampler). - Применение методов переобучения (RandomOverSampler). - Применение SMOTE (SMOTE, ADASYN). - Сравнение эффективности методов на одной модели.
7	<p>Тема 7. Техники аугментации данных на практике.</p> <p>Рассматриваемые вопросы:</p> <ul style="list-style-type: none"> - Применение аугментации для табличных данных (добавление шума, масштабирование). - Применение аугментации для текстовых данных (синонимическая замена). - Оценка влияния аугментации на качество моделей. - Сравнение с моделями без аугментации (АВ-тестирование).
8	<p>Тема 8. Логистическая регрессия с регуляризацией.</p> <p>Рассматриваемые вопросы:</p> <ul style="list-style-type: none"> - Построение модели логистической регрессии с L1 и L2 регуляризацией.

№ п/п	Тематика практических занятий/краткое содержание
	<ul style="list-style-type: none"> - Обучение и оценка качества на несбалансированных данных. - Интерпретация коэффициентов и отбор признаков. - Построение матрицы ошибок и ROC-кривой.
9	<p>Тема 9. Метод опорных векторов и сравнение с логистической регрессией.</p> <p>Рассматриваемые вопросы:</p> <ul style="list-style-type: none"> - Построение модели SVM с линейным и RBF-ядром. - Настройка гиперпараметров (C, gamma) с использованием Grid Search. - Сравнение SVM и логистической регрессии на одном датасете. - Оценка качества и интерпретация результатов.
10	<p>Тема 10. Сравнение ансамблевых методов: случайный лес и градиентный бустинг.</p> <p>Рассматриваемые вопросы:</p> <ul style="list-style-type: none"> - Построение моделей случайного леса и градиентного бустинга (XGBoost). - Настройка гиперпараметров (n_estimators, max_depth, learning_rate). - Сравнение с логистической регрессией и SVM. - Оценка важности признаков и интерпретация.
11	<p>Тема 11. Линейная, гребневая регрессия и LASSO.</p> <p>Рассматриваемые вопросы:</p> <ul style="list-style-type: none"> - Построение моделей линейной регрессии, Ridge и LASSO. - Обучение и оценка качества (MSE, MAE, R²). - Интерпретация коэффициентов и влияние регуляризации. - Сравнение методов на данных с мультиколлинеарностью.
12	<p>Тема 12. Кластеризация методом K-means и оценка качества.</p> <p>Рассматриваемые вопросы:</p> <ul style="list-style-type: none"> - Реализация алгоритма K-means. - Определение оптимального числа кластеров (метод локтя, силуэт). - Визуализация результатов кластеризации. - Интерпретация полученных кластеров для бизнес-задач.
13	<p>Тема 13. Комплексная оценка качества моделей классификации.</p> <p>Рассматриваемые вопросы:</p> <ul style="list-style-type: none"> - Расчёт метрик качества для сбалансированных и несбалансированных данных. - Построение матрицы ошибок и отчёта классификации. - Построение ROC-кривой и PR-кривой. - Интерпретация метрик для бизнес-задач и выбора модели.
14	<p>Тема 14. Комплексная оценка качества моделей регрессии.</p> <p>Рассматриваемые вопросы:</p> <ul style="list-style-type: none"> - Расчёт метрик качества: MSE, MAE, RMSE, MAPE, R². - Визуализация остатков и анализ гетероскедастичности. - Интерпретация метрик и сравнение моделей. - Выбор лучшей модели для прогнозирования.
15	<p>Тема 15. Подбор гиперпараметров и построение пайплайнов.</p> <p>Рассматриваемые вопросы:</p> <ul style="list-style-type: none"> - Применение кросс-валидации с использованием Grid Search и Random Search. - Сравнение эффективности методов подбора. - Построение полного пайплайна (предобработка > балансировка > модель > оценка). - Выбор оптимальных гиперпараметров для итоговой модели.
16	<p>Тема 16. Комплексный проект: защита и презентация результатов.</p> <p>Рассматриваемые вопросы:</p> <ul style="list-style-type: none"> - Формулировка прикладной задачи на основе реального кейса. - Сбор, предобработка и анализ данных. - Выбор и обучение моделей.

№ п/п	Тематика практических занятий/краткое содержание
	<ul style="list-style-type: none"> - Оценка качества и интерпретация результатов. - Подготовка итогового отчёта и презентации. - Защита проекта.

4.3. Самостоятельная работа обучающихся.

№ п/п	Вид самостоятельной работы
1	Работа с лекционным материалом и дополнительной литературой.
2	Изучение документации по библиотекам для анализа данных и машинного обучения.
3	Подготовка к практическим занятиям и выполнение домашних заданий.
4	Решение задач по подготовке данных и построению моделей.
5	Работа с реальными наборами данных.
6	Выполнение комплексного проекта.
7	Подготовка к промежуточной аттестации и текущему контролю.
8	Подготовка к промежуточной аттестации.
9	Подготовка к текущему контролю.

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

№ п/п	Библиографическое описание	Место доступа
1	Машинное обучение : методические указания / составитель И. А. Андрианов. — Вологда : ВоГУ, 2023. — 33 с.	https://e.lanbook.com/book/518045
2	Артемьев, М. А. Машинное обучение : учебно-методическое пособие / М. А. Артемьев, С. В. Золотарев, Е. С. Барановский. — Воронеж : ВГУ, 2021. — 22 с.	https://e.lanbook.com/book/455024

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).

Python 3.x
Jupyter Notebook
Система контроля версий Git
Офисный пакет приложений
Браузер

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 во 2 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

Директор ЦС IT - программ ВИШ

О.Б. Проневич

Председатель учебно-методической
комиссии

Д.В. Паринов