

МИНИСТЕРСТВО ТРАНСПОРТА РОССИЙСКОЙ ФЕДЕРАЦИИ
ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ ОБРАЗОВАТЕЛЬНОЕ
УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ
«РОССИЙСКИЙ УНИВЕРСИТЕТ ТРАНСПОРТА»
(РУТ (МИИТ))



Рабочая программа дисциплины (модуля),
как компонент образовательной программы
высшего образования - программы бакалавриата
по направлению подготовки
09.03.01 Информатика и вычислительная техника,
утвержденной первым проректором 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): 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 РУТ (МИИТ)
ID подписи: 937226
Подписал: руководитель образовательной программы
Проневич Ольга Борисовна
Дата: 24.10.2024

1. Общие сведения о дисциплине (модуле).

Целями освоения учебной дисциплины являются формирование у обучающихся профессиональных знаний и навыков в области использования моделей и методов систем искусственного интеллекта (ИИ); усвоение основ разработки и применения методов систем ИИ при решении прикладных задач обработки информации и оптимального управления, а также формирование у обучающихся компетенций в области построения и функционирования систем ИИ, методов оптимального представления и обработки знаний в этих системах, методов и алгоритмов обработки информации и оптимального управления.

Дисциплина предназначена для формирования и закрепления у обучающихся компетенций, необходимых для решения следующих профессиональных задач (в рамках проектной деятельности):

- Проверка реализации запросов на изменение (верификация) в соответствии с полученным планом;
- Обеспечение качества в проектах в области ИТ в соответствии с установленными регламентами;
- Организация приемо-сдаточных испытаний (валидация) в проектах малого и среднего уровня сложности в области ИТ в соответствии с установленными регламентами.

2. Планируемые результаты обучения по дисциплине (модулю).

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

ОПК-8 - Способен разрабатывать алгоритмы и программы, пригодные для практического применения;

ПК-1 - Способен анализировать большие данные с использованием существующей в организации методологической и технологической инфраструктуры.

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

Знать:

- требования к персональным компьютерам и серверным ресурсам, реализующим алгоритмы машинного обучения
- основные методы оптимального представления и обработки больших данных в системах ИИ

Уметь:

- разрабатывать и проводить сравнительный анализ алгоритмов распознавания и обработки больших данных
- проектировать вычислительную инфраструктуру для хранения больших объемов данных

Владеть:

- навыками применения методов машинного обучения при решении профессиональных задач
- навыками разработки алгоритмов машинного обучения

3. Объем дисциплины (модуля).**3.1. Общая трудоемкость дисциплины (модуля).**

Общая трудоемкость дисциплины (модуля) составляет 3 з.е. (108 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 №4
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 44 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. Жизненный цикл разработки моделей машинной обучения. Рассматриваемые вопросы: <ul style="list-style-type: none">- машинное обучение на транспорте. Виды данных- нормативные документы в области машинного обучения- жизненный цикл машинного обучения- разведочный анализ данных
2	Тема 2. Разведочный анализ данных (EDA). Часть 1 - Анализ. Рассматриваемые вопросы: <ul style="list-style-type: none">- Макроуровень EDA- Микроуровень EDA- описательная статистика для EDA- Проверка статистических гипотез для определения количества массивов данных
3	Тема 3. Разведочный анализ данных. Часть 2 - Преобразование. Рассматриваемые вопросы: <ul style="list-style-type: none">- переход к целевой структуре данных- анализ аномалий
4	Тема 4. Инженерия данных. Обработка числовых и категориальных данных Рассматриваемые вопросы: <ul style="list-style-type: none">- масштабирование данных- преобразование данных категориальных типов- снижение размерности- генерация новых признаков
5	Тема 5. Обработка текстовых данных Рассматриваемые вопросы: <ul style="list-style-type: none">- токенизация текста- переход от текста к числовой форме TF-IDF- переход от текста к числовой форме мешок-слов- лемматизация
6	Тема 6. Виды машинного обучения. Формальная постановка задачи ML Рассматриваемые вопросы: <ul style="list-style-type: none">- машинное обучение с учителем- машинное обучение без учителя- гибридные методы
7	Тема 7. Модели классификации Рассматриваемые вопросы: <ul style="list-style-type: none">- дерево решений- опорные вектора- градиентный бустинг- логистическая регрессия
8	Тема 8. Модели регрессии Рассматриваемые вопросы: <ul style="list-style-type: none">- линейная регрессия

№ п/п	Тематика лекционных занятий / краткое содержание
	- гребневая регрессия - модели прогнозирования временных рядов
9	Тема 9. Гиперпараметры моделей машинного обучения Рассматриваемые вопросы: - основные гиперпараметры - gridsearchcv - k-fold
10	Тема 10. Метрики качества моделей и визуализация результатов Рассматриваемые вопросы: - метрики качества регрессионных моделей - метрики качества классификация - визуализация результатов
11	Тема 11. Кластеризация и балансировка классов Рассматриваемые вопросы: - методы поиска ближайших соседей - методы кластеризации - методы балансировки классов
12	Тема 12. Погружение в технологии машинного обучения Рассматриваемые вопросы: - градиентный спуск - виды градиентного бустинга
13	Тема 13. Работа с временными рядами Рассматриваемые вопросы: - avtoArima - динамическая трансформация временных рядов

4.2. Занятия семинарского типа.

Практические занятия

№ п/п	Тематика практических занятий/краткое содержание
1	Тема 1. Что такое машинное обучение (МО) Рассматриваемые вопросы: - Задачи машинного обучения (постановка) - Жизненный цикл - Классификация
2	Тема 2. EDA\Feature engineering Рассматриваемые вопросы: - Место в жизненном цикле интеллектуального анализа данных. - Задачи микроуровня и макроуровня. - Обязательные анализы, необходимы при проведении EDA для различных типов данных. - Инженерия данных
3	Тема 3. Кластеризация Рассматриваемые вопросы: - Виды кластеризации - k-means
4	Тема 4. Поиск аномалий Рассматриваемые вопросы: - Что такое аномалии?

№ п/п	Тематика практических занятий/краткое содержание
	- Методы выявления и обработки аномалий. - 3-сигма, автоматизированные методы, методы с помощью нейронных сетей
5	Тема 5. Метрики качества Рассматриваемые вопросы: - Для чего нужны метрики? - RMSE, MAE, MSE - Accuracy, Precision,
6	Тема 6. Регрессия Рассматриваемые вопросы: - Линейная регрессия - Дерево решений - Нейронные сети - Случайный лес
7	Тема 7. Классификация Рассматриваемые вопросы: - Логистическая регрессия - Дерево решений - Случайный лес - Градиентный бустинг
8	Тема 8. Интерпретация моделей машинного обучения Рассматриваемые вопросы: - LIME - SHAP - ICE - PD
9	Тема 9. Подбор гиперпараметров и Auto ML Рассматриваемые вопросы: - Кросс-валидация - Grid Search - Optuna - Auto ML-библиотеки - Ансамбли моделей
10	Тема 10. Low Code инструменты для моделирования Рассматриваемые вопросы: - Open Source инструменты - Российский рынок ПО для ML
11	Тема 11. MLOps Рассматриваемые вопросы: - ModelOps и MLOps - Из чего собрать платформу для работы с машинным обучением? - Подводные камни, регламентные процессы, проектный опыт

4.3. Самостоятельная работа обучающихся.

№ п/п	Вид самостоятельной работы
1	Изучение литературы
2	Работа с документацией библиотек pandas, matplotlib, seaborn, pandas_profiling
3	Выполнение курсовой работы.

4	Подготовка к промежуточной аттестации.
5	Подготовка к текущему контролю.

4.4. Примерный перечень тем курсовых работ

1. Использование метода опорных векторов в задачах машинного обучения.
2. Использование метода наименьших квадратов в задачах машинного обучения.
3. Использование нейронных сетей долгой краткосрочной памяти в задачах анализа временных рядов.
4. Использование библиотеки TensorFlow в задачах машинного обучения.
5. Использование IBM Watson Studio для создания когнитивных приложений.
6. Использование IBM Watson Studio в задачах распознавания образов.
7. Обработка естественного языка в когнитивных приложениях.
8. Алгоритмы поиска ассоциативных правил.
9. Рекомендательные системы.
10. Распознавание образов с использованием искусственных нейронных сетей глубокого обучения.
11. Обучение с подкреплением в системах искусственного интеллекта.
12. Сегментация изображений с использованием методов кластерного анализа.

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

№ п/п	Библиографическое описание	Место доступа
1	Габдуллин, Н. М. Развитие человеческого капитала и цифровой экономики в регионах России: факторный и кластерный анализ : монография / Н. М. Габдуллин. — Казань : КФУ, 2019. — 268 с. — ISBN 978-5-00130-291-9	https://e.lanbook.com/book/173018
2	Гласснер, Э. Глубокое обучение без математики. Том 2. Практика : руководство / Э. Гласснер ; перевод с английского В. А. Яроцкого. — Москва : ДМК Пресс, 2020. — 610 с. — ISBN 978-5-97060-767-1	https://e.lanbook.com/book/131710

3	Гулятьева, Т. А. Методы статистического обучения в задачах регрессии и классификации : монография / Т. А. Гулятьева, А. А. Попов, А. С. Саутин. — Новосибирск : НГТУ, 2016. — 323 с. — ISBN 978-5-7782-2817-7	https://e.lanbook.com/book/118291
4	Кук, Д. Машинное обучение с использованием библиотеки H2O / Д. Кук ; перевод с английского А. Б. Огурцова. — Москва : ДМК Пресс, 2018. — 250 с. — ISBN 978-5-97060-508-0	https://e.lanbook.com/book/97353
5	Флах, П. Машинное обучение. Наука и искусство построения алгоритмов, которые извлекают знания из данных / П. Флах. — Москва : ДМК Пресс, 2015. — 400 с. — ISBN 978-5-97060-273-7	https://e.lanbook.com/book/69955
6	Шалев-Шварц, Ш. Идеи машинного обучения : учебное пособие / Ш. Шалев-Шварц, Бен-Давид Ш. ; перевод с английского А. А. Слинкина. — Москва : ДМК Пресс, 2019. — 436 с. — ISBN 978-5-97060-673-5	https://e.lanbook.com/book/131686

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

Документация библиотеки pandas. <https://pandas.pydata.org/docs/>

Документация библиотеки matplotlib.

<https://matplotlib.org/stable/users/index>

Документация библиотеки seaborn. <https://seaborn.pydata.org/tutorial.html>

Документация библиотеки pandas_profiling. <https://pandas-profiling.github.io/pandas-profiling/docs/master/index.html#pandas-profiling>

Документация scikit-learn https://scikit-learn.org/stable/user_guide.html

Документация библиотеки lime <https://github.com/marcotcr/lime>

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

<https://journal.tinkoff.ru/short/ai-for-all/> - база данных нейронных сетей

<https://vc.ru/services/916617-luchshie-neyroseti-bolshaya-podborka-iz-top-200-ii-generatorov-po-kategoriyam> - база данных нейронных сетей

<https://github.com/abalmumcu/bert-rest-api> - профессиональная платформа для командой работы над проектов (нейронная сеть bert)

<http://library.miit.ru/> - электронно-библиотечная система Научно-технической библиотеки МИИТ

<https://proglib.io/p/raspoznavanie-obektov-s-pomoshchyu-yolo-v3-na-tensorflow-2-0-2020-11-08> - профессиональная библиотека программистов

https://yandex.cloud/ru/blog/posts/2022/12/andrey-berger-and-yandex-cloud?utm_referrer=https%3A%2F%2Fyandex.ru%2F – библиотека профессиональных статей разработчиков Яндекс

<https://yandex.cloud/ru/blog> - библиотека профессиональных статей разработчиков Яндекс

<https://tproger.ru/translations/opencv-python-guide> - библиотека основных команд OpenCV

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

MS Office (Excel, Word)

Notepad++

Браузер Chrome

Anaconda

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

Компьютер преподавателя

Компьютеры студентов

Проектор

Экран для проектора

Маркерная доска

9. Форма промежуточной аттестации:

Зачет в 4 семестре.

Курсовая работа в 4 семестре.

10. Оценочные материалы.

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

Авторы:

доцент, к.н. Академии "Высшая
инженерная школа"

Б.В. Игольников

Согласовано:

Директор

Б.В. Игольников

Руководитель образовательной
программы

О.Б. Проневич

Председатель учебно-методической
комиссии

Д.В. Паринов