Документ подписан простой электронной подписью
Информация о владельце:
ФИО: Артеменков Михаил Николаевич
Должность: Ректор
Дата подписания: 21.10.2026 15:51:46
Уникальный программный ключ:
a63b74166e8f4fc17bd1e0462725f79e393bfdf9

Министерство науки и высшего образования Российской Федерации
Федеральное государственное бюджетное образовательное учреждение
высшего образования
«Смоленский государственный университет»

«Утверждаю»
Проректор по учебно-
методической деятельности
Ю.А. Устименко
«25» декабря 2025 г.

Рабочая программа дисциплины
ОП.13 Инжиниринг данных

Специальность: 09.02.13 Интеграция решений с применением технологий искусственного интеллекта
Квалификация выпускника: специалист по работе с искусственным интеллектом
Нормативный срок обучения: 3 года 10 месяцев
Форма обучения: очная

Смоленск
2025

Содержание:

1. Паспорт рабочей программы дисциплины
1.1. Область применения программы
1.2 Место дисциплины в структуре ППССЗ
1.3 Общая характеристика учебной дисциплины
2. Структура и содержание дисциплины
2.1 Объем и виды учебной работы
2.2 Тематический план и содержание дисциплины
3. Условия реализации дисциплины
3.1 Требования к материально-техническому обеспечению
3.2. Информационное обеспечение реализации дисциплины
3.3 Программное обеспечение реализации дисциплины
4. Контроль и оценка результатов освоения дисциплины

1. Паспорт рабочей программы предмета

Целью реализации программы является достижение выпускниками планируемых результатов: компетенций и компетентностей, определяемых личностными, семейными, общественными, государственными потребностями и возможностями обучающегося старшего школьного возраста, индивидуальной образовательной траекторией его развития и состоянием здоровья.

1.1 Область применения программы

Реализация среднего профессионального образования по программе подготовки специалистов среднего звена в соответствии с ФГОС СПО по специальности 09.02.13 Интеграция решений с применением технологий искусственного интеллекта, утвержденного приказом Минпросвещения России от 24.12.2024 г. N 1025.

1.2. Место дисциплины в структуре образовательной программы

Дисциплина ОП.13 Инжиниринг данных входит в часть, формируемую участниками образовательных отношений (вариативная часть), общепрофессионального цикла учебного плана образовательной программы в соответствии с ФГОС СПО по специальности 09.02.13 Интеграция решений с применением технологий искусственного интеллекта. Особое значение дисциплина имеет при формировании и развитии следующих компетенций в соответствии с ФГОС СПО: ОК 01.; ОК 02.; ПК 2.2.; ПК 2.4.; ПК 2.5.

1.3 Общая характеристика дисциплины

Цель дисциплины: формирование знаний о принципах, методах и инструментах инжиниринга данных, освоение полного цикла работы с данными — от получения и очистки до преобразования и загрузки (ETL/ELT), развитие навыков построения надежных и эффективных конвейеров данных (data pipelines) для обеспечения качественными данными систем анализа и машинного обучения.

Задачи дисциплины:

– Сформировать системное понимание полного жизненного цикла данных в контексте проектов с использованием ИИ, роли инженера данных и места инжиниринга данных в процессе Data Science.

– Освоить методы и инструменты извлечения данных из разнородных источников: реляционные и NoSQL БД, файлы (CSV, JSON, XML), веб-API, потоковые источники.

– Развить умение проводить комплексную очистку, валидацию и трансформацию данных для приведения их к виду, пригодному для анализа и обучения моделей.

– Научить проектировать и реализовывать простые конвейеры данных (data pipelines) с использованием современных фреймворков и облачных сервисов.

– Сформировать навыки формирования требований к хранилищам данных (Data Warehouses, Data Lakes) для задач машинного обучения, а также подготовки и передачи curated data (курированных данных) в базы знаний и для обучения моделей.

Планируемые результаты освоения дисциплины

В результате освоения учебной дисциплины обучающийся должен **знать**:

– З1 – основные концепции и этапы инжиниринга данных: ETL (Extract, Transform, Load) и ELT, роль data pipeline в инфраструктуре данных для ИИ;

– З2 – разнообразие источников данных и методы их извлечения: запросы к SQL/NoSQL БД, чтение файлов различных форматов, работа с REST API, основы потоковой передачи данных (streaming);

– З3 – типовые проблемы качества данных (пропуски, выбросы, дубликаты, несогласованность) и методы их обработки: импутация, фильтрация, нормализация, стандартизация;

– З4 – основные операции трансформации данных: агрегация, джойны (joins), оконные функции, работа с временными рядами, создание признаков (feature engineering) для ML;

– З5 – архитектурные паттерны и технологии хранения данных для аналитики и ML: Data Warehouses, Data Lakes, их отличия, принципы организации витрин данных (data marts);

– З6 – инструменты и фреймворки для инжиниринга данных: Apache Airflow для оркестрации, Pandas для обработки, Spark (PySpark) для больших данных, облачные сервисы (Google Dataflow, AWS Glue);

– З7 – принципы мониторинга и обеспечения надежности конвейеров данных, версионирования данных и кода (Data Version Control), основы документирования процессов.

В результате освоения дисциплины обучающийся должен **уметь:**

– У1 – проектировать схему процесса обработки данных (конвейер) под конкретную задачу анализа или обучения модели;

– У2 – извлекать данные из различных источников с использованием SQL-запросов, библиотек (pandas, requests) и коннекторов;

– У3 – проводить комплексную очистку и предобработку наборов данных: обработка пропусков и аномалий, приведение типов, кодирование категориальных признаков;

– У4 – выполнять преобразования данных: агрегирование, объединение таблиц, создание новых производных признаков, ресемплирование временных рядов;

– У5 – формировать технические требования к структуре и характеристикам хранилища данных, оптимизированного для последующего доступа ML-алгоритмов;

– У6 – реализовывать и выполнять простые конвейеры данных с помощью инструментов оркестрации (например, Apache Airflow) или скриптов на Python;

– У7 – подготавливать и передавать окончательные, готовые к использованию (production-ready) наборы данных в хранилища или напрямую для обучения моделей ИИ.

В результате освоения дисциплины обучающийся должен иметь **практический опыт:**

– П1 – сквозной обработки данных от "сырого" состояния до пригодного для анализа формата на реальных или учебных наборах;

– П2 – построения отказоустойчивых и воспроизводимых конвейеров данных с использованием современных инструментальных средств;

– П3 – администрирования процессов загрузки и обновления данных в хранилищах, мониторинга их качества и выполнения;

– П4 – взаимодействия с аналитиками и ML-инженерами для понимания требований к данным и поставки качественных датасетов для решения бизнес-задач.

Результатом освоения дисциплины является овладение обучающимися общими (ОК) и профессиональными (ПК) компетенциями:

Код	Наименование результата обучения
ОК 01	Выбирать способы решения задач профессиональной деятельности применительно к различным контекстам
ОК 02	Использовать современные средства поиска, анализа и интерпретации информации и информационные технологии

	для выполнения задач профессиональной деятельности
ПК 2.2	Осуществлять процедуры администрирования баз данных
ПК 2.4	Формировать требования хранилищ банка данных для обучения
ПК 2.5	Подготавливать данные для базы знаний

2. Структура и содержание предмета

2.1 Объем и виды учебной работы

Вид учебной работы	Объем, академические часы
7 семестр	
Максимальная учебная нагрузка (всего)	60
Обязательная аудиторная учебная нагрузка (всего)	54
Лекции	16
Лабораторные занятия	38
Самостоятельная работа обучающегося	6
Форма промежуточной аттестации: зачет с оценкой	
Всего часов по предмету	60

2.2 Тематический план и содержание предмета

Наименование разделов и тем	**Содержание учебного материала и формы организации деятельности обучающихся**	**Объем, акад. ч. / в том числе в форме практической подготовки, ак.ч.**	**Коды компетенций, формированию которых способствует элемент программы**
7 семестр			
Раздел 1. Основы инжиниринга данных и управление источниками			
Тема 1.1. Введение в инжиниринг данных. Конвейеры (Data Pipelines)	1. Роль и задачи инженера данных в цикле Data Science и ML. 2. Концепции ETL и ELT. Жизненный цикл данных. 3. Архитектура и компоненты современного конвейера данных.	**6** **из них:**	З1, У1, П4, ОК 01
	Лекционные занятия	2	
	Лабораторные занятия	4	
Тема 1.2. Источники данных и методы их извлечения (Extract)	1. Типы источников: реляционные БД, NoSQL, файлы (CSV, JSON, Parquet), REST API, потоковые данные. 2. Извлечение данных из SQL-баз с помощью библиотек (SQLAlchemy, psycopg2).	**6** **из них:**	З2, У2, П1, ОК 02, ПК 2.2

	3. Работа с файлами и веб-API (requests, pandas). Основы Apache Kafka для потоков.		
	Лекционные занятия	2	
	Лабораторные занятия	4	
Тема 1.3. Организация процессов и администрирование потоков данных	1. Основы администрирования заданий по загрузке и обновлению данных. 2. Мониторинг процессов ETL/ELT, обработка ошибок и повторные попытки. 3. Введение в инструменты оркестрации (Apache Airflow) и концепцию DAG.	**6** **из них:**	З7, У6, П3, ОК 02, ПК 2.2
	Лекционные занятия	2	
	Лабораторные занятия	4	
Раздел 2. Обработка и трансформация данных для анализа и ML			
Тема 2.1. Обеспечение качества данных: очистка и валидация	1. Типовые проблемы данных: пропуски, выбросы, дубликаты, некорректные форматы. 2. Методы обработки: импутация, фильтрация, стандартизация. 3. Валидация данных с использованием схем (Great Expectations, Pandera).	**8** **из них:**	З3, У3, П1, ОК 01, ПК 2.5
	Лекционные занятия	2	
	Лабораторные занятия	6	
Тема 2.2. Трансформация данных и feature engineering	1. Основные операции трансформации: агрегация, объединение (joins), оконные функции. 2. Работа с датами и временными рядами. 3. Создание признаков (feature engineering) для задач машинного обучения.	**8** **из них:**	З4, У4, У7, П1, ПК 2.5
	Лекционные занятия	2	
	Лабораторные занятия	4	
	Самостоятельная работа обучающихся: изучение теоретического материала, подготовка к практическим занятиям, выполнение индивидуального/группового задания.	2	
Тема 2.3. Инструменты для работы с большими объемами данных	1. Ограничения pandas и переход к распределенным вычислениям. 2. Введение в Apache Spark (PySpark) для обработки больших данных. 3. Основные операции с RDD и DataFrames в PySpark.	**8** **из них:**	З6, У4, П2, ОК 02
	Лекционные занятия	2	

	Лабораторные занятия	4	
	Самостоятельная работа обучающихся: изучение теоретического материала, подготовка к практическим занятиям, выполнение индивидуального/группового задания.	2	
Раздел 3. Хранилища данных, оркестрация и итоговый проект			
Тема 3.1. Хранилища данных для аналитики и ML (Data Warehouse, Data Lake)	1. Архитектура и принципы работы хранилищ данных (Data Warehouse) и озер данных (Data Lake). 2. Модели данных: звезда, снежинка. Витрины данных (Data Marts). 3. Формирование требований к хранилищу для обучения моделей ИИ (доступность, структура, версионность).	**6** **из них:**	З5, У5, П3, П4, ОК 01, ПК 2.4
	Лекционные занятия	2	
	Лабораторные занятия	4	
Тема 3.2. Построение и оркестрация конвейера данных (Airflow)	1. Проектирование Directed Acyclic Graph (DAG) в Apache Airflow. 2. Создание задач (Operators) для извлечения, трансформации и загрузки. 3. Запуск, планирование и мониторинг конвейера.	**6** **из них:**	З1, З6, З7, У6, П2, П3, ОК 02
	Лекционные занятия	2	
	Лабораторные занятия	4	
Тема 3.3. Интеграционный проект: Подготовка датасета для обучения модели	1. Сквозная задача: от получения ТЗ от ML-инженера до поставки подготовленного датасета. 2. Реализация конвейера: извлечение из нескольких источников, очистка, трансформация, загрузка в staging-зону. 3. Документирование процесса и метаданных, финальная подготовка данных для базы знаний или обучения.	**6** **из них:**	У1, У2, У3, У4, У5, У7, П1, П2, П3, П4, ОК 01, ОК 02, ПК 2.4, ПК 2.5
	Лабораторные занятия	4	
	Самостоятельная работа обучающихся: изучение теоретического материала, подготовка к практическим занятиям, выполнение индивидуального/группового задания.	2	
Всего:			

Лекции	**16**	
Лабораторные занятия	**38**	
Самостоятельная работа обучающегося	**6**	
Промежуточная аттестация – зачет с оценкой		
Всего	**60**	

3. Условия реализации дисциплины

3.1 Требования к материально-техническому обеспечению

Для реализации программы учебной дисциплины предусмотрены следующие специальные помещения:

1. Специальное помещение: учебная аудитория для проведения занятий лекционного типа, занятий семинарского типа, групповых и индивидуальных консультаций, текущего контроля, промежуточной и государственной итоговой аттестации, оснащенное:

- стандартная учебная мебель (30 посадочных мест);
- стол и стул для преподавателя (по 1 шт.);
- кафедра для лектора (1 шт.);
- доска настенная трехэлементная (1 шт.);
- интерактивная доска (напольная) SMART BOARD (1 шт.);
- мультимедийный проектор NEC (1 шт.);
- ноутбук Asus(1 шт.);
- колонки Genius(1 пара);
- компьютерный студенческий стол (15 шт.);
- компьютерный стол
преподавателя (1 шт.);
- монитор Acer (15 шт.);
- системный блок Kraftway (15 шт.);
- принтер Canon (1 шт.);
- наушники с микрофоном (15 шт.)

2. Помещение для организации самостоятельной и воспитательной работы, оснащенное компьютерной техникой с возможностью подключения к информационно-телекоммуникационной сети «Интернет» и доступом в электронную информационно-образовательную среду организации, оснащенное:

- персональные компьютеры с выходом в Интернет (15 шт.);
- компьютерный студенческий стол (15 шт.);
- стол и стул для преподавателя (по 1 шт.);
- доска настенная трехэлементная (1 шт.);
- интерактивная доска IQBoard;
- мультимедиа проектор Optoma PX 329 DLP;
- акустическая система 2.0 (3 шт.);
- наушники с микрофоном (15 шт.)

3.2. Информационное обеспечение реализации программы

3.2.1. Обязательные печатные и электронные издания

1. Стружкин, Н. П. Базы данных: проектирование : учебник для среднего профессионального образования / Н. П. Стружкин, В. В. Годин. — Москва : Издательство Юрайт, 2025. — 477 с. — (Профессиональное образование). — ISBN 978-5-534-11635-9. — Текст : электронный // Образовательная платформа Юрайт [сайт]. — URL: https://urait.ru/bcode/566509

3.2.2. Дополнительные источники

Тарасов, С. В. СУБД для программиста. Базы данных изнутри / С. В. Тарасов. - СУБД для программиста. Базы данных изнутри,2022-05-25. - Электрон. дан. (1 файл). - Москва :

СОЛОН-Пресс, 2018. - 320 с. - электронный. – Книга находится в премиум-версии ЭБС IPR BOOKS. - ISBN 978-2-7466-7383-0

Бродовская,, Е. В. Большие данные в исследовании политических процессов : учебное пособие / Е. В. Бродовская, А. Ю. Домбровская. - Большие данные в исследовании политических процессов, Весь срок охраны авторского права. - Электрон. дан. (1 файл). - Москва : Московский педагогический государственный университет, 2018. - 88 с. - электронный. - Книга находится в премиум-версии ЭБС IPR BOOKS. - ISBN 978-5-4263-0712-4

3.2.3. Перечень ресурсов информационно-телекоммуникационной сети «Интернет»

1. Национальный открытый университет (intuit.ru).
2. Национальная платформа открытого образования (opened.ru)

3.3 Программное обеспечение

KasperskyEndpointSecurity для бизнеса Стандартный АО «Лаборатория Касперского», лицензия 1FB6-161215-133553-1-6231.

Microsoft Open License, лицензия 49463448 в составе:

1. Microsoft Windows Professional 7 Russian;
2. Microsoft Office 2010 Russian.

Обучающимся обеспечен доступ к ЭБС «Юрайт», ЭБС «IPRbooks», доступ в электронную информационно-образовательную среду университета, а также доступ к современным профессиональным базам данных и информационным справочным системам.

4. Контроль и оценка результатов освоения дисциплины

Контроль и оценка результатов освоения учебной дисциплины осуществляется преподавателем в процессе проведения лабораторных занятий, тестирования, а также выполнения обучающимися индивидуальных заданий, проектов, исследований.

Курсовые работы по данной дисциплине учебным планом не предусмотрены.

Результаты обучения	Формы и методы контроля и оценки результатов обучения
Предметные результаты обучения Обучающийся должен **знать/иметь представление:** – З1 – основные концепции и этапы инжиниринга данных: ETL (Extract, Transform, Load) и ELT, роль data pipeline в инфраструктуре данных для ИИ; – З2 – разнообразие источников данных и методы их извлечения: запросы к SQL/NoSQL БД, чтение файлов различных форматов, работа с REST API, основы потоковой передачи данных (streaming); – З3 – типовые проблемы качества данных (пропуски, выбросы, дубликаты, несогласованность) и методы их обработки: импутация, фильтрация, нормализация, стандартизация; – З4 – основные операции трансформации данных: агрегация, джойны (joins), оконные функции, работа с временными рядами, создание признаков (feature engineering) для ML; – З5 – архитектурные паттерны и технологии хранения данных для аналитики и ML: Data Warehouses,	Текущий контроль: - устный опрос; - оценка письменных работ; - защита лабораторных работ; - тестирование (письменное или компьютерное) по разделам; - проверка конспектов; - проектные задания; - решение кейсов; Промежуточная аттестация: - зачет с оценкой

Data Lakes, их отличия, принципы организации витрин данных (data marts); – З6 – инструменты и фреймворки для инжиниринга данных: Apache Airflow для оркестрации, Pandas для обработки, Spark (PySpark) для больших данных, облачные сервисы (Google Dataflow, AWS Glue); – З7 – принципы мониторинга и обеспечения надежности конвейеров данных, версионирования данных и кода (Data Version Control), основы документирования процессов. **уметь/владеть:** – У1 – проектировать схему процесса обработки данных (конвейер) под конкретную задачу анализа или обучения модели; – У2 – извлекать данные из различных источников с использованием SQL-запросов, библиотек (pandas, requests) и коннекторов; – У3 – проводить комплексную очистку и предобработку наборов данных: обработка пропусков и аномалий, приведение типов, кодирование категориальных признаков; – У4 – выполнять преобразования данных: агрегирование, объединение таблиц, создание новых производных признаков, ресемплирование временных рядов; – У5 – формировать технические требования к структуре и характеристикам хранилища данных, оптимизированного для последующего доступа ML-алгоритмов; – У6 – реализовывать и выполнять простые конвейеры данных с помощью инструментов оркестрации (например, Apache Airflow) или скриптов на Python; – У7 – подготавливать и передавать окончательные, готовые к использованию (production-ready) наборы данных в хранилища или напрямую для обучения моделей ИИ. – П1 – сквозной обработки данных от "сырого" состояния до пригодного для анализа формата на реальных или учебных наборах; – П2 – построения отказоустойчивых и воспроизводимых конвейеров данных с использованием современных инструментальных средств; – П3 – администрирования процессов загрузки и обновления данных в хранилищах, мониторинга их качества и выполнения; – П4 – взаимодействия с аналитиками и ML-инженерами для понимания требований к данным и поставки качественных датасетов для решения бизнес-задач.	