

ОБЩЕСТВО С ОГРАНИЧЕННОЙ ОТВЕТСТВЕННОСТЬЮ «IT»



УТВЕРЖДАЮ

Генеральный директор

В.В. Кармаза

» 2022 г.

ДОПОЛНИТЕЛЬНАЯ ПРОФЕССИОНАЛЬНАЯ ПРОГРАММА
ПОВЫШЕНИЯ КВАЛИФИКАЦИИ
«Инженер данных»

Специальность: Инженер данных (Data Engineer)

Объем: 260 академических часов.

Форма обучения: очно-заочная, с применением ЭО и ДОТ

Москва, 2022

АННОТАЦИЯ ПРОГРАММЫ

Программа направлена на получение профессиональных компетенций по профессии Инженер данных.

#инженер_данных #data_engineer #большие_данные #big_data

Программа рассчитана на слушателей, имеющих базовое математическое или техническое образование.

Результаты обучения: у слушателей будут сформированы необходимые компетенции для профессиональной деятельности специалиста (Инженера данных) в области искусственного интеллекта и больших данных при реализации проектов и разработки новых решений на основе данных.

Количество часов на практические задания 185 часов – более 70%.

В результате освоения программы «Инженер данных» выпускник должен:

Знать:

- основные принципы, процесс, стадии и методологии разработки решений на основе ИИ;
- основные платформы данных (облачные и внутрикорпоративные), применяемые при разработке решений на основе ИИ;
- основные технологии Data Science и BigData, используемые для решения практических задач;
- архитектуру и принципы работы специализированных решений, созданных на основе ИИ;
- основные виды представления данных: табличные, графовые, временные ряды;
- основы массово-параллельной обработки и анализ данных;
- основы методов оптимизации;
- принципы работы NoSQL баз данных;
- принципы работы реляционных баз данных;
- язык запросов к СУБД - SQL
- области применений массово-параллельных вычислений для ускорения машинного обучения;
- технологии распределенной обработки данных;

- основы работы с распределенными кластерными системами;
- синтаксис и семантику основных языков искусственного интеллекта, и основные приемы программирования на них;
- основные инструменты анализа данных и ML (RapidMiner);
- основные уровни представления данных;
- основы потоковой обработки данных;
- основные ETL процессы и инструменты.

Уметь:

- применять системы обработки и анализа больших массивов данных;
- анализировать требования по сбору данных
- применять SQL базы данных для прикладных решений;
- применять методы оценки качества данных;
- применять инструменты, технологии и библиотеки Data Science при разработке решений на основе ИИ;
- применять язык программирования Python и библиотеки при разработке решений на основе ИИ;
- использовать основные статистические методы анализа данных при решении профессиональных задач;
- реализовывать методы машинного обучения;
- осуществлять поиск и структурирование данных;
- визуализировать анализируемые данные;
- применять методы машинного обучения для работы с большими данными;
- использовать шины данных (kafka) при обработке больших данных в рамках решения профессиональных задач;
- использовать цифровые платформы анализа данных для решения профессиональных задач.

Владеть:

- языками запросов для обращения к СУБД;
- языками программирования для реализации методов подготовки данных;
- основными принципами, возможностями и порядком применения метода ближайших соседей (kNN) для решения задач классификации данных;
- методами машинного обучения и нейронных сетей;

- математическими методами анализа данных;
- культурой постановки и планирования последовательности решения задач анализа данных и классификации;
- навыками анализа реальных задач из различных предметных областей на уровне отдельных подходов и коллективами алгоритмов.
- методами случайных лесов (Random forest) для решения задач классификации данных;
- библиотеками, программными платформами (фреймворки) и программными комплексами машинного обучения.

1 ОБЩАЯ ХАРАКТЕРИСТИКА ПРОГРАММЫ

Целью программы повышения квалификации «Инженер данных» является совершенствование и (или) получение новых профессиональных компетенций, необходимых для профессиональной деятельности, и (или) повышение профессионального уровня в области искусственного интеллекта и больших данных при реализации проектов и разработки новых решений на основе данных.

Программа является программой повышения квалификации и разработана с учетом потребностей слушателей, которые хотят продолжить карьеру в IT в качестве инженера данных.

Программа предназначена для изучения основ профессии Data Engineer.

Программа направлена на формирование следующих профессиональных компетенций:

- способен классифицировать и идентифицировать задачи искусственного интеллекта, выбирать адекватные методы и инструментальные средства решения задач искусственного интеллекта;
- способен разрабатывать и применять методы машинного обучения для решения задач;
- способен использовать инструментальные средства для решения задач машинного обучения;
- способен создавать и поддерживать системы искусственного интеллекта на основе нейросетевых моделей и методов;
- способен разрабатывать системы анализа больших данных;
- способен создавать и внедрять одну или несколько сквозных цифровых субтехнологий искусственного интеллекта.

Основной формой реализации Программы является очно-заочная форма с использованием электронного обучения и дистанционных образовательных технологий.

Количество часов, отведённое на выполнение практических заданий, в т.ч. в режиме самостоятельной работы, составляет 185 часов (более 70%).

1.1 Актуальность

Стремительность проникновения цифровых технологий во все сферы деятельности Российской Федерации служит причиной существенных изменений привычных моделей экономического и социального уклада государства и условием его инклюзивного устойчивого

развития, что и определило приоритетность тенденции цифровизации. В частности, цифровизация экономики Российской Федерации на сегодняшний день в рамках национального проекта «Национальная программа «Цифровая экономика Российской Федерации» (утв. президиумом Совета при Президенте Российской Федерации по стратегическому развитию и национальным проектам, протокол от 04.06.2019 № 7) определена одним из приоритетных направлений государственной политики и рассматривается в качестве инновационной составляющей конкурентного поступательного движения и крупномасштабных качественных преобразований, центрального звена стратегического управления и фундаментального компонента экономического суверенитета государства.

Интеграция физических и цифровых ресурсов в контексте цифровизации выступает основным средством трансформации системы образования, которая посредством использования инновационных цифровых методов, технологий и инструментов направлена на актуализацию образовательного процесса в аспекте соответствия мировым тенденциям информатизации общества для повышения качественного уровня профессиональной подготовки специалистов для стремительно меняющейся и продолжающей свое формирование цифровой экономики. Этому процессу в образовательной среде присущи форсированная автоматизация и роботизация производства и управления, коренные изменения социальных структур – расширение сферы информационной деятельности и услуг, что в совокупности, согласно Указу Президента Российской Федерации от 7 мая 2018 г. № 204 «О национальных целях и стратегических задачах развития Российской Федерации на период до 2024 года», является одной из национальных целей развития Российской Федерации на период до 2024 года.

Формирование единого образовательного пространства, обеспечение высококачественного и доступного образования всех видов и уровней на территории Российской Федерации являются основной миссией федерального проекта «Цифровая образовательная среда» национального проекта «Образование» (утв. президиумом Совета при Президенте Российской Федерации по стратегическому развитию и национальным проектам, протокол от 3.09.2018 №10), при этом функционирование цифровой образовательной среды невозможно без использования соответствующих образовательных информационных ресурсов (электронных и интерактивных обучающих материалов), создание и внедрение в образовательный процесс которых определено, как один из ожидаемых результатов реализации мероприятий государственной программы «Развитие образования» (утв. постановлением Правительства Российской Федерации от 26.12.2017 № 1642).

На современном этапе системе образования отведена роль первостепенной важности по обеспечению устойчивой трансформации общества в цифровую эпоху, ориентированную на повышение производительности, новые типы трудовой деятельности, потребности человека, что осуществимо исключительно путем включения в образовательный процесс всех групп населения, построения индивидуальных траекторий обучения, управления собственными результатами обучения, виртуальной и дополненной реальности.

Роль и характер работы в разных сферах экономики Российской Федерации продолжают трансформироваться – появление новых компетенций и навыков применения инновационных инструментов позволит результативно и системно планировать карьерное развитие, что актуализирует задачу обеспечения непрерывного развития кадров в соответствии с глобальными трендами цифровизации. Эффективная реализация национальных стратегических целей предполагает широкое внедрение искусственного интеллекта, как генерального драйвера технологических перемен, а это значит, что анализ данных и машинное обучение должны стать элементом массового образования.

Под искусственным интеллектом в Программе развития в соответствии с «Национальной стратегией развития искусственного интеллекта на период до 2030 года», утверждённой Указом Президента РФ от 10.10.2019 г. № 409 (далее – Стратегия), понимается комплекс технологических решений, позволяющий имитировать когнитивные функции человека (включая самообучение и поиск решений без заранее заданного алгоритма) и получать при выполнении конкретных задач результаты, сопоставимые, как минимум, с результатами интеллектуальной деятельности человека. Для продуктивной реализации задач указанной Стратегии на основе технологий искусственного интеллекта необходимо повышение эффективности процессов планирования, прогнозирования и принятия управленческих решений, использование автономного интеллектуального оборудования и робототехнических комплексов, интеллектуальных систем управления логистикой.

Одним из приоритетных направлений научно-технического развития Российской Федерации согласно Стратегии научно-технологического развития Российской Федерации, утвержденной Указом Президента Российской Федерации от 1 декабря 2016 г. № 642, обозначено создание технологий, являющихся основой инновационного развития внутреннего рынка продуктов и услуг, что может быть осуществлено за счет перехода к цифровым, интеллектуальным производственным технологиям, роботизированным системам, новым материалам и способам конструирования, создания систем обработки больших объемов данных, машинного обучения и искусственного интеллекта.

Федеральный проект «Искусственный интеллект» предусматривает повышение уровня кадрового обеспечения российского рынка технологий искусственного интеллекта, внедрение инновационных решений в разные сферы хозяйствования, повышение доступности и качества данных, необходимых для развития технологий искусственного интеллекта.

Следовательно, первоочередным на нынешнем этапе является формирование принципиально новых направлений деятельности во всех отраслях экономики на основе технологий искусственного интеллекта, что предполагает одновременно дальнейшее развитие целостной системы устойчивого воспроизводства и привлечения кадров при тесном сотрудничестве и взаимодействии научных и образовательных организаций, предпринимательского сообщества, государственных корпораций и компаний, создание новых высокопроизводительных рабочих мест и повышение уровня занятости населения, оптимизация процессов подбора и обучения кадров.

На сегодняшний день данные – это жизнеобеспечение любой компании, «кровь» предприятия, без чего невозможно его существование. Поэтому в любой сфере национальной экономики функционирование организации должно быть реализовано на основе гибкого управления данными, в том числе большими (Big Data), и оно станет успешным, если все её бизнес-процессы и организационная структура построены на основе сквозной интеграции информационных потоков, их непрерывной и прогнозной предиктивной аналитики.

Специалисты – Data Professionals отвечают за сбор, преобразование, управление данными, а также извлечение полезных для бизнеса сведений из массивов Big Data и мониторинг их жизненного цикла. При этом весьма важными навыками являются такие, как создание и поддержка инфраструктуры Big Data, обеспечение сбора, хранения и управление потоками данных в реальном времени, анализ структурированных и неструктурированных данных с помощью методов статистики, алгоритмов машинного обучения (Machine Learning) и предиктивной аналитики, что актуализирует задачу подготовки кадров соответствующей квалификации (инженеров данных) с учетом потребностей бизнеса к квалификации специалистов в области искусственного интеллекта.

С каждым годом объемы информации, поступающие из множества разнородных источников (Интернет, технологическое оборудование, корпоративные системы и хранилища данных), неумолимо растут, представляя собой данные о рабочих процессах и продуктах, текущем и прогнозном состояниях внешней среды, настоящих и будущих клиентских потребностях, при этом совершенствуются и программные инструменты Big Data, обеспечивающие обработку и анализ собранной информации для получения конечных вариантов всевозможных ситуативных решений. В дальнейшем на основе предложенных

аналитических и прогнозных моделей машинного обучения (Machine Learning) менеджмент компании принимает управленческие решения, которые, например, позволят ей увеличить прибыль или сократить затраты.

Одна из ключевых задач инженера данных – сделать процесс анализа данных в компании максимально удобным для аналитиков, обеспечить их очищенными достоверными данными в должном количестве и в должный срок, а это, в свою очередь, будет обеспечено посредством ряда следующих шагов:

- определение основных источников данных и их сбор;
- выделение внутренней информации (финансовые отчеты, корпоративные базы данных, заявки на сайте и внешние источники – социальные сети, новостные заголовки, Интернет вещей);
- «очистка» данных;
- построение архитектуры процесса обработки данных и визуализация процесса разработки продукта;
- разработка моделей процессов;
- использование необходимого инструментария хранения и обработки больших массивов данных;
- превращение моделей в готовый продукт или сервис.

За последние годы появилось большое число новых технологий, и бизнес нуждается в профессионалах, которые в них разбираются. Система подготовка специалистов профильными образовательными организациями не в полной мере успевает за прогрессом в IT-сфере с учетом ускорения темпов цифровизации, поэтому они не могут достаточно быстро обучиться новым инструментам, а поскольку спрос на таких специалистов, как инженеры данных, очень высокий, остроактуальной становится подготовка кадров в рамках дополнительного профессионального образования. Образовательная программа в рамках специальности «Инженер данных», направленная на обеспечение получения гражданами дополнительного профессионального образования в области искусственного интеллекта и в смежных областях с использованием механизма персональных цифровых сертификатов, обеспечит приобретение следующих профессиональных компетенций:

- способность классифицировать и идентифицировать задачи искусственного интеллекта, выбирать адекватные методы и инструментальные средства решения задач искусственного интеллекта;
- способность разрабатывать и применять методы машинного обучения для решения задач;

- способность использовать инструментальные средства для решения задач машинного обучения;
- способность создавать и поддерживать системы искусственного интеллекта на основе нейросетевых моделей и методов;
- способность разрабатывать системы анализа больших данных;
- способность создавать и внедрять одну или несколько сквозных цифровых субтехнологий искусственного интеллекта.

1.2 Требования к уровню подготовки слушателя

Образование: высшее (Математические науки, Инженерное дело, технологии и технические науки)

Квалификация: специалист (Математические науки, Инженерное дело, технологии и технические науки), бакалавр (Математические науки, Инженерное дело, технологии и технические науки) или магистр (Математические науки, Инженерное дело, технологии и технические науки)

Наличие опыта профессиональной деятельности: без опыта.

Требования к уровню подготовленности, определяемому контрольно-измерительными материалами.

Слушатели должны обладать следующими знаниями, умениями и владеть навыками:

PYTHON

- Знание синтаксиса языка
- Понимание базовых структур данных
- Владение основами ООП (класс, объект)

SQL

- Знание базового синтаксиса (SELECT, WHERE, GROUP BY, HAVING)
- Умение составлять подзапросы и делать все виды JOIN
- Навык работы с оконными функциями

ИНФРАСТРУКТУРА

- Умение работать с командной строкой
- Навыки работы с Docker

- Знание базовых команд Linux
- Опыт работы с Git

Требования к компетенциям, которыми должен обладать гражданин при поступлении на Образовательную программу:

Способен использовать и адаптировать существующие математические методы и системы программирования для разработки и реализации алгоритмов решения прикладных задач;

Способен применять и модифицировать математические модели для решения задач в области профессиональной деятельности;

Владеет широкой общей подготовкой (базовыми знаниями) для решения практических задач в области информационных систем и технологий;

Способен использовать современные компьютерные технологии поиска информации для решения поставленной задачи, критического анализа этой информации и обоснования принятых идей и подходов к решению;

Способен выбирать и оценивать способ реализации информационных систем и устройств (программно-, аппаратно- или программно-аппаратно-) для решения поставленной задачи.

1.3 Планируемые результаты обучения

Программа повышения квалификации разработана с учетом профессионального стандарта 06.042 Специалист по большим данным

Программа повышения квалификации разработана с учетом:

ФГОС 01.04.02 ПРИКЛАДНАЯ МАТЕМАТИКА И ИНФОРМАТИКА,

ФГОС 09.04.02. ИНФОРМАЦИОННЫЕ СИСТЕМЫ И ТЕХНОЛОГИИ.

Совершенствуемые и/или формируемые компетенции	Тип компетенции	Планируемые результаты обучения (знать, уметь, владеть - использовать конкретные инструменты)
--	-----------------	---

Способен разрабатывать оригинальные алгоритмы и программные средства, в том числе с использованием современных интеллектуальных технологий, для решения профессиональных задач	Общепрофессиональная	Знать: Содержание профессии «Инженер данных»: потребность и ценность, задачи, навыки, инструменты в классификации данных. Обязанности и функция в команде. Технологии DataScience и BigData для решения практических задач. Данные и источники, характеристики, корреляция. Синтаксис и структуры в Python. Использование библиотек NumPy, SciPy для вычислений, их отличия. Обработка данных в Python. Библиотека Pandas. Библиотеки визуализации Python Matplotlib, Seaborn, Plotly, Bokeh, Altair, Folium Синтаксис языка программирования C++ Сопоставление хэш-таблицы и map в C++ Уметь: Формулировать задачи инженера данных, основные вызовы, стоящие перед ним. Различать структурированные, полуструктурированные и неструктурированные данные. Решать практические задачи используя базовые конструкции и структуры языка Python (основные функции для работы со списками и кортежами, структуру словарей и множеств). Решать практические задачи на Python используя библиотеки Pandas
---	-----------------------------	---

		Решать практические задачи на Python используя библиотеки Matplotlib, Seaborn, Plotly Решать практические задачи на C++ используя циклы, условия, функции и классы Владеть навыками: Установки и настройки среды разработки Python (Anaconda или др.) Решения задач на Python используя библиотеки NumPy и Scipy Загрузки датасета, преобразования и осуществления срезов данных, проведения описательного анализа, построения графиков распределения, визуализации разных признаков, их распределения, агрегирования признаков, выявления топа коррелируемых признаков, оценки взаимосвязи
Способен классифицировать и идентифицировать задачи искусственного интеллекта, выбирать адекватные методы и инструментальные средства решения задач искусственного интеллекта	профессиональная	Знать: Определения, история развития и главные тренды ИИ. Формирование концепции искусственного интеллекта, основные направления исследований, этапы развития и проблемы. Рынок технологий искусственного интеллекта. Рынок ИИ в России. Мировой рынок ИИ. Процесс, стадии и методологии разработки решений на основе ИИ. Архитектура и принципы работы промышленных решений, созданных на основе ИИ. Области применения искусственного интеллекта.

		Уметь: Проводить классификацию знаний. Строить модель знаний в виде графа.
Способен разрабатывать и применять методы машинного обучения для решения задач	профессиональная	Знать: Методы машинного обучения; Постановка основных классов задач в машинном обучении. Основной аппарат комбинаторики и мат. статистики. Регрессионный анализ и сжатие данных. Методы оптимизации; SQL базы данных (GreenPlum, Postgres, Oracle); NoSQL базы данных (Cassandra, MongoDB, ElasticSearch, Neo4J, Hbase); Виды представления данных: табличные, графовые, временные ряды; Уметь: Работать с программным инструментарием Data Mining и Machine Learning Проводить анализ качества построенной модели линейной регрессии с помощью библиотеки Scikit-Learn Сравнивать различные алгоритмы по эффективности Применять методы оптимизации с использованием глобальных верхних оценок, зависящих от параметра Владеть навыками: Работы с библиотекой машинного обучения Scikit-Learn

		Построения модели линейной регрессии с помощью библиотеки Scikit-Learn Решения оптимизационных задач
Способен использовать инструментальные средства для решения задач машинного обучения	профессиональная	Знать: Технологии кластеризации и классификации. Деревья решений. Потоковая обработка данных (data streaming, event processing); Шины данных (kafka); Языки программирования и библиотеки (Python, R); Языки программирования и библиотеки (C++); Платформы данных (облачные и внутрикорпоративные); Качество данных, подходы и инструменты; Инструменты анализа данных и Machine Learning (Rapid Miner); Массово параллельные вычисления для ускорения машинного обучения (GPU) в задачах машинного обучения. Преимущества графических процессоров для глубокого обучения Уметь: Проводить анализ качества построенной модели логистической регрессии Применять алгоритмы кластеризации данных Производить импорт библиотек RAPIDS в Python Владеть навыками:

		Построения модели логистической регрессии с помощью библиотеки Scikit-Learn. Реализации алгоритма построения дерева с критерием информационного выигрыша и критерием Джини и определению класса по мажоритарному классу в листе Работы с RAPIDS – открытыми библиотеками NVIDIA для Python
Способен создавать и поддерживать системы искусственного интеллекта на основе нейросетевых моделей и методов	профессиональная	Знать: Нейронные сети. Генетические алгоритмы. Онлайн подход к обучению на больших данных на примере линейных моделей. Архитектура нейронных сетей. Эффективное использование нескольких GPU. LSH. Кластеризация больших данных. Параллельные алгоритмы для обработки BigData. Архитектура и принципы работы промышленных решений, созданных на основе искусственного интеллекта Уметь: Обучать модели на выборке (в несколько гигабайт) Применять Apache Spark в оптимизации гиперпараметров Работать с системой поддержки принятия решений

		Сравнивать экспериментальные графики (обучающая выборка и тестовая выборка) Работать с Keras (Deep Learning и Python) Использовать кластеризацию k-средних для построения модели Владеть навыками: Использования Vowpal Wabbit в задаче классификации Использования метода стохастического градиентного спуска SGD Использования алгоритмов K-Means
Способен разрабатывать системы анализа больших данных	профессиональная	Знать: Обработка данных с помощью (Hadoop/Hive/Spark) Конвейер данных (Storm, Kafka) Назначение, область применения, структура, принципы работы MapReduce Распределенные файловые системы (HDFS, Object Storage): структура, достоинства, недостатки и сфера применения Статистические методы анализа данных Технологии анализа данных. Платформы данных (облачные и внутрикорпоративные). Цифровая платформа анализа данных. Экосистема Hadoop и элементы Системы Обработки Данных. Аналоги из экосистем GCP, AWS. Инструменты с открытым исходным кодом для анализа больших данных Процессы ETL в анализе BigData

		Язык запросов MDX Моделирование данных Совместное использование базы данных. Безопасность данных. Реляционные базы данных SQL: GreenPlum, Postgres, Oracle Базы данных NoSQL: Cassandra, MongoDB, ElasticSearch, Neo4J, Hbase. Массово параллельная обработка и анализ данных. Особенности организации СУБД. Облачные хранилища данных (Data Warehouses) Data Lake (озеро данных): назначение, компоненты, проблемы. Поточковая обработка данных (data streaming, event processing). Уметь: Применять методы бакетирования и партиционирования в Apache Hive и Spark в зависимости от поставленной задачи Строить архитектуру обработки данных в реальном времени – Apache Kafka Применять MapReduce для решения практических задач. Применять средства анализа данных на персональных компьютерах Проводить анализ данных используя методологию CRISP-DM. Применять методы анализа на графах Выбирать облачные технологии или on-premises инфраструктуру в зависимости от задачи Применять Rapid Miner для решения практических задач
--	--	--

		Создавать логическую и физическую модель базы данных Разрабатывать запросы для загрузки данных в формате JSON в СУБД MongoDB Моделировать данные в хранилище Data Warehouses Владеть навыками: Настройки кластера Apache Spark и Hive на Hadoop Обработки данных с помощью (Hadoop/Hive/Spark) Установки и настройки Storm и Kafka Установки, настройки, запуска MapReduce Обработки данных с помощью библиотеки tidyverse. Статистического анализа данных в R. Извлечения данных из источника, очистки данных, обогащения, трансформирования и проведения интеграции в единую целевую модель Установки базы данных PostgreSQL и PGAdmin, создания базы данных для решения практической задачи, приведения ее к 3NF, проведения ее денормализации Использования СУБД MongoDB для решения практической задачи Разработки консольной утилиты для преобразования информации в формате CSV в формат JSON
--	--	--

Способен создавать и внедрять одну или несколько сквозных цифровых субтехнологий искусственного интеллекта	профессиональная	Знать: Особенности построения рекомендательных систем на больших данных. Content-based, collaborative filtering, ALS, iALS. Цифровая платформа анализа данных; Работа с распределенной кластерной системой; ETL процессы и инструменты; Уровни предоставления данных (ODS, DDL, семантический слой, модель данных); Системы поддержки принятия решений Уметь: Создавать систему рекомендаций (по продуктам) Владеть навыками: Обучения модели content-based и collaborative фильтрации, проверки работы модели, написания выводов
--	------------------	---

2 КРИТЕРИИ ДЛЯ ОЦЕНКИ УРОВНЯ СФОРМИРОВАННОСТИ УКАЗАННЫХ КОМПЕТЕНЦИЙ И СООТВЕТСТВИЯ ИНДИКАТОРАМ ДОСТИЖЕНИЯ КОМПЕТЕНЦИЙ

Компетенция	Критерии для оценки уровня сформированности компетенций
Способен разрабатывать оригинальные алгоритмы и программные средства, в том числе с использованием современных интеллектуальных технологий, для решения профессиональных задач	Разрабатывает оригинальные алгоритмы и программные средства, в том числе с использованием современных интеллектуальных технологий, для решения профессиональных задач
Способен классифицировать и идентифицировать задачи искусственного интеллекта, выбирать адекватные методы и инструментальные средства решения задач искусственного интеллекта	Классифицирует и идентифицирует задачи систем искусственного интеллекта в зависимости от особенностей проблемной и предметной областей
Способен разрабатывать и применять методы машинного обучения для решения задач	Проводит анализ требований и определяет необходимые классы задач машинного обучения
Способен использовать инструментальные средства для решения задач машинного обучения	Осуществляет оценку и выбор инструментальных средств для решения поставленной задачи; Разрабатывает модели машинного обучения для решения задач; Создает, поддерживает и использует системы искусственного интеллекта, включающие разработанные модели и методы, с применением выбранных инструментов машинного обучения (рекомендательные системы)
Способен создавать и поддерживать системы искусственного интеллекта на основе нейросетевых моделей и методов	Осуществляет оценку и выбор моделей искусственных нейронных сетей и инструментальных средств для решения поставленной задачи

Способен разрабатывать системы анализа больших данных	Разрабатывает программные компоненты извлечения, хранения, подготовки больших данных с учетом вариантов использования больших данных, определений, словарей и эталонной архитектуры больших данных; Разрабатывает программные компоненты обработки, удаленной, распределенной и объединенной аналитики, использования результатов анализа, описания и управления качеством и достоверностью больших данных
Способен создавать и внедрять одну или несколько сквозных цифровых субтехнологий искусственного интеллекта	Участствует в реализации проектов в области сквозной цифровой субтехнологии «Рекомендательные системы и системы поддержки принятия решений»

3 УЧЕБНЫЙ (ТЕМАТИЧЕСКИЙ) ПЛАН

№ п/ п	Наименование модулей/тем программы	Всего, час	Виды учебных занятий			Формы контроля
			лекции	практические занятия	самостоятельная работа	
1	Модуль 1. Введение в инженерии больших данных. Языки программирования (Python/C++)					
2	Тема 1.1 Инженер данных. Цели, направления, задачи, обязанности и функция в команде	3	1	1	1	тест, практический кейс
3	Тема 1.2 Технологии DataScience и BigData для решения практических задач.	4	2	0	2	тест
4	Тема 1.3 Основы синтаксиса и структур в Python	8	2	4	2	тест, практический кейс
5	Тема 1.4 Использование библиотек NumPy, SciPy для вычислений	8	2	4	2	тест, практический кейс
6	Тема 1.5 Обработка данных в Python. Библиотека Pandas.	9	2	6	1	тест, практический кейс
7	Тема 1.6 Библиотеки Python для визуализации данных	7	2	4	1	тест, практический кейс

8	Тема 1.7 Синтаксис языка программирования C++	10	2	6	2	тест, практический кейс
9	Модуль 2. Системы обработки и анализа больших данных.					
10	Тема 2.1 Обработка данных с помощью (Hadoop/Hive/Spark)	10	2	6	2	тест, практический кейс
11	Тема 2.2 Конвейер данных (Storm, Kafka)	8	2	4	2	тест, практический кейс
12	Тема 2.3 Базовое представление о Map Reduce	4,5	1	2	1,5	тест, практический кейс
13	Тема 2.4 Распределенная файловая система HDFS	4	2	0	2	тест
14	Тема 2.5 Статистические методы анализа данных	3	2	0	1	тест
15	Тема 2.6 Технологии анализа данных	8	2	4	2	тест, практический кейс
16	Тема 2.7 Использование платформ данных	4	2	0	2	тест
17	Тема 2.8 Инструменты с открытым исходным кодом для анализа больших данных	7,5	2	4	1,5	тест, практический кейс

18	Тема 2.9 Что такое ETL: как справиться с анализом BigData	8	2	4	2	тест, практический кейс
19	Промежуточная аттестация	2		2		практический кейс
20	Модуль 3. Хранение данных на основе БД. Моделирование данных. Хранилища данных. Data Lake					
21	Тема 3.1 Введение в моделирование данных	8	2	4	2	тест, практический кейс
22	Тема 3.2 Реляционные базы данных	8	2	4	2	тест, практический кейс
23	Тема 3.3 Базы данных NoSQL	10	2	6	2	тест, практический кейс
24	Тема 3.4 Массово параллельная обработка и анализ данных	4	2	0	2	тест
25	Тема 3.5 Облачные хранилища данных (Data Warehouses)	10	2	6	2	тест, практический кейс
26	Тема 3.6 Data Lake (озеро данных)	4	2	0	2	тест, практический кейс
27	Модуль 4. Машинное обучение					
28	Тема 4.1 Введение в машинное обучение и обработку данных.	8	2	4	2	тест, практический кейс

	Программный инструментарий Data Mining и Machine Learning.					
29	Тема 4.2 Основной аппарат комбинаторики и мат. статистики. Регрессионный анализ и сжатие данных.	8	2	4	2	тест, практический кейс
30	Тема 4.3 Технологии кластеризации и классификации. Деревья решений. Нейронные сети. Генетические алгоритмы.	10	2	6	2	тест, практический кейс
31	Тема 4.4 Деревья решений	8	2	4	2	тест, практический кейс
32	Тема 4.5 Методы оптимизации машинного обучения	9,5	2	6	1,5	тест, практический кейс
33	Тема 4.6 GPU в задачах машинного обучения	4,5	1	2	1,5	тест, практический кейс
34	Тема 4.7 Машинное обучение на больших данных. Онлайн обучение и линейные модели	8	2	4	2	тест, практический кейс
35	Тема 4.8 Рекомендательные системы	10	2	6	2	тест, практический кейс

36	Тема 4.9 Нейронные сети	8	2	4	2	тест, практический кейс
37	Тема 4.10 LSH. Кластеризация больших данных	8	2	4	2	тест, практический кейс
38	Модуль 5. Искусственный интеллект					
39	Тема 5.1 Введение в искусственный интеллект	4	2	0	2	тест
40	Тема 5.2 Роль искусственного интеллекта в бизнесе	8	2	4	2	тест, практический кейс
41	Промежуточная аттестация	2		2		практический кейс
42	Практика на базе привлекаемой организации	10				практический кейс
	Итоговая аттестация	2		2		Решение практических кейсов
	Всего часов	260	65	122	63	

4 КАЛЕНДАРНЫЙ УЧЕБНЫЙ ГРАФИК

Начало обучения: 01.09.2022.

Завершение обучения: 15.12.2022.

№ п/п	Наименование учебных модулей/ практик/ аттестации	Трудоёмкость (час)	Срок освоения
1	Модуль 1. Введение в инженерию больших данных. Языки программирования (Python, C++)	49	01.09.2022 – 17.09.2022
2	Модуль 2. Системы обработки и анализа больших данных	57	19.09. – 15.10.2022
3	Промежуточная аттестация. Решение практико-ориентированного кейса	2	15.10.2022
4	Модуль 3. Хранение данных на основе БД. Моделирование данных. Хранилища данных. Data Lake	44	17.10. – 05.11.2022
5	Модуль 4. Машинное обучение	82	07.11. – 01.12.2022
6	Модуль 5. Искусственный интеллект	12	02.12. – 08.12.2022
7	Промежуточная аттестация. Решение практико-ориентированного кейса	2	08.12.2022
8	Практика на базе привлекаемой организации с решением задач на стендах	10	09.12. – 14.12.2022
9	Итоговая аттестация	2	15.12.2022

5 УЧЕБНАЯ (РАБОЧАЯ) ПРОГРАММА

Наименование учебных модулей и тем	Виды учебных занятий	Содержание учебных занятий
Входная диагностика		Входное тестирование с целью уточнения уровня подготовки слушателей и требуемых компетенций
Модуль 1. Введение в инженерии больших данных. Языки программирования (Python/C++)		
Тема 1.1 Инженер данных. Цели, направления, задачи, обязанности и функция в команде	Лекция (1 ч)	Инженер данных: потребность и ценность. Задачи, навыки, инструменты в классификации данных. Обязанности и функция в команде.
	Практическое занятие (1 ч)	Провести анализ рынка Инженер Данных: РФ, USA, EU
	Самостоятельная работа (1 ч)	Тест 1.1. Функции Инженера данных. Классификация данных. Сформулировать задачи инженера данных по формированию ценностей для бизнеса, основные вызовы (challenges).
Тема 1.2 Технологии DataScience и BigData для решения практических задач.	Лекция (2 ч)	Что такое большие данные? Ключевые технологии и подходы. Функции и задачи больших данных. Принципы работы с большими данными. Хранение больших данных. Большие данные в бизнесе, маркетинге. Что такое Data Science? Инструменты, библиотеки и технологии Data Science.
	Самостоятельная работа (2 ч)	Тест 1.2. Технологии DataScience и BigData Изучить материал: Данные и источники, характеристики, корреляция. Структурированные, полуструктурированные и неструктурированные данные.

Тема 1.3 Основы синтаксиса и структур в Python	Лекция (2 ч)	Введение в Python. Операторы языка. Типы данных и переменные, мутабельные/иммутабельные и простые/составные типы данных, приемы отладки. Базовые конструкции и структуры языка. Циклы и условия. Функции. Сетевые запросы. Словари и множества. Библиотеки.
	Практическое занятие (4 ч)	Решить практические задачи используя базовые конструкции и структуры языка. Решить практические задачи на Python, используя основные функции для работы со списками и кортежами, структуру словарей и множеств.
	Самостоятельная работа (2 ч)	Произвести установку и настройку среды разработки Python (Anaconda или др.). Тест 1.3. Основы синтаксиса и структур в Python
Тема 1.4 Использование библиотек NumPy, SciPy для вычислений	Лекция (2 ч)	Математический и числовой анализ с помощью библиотек NumPy, SciPy. Использование NumPy, SciPy для научных вычислений. Практическое применение библиотек. Работа и тестирование данных.
	Практическое занятие (4 ч)	Решить практические задачи на Python, используя библиотеки NumPy и Scipy. ноутбук для решения: python_jupyternotebook.ipynb  python_jupyter notebook.ipynb
	Самостоятельная работа (2 ч)	Изучить отличия SciPy от NumPy Тест 1.4. Использование библиотек NumPy, SciPy в Python.

Тема 1.5 Обработка данных в Python. Библиотека Pandas.	Лекция (2 ч)	Работа с табличными данными в Python. Библиотеки Pandas, как стандарт обработки данных. Манипуляции с данными, обработка пропусков, фильтрация и агрегация. Описательный анализ данных.
	Практическое занятие (6 ч)	Решить практические задачи на Python, используя библиотеки Pandas. Загрузить датасет, преобразовать и сделать срезы данных. Провести описательный анализ, построить графики распределения, выполнить агрегирование признаков, выявить топ коррелируемых признаков, оценить взаимосвязи. Датасет customers_orders_data.7z  customers_orders_data.7z ноутбук для решения: python_jupyternotebook.ipynb
	Самостоятельная работа (1 ч)	Определить преимущества и недостатки библиотеки Pandas. Тест 1.5. Методы библиотеки Pandas
Тема 1.6 Библиотеки Python для визуализации данных	Лекция (2 ч)	Библиотеки визуализации Python Matplotlib, Seaborn, Plotly, Bokeh, Altair, Folium. Варианты отрисовки графиков и изображений. Визуализация парных взаимосвязей, диаграммы, тепловая карта корреляции. Продвинутая визуализация. Динамические графики. Виды представления данных: табличные, графовые, временные ряды)

	Практическое занятие (4 ч)	Решить практические задачи на Python, используя библиотеки Matplotlib, seaborn, plotly. Загрузить датасет, сделать визуализацию разных признаков, их распределения, корреляции и взаимосвязей. Датасет customers_orders_data.7z ноутбук для решения: python_jupyternotebook.ipynb
	Самостоятельная работа (1 ч)	Изучить библиотеку Seaborn для визуализация парных взаимосвязей, heatmap, диаграммы, тепловая карта корреляции. Тест 1.6. Способы отрисовки данных и их связей с помощью библиотек Matplotlib, Seaborn, Plotly, Bokeh, Altair, Folium.
Тема 1.7 Синтаксис языка программирования C++	Лекция (2 ч)	Ознакомление с синтаксисом языка C++. Переменные, типы данных. Ветвления, циклы. Массивы. Строки. Указатели и ссылки. Функции. Введение в алгоритмы. Сбалансированные деревья поиска. Библиотеки C++.
	Практическое занятие (6 ч)	Решить практические задачи на C++, используя циклы, условия, функции и классы.
	Самостоятельная работа (2 ч)	Изучить материал: Сопоставление хэш-таблицы и map в C++ Тест 1.7. Синтаксис языка программирования C++
Модуль 2. Системы обработки и анализа больших данных.		
Тема 2.1 Обработка данных с помощью (Hadoop/Hive/Spark)	Лекция (2 ч)	Основы Hadoop. Архитектура Apache Hadoop. Apache Spark. Использование Hadoop. Создание масштабируемых решения для обработки данных с помощью

		Apache Hive и Apache Spark. Работа с распределенной кластерной системой.
	Практическое занятие (6 ч)	Настройка кластера Apache Spark и Hive на Hadoop
	Самостоятельная работа (2 ч)	Изучить материал: Бакетирование vs партиционирование в Apache Hive и Spark. Тест 2.1. Обработка данных с помощью (Hadoop/Hive/Spark)
Тема 2.2 Конвейер данных (Storm, Kafka)	Лекция (2 ч)	Системы обмена сообщениями Storm и Kafka. Создание конвейеры данных с помощью Storm и Kafka. Шины данных Kafka.
	Практическое занятие (4 ч)	Установка и настройка Storm и Kafka.
	Самостоятельная работа (2 ч)	Изучить материал: Архитектура обработки данных в реальном времени – Apache Kafka. Тест 2.2. Конвейер данных (Storm, Kafka)
Тема 2.3 Базовое представление о Map Reduce	Лекция (1 ч)	Назначение и область применения MapReduce. Как устроена MapReduce-принципы работы. Достоинства и недостатки MapReduce
	Практическое занятие (2 ч)	Установка, настройка, запуск MapReduce. Произвести подсчет слов в тексте, с помощью MapReduce.
	Самостоятельная работа (1,5 ч)	Изучить материал: Как используется модель MapReduce компаниями? Тест 2.3. Базовое представление о Map Reduce
Тема 2.4 Распределенная файловая система HDFS	Лекция (2 ч)	Архитектура HDFS. Структура файлов. Алгоритмы чтения и записи в HDFS. Обзор API для работы с HDFS. Object Storage, сходство и отличия от HDFS.
	Самостоятельная работа (2 ч)	Изучить материал: Распределённые файловые системы. Их составляющие. Их

		достоинства, недостатки и сфера применения Тест 2.4. Распределенная файловая система HDFS
Тема 2.5 Статистические методы анализа данных	Лекция (2 ч)	Этапы анализа данных: Описание данных, изучение сходств и различий, исследование зависимостей, снижение размерности, классификация и прогноз. Статистические методы исследования: Описательная статистика, статистические критерии: Крамера-Уэлча, Вилкоксона-Манна-Уитни, хи-квадрат, Фишера и др. Корреляционный анализ, дисперсионный анализ, регрессионный анализ. Факторный анализ, метод главных компонент. Дискриминантный анализ, кластерный анализ, группировка.
	Самостоятельная работа (1 ч)	Изучить материал: Средства анализа данных на персональных компьютерах. Тест 2.5. Статистические методы анализа данных
Тема 2.6 Технологии анализа данных	Лекция (2 ч)	Стандарты жизненного цикла Big Data: CRISP-DM. Когнитивный анализ данных. Data Mining. Методы Data Mining. Предобработка данных (Data Preprocessing). Data quality качество данных, подходы и инструменты Weka, RapidMiner, Knime, Orange IBM SPSS Modeler, Tableau и др.
	Практическое занятие (4 ч)	Провести анализ данных используя методологию CRISP-DM.
	Самостоятельная работа (2 ч)	Изучить материал: Методы анализа на графах. Случайные графы, без масштабные графы, социальные сети – сети тесного

		мира. Закономерности, методы кластеризации на графах. Тест 2.6. Технологии анализа данных
Тема 2.7 Использование платформ данных	Лекция (2 ч)	Платформы данных (облачные и внутрикорпоративные). Цифровая платформа анализа данных.
	Самостоятельная работа (2 ч)	Изучить материал: On premises /Cloud solutions. Облака в сравнении с on-premises инфраструктурой: возможности, преимущества, особенности. Экосистема Hadoop и элементы Системы Обработки Данных. Аналоги из экосистем GCP, AWS. Тест 2.7. Использование платформ данных
Тема 2.8 Инструменты с открытым исходным кодом для анализа больших данных	Лекция (2 ч)	Интеллектуальный анализа данных с помощью языка программирования R. Манипуляция, анализ и моделирование данных с помощью RapidMiner. Инструменты анализа данных и ML Rapid Miner.
	Практическое занятие (4 ч)	Ознакомление с синтаксисом языка R для анализа данных. Обработка данных с помощью библиотеки tidyverse. Статистический анализ данных в R.
	Самостоятельная работа (1,5 ч)	Изучить материал: Применение Rapid Miner для анализа временных рядов и анализа выделенных параметров. Тест 2.8. Инструменты с открытым исходным кодом для анализа больших данных
Тема 2.9 Что такое ETL: как справиться с анализом BigData	Лекция (2 ч)	Основы ETL. ETL процессы и инструменты. Извлечение и очистка данных. Аудит исходных данных. Анализ извлечения

		данных. Проблемы, возникающие в процессе добычи.
	Практическое занятие (4 ч)	Используя процессы ETL необходимо провести: извлечение данных из источника, очистку данных, обогащение и трансформирование. Провести интеграцию в единую целевую модель. Рассмотреть частный случай импорта данных из внешнего файла. Например, из Excel или csv.
	Самостоятельная работа (2 ч)	Изучить материал: Введение в MDX. Тест 2.9. Основы ETL
Промежуточная аттестация	2 ч.	Выполнение практического задания
Модуль 3. Хранение данных на основе БД. Моделирование данных. Хранилища данных. Data Lake		
Тема 3.1 Введение в моделирование данных	Лекция (2 ч)	Основные цели моделирования данных. Типы моделей данных. Методы моделирования данных. Определение сильных и слабых сторон различных типов баз данных и методов хранения данных. Уровни представления данных (ODS, DDL, семантический слой, модель данных).
	Практическое занятие (4 ч)	Создать Логическую и физическую модель базы данных продуктового магазина.
	Самостоятельная работа (2 ч)	Изучить материал: Совместное использование базы данных. Безопасность данных. Тест 3.1. Основные цели моделирования данных.
Тема 3.2 Реляционные базы данных	Лекция (2 ч)	Реляционные базы данных, для хранения структурированных данных. Сравнение OLAP и OLTP систем. Создание таблиц

		нормализованных данных. Внедрение денормализованных схем (например, STAR, Snowflake). Использование реляционных баз данных SQL: GreenPlum, Postgres, Oracle. Использование SQL в Big Data.
	Практическое занятие (4 ч)	Установить базу данных PostgreSQL и PGAdmin. Создать базу данных продуктового магазина. Провести ее к 3NF, при необходимости провести ее денормализацию.
	Самостоятельная работа (2 ч)	Изучить материал: Функционал PGAdmin для PostgreSQL. Тест 3.2. Реляционные базы данных
Тема 3.3 Базы данных NoSQL	Лекция (2 ч)	Хранилища данных NoSQL, назначение и особенности. Отличия от реляционных баз данных. NoSQL в BigData. Классы NoSQL-СУБД с точки зрения CAP-теоремы и их значимость для Big Data. Использование базы данных NoSQL: Cassandra, MongoDB, Elasticsearch, Neo4J, Hbase.
	Практическое занятие (6 ч)	1.Используя СУБД MongoDB, разработайте базу данных, предназначенную для хранения логов веб-сервера. Лог включает в себя следующие поля: адрес ресурса (URL), IP-адрес пользовательского компьютера, отметка времени начала просмотра ресурса, длительность просмотра ресурса. 2. Разработайте консольную утилиту для преобразования лога веб-сервера в формате CSV (Comma Separated Values), в формат JSON. Лог должен содержать поля со следующими названиями: URL, IP, timeStamp, timeSpent.

		3. Разработайте запросы для загрузки полученных данных в формате JSON в СУБД MongoDB.
	Самостоятельная работа (2 ч)	Изучить материал: Документно-ориентированная моделью данных MongoDB. Тест 3.3. Базы данных NoSQL
Тема 3.4 Массово параллельная обработка и анализ данных	Лекция (2 ч)	Массово параллельные базы данных в BigData. MPP-СУБД для хранения и аналитики Big Data на примере GreenPlum
	Самостоятельная работа (2 ч)	Изучить материал: Особенности организации СУБД в MPP-системе. Тест 3.4. Массово параллельная обработка и анализ данных
Тема 3.5 Облачные хранилища данных (Data Warehouses)	Лекция (2 ч)	Введение в хранилища данных. Архитектура хранилища данных. ETL процессы и инструменты. Использование процесса ETL для денормализации базы данных (от 3NF до Star). Создание куба OLAP из фактов и измерений.
	Практическое занятие (6 ч)	Моделирование данных в Хранилище. Произвести развертывание окружения: СУБД + dbt. Конфигурация проекта dbt, установка модуля dbtVault. Выполнить проектирование детального слоя на базе подхода Data Vault, подготовка метаданных для кодогенерации. Автоматизовать наполнения детального слоя данных с помощью dbt + dbtVault Сформировать витрины данных на основе детального слоя.
	Самостоятельная работа (2 ч)	Изучить принципы моделирования данных в Data Warehouse: нормализация, обогащение,

		интеграция, единая логическая модель, витрины данных и расчет метрик. Тест 3.5. Основы облачных хранилищ данных (Data Warehouses)
Тема 3.6 Data Lake (озеро данных)	Лекция (2 ч)	Назначения Data Lake (озеро данных). Внедрение озер данных. Использование Spark для запуска процессов ELT и аналитики данных с различных источников, структур и данных. Поточковая обработка данных (data streaming, event processing). Основные компоненты и проблемы.
	Самостоятельная работа (2 ч)	Изучить материал: Нужно ли нам озеро данных? А что делать с хранилищем данных? Тест 3.6. Data lake (озеро данных)
Модуль 4. Машинное обучение		
Тема 4.1 Введение в машинное обучение и обработку данных.	Лекция (2 ч)	Введение в машинное обучение и обработку данных. Методы машинного обучения. Постановка основных классов задач в машинном обучении. Обучение с учителем; регрессия и классификация; обучение без учителя; кластеризация, снижение размерности; рекомендательные системы; обработка текстов; обработка изображений; обучение с подкреплением. Примеры задач.
	Практическое занятие (4 ч)	Познакомится с библиотекой машинного обучения Scikit-Learn. Изучить разделы способы работы. Произвести импорт библиотеки в Python. Гиперпараметры и проверка качества модели
	Самостоятельная работа (2 ч)	Изучить материал: Программный инструментарий Data Mining и Machine Learning.

		Тест 4.1. Введение в машинное обучение и обработку данных.
Тема 4.2 Основной аппарат комбинаторики и мат. статистики. Регрессионный анализ и сжатие данных.	Лекция (2 ч)	Основной аппарат комбинаторики и мат. статистики. Основные понятия математической статистики: статистические оценки, их свойства, проверка гипотез. Регрессионный анализ и сжатие данных. Задача регрессии. Минимизация квадрата отклонения. Регрессионная функция: условное мат.ожидание. Линейная регрессия и метод k ближайших соседей. Переобучение и недообучение. Разложение ошибки на шум, смещение и разброс.
	Практическое занятие (4 ч)	Построение модели линейной регрессии для задачи предсказания автомобильного трафика. Отбор и кодирование признаков. Визуальное сравнение общего и предсказанного моделью трафика. Проверка качества. Построить модель линейной регрессии с помощью библиотеки Scikit-Learn для заданного набора данных. Провести анализ качества построенной модели.
	Самостоятельная работа (2 ч)	Изучить материал: Семь методов регрессионного анализа, которые должны освоить инженер данных. Тест 4.2. Основной аппарат комбинаторики и мат. статистики. Регрессионный анализ и сжатие данных.
Тема 4.3 Технологии кластеризации и классификации. Деревья	Лекция (2 ч)	Технологии кластеризации и классификации. K-means. ЕМ-алгоритм. Другие методы кластеризации. Задачи

решений. Нейронные сети. Генетические алгоритмы.		классификации. Байесовский классификатор. Линейные методы для классификации. Логистическая регрессия, максимизация правдоподобия. Нейронные сети: общая архитектура. Многослойные сети. Обратное распространение ошибки. Стохастический градиентный спуск. Генетические алгоритмы.
	Практическое занятие (6 ч)	Построение модели логистической регрессии для задачи предсказания оттока клиентов мобильного оператора. Отбор и кодирование признаков. Проверка качества модели с помощью перекрёстной проверки. Построить модель логистической регрессии с помощью библиотеки Scikit-Learn. Провести анализ качества построенной модели.
	Самостоятельная работа (2 ч)	Изучить тему: Применение алгоритмов кластеризации данных. Тест 4.3. Основы технологии кластеризации и классификации.
Тема 4.4 Деревья решений	Лекция (2 ч)	Правила и анализ качества (точность, полнота). Анализ с помощью ROC кривой. Алгоритм построения деревьев решений. Критерий информационного выигрыша и критерий Джини. Леса решающих деревьев.
	Практическое занятие (4 ч)	Реализовать алгоритм построения дерева с критерием информационного выигрыша и критерием Джини и определению класса по мажоритарному классу в листе. Найти оптимальную глубину дерева в обоих случаях (в отрезке 2-10).

	Самостоятельная работа (2 ч)	Изучить материал: Деревья решений в Python с Scikit-Learn. Тест 4.4. Деревья решений
Тема 4.5 Методы оптимизации машинного обучения	Лекция (2 ч)	Постановка задач оптимизации. Классификация задач оптимизации. Классификация методов оптимизации Методы одномерной оптимизации. Методы многомерной оптимизации.
	Практическое занятие (6 ч)	1. Найти аналитическое решение задачи $\min_{x \in [a,b]} f(x)$ с методами одномерного поиска. Сравнить различные алгоритмы по эффективности на тестовых примерах. 2. Найти аналитическое решение оптимизационной задачи с методами поиска минимума функции двух переменных в оптимизационных задачах без ограничений.
	Самостоятельная работа (1,5 ч)	Изучить материал: Методы оптимизации с использованием глобальных верхних оценок, зависящих от параметра. Тест 4.5. Методы оптимизации машинного обучения
Тема 4.6 GPU в задачах машинного обучения	Лекция (1 ч)	Массово параллельные вычисления для ускорения машинного обучения с помощью GPU. Архитектура GPU, ограничения. Вычисления на GPU. Особенности архитектуры GPU. Решения на базе вычислителей Nvidia. Решения на базе вычислителей AMD. Изделия CPU и GPU в одном чипе. Гибридные вычислительные системы.
	Практическое занятие (2 ч)	Познакомится с RAPIDS — открытыми библиотеками NVIDIA для Python. Изучить

		разделы способы работы. Произвести импорт библиотеки в Python.
	Самостоятельная работа (1,5 ч)	Изучить материал: Преимущества графических процессоров для глубокого обучения. Тест 4.6. GPU в задачах машинного обучения
Тема 4.7 Машинное обучение на больших данных. Онлайн обучение и линейные модели	Лекция (2 ч)	Онлайн подход к обучению на больших данных на примере линейных моделей. Разбор принципов работы vowpal wabbit. Progressive validation, трюк с хэшированием. Запуск обучения на кластере.
	Практическое занятие (4 ч)	Использование Vowpal Wabbit в задаче классификации тегов вопросов на сайте. Обучить модели на выборке в несколько гигабайт.
	Самостоятельная работа (2 ч)	Изучить материал: Apache Spark и оптимизация гиперпараметров. Тест 4.7. Онлайн обучение и линейные модели
Тема 4.8 Рекомендательные системы	Лекция (2 ч)	Особенности построения рекомендательных систем на больших данных. Content-based, collaborative filtering, ALS, iALS.
	Практическое занятие (6 ч)	Создать систему рекомендаций по продуктам наподобие Amazon. Система рекомендаций может предложить вам продукты, фильмы и т. д. на основе ваших интересов и вещей, которые вам нравятся и которые вы использовали ранее. 1. Загрузить датасет, обучить модель Content-based фильтрации, проверить работу модели, написать выводы. Датасет: IMDB-Dataset.zip

		2. Загрузить датасет, обучить модель collaborative фильтрации, проверить работу модели, написать выводы. Датасет: IMDB-Dataset.zip
	Самостоятельная работа (2 ч)	Изучить материал: Рекомендательные системы и системы поддержки принятия Решений. Тест 4.8. Рекомендательные системы
Тема 4.9 Нейронные сети	Лекция (2 ч)	Архитектура нейронных сетей. Распараллеливание SGD: async sgd, delayed sgd. Parameter Server. Эффективное использование нескольких GPU, использование нескольких машин с GPU.
	Практическое занятие (4 ч)	Используйте метод стохастического градиентного спуска SGD, чтобы нарисовать график количества итераций (каждое K раз, здесь K вы задаете сами) и ошибки, соответствующей обучающей выборке и тестовой выборке. Сравните экспериментальные графики и обобщите свои выводы.
	Самостоятельная работа (2 ч)	Изучить материал: Как начать работу с Keras (Deep Learning и Python). Тест 4.9. Основные понятия нейронных сетей
Тема 4.10 LSH. Кластеризация больших данных	Лекция (2 ч)	Нахождение похожих объектов на примере меры Жаккара. Распределенный вариант алгоритма K-Means.
	Практическое занятие (4 ч)	Использовать кластеризацию k-средних для построения модели обнаружения мошеннических действий. Загрузить датасет выполнить кластеризацию используя алгоритмы K-Means. Датасет: enron_mail.zip

	Самостоятельная работа (2 ч)	Изучить материал: Параллельные алгоритмы для обработки BigData: подводные камни и непростые решения. Тест 4.10. LSH. Кластеризация больших данных
Модуль 5. Искусственный интеллект		
Тема 5.1 Введение в искусственный интеллект	Лекция (2 ч)	Определения, история развития и главные тренды ИИ. Формирование концепции искусственного интеллекта, основные направления исследований, этапы развития и проблемы. Рынок технологий искусственного интеллекта. Рынок ИИ в России. Мировой рынок ИИ. Процесс, стадии и методологии разработки решений на основе ИИ
	Самостоятельная работа (2 ч)	Изучить материал: Национальная стратегия развития искусственного интеллекта в России. Тест 5.1. Основы искусственного интеллекта
Тема 5.2 Роль искусственного интеллекта в бизнесе	Лекция (2 ч)	Архитектура и принципы работы промышленных решений, созданных на основе ИИ. Области применения искусственного интеллекта.
	Практическое занятие (4 ч)	1.Провести классификацию знаний. Изучить заданную предметную область и построить модель знаний в виде графа. Список:  Список предметных област
	Самостоятельная работа (2 ч)	Изучить материал: Практика внедрения искусственного интеллекта в бизнес российских компаний.

		Тест 5.2. Роль искусственного интеллекта в бизнесе
<i>Промежуточная аттестация</i>	2 ч.	<i>Выполнение практического задания</i>
<i>Практика на базе привлекаемой организации</i>	10 ч	<i>Решение практических кейсов</i>
<i>Итоговая аттестация</i>	2 ч	<i>Решение практических кейсов</i>
Итого	260 ч	

Формы текущего контроля – тесты, выполнение практических кейсов - представлены в разделе «Формы аттестации и оценочные материалы».

Учебно-методическое и информационное обеспечение программы представлено в разделе «Учебно-методическое обеспечение и информационное сопровождение».

6 ФОРМЫ АТТЕСТАЦИИ И ОЦЕНОЧНЫЕ МАТЕРИАЛЫ

6.1 Входная диагностика

Входная диагностика состоит из решения тестового задания из 45 вопросов. Слушатель выбирает утверждения, которые он считает верными. Выбранные утверждения демонстрируют степень его владения компетенциями в области искусственного интеллекта и методов машинного обучения.

Каждому утверждению присваивается балльная оценка (от 0 до 1 балла). Оценивание производится на основании суммы баллов.

При отсутствии ответов оценка 0 (ноль) баллов присваивается по умолчанию.

Шкала оценивания:

От 0 до 29 баллов – входная диагностика считается не пройденной;

От 30 до 45 баллов – входная диагностика считается пройденной.

Таким образом, слушатель считается успешно прошедшим входное тестирование при наборе не менее 65% правильных ответов.

Тестовое задание.

**Правильный ответ выделен полужирным.*

1. Как удалить файл DataEngineer.txt из папки /tmp ?
 - a. rm /tmp
 - b. rm -f /tmp
 - c. **rm /tmp/DataEngineer.txt**
 - d. tmp / rm -f
2. Как создать файлы file1, file2, file3 в папке /tmp
 - a. **touch /tmp/file1 /tmp/file2 /tmp/file3**
 - b. file /tmp/file1 && /tmp/file2 && /tmp/file3
 - c. **cd /tmp && touch file1 file2 file3**
 - d. cd /tmp/file1 && /tmp/file2 && /tmp/file3
3. Какая команда позволит отсортировать файлы по дате изменения?
 - a. sort -all
 - b. **ls -lt**
 - c. ls -s

- d. `ls -la`
4. Что делает команда `git add -u`
- a. добавляет в индекс все обновленные файлы
 - b. добавляет в индекс все удаленные файлы
 - c. добавляет в индекс все новые и обновленные файлы
 - d. добавляет в индекс все обновленные и удаленные файлы**
5. Как удалить stash по имени?
- e. `git stash clear <name>`
 - f. `git stash pop <name>`
 - g. `git stash drop <name>`**
 - h. `git stash delete <name>`
6. Предположим, вам нужно увидеть список всех файлов, которые были изменены или добавлены в конкретный коммит. Как сделать это?
- i. найти коммит в удаленном репозитории, так как это единственное место, где хранится такая информация
 - j. использовать команду `diff-tree` с хэшем коммита**
 - k. запустить `git commit --info` с хэшем коммита
 - l. использую команду `git stash` с хэшем коммита
7. Какая команда позволяет добавить файл для отслеживания
- m. `git push <file>`
 - n. `git add <file>`**
 - o. `git put <file>`
 - p. `git apply <file>`
8. Укажите все запросы, которые эквивалентны следующему: `select * from numbers where textvalue = 'one'`
- q. `select * from numbers where textvalue like '%one%'`
 - r. `select * from numbers where textvalue like '%one'`
 - s. `select * from numbers where textvalue like 'one'`**
 - t. `select * from numbers where textvalue like one`
9. Для какого из ключевых слов можно добавить строку "TABLE employees", чтобы запрос выполнялся для таблицы employees?
- u. DROP**
 - v. UPDATE
 - w. INSERT INTO
 - x. SELECT
10. Что верно про запрос `SELECT firstName, lastName FROM Employee WHERE lastName BETWEEN 'A%' AND 'D%'`,

- у. будут отображаться все сотрудники, имеющие фамилии начинающиеся с 'A' до 'D' по алфавиту, включая A и исключая D
 - z. выдаст ошибку, так как BETWEEN может использоваться только для чисел, а не для строк
 - aa. будут отображаться все сотрудники с фамилиями, начиная с «A» и заканчивая «D»
 - bb. будут отображаться все сотрудники, имеющие фамилии начинающиеся с 'A' до 'D' по алфавиту, исключая A и D
11. Как выбрать все записи из таблицы "Customers", где значение поля "City" начинается с "V"?
- cc. `SELECT * FROM Customers WHERE City='V%'`
 - dd. `SELECT * FROM Customers WHERE City LIKE 'V%'`**
 - ee. `SELECT * FROM Customers WHERE City LIKE '%V'`
 - ff. `SELECT * FROM Customers WHERE City='%V%'`
12. Как создать инициализатор класса A в python?
- gg. A(параметры инициализатора)
 - hh. `def __init__(параметры инициализатора)`**
 - ii. `def __A__(параметры инициализатора)`
 - jj. `def init(параметры инициализатора)`
13. Какой принцип ООП описывает следующее предложение - Этот принцип является способностью использовать общий интерфейс для нескольких форм (типов данных)?
- kk. Инкапсуляция
 - ll. Абстракция
 - mm. Наследование
 - nn. Полиморфизм**
14. Что будет выведено на экран:
- ```
L = [a * 4 for a in 'abc']
print(L)
```
- oo. ['aaaa']
  - pp. ['4a']
  - qq. ['aaaa', 'bbbb', 'cccc']**
  - rr. ['abc', 'abc', 'abc', 'abc']
15. Какой метод добавляет элемент в конец списка?
- ss. `insert()`
  - tt. `append()`**
  - uu. `pop()`

vv. extend()

16. Что такое кортеж?

ww. неупорядоченная неизменяемая коллекция объектов произвольных типов

xx. упорядоченная изменяемая коллекция объектов произвольных типов

**yy. упорядоченная неизменяемая коллекция объектов произвольных типов**

zz. неупорядоченная неизменяемая коллекция объектов одного типа

17. Как будет выглядеть кортеж d:

a = (1, 2, 3)

b = [4, 5, 6]

c = 2

d = ( a + tuple(b) ) \* c

print(d)

aaa. TypeError: can only concatenate tuple (not "list") to tuple

bbb. (1, 2, 3, 1, 2, 3)

ccc. (1, 2, 3, 1, 2, 3, 1, 2, 3, 1, 2, 3)

**ddd. (1, 2, 3, 4, 5, 6, 1, 2, 3, 4, 5, 6)**

18. Что выведет этот код

a = float("nan")

dict = {5: 'very good', a: 'it is empty'}

print(len(dict))

eee. 0

fff. 1

**ggg. 2**

hhh. Ошибку

19. Что выведет данный скрипт:

dict = {'Lermontov': 'Poets death', 'Pushkin': 'Eugene Onegin'}

dict['Lermontov'], dict['Pushkin'], = dict['Pushkin'], dict['Lermontov']

print(dict)

**iii. {'Lermontov': 'Eugene Onegin', 'Pushkin': 'Poets death'}**

jjj. {'Lermontov': 'Poets death', 'Pushkin': 'Eugene Onegin'}

kkk. {'Lermontov': 'Poets death'}

lll. Ошибка

20. Выберите верные утверждения про словари:

**mmm. словари изменяемые**

nnn. поиск ключа в словаре происходит за линейное время

**ooo. поиск ключа в словари происходит за константное время**

ppp. словари неизменяемые

21. К чему приведет обращение к непустому списку по индексу -1  
qqq. будет ошибка IndexError  
**rrr. вернется последний элемент**  
sss. вернется первый элемент  
ttt. ошибка KeyError
22. В каком случае правильно создана анонимная функция?  
uuu. `lambda x+1 : x`  
vvv. `x + 1 = lambda x`  
**www. `lambda x : x + 1`**  
xxx. `def lambda():`
23. Что такое функция?  
**ууу. объект, принимающий аргументы и возвращающий значение**  
zzz. структура, определяющая поведение объекта  
aaaa. любой код в Python  
bbbb. код, заключенный в круглые скобки
24. Чем отличаются методы `remove()` и `discard()`, применяемые к множеству  
cccc. ничем  
**dddd. `remove()` удаляет элемент если он есть, но бросает ошибку если элемента нет, `discard()` просто удаляет элемент если он есть**  
eeee. `discard()` удаляет элемент если он есть, но бросает ошибку если элемента нет, `remove()` просто удаляет элемент если он есть  
ffff. метода `discard()` для множеств не существует
25. В каких вариантах правильно объявляется множество?  
**gggg. `a = {}`**  
hhhh. `a = []`  
**iiii. `a = set()`**  
jjjj. `a = ()`
26. Какой метод используется для добавления элемента в множество  
**kkkk. `add()`**  
llll. `append()`  
mmmm. `new()`  
nnnn. `get()`
27. Apache NiFi используется для  
**oooo. маршрутизации потоков Big Data и построения ETL-конвейеров**  
pppp. эффективного хранения больших данных  
qqqq. визуализации результатов аналитики

- rrrr. оптимизации SQL-запросов к DWH
28. Для машинного обучения подходят данные
- ssss. Бинарные
  - tttt. Числовые типа int
  - uuuu. Любых форматов в цифровом виде
  - vvvv. Предварительно подготовленные, очищенные от ошибок, пропусков и выбросов, а также нормализованные и представленные в виде числовых векторов**
29. Для распределенного глубокого машинного обучения (Deep Learning) больше подходит фреймворк?
- www. TensorFlow
  - xxxx. Scikit-learn
  - yyyy. PyTorch**
  - zzzz. Flask
30. Автоматизировать запуск пакетных задач в рамках конвейера обработки больших данных по расписанию можно с помощью
- aaaa. Apache Kafka
  - bbbb. Apache Hive
  - cccc. Apache Hadoop
  - dddd. Apache AirFlow**
31. Формат Parquet считается
- eeee. полуструктурированным
  - ffff. неструктурированным
  - gggg. строковым
  - hhhh. колоночным (столбцовым)**
32. Повысить производительность Apache Kafka можно с помощью:
- iiii. Увеличения плотности разделов на каждом брокере
  - jjjj. Повышения коэффициента репликации
  - kkkk. Замены HDD-дисков на SSD**
  - llll. Увеличения размера сообщений
33. Для полнотекстового интеллектуального поиска и аналитики по полуструктурированным данным в формате JSON отлично подходит СУБД
- mmmm. Hbase
  - nnnn. Hive
  - oooo. Elasticsearch**
  - pppp. Cassandra



34. Для реализации микросервисной архитектуры и интеграции разрозненных систем подходит
- qqqqq. Apache Hadoop
  - rrrrr. Apache AirFlow
  - sssss. Apache Spark
  - ttttt. Apache Kafka**
35. Выберите технологию потоковой обработки событий в режиме реального времени
- uuuuu. MapReduce
  - vvvvv. Spark Streaming
  - wwwww. Apache Hadoop
  - xxxxx. Apache Kafka**
36. Какие из следующих технологий СУБД не используют принцип MapReduce?
- yyyyy. Hadoop
  - zzzzz. Cassandra
  - aaaaaa. Redis**
  - bbbbbb. HDInsight
37. Отметьте те из вариантов, в которых данные структурированы:
- ssssss. данные о продажах компании, представленные в виде помесечных отчётов в формате MS Word
  - dddddd. библиотека фильмов, представленных в формате mpeg4 на одном жестком диске
  - eeeeee. таблица с ежедневными показаниями температуры помещения за год в файле формата csv**
  - ffffff. текст педагогической поэмы Макаренко, представленный в формате PDF
38. В чём состоит свойство расширяемости записей СУБД
- gggggg. СУБД не имеет чёткой структуры, поэтому любую запись можно расширить
  - hhhhhh. повышение отказоустойчивости системы при добавлении новых записей в СУБД
  - iiiiii. в любую таблицу СУБД можно добавить новую колонку, предварительно изменив структуру этой таблицы
  - jjjjjj. СУБД имеет чёткую, но расширяемую структуру, в каждую запись можно добавить новую колонку, также, как и узнать значение любой записи по добавленной колонке**
39. Отметьте причины создания NoSQL баз данных
- kkkkkk. дороговизна лицензий RDBMS

**III. высокая стоимость горизонтальной масштабируемости RDBMS при  
сохранении требования высокой доступности**

mmmmmm. недостаточная гибкость языка запросов SQL

nnnnnn. невозможность хранить большие объёмы данных

40. Начиная с каких размеров данных обоснованно применение кластера Hadoop для хранения данных?

oooooo. **100Тб**

pppppp. **1Пб**

qqqqqq. 100Гб

rrrrrr. 1Тб

41. Вы хотите предсказать суммы, которые клиенты потратят на оплату трафика в разные месяцы, исходя из истории их предыдущего потребления. Это задача:

ssssss. **Регрессии**

tttttt. Классификации

uuuuuu. Классификации и регрессии

42. Какие из этих задач типичны для машинного обучения с учителем?

vvvvvv. Группировка сообщений от пользователей

wwwww. **Оценка тона комментария: положительный или отрицательный**

xxxxxx. Группировка изображений по визуальным признакам на размеченных данных

yyyyyy. **Оценка вероятности, кликнет ли человек на рекламный баннер**

43. Выберите все задачи, которые характерны для обучения без учителя?

zzzzzz. Прогноз стоимости недвижимости

aaaaaa. Предсказание пола автора комментария

bbbbbb. **Рекомендация друзей, контента и пабликов в социальных сетях**

ssssss. **Сегментация пользователей интернет-магазина по неявным интересам**

44. В базе данных есть следующие записи: длительность звонков, общее число звонков, общее число переданных сообщений, количество потраченных гигабайтов трафика, Вы хотите предсказывать объем трафика, который потратят клиенты. Что будет объектом модели в этой задаче?

dddddd. Длительность звонков

eeeeee. Общее число звонков

ffffff. **Клиент**

gggggg. Количество трафика

45. Какой алгоритм не подходит для решения задачи, объекты в которой нужно разделить на классы?

hhhhhhh. Случайный лес

iiiiiii. **Линейная регрессия**

jjjjjjj. Дерево принятия решений

kkkkkkk. Логистическая регрессия

## 6.2 Текущий контроль

Текущий контроль предусмотрен в ходе изучения каждой темы.

**Формами** текущего контроля являются тесты и выполнение практических кейсов. Тесты содержат не менее 10 вопросов с одним или несколькими правильными ответами. За каждый правильный ответ ставится 1 балл. Критерием прохождения теста является получение не менее 75% правильных ответов.

Тематика тестов:

Тест 1.1. Функции Инженера данных. Классификация данных.

Тест 1.2. Технологии DataScience и BigData

Тест 1.3. Основы синтаксиса и структур в Python

Тест 1.4. Использование библиотек NumPy, SciPy в Python.

Тест 1.5. Методы библиотеки Pandas

Тест 1.6. Способы отрисовки данных и их связей с помощью библиотек Matplotlib, Seaborn, Plotly, Bokeh, Altair, Folium.

Тест 1.7. Синтаксис языка программирования C++

Тест 2.1. Обработка данных с помощью (Hadoop/Hive/Spark)

Тест 2.2. Конвейер данных (Storm, Kafka)

Тест 2.3. Базовое представление о Map Reduce

Тест 2.4. Распределенная файловая система HDFS

Тест 2.5. Статистические методы анализа данных

Тест 2.6. Технологии анализа данных

Тест 2.7. Использование платформ данных

Тест 2.8. Инструменты с открытым исходным кодом для анализа больших данных

Тест 2.9. Основы ETL

Тест 3.1. Основные цели моделирования данных.

Тест 3.2. Реляционные базы данных

Тест 3.3. Базы данных NoSQL

- Тест 3.4. Массово параллельная обработка и анализ данных
- Тест 3.5. Основы облачных хранилищ данных (Data Warehouses)
- Тест 3.6. Data lake (озеро данных)
- Тест 4.1. Введение в машинное обучение и обработку данных.
- Тест 4.2. Основной аппарат комбинаторики и мат. статистики. Регрессионный анализ и сжатие данных.
- Тест 4.3. Основы технологии кластеризации и классификации.
- Тест 4.4. Деревья решений
- Тест 4.5. Методы оптимизации машинного обучения
- Тест 4.6. GPU в задачах машинного обучения
- Тест 4.7. Онлайн обучение и линейные модели
- Тест 4.8. Рекомендательные системы
- Тест 4.9. Основные понятия нейронных сетей
- Тест 4.10. LSH. Кластеризация больших данных
- Тест 5.1. Основы искусственного интеллекта
- Тест 5.2. Роль искусственного интеллекта в бизнесе

**Выполнение практических кейсов** оценивается по следующей методике.

Критерии и шкала оценивания.

0-4 балла: имеются содержательные и логические ошибки, решение кейса не найдено

5-6 баллов: решение кейса в целом найдено, но оно неоптимально и/или имеются логические ошибки

7-8 баллов: решение кейса найдено, но имеются неточности в решении

9-10 баллов: решение кейса найдено, ошибки отсутствуют

Максимально возможное число баллов за работу – 10.

не менее 9 баллов – «отлично»

7-8 баллов – «хорошо»

5-6 баллов – «удовлетворительно»

0-4 балла – «неудовлетворительно»

## 6.3 Промежуточная аттестация

Промежуточная аттестация предусмотрена после 2 и 5 модулей. Она заключается в решении практико-ориентированных задач (кейсов).

### Модуль 2

#### Задание 1

Развернуть простое облако на платформе GCP с одной виртуальной машиной и одним хранилищем SSD. Сделать скриншоты рабочего instance.

#### Задание 2

Собрать статистику продаж в филиалах супермаркета и их зависимость от площади магазина и ежедневного количества покупателей используя Apache Spark.

В качестве исходных данных используется датасет

<https://www.kaggle.com/datasets/surajjha101/stores-area-and-sales-data>

Выполнить обработку и анализ данных.

1. Произвести импорт библиотек и загрузить данные из датасета.
2. Проверить данные на корректность, наличие дубликатов. Очистить.
3. Провести кластерный анализ данных.
4. Выполнить предварительную обработку данных в ML.
5. Написать рекомендации.

Необходимо загрузить на платформу, где проходит обучение, ссылку на репозиторий с кодом.

Критерии и шкала оценивания.

0-4 балла: имеются содержательные и логические ошибки, решение кейса не найдено

5-6 баллов: решение кейса в целом найдено, но оно неоптимально и/или имеются логические ошибки

7-8 баллов: решение кейса найдено, но имеются неточности в решении

9-10 баллов: решение кейса найдено, ошибки отсутствуют

Максимально возможное число баллов за работу – 10.

не менее 9 баллов – «отлично»

7-8 баллов – «хорошо»

5-6 баллов – «удовлетворительно»

0-4 балла – «неудовлетворительно»

## **Модуль 5**

### **Задание 1**

Выгрузка данных из внешних систем. Загрузка сырых данных.

Необходимо настроить data pipeline в Apache NiFi, который будет:

1. Получать данные, отправляемые по HTTP (по сути, работать как API);
2. Анализировать содержимое того, что ему прислали;
3. Выполнять преобразования (склейку);
4. Сохранять результат в отдельную директорию в соответствии с контентом.
- 5.

В результате должны получить:

1. Работающий pipeline NiFi в контейнерном окружении;
2. Примеры curl-команд, приводящих к разным результатам;
3. Текстовые данные в соответствующих директориях контейнерной файловой системы.

### **Задание 2**

Спроектируете схему данных и постройте аналитическую витрину

СУБД: использовать Vertica (Docker) либо PostgreSQL (Docker) либо BigQuery (GCP)

Датасет: Захват данных из Divolte, либо любой из GCP Public Datasets.

Definition of Done:

- DDL объектов
- DML шагов преобразований
- Опционально: Тестирование на наличие ошибок в данных

Необходимо загрузить на платформу, где проходит обучение, ссылку на репозиторий с кодом.

Критерии и шкала оценивания.

0-4 балла: имеются содержательные и логические ошибки, решение кейса не найдено

5-6 баллов: решение кейса в целом найдено, но оно неоптимально и/или имеются логические ошибки

7-8 баллов: решение кейса найдено, но имеются неточности в решении

9-10 баллов: решение кейса найдено, ошибки отсутствуют

Максимально возможное число баллов за работу – 10.

не менее 9 баллов – «отлично»

7-8 баллов – «хорошо»

5-6 баллов – «удовлетворительно»

0-4 балла – «неудовлетворительно»

## 6.4 Итоговая аттестация

Итоговая аттестация проходит в форме решения практико-ориентированных задач (кейсов).

### Описание практико-ориентированных задач (кейсов).

Примеры кейсов:



## ПОСТРОЕНИЕ ЦИФРОВОЙ МОДЕЛИ ФАБРИКИ

### ЗАДАЧА:

Моделирование бизнес и производственных процессов действующего производства, поиск 'узких мест' в производстве для их устранения, поиск параметров влияющих на выход годной продукции

### РЕШЕНИЕ:

- Построение датацентричной системы сбора инженерной и производственной информации;
- Создание и визуализация графа зависимостей между собой процессов, оборудования;
- Построение предсказательной модели выхода годной продукции. На основе данных выпуска прошлых партий, где мы знаем входные параметры (т.е. параметры сырья, компетенции операторов, технологий, параметры технологического оборудования и т.д.) и процент выхода годной продукции по завершению производственного процесса – производится выбор и настройка параметров классификатора с использованием алгоритмов машинного обучения.

### РЕЗУЛЬТАТ:

- Моделирование процессов фабрики, снижение брака;
- Создана система поддержки принятия решений;
- Повышение управляемости сложного технологического производства

### ИСПОЛЬЗУЕМЫЕ ТЕХНОЛОГИИ

Анализ временных рядов, XGBoost, CatBoost, нейронные сети

### ТЕХНИЧЕСКИЙ СТЕК:

TensorFlow, PyTorch, Cuda



[ПЕРЕХОД НА ПЛАТФОРМУ](#)

## КЛАСТЕРИЗАЦИЯ КЛИЕНТСКОЙ БАЗЫ. РЕКОМЕНДАТЕЛЬНАЯ СИСТЕМА БАНКОВСКИХ ПРОДУКТОВ

### ПРОБЛЕМА

Неэффективность проводимых рекламных кампаний. Приведенные клиенты не возвращаются за дополнительными услугами, либо возвращаются, но через очень длительный период времени.

### РЕЗУЛЬТАТ

Разработана система по рекомендации услуг банка, на вход которой подается информация о транзакциях клиента, его доходе, социально-демографических характеристиках и др. Рекомендации выдаются, основываясь на поведенческих характеристиках одного человека или группы людей (группы определяются в результате работы алгоритма по сегментации клиентов).

### ЗАДАЧА

Разработать алгоритм для сегментации клиентской базы. Построить рекомендательную систему по услугам банка.



### ВЫГОДА

1. Повышение доходности банковских продуктов.
2. Снижение расходов на маркетинговые кампании и повышение их эффективности.
3. Увеличение продаж за счет cross-sell и up-sell.

## АТМ-СЕРВИС

### ИСПОЛЬЗУЕМЫЕ ТЕХНОЛОГИИ

Random Forest (бинарный классификатор) и счетчик устойчивости прогноза (количество дней, когда модель оценивает риск наступления поломки как высокий)



### ЗАДАЧА:

Прогнозирование поломок банкоматов (изменить ситуацию: реакция на поломку АТМ осуществляется по факту возникновения сбоя, превентивная реакция отсутствует)

### РЕШЕНИЕ:

Результаты проверки гипотезы о вероятности возникновения поломки узла CashHandler от количества транзакций, количества ошибок датчиков и статистики сбойности АТМ позволяют сделать следующие выводы: Качество работы модели существенно выше для поломок с низкой частотой повторяемости (от 4 недель). Т.е. для низкочастотных событий гипотеза о влиянии фактора нормы выработки узла в целом подтверждается. Для поломок с высокой частотой повторяемости (5 дней и чаще) – определяющими являются иные факторы, не отраженные в данных (т.н. случайные факторы).

### РЕЗУЛЬТАТ:

Метрики качества модели существенно улучшаются при увеличении горизонта прогнозирования. Установление оценки вероятности возникновения фатальной ошибки





## ПРОГНОЗИРОВАНИЕ ПРОСРОЧКИ КЛИЕНТОВ-КОМПАНИЙ ПО КРЕДИТАМ

### ПРОБЛЕМА

На текущий момент в большинстве банков кредитные риски (невозможность клиентом выполнения обязательств по кредиту) компаний анализируются ежемесячно вручную, причем выборочно для некоторой части компаний из всего списка.

### РЕЗУЛЬТАТ

Раннее обнаружение компаний (клиентов), наиболее близких к риску выхода в просрочку по кредиту (еще до самого факта наступления просрочки).

### ЗАДАЧА

Разработка модели по прогнозированию вероятности просрочки клиентами банка выплат по кредитам, основываясь на данных о компаниях (информация о компании, её транзакциях, заключенных договорах и др.)



### ВЫГОДА

Прогноз модели направлен на помощь специалистам по мониторингу для более эффективного нахождения компаний с повышенным кредитным риском и быстрого оповещения собственников и менеджмента компаний о возможном риске с целью его минимизации.

## AI РЕКОМЕНДАТЕЛЬНЫЕ СИСТЕМЫ ДЛЯ ДИАГНОСТИКИ ЗАБОЛЕВАНИЙ (ЗЛОКАЧЕСТВЕННЫЕ ОБРАЗОВАНИЯ)

### ЗАДАЧА:

- Разработать, обучить и объединить в общий ансамбль алгоритмов по обработке медицинских данных (включая DICOM-изображения, DICOM-тэги, метаданные пациента (в т.ч. формализованные протоколы))
- Достижение точности классификации не менее 95% на класс (для задачи классификации снимков/исследований)
- Достижение точности выделения сегмента патологии на снимке выше чем 90% (для задачи сегментации)

### ИИ МОДУЛЬ ПО ДИАГНОСТИКЕ ЗАБОЛЕВАНИЙ

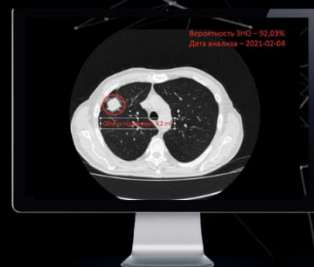
Модель на базе ИНС по обработке метаданных (На основании анализа DICOM-тэгов, На основании анализа метаданных пациента (формализованных протоколов))  
Модель на базе ИНС по обработке DICOM-изображений (Сегментаций патологий на снимке мед исследования, Диагностика (классификация снимка))

### РЕЗУЛЬТАТ:

Вероятность ЗНО по всему снимку - 92,03%  
Найден 1 очаг с объемом: - 7,2 ml  
Дата анализа - 2021-02-04

### ТЕХНИЧЕСКИЙ СТЕК:

Frameworks: Tensorflow, PyTorch  
Segmentation NN: U-Net, W-Net, Mask R-CNN, MeshNet, CNN+CRF  
Language: Python 3



Данные кейсы предоставлены Ассоциацией компаний по развитию искусственного интеллекта.

AI & IT Lab Z-union (z-union.ru) - команда инженеров в области машинного обучения и глубокого обучения, основатели лаборатории ИИ (МФТИ, Сколтех), основатели Ассоциации Лабораторий по развитию ИИ (alrui.ru). Команда занимается прикладным внедрением технологий ИИ в различные сферы. Основные направления кейсов:

Fintech (банковские кейсы);

RetailTech;

ProdTech (производство);

MedTech;

LegalTech;

R&D based on AI & Blockchain technologies.

Основные направления разработок команды:

- разработка прогностических моделей, рекомендательных моделей;
- разработка моделей по работе с изображениями (распознавание, классификация, сегментация, детектирование событий, генерация изображений – много работали с генеративно-состязательными нейронными сетями, face swap в том числе); <https://deepfake.z-union.ru> , <https://zmed.z-union.ru>
- разработка моделей по работе с текстами (распознавание текста и OCR, выделение именованных сущностей); <http://legaltech.z-union.ru>
- НИРы и НИОКРы на стыке разных технологий. Основные направления – AI + Blockchain, Reinforcement learning; <https://gbc.ai>

В ходе итоговой аттестации обучающиеся должны продемонстрировать следующие знания, умения и навыки:

**Знать:**

- основные принципы, процесс, стадии и методологии разработки решений на основе ИИ;
- основные платформы данных (облачные и внутрикорпоративные), применяемые при разработке решений на основе ИИ;
- основные технологии Data Science и BigData, используемые для решения практических задач
- архитектуру и принципы работы специализированных решений, созданных на основе ИИ;
- основные виды представления данных: табличные, графовые, временные ряды;
- основы массово-параллельной обработки и анализ данных;
- основы методов оптимизации;
- принципы работы NoSQL баз данных;
- принципы работы реляционных баз данных;
- язык запросов к СУБД - SQL
- области применений массово-параллельных вычислений для ускорения машинного обучения;
- технологии распределенной обработки данных;
- основы работы с распределенными кластерными системами;

- синтаксис и семантику основных языков искусственного интеллекта, и основные приемы программирования на них;

- основные инструменты анализа данных и ML (RapidMiner);

- основные уровни представления данных;

- основы потоковой обработки данных;

- основные ETL процессы и инструменты.

**Уметь:**

- применять системы обработки и анализа больших массивов данных;

- анализировать требования по сбору данных

- применять SQL базы данных для прикладных решений;

- применять методы оценки качества данных;

- применять инструменты, технологии и библиотеки Data Science при разработке решений на основе ИИ;

- применять язык программирования Python и библиотеки при разработке решений на основе ИИ;

- использовать основные статистические методы анализа данных при решении профессиональных задач;

- реализовывать методы машинного обучения;

- осуществлять поиск и структурирование данных;

- визуализировать анализируемые данные;

- применять методы машинного обучения для работы с большими данными;

- использовать шины данных (kafka) при обработке больших данных в рамках решения профессиональных задач;

- использовать цифровые платформы анализа данных для решения профессиональных задач.

**Владеть:**

- языками запросов для обращения к СУБД;

- языками программирования для реализации методов подготовки данных;

- основными принципами, возможностями и порядком применения метода ближайших соседей (kNN) для решения задач классификации данных;

- методами машинного обучения и нейронных сетей;

- математическими методами анализа данных;

- культурой постановки и планирования последовательности решения задач анализа данных и классификации;
- навыками анализа реальных задач из различных предметных областей на уровне отдельных подходов и коллективами алгоритмов.
- методами случайных лесов (Random forest) для решения задач классификации данных;
- библиотеками, программными платформами (фреймворки) и программными комплексами машинного обучения.

### **Критерии оценивания, шкала оценивания.**

0-4 балла: имеются содержательные и логические ошибки, решение кейса не найдено

5-6 баллов: решение кейса в целом найдено, но оно неоптимально и/или имеются логические ошибки

7-8 баллов: решение кейса найдено, но имеются неточности в решении

9-10 баллов: решение кейса найдено, ошибки отсутствуют

Максимально возможное число баллов – 10.

не менее 9 баллов – «отлично»

7-8 баллов – «хорошо»

5-6 баллов – «удовлетворительно»

0-4 балла – «неудовлетворительно»

## 7 ОРГАНИЗАЦИОННО-ПЕДАГОГИЧЕСКИЕ УСЛОВИЯ РЕАЛИЗАЦИИ ПРОГРАММЫ

### 7.1 Специфика организационных действий и педагогических условий

Для достижения планируемых результатов обучение строится с использованием следующих:

**методов:** case-study, метод проектов, модульное обучение, проблемное обучение, контекстное обучение;

**форм:** лекции с использованием мультимедиа, практические занятия, самостоятельная работа, практика на базе привлекаемой организации.

Кроме того, обучение строится с применением **технологий** электронного обучения и дистанционных образовательных технологий.

Для реализации программы в состав **организационной группы** входят:

1. Директор образовательной программы.
2. Методист образовательной программы.
3. education data engineer (EDE) аккаунт — лицо, отвечающее за адаптацию образовательных программ для работы с данными, необходимыми для выявления потенциала и формирования компетентностного профиля обучающихся, поиска ответов на вопросы, связанных со спецификой обучения. Функции EDE-аккаунта включают разработку педагогического дизайна образовательной программы, диагностику участников образовательной программы, сбор и анализ Цифрового следа, полученного в ходе деятельностных форматов обучения.
4. tech аккаунт — лицо, отвечающее за работу с сервисами Платформы Университета 2035, в том числе обеспечивает сборку образовательного пространства, формирование диагностики, конструирование расписания, выгрузку данных и др.

### 7.2 Кадровое обеспечение программы (преподавательский состав)

К реализации программы привлечены представители образовательных организаций высшего образования, научных организаций и представители компаний со стажем работы в области искусственного интеллекта и в смежных областях.

Представители образовательной организации высшего образования и научной организации имеют высшее образование, ученую степень кандидата наук, стаж научно-педагогической работы более трех лет, а также не выполняют функции иностранного агента.

Представители компаний со стажем работы в области искусственного интеллекта и в смежных областях имеют опыт решения практических задач с использованием технологий искусственного интеллекта более 3 лет в течение последних 10 лет в профильной компании или в профильном подразделении, а также не выполняют функции иностранного агента.

Члены преподавательского состава имеет за последние 3 года научные публикации, соответствующие направлению данной программы, в журналах, включенных в перечень ВАК и в журналах Web of Science (WOS).

| №<br>п/п | Фамилия, имя,<br>отчество<br>(при наличии) | Место основной работы и должность,<br>ученая степень и ученое звание<br>(при наличии)                                                                                 | Ссылки на веб-страницы<br>с портфолио<br>(при наличии)                                  |
|----------|--------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------|
| 1        | Рабин Алексей<br>Владимирович              | Санкт-Петербургский государственный университет аэрокосмического приборостроения (ГУАП),<br>Директор центра координации научных исследований, доктор технических наук |                                                                                         |
| 2        | Васильев<br>Роберт<br>Андреевич            | Компания «Z-Union», CEO,<br>управляющий партнер,<br>ООО «Реал-В», генеральный директор                                                                                | <a href="http://course.alrii.ru/#rec430127293">http://course.alrii.ru/#rec430127293</a> |
| 3        | Душкин Роман<br>Викторович                 | «Агентство по развитию искусственного интеллекта», директор по науке и технологиям                                                                                    | <a href="http://course.alrii.ru/#rec430127293">http://course.alrii.ru/#rec430127293</a> |
| 4        | Постный Игнат<br>Алексеевич                | TAG Consulting,<br>Исполнительный директор                                                                                                                            | <a href="http://course.alrii.ru/#rec430127293">http://course.alrii.ru/#rec430127293</a> |
| 5        | Юров Юрий<br>Аркадьевич                    | Z-union,<br>ML/DL разработчик                                                                                                                                         | <a href="http://course.alrii.ru/#rec430127293">http://course.alrii.ru/#rec430127293</a> |

|   |                                  |                             |  |
|---|----------------------------------|-----------------------------|--|
| 6 | Шатский<br>Владимир<br>Сергеевич | Разработчик, IBS Infinisoft |  |
|---|----------------------------------|-----------------------------|--|

### 7.3 Учебно-методическое обеспечение и информационное сопровождение

#### Учебно-методические материалы

Методические разработки, материалы курса, учебная литература

Обучающие материалы представлены в виде видеолекций, текстовых и графических материалов, размещенных на платформе sprint.IT.

Все слушатели получают авторизованный доступ к системе дистанционного обучения, расположенной в сети Интернет, к учебно-методическим материалам курса

#### *Список учебной и учебно-методической литературы*

##### **Иностранные оригинальные издания**

1. Kuhn, M., & Johnson, K. (2019). Feature Engineering and Selection: A Practical Approach for Predictive Models (1st ed.). Chapman and Hall/CRC. <https://doi.org/10.1201/9781315108230>.
2. Фаулер, Мартин, Садаладж, Прамодкумар Дж. «NoSQL: новая методология разработки нереляционных баз данных» Вильямс, 2013 год. – 192 с.
3. Sundaram, R. K. A first course in optimization theory. – Cambridge University Press, 2011 – 351 с. ISBN 978-0-521-49770-1.

##### **Иностранные переводные издания**

1. Томас Дэвенпорт. Внедрение искусственного интеллекта в бизнес-практику. Преимущества и сложности. – Альпина Паблишер, 2021. – 316 с. ISBN 978-5-9614-3952-6.
2. Машинное обучение с использованием Python. Сборник рецептов: Пер. с англ. Front Cover. Элбон Крис. БХВ-Петербург, 2020. – 384 с. ISBN 978-5-9775-4056-8.

3. Прадик Джоши. Искусственный интеллект с примерами на Python. – Диалектика, 2019. – 448 с. ISBN 978-5-907114-41-8.

4. Джон Д. Келлехер, Брайан Мак-Нейми, Аоифе д'Арси. Основы машинного обучения для аналитического прогнозирования: алгоритмы, рабочие примеры и тематические исследования. – Диалектика-Вильямс, 2019 – 656 с. ISBN: 978-5-6040044-9-4.

5. Дэви Силен, Арно Мейсман, Мохамед Али. Основы Data Science и Big Data. Python и наука о данных. – СПб.: Питер, 2017 – 336 с.: ил – (Серия «Библиотека программиста»). ISBN 978-5-496-02517-1.

### **Российские издания**

1. Бурков А. Машинное обучение без лишних слов. Изд-во: Питер, 2020 – 192 с. ISBN 978-5-4461-1560-0.

2. Кухарев Г. А. Методы обработки и распознавания изображений лиц в задачах биометрии / Г. А. Кухарев, Е. И. Каменская, Ю. Н. Матвеев, Н. Л. Щеголева. Монография СПб: Политехника, 2013. – 388 с.

3. Барсегян, А. А. Анализ данных и процессов: учеб. пособие / А. А. Барсегян, М. С. Куприянов, И. И. Холод, М. Д. Тесс, С. И. Елизаров. – 3-е изд., перераб. и доп. – СПб.: БХВ-Петербург, 2009. – 512 с.: ил. + CD-ROM — (Учебная литература для вузов). ISBN 978-5-9775-0368-6.

4. Лесин В. В. Основы методов оптимизации: учебное пособие для вузов / Лесин Виктор Васильевич, Лисовец Юрий Павлович. - 3-е изд., испр. – Санкт-Петербург : Лань, 2011. – 352с. : ил.

5. Аттетков, А.В. Методы оптимизации: учебное пособие / А. В. Аттетков, В. С. Зарубин, А. Н. Канатников. – Москва: РИОР: ИНФРА-М, 2013. – 269 с.

6. Косников, С.Н. Математические методы в экономике : учеб. пособие для вузов / С. Н. Косников. – 2-е изд., испр. и доп. — М. : Издательство Юрайт, 2019. – 172 с. – (Серия : Университеты России). – ISBN 978-5-534-04098-2.

7. Методы оптимизации: практикум / Б. В. Соболев, Б. Ч. Месхи, Г. И. Каныгин. - Ростов-на-Дону : Феникс, 2009. – 377 с. ISBN. 978-5-222-16161-6.

### **Учебно-методические и информационные материалы**

1. Большие данные в социальных и гуманитарных науках: Сб. обзоров и рефератов / РАН. ИНИОН. Центр науч.-информ. исслед. по науке, образованию и технологиям; отв.



ред. – Гребенщикова Е.Г. – М., 2019. – 193 с. – (Сер.: Наука, образование и технологии). ISBN 978-5-248-00912-1.

2. Цифровая экономика от теории к практике: как российский бизнес использует искусственный интеллект / исслед. РАЭК / НИУ ВШЭ при поддержке Microsoft. – 2019. – 66 с. - URL: <http://raec.ru/upload/files/190715-ii.pdf>.

3. Радченко И.А, Николаев И.Н. Технологии и инфраструктура Big Data. – СПб: Университет ИТМО, 2018. – 52 с.

4. Шаг в будущее: искусственный интеллект и цифровая экономика : материалы 1-й Меж-дунар. науч.-практ. конф., 4-5 дек. 2017, г. Москва : в 4 вып. - М.: Изд. дом ГУУ, 2017. - Вып. 1. – 289 с.

5. Вьюгин В. В. Математические основы теории машинного обучения и прогнозирования. – М. : МЦНМО, 2013. – 387 с.

6. Каклаускас А., Завадскас Э.-К. Биометрическая и интеллектуальная поддержка решений: монография. Вильнюс: Техника, 2012. – С. 344.

7. Неделько В. М. Основы статистических методов машинного обучения : учеб. пособие. – Новосибирск : Новосиб. гос. техн. ун-т, 2010. – 79 с.

8. Методы одномерной оптимизации: метод. указания / сост. И.А. Власова – Самара, 2015. –90 с.: ил.

9. Попова Т.М. Методы безусловной оптимизации : Тексты лекций. / Т. М. Попова; [науч. ред. Р. В. Намм]. - Хабаровск: Изд-во Тихоокеан. гос. ун-та, 2013. – 76 с. ISBN 987-5-7389-1245-0.

10. Бирюков А. Г. Методы оптимизации. Условия оптимальности в экстремальных задачах : учеб. пособие для вузов (направление «Прикладные мат. и физика») / Бирюков Александр Гаврилович. – Москва : МФТИ, 2010. – 226 с.

11. Щетинин, Е.Ю. Математические методы оптимизации и оптимальных процессов : учебное пособие / Е.Ю. Щетинин. – Москва : СТАНКИН, 2015.

12. Сухарев, А. Г. Курс методов оптимизации : учебное пособие / А. Г. Сухарев, А. В. Тимохов, В. В. Федоров. – 2-е изд. – Москва : Физматлит, 2011. – 368 с. – ISBN 978-5-9221-0559-0.

### **Статьи**

1. Shurugin, Sergei & Enchinov, Aivengo & Gritskevich, Evgeny. (2021). USING THE METHOD OF STATISTICAL MODELING FOR IDENTIFICATION OF AN OBJECT BY GEOMETRIC PARAMETERS OF ITS IMAGE. Interexpo GEO-Siberia. 6. 314-320. 10.33764/2618-981X-2021-6-314-320.

2. Саханевич Д. Ю., Кремин А. Е. Систематизация методов машинного обучения в целях внедрения искусственного интеллекта в социально-экономические процессы региона // Вестник Владимирского государственного университета имени Александра Григорьевича и Николая Григорьевича Столетовых. Серия: Экономические науки. - 2019. - № 4 (22). - С. 57-65.
3. Щурина С. В., Данилов А. С. Искусственный интеллект как технологическая инновация развития экономики // Экономика. Налоги. Право. – 2019. – № 12 (3). – С. 125-133.
4. Аналитика Больших данных как инструмент бизнес-инноваций. [Электронный ресурс]. – Режим доступа: <https://filearchive.cnews.ru/img/files/2019/05/27/20190424idchitchiwpbdafin.pdf> (дата обращения: 17.07.2022).
5. Баева В.Р., Дроздов А.Ю. ETL: АКТУАЛЬНОСТЬ И ПРИМЕНЕНИЕ. ПРЕИМУЩЕСТВА И НЕДОСТАТКИ ETL ИНСТРУМЕНТОВ // Вестник науки. 2019. №5 (14). URL: <https://cyberleninka.ru/article/n/etl-aktualnost-i-primenenie-preimuschestva-i-nedostatki-etl-instrumentov> (дата обращения: 17.07.2022).
6. Красов Андрей Владимирович, Штеренберг Станислав Игоревич, Фахрутдинов Ринат Маратович, Рыжаков Данила Владиславович, Пестов Игорь Евгеньевич Анализ информационной безопасности предприятия на основе сбора данных пользователей с открытых ресурсов и мониторинга информационных ресурсов с использованием машинного обучения // Т-Comm. 2018. №10. URL: <https://cyberleninka.ru/article/n/analiz-informatsionnoy-bezopasnosti-predpriyatiya-na-osnove-sbora-dannyh-polzovateley-s-otkrytyh-resursov-i-monitoringa> (дата обращения: 17.07.2022).
7. Бутенко Е. К. Искусственный интеллект в банках сегодня: опыт и перспективы // Финансы и кредит. – 2018. – № 24 (3). – С. 143-153.
8. Коротеев М. В. Обзор некоторых современных тенденций в технологии машинного обучения // E-Managment. – 2018. – Т. 1, № 1. – С. 26-35. – DOI: 10.26425/2658-3445-2018-1-26-35.
9. Hilbert, M. (2016). Big Data for Development: A Review of Promises and Challenges. Development Policy Review, 34(1), 135–174. <http://doi.org/10.1111/dpr.12142>
10. Синяк Николай Георгиевич, Каклаускас Артурас, Зинькина Диана Васильевна, Хабиб Авада, Шариф Нуреддин, Бондаренко Анна Владимировна Использование больших данных в оценке экономического потенциала организаций // Труды БГТУ. Серия 5: экономика и управление. 2015. №7. URL: <https://cyberleninka.ru/article/n/ispolzovanie-bolshikh-dannykh-v-otsenke-ekonomicheskogo-potentsiala-organizatsiy>

11. Цветкова Л. А., Черченко О. В. Внедрение технологий Big Data в здравоохранение: оценка технологических и коммерческих перспектив // Экономика науки. 2016. №2. URL: <https://cyberleninka.ru/article/n/vnedrenie-tehnologiy-big-data-v-zdravoohranenie-otsenka-tehnologicheskikh-i-kommercheskikh-perspektiv> (дата обращения: 17.07.2022).

13. Радченко И.А. Особенности использования программно-аппаратного комплекса, предназначенного для управления предприятиями и отраслями экономики, в рамках концепции Linked Enterprise Data. Сборник трудов конференции «Инженерия знаний и технологии семантического веба – 2012». – СПб: НИУ ИТМО, 2012. С. 165-167.

14. Ахмедова М.Р., Гусейнов Р.В. Использование методов оптимизации для анализа и обработки информации // Вестник ДГТУ. Технические науки. 2016. №2. URL: <https://cyberleninka.ru/article/n/ispolzovanie-metodov-optimizatsii-dlya-analiza-i-obrabotki-informatsii> (дата обращения: 18.07.2022).

| Электронные образовательные ресурсы                                                                                                                                                                                                                                                                                                                                                                                                                | Электронные<br>информационные ресурсы                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                   |
|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <p>1. Воронцов К. В. Машинное обучение: курс лекций // MachineLearning.ru. - URL: <a href="http://www.recognition.su/wiki/index.php?title=Машинное_обучение_(курс_лекций%2C_K.В.Воронцов)">http://www.recognition.su/wiki/index.php?title=Машинное_обучение_(курс_лекций%2C_K.В.Воронцов)</a>).</p> <p>2. Проектирование баз данных: Распределенные базы и хранилища данных. Агрегирование // Национальный Открытый Университет «ИНТУИТ». URL:</p> | <p>1. <a href="https://ai.2035.university/">https://ai.2035.university/</a> - сайт УНТИ 2035 «Обучение в области искусственного интеллекта»</p> <p>2. <a href="https://xn--80aapampemcchfmo7a3c9ehj.xn--plai/projects/tsifrovaya-ekonomika">https://xn--80aapampemcchfmo7a3c9ehj.xn--plai/projects/tsifrovaya-ekonomika</a> - сайт Национального проекта «Цифровая экономика»</p> <p>3. <a href="https://xn--80aapampemcchfmo7a3c9ehj.xn--plai/projects/tsifrovaya-ekonomika/p-">https://xn--80aapampemcchfmo7a3c9ehj.xn--plai/projects/tsifrovaya-ekonomika/p-</a></p> |

|                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                  |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                   |
|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <p><a href="http://www.intuit.ru/studies/professional_retraining/953/courses/214/lecture/5508/">http://www.intuit.ru/studies/professional_retraining/953/courses/214/lecture/5508/</a> (Дата обращения: 22.05.2019).</p> <p>3. Бесплатные материалы по Data Engineering от преподавателей МФТИ<br/><a href="https://fpmi-edu.ru/free-de">https://fpmi-edu.ru/free-de</a></p> <p>4. <a href="https://ru.coursera.org/specializations/gcp-data-machine-learning#courses">https://ru.coursera.org/specializations/gcp-data-machine-learning#courses</a></p> <p>5. Big Data Specialization.<br/><a href="https://www.coursera.org/specializations/big-data">https://www.coursera.org/specializations/big-data</a></p> <p>6. <a href="https://stepik.org/course/101687/promo">https://stepik.org/course/101687/promo</a></p> <p>7. Курс «Big Data от Sergey Petrovich»<br/><a href="https://www.youtube.com/playlist?list=PL15mR4o-R9Ng3Fh8Z2HpLdQeJQHSoNKcp">https://www.youtube.com/playlist?list=PL15mR4o-R9Ng3Fh8Z2HpLdQeJQHSoNKcp</a></p> <p>8. <a href="https://mlcourse.ai/book/index.html">https://mlcourse.ai/book/index.html</a></p> <p>9. Дашборды: интерактивная визуализация данных / Алексей Колоколов (Институт бизнес-аналитики)<br/><a href="https://www.youtube.com/watch?v=ve8aQmE-vBY&amp;list=PLCaAlZXDo4bVQRuOpZTs4arnJHzOIrZ7s">https://www.youtube.com/watch?v=ve8aQmE-vBY&amp;list=PLCaAlZXDo4bVQRuOpZTs4arnJHzOIrZ7s</a></p> <p>10. Введение в системы баз данных от Университета Карнеги-Меллона.<br/><a href="https://www.youtube.com/playlist?list=PLSE8ODhjZXjYutVzTeAds8xUt1rcmyT7x">https://www.youtube.com/playlist?list=PLSE8ODhjZXjYutVzTeAds8xUt1rcmyT7x</a></p> <p>11. Продвинутое системы баз данных от Университета Карнеги-Меллона.<br/><a href="https://www.youtube.com/playlist?list=PLSE8ODhjZXja7K1hjZ01UTVDnGQdx5v5U">https://www.youtube.com/playlist?list=PLSE8ODhjZXja7K1hjZ01UTVDnGQdx5v5U</a></p> <p>12. Распределенные компьютерные системы.<br/><a href="https://www.youtube.com/watch?v=w8KFPWk">https://www.youtube.com/watch?v=w8KFPWk</a></p> | <p><a href="#">iskusstvennyy-intellekt-p</a> – сайт федерального проекта «Искусственный интеллект» Национального проекта «Цифровая экономика»</p> <p>4. <a href="https://sprint.1t.ru/">https://sprint.1t.ru/</a> - сайт образовательной платформы ООО «1Т»</p> <p>5. <a href="https://trends.rbc.ru/trends/tag/ai">https://trends.rbc.ru/trends/tag/ai</a> - раздел «Искусственный интеллект» на сайте РБК.</p> <p>6. <a href="https://www.tadviser.ru/index.php/%D0%9F%D1%80%D0%BE%D0%B4%D1%83%D0%BA%D1%82:%D0%98%D1%81%D0%BA%D1%83%D1%81%D1%81%D1%82%D0%B2%D0%B5%D0%BD%D0%BD%D1%8B%D0%B9_%D0%B8%D0%BD%D1%82%D0%B5%D0%BB%D0%BB%D0%B5%D0%BA%D1%82_(%D0%98%D0%98,_Artificial_intelligence,_AI)">https://www.tadviser.ru/index.php/%D0%9F%D1%80%D0%BE%D0%B4%D1%83%D0%BA%D1%82:%D0%98%D1%81%D0%BA%D1%83%D1%81%D1%81%D1%82%D0%B2%D0%B5%D0%BD%D0%BD%D1%8B%D0%B9_%D0%B8%D0%BD%D1%82%D0%B5%D0%BB%D0%BB%D0%B5%D0%BA%D1%82_(%D0%98%D0%98,_Artificial_intelligence,_AI)</a> – раздел «Искусственный интеллект» на сайте «TAdviser». Государство. Бизнес. Технологии.</p> <p>7. <a href="https://rb.ru/opinion/inzhener-dannyh/">https://rb.ru/opinion/inzhener-dannyh/</a> - «Инженер данных: что это за специалист и как им стать?» на сайте - RB.RU</p> <p>8. <a href="https://rb.ru/opinion/data-engineer-modeli-data-scientist/">https://rb.ru/opinion/data-engineer-modeli-data-scientist/</a> - «В России катастрофически не хватает инженеров данных – кому и зачем они нужны» на сайте - RB.RU</p> |
|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|

|                                                                              |  |
|------------------------------------------------------------------------------|--|
| <a href="#">K0bI&amp;list=PLawkBQ15NDEkDJ5IyLIJUTZ1rRM9YQq6N&amp;index=3</a> |  |
|------------------------------------------------------------------------------|--|

#### 7.4 Материально-технические условия реализации программы

| Вид занятий                                                                       | Наименование оборудования, программного обеспечения                                                                                                                                                                                                                                                                                |
|-----------------------------------------------------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Лекционные занятия                                                                | Персональный компьютер с установленным на нем: Windows 10-11, x64/x86; от 8 Gb RAM; от 128 Gb SSD/HDD, монитор от 15”; сетевой интерфейс Fast Ethernet 100Мбит; веб-браузеры Google Chrome, Mozilla Firefox, Opera, Microsoft Edge, Яндекс.Браузер и др.                                                                           |
| Практические занятия, самостоятельная работа, промежуточная и итоговая аттестация | Персональный компьютер с установленным на нем: Windows 10 и выше, x64/x86; от 8 Gb RAM; от 128 Gb SSD/HDD, монитор от 15”; сетевой интерфейс Fast Ethernet 100Мбит; веб-браузеры Google Chrome, Mozilla Firefox, Opera, Microsoft Edge, Яндекс.Браузер и др.<br>Anaconda 2.7 или 3.5<br>Доступ к облачным вычислительным ресурсам. |