МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ
Федеральное государственное автономное образовательное учреждение высшего образования
«Дальневосточный федеральный университет»
(ДВФУ)
ИНСТИТУТ МАТЕМАТИКИ И КОМПЬЮТЕРНЫХ ТЕХНОЛОГИЙ (ШКОЛА)

СОГЛАСОВАНО
Руководитель ОП
_____Р.И. Дремлюга
Подпись (ФИО)

УТВЕРЖДАЮ
И.о директора Академии цифровой трансформации
_____Еременко А.С.
(подпись) (ФИО)
«03» марта 2023 г.

РАБОЧАЯ ПРОГРАММА ДИСЦИПЛИНЫ

Прикладные методы машинного обучения и анализа больших данных
Направление подготовки 09.04.01 Информатика и вычислительная техника
(Искусственный интеллект и большие данные (совместно с ПАО Сбербанк))
Форма подготовки: очная

Рабочая программа дисциплины разработана в соответствии с федеральным государственным образовательным стандартом (ФГОС) высшего образования (ВО) – магистратура по направлению подготовки 09.04.01 Информатика и вычислительная техника, утвержденного приказом Министерства образования и науки Российской Федерации от 19.09.2017 г. № 918 (с изменениями и дополнениями).

Рабочая программа обсуждена на заседании Академии цифровой трансформации от «16» декабря 2022 г., протокол №4

И.о директора Академии цифровой трансформации: к.т.н., профессор Еременко А.С.

Составители:
к.т.н. Еременко А.С., Кленин А.С., ассистент Синягина А.Д.

Владивосток
2023

Оборотная сторона титульного листа РПУД

I. *Рабочая программа рассмотрена и утверждена на заседании*
_____,
протокол от «____»_____202__г. № _____.

II. *Рабочая программа пересмотрена на заседании*
_____
и утверждена на заседании
_____,
протокол от «____»_____202__г. № _____.

III. *Рабочая программа пересмотрена на заседании*
_____
и утверждена на заседании
_____,
протокол от «____»_____202__г. № _____.

IV. *Рабочая программа пересмотрена на заседании*
_____
и утверждена на заседании
_____,
протокол от «____»_____202__г. № _____.

V. *Рабочая программа пересмотрена на заседании*
_____
и утверждена на заседании
_____,
протокол от «____»_____202__г. № _____.

VI. ЦЕЛИ И ЗАДАЧИ ОСВОЕНИЯ ДИСЦИПЛИНЫ

Цель курса – дать понимание внутреннего устройства, механики работы, области применимости существующих решений, осветить сильные и слабые стороны, научить практическим навыкам анализа больших массивов информации. Курс посвящен методам построения систем обработки больших данных и существующим инструментам в этой области.

Структурно курс состоит из четырех разделов: пакетная обработка данных, потоковая обработка данных, хранение данных и особенности анализа данных в разных прикладных сферах (медицине, финансах, государственном и муниципальном управлении и т.п.).

В лекционной части курса рассматриваются такие технологии как HDFS, Hadoop MapReduce, HBase, Cassandra, Spark, Kafka, Spark Streaming, Storm.

Для успешного изучения дисциплины «Прикладные методы машинного обучения и анализа данных» обучающиеся должны обладать базовыми знаниями в следующих теоретических дисциплинах:

- специальные разделы математики, в том числе линейная алгебра, основы статистики, основы дискретной математики, исследование операций и оптимизация;
- технологии и методы программирования, в том числе объектно-ориентированного и начал функционального программирования;
- основы теории автоматов, основы теории вычислений;
- прикладные алгоритмы, а именно алгоритмы на графах и сетях, алгоритмы компьютерной графики, алгоритмы извлечения, обработки и классификации данных.

В результате данной дисциплины у обучающихся формируются следующие профессиональные компетенции:

Код и наименование профессиональной компетенции (результат освоения)	Код и наименование индикатора достижения компетенции
ПК-1 Способен проектировать и разрабатывать системные и прикладные решения по анализу больших данных	ПК-1.1 Владеет инструментарием получения, хранения, передачи и обработки больших данных ПК-1.2 Формулирует и решает системные и прикладные задачи анализа больших данных для конкретных предметных областей ПК-1.3 Способен управлять разработкой продуктов, услуг и решений на основе больших данных
ПК-2 Способен разрабатывать методики выполнения аналитических работ	ПК-2.1 Умеет выявлять проблемы и сложности в существующих практиках выполнения аналитических работ в организации; описывать методики выполнения аналитических работ

Код и наименование профессиональной компетенции (результат освоения)	Код и наименование индикатора достижения компетенции
	ПК-2.2 Владеет навыками выполнения аналитических работ, их апробации и доработки на выбранных проектах
ПК-3 Способен осуществлять планирование, организацию и контроль аналитических работ в IT-проекте	ПК-3.1 Владеет навыками работы с инструментами анализа данных как системного, так и прикладного уровня ПК-3.2 Применяет технологии и методы, используемые в управлении IT-проектами; осуществляет выбор программных и аппаратных средств для аналитических работ ПК-3.3 Управляет процессом аналитических работ, в том числе осуществляет сбор информации, определяет причины отклонений от планов, умеет выявлять и разрешать проблемные ситуации в ходе выполнения аналитических работ
ПК-4 Способен ставить цели и принимать управленческие решения, основанные на анализе больших данных	ПК-4.1 Владеет навыками стратегического управления развитием методологической и технологической инфраструктуры анализа больших данных в организации ПК-4.2 Определяет необходимые ресурсы и инструменты для решения задач с использованием анализа данных; руководит работой команды, вырабатывая командную стратегию на основе анализа данных

Код и наименование индикатора достижения компетенции	Наименование показателя оценивания (результата обучения)
ПК-1.1 Владеет инструментарием получения, хранения, передачи и обработки больших данных	Знает технологии, методы и инструментальные средства обработки больших данных
	Умеет использовать архитектуры и модели баз и хранилищ данных, адаптированные к технологиям больших данных
	Владеет навыками разработки предложений по развитию и совершенствованию системы получения, хранения, передачи, обработки больших данных
ПК-1.2 Формулирует и решает системные и прикладные задачи анализа больших данных для конкретных предметных областей	Знает существующий опыт разработки и использования продуктов и услуг на основе технологий больших данных
	Умеет разрабатывать проектную и рабочую документацию на разработку аналитических услуг на основе технологий больших данных
	Владеет навыками решения прикладных задач анализа больших данных для конкретных предметных областей
ПК-1.3 Способен управлять разработкой продуктов, услуг	Знает существующий опыт разработки и использования продуктов и услуг на основе технологий больших данных

Код и наименование индикатора достижения компетенции	Наименование показателя оценивания (результата обучения)
и решений на основе больших данных	Умеет управлять исполнением проектных работ в области больших данных
	Владеет навыками создания прототипа сервиса на основе аналитики больших данных
ПК-2.1 Умеет выявлять проблемы и сложности в существующих практиках выполнения аналитических работ в организации; описывать методики выполнения аналитических работ	Знает основные методики и практики выполнения аналитических работ
	Умеет выявлять проблемы и сложности в существующих практиках выполнения аналитических работ в организации
	Владеет навыками выполнения аналитических работ, их апробации и доработки на выбранных проектах
ПК-2.2 Владеет навыками выполнения аналитических работ, их апробации и доработки на выбранных проектах	Знает методы, применяемые для функционального и оперативного управления предприятиями
	Планировать проектные работы
	Владеет навыками выполнения аналитических работ, их апробации и доработки на выбранных проектах
ПК-3.1 Владеет навыками работы с инструментами анализа данных как системного, так и прикладного уровня	Знает существующие и перспективные методы и программный инструментарий технологий больших данных
	Умеет разрабатывать программно-аппаратные компоненты и системы на основе технологий и аналитики больших данных
	Владеет методами проведения анализ больших данных в соответствии с утвержденными требованиями к результатам аналитического исследования
ПК-3.2 Применяет технологии и методы, используемые в управлении IT-проектами; осуществляет выбор программных и аппаратных средств для аналитических работ	Знает методы управления IT-проектами
	Умеет описывать задачи и составлять график выполнения работ IT-проекта, исходя из его целей и методов их достижения
	Владеет требуемыми технологиями проектирования
ПК-3.3 Управляет процессом аналитических работ, в том числе осуществляет сбор информации, определяет причины отклонений от планов, умеет выявлять и разрешать проблемные ситуации в ходе выполнения аналитических работ	Знает методы, применяемые для функционального и оперативного управления предприятиями
	Умеет осуществлять сбор информации, определять причины отклонений от планов, выявлять и разрешать проблемные ситуации в ходе выполнения аналитических работ
	Владеет навыками управления проектными рисками в IT-проекте
ПК-4.1 Владеет навыками стратегического управления развитием методологической и технологической инфраструктуры анализа больших данных в организации	Знает существующие и перспективные методы и программный инструментарий технологий больших данных
	Умеет управлять развитием технологической инфраструктуры анализа больших данных
	Владеет навыками стратегического управления

Код и наименование индикатора достижения компетенции	Наименование показателя оценивания (результата обучения)
ПК-4.2 Определяет необходимые ресурсы и инструменты для решения задач с использованием анализа данных; руководит работой команды, вырабатывая командную стратегию на основе анализа данных	Знает методы создания программного обеспечения для анализа и обработки данных
	Умеет использовать методы проектирования систем анализа и обработки данных
	Владеет навыками работы в распределенных командах

II. ТРУДОЁМКОСТЬ ДИСЦИПЛИНЫ И ВИДОВ УЧЕБНЫХ ЗАНЯТИЙ ПО ДИСЦИПЛИНЕ

Общая трудоемкость дисциплины составляет 8 зачётных единиц (288 академических часов).

Видами учебных занятий и работы обучающегося по дисциплине могут являться:

Обозначение	Виды учебных занятий и работы обучающегося
Лек	Лекции
Пр	Практические занятия
СР	Самостоятельная работа обучающегося в период теоретического обучения

III. СТРУКТУРА ДИСЦИПЛИНЫ

Форма обучения – очная.

№	Наименование раздела дисциплины	Семестр	Количество часов по видам учебных занятий и работы обучающегося						Формы текущего контроля успеваемости и промежуточной аттестации
			Лек	Лаб	Пр	ОК	СР	Контроль	
1	Модуль 1. Методы и системы обработки больших данных	3	18		36		131	63	УО-1, ПР-7; ПР-11
1.1	Тема 1	3	2		4				
1.2	Тема 2	3	3		6				
1.3	Тема 3	3	3		6				
1.4	Тема 4	3	3		6				
1.5	Тема 5	3	3		6				
1.6	Тема 6	3	2		4				

№	Наименование раздела дисциплины	Семестр	Количество часов по видам учебных занятий и работы обучающегося						Формы текущего контроля успеваемости и промежуточной аттестации
			Лек	Лаб	Пр	ОК	СР	Контроль	
1.7	Тема 7	3	2		4				
2	Модуль 2. Особенности анализа данных в прикладных сферах	4	20		20				
2.1	Тема 1	4	8		4				
2.2	Тема 2	4	6		4				
2.3	Тема 3	4	6		4				
	Итого:		38		56		131	63	Экзамен

IV. СТРУКТУРА И СОДЕРЖАНИЕ ТЕОРЕТИЧЕСКОЙ ЧАСТИ КУРСА
(38 часов)

МОДУЛЬ 1: МЕТОДЫ И СИСТЕМЫ ОБРАБОТКИ БОЛЬШИХ ДАННЫХ (18 часов)

Тема 1: Вступление, распределенные файловые системы

Понятие «большие данные». Постановка задачи обработки и хранения больших данных. Примеры применения больших данных в IT индустрии. Архитектура распределенных файловых систем. Основные проблемы в работе распределенных систем. Виды отказов узлов, связей между узлами. Устройство GFS, HDFS. Процесс восстановления HDFS.

Тема 2: Модель вычислений MapReduce

Математическая модель парадигмы MapReduce. Модель вычислений Map, Shuffle и Reduce фазы. Формальная модель парадигмы MapReduce. Задача подсчета слов в датасете (WordCount). Hadoop MapReduce. Обеспечение отказоустойчивости в MapReduce. Сравнение MapReduce v1 и YARN. История развития MapReduce. MapReduce Streaming на примере Python Расширения модели. Comparator, partitioner, combiner, зачем нужны и когда используются. Часто применяемые техники в обработке данных о Map-side join, reduce-side join. Salting. Способы тюнинга MapReduce. Способы семплирования данных. Итеративные задачи.

Тема 3. SQL over BigData. Hive.

Hive: мотивация, языковая модель. Проблема смещения данных в обработке больших данных. Применение SQL в IT индустрии. Сравнение решений Hive и MapReduce на примере задач анализа логов. Практика SQL: агрегация данных, фильтрация данных, сортировка, объединение таблиц. Архитектура Hive: Metastore + Hadoop + HDFS. Язык определения данных в Hive (Hive DDL): типы таблиц, разделители. Язык управления данными в Hive (Hive DML): загрузка данных, перезапись данных, CTAS. Hive: расширенные возможности. Парсер данных SerDe. Hive View: особенности, преимущества и недостатки. Пользовательские функции (UDF), пользовательские агрегирующие функции (UDAF), пользовательские функции для генерации таблиц (UDTF). История развития MapReduce. MapReduce Streaming на примере Python. Hive Streaming. Hive Partitioning, Bucketing and Sampling. Особенности Join в Hive. Исправление проблемы смещения в Hive. Поколоночное хранение в Hive (RCFile, ORC, Parquet).

Тема 4. Beyond MapReduce. Spark

Недостатки MapReduce. Costly disk spill, write barrier, job launch overhead. Перекосы в данных и перекосы в планировании. От MR к DAG-ам вычислений: почему это удобней?

Spark. Понятие RDD и Source RDD. MR over Spark, Pregel over Spark. Кеширование RDD, итеративные вычисления о Преобразования и действия. Spark UI и работа в режиме YARN. Spark SQL. Spark DataFrame: особенности и сравнение с Pandas DataFrame о Spark SQL + Hive. Агрегирование данных в Spark DataFrame. Обработка графов при помощи Spark. Задача подсчета количества общих друзей. Задача подсчета числа треугольников. Пакет GraphFrames. Понятие motif. Использование motif для решения задачи. Решение задачи PageRank при помощи GraphFrames и Spark API. Оптимизация Spark. Управление памятью. Оптимизация UDF. Оптимизация объединений.

Тема 5. Машинное обучение на больших данных

Алгоритмы для работы с большими данными. Методы онлайн обучения. Градиентный спуск. Решение задач кластеризации на больших данных о Задача подсчета слов в датасете (WordCount). API для обучения алгоритмов на больших данных о Библиотеки Spark Mllib и Spark ML. Обработка текстов при помощи Spark ML о Ансамблевые модели на Spark ML. Map-side join, reduce-side join

Тема 6. Потоковая обработка данных

Обработка больших данных в режиме реального времени. Подходы к обработке больших данных в режиме реального времени. Семантика доставки (Devilery semantics). Архитектуры Lambda и Kappa. Входные и выходные данные для обработки в режиме реального времени. Apache Spark Streaming:

объяснение концепции на практической задаче. Apache Spark Structured Streaming: объяснение концепции на практической задаче. Модель парадигмы Kafka. Понятие интервала в Kafka. Особенности Kafka. Интерфейс командной строки Kafka. Связь Kafka и семантики доставки о Потоки Kafka (Kafka Streams).

Тема 7. Key-value хранилища в больших данных

HBase. NoSQL подходы к реализации распределенных баз данных, key-value хранилища. Основные компоненты BigTable-подобных систем и их назначение, отличие от реляционных БД. Чтение, запись и хранение данных в HBase. Minor- и major- компактификация. Надёжность и отказоустойчивость в HBase. Cassandra. Основные особенности. Чтение и запись данных. Отказоустойчивость. Примеры применения HBase и Cassandra. Отличие архитектуры HBase от архитектуры Cassandra.

МОДУЛЬ 2: ОСОБЕННОСТИ АНАЛИЗА ДАННЫХ В ПРИКЛАДНЫХ СФЕРАХ (20 часов)

Тема 1. Биоинформатика и анализ данных в медицине и здравоохранении

Предмет, задачи и объекты биоинформатики. Новейшие достижения в области молекулярной биологии и генетики, вызвавшие необходимость развития биоинформатики. Информационные технологии, нашедшие применение в биоинформатике. Системный подход в биоинформатике. Принципы создания математических моделей фармакокинетических, физиологических и других процессов, протекающих в организме человека, для последующего их использования в составе автоматизированных систем поддержки принятия врачебных решений. Виды математических моделей.

Тема 2. Анализ социальных сетей (Social Network Analysis)

Основные понятия в теории сетей. Основные измеряемые свойства сетей. Примеры сетей. История исследования социальных сетей. Методы анализа компьютерных социальных сетей. Степенное распределение. Масштабно-инвариантные сети (scale-free networks). Распределение Парето, нормализация, моменты. Закон Ципфа. Граф ранк-частота. Методы измерений параметров сетей. Параметры сложных сетей. Параметры узлов сети. Общие параметры сети. Распределение степеней узлов. Путь между узлами. Коэффициент кластерности. Посредничество. Эластичность сети. Структура сообщества. Модельные графы. Degree centrality, closeness centrality, betweenness centrality, статус/rank prestige (eigenvector centrality). Центральность сети. Анализ связей. Алгоритм PageRank. Стохастические

матрицы. Теорема Perrona-Frobenius. Степенные итерации. Нахождение собственного вектора. Hubs и Authorities. Модель «слабых связей». Модель Уаттса-Строгатца. Графовые модели. Стохастические блоковые модели. Вероятностные графовые модели. Анализ центральности и других локальных свойств. Понятие сетевых сообществ (network communities). Плотность связей. Метрики. Разделение графа на части (graph partitioning). Разрезы (cuts) в графе. Min-cut, quotent and normalized cuts метрики. Divisive and agglomerative algorithms. Repeated bisection. Корреляционная матрица. Классификация алгоритмов нахождения сообществ. Edge Betweenness. NewmanGirvin

Тема 3 Анализ финансовых потоков и финансовая аналитика

Оценка эффективности работы аналитиков. Понятия «финансовый аналитик», «аналитическое покрытие», «консенсус прогнозы». Виды финансовых аналитиков (buy-side, sell-side). В чем причины ошибок финансовых аналитиков? Виды и оценка эффективности прогнозов. Причины и последствия конфликта интересов в области финансовой аналитики. Источники получения данных о прогнозах аналитических команд. Рейтингование аналитиков- проблемы выявления качественных прогнозов. Типичное построение аналитических отчетов и рекомендаций российских команд аналитиков. Влияние аналитических рекомендаций на поведение цен финансовых активов. Виды эффективности прогнозов аналитиков. Реакция финансовых рынков на информационные события.. Принципы проведения событийного анализа (event study). Классификация информационных событий. Объявления о прибыли, раскрытие отчетности pro forma и рыночные реакции. Объявления об инвестициях (капитальные вложения, ценные бумаги и инновации). Объявления о слияниях и поглощениях. Принципы проведения событийного анализа (event study), тестирование гипотезы о систематической аномальной доходности. Значимость событийного ряда для внутридневной доходности.

V. СТРУКТУРА И СОДЕРЖАНИЕ ПРАКТИЧЕСКОЙ ЧАСТИ КУРСА

Практические занятия (56 часов)

Практическое задание № 1 «Основы работы с аналитической платформой Deductor studio»

1. Проработка теоретических вопросов по теме лабораторной работы.
2. Постановка задач и компьютерное моделирование по вопросам практической части лабораторной работы.
3. Обработка результатов, составление отчета, защита лабораторной работы.

Практическое задание № 2 «Трансформация данных в Deductor Studio»

1. Проработка теоретических вопросов по теме лабораторной работы.
2. Постановка задач и компьютерное моделирование по вопросам практической части лабораторной работы.
3. Обработка результатов, составление отчета, защита лабораторной работы.

Практическое задание № 3 «Создание, заполнение и использование хранилища данных Deductor Warehouse на базе Firebird»

1. Проработка теоретических вопросов по теме лабораторной работы.
2. Постановка задач и компьютерное моделирование по вопросам практической части лабораторной работы.
3. Обработка результатов, составление отчета, защита лабораторной работы.

Практическое задание № 4 «Определение представления источника данных в проекте служб Analysis Services»

1. Проработка теоретических вопросов по теме лабораторной работы.
2. Постановка задач и компьютерное моделирование по вопросам практической части лабораторной работы.
3. Обработка результатов, составление отчета, защита лабораторной работы.

Практическое задание № 5 «Определение и развертывание куба»

1. Проработка теоретических вопросов по теме лабораторной работы.
2. Постановка задач и компьютерное моделирование по вопросам практической части лабораторной работы.
3. Обработка результатов, составление отчета, защита лабораторной работы.

Практическое задание № 6 «Изменение мер, атрибутов и иерархий»

1. Проработка теоретических вопросов по теме лабораторной работы.
2. Постановка задач и компьютерное моделирование по вопросам практической части лабораторной работы.
3. Обработка результатов, составление отчета, защита лабораторной работы.

Практическое задание № 7. «Ассоциативные правила»

1. Проработка теоретических вопросов по теме лабораторной работы.
2. Постановка задач и компьютерное моделирование по вопросам практической части лабораторной работы.
3. Обработка результатов, составление отчета, защита лабораторной работы.

Практическое задание № 8. «Основы работы с пакетом STATISTICA»

1. Проработка теоретических вопросов по теме лабораторной работы.
2. Постановка задач и компьютерное моделирование по вопросам практической части лабораторной работы.
3. Обработка результатов, составление отчета, защита лабораторной работы.

Практическое задание № 9. «Кластерный анализ»

1. Проработка теоретических вопросов по теме лабораторной работы.
2. Постановка задач и компьютерное моделирование по вопросам практической части лабораторной работы.
3. Обработка результатов, составление отчета, защита лабораторной работы.

Практическое задание № 10. «Искусственные нейронные сети»

1. Проработка теоретических вопросов по теме лабораторной работы.
2. Постановка задач и компьютерное моделирование по вопросам практической части лабораторной работы.
3. Обработка результатов, составление отчета, защита лабораторной работы.

VI. УЧЕБНО-МЕТОДИЧЕСКОЕ ОБЕСПЕЧЕНИЕ САМОСТОЯТЕЛЬНОЙ РАБОТЫ ОБУЧАЮЩИХСЯ

План-график выполнения самостоятельной работы по дисциплине, в том числе примерные нормы времени на выполнение по каждому заданию.

План-график выполнения самостоятельной работы по дисциплине в семестр

№ п/п	Дата/сроки выполнения	Вид самостоятельной работы	Примерные нормы времени на выполнение	Форма контроля
1	1-18 неделя обучения	Подготовка лабораторных работ	36	Отчет по практической работе
2	Сессия	Подготовка к экзамену	18	Экзамен
	Всего в семестр		54	

Методические рекомендации к работе с литературными источниками

В процессе подготовки к практическим занятиям, студентам необходимо обратить особое внимание на самостоятельное изучение рекомендованной учебно-методической (а также научной и популярной) литературы. Самостоятельная работа с учебниками, учебными пособиями, научной, справочной и популярной литературой, материалами периодических изданий и Интернета, статистическими данными является наиболее эффективным методом получения знаний, позволяет значительно активизировать процесс овладения информацией, способствует более глубокому усвоению изучаемого материала, формирует у студентов свое отношение к конкретной проблеме. Более глубокому раскрытию вопросов способствует знакомство с дополнительной литературой, рекомендованной преподавателем по каждой теме практического занятия, что позволяет студентам проявить свою индивидуальность в рамках выступления на данных занятиях, выявить широкий спектр мнений по изучаемой проблеме.

Критерии оценки выполнения самостоятельной работы

Контроль самостоятельной работы студентов предусматривает:

- соотнесение содержания контроля с целями обучения;
- объективность контроля;
- валидность контроля (соответствие предъявляемых заданий тому, что предполагается проверить);
- дифференциацию контрольно-измерительных материалов.

Формы контроля самостоятельной работы

1. Просмотр и проверка выполнения самостоятельной работы преподавателем.
2. Самопроверка, взаимопроверка выполненного задания в группе.
3. Обсуждение результатов выполненной работы на занятии.
4. Текущее тестирование.

Критерии оценки результатов самостоятельной работы

Критериями оценок результатов внеаудиторной самостоятельной работы студента являются:

- уровень освоения студентов учебного материала;
- умения студента использовать теоретические знания при выполнении практических задач;
- умения студента активно использовать электронные образовательные ресурсы, находить требующуюся информацию, изучать ее и применять на практике;
- обоснованность и четкость изложения ответа;
- оформление материала в соответствии с требованиями;

- умение ориентироваться в потоке информации, выделять главное;
- умение четко сформулировать проблему, предложив ее решение, критически оценить решение и его последствия;
- умение показать, проанализировать альтернативные возможности, варианты действий;
- умение сформировать свою позицию, оценку и аргументировать ее

Критерии оценки выполнения контрольных заданий для самостоятельной работы

Процент правильных ответов	**Оценка**
От 95% до 100%	отлично
От 76% до 95%	хорошо
От 61% до 75%	удовлетворительно
Менее 61 %	неудовлетворительно

Самостоятельная работа при подготовке к экзамену включает изучение теоретического материала с использованием лекционных материалов, рекомендуемых источников, материалов по практическим занятиям и лабораторным работам.

Контрольные вопросы для самостоятельной оценки качества освоения учебной дисциплины:

1. Определите сущность понятия «большие данные».
2. Опишите методики анализа больших данных.
3. Процесс аналитики анализа больших данных.
4. Дайте характеристику Big Data на мировом рынке.
5. Охарактеризуйте Big Data в России.
6. Определите понятие Data Mining.
7. Вопросы безопасности больших данных.
8. В чем состоит когнитивный анализ данных.
9. Какие модели данных вы знаете?
10. Концепция MapReduce.
11. Основные методы анализа больших данных.
12. Продемонстрировать использование окна Explore для изучения распределения переменной. Выявить, какие отклонения выделяются в распределении переменной? Как можно исправить эту ситуацию?
13. Провести первоначальное исследование данных с использованием узлов Data Partition и
14. Decision Tree. Сделать выводы о модели.

15. Провести прогнозное моделирование с использованием регрессии Regression. Объяснить, какие переменные являются важными в модели. Какое значение имеет статистика среднеквадратичной ошибки, посчитанная на проверочной выборке?
16. Нужно ли делать преобразования входных переменных перед их использованием в модели нейронной сети?
17. Объяснить результаты, полученные на основании Model Comparison. Сделать выводы.
18. Регрессия. Логистическая регрессия. Полиномиальные регрессии.
19. Задания по обработке данных и созданию моделей выполняются с использованием данных из набора SAMPSIO

Темы рефератов

1. Роль отечественных и зарубежных ученых в становлении и развитии современной биоинформатике.
2. Нанобиотехнологии и биоинформатика.
3. Проект «Геном человека» и его роль в становлении современной биоинформатики.
4. Физико-химические и биоинформационные методы исслдеования биополимеров: сравнительные аспекты.
5. Базы данных последовательностей и структур белков
6. Базы данных последовательностей и структур нуклеиновых кислот.
7. Виды баз данных, используемых в биологических исследованиях.
8. Современное значение и перспективы применения биоинформатики в медицине.
9. Прикладное значение биоинформатики: сельское хозяйство.
10. Прикладное значение биоинформатики: пищевая промышленность.
11. Математические методы, используемые в биоинформатике.
12. Вопросы патентования в биоинформатике.
13. Место биоинформатики в избранной научной тематике.
14. Применение биоинформационных технологий в небиологических отраслях.
15. Конструирование модифицированных и новых биологических объектов.
16. Внутриклеточный транспорт токсичных молекул.
17. Особенности создания генно-инженерных конструкций.
18. Генетические маркеры выносливости и работоспособности человека.
19. Приоритеты компьютерного программирования в биоинженерии и биоинформатике.

20. Сигнальные каскады: регуляция экспрессии генов, пролиферации и апоптоза.

21. Мутационный процесс: изменения в последовательности ДНК.

22. Использование методов биоинформатики и молекулярного

23. Геномный браузер: поиск информации о геноме человека. 2

24. Построение выравниваний , реконструкция филогенетических деревьев (сравнение локальных и глобальных выравниваний, зависимость выравнивания от параметров, оценка статистической значимости).

Перечень тем для самостоятельного изучения по дисциплине

Тема 1. Метрики

Примеси Джини (Gini impurity), добавленная информация (information gain). Деревья регрессии. Метрика вариации. Непрерывные признаки. Использование главных компонент вместо признаков. Сокращение дерева (pruning). Метрики, понятие центроида и представителя класса. Центроидные алгоритмы: k-means, k-medoid. Алгоритмы, основанные на плотности: DBSCAN, OPTICS. Алгоритмы, основанные на распределении: сумма гауссиан. Нечеткая кластеризация, алгоритм c-means. Метрики качества: leave-one-out, силуэт, индекс Дэвиса-Болдина (Davies-Bouldin), индекс Данна (Dunn).

Тема 2. Последовательности изображений

Анализ последовательностей изображений в Google EarthEngine. Классификатора на базе многослойного перцептрона для обработки последовательности изображений. Рекуррентные нейронные сети.

Тема 3. Алгоритмы компьютерного зрения

Классические алгоритмы обработки изображений.

Тема 4. Алгоритмы машинного обучения

Сверхточные нейронные сети. Классификация изображений. Сегментация. Локализация.

Вопросы к темам для самостоятельного изучения

1. Задачи обработки текста: извлечение, поиск, классификация (тематическая, эмоциональная), перевод
2. Разбиение на слова, пунктуация, лексический и морфологический анализ
3. Определение частей речи, имён, основ слов
4. Частотный анализ, представление bag-of-words, TF-IDF и его варианты
5. N-грамы, byte-pair encoding.
6. Векторные представления, семантическая интерпретация алгебраических операций

7. Унитарный код (One-hot encoding).
8. Алгоритмы Word2Vec и FastText.
9. Алгоритм GloVe*.
10. Постановка задачи, причины и цели снижения размерности.
11. Выбор и извлечение признаков.
12. Подходы к выбору признаков: filtering, wrapping, embedding.
13. Расстояние между распределениями. Расстояние Кульбака-Лейблера. Взаимная информация.
14. Алгоритмы выбора признаков: на основе корреляции (CFS), взаимной информации, Relief.
15. Метод главных компонент (PCA).
16. Нелинейные обобщения метода главных компонент. Kernel PCA.*
17. Неотрицательное матричное разложение (NMF).*
18. Стохастическое вложение соседей с t-распределением (t-SNE).

VII. КОНТРОЛЬ ДОСТИЖЕНИЯ ЦЕЛЕЙ КУРСА

Для текущей аттестации при изучении дисциплины «Прикладные методы машинного обучения и анализа данных» используются следующие оценочные средства:

1) Устный опрос (УО):

Собеседование (консультация с преподавателем) (УО-1)

2) Письменные работы (ПР):

Конспект (ПР-7)

Практическая работа (ПР-11)

№ п/п	Контролируемые разделы / темы дисциплины	Коды и этапы формирования компетенций		Оценочные средства	
				текущий контроль	промежуточная аттестация
1	Модуль I	ПК-1, ПК-2, ПК-3, ПК-4	знает	УО-1 Собеседование ПР-7 Конспект ПР-11 Практическая работа	Вопросы к экзамену
			умеет		
			владеет		
2	Модуль II	ПК-1, ПК-2, ПК-3, ПК-4	знает	УО-1 Собеседование	Вопросы к экзамену
			умеет	ПР-7 Конспект	
			владеет	ПР-11 Практическая работа	

VIII. СПИСОК УЧЕБНОЙ ЛИТЕРАТУРЫ И ИНФОРМАЦИОННО-МЕТОДИЧЕСКОЕ ОБЕСПЕЧЕНИЕ ДИСЦИПЛИНЫ

Основная литература

(электронные и печатные издания)

1. Анализ данных / В.С. Мхитарян и др. – М: Издательство Юрайт, 2019. – Текст : электронный // ЭБС Юрайт [сайт]. – URL: https://biblio-online.ru/bcode/432178
2. Просто о больших данных : пер. с англ. / Джудит Гурвиц, Алан Ньюджент, Ферн Халпер [и др.]. - Москва : Сбербанк, : [Эксмо], 2015. - 395 с. - http://lib.dvfu.ru:8080/lib/item?id=chamo:826169&theme=FEFU
3. Пальмов С.В. Интеллектуальный анализ данных [Электронный ресурс] : учебное пособие / С.В. Пальмов. – Самара: Поволжский государственный университет телекоммуникаций и информатики, 2017. – 127 с. – 2227-8397. – Режим доступа: http://www.iprbookshop.ru/75376.html
4. Методы и средства комплексного статистического анализа данных: учеб. пособие / А.П. Кулаичев. — 5-е изд., перераб. и доп. — М. : ИНФРА-М, 2018. — 484 с. — (Высшее образование: Бакалавриат). — www.dx.doi.org/10.12737/25093. - Режим доступа: http://znanium.com/catalog/product/975598

Дополнительная литература

(печатные и электронные издания)

1. SPSS 19: профессиональный статистический анализ данных : [практическое руководство] / А. Наследов. - Санкт-Петербург : Питер, 2011. - 399 с.
2. Гулай, Т.А. Теория вероятностей и математическая статистика [Электронный ресурс] : учебное пособие / Т.А. Гулай, А.Ф. Долгополова, Д.Б. Литвин, С.В. Мелешко. - 2-е изд., доп. – Ставрополь: АГРУС, 2013. - 260 с. - Режим доступа: http://znanium.com/catalog.php?bookinfo=514780
3. Петрунин, Ю. Ю. Информационные технологии анализа данных. Data Analysis : учебное пособие для вузов по управленческим и экономическим специальностям и направлениям / Ю. Ю. Петрунин ; Московский государственный университет, Факультет государственного управления. – 3-е изд. – М. : Университет, 2014 – 291 с. – Каталог НБ ДВФУ: http://lib.dvfu.ru:8080/lib/item?id=chamo:734307&theme=FEFU
4. Туманов, В.Е. Проектирование хранилищ данных для систем бизнес-аналитики [Электронный ресурс] : учеб. пособие / Туманов В.Е. – М. : БИНОМ. Лаборатория знаний, Интернет-Университет Информационных Технологий

(ИНТУИТ), 2010. – 615 с. – Режим доступа : http://www.iprbookshop.ru/16096.html

5. Интеллектуальные системы принятия решений и управления : учебное пособие для вузов / Ю. И. Еременко. – Старый Оскол : ТНТ, 2015. – 401 с. – Каталог НБ ДВФУ:.http://lib.dvfu.ru:8080/lib/item?id=chamo:813810&theme=FEFU

6. Интеллектуальный анализ данных и систем управления бизнес-правилами в телекоммуникациях: Монография / Р.Р. Вейнберг. – М.: НИЦ ИНФРА-М, 2016. – 173 с.: – Режим доступа: http://znanium.com/catalog/product/520998

7. Нестеров, С.А. Интеллектуальный анализ данных средствами MS SQL Server [Электронный ресурс] / Нестеров С.А. – М.: Интернет-Университет Информационных Технологий (ИНТУИТ), 2012. – 189 с. – Режим доступа : http://www.iprbookshop.ru/16702.html

8. Чубукова И.А. Data Mining [Электронный ресурс]/ Чубукова И.А. – М.: Интернет-Университет Информационных Технологий (ИНТУИТ), 2016. – 470 с. – Режим доступа: http://www.iprbookshop.ru/56315.html.

9. Чубукова, И.А. Data Mining : учеб. пособие для вузов / И.А. Чубукова / М.Р. Мидлтон ; пер. с англ. [С.Г. Кобелькова]. – М. : Интернет-Университет Информационных Технологий БИНОМ. Лаборатория знаний, 2008. – 282 с. : – Каталог НБ ДВФУ: http://lib.dvfu.ru:8080/lib/item?id=chamo:274659&theme=FEFU

10. Порозов Ю.Б. Биоинформатика [Электронный ресурс] / Ю.Б. Порозов. — Электрон. текстовые данные. — СПб. : Университет ИТМО, 2012. — 54 с. — 2227-8397. — Режим доступа: http://www.iprbookshop.ru/65798.html

11. Микони С. В. Дискретная математика: множества, отношения, функции, графы. — СПб. : Лань, 2012. — 187 с. - Режим доступа: http://e.lanbook.com/books/element.php?pl 1 id=4316.

12. Ржевский, С.В. Исследование операций. — СПб. : Лань, 2013. — 476 с. - Режим доступа: http://e.lanbook.com/books/element.php?pl1 id=32821.

13. Губанов, Д. А. Социальные сети. Модели информационного влияния, управления и противоборства [Электронный ресурс] : учебное пособие / Д. А. Губанов, Д. А. Новиков, А. Г. Чхартишвили ; под ред. Д. А. Новиков. — Электрон. текстовые данные. — М. : Издательство физико-математической литературы, 2010. — 228 с. — 9875-94052-194-5. — Режим доступа: http://www.iprbookshop.ru/8531.html

14. Ничего личного: Как социальные сети, поисковые системы и спецслужбы используют наши персональные данные / Кин Э. - М.:Альпина

Пабл., 2016. - 224 с.: ISBN 978-5-9614-5128-3 - Режим доступа: http://znanium.com/catalog/product/915406

15. Соловьев, В.И. Анализ данных в экономике / В.И. Соловьев. – М: КНОРУС, 2019. Миркин, Б. Г. Введение в анализ данных / Б. Г. Миркин. – М: Издательство Юрайт, 2019. – Текст: электронный // ЭБС Юрайт [сайт]. – URL: https://biblio-online.ru/bcode/432851 (дата обращения: 19.01.2020).

16. Макаров, А.А. Анализ данных на компьютере / Макаров А.А., Тюрин Ю.Н. – М.: МЦНМО, 2016.

17. Шитиков, В.К. Классификация, регрессия, алгоритмы Data Mining с использованием R [Электронный ресурс] / Шитиков В. К., Мастицкий С. Э. – URL: https://github.com/ranalytics/data-mining (дата обращения: 19.01.2020), https://ranalytics.github.io/data-mining/index.html (дата обращения: 19.01.2020).

18. Анализ данных в R. [Электронный ресурс]. – URL: https://stepik.org/course/129/promo (дата обращения: 19.01.2020).

19. MachineLearning: профессиональный информационно-аналитический ресурс, посвященный машинному обучению, распознаванию образов и интеллектуальному анализу данных. [Электронный ресурс]. – URL: http://www.machinelearning.ru (дата обращения: 19.01.2020).

20. Анализ данных на Python [Электронный ресурс]. – URL: https://www.pvsm.ru/python/310645 (дата обращения: 19.01.2020).

21. Анализ данных на Python в примерах и задачах [Электронный ресурс]. – URL: https://compscicenter.ru/courses/data-mining-python/2018-spring/classes/

Перечень ресурсов информационно-телекоммуникационной сети «Интернет»

1. Linux. Карманный справочник. Скотт Граннеман
2. Unix и Linux. Руководство системного администратора. Эви Немет, Гарт Снайдер,Трент Р. Хейн, Бен Уэйли
3. The Official Ubuntu Book Matthew Helmke, Elizabeth K. Joseph, José Antonio Rey, Philip

Ballew, Benjamin Mako Hill

4. Hadoop: The Definitive Guide 3e. Tom White
5. Professional Hadoop Solutions. Boris Lublinsky

Перечень информационных технологий и программного обеспечения

При осуществлении образовательного процесса по дисциплине используется свободно распространяемое программное обеспечение MS Excel, GNU R, Python, а также автоматическая тестирующая система CATS ДВФУ https://imcs.dvfu.ru/cats/.

IX.МЕТОДИЧЕСКИЕ УКАЗАНИЯ ПО ОСВОЕНИЮ ДИСЦИПЛИНЫ

Дисциплина «Прикладные методы машинного обучения и анализа больших данных» является базисом для магистрантов, обучающегося по профилю подготовки «Искусственный интеллект и большие данные».

Процесс изучения дисциплины осуществляется в следующих организационных формах:

- выполнение аудиторных лабораторных работ;
- выполнение аудиторных практических заданий
- самостоятельное изучение материала;
- выполнение контрольных работ;
- подготовка и сдача экзамена.

В дисциплине можно выделить две области:

- базовые знания, относительно стабильные, составляющие ядро дисциплины;
- технологические знания, связанные с освоением конкретных статистических методов, сред и алгоритмов.

Базовые знания основных понятий и принципов теории вероятности и математической статистики, понимание процесса обработки статистических данных, обработки компьютером больших данных образуют понятийное ядро дисциплины и служат основой для изучения многих дисциплин специальности. Эта область включает в себя системный подход к решению задач анализа, прогнозирования и проектирования, математическое мышление, знание терминологии и современных средств обработки данных.

Технологическая часть дисциплины связана с практическим освоением методов статистического анализа, описания средств анализа данных (пакет анализа) и статистических функций, входящих в MS Excel. Отдельное внимание на занятиях уделяется различным способам организации данных в программе, решению стандартных алгоритмических задач. Обучающим инструментом для практического освоения излагаемых методов является универсальный российский статистический пакет STADIA, ставший в данной области своеобразным стандартом де-факто методов статистического анализа.

Лабораторные работы проводится в компьютерных классах и подкреплены методическими указаниями, рекомендациями и требованиями к представлению и оформлению результатов работы.

Самостоятельная работа включает изучение теоретического материала дисциплины и выполнение индивидуальных работ.

Для изучения дисциплины приводится перечень рекомендуемой литературы, методические указания и вопросы к контрольным заданиям и экзамену.

В качестве основы для изучения дисциплины можно взять учебники, учебные пособия, электронные материалы и методические указания, приведенные в списке литературы.

При изучении теоретического материала следует по методическим указаниям ознакомиться с планом темы. Освоив теоретический материал, необходимо самостоятельно, без помощи литературы, сделать попытку ответить на вопросы по теме. С каждой темой связан перечень ключевых понятий. После изучения темы необходимо уметь самостоятельно давать определение понятий.

Работа с теоретическими материалами. Изучение дисциплины следует начинать с проработки тематического плана лекций, уделяя особое внимание структуре и содержанию темы и основных понятий. Изучение «сложных» тем следует начинать с составления логической схемы основных понятий, категорий, связей между ними. Целесообразно прибегнуть к классификации материала, в частности при изучении тем, в которых присутствует большое количество незнакомых понятий, категорий, теорий, концепций, либо насыщенных информацией типологического характера. Студенты должны составлять конспекты лекций, систематически готовиться к практическим занятиям, вести глоссарий и быть готовы ответить на контрольные вопросы в ходе лекций и аудиторных занятий. Успешное освоение программы курса предполагает прочтение ряда оригинальных работ и выполнение практических заданий.

X. МАТЕРИАЛЬНО-ТЕХНИЧЕСКОЕ ОБЕСПЕЧЕНИЕ ДИСЦИПЛИНЫ

Учебные занятия по дисциплине проводятся в помещениях, оснащенных соответствующим оборудованием и программным обеспечением.

Перечень материально-технического и программного обеспечения дисциплины приведен в таблице.

Материально-техническое и программное обеспечение дисциплины

Наименование специальных помещений и помещений для самостоятельной работы	Оснащенность специальных помещений и помещений для самостоятельной работы	Перечень лицензионного программного обеспечения. Реквизиты подтверждающего документа
Мультимедийная аудитория: G467	Проектор DLP, 3000 ANSI Lm, WXGA 1280x800,	Techdesigner, MAX8, VVVV, Adobe Photoshop, Adobe Premier, Adobe

	2000:1 EW330U Mitsubishi,; Моноблок HP ProOne 440 G3 23.8" All-in-One, диагональ экрана 23.8", разрешение экрана 1920x1080, Bluetooth, Wi-Fi, операционная система: Windows 10 Enterprise, оптический привод DVD, процессор: Intel Core i5-7500T, размер оперативной памяти: 8 ГБ, видеопроцессор: Intel HD Graphics 630, объем жесткого диска: 1Tb. Беспроводные ЛВС для обучающихся обеспечены системой на базе точек доступа 802.11a/b/g/n 2x2 MIMO(2SS). AfterEffects	
Мультимедийная аудитория: G469	Проектор DLP, 4000 ANSI Lm, 1920x1080, 2000:1 FD630u Mitsubishi; Проектор DLP, 2800 ANSI Lm, 1920x1080, 2000:1 GT1080 Optoma; Проектор DLP, 3000 ANSI Lm, WXGA 1280x800, 2000:1 EW330U Mitsubishi; Беспроводные ЛВС для обучающихся обеспечены системой на базе точек доступа 802.11a/b/g/n 2x2 MIMO(2SS). Специализированное оборудование: Платформа Adruino UNO, Бесконтактный сенсорный Microsoft Kinnect 2.0, Аудио система Dialog 2.0, MIDI контроллер Playtron, Одноплатный компьютер Raspberry PI	Techdesigner, MAX8, VVVV, Adobe Photoshop, Adobe Premier, Adobe

Рабочие места для людей с ограниченными возможностями здоровья оснащены дисплеями и принтерами Брайля; оборудованы: портативными устройствами для чтения плоскопечатных текстов, сканирующими и читающими машинами, видеоувеличителем с возможностью регуляции цветовых спектров; увеличивающими электронными лупами и ультразвуковыми маркировщиками.

В целях обеспечения специальных условий обучения инвалидов и лиц с ограниченными возможностями здоровья в ДВФУ все здания оборудованы пандусами, лифтами, подъемниками, специализированными местами, оснащенными туалетными комнатами, табличками информационно-навигационной поддержки.