Федеральное государственное бюджетное учреждение науки
Ордена Трудового Красного Знамени
Институт солнечно-земной физики
Сибирского отделения Российской академии наук
(ИСЗФ СО РАН)

ДОКУМЕНТ ПОДПИСАН
ЭЛЕКТРОННОЙ ПОДПИСЬЮ

Сертификат: 419490770E381130FC93D3E8F4B8E013
Владелец: Медведев Андрей Всеволодович
Действителен: с 01.11.2023 до 24.01.2025

УТВЕРЖДАЮ:
Врио директора ИСЗФ СО РАН
чл.– корр. РАН_____А.В. Медведев
«15» марта 2024 г.

Рабочая программа дисциплины

Б1.В.ДВ.1.1 Введение в технологии Больших Данных

Направление подготовки **03.04.02 Физика**

Направленность (профиль): **Физика солнечно-земных связей**

Квалификация выпускника: **МАГИСТР**

Тип профессиональной деятельности**: научно-исследовательский, педагогический**

Форма обучения: **очная**

Иркутск 2024

Рабочая программа составлена на основании Федерального государственного образовательного стандарта высшего образования по направлению подготовки 03.04.02 Физика (уровень магистратуры), утвержденного приказом Минобрнауки России от 07.08.2020 № 914

РАБОЧУЮ ПРОГРАММУ разработал кандидат физико-математических наук	О.И. Бернгардт

1. Место и роль дисциплины (модуля) в структуре ОПОП

Дисциплина «Введение в технологии Больших Данных» относится к части, формируемой участниками образовательных отношений, блока 1 «Дисциплины (модули)» основной образовательной программы по направленности (профилю) подготовки Физика солнечно-земных связей направления подготовки 03.04.02 Физика, и является дисциплиной по выбору.

Предшествующие дисциплины, на которые данная дисциплина опирается: Компьютерные технологии.

2. Цели и задачи дисциплины (модуля)

Целью дисциплины «Введение в технологии Больших Данных» является формирование у обучающихся комплекса теоретических знаний и практических навыков по работе с большими данными. Знания, полученные в результате освоения дисциплины, помогут в понимании принципов работы различного программного обеспечения в этой области, а также помогут при сборе, разработке методов анализа и самом анализе структурированной или неструктурированной информации существенных объемов

Задачами дисциплины «Введение в технологии Больших Данных» является:

- получение начальных знаний о предмете Big Data
- получение начальных знаний и умений по получению Big Data
- получение начальных знаний и умений по простейшей обработке Big Data
- получение начальных знаний и умений по созданию систем обработки Big Data и использованию BigData в системах принятия решений
- получение начальных знаний и умений по созданию систем обработки Big Data и использованию BigData в системах реального времени

3. Требования к результатам освоения дисциплины (модуля)

Процесс изучения дисциплины «Введение в технологии Больших Данных» направлен на формирование следующих компетенций в соответствии с ОПОП по направлению подготовки 03.04.02 Физика:

Компетенции	**Индикаторы достижения компетенции**	**Планируемые результаты обучения по дисциплине**
ОПК-3. Способен применять знания в области информационных технологий, использовать современные компьютерные сети, программные продукты и ресурсы информационно-телекоммуникационной сети "Интернет" (далее - сеть "Интернет") для решения задач профессиональной деятельности, в том числе находящихся за пределами профильной подготовки	ИД 1. Сбор и систематизация научно-исследовательской информации о рассматриваемом объекте или явлении с использованием информационных технологий в рамках задач предметной области	**Знать** основные понятия, подходы и алгоритмы в приложении к задачам Big Data; **Уметь** применять основные понятия, подходы и алгоритмы в приложении к задачам Big Data; **Владеть** подходами и алгоритмами в приложении к задачам Big Data, выбирать оптимальный метод в зависимости от условий задачи
	ИД 2. Критическая оценка достоверности полученной научно-исследовательской информации о рассматриваемом объекте или явлении	**Знать** достоинства и недостатки современных подходов и алгоритмы в приложении к задачам Big Data, теоретические основы их работы;

		Уметь анализировать эффективность подходов и алгоритмы в приложении к конкретным задачам Big Data; **Владеть** умением определить наиболее эффективные подходы и алгоритмы в приложении к задачам Big Data, аргументировать их выбор
	ИД 4. Применение на практике методов и алгоритмов разработки программного обеспечения для решения проблем в рамках научно-исследовательских задач в том числе задач обработки наблюдательных данных.	**Знать** особенности реализации современных программно-аппаратные решения в приложении к задачам Big Data; **Уметь** применять современные программно-аппаратные решения в приложении к задачам Big Data; **Владеть** современными программно-аппаратными решениями для решения практических задач Big Data
ПКА-2. Способен проводить научные исследования в области физики солнечно-земных связей, используя необходимые знания теоретических и экспериментальных разделов физики	ИД 3. Использует современные теоретические и экспериментальные методы, включая методы обработки и анализа данных, при проведении научных исследований и реализации научных проектов в области физики солнечно-земных связей	**Знать** возможности современных программно-аппаратных решений в приложении к задачам Big Data; **Владеть** умением проводить аргументированный выбор основных подходов и алгоритмов в приложении к решению конкретных задач Big Data;

4. Объем дисциплины (модуля) и виды учебной работы

Общая трудоемкость дисциплины составляет 3 зачетных единицы, 108 часов.

Вид учебной работы	**Всего часов / зачетных единиц**
Аудиторные занятия (всего)	**36/1**
В том числе:	
Лекции	18/0,5
Лабораторные работы	
Практические занятия	18/0,5
Самостоятельная работа (всего)	**72/2**
Вид промежуточной аттестации **(зачет)**	
Контактная работа (всего)	**36/1**
Общая трудоёмкость (часы/зачетные единицы)	**108/3**

Содержание дисциплины

5.1. Содержание разделов и тем дисциплины (модуля).

Раздел 1. Введение

Тема 1. Проверка базовых знаний и вводная лекция

Введение в Big Data и машинное обучение. Объем данных, скорость данных, различность данных, качество данных, значимость данных. Получение данных. Хранение данных. Доступ к данным. Обработка данных. Проверка гипотез и выявление скрытых зависимостей. Параллельные и облачные вычисления. Машинное обучение. Искусственные нейронные сети. Входной тест.

Тема 2. Введение в Питон

Элементы языка Питон: базовые операторы, структуры и функции. Работа с файлами. Работа с модулями. Создание модулей.

Раздел 2. Дата майнинг

Тема 3. Протоколы и форматы интернета

Структура сервер-клиент. Работа с сокетами. Telnet. Организация протокола HTTP, команды, авторизация. Организация протокола FTP, команды, авторизация. Шифрование. Организация протокола SSH, авторизация. Протокол HTTPS.

Тема 4. Протоколы и форматы баз данных

Структура реляционной базы данных на примере MySQL. SQL - команды. Создание базы данных. Добавление, удаление и редактирование записей. Поиск и выбор записей. Объединения таблиц. Индексы.

Тема 5. Подготовка к анализу текстов

Команды и элементы языка regexp. Шаблоны. Контекстный поиск. Контекстный поиск и замена. Многовариантный поиск. Переменные regexp. Особенности использования на Питоне. Ассоциированные массивы - реализация и использование. Задача сбора статистики.

Тема 6. Типы данных

Числовые и категориальные данные. Особенности операций над ними. Упорядоченные и неупорядоченные данные.

Раздел 3. Предварительная обработка больших данных

Тема 7. Принятие простых решений

Статистические распределения, вероятность. Условная вероятность. Функция распределения и плотность распределения. Интегральная оценка вероятности события. Принятие решения на основе статистических данных. Байесовский подход. Проверка статистических гипотез. Ошибки первого и второго рода.

Тема 8. Выделение явных закономерностей

Корреляция. Регрессионная зависимость. Линейная и нелинейная регрессия. Метод наименьших квадратов и случаи его сведения к системе линейных уравнений. Метод максимального правдоподобия. Периодичность процессов и Фурье-анализ.

Тема 9. Выявление скрытых закономерностей.

Метод наименьших квадратов и случаи его сведения к анализу матриц. Матрицы: определители и ранги, собственные числа и собственные вектора. Уменьшение числа параметров. Линейно-зависимые и линейно-независимые параметры. Метод главных компонент. Матричный анализ. Сингулярное разложение.

Раздел 4. Экспертные системы

Тема 10. Линейные модели и деревья принятия решений.

Таблицы принятия решений. Деревья принятия решений. Листья и ветви. Обучение (построение) дерева. Условия ветвления. Отсечение ветвей. Глубина дерева и остановка.

Тема 11. Нейронные сети.

Искусственный нейрон. Искусственные нейронные сети. Входные, скрытые и выходные нейроны. Типы и структура нейронных сетей. Сети прямого распространения и рекуррентные сети. Архитектуры нейронных сетей. Обучение с учителем, без учителя и с подкреплением. Генетические алгоритмы.

Раздел 5. Технологии работы с категориальными данными

Тема 12. Векторные представления и метод мешка слов

Векторные представления и метод мешка слов. Решение задач обработки текстов методом мешка слов.

Тема 13. Оптимальные векторные представления

Оптимальные векторные представления. Слой эмбеддинга в нейронных сетях. Решение задач обработки текстов методом оптимальных векторных представлений.

Тема 14. Трансфер знаний

Трансфер знаний. Решение задач обработки текстов и изображений методом трансфера знаний с использованием больших нейронных сетей, обученных на других задачах.

Тема 15. Финальный проект.

Разработка собственной системы анализа BigData на доступных наборах больших данных (например http://www.datasciencecentral.com/profiles/blogs/big-data-sets-available-for-free; https://catalog.data.gov/dataset).

5.2. Разделы дисциплины (модуля) и виды занятий

№п/п	Раздел	Всего часов	Аудиторные занятия				СРС
			Лекции	Лаб. занятия	Практические занятия	Семинары	
1.	**Раздел 1. Введение**	**8**	**2**		**2**		**4**
2.	Тема 1. Проверка базовых знаний и вводная лекция	4	1		1		2
3.	Тема 2. Введение в Питон	4	1		1		2
4.	**Раздел 2. Дата майнинг**	**24**	**4**		**4**		**16**
5.	Тема 3. Протоколы и форматы интернета	6	1		1		4
6.	Тема 4. Протоколы и форматы баз данных	6	1		1		4
7.	Тема 5. Подготовка к анализу текстов	6	1		1		4
8.	Тема 6. Типы данных	6	1		1		4
9.	**Раздел 3. Предварительная обработка больших данных**	**19**	**3**		**4**		**12**
10.	Тема 7. Принятие простых решений	6	1		1		4
11.	Тема 8. Выделение явных закономерностей	6	1		1		4
12.	Тема 9. Выявление скрытых закономерностей.	7	1		2		4
13.	**Раздел 4. Экспертные системы**	**18**	**3**		**3**		**12**

14.	Тема 10. Линейные модели и деревья принятия решений.	6	1		1		4
15.	Тема 11. Нейронные сети.	12	2		2		8
16.	**Раздел 5. Технологии работы с категориальными данными**	**39**	**6**		**5**		**28**
17.	Тема 12. Векторные представления и метод мешка слов	6	1		1		4
18.	Тема 13. Оптимальные векторные представления	6	1		1		4
19.	Тема 14. Трансфер знаний	7	2		1		4
20.	Тема 15. Финальный проект	20	2		2		16
Итого (часы)		**108**	**18**		**18**		**72**
Итого (з.е.)		**3**	**0,5**		**0,5**		**2**

5.3. Разделы и темы дисциплины (модуля) и междисциплинарные связи

№ п/п	Наименование обеспечиваемых (последующих) дисциплин и практик	№ № разделов и/или тем данной дисциплины, необходимых для изучения обеспечиваемых (последующих) дисциплин
1.	Производственная практика (научно-исследовательская работа)	Все

5.4. Перечень лекционных занятий

№ п/п	№ раздела и темы дисциплины (модуля)	Наименование используемых технологий	Трудоемкость (часы)	Оценочные средства
1.	Р1.Т1	Тема 1. Проверка базовых знаний и вводная лекция	1	Устный опрос
2.	Р1.Т2.	Тема 2. Введение в Питон	1	Устный опрос
3.	Р2.Т3.	Тема 3. Протоколы и форматы интернета	1	Устный опрос
4.	Р2.Т4.	Тема 4. Протоколы и форматы баз данных	1	Устный опрос
5.	Р2.Т5.	Тема 5. Подготовка к анализу текстов	1	Устный опрос
6.	Р2.Т6.	Тема 6. Типы данных	1	Устный опрос
7.	Р3.Т7.	Тема 7. Принятие простых решений	1	Устный опрос
8.	Р3.Т8.	Тема 8. Выделение явных закономерностей	1	Устный опрос
9.	Р3.Т9.	Тема 9. Выявление скрытых закономерностей.	1	Устный опрос
10.	Р4.Т10.	Тема 10. Линейные модели и деревья принятия решений.	1	Устный опрос
11.	Р4.Т11.	Тема 11. Нейронные сети.	2	Устный опрос
12.	Р5.Т12.	Тема 12. Векторные представления и метод мешка слов	1	Устный опрос
13.	Р5.Т13.	Тема 13. Оптимальные векторные представления	1	Устный опрос
14.	Р5.Т14.	Тема 14. Трансфер знаний	2	Устный опрос, Итоговое тестирование
15.	Р5.Т15.	Тема 15. Финальный проект	2	Доклад

5.5. Перечень семинарских, практических занятий и лабораторных работ

№ п/п	№ раздела и темы дисциплины (модуля)	Наименование семинаров, практических и лабораторных	Трудоемкость (часы)	Оценочные средства

		работ		
1.	Р1.Т1	Тема 1. Проверка базовых знаний и вводная лекция	1	Практическое задание
2.	Р1.Т2.	Тема 2. Введение в Питон	1	Практическое задание
3.	Р2.Т3.	Тема 3. Протоколы и форматы интернета	1	Практическое задание
4.	Р2.Т4.	Тема 4. Протоколы и форматы баз данных	1	Практическое задание
5.	Р2.Т5.	Тема 5. Подготовка к анализу текстов	1	Практическое задание
6.	Р2.Т6.	Тема 6. Типы данных	1	Практическое задание
7.	Р3.Т7.	Тема 7. Принятие простых решений	1	Практическое задание
8.	Р3.Т8.	Тема 8. Выделение явных закономерностей	1	Практическое задание
9.	Р3.Т9.	Тема 9. Выявление скрытых закономерностей.	2	Практическое задание
10.	Р4.Т10.	Тема 10. Линейные модели и деревья принятия решений.	1	Практическое задание
11.	Р4.Т11.	Тема 11. Нейронные сети.	2	Практическое задание
12.	Р5.Т12.	Тема 12. Векторные представления и метод мешка слов	1	Практическое задание
13.	Р5.Т13.	Тема 13. Оптимальные векторные представления	1	Практическое задание
14.	Р5.Т14.	Тема 14. Трансфер знаний	1	Практическое задание
15.	Р5.Т15.	Тема 15. Финальный проект	2	Практическое задание

5.6. Тематика заданий для самостоятельной работы

Раздел	Тема	Вид самостоятельной работы	Задание	Рекомендуемая литература	Кол-во часов
1	Р1.Т1	Выполнение практических заданий	Решение задач по теме 1.	СИР [1-6]	2
2	Р1.Т2.	Выполнение практических заданий	Решение задач по теме 2.	СИР [1-6]	2
3	Р2.Т3.	Выполнение практических заданий	Решение задач по теме 3.	СИР [1-6]	4
4	Р2.Т4.	Выполнение практических заданий	Решение задач по теме 4.	СИР [1-6]	4
5	Р2.Т5.	Выполнение практических заданий	Решение задач по теме 5.	СИР [1-6]	4
6	Р2.Т6.	Выполнение практических заданий	Решение задач по теме 6.	СИР [1-6]	4
7	Р3.Т7.	Выполнение практических заданий	Решение задач по теме 7.	СИР [1-6]	4
8	Р3.Т8.	Выполнение практических заданий	Решение задач по теме 8.	СИР [1-6]	4
9	Р3.Т9.	Выполнение практических	Решение задач по	СИР [1-6]	4

		заданий	теме 9.		
10	Р4.Т10.	Выполнение практических заданий	Решение задач по теме 10.	СИР [1-6]	4
11	Р4.Т11.	Выполнение практических заданий	Решение задач по теме 11.	СИР [1-6]	8
12	Р5.Т12.	Выполнение практических заданий	Решение задач по теме 12.	СИР [1-6]	4
13	Р5.Т13.	Выполнение практических заданий	Решение задач по теме 13.	СИР [1-6]	4
14	Р5.Т14.	Выполнение практических заданий	Решение задач по теме 14.	СИР [1-6]	4
15	Р5.Т15.	Выполнение практических заданий	Выполнение проекта	СИР [1-6]	16

5.7. Методические рекомендации по организации самостоятельной работы обучающихся

Самостоятельная работа студентов в рамках изучения дисциплины «Введение в технологии больших данных» регламентируется общим графиком учебной работы, предусматривающим посещение практических занятий и регулярное выполнение заданий по ним, выполнение домашних заданий.

При организации самостоятельной работы по дисциплине «Введение в технологии больших данных» студенту следует:

1. Внимательно изучить материалы, характеризующие курс и тематику самостоятельного изучения дисциплины. Это позволит четко представить, как круг изучаемых тем, так и глубину их постижения.

2. Составить подборку литературы и источников, достаточную для изучения предлагаемых тем. Следует заметить, что данный курс является крайне современным, и доступная печатная литература в настоящее время отсутствует, поэтому необходимо использовать литературу из интернет-источников. В программе дисциплины представлены основной и дополнительный списки литературы. Они носят рекомендательный характер, это означает, что всегда есть литература, которая может не входить в данный список, но является необходимой для освоения темы.

3. Основное содержание той или иной проблемы следует уяснить, изучая учебную литературу и источники, с опорой на конспекты лекций.

4. Абсолютное большинство задач носит практический характер, и они могут быть решены студентом только с привлечением компьютерной обработки данных. Это предполагает наличие у студентов не только знания категорий и понятий, но и умения использовать их в качестве инструмента для анализа и выполнения практических задач.

5. Финалом практической работы студента является выполнение им финального проекта, решающего некую практическую задачу, связанную с тематикой курса (в паре или самостоятельно). Использование парного выполнения проекта поощряется, поскольку во-первых, связано с современными методиками эффективного (экстремального) программирования, позволяющим глубже разобраться в проблеме и используемых для ее решения методах, а во-вторых, учит эффективной командной работе.

Пояснительная записка к финальному проекту должна включать:

- Постановку задачи, включающую требования к функционалу системы, к программному и программному обеспечению, необходимому для ее работы, описанию исходных данных и ожидаемых результатов
- Описание проекта разрабатываемой системы в виде структуры программы, идеи решения проблемы, используемых технических решений, метода отладки.
- Исходный код программной реализации программной системы на выбранном языке программирования.
- Описание метода проверки решения, оценочные точности (если возможно).

К отчету по финальному проекту допускаются студенты, продемонстрировавшие работу программной системы. На отчете студент должен ответить на вопросы преподавателя по алгоритмам и программной реализации предложенного решения.

6. Учебно-методическое и информационное обеспечение дисциплины

6.1. Основная литература

№ п/п	Автор, название, место издания, издательство, год издания учебной и учебно-методической литературы	Количество экземпляров
1.	Джонсон, Н., Статистика и планирование эксперимента в технике и науке: Методы обработки данных [Текст] / Н. Джонсон, Ф. Лион. - М. : Мир, 1980. - 610 с.	2
2.	Водолазкий, В., Энциклопедия Perl / В. Водолазкий, В. Семериков. - СПб. : Питер, 2002. - 576 с.	2
3.	Бернгардт О.И. Введение в Большие Данные и методы машинного обучения (конспекты лекций). Часть 1. Классические методы и базовые алгоритмы	ЭБ http://irbis.iszf.irk.ru неограниченный доступ

6.2. Дополнительная литература

№ п/п	Автор, название, место издания, издательство, год издания учебной и учебно-методической литературы	Количество экземпляров
1.	Колемаев, В. А. Теория вероятностей и математическая статистика [Текст] : учебник для вузов / В. А. Колемаев, В. Н. Калинина. - 2-е изд., испр. и доп. - М. : ЮНИТИ-ДАНА, 2017. - 352 с.	ЭБ http://irbis.iszf.irk.ru неограниченный доступ
2.	Секей, Г. Парадоксы в теории вероятностей и математической статистике [Текст] : пер. с англ. / Г. Секей, В.М. Калинина - М.: Мир, 1990. - 240 с.	ЭБ http://irbis.iszf.irk.ru неограниченный доступ
3.	Доусон, Майкл. Программируем на Python : пер. с англ. / М. Доусон. - 2-е изд. - СПб : Питер, 2014. - 416 с.	ЭБ http://irbis.iszf.irk.ru неограниченный доступ
4.	Петрович, М. Л. Статистическое оценивание и проверка гипотез на ЭВМ / М. Л. Петрович, М. И. Давидович. - М.: Финансы и статистика, 1989. - 191 с.	ЭБ http://irbis.iszf.irk.ru неограниченный доступ

6.3. Профессиональные базы данных, используемые при осуществлении образовательного процесса по дисциплине:

- http://sdrus.iszf.irk.ru/

6.4. Информационные справочные системы, используемые при осуществлении образовательного процесса по дисциплине:

- Статьи по машинному обучению портала https://habrahabr.ru
- Библиотека разработчика IBM https://www.ibm.com/developerworks/analytics/library/

6.5. Ресурсы информационно-телекоммуникационной сети «Интернет», необходимые для освоения дисциплины:

- Курс MIT https://www.edx.org/course/introduction-computational-thinking-data-mitx-6-00-2x-5
- Курс Стенфорда https://www.youtube.com/watch?v=UzxYlbK2c7E
- Курс TeamDEV https://www.youtube.com/watch?v=fIZ64zHC6sU
- Открытый курс машинного обучения. https://habrahabr.ru/company/ods/blog/322626/
- Открытый курс машинного обучения http://jsman.ru/mongo-book/ (https://github.com/karlseguin/the-little-mongodb-book)

- Открытый курс машинного обучения
  https://www.ibm.com/developerworks/ru/library/l-hadoop-1/index.html

6.6. Программное обеспечение

Лицензионное и свободно распространяемое программное обеспечение, в том числе отечественного производства используемое при осуществлении образовательного процесса по дисциплине:

- Операционная система Ubuntu 18.04 (свободно распространяемое ПО)
- Офисный пакет Libre Office (свободно распространяемое ПО)
- 7-Zip (свободно распространяемое ПО)
- Adobe Acrobat Reader DC (свободно распространяемое ПО)
- Mozilla Firefox 1 (свободно распространяемое ПО)
- VLC Mediaplayer (свободно распространяемое ПО)
- K-Lite Codec Pack (свободно распространяемое ПО)
- Дистрибутив Python Anaconda (свободно распространяемое ПО)
- Набор компиляторов GCC (свободно распространяемое ПО)
- Операционная система Microsoft Windows 10 Pro
- Система ВКС VideoMost Proton

7. Образовательные технологии

В учебном процессе используются как активные, так интерактивные формы проведения занятий.

Аудиторные занятия (АЗ) проводятся в интерактивной форме с использованием мультимедийного обеспечения (ноутбук, проектор).

Практические занятия (ПЗ) включают в себя выполнение поставленных преподавателем заданий в индивидуальном и групповом порядке, заключительным этапом является выполнение финального проекта (парное или индивидуальное).

8. Практическая подготовка

Практическая подготовка обучающихся в рамках реализации данной учебной дисциплины осуществляется на практических занятиях.

9. Материально-техническое обеспечение дисциплины (модуля)

Учебная аудитория для проведения занятий лекционного типа, занятий семинарского типа, курсового проектирования, групповых и индивидуальных консультаций, текущего контроля и промежуточной аттестации	Аудитория укомплектована специализированной мебелью на 30 посадочных мест, оснащена оборудованием и техническими средствами обучения, служащими для представления учебной информации большой аудитории: • доска магнитно-маркерная Branberg • экран для проектора Projecta • проектор BenQ MH733 1920 x 1080 • ноутбук ASUS L1500CDA Windows 10 Pro • система акустическая Electro Voice EVID 6.2
Учебная аудитория для групповых и индивидуальных консультаций и самостоятельной работы	Аудитория укомплектована специализированной мебелью на 7 посадочных мест, оснащена компьютерной техникой с возможностью подключения к сети «Интернет» и обеспечением доступа к электронной информационно-образовательной среде: • персональные компьютеры Нettop Think Center Lenovo M710Q

	• мониторы IIYAMA PL2283H, Dell CRHX9K2 • доска магнитно-маркерная Branberg • экран для проектора Projecta • проектор BenQ MH733 1920 x 1080

10. Фонд оценочных средств

В результате освоения дисциплины обучающийся должен:

Знать:

1. основные понятия, подходы и алгоритмы в приложении к задачам Big Data;
2. достоинства и недостатки современных подходов и алгоритмы в приложении к задачам Big Data, теоретические основы их работы;
3. особенности реализации современных программно-аппаратные решения в приложении к задачам Big Data;
4. возможности современных программно-аппаратных решений в приложении к задачам Big Data.

Уметь:

1. применять основные понятия, подходы и алгоритмы в приложении к задачам Big Data;
2. анализировать эффективность подходов и алгоритмы в приложении к конкретным задачам Big Data;
3. применять современные программно-аппаратные решения в приложении к задачам Big Data.

Владеть:

1. подходами и алгоритмами в приложении к задачам Big Data, выбирать оптимальный метод в зависимости от условий задачи;
2. умением определить наиболее эффективные подходы и алгоритмы в приложении к задачам Big Data, аргументировать их выбор;
3. современными программно-аппаратными решениями для решения практических задач Big Data;
4. умением проводить аргументированный выбор основных подходов и алгоритмов в приложении к решению конкретных задач Big Data.

Перечень компетенций с указанием этапов их формирования в процессе освоения образовательной программы

Код компетенции	*Разделы дисциплины, направленные на формирование компетенции*				
	1	*2*	*3*	*4*	*5*
ОПК-3	+	+	+	+	+
ПКА-2	+	+	+	+	+

Описание показателей и критериев оценивания компетенций на различных этапах их формирования, описание шкал оценивания

Код компетенции	**Показатели (индикаторы)**	**Формы оценивания**			
		Текущий контроль			**Промежуточная аттестация**
		Устный опрос	**Контроль самостоятельной работы**	**Тестирование**	**Зачет/экзамен**
ОПК-3	**Знать** основные понятия, подходы и алгоритмы в приложении к задачам Big Data; достоинства	вопросы 1-10	задачи для решения на практических занятиях	Итоговый тест	зачет

	и недостатки современных подходов и алгоритмы в приложении к задачам Big Data, теоретические основы их работы; особенности реализации современных программно-аппаратные решения в приложении к задачам Big Data; **Уметь** применять основные понятия, подходы и алгоритмы в приложении к задачам Big Data; анализировать эффективность подходов и алгоритмы в приложении к конкретным задачам Big Data; применять современные программно-аппаратные решения в приложении к задачам Big Data. **Владеть** подходами и алгоритмами в приложении к задачам Big Data, выбирать оптимальный метод в зависимости от условий задачи подходами и алгоритмами в приложении к задачам Big Data, выбирать оптимальный метод в зависимости от условий задачи; умением определить наиболее эффективные подходы и алгоритмы в приложении к задачам Big Data, аргументировать их выбор; современными программно-аппаратными решениями для решения практических задач Big Data;				
ПКА-2	**Знать** возможности современных программно-аппаратных решений в приложении к задачам	Вопросы 11-20	задачи для решения на практических занятиях	Итоговый тест	

	Big Data; **Владеть** умением проводить аргументированный выбор основных подходов и алгоритмов в приложении к решению конкретных задач Big Data;				

Программа оценивания контролируемой компетенции

Тема или раздел дисциплины	Формируемый признак компетенции	Показатель	Критерий оценивания	Наименование ОС	
				ТК	ПА
Раздел 1. Введение	ОПК-3, ПКА-2	Отвечает на вопросы по изученному материалу	Владеет материалом раздела 1.	устный групповой опрос, решение задач	зачет
Раздел 2. Дата майнинг	ОПК-3, ПКА-2	Отвечает на вопросы по изученному материалу	Владеет материалом раздела 2.	устный групповой опрос, решение задач	зачет
Раздел 3. Предварительная обработка больших данных	ОПК-3, ПКА-2	Отвечает на вопросы по изученному материалу	Владеет материалом раздела 3.	устный групповой опрос, решение задач	зачет
Раздел 4. Экспертные системы	ОПК-3, ПКА-2	Отвечает на вопросы по изученному материалу	Владеет материалом раздела 4.	устный групповой опрос, решение задач	зачет
Раздел 5. Технологии работы с категориальными данными	ОПК-3, ПКА-2	Отвечает на вопросы по изученному материалу	Владеет материалом раздела 5.	устный групповой опрос, решение задач	зачет

Текущая и промежуточная аттестация

Цель контроля - получение информации о результатах обучения и степени их соответствия результатам обучения.

Текущий контроль

Текущий контроль успеваемости студента, т.е. проверка усвоения учебного материала, регулярно осуществляется на протяжении семестра. Текущий контроль знаний обучающихся организован как устный групповой опрос, выполнение практических заданий и самостоятельной работы.

Раздел / Тема	Индекс и уровень формируемой компетенции или дескриптора	ОС	Содержание задания
Разделы 1-5	ОПК-3, ПКА-2	устный опрос, выполнение практических заданий, итоговое тестирование	Изложить свое мнение по проблемным вопросам по изученным разделам.
Разделы 1-5	ОПК-3, ПКА-2	устный опрос, выполнение практических заданий, финальный	Разработать и защитить финальный проект, решающий практическую задачу на основе знаний,

		проект	полученных на курсе.

Вопросы для устного опроса

Раздел 1

1. Какие из перечисленных систем являются системами, описываемыми термином 'Big Data', почему:

а) Файл-лог доступа к популярному сайту (например lenta.ru) за 7 дней - время, адрес клиента, адрес просмотренной страницы

б) Полнотекстовая необновляемая или редкообновляемая библиотека электронных текстов

в) Полнотекстовая частообновляемая библиотека электронных текстов

г) Система поиска по автору в полнотекстовой необновляемой библиотеке электронных текстов

д) Система поиска по автору в полнотекстовой обновляемой библиотеке электронных текстов

е) Система контекстного поиска по полнотекстовой необновляемой библиотеке электронных текстов

ж) Система контекстного поиска по полнотекстовой обновляемой библиотеке электронных текстов

2. Словари на языке Питон - реализация и использование.

3. Понятие Big Data и машинное обучение. Что понимается под объемом, скоростью и различностью данных?

Раздел 2

1. Что такое regexp, какие команды и элементы языка regexp вы знаете?
2. Чем категориальные данные отличаются от числовых, почему их надо различать?

Раздел 3

1. Чем корреляция Пирсона отличатеся от корреляции Спирмена?
2. Для чего используется линейная регрессия?

Раздел 4

1. Что такое дерево принятия решений?
2. Сколько нейронов должно быть на выходном слое нейронной сети, выполняющей классификацию изображения на 3 класса?

Раздел 5

1. Что такое трансфер знаний, для чего его используют?
2. С какого слоя сети желательно брать признаки при трансфере знаний: с первых или с последних?

Задания для практических занятий

1. Дана строка s:

s='Мама мыла раму. Раму мыла мама.'

Производится поиск по регулярному выражению:

([\w\s]*[\.\?])

Какие символы в этом регулярном выражении не используется при анализе этой строки?

2. Даны 2 числовых ряда: 1,8,7,2,5,4,3 и 0.4, 3, 2.5, 0.5, 1.5, 1.5, 1.1
Рассчитать методом наименьших квадратов регрессионную зависимость одного ряда от другого. Определить, чему будет равно следующее значение второго числового ряда, если следующее значение первого числового ряда 12. А каких пределах может находится следующее значение?

3. Выявление скрытых закономерностей.
Даны 5 рядов:

13,27,4,-6,-20,17,-16,-2,9,4
7,9,6,-8,-13,5,-7,-6,8,-3
11,-17,-9,14,9,-12,16,9,-1,7
4,-26,-15,22,22,-17,23,15,-9,10
1,-9,8,-10,-6,-7,2,-10,7,-10

Выяснить, сколько и какие из них являются линейно независимыми от других.

4. Вашей экспертной системе необходимо проверить, выдадут-ли человеку кредит в банке. Постройте примерное дерево принятия решения для решения этой задачи.

5. Вам необходимо построить генератор на нейронной сети, а именно: после подачи на ее вход двух входных импульсов, вы должны получить периодически изменяющийся, бесконечный сигнал на выходе. Какого типа нейронная сеть подходит для этой задачи. Почему?

6. Вам необходимо построить нейронную сеть, отличающую изображения собак от изображений кошек. Предложите вариант ее реализации с помощью нейронных сетей. Сколько нейронов должно быть на выходном слое и какая у них должна быть функция активации?

7. Вам необходимо построить нейронную сеть, прогнозирующую число пятен на Солнце (числа Вольфа). Предложите вариант ее реализации с помощью нейронных сетей. Сколько нейронов должно быть на выходном слое и какая у них должна быть функция активации?

Пример итогового тестирования

Продолжить утверждения:

1. Большие данные отличаются от обычных данных тем, что .....

2. Основной задачей анализа больших данных является .....

3. Основной проблемой при анализе больших данных является .....

4. Одним из основных способов уменьшения количества анализируемых параметров является .....

5. Задача определения коэффициентов A-F при аппроксимации неизвестной функции функцией вида A+Bx+Cx^2+Dx^3+E cos(x)+F sin(2x) может быть сведена к решению .....

6. Основным условием, определяющим последовательность выбора параметра, по которому каждый раз проводится ветвление при построении решающего дерева является .....

7. Отсечение ветвей в решающем дереве используют, чтобы .....

8. Если в данных много параметров со случайными значениями, построение решающего дерева без предварительных преобразований переменных приводит к .....

9. Для выполенения замены всех слов в строке, состоящих из 7-14 букв на слово «длинноеслово» на языке regexp необходимо .....

10. Наиболее затратной с точки зрения времени при использовании метода нейронных сетей является процесс ее .....

11. Метод главных компонент позволяет выделить в данных .....

12. Для дальнейшего эффективного использования нейронной сети или решающего дерева необходимо после их обучения .....

13. Основная последовательность действий при работе с большими данными имеет вид:.....

14. Обучение случайного леса решений проводится в несколько этапов:.....

15. Для работы с большими данными при обработке файлов программа должна .....

16. Трансфер знаний используется для того, чтобы .....…

17. Переобученная нейронная сеть дает …. качество на обучающем датасете и … качество на тестовом датасете

18. Для обучения нейронной сети мы обычно делим данные нам датасет на … части: (...перечислить…)

19. Задача кластеризации - это задача машинного обучения (без учителя/с учителем — выбрать)

20. Задача регрессии — это задача машинного обучения (без учителя/с учителем — выбрать)

Промежуточная аттестация

Промежуточная аттестация студентов по дисциплине осуществляется по окончанию дисциплины, в виде зачета в соответствии с графиком учебного процесса. Проверка наличия конспектов по дисциплине является допуском к зачёту. В случае наличия учебной задолженности (пропущенных занятий или невыполненных заданий), студент отрабатывает пропущенные занятия и выполняет задания.

Вопросы к зачету

1. Какие из перечисленных систем являются системами, описываемыми термином 'Big Data':

а) Файл-лог доступа к популярному сайту (например lenta.ru) за 7 дней - время, адрес клиента, адрес просмотренной страницы

б) Полнотекстовая необновляемая или редкообновляемая библиотека электронных текстов

в) Полнотекстовая частообновляемая библиотека электронных текстов

г) Система поиска по автору в полнотекстовой необновляемой библиотеке электронных текстов

д) Система поиска по автору в полнотекстовой обновляемой библиотеке электронных текстов

е) Система контекстного поиска по полнотекстовой необновляемой библиотеке электронных текстов

ж) Система контекстного поиска по полнотекстовой обновляемой библиотеке электронных текстов

Пояснить выбор.

2. Вероятность какого события больше - выпадение орла на стандартной монете или выпадение четного числа на стандартной игральной кости?
3. Указать сходства и отличия случайного леса деревьев и нейронной сети
4. Понятие Big Data и машинное обучение. Объем данных, скорость данных, различность данных, качество данных, значимость данных. Получение данных. Хранение данных. Доступ к данным.
5. Команды и элементы языка regexp. Шаблоны. Контекстный поиск. Контекстный поиск и замена. Многовариантный поиск. Переменные regexp. Особенности использования на Питоне.
6. Ассоциированные массивы (словари) на Питон - реализация и использование.

7. Принятие решения на основе статистических данных. Байесовский подход. Проверка статистических гипотез. Ошибки первого и второго рода.
8. Корреляция. Регрессионная зависимость. Линейная и нелинейная регрессия. Метод наименьших квадратов и случаи его сведения к системе линейных уравнений.
9. Уменьшение числа параметров. Линейно-зависимые и линейно-независимые параметры. Метод главных компонент.
10. Обучение без учителя. Кластеризация данных.
11. Деревья принятия решений. Обучение (построение) дерева. Отсечение ветвей. Глубина дерева и остановка.
12. Искусственный нейрон. Искусственные нейронные сети. Входные, скрытые и выходные нейроны. Обучение с учителем. Классификация и регрессия.
13. Построение нейронных сетей с оптимальными векторными представлениями. Слой эмбеддинга
14. Трансфер знаний. Использование нейронных сетей, обученных на других задачах, для решения своих задач

Задание для финального проекта

Пример 1: Обучить нейронную сеть, отличающую изображения кошек от изображений собак.
Пример 2: Обучить нейронную сеть, прогнозирующую солнечную активность (числа Вольфа).

Оценочные средства сформированности компетенций

Компетенция	**Индекс достижения компетенции**	**Задание**
ОПК-3 Способен применять знания в области информационных технологий, использовать современные компьютерные сети, программные продукты и ресурсы информационно-телекоммуникационной сети "Интернет" (далее - сеть "Интернет") для решения задач профессиональной деятельности, в том числе находящихся за пределами профильной подготовки	**ИД 1. Знать** основные понятия, подходы и алгоритмы в приложении к задачам Big Data; **Уметь** применять основные понятия, подходы и алгоритмы в приложении к задачам Big Data;	Вопросы для зачета 1-14
	ИД 1. Владеть подходами и алгоритмами в приложении к задачам Big Data, выбирать оптимальный метод в зависимости от условий задачи	Финальный проект
	ИД-2. Знать достоинства и недостатки современных подходов и алгоритмы в приложении к задачам Big Data, теоретические основы их работы; **Уметь** анализировать эффективность подходов и алгоритмы в приложении к конкретным задачам Big Data;	Вопросы для зачета 1-14
	ИД-2. Владеть умением определить наиболее эффективные подходы и алгоритмы в приложении к задачам Big Data, аргументировать их выбор	Финальный проект
	ИД 4. Знать особенности реализации современных программно-аппаратные решения в приложении к задачам Big Data;	Вопросы для зачета 1-14

	ИД 4. Уметь применять современные программно-аппаратные решения в приложении к задачам Big Data; **Владеть** современными программно-аппаратными решениями для решения практических задач Big Data	Финальный проект
ПКА-2 Способен проводить научные исследования в области физики солнечно-земных связей, используя необходимые знания теоретических и экспериментальных разделов физики	**ИД 3. Знать** возможности современных программно-аппаратных решений в приложении к задачам Big Data;	Вопросы для зачета 1-14
	ИД 3. Владеть умением проводить аргументированный выбор основных подходов и алгоритмов в приложении к решению конкретных задач Big Data;	Финальный проект

Критерии оценки:

- оценка «зачтено» выставляется студенту, если основной материал усвоен, студент приобрел необходимые знания и умения;
- оценка «не зачтено» - если основной материал усвоен недостаточно, студент не приобрел необходимых знаний и умений

Оценочные средства, обеспечивающие диагностику сформированности компетенций, заявленных в рабочей программе дисциплины (модуля)

Результат диагностики сформированности компетенций	**Показатели**	**Критерии**	**Соответствие / несоответствие**	**Зачет / экзамен**
Положительные результаты устного промежуточного контроля	подготовка к устному промежуточному контролю, знание основных тем дисциплины, указанных в Программе оценивания контролируемой компетенции	Дал грамотный и развернутый ответ на вопросы для подготовки по теоретическим вопросам курса Не ответил или ответил неправильно на вопросы для подготовки по теоретическим вопросам курса	Соответствие Несоответствие	зачет
Положительные результаты решения задач	Решение предложенных преподавателем задач, знание основных тем дисциплины, в Программе оценивания контролируемой компетенции	Положительные результаты решения задач Не решил или неправильно решил предложенные задачи	Соответствие Несоответствие	зачет
Положительные результаты зачета	Подготовка к зачету, знание вопросов , ответы на зачете	Полностью раскрыты все вопросы, даны все правильные определения Не полностью раскрыт один из вопросов и (или) в определениях есть неточности	Соответствие Соответствие	зачет

		Не полностью раскрыты два вопроса и (или) определения неверны	Несоответствие	