МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ

федеральное государственное бюджетное образовательное учреждение высшего образования

«ИРКУТСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ»

(ФГБОУ ВО «ИГУ»)

Институт математики и информационных технологий

Кафедра вычислительной математики и оптимизации

«УТВЕРЖДАЮ»
Директор ИМИТ ИГУ
_____ М. В. Фалалеев
«19» мая 2021 г.

Рабочая программа дисциплины (модуля)

Б1.В.15 Обработка больших объемов данных

Направление подготовки	01.03.02 Прикладная математика и информатика
Направленность (профиль) подготовки	Прикладная математика и информатика
Квалификация выпускника	бакалавр
Форма обучения	очная

Иркутск 2021 г.

1. ЦЕЛИ И ЗАДАЧИ ДИСЦИПЛИНЫ

Цели: формирование компетенций специалиста по направлению "Прикладная математика и информатика" в предметной области, связанной с решением задач сбора и анализа огромных объемов структурированной или слабоструктурированной информации, разработке на ее основе моделей данных и извлечении новых знаний.
Задачи: приобретение студентами знаний о технологиях подготовки, хранения, обработки и анализа больших данных; применение статистических и математических методов для анализа больших объемов информации; приобретение практических навыков работы с большими данными с помощью сред на базе языков программирования Python (R)

2. МЕСТО ДИСЦИПЛИНЫ В СТРУКТУРЕ ОПОП ВО

Учебная дисциплина Б1.В.15 Обработка больших объемов данных относится к части Блока 1 образовательной программы, формируемой участниками образовательных отношений.

3. ТРЕБОВАНИЯ К РЕЗУЛЬТАТАМ ОСВОЕНИЯ ДИСЦИПЛИНЫ

Процесс освоения дисциплины направлен на формирование следующих компетенций в соответствии с ФГОС ВО и ОП ВО по направлению подготовки 01.03.02 Прикладная математика и информатика:
ПК-3 Способен создавать, модифицировать и сопровождать информационные системы, автоматизирующие задачи организационного управления и процессов функционирования производственных организаций, социальных институтов и структур.

4. СОДЕРЖАНИЕ И СТРУКТУРА ДИСЦИПЛИНЫ

Объем дисциплины составляет 4 зачетных ед., 144 час.

Форма промежуточной аттестации: зачет с оценкой.

4.1. Содержание дисциплины, структурированное по темам, с указанием видов учебных занятий и отведенного на них количества академических часов

Раздел дисциплины / тема	Виды учебной работы				Формы текущего контроля; Формы промежут. аттестации
	Контактная работа преподавателя с обучающимися			Самост. работа	
	Лекции	Лаб. занятия	Практ. занятия		
Тема 1. Введение в большие данные	4	4		10	
Тема 2. Жизненный цикл анализа больших данных	4	4		10	
Тема 3. Корреляция и регрессионный анализ.	4	4		10	
Тема 4. Технологии хранения и обработки больших данных	4	4		10	
Тема 5. Языки Python и R. Синтаксис языка R, основные типы данных	4	4		10	
Тема 6. Подготовка данных. Визуализация данных. Понимание данных.	4	4		10	
Тема 7. Парадигма Map Reduce. Ее реализация Hadoop	8	8		12	
Итого (8 семестр):	32	32		72	зач.с оц.

4.2. Содержание учебного материала

Тема 1. Введение в большие данные

Основные определения, термины, задачи анализа больших данных. Понятие Data Mining. Системный анализ и методы его проведения. Методики анализа больших данных

Тема 2. Жизненный цикл анализа больших данных

Создание данных (Data Generation/Data Capture). Создание данных (Data Generation/Data Capture). Создание данных (Data Generation/Data Capture). Использование данных (Data Usage). Публикация данных (Data Publication). Публикация данных (Data Publication). Уничтожение данных (Data Purging). Жизненный цикл метаданных.

Тема 3. Корреляция и регрессионный анализ.

Корреляция и регрессионный анализ. Коэффициент корреляции. Графическое представление. Постановка задачи регрессионного анализа. Линейная регрессия. Метод наименьших квадратов. Их роль в аналитике больших данных.

Тема 4. Технологии хранения и обработки больших данных

Обзор технологий хранения больших данных. Базы данных. Системы управления базами данных. Модели данных. Подготовка исходных данных для анализа: первичная обработка и визуализация имеющихся данных. Базы данных NoSQL.

Тема 5. Языки Python и R. Синтаксис языка R, основные типы данных

Роль языков программирования Python и R в аналитике больших данных. Необходимый набор библиотек. Готовые решения анализа данных и их роль в области больших данных.

Тема 6. Подготовка данных. Визуализация данных. Понимание данных.

Методы предварительной подготовки данных. Инструменты и методы визуализации данных.

Тема 7. Парадигма Map Reduce. Ее реализация Hadoop

Парадигма Map Reduce. Роль Map Reduce в аналитике больших данных. Оператор Map. Лямбда-архитектура.

4.3. Методические указания по организации самостоятельной работы студентов

Самостоятельная работа студентов всех форм и видов обучения является одним из обязательных видов образовательной деятельности, обеспечивающей реализацию требований Федеральных государственных стандартов высшего образования. Согласно требованиям нормативных документов самостоятельная работа студентов является обязательным компонентом образовательного процесса, так как она обеспечивает закрепление получаемых на лекционных занятиях знаний путем приобретения навыков осмысления и расширения их содержания, навыков решения актуальных проблем формирования общекультурных и профессиональных компетенций, научно-исследовательской деятельности, подготовки к семинарам, лабораторным работам, сдаче зачетов и экзаменов. Самостоятельная работа студентов представляет собой совокупность аудиторных и внеаудиторных занятий и работ. Самостоятельная работа в рамках образовательного процесса в вузе решает следующие задачи:

– закрепление и расширение знаний, умений, полученных студентами во время аудиторных и внеаудиторных занятий, превращение их в стереотипы умственной и физической деятельности;

– приобретение дополнительных знаний и навыков по дисциплинам учебного плана;

– формирование и развитие знаний и навыков, связанных с научно-исследовательской деятельностью;

– развитие ориентации и установки на качественное освоение образовательной программы;

– развитие навыков самоорганизации;

– формирование самостоятельности мышления, способности к саморазвитию, самосовершенствованию и самореализации;

– выработка навыков эффективной самостоятельной профессиональной теоретической, практической и учебно-исследовательской деятельности.

Подготовка к лекции. Качество освоения содержания конкретной дисциплины прямо зависит от того, насколько студент сам, без внешнего принуждения формирует у себя установку на получение на лекциях новых знаний, дополняющих уже имеющиеся по данной дисциплине. Время на подготовку студентов к двухчасовой лекции по нормативам составляет не менее 0,2 часа.

Подготовка к практическому занятию. Подготовка к практическому занятию включает следующие элементы самостоятельной деятельности: четкое представление цели и задач его проведения; выделение навыков умственной, аналитической, научной деятельности, которые станут результатом предстоящей работы. Выработка навыков осуществляется с помощью получения новой информации об изучаемых процессах и с помощью знания о том, в какой степени в данное время студент владеет методами исследовательской деятельности, которыми он станет пользоваться на практическом занятии. Подготовка к практическому занятию нередко требует подбора материала, данных и специальных источников, с которыми предстоит учебная работа. Студенты должны дома подготовить к занятию 3–4 примера формулировки темы исследования, представленного в монографиях, научных статьях, отчетах. Затем они самостоятельно

осуществляют поиск соответствующих источников, определяют актуальность конкретного исследования процессов и явлений, выделяют основные способы доказательства авторами научных работ ценности того, чем они занимаются. В ходе самого практического занятия студенты сначала представляют найденные ими варианты формулировки актуальности исследования, обсуждают их и обосновывают свое мнение о наилучшем варианте. Время на подготовку к практическому занятию по нормативам составляет не менее 0,2 часа.

Подготовка к семинарскому занятию. Самостоятельная подготовка к семинару направлена: на развитие способности к чтению научной и иной литературы; на поиск дополнительной информации, позволяющей глубже разобраться в некоторых вопросах; на выделение при работе с разными источниками необходимой информации, которая требуется для полного ответа на вопросы плана семинарского занятия; на выработку умения правильно выписывать высказывания авторов из имеющихся источников информации, оформлять их по библиографическим нормам; на развитие умения осуществлять анализ выбранных источников информации; на подготовку собственного выступления по обсуждаемым вопросам; на формирование навыка оперативного реагирования на разные мнения, которые могут возникать при обсуждении тех или иных научных проблем. Время на подготовку к семинару по нормативам составляет не менее 0,2 часа.

Подготовка к коллоквиуму. Коллоквиум представляет собой коллективное обсуждение раздела дисциплины на основе самостоятельного изучения этого раздела студентами. Подготовка к данному виду учебных занятий осуществляется в следующем порядке. Преподаватель дает список вопросов, ответы на которые следует получить при изучении определенного перечня научных источников. Студентам во внеаудиторное время необходимо прочитать специальную литературу, выписать из нее ответы на вопросы, которые будут обсуждаться на коллоквиуме, мысленно сформулировать свое мнение по каждому из вопросов, которое они выскажут на занятии. Время на подготовку к коллоквиуму по нормативам составляет не менее 0,2 часа.

Подготовка к контрольной работе. Контрольная работа назначается после изучения определенного раздела (разделов) дисциплины и представляет собой совокупность развернутых письменных ответов студентов на вопросы, которые они заранее получают от преподавателя. Самостоятельная подготовка к контрольной работе включает в себя: — изучение конспектов лекций, раскрывающих материал, знание которого проверяется контрольной работой; повторение учебного материала, полученного при подготовке к семинарским, практическим занятиям и во время их проведения; изучение дополнительной литературы, в которой конкретизируется содержание проверяемых знаний; составление в мысленной форме ответов на поставленные в контрольной работе вопросы; формирование психологической установки на успешное выполнение всех заданий. Время на подготовку к контрольной работе по нормативам составляет 2 часа.

Подготовка к зачету. Самостоятельная подготовка к зачету должна осуществляться в течение всего семестра. Подготовка включает следующие действия: перечитать все лекции, а также материалы, которые готовились к семинарским и практическим занятиям в течение семестра, соотнести эту информацию с вопросами, которые даны к зачету, если информации недостаточно, ответы находят в предложенной преподавателем литературе. Рекомендуется делать краткие записи. Время на подготовку к зачету по нормативам составляет не менее 4 часов.

Подготовка к экзамену. Самостоятельная подготовка к экзамену схожа с подготовкой к зачету, особенно если он дифференцированный. Но объем учебного материала, который нужно восстановить в памяти к экзамену, вновь осмыслить и понять, значительно больше, поэтому требуется больше времени и умственных усилий. Важно сформировать целостное представление о содержании ответа на каждый вопрос, что

предполагает знание разных научных трактовок сущности того или иного явления, процесса, умение раскрывать факторы, определяющие их противоречивость, знание имен ученых, изучавших обсуждаемую проблему. Необходимо также привести информацию о материалах эмпирических исследований, что указывает на всестороннюю подготовку студента к экзамену. Время на подготовку к экзамену по нормативам составляет 36 часов для бакалавров.

В ФБГОУ ВО «ИГУ» организация самостоятельной работы студентов регламентируется Положением о самостоятельной работе студентов, принятым Ученым советом ИГУ 22 июня 2012 г.

5. УЧЕБНО-МЕТОДИЧЕСКОЕ И ИНФОРМАЦИОННОЕ ОБЕСПЕЧЕНИЕ ДИСЦИПЛИНЫ

5.1. Литература, базы данных, информационно-справочные и поисковые системы

1. Макшанов, А. В. Большие данные. Big Data / А. В. Макшанов, А. Е. Журавлев, Л. Н. Тындыкарь. - 2-е изд., стер. - Санкт-Петербург : Лань, 2022. - 188 с. - ISBN 978-5-8114-9690-7. - Текст : электронный // Лань : электронно-библиотечная система. - URL: https://e.lanbook.com/book/198599. - Режим доступа: для авториз. пользователей.
2. Миркин, Б. Г. Введение в анализ данных : учебник и практикум / Б. Г. Миркин. - Москва : Издательство Юрайт, 2020. - 174 с. - (Высшее образование). - ISBN 978-5-9916-5009-0. - Текст : электронный // ЭБС Юрайт [сайт]. - URL: http://www.biblio-online.ru/bcode/450262.

6. МАТЕРИАЛЬНО-ТЕХНИЧЕСКОЕ ОБЕСПЕЧЕНИЕ ДИСЦИПЛИНЫ

6.1. Учебная аудитория для проведения:

– занятий лекционного типа,
– занятий семинарского (практического) типа,
– групповых и индивидуальных консультаций,
– текущего контроля и промежуточной аттестации.

Оснащение:

Учебная аудитория, укомплектованная специализированной мебелью и техническими средствами обучения, служащими для представления учебной информации большой аудитории, для проведения занятий лекционного типа, практических занятий (семинарского типа), курсового проектирования (выполнения курсовых работ), групповых и индивидуальных консультаций, текущего контроля и промежуточной аттестации. Для проведения занятий лекционного типа обучающимся предлагаются наборы демонстрационного оборудования и учебно-наглядные пособия, обеспечивающие тематические иллюстрации.

6.2. Помещения для самостоятельной работы обучающихся.

Оснащение:

Помещения для самостоятельной работы обучающихся, оснащенные учебной мебелью. Рабочие места обучающихся оборудованы компьютерной техникой и подключены в локальную вычислительную сеть, в т.ч. с использованием беспроводного Wi-Fi подключения, с возможностью выхода в глобальную сеть Интернет и с доступом в электронную информационно-образовательную среду.

6.3. Программное обеспечение

Приложение для чтения PDF-файлов, браузер для просмотра интернет контента, приложение для создания PDF-файлов.

7. ОЦЕНОЧНЫЕ МАТЕРИАЛЫ ДЛЯ ТЕКУЩЕГО КОНТРОЛЯ И ПРОМЕЖУТОЧНОЙ АТТЕСТАЦИИ

7.1. Оценочные средства для промежуточной аттестации

Список вопросов для промежуточной аттестации:

1. Понятие Большие данные. Роль цифровой информации в 21 веке
2. Виды массивов данных.
3. Корреляция и регрессионный анализ. Коэффициент корреляции. Графическое представление.
4. Постановка задачи регрессионного анализа. Линейная регрессия. Метод наименьших квадратов. Привести примеры использования регрессионного анализа.
5. Классификация. Признаковое описание объекта и таблица объектсвойства. Постановка задачи. Отличия задачи классификации от задачи регрессии.
6. Определение модели и алгоритма. Процесс обучения. Проблема переобучения. Р
7. Регуляризация. Cross validation. Привести примеры использования алгоритмов классификации.
8. Кластеризация. Метрики. Матрица парных расстояний. Постановка задачи кластеризации. Отличие от задачи классификации.
9. Ассоциативные правила. Определение. Достоверность и поддержка. Отличия построения ассоциативного правила от решающего правила задачи классификации.
10. Парадигма Map Reduce. Описать принцип работы. Нарисовать диаграмму. Перечислить слабые и сильные стороны. Обозначить области применимости.
11. Визуализация. Дать определение визуализации. Показать важность визуализации в аналитике больших данных. Привести примеры использования визуализации.
12. "Жизненный цикл" проекта по аналитике больших данных.
13. Типовая архитектура проекта в области больших данных. Перечислить используемые технологии, указать степень вовлеченности каждой из технологий на каждом этапе работы над проектом.
14. Современные научные проблемы больших данных. Показать значимость проблем, актуальность, связь с областями математики и инженерии.
15. Определение больших данных, ключевые характеристики. Примеры задач больших данных. Основные виды данных
16. Роль аналитика по данным (Data Scientist). Ключевые компетенции аналитика. Отличия BI от Data Science.
17. Использование модели множественной линейной регрессии для прогнозирования экономических показателей.
18. Доверительные интервалы для зависимой переменной.
19. Сглаживание временных рядов. Динамические модели с распределенными лагами.
20. Стационарные временные ряды. Тестирование стационарности.
21. Коинтеграция. Анализ временных рядов.
22. Адаптивные и мультипликативные методы прогнозирования. Экспоненциальное сглаживание.
23. Авторегрессионные модели. Модели скользящего среднего.
24. Интегрированные процессы. Идентификация авторегрессионной модели скользящего среднего.

25. Прогнозирование с моделями временных рядов. Доверительные интервалы прогноза.
26. Предсказание и прогнозирование социально-экономических прогнозов.
27. Дисперсионный анализ влияния качественных факторов. Ранговые методы.
28. Факторный анализ. Метод главных факторов.
29. Многомерное шкалирование. Классическая модель многомерного шкалирования.
30. Неметрические методы. Кластерный анализ. Дискриминантный анализ.
31. Многомерный статистический анализ