МИНОБРНАУКИ РОССИЙСКОЙ ФЕДЕРАЦИИ
НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
ТОМСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ

Институт прикладной математики и компьютерных наук

УТВЕРЖДАЮ
Директор института прикладной
математики и компьютерных наук
_____ А.В. Замятин
«_11_» _____ 2021 г.

Анализ больших массивов данных (Big Data)

рабочая программа дисциплины

Закреплена за кафедрой	*теоретических основ информатики*
Учебный план	*01.03.02 Прикладная математика и информатика, профиль «Прикладная математика и информатика»*
Форма обучения	*очная*
Общая трудоёмкость	*2 з.е.*
Часов по учебному плану	*72*
в том числе:	
аудиторная контактная работа	*33,85*
самостоятельная работа	*38,15*
Вид(ы) контроля в семестрах	
экзамен/зачет/зачет с оценкой	*Семестр 8 – зачет*

Томск-2021

Программу составил:
д-р техн. наук, профессор,
заведующий кафедрой теоретических основ информатики А.В. Замятин

Рецензент:
д-р техн. наук, профессор,
заведующий кафедрой прикладной информатики С.П. Сущенко

Рабочая программа дисциплины «Анализ больших массивов данных (Big Data)» разработана в соответствии с образовательным стандартом высшего образования – бакалавриат, самостоятельно устанавливаемым федеральным государственным автономным образовательным учреждением высшего образования «Национальный исследовательский Томский государственный университет» по направлению подготовки 01.03.02 Прикладная математика и информатика (утвержден Ученым советом НИ ТГУ, протокол от 27.10.2021 г. № 08).

Рабочая программа одобрена на заседании кафедры теоретических основ информатики

Протокол от 04 июня 2021 г. № 05

Заведующий кафедрой теоретических основ информатики,
д-р техн. наук, профессор А.В. Замятин

Рабочая программа одобрена на заседании учебно-методической комиссии института прикладной математики и компьютерных наук (УМК ИПМКН)

Протокол от 17 июня 2021 г. № 05

Председатель УМК ИПМКН,
д-р техн. наук, профессор С.П. Сущенко

Цель освоения дисциплины

Цель – получение знаний в области моделей и методов интеллектуального анализа данных в задачах поиска информации, обработки и анализа данных, а также приобретение навыков исследователя данных (data scientist) и разработчика математических моделей, методов и алгоритмов анализа данных.

1. Место дисциплины в структуре ОПОП

Дисциплина «Анализ больших массивов данных (Big Data)» относится к обязательной части Блока 1 «Дисциплины (модули)», входит в модуль «Математика».

Пререквизиты: «Интеллектуальные информационные системы».

Постреквизиты: нет.

2. Компетенции и результаты обучения, формируемые в результате освоения дисциплины

Таблица 1.

Компетенция	**Индикатор компетенции**	**Код и наименование результатов обучения** (планируемые результаты обучения по дисциплине, характеризующие этапы формирования компетенций)
ОПК-2. Способен использовать и адаптировать существующие математические методы и системы программирования для разработки и реализации алгоритмов решения прикладных задач	ИОПК-2.1. Обладает навыками объектно-ориентированного программирования для решения прикладных задач в профессиональной деятельности.	Обучающийся сможет: ОР-2.1.1. Уметь применять полученные знания объектно-ориентированного программирования при разработке программ. ОР-2.2.1.Знать языки программирования C# и Python. ОР-2.2.2. Знать библиотеки numpy, pandas, matplotlib для работы с искусственным интеллектом на языке Python. ОР-2.2.3.Уметь работать с online компиляторами как средствами редактирования, отладки, компиляции и выполнения программ. ОР-2.3.1. Знает основные методы численного анализа математических моделей и границы применимости данных методов. ОР-2.3.2. Умеет осуществлять выбор метода численного решения задачи математического моделирования. ОР-2.3.3. Владеет цифровыми инструментами, необходимыми для численного анализа математических моделей при решении конкретной прикладной задачи. ОР-2.4.1. Владеет опытом адаптации существующих математических методов для решения конкретной прикладной задачи.
	ИОПК-2.2. Проявляет навыки использования основных языков программирования, основных методов разработки программ, стандартов оформления программной документации.	
	ИОПК-2.3. Демонстрирует умение отбора среди существующих математических методов, наиболее подходящих для решения конкретной прикладной задачи.	
	ИОПК-2.4. Демонстрирует умение адаптировать существующие математические методы для решения конкретной прикладной задачи.	
ОПК-4. Способен понимать принципы работы современных информационных технологий и использовать их для решения задач профессиональной деятельности.	ИОПК-4.1. Обладает необходимыми знаниями в области информационных технологий, в том числе понимает принципы их работы.	Обучающийся сможет: ОР-4.1.1. Знать принципы работы, используемые в информационных технологиях ОР-4.2.1. Уметь применять знания, полученные в области информационных технологий, при решении задач профессиональной деятельности.
	ИОПК-4.2. Применяет знания, полученные в области информационных технологий, при решении задач профессиональной деятельности.	

ПК-1. Способен осуществлять научно-исследовательские и опытно-конструкторские разработки как по отдельным разделам темы, так и при исследовании самостоятельных тем	ИПК-1.1. Осуществляет проведение работ по обработке и анализу научно-технической информации и результатов исследований	Обучающийся сможет: ОР-1.1.1: Знать основные методы научно-практического поиска в задачах интеллектуального анализа данных и других областях с использованием информационных технологий. ОР-1.1.2: Знать существующие методы и подходы к интеллектуальному анализу данных различной природы. ОР-1.1.3: Уметь применять существующие методы интеллектуального анализа данных, обоснованно адаптируя и модифицируя их с учетом особенностей задачи предметной области. ОР-1.1.4: Уметь формулировать научно-практическую задачу, планировать ее решение и выполнить в соответствии с планом.
ПК-2. Способен анализировать и оценивать риски, разрабатывать отдельные функциональные направления управления рисками.	ИПК-2.2. Собирает и обрабатывает аналитическую информацию для анализа и оценки рисков.	Обучающийся сможет: ОР-2.2.4. Научиться использовать библиотеки для работы с большими данными и искусственным интеллектом. ОР-2.2.5. Применять на практике структуры данных для хранения и обработки данных.

3. Структура и содержание дисциплины

3.1 Структура и трудоёмкость видов учебной работы по дисциплине

Общая трудоёмкость дисциплины составляет 2 зачётные единицы, 72 часа.

Таблица 2.

Вид учебной работы	Трудоемкость в академических часах	
	8 семестр	всего
Общая трудоемкость	**72**	**72**
Контактная работа:	**33,85**	**33,85**
Лекции (Л):	16	16
Практики (ПЗ)		
Лабораторные работы (ЛР)	16	16
Семинары (СЗ)		
Групповые консультации		
Индивидуальные консультации	1,6	1,6
Промежуточная аттестация	0,25	0,25
Самостоятельная работа обучающегося:	**38,15**	**38,15**
- выполнение контрольных заданий	8	8
- изучение учебного материала	14	14
- подготовка к лабораторным занятиям	14	14
- подготовка к рубежному контролю по теме/разделу	2,15	2,15
Вид промежуточной аттестации (зачет, зачет с оценкой, экзамен)	**Зачет**	**Зачет**

3.2. Содержание и трудоемкость разделов дисциплины

Таблица 3.

Код занятия	Наименование разделов и тем и их содержание	Вид учебной работы, занятий, контроля	Семестр	Часы в электронной форме	Всего (час.)	Литература	Коды результатов обучения
	Раздел 1. Основные проблемы построения систем		**5**		**17**	**1**	ОР-2.2.1 ОР-2.2.2. ОР-2.3.1. ОР-4.1.1. ОР-4.2.1. ОР-4.1.1. ОР-1.1.1: ОР-1.1.2: ОР-2.2.4. ОР-2.2.5.
1.1.	Актуальность, базовая терминология и тенденции развития. Основные задачи, этапы и классификация методов анализа данных.	Лекции	5		2		
1.2.	Предварительная обработка данных. Классификация.	Лекции	5		2		
1.3.	Регрессия. Ассоциация, последовательная ассоциация, аномалии и визуализация.	Лаб. работы	5		2		
1.4.	Высокопроизводительная обработка данных. Программные среды для интеллектуального анализа данных.	Лаб. работы	5		2		
1.5	Изучение учебного материала, подготовка к лабораторным работам, подготовка к рубежному контролю по дисциплине	СРС	5		9		
	Текущий контроль успеваемости		5				
	Раздел 2. Предварительная обработка данных. Классификация.		**5**		**17**	**1, 2, 3, 4, 5**	ОР-2.1.1. ОР-2.2.1 ОР-2.2.2. ОР-2.2.3 ОР-2.3.1. ОР-2.3.2. ОР-2.3.3. ОР-2.4.1. ОР-4.1.1. ОР-4.2.1. ОР-4.1.1. ОР-4.2.1. ОР-1.1.1: ОР-1.1.2: ОР-1.1.3: ОР-1.1.4 ОР-2.2.4. ОР-2.2.5.
2.1	Основные методы и предварительная обработка данных.	Лекции	5		2		
2.2	Оптимизация признакового пространства без трансформации пространства признаков.	Лекции	5		2		
2.3	Контролируемая непараметрическая нейросетевая классификация.	Лаб. работы	5		2		
2.4	Классификация по методу машины опорных векторов. Деревья решений.	Лаб. работы	5		2		
2.5	Изучение учебного материала, подготовка к лабораторным работам, подготовка к	СРС	5		9		

	рубежному контролю по дисциплине						
	Текущий контроль успеваемости		5				
	Раздел 3. Регрессия. Ассоциация, последовательная ассоциация, аномалии и визуализация		**5**		**17**	**1, 2, 3, 4, 5**	ОР-2.1.1. ОР-2.2.1 ОР-2.2.2. ОР-2.2.3 ОР-2.3.1. ОР-2.3.2. ОР-2.3.3. ОР-2.4.1. ОР-4.1.1. ОР-4.2.1. ОР-4.1.1. ОР-4.2.1. ОР-1.1.1: ОР-1.1.2: ОР-1.1.3: ОР-1.1.4 ОР-2.2.4. ОР-2.2.5.
3.1	Понятие регрессии. Основные этапы регрессионного анализа.	Лекции	5		2		
3.2	Описание алгоритма ассоциации.	Лекции	5		2		
3.3	Алгоритмы семейства «Априори». Алгоритм GSP.	Лаб. работы	5		2		
3.4	Обнаружение аномалий и методы визуализации.	Лаб. работы	5		2		
3.5	Изучение учебного материала, подготовка к лабораторным работам, подготовка к рубежному контролю по дисциплине	СРС	5		9		
	Текущий контроль успеваемости		5				
	Раздел 4. Высокопроизводительная обработка данных. Программные среды для интеллектуального анализа данных		**5**		**17**	**1, 2, 3, 4, 5**	ОР-2.1.1. ОР-2.2.1 ОР-2.2.2. ОР-2.2.3 ОР-2.3.1. ОР-2.3.2. ОР-2.3.3. ОР-2.4.1. ОР-4.1.1. ОР-4.2.1. ОР-4.1.1. ОР-4.2.1. ОР-1.1.1: ОР-1.1.2: ОР-1.1.3: ОР-1.1.4 ОР-2.2.4. ОР-2.2.5.
4.1	Принципы организации высокопроизводительных вычислений. SMP-системы.	Лекции	5		2		
4.2	Модели параллельных вычислений MPMD, SPMD.	Лекции	5		2		
4.3	Платформа программирования и выполнения распределённых вычислений Hadoop MapReduce, Mahout, Cassandra, Spark. Нереляционные базы данных HBase и язык NoSQL. Среды и языки программирования Python, R.1	Лаб. работы	5		4		
4.5	Изучение учебного материала, подготовка к лабораторным работам, подготовка к рубежному контролю по дисциплине	СРС	5		9		
	Консультации в период теоретического обучения и промежуточной аттестации	К	**5**		**1,6**		
	Подготовка к промежуточной аттестации в форме зачета	СРС	**5**		**2,15**		

	Прохождение промежуточной аттестации в форме зачета	3	**5**		**0,25**		

4. Образовательные технологии, учебно-методическое и информационное обеспечение для освоения дисциплины

Каждый студент реализует индивидуальный или групповой проект как последовательность лабораторных работ. Темы проектов имеют следующий шаблон:

1. Реализовать алгоритм анализа данных.
2. Предложить и реализовать технологии повышения производительности вычислений, выполняемых алгоритмом.

Самостоятельная работа студентов по предмету организуется в следующих формах:

1) самостоятельное изучение основного теоретического материала, ознакомление с дополнительной литературой, Интернет-ресурсами;
2) выполнение индивидуальных проектов, решение профессиональных задач из реальной предметной области.

Темы для изучения	**Формы выполнения заданий**	**Количество часов**
Актуальность, базовая терминология и тенденции развития. Основные задачи, этапы и классификация методов анализа данных	Самостоятельное изучение темы, предложенной преподавателем. Самостоятельное выполнение лабораторной работы №1	8,15
Предварительная обработка данных. Классификация.	Самостоятельное изучение темы, предложенной преподавателем. Самостоятельное выполнение лабораторной работы №2	10
Регрессия. Ассоциация, последовательная ассоциация, аномалии и визуализация	Самостоятельное изучение темы, предложенной преподавателем. Самостоятельное выполнение лабораторной работы №3	10
Высокопроизводительная обработка данных. Программные среды для интеллектуального анализа данных	Самостоятельное изучение темы, предложенной преподавателем. Самостоятельное выполнение лабораторной работы №4	10
Итого		**38,15**

В качестве учебно-методического обеспечения самостоятельной работы используется основная и дополнительная литература по предмету, Интернет-ресурсы, материал лекций, указания, выданные преподавателем при проведении лабораторных работ.

Типовые задания или иные материалы, необходимые для оценки результатов обучения, характеризующих этапы формирования компетенций, и методические материалы, определяющие процедуры оценивания результатов обучения, приведены в Приложении 1 к рабочей программе «Фонд оценочных средств».

4.1. Рекомендуемая литература и учебно-методическое обеспечение

№ п/п	Авторы / составители	Заглавие	Издательство	Год издания, количество страниц
Основная литература				
1.	Замятин А.В.	А.В. Введение в интеллектуальный анализ данных	Издательский Дом государственного университета	2016 г., 194с.

2.	Mohamed Medhat Gaber, Frederic Stahl, João Bártolo Gomes.	Pocket Data Mining electronic resource : Big Data on Small Devices	Springer International Publishing : : Imprint: Springer	2014 г., 177 с.
3.	Max Bramer	Principles of Data Mining electronic resource	Springer London : Imprint: Springer	2013 г., 339 с.
Дополнительная литература				
4.	Миркин Б. Г.	Введение в анализ данных: учебник и практикум для бакалавриата и магистратуры : [для студентов вузов, обучающихся по инженерно-техническим, естественно-научным и экономическим направлениям и специальностям]	"Высшая школа экономики" Национальный исследовательский университет. – Москва: Юрайт	2015 г., 173 с.
5.	Кулаичев А.П.	Методы и средства комплексного анализа данных: учебное пособие.	Москва: Форум	2014 г., 511 с.

4.2. Базы данных и информационно-справочные системы, в том числе зарубежные

1. Data Mining for Service electronic. Berlin, Heidelberg, Imprint: Springer, Springer eBooks VIII, 291 p. 2014 (edited by Katsutoshi Yada) [Электронный ресурс]. – Режим доступа: http://dx.doi.org/10.1007/978-3-642-45252-9

2. Data Mining for Geoinformatics electronic resource : Methods and Applications / /edited by Guido Cervone, Jessica Lin, Nigel Waters. New York, NY : : Springer New York : : Imprint: Springer, , 2014, 166 p. [Электронный ресурс]. – Режим доступа: http://dx.doi.org/10.1007/978-1-4614-7669-6

4.3. Перечень лицензионного и программного обеспечения

Средства и среды программирования C, C++, C#, Python, R-Studio, Rapid Miner, MS Azure.

4.4. Оборудование и технические средства обучения

Для материально-технического обеспечения дисциплины требуется наличие компьютерной техники с установленным соответствующим программным обеспечением и другого оборудования, поддерживающего проведение презентаций, построение проектной документации, выход в сеть Интернет.

5. Методические указания обучающимся по освоению дисциплины

Для успешного освоения дисциплины студенты должны посещать занятия, прорабатывать указанные материалы для самостоятельной работы студентов, выполнять лабораторные работы.

6. Преподавательский состав, реализующий дисциплину

Замятин Александр Владимирович, д-р техн. наук, профессор, заведующий кафедрой теоретических основ информатики ТГУ.

7. Язык преподавания – русский язык.