

МИНОБРНАУКИ РОССИИ
Ярославский государственный университет им. П.Г. Демидова

Кафедра дифференциальных уравнений

УТВЕРЖДАЮ

Декан математического факультета



Нестеров П.Н.

20 июня 2023 г.

Рабочая программа дисциплины

Введение в анализ big data

Направление подготовки (специальности)
01.04.02 Прикладная математика и информатика

Направленность (профиль)
«Математическое моделирование и численные методы»

Форма обучения очная

Программа рассмотрена
на заседании кафедры
от 19 апреля 2023 г., протокол № 8

Программа одобрена НМК
математического факультета
протокол № 9 от 3 мая 2023 г.

1. Цели освоения дисциплины

Данная дисциплина предполагает изучение моделей машинного обучения, математических основ машинного обучения и анализа данных.

Целью освоения дисциплины "Введение в анализ big data" является формирование у студентов навыков, соответствующих видам профессиональной деятельности, необходимых для решения следующих профессиональных задач:

- разработка и применение современных высокопроизводительных вычислительных технологий, применение современных суперкомпьютеров в проводимых исследованиях;
- разработка архитектуры, алгоритмических и программных решений системного и прикладного программного обеспечения;
- развитие и использование математических и информационных инструментальных средств, автоматизированных систем в научной и практической деятельности;

2. Место дисциплины в структуре образовательной программы

Изучение дисциплины продолжает курс информатики старших классов школьной программы и начальных курсов вуза. В ходе программы закрепляются полученные знания изученных ранее курсов «Теория вероятности», «Программирование на языке Python». Полученные знания в данном курсе дают очень важные, базисные навыки, в дальнейшем будут использоваться для написания курсовых и дипломных работ и развития программистских навыков обучающихся.

3. Планируемые результаты обучения по дисциплине, соотнесенные с планируемыми результатами освоения образовательной программы

Процесс изучения дисциплины направлен на формирование следующих компетенций в соответствии с ФГОС ВО, ООП ВО и приобретения следующих знаний, умений, навыков и (или) опыта деятельности:

Формируемая компетенция (код и формулировка)	Индикатор достижения компетенции (код и формулировка)	Перечень планируемых результатов обучения
Общепрофессиональные компетенции		

<p>ПК-2 (Способен разрабатывать и анализировать концептуальные и теоретические модели решаемых научных проблем и задач)</p>	<p>ПК-2.1 (Обладает устойчивыми знаниями в теоретических вопросах, связанных с профессиональной деятельностью)</p> <p>ПК-2.2 (Имеет опыт разработки теоретических моделей решаемых задач)</p> <p>ПК-2.3 (Имеет представление о концептуальных моделях в области решаемых научных проблем и задач)</p>	<p>Знать:</p> <ul style="list-style-type: none"> • методы решения задач обработки и анализа больших данных, возможности высокопроизводительных вычислительных систем, технологии распределенных вычислений, методы и модели Data Mining; <p>Уметь:</p> <ul style="list-style-type: none"> • разрабатывать и анализировать концептуальные и теоретические модели прикладных задач анализа больших данных; • использовать и применять углубленные знания в области обработки и анализа больших данных; • оценивать время и необходимые аппаратные ресурсы для решения задач анализа и обработки данных; • создавать алгоритмы анализа и обработки большого объема данных с применением моделей Data Mining; <p>Владеть:</p> <ul style="list-style-type: none"> • навыками применения программных систем, предназначенных для анализа больших данных;
--	--	---

4. Объем, структура и содержание дисциплины

Общая трудоемкость дисциплины составляет 2 зачетных единиц, 72 акад. часов.

№ п/п	Темы (разделы) дисциплины, их содержание	Семестр	Виды учебных занятий, включая самостоятельную работу студентов, и их трудоемкость (в академических часах)						Формы текущего контроля успеваемости Форма промежуточной аттестации (по семестрам) Формы ЭО и ДОТ (при наличии)
			Контактная работа					самостоятельная работа	
			лекции	практические	лабораторные	консультации	аттестационные испытания		

1	Большие данные (Big Data): современные подходы к обработке и хранению. Проблема множественного сравнения данных. Процесс анализа. Общая схема анализа. Извлечение и визуализация данных. Этапы моделирования. Процесс построения моделей. Формы представления данных, типы и виды данных. Представления наборов данных.	3	2	2				4	
	<i>в том числе с ЭО и ДОТ</i>							2	
2	Технологии KDD и Data Mining. Подготовка данных к анализу. Методика извлечения знаний. Data Mining. Мультидисциплинарный характер Data Mining. Причины распространения KDD и Data Mining. Актуальность технологий Data Mining как средств обработки больших объемов информации.	3	2	2				4	
	<i>в том числе с ЭО и ДОТ</i>							2	
3	Программное обеспечение в области анализа данных. Аналитические платформы: классификация и особенности применения. Языки визуального моделирования.	3	2	2				4	
	<i>в том числе с ЭО и ДОТ</i>							2	
4	Ассоциативные правила. Аффинитивный анализ, предметный набор. Поддержка и достоверность ассоциативного правила. Значимость ассоциативных правил, лифт и левередж. Поиск ассоциативных правил. Частые предметные наборы и их обнаружение. Алгоритм генерации ассоциативных правил. Иерархические ассоциативные правила. Методы поиска иерархических ассоциативных правил.	3	2	2				4	
	<i>в том числе с ЭО и ДОТ</i>							2	
5	Определение кластеризации. Постановка задачи кластеризации. Цели кластеризации в Data Mining. Примеры кластеризации в различных областях. Виды метрик. Шаги алгоритма. Меры расстояний. Пример работы алгоритма k-means.	3	2	2				4	

	Проблемы алгоритмов кластеризации.								
	<i>в том числе с ЭО и ДОТ</i>							2	
6	Применение классификации и регрессии. Обзор методов классификации и регрессии. Статистические методы. Методы, основанные на обучении, разнообразие подходов.	3	2	2				4	
	<i>в том числе с ЭО и ДОТ</i>							2	
7	Основные понятия теории нейронных сетей. Основные парадигмы нейронных сетей. Многослойный персептрон: класс решаемых задач, архитектура.	3	2	2				4	
	<i>в том числе с ЭО и ДОТ</i>							2	
8	Определение дерева решений. Причины популярности и условия применимости. Структура дерева решений. Выбор атрибута разбиения в узле. Алгоритм ID3, критерий выбора атрибута разбиения ID3, пример работы алгоритма. Проблема переобучения.	3	2	2				4	
	<i>в том числе с ЭО и ДОТ</i>							2	
						6	0,3	6,3	Зачет
	<i>в том числе с ЭО и ДОТ</i>							2	
	ИТОГО	3	16	16		4	0,3	36,3	
	<i>в том числе с ЭО и ДОТ</i>								

5. Образовательные технологии, в том числе технологии электронного обучения и дистанционные образовательные технологии, используемые при осуществлении образовательного процесса по дисциплине

- лекции (обзорные, обобщающие);
- практические и лабораторные занятия (семинарские занятия, компьютерный практикум);

При реализации различных видов учебной работы используются образовательные технологии, направленные на развитие у студентов творческих способностей и самостоятельности. К ним относятся активные формы проведения занятий, методы проблемного обучения, исследовательские методы и групповые дискуссии.

Дисциплина «Введение в анализ big data» предполагает чтение научной литературы отечественных и зарубежных авторов.

В процессе обучения используются следующие технологии электронного обучения и дистанционные образовательные технологии:

Электронный учебный курс «Введение в анализ big data» в LMS Электронный университет Moodle ЯрГУ, в котором:

- представлены задания для самостоятельной работы обучающихся при подготовке к экзамену;

- осуществляется проведение отдельных мероприятий текущего контроля успеваемости студентов;
- презентации лекций по темам дисциплины;
- представлены правила прохождения промежуточной аттестации по дисциплине;
- посредством форума осуществляется синхронное и (или) асинхронное взаимодействие между обучающимися и преподавателем в рамках изучения дисциплины.

6. Перечень лицензионного и (или) свободно распространяемого программного обеспечения, используемого при осуществлении образовательного процесса по дисциплине

В процессе осуществления образовательного процесса используются:

- для формирования текстов материалов для промежуточной и текущей аттестации – программы Microsoft Office, издательская система LaTeX;
- для поиска учебной литературы библиотеки ЯрГУ – Автоматизированная библиотечная информационная система «БУКИ-NEXT» (АБИС «Буки-Next»).

7. Перечень современных профессиональных баз данных и информационных справочных систем, используемых при осуществлении образовательного процесса по дисциплине (при необходимости)

8. Перечень основной и дополнительной учебной литературы, ресурсов информационно-телекоммуникационной сети «Интернет» (при необходимости), рекомендуемых для освоения дисциплины

а) основная литература

1. Жуковский О.И. Информационные технологии и анализ данных: учебное пособие 5-е изд., Томск: Эль Контент, 2014г., 130с.
2. Боровков А.А. Математическая статистика 4-е изд., СПб: Лань, 2010г., 704с.

б) дополнительная литература

3. Кормен Томас. Алгоритмы. Построение и анализ 2-е изд., Диалектика, 2019г., 1328с.

9. Материально-техническая база, необходимая для осуществления образовательного процесса по дисциплине

Материально-техническая база, необходимая для осуществления образовательного процесса по дисциплине включает в свой состав специальные помещения:

- учебные аудитории для проведения практических занятий (семинаров);
- учебные аудитории для проведения групповых и индивидуальных консультаций;
- учебные аудитории для проведения текущего контроля и промежуточной аттестации;
- помещения для самостоятельной работы;

- помещения для хранения и профилактического обслуживания технических средств обучения.

Специальные помещения укомплектованы средствами обучения, служащими для представления учебной информации большой аудитории.

Помещения для самостоятельной работы обучающихся оснащены компьютерной техникой с возможностью подключения к сети «Интернет» и обеспечением доступа к электронной информационно-образовательной среде ЯрГУ.

Автор(ы):

Доцент кафедры дифференциальных уравнений
должность, ученая степень

подпись

М.В. Смирнова
И.О. Фамилия

Приложение № 1 к рабочей программе дисциплины
«Введение в анализ big data»
наименование дисциплины

Фонд оценочных средств
для проведения текущего контроля успеваемости
и промежуточной аттестации студентов
по дисциплине

1. Типовые контрольные задания и иные материалы, используемые в процессе текущего контроля успеваемости
 1. Понятие Большие данные. Роль цифровой информации в 21 веке.
 2. Проблемы анализа и обработки большого объема данных.
 3. Базовые принципы обработки больших данных.
 4. Определение модели. Свойства модели.
 5. Аналитический подход к моделированию.
 6. Информационный подход к моделированию.
 7. Лица, участвующие в информационном моделировании.
 8. Общая схема анализа.
 9. Определение тиражирования знаний. Процесс построения модели.
 10. Технологии обработки больших данных: NoSQL,
 11. Технологии обработки больших данных: MapReduce,
 12. Технологии обработки больших данных: Hadoop, R.
 13. Методика извлечения знаний Knowledge Discovery in Databases (KDD). Этапы KDD.
 14. Data Mining. Постановка основных задач.
 15. Машинное обучение.
 15. Бизнес-решения с помощью алгоритмов Data Mining.
 16. Классификация ПО в области Data Mining и KDD.
 17. Типовая схема системы на базе аналитической платформы.
 18. Понятие ассоциативного правила и транзакции.
 19. Определение поддержки и достоверности.
 20. Определение значимости и полезности ассоциативных правил, показатели их характеризующие.
 21. Формальная постановка задачи кластеризации.
 22. Цели кластеризации.
 23. Основные шаги алгоритма k-means. Условие остановки алгоритма k-means. Преимущества и недостатки алгоритма k-means.
 24. Кластеризация с помощью самоорганизующейся карты Кохонена
 25. Этапы проведения классификации.
 26. Обзор методов классификации и регрессии.
 27. Задачи линейной и логистической регрессии.
 28. Определение дерева решений. Структура дерева решений. Выбор атрибута разбиения в узле.
2. Список вопросов и (или) заданий для проведения промежуточной аттестации
 1. Понятие Большие данные. Роль цифровой информации в 21 веке.
 2. Проблемы анализа и обработки большого объема данных.
 3. Базовые принципы обработки больших данных.
 4. Определение модели. Свойства модели.
 5. Аналитический подход к моделированию.
 6. Информационный подход к моделированию.

7. Лица, участвующие в информационном моделировании.
8. Общая схема анализа.
9. Определение тиражирования знаний. Процесс построения модели.
10. Технологии обработки больших данных: NoSQL
11. Технологии обработки больших данных: MapReduce
12. Технологии обработки больших данных: Hadoop, R.
13. Методика извлечения знаний Knowledge Discovery in Databases (KDD). Этапы KDD.
14. Data Mining. Постановка основных задач

Приложение № 2 к рабочей программе дисциплины
«Введение в анализ big data»
наименование дисциплины

Методические указания для студентов по освоению дисциплины

Целью процедуры оценивания является определение степени овладения студентом ожидаемыми результатами обучения (знаниями, умениями, навыками и (или) опытом деятельности).

Процедура оценивания степени овладения студентом ожидаемыми результатами обучения осуществляется с помощью методических материалов, представленных в разделе «Типовые контрольные задания или иные материалы, необходимые для оценки знаний, умений, навыков и (или) опыта деятельности, характеризующих этапы формирования компетенций», или других заданий аналогичного уровня сложности