

МИНОБРНАУКИ РОССИИ
федеральное государственное бюджетное образовательное учреждение
высшего образования
«Нижегородский государственный технический университет
им. Р.Е. Алексеева» (НГТУ)

Дзержинский политехнический институт (филиал)

УТВЕРЖДАЮ:
Директор института
_____ А.М. Петровский
“08” ____ июня ____ 2023 г.

РАБОЧАЯ ПРОГРАММА ДИСЦИПЛИНЫ

Б1.В.ОД.8 «Большие данные»

(индекс и наименование дисциплины по учебному плану)

для подготовки бакалавров

Направление подготовки: **01.03.04 Прикладная математика**

Направленность: Математические и компьютерные методы для современных технологий

Форма обучения: очная

Год начала подготовки 2023

Выпускающая кафедра Автоматизация, энергетика, математика и информационные системы

Кафедра-разработчик Автоматизация, энергетика, математика и информационные системы

Объем дисциплины 108 / 3

Промежуточная аттестация зачет с оценкой

Разработчик: Наумова Е.Г., к.т.н.

Дзержинск 2023

Рабочая программа дисциплины: разработана в соответствии с Федеральным государственным образовательным стандартом высшего образования (ФГОС ВО 3++) по направлению подготовки 01.03.04 Прикладная математика, утвержденного приказом МИНОБРНАУКИ РФ от 10 января 2018 года № 11 на основании учебного плана, принятого УС ДПИ НГТУ

на основании учебного плана принятого УС ДПИ НГТУ
протокол от 02.06.2023 № 9

Рабочая программа одобрена на заседании кафедры-разработчика РПД Автоматизация, энергетика, математика и информационные системы
протокол от 08.06.2023 № 8

Заведующий кафедрой АЭМИС, к.т.н., доцент

Л.Ю. Вадова

(подпись)

(расшифровка подписи)

СОГЛАСОВАНО:

Заведующий выпускающей кафедрой «Автоматизация, энергетика, математика и информационные системы»,
к.т.н., доцент

Л.Ю. Вадова

(подпись)

(расшифровка подписи)

Начальник ОУМБО

И.В. Старикова

(подпись)

(расшифровка подписи)

Рабочая программа зарегистрирована в ОУМБО: 01.03.04 - 41

СОДЕРЖАНИЕ

1. Цели и задачи освоения дисциплины	4
2. Место дисциплины в структуре образовательной программы.....	5
3. Компетенции обучающегося, формируемые в результате освоения дисциплины.....	5
4. Структура и содержание дисциплины.....	6
5. Текущий контроль успеваемости и промежуточная аттестация по итогам освоения дисциплины.....	11
6. Учебно-методическое обеспечение дисциплины.....	15
7. Информационное обеспечение дисциплины.....	16
8. Образовательные ресурсы для инвалидов и лиц с ОВЗ.....	17
9. Материально-техническое обеспечение, необходимое для осуществления образовательного процесса по дисциплине.....	18
10. Методические рекомендации обучающимся по освоению дисциплины.....	19
11. Оценочные средства для контроля освоения дисциплины.....	21

1. ЦЕЛИ И ЗАДАЧИ ОСВОЕНИЯ ДИСЦИПЛИНЫ (МОДУЛЯ)

1.1. Цель освоения дисциплины: изучение теоретических основ анализа больших данных; изучение математических методов и подходов, используемых в программных системах обработки и анализа больших данных.

1.2. Задачи освоения дисциплины (модуля):

- изучение методов анализа и хранения больших объемов данных,
- изучение этапов жизненного цикла обработки больших данных,
- получение навыков применения технологий обработки данных (в том числе big data) и машинного обучения к решению прикладных задач;
- знакомство и применение элементов анализа данных и интерпретации результатов.

2. МЕСТО ДИСЦИПЛИНЫ В СТРУКТУРЕ ОБРАЗОВАТЕЛЬНОЙ ПРОГРАММЫ

Учебная дисциплина «Большие данные» включена в перечень дисциплин вариативной части (формируемой участниками образовательных отношений), определяющий направленность ОП. Дисциплина реализуется в соответствии с требованиями ФГОС, ОП ВО и УП.

Дисциплина «Большие данные» базируется на следующих дисциплинах: Линейная алгебра, Дискретная математика, Основы машинного обучения, Методы оптимизации, Методы искусственного интеллекта.

Дисциплина «Большие данные» является основополагающей для выполнения и защиты выпускной квалификационной работы.

Рабочая программа дисциплины «Большие данные» для инвалидов и лиц с ограниченными возможностями здоровья разрабатывается индивидуально с учетом особенностей психофизического развития, индивидуальных возможностей и состояния здоровья таких обучающихся, по их личному заявлению.

3. КОМПЕТЕНЦИИ ОБУЧАЮЩЕГОСЯ, ФОРМИРУЕМЫЕ В РЕЗУЛЬТАТЕ ОСВОЕНИЯ ДИСЦИПЛИНЫ (МОДУЛЯ)

Таблица 3.1

Формирование компетенций ПКС-2 дисциплинами
для обучающихся очной формы обучения

Наименование дисциплин, формирующих компетенцию совместно	Семестры формирования компетенций дисциплинами							
	1	2	3	4	5	6	7	8
Код компетенции ПКС-2.								
Ознакомительная практика				х				
Методы искусственного интеллекта						х		
Проектно-технологическая практика						х		
Большие данные							х	
Основы программирования в 1С: Предприятие								х
Преддипломная практика								х
Выполнение и защита выпускной квалификационной работы								х

ПЕРЕЧЕНЬ ПЛАНИРУЕМЫХ РЕЗУЛЬТАТОВ ОБУЧЕНИЯ ПО ДИСЦИПЛИНЕ, СООТНЕСЕННЫХ С ПЛАНИРУЕМЫМИ РЕЗУЛЬТАТАМИ ОСВОЕНИЯ ОП

Таблица 3.2

Перечень планируемых результатов обучения по дисциплине, соотнесенных с планируемыми результатами освоения

Код и наименование компетенции	Код и наименование индикатора достижения компетенции	Планируемые результаты обучения по дисциплине			Оценочные средства	
					Текущего контроля	Промежуточной аттестации
ПКС-2. Способен обрабатывать, анализировать данные и делать выводы, используя соответствующий математический аппарат и современные прикладные программные средства.	ИПКС-2.1- Выбирает и использует компьютерные системы с целью интерпретации внешних данных, извлечения из них полезной информации и использования полученных знаний для достижения конкретных целей	Знать: - основные понятия и терминологию в области технологий обработки и анализа больших данных; - методы и техники анализа больших данных; - OLAP и многомерные базы данных. - основное программное обеспечение для работы с большими наборами данных; - модель обработки информации MapReduce; - компоненты кластерной системы Hadoop; - технологии Data mining.	Уметь: - обрабатывать и управлять большими объемами постоянно обновляющейся информации; - использовать и применять углубленные знания в области обработки и анализа больших наборов данных; - оценивать время и необходимые аппаратные ресурсы для решения задач анализа и обработки данных; - использовать программные инструменты для работы с большими наборами данных;	Владеть: - методологией и методикой прогнозирования. - навыками оптимизации параллельного доступа и обработки больших данных - программными инструментами для работы с большими наборами данных - современными технологиями создания и обслуживания больших данных	Тестирование в системе MOODLE (2 тестирования, в базе каждого тестирования около 100 вопросов), собеседование и отчеты при сдаче лабораторных работ	По результатам накопительного рейтинга или в форме тестирования для обучающихся очной формы обучения

4. СТРУКТУРА И СОДЕРЖАНИЕ ДИСЦИПЛИНЫ

4.1. Распределение трудоёмкости дисциплины по видам работ по семестрам

Общая трудоёмкость дисциплины составляет 3 зач.ед. / 108 часов, распределение часов по видам работ и семестрам представлено в таблицах 4.1.

Формат изучения дисциплины: с использованием элементов электронного обучения

Таблица 4.1

Распределение трудоёмкости дисциплины по видам работ по семестрам для обучающихся очной формы обучения

Вид учебной работы	Всего часов	Семестр
		7
1. Контактная работа обучающихся с преподавателем (по видам учебных занятий) (всего), в том числе:	55	55
1.1. Аудиторные занятия (всего), в том числе:	51	51
- лекции (Л)	17	17
- лабораторные работы (ЛР)	34	34
- практические занятия (ПЗ)		
- практикумы (П)		
1.2. Внеаудиторные занятия (всего), в том числе:	4	4
- групповые консультации по дисциплине	4	4
- групповые консультации по промежуточной аттестации (экзамен)		
- индивидуальная работа преподавателя с обучающимся: - по проектированию: проект (работа) - по выполнению РГР - по выполнению КР - по составлению реферата, доклада, эссе		
2. Самостоятельная работа обучающихся (СРС) (всего)	53	53
Вид промежуточной аттестации зачёт с оценкой	Зачёт с оценкой	Зачёт с оценкой
Общая трудоемкость, часы/зачетные единицы	108 / 3	108 / 3

4.2. Содержание дисциплины, структурированное по темам

Содержание дисциплины, структурированное по темам, приведено в таблицах 4.2.

Таблица 4.3

Содержание дисциплины, структурированное по темам для обучающихся очной формы обучения

Планируемые (контролируемые) результаты освоения: код УК; ОПК; ПК и индикаторы достижения компетенций	Наименование разделов, тем	Виды учебной работы				Вид СРС	Наименование используемых активных и интерактивных образовательных технологий	Реализация в рамках практической подготовки (трудоемкость в часах)	Наименование разработанного электронного курса (трудоемкость в часах)
		Контактная работа			Самостоятельная работа обучающихся (СРС), час				
		Лекции, час	Лабораторные работы, час	Практические занятия, час					
6 семестр									
ПКС-2, ИПКС-2.1	Раздел 1 Введение в дисциплину					Подготовка к лекциям, тестированию, выполнение заданий для самостоятельной работы. 6.1.1: раздел 1 6.1.2: раздел 13			
	Тема 1.1. Основные понятия Big Data	0,5			2		Тестирование в системе MOODLE (Тест 1)		
	Тема 1.2. Жизненный цикл Big Data	0,5			2				
	Тема 1.3. Программное и аппаратное обеспечение	0,5			2				
	Тема 1.4. Область применения, специалисты в области Big Data	0,5			2				
	Итого по разделу 1	2			8				
ПКС-2, ИПКС-2.1	Раздел 2 Предобработка информации					Подготовка к лекциям, тестированию, выполнение заданий для самостоятельной работы. 6.1.3: раздел 3, 4, 5; 6.1.4: раздел 1.8			

Планируемые (контролируемые) результаты освоения: код УК; ОПК; ПК и индикаторы достижения компетенций	Наименование разделов, тем	Виды учебной работы				Вид СРС	Наименование используемых активных и интерактивных образовательных технологий	Реализация в рамках практической подготовки (трудоемкость в часах)	Наименование разработанного электронного курса (трудоемкость в часах)
		Контактная работа							
		Лекции, час	Лабораторные работы, час	Практические занятия, час	Самостоятельная работа обучающихся (СРС), час				
	Тема 2.1. Сбор и хранение информации, песочница	1			2		Тестирование в системе MOODLE (Тест 1)		
	Тема 2.2. Очистка данных и их трансформация	1			3				
	Тема 2.3. Получение статистических показателей	1			3				
	Тема 2.4. Графическое представление данных	1			2				
	Лабораторная работа № 1 Предварительная обработка данных		4		2	Подготовка отчёта по ЛР № 1 и подготовка к собеседованию по отчёту 6.1.3: раздел 3, 4, 5; 6.2.1	Собеседование по отчёту		
	Итого по разделу 2	4	4		12				
ПКС-2, ИПКС-2.1	Раздел 3 Постановка и решение задач Data mining. Стандарты процесса KDD&DM					Подготовка к лекциям, тестированию, выполнение заданий для самостоятельной работы. 6.1.1: разделы 2, 3, 5 6.1.4: раздел 1			
	Тема 3.1. Задачи Data mining	2			3		Тестирование в системе MOODLE (Тест 2)		
	Тема 3.2. Классификация	1			3				
	Тема 3.3. Кластеризация	1			3				
	Тема 3.4. Задачи прогнозирования, регрессия	1			3				
	Тема 3.5. Поиск ассоциативных правил	1			3				
	Тема 3.6. Нейронная сеть	1			3				
	Тема 3.7. Стандарты процесса KDD&DM	2			3				

Планируемые (контролируемые) результаты освоения: код УК; ОПК; ПК и индикаторы достижения компетенций	Наименование разделов, тем	Виды учебной работы				Вид СРС	Наименование используемых активных и интерактивных образовательных технологий	Реализация в рамках практической подготовки (трудоемкость в часах)	Наименование разработанного электронного курса (трудоемкость в часах)
		Контактная работа			Самостоятельная работа обучающихся (СРС), час				
		Лекции, час	Лабораторные работы, час	Практические занятия, час					
	Лабораторная работа № 2 Классификация		8		2	Подготовка отчёта по ЛР № 2 и подготовка к собеседованию по отчёту 6.1.1: разделы 3.2, 3.3 6.1.4: разделы 1.1-1.4 6.2.3	Собеседование по отчёту		
	Лабораторная работа № 3 Кластеризация		8		2	Подготовка отчёта по ЛР № 3 и подготовка к собеседованию по отчёту 6.1.1: разделы 3.2, 3.3 6.1.4: разделы 1.1-1.4 6.2.4	Собеседование по отчёту		
	Лабораторная работа № 4 Поиск ассоциативных правил		6		2	Подготовка отчёта по ЛР № 4 и подготовка к собеседованию по отчёту 6.1.1: раздел 3.5 6.2.5	Собеседование по отчёту		
	Лабораторная работа № 5 Задачи прогнозирования и регрессия		8		2	Подготовка отчёта по ЛР № 5 и подготовка к собеседованию по отчёту 6.1.1: раздел 3.4 6.1.4: раздел 2.3 6.2.2	Собеседование по отчёту		
	Итого по разделу 3	9	30		29				
ПКС-2, ИПКС-2.1	Раздел 4 Технология MapReduce.					Подготовка к лекциям, тестированию, выполнение заданий для самостоятельной работы. 6.1.2: разделы 14, 15			

Планируемые (контролируемые) результаты освоения: код УК; ОПК; ПК и индикаторы достижения компетенций	Наименование разделов, тем	Виды учебной работы				Вид СРС	Наименование используемых активных и интерактивных образовательных технологий	Реализация в рамках практической подготовки (трудоемкость в часах)	Наименование разработанного электронного курса (трудоемкость в часах)
		Контактная работа			Самостоятельная работа обучающихся (СРС), час				
		Лекции, час	Лабораторные работы, час	Практические занятия, час					
	Тема 4.1. Принципы технологии MapReduce	1			2		Тестирование в системе MOODLE (Тест 2)		
	Тема 4.2. Приемы и стратегии разработки MapReduce-приложений	1			2				
	<i>Итого по разделу 4</i>	2			4				
	ИТОГО по дисциплине	17	34		53				

5. ТЕКУЩИЙ КОНТРОЛЬ УСПЕВАЕМОСТИ И ПРОМЕЖУТОЧНАЯ АТТЕСТАЦИЯ ПО ИТОГАМ ОСВОЕНИЯ ДИСЦИПЛИНЫ.

5.1. Типовые контрольные задания или иные материалы, необходимые для оценки знаний, умений и навыков и (или) опыта деятельности

Тесты для текущего и промежуточного контроля знаний обучающихся

Тесты проводятся на электронной платформе Moodle на сайте ДПИ НГТУ по адресу: <http://dpingtu.ru/Moodle>. Примеры типовых тестовых заданий приведены в разделе 11.1.1 настоящей рабочей программы.

Вопросы для подготовки к контрольным мероприятиям и к защите отчётов по лабораторным работам (текущий контроль)

Раздел 1.

1. Основные понятия курса.
2. Основные признаки BigData. Жизненный цикл больших данных
3. Проблемы анализа и обработки большого объема данных.
4. Типы информации. Принципы обработки больших данных
5. Архитектура информационных систем анализа данных
6. Теоремы CAP и ACID
7. Организационные и человеческие факторы при внедрении ИС, работающей с большими данными.
8. Сферы применения BigData
9. Профессии, связанные с BigData
10. Стандарты в области работы с большими данными
11. СУБД для хранения больших данных. NoSQL.
12. Программное и аппаратное обеспечение, используемое для обработки и анализа больших данных

Раздел 2.

1. Источники больших данных
2. Методы сбора информации. Песочница проекта
3. Первичная очистка данных.
4. Выбросы и пропуски в данных
5. Основные понятия статистики. Статистические показатели
6. Корреляционный анализ
7. Визуализация данных на разных этапах анализа больших данных

Раздел 3.

1. Методы Data Mining
2. Основные задачи Data Mining
3. Регрессионный анализ
4. Задача классификации
5. Задача кластеризации. Методы кластеризации
6. Методы поиска ассоциативных правил
7. Определение дерева решений
8. Нейронные сети
9. Методика извлечения данных KDD
10. Этапы KDD

Раздел 4.

1. MapReduce. Состав и назначение.
2. Принципы технологии MapReduce
3. Приемы и стратегии разработки MapReduce-приложений
4. Фильтр Блума

Перечень вопросов, выносимых на промежуточную аттестацию (зачёт с оценкой)

1. Основные понятия курса.
2. Основные признаки BigData. Жизненный цикл больших данных
3. Проблемы анализа и обработки большого объема данных.
4. Типы информации. Принципы обработки больших данных
5. Архитектура информационных систем анализа данных
6. Теоремы CAP и ACID
7. Организационные и человеческие факторы при внедрении ИС, работающей с большими данными.
8. Сферы применения BigData
9. Профессии, связанные с BigData
10. Стандарты в области работы с большими данными
11. СУБД для хранения больших данных. NoSQL.
12. Программное и аппаратное обеспечение, используемое для обработки и анализа больших данных
13. Источники больших данных
14. Методы сбора информации. Песочница проекта
15. Первичная очистка данных.
16. Выбросы и пропуски в данных
17. Основные понятия статистики. Статистические показатели
18. Корреляционный анализ
19. Визуализация данных на разных этапах анализа больших данных
20. Методы Data Mining
21. Основные задачи Data Mining
22. Регрессионный анализ
23. Задача классификации
24. Задача кластеризации. Методы кластеризации
25. Методы поиска ассоциативных правил
26. Определение дерева решений
27. Нейронные сети
28. Методика извлечения данных KDD
29. Этапы KDD
30. MapReduce. Состав и назначение.
31. Принципы технологии MapReduce
32. Приемы и стратегии разработки MapReduce-приложений
33. Фильтр Блума

5.2. Описание показателей и критериев контроля успеваемости, описание шкал оценивания

Для оценки знаний, умений, навыков и формирования компетенции по дисциплине может применяться балльно-рейтинговая система контроля и оценки успеваемости обучающихся очной формы. Основные требования балльно-рейтинговой системы по дисциплине и шкала оценивания приведены в таблицах 5.1 – 5.3.

Таблица 5.1

Требования балльно-рейтинговой системы по дисциплине
для обучающихся очной формы обучения

Вид работ	Количество подвидов работы	Макс. баллы за подвид работы	Штрафные баллы	Макс. количество баллов по виду работ
1. Тестирование	2	20	0* -2 за повтор (3 попытки)	40
2. Выполнение лабораторных работ, в т.ч. на 1 работу	5	6	+1 ÷ +3 – за выполнение доп. заданий -1 – за нарушение срока	36
- выполнение		2	-1 ÷ -3 при наличии ошибок	10
- оформление и защита отчётов		4	-1 ÷ -3 отсутствие ответов на вопросы	20
3. Посещение занятий	51	0,5		26
- лекции	17		+0,5 за ответ на контр. вопрос	9
- лабораторные работы	34			17
Итого**				100

* Если при тестировании верно выполнено менее 55% заданий, то количество баллов за работу приравнивается к «0»

** Освобождение от зачёта с оценкой возможно при условии выполнения всех лабораторных работ, положительных оценок за тестирование, выполнения реферата, посещения не менее 50 % занятий.

Критерии оценивания результата обучения по дисциплине и шкала оценивания

Код и наименование компетенции	Код и наименование индикатора достижения компетенции	Критерии оценивания результатов обучения			
		Оценка «неудовлетворительно» 0-54% от тах рейтинговой оценки контроля	Оценка «удовлетворительно» 55-70% от тах рейтинговой оценки контроля	Оценка «хорошо» 71-85% от тах рейтинговой оценки контроля	Оценка «отлично» 86-100% от тах рейтинговой оценки контроля
<p>ПКС-2. Способен обрабатывать, анализировать данные и делать выводы, используя соответствующий математический аппарат и современные прикладные программные средства.</p>	<p>ИПКС-2.1- Выбирает и использует компьютерные системы с целью интерпретации внешних данных, извлечения из них полезной информации и использования полученных знаний для достижения конкретных целей</p>	<p>Изложение учебного материала бессистемное, неполное, не знает основ анализа больших данных</p>	<p>Фрагментарные, поверхностные знания по методам и техникам анализа больших данных. Изложение полученных знаний неполное, но это не препятствует усвоению последующего материала. Допускаются отдельные существенные ошибки, исправленные с помощью преподавателя. Затруднения при формулировании результатов и их решений</p>	<p>Знает материал на достаточно хорошем уровне; обрабатывать и управлять большими объемами постоянно обновляющейся информации. в рамках постановки целей и выбора оптимальных способов их достижения.</p>	<p>Имеет глубокие знания всего материала структуры дисциплины; освоил новации лекционного курса по сравнению с учебной литературой; изложение полученных знаний полный, системное; допускаются единичные ошибки, самостоятельно исправляемые при собеседовании</p>

Критерии оценивания

Оценка	Критерии оценивания
Высокий уровень «5» (отлично)	оценку « отлично » заслуживает обучающийся, освоивший знания, умения, компетенции и теоретический материал без пробелов; выполнивший все задания, предусмотренные учебным планом на высоком качественном уровне; практические навыки профессионального применения освоенных знаний сформированы.
Средний уровень «4» (хорошо)	оценку « хорошо » заслуживает обучающийся, практически полностью освоивший знания, умения, компетенции и теоретический материал, учебные задания не оценены максимальным числом баллов, в основном сформировал практические навыки.
Пороговый уровень «3» (удовлетворительно)	оценку « удовлетворительно » заслуживает обучающийся, частично с пробелами освоивший знания, умения, компетенции и теоретический материал, многие учебные задания либо не выполнил, либо они оценены числом баллов близким к минимальному, некоторые практические навыки не сформированы.
Минимальный уровень «2» (неудовлетворительно)	оценку « неудовлетворительно » заслуживает обучающийся, не освоивший знания, умения, компетенции и теоретический материал, учебные задания не выполнил, практические навыки не сформированы.

6. УЧЕБНО-МЕТОДИЧЕСКОЕ ОБЕСПЕЧЕНИЕ ДИСЦИПЛИНЫ

6.1. Учебная литература

6.1.1. Афанасьева, Т. В. Введение в проектирование систем интеллектуального анализа данных : учебное пособие / Т. В. Афанасьева. — Ульяновск : УлГТУ, 2017. — 64 с. — ISBN 978-5-9795-1686-8. — Текст : электронный // Лань : электронно-библиотечная система. — URL: <https://e.lanbook.com/book/165064>

6.1.2. Григорьев, Ю. А. Реляционные базы данных и системы NoSQL : учебное пособие / Ю. А. Григорьев, А. Д. Плутенко, О. Ю. Плужникова. — Благовещенск : АмГУ, 2018. — 424 с. — ISBN 978-5-93493-308-2. — Текст : электронный // Лань : электронно-библиотечная система. — URL: <https://e.lanbook.com/book/156492>

6.1.3. Бельчик, Т. А. Основы математической обработки информации с помощью SPSS : учебное пособие / Т. А. Бельчик. — Кемерово : КемГУ, 2013. — 232 с. — ISBN 978-5-8353-1265-8. — Текст : электронный // Лань : электронно-библиотечная система. — URL: <https://e.lanbook.com/book/44312>.

6.1.4. Кабанов, А. Н. Технологии обработки информации : учебное пособие / А. Н. Кабанов, Д. Н. Фоломкин. — Рязань : РГРТУ, 2017. — 48 с. — Текст : электронный // Лань : электронно-библиотечная система. — URL: <https://e.lanbook.com/book/168325>.

6.2. Методические указания, рекомендации и другие материалы к занятиям

6.2.1. Предварительная обработка данных: метод. указания по дисциплине «Большие данные» к выполнению лабораторных работ для обучающихся направлений подготовки 09.03.02 «Информационные системы и технологии», 01.03.04 «Прикладная математика» очной формы обучения и практических работ для обучающихся направления подготовки 09.03.02 «Информационные системы и технологии» заочной формы обучения / ДПИ НГТУ; сост.: Е.Г. Наумова.– Дзержинск, 2020. - 14 с.

6.2.2. Задачи прогнозирования и регрессия: метод. указания по дисциплине «Большие данные» к выполнению лабораторных работ для обучающихся направлений подготовки 09.03.02 «Информационные системы и технологии», 01.03.04 «Прикладная математика» очной формы обучения и практических работ для обучающихся направления подготовки 09.03.02 «Информационные системы и технологии» заочной формы обучения / ДПИ НГТУ; сост.: Е.Г. Наумова.– Дзержинск, 2020. - 14 с.

6.2.3. Классификация: метод. указания по дисциплине «Большие данные» к выполнению лабораторных работ для обучающихся направлений подготовки 09.03.02 «Информационные системы и технологии», 01.03.04 «Прикладная математика» очной формы обучения и практических работ для обучающихся направления подготовки 09.03.02 «Информационные системы и технологии» заочной формы обучения / ДПИ НГТУ; сост.: Е.Г. Наумова.– Дзержинск, 2020. - 14 с.

6.2.4. Кластеризация: метод. указания по дисциплине «Большие данные» к выполнению лабораторных работ для обучающихся направлений подготовки 09.03.02 «Информационные системы и технологии», 01.03.04 «Прикладная математика» очной формы обучения и практических работ для обучающихся направления подготовки 09.03.02 «Информационные системы и технологии» заочной формы обучения / ДПИ НГТУ; сост.: Е.Г. Наумова.– Дзержинск, 2020. - 14 с.

6.2.5. Поиск ассоциативных правил: метод. указания по дисциплине «Большие данные» к выполнению лабораторных работ для обучающихся направлений подготовки 09.03.02 «Информационные системы и технологии», 01.03.04 «Прикладная математика» очной формы обучения и практических работ для обучающихся направления подготовки 09.03.02 «Информационные системы и технологии» заочной формы обучения / ДПИ НГТУ; сост.: Е.Г. Наумова.– Дзержинск, 2020. - 14 с.

7. ИНФОРМАЦИОННОЕ ОБЕСПЕЧЕНИЕ ДИСЦИПЛИНЫ

Учебный процесс по дисциплине обеспечен необходимым комплектом лицензионного и свободно распространяемого программного обеспечения, в том числе отечественного производства (состав по дисциплине определен в настоящей РПД и подлежит обновлению при необходимости).

Дисциплина, относится к группе дисциплин, в рамках которых предполагается использование информационных технологий как вспомогательного инструмента для выполнения задач, таких как:

- оформление отчетов по лабораторному занятию;
- использование электронной образовательной среды института;
- использование специализированного программного обеспечения;
- организация взаимодействия с обучающимися посредством электронной почты;
- использование видеоконференцсвязи;
- компьютерное тестирование.

7.1. Перечень информационных справочных систем

Таблица 7.1

Перечень электронных библиотечных систем

№	Наименование ЭБС	Ссылка к ЭБС
1	Консультант студента	http://www.studentlibrary.ru/
2	Лань	https://e.lanbook.com/

7.2. Перечень лицензионного и свободно распространяемого программного обеспечения, в том числе отечественного производства необходимого для освоения дисциплины

Таблица 7.2

Программное обеспечение

№ п/п	Программное обеспечение, используемое в университете на договорной основе	Программное обеспечение свободного распространения
1	Microsoft Windows 10 (подписка MSDN 700593597, подписка DreamSparkPremium, 19.06.19)	Adobe Acrobat Reader https://acrobat.adobe.com/ru/ru/acrobat/pdf-reader.html
2	Microsoft VISUAL STUDIO 2008 (подписка MSDN 700593597, подписка DreamSparkPremium, 19.06.19)	Visual Studio Code https://code.visualstudio.com/download
3	Microsoft office 2010 (Лицензия № 49487295 от 19.12.2011)	OpenOffice https://www.openoffice.org/ru/
4	Консультант Плюс	Python https://www.python.org

7.3. Перечень современных профессиональных баз данных и информационных справочных систем

В таблице 7.3 указан перечень профессиональных баз данных и информационных справочных систем, к которым обеспечен доступ (удаленный доступ). Данный перечень подлежит обновлению в соответствии с требованиями ФГОС ВО.

Таблица 7.3

Перечень современных профессиональных баз данных и информационных справочных систем

№ п/п	Наименование профессиональной базы данных, информационно-справочной системы	Доступ к ресурсу (удаленный доступ с указанием ссылки/доступ из локальной сети университета)
1	База данных стандартов и регламентов РОССТАНДАРТ	https://www.gost.ru/portal/gost_//home/standarts
2	Перечень профессиональных баз данных и информационных справочных систем	https://cyberpedia.su/21x47c0.html
3	Инструменты и веб-ресурсы для веб-разработки – 100+	https://techblog.sdstudio.top/blog/instrumenty-i-veb-resursy-dlia-veb-razrabotki-100-plus
4	Справочная правовая система «КонсультантПлюс»	доступ из локальной сети

8. ОБРАЗОВАТЕЛЬНЫЕ РЕСУРСЫ ДЛЯ ИНВАЛИДОВ И ЛИЦ С ОВЗ

В таблице 8.1 указан перечень образовательных ресурсов, имеющих формы, адаптированные к ограничениям их здоровья, а также сведения о наличии специальных технических средств обучения коллективного и индивидуального пользования.

Таблица 8.1

Образовательные ресурсы для инвалидов и лиц с ОВЗ

№	Перечень образовательных ресурсов, приспособленных для использования инвалидами и лицами с ОВЗ	Сведения о наличии специальных технических средств обучения коллективного и индивидуального пользования
1	ЭБС «Консультант студента»	озвучка книг и увеличение шрифта
2	ЭБС «Лань»	специальное мобильное приложение - синтезатор речи, который воспроизводит тексты книг и меню навигации
3	ЭБС «Юрайт»	версия для слабовидящих

Согласно Федеральному Закону об образовании 273-ФЗ от 29.12.2012 г. ст. 79, п.8 "Профессиональное обучение и профессиональное образование обучающихся с ограниченными возможностями здоровья осуществляются на основе образовательных программ, адаптированных при необходимости для обучения указанных обучающихся". АОП разработа-

тывается по каждой направленности при наличии заявлений от обучающихся, являющихся инвалидами или лицами с ОВЗ и изъявивших желание об обучении по данному типу образовательных программ.

9. МАТЕРИАЛЬНО-ТЕХНИЧЕСКОЕ ОБЕСПЕЧЕНИЕ, НЕОБХОДИМОЕ ДЛЯ ОСУЩЕСТВЛЕНИЯ ОБРАЗОВАТЕЛЬНОГО ПРОЦЕССА ПО ДИСЦИПЛИНЕ

Учебные аудитории для проведения занятий по дисциплине, оснащены оборудованием и техническими средствами обучения.

В таблице 9.1 перечислены:

- учебные аудитории для проведения учебных занятий, оснащенные оборудованием и техническими средствами обучения;

- помещения для самостоятельной работы обучающихся, которые должны быть оснащены компьютерной техникой с возможностью подключения к сети «Интернет» и обеспечением доступа в электронную информационно-образовательную среду ДПИ НГТУ.

Таблица 9.1

Оснащенность аудиторий и помещений
для самостоятельной работы обучающихся по дисциплине

№	Наименование аудиторий и помещений для самостоятельной работы	Оснащенность аудиторий помещений и помещений для самостоятельной работы	Перечень лицензионного программного обеспечения. Реквизиты подтверждающего документа
1	1347 Аудитория лекционных занятий	-	-
2	1440 Компьютерный класс; Нижегородская обл., г. Дзержинск, ул. Гайдара, д. 49	Оснащён ПК, CPU Intel core i5-10400/Ram 16 Gb/SSD 500 Gb/ Intel UHD Graphics 630 – 16 шт.	<ul style="list-style-type: none"> • Microsoft Windows 10 (подписка DreamSpark Premium) • Apache OpenOffice 4.1.8 (свободное ПО); • Mozilla Firefox (свободное ПО); • Adobe Acrobat Reader (свободное ПО); • 7-zip для Windows (свободное ПО); • КонсультантПлюс (ГПД № 0332100025418000079 от 21.12.2018)
3	1234 Научно-техническая библиотека ДПИ НГТУ, студенческий читальный зал; Нижегородская обл., г. Дзержинск, ул. Гайдара, д. 49	Комплект демонстрационного оборудования: <ul style="list-style-type: none"> • ПК, с выходом на мультимедийный проектор, на базе Intel Pentium G4560 3.5 ГГц, 4 Гб ОЗУ, монитор 20' – 1шт. • Мультимедийный проектор Epson- 1 шт; • Экран – 1 шт.; Набор учебно-наглядных пособий	<ul style="list-style-type: none"> • Microsoft Windows 10 Домашняя (поставка с ПК) • LibreOffice 6.1.2.1. (свободное ПО) • Foxit Reader (свободное ПО); • 7-zip для Windows (свободное ПО)
4	1443а компьютерный класс - помещение для СРС, курсового проектирования (выполнения курсовых работ), Нижегородская обл., г. Дзержинск, ул. Гайдара, д. 49	ПК на базе Intel Celeron 2.67 ГГц, 2 Гб ОЗУ, монитор Acer 17' – 4 шт. ПК подключены к сети «Интернет» и обеспечивают доступ в электронную информационно-образовательную среду университета	<ul style="list-style-type: none"> • Microsoft Windows 7 (подписка DreamSpark Premium) • Apache OpenOffice 4.1.8 (свободное ПО); • Mozilla Firefox (свободное ПО); • Adobe Acrobat Reader (свободное ПО); • 7-zip для Windows (свободное ПО); • КонсультантПлюс (ГПД № 0332100025418000079 от 21.12.2018)

10. МЕТОДИЧЕСКИЕ РЕКОМЕНДАЦИИ ОБУЧАЮЩИМСЯ

ПО ОСВОЕНИЮ ДИСЦИПЛИНЫ

10.1. Общие методические рекомендации для обучающихся по освоению дисциплины, образовательные технологии

Дисциплина реализуется посредством проведения контактной работы с обучающимися (включая проведение текущего контроля успеваемости), самостоятельной работы обучающихся и промежуточной аттестации.

Контактная работа: аудиторная, внеаудиторная, а также проводится в электронной информационно-образовательной среде университета (далее - ЭИОС).

Преподавание дисциплины ведется с применением следующих видов образовательных технологий:

- балльно-рейтинговая технология оценивания;
- текущий контроль знаний в форме тестирования в среде MOODLE.

При преподавании дисциплины «Большие данные» используются современные образовательные технологии, позволяющие повысить активность обучающихся при освоении материала курса и предоставить им возможность эффективно реализовать часы самостоятельной работы.

Лекционный материал курса может сопровождаться компьютерными презентациями, в которых наглядно преподносятся материал различных разделов курса и что дает возможность обсудить материал с обучающимися во время чтения лекций, активировать их деятельность при освоении материала. Материалы лекций в виде слайдов или тезисов находятся в свободном доступе в системе MOODLE и могут быть получены до чтения лекций и проработаны обучающимися в ходе самостоятельной работы.

На лекциях, лабораторных занятиях реализуются интерактивные технологии, приветствуются вопросы и обсуждения, используется личностно-ориентированный подход, технология работы в малых группах, что позволяет обучающимся проявить себя, получить навыки самостоятельного изучения материала, выровнять уровень знаний в группе.

Все вопросы, возникшие при самостоятельной работе над домашним заданием, подробно разбираются на лабораторных занятиях и лекциях. Проводятся индивидуальные и групповые консультации с использованием, как встреч с обучающимися, так и современных информационных технологий (электронная почта, Zoom).

Иницируется активность обучающихся, поощряется задание любых вопросов по материалу, практикуется индивидуальный ответ на вопросы обучающегося, рекомендуются методы успешного самостоятельного усвоения материала в зависимости от уровня его базовой подготовки.

Для оценки знаний, умений, навыков и уровня сформированности компетенции применяется балльно-рейтинговая система контроля и оценки успеваемости обучающихся очной формы обучения в процессе текущего контроля.

Промежуточная аттестация проводится в форме зачёта с оценкой с учетом текущей успеваемости.

Результат обучения считается сформированным на повышенном уровне, если теоретическое содержание курса освоено полностью. При устных собеседованиях обучающийся исчерпывающе, последовательно, четко и логически излагает учебный материал; свободно справляется с задачами, вопросами и другими видами заданий, использует в ответе дополнительный материал. Все предусмотренные рабочей учебной программой задания выполнены в соответствии с установленными требованиями, обучающийся способен анализировать полученные результаты, проявляет самостоятельность при выполнении заданий.

Результат обучения считается сформированным на пороговом уровне, если теоретическое содержание курса освоено полностью. При устных собеседованиях обучающийся последовательно, четко и логически стройно излагает учебный материал; справляется с задачами, вопросами и другими видами заданий, требующих применения знаний; все предусмотренные рабочей учебной программой задания выполнены в соответствии с

установленными требованиями, обучающийся способен анализировать полученные результаты; проявляет самостоятельность при выполнении заданий.

Результат обучения считается несформированным, если обучающийся при выполнении заданий не демонстрирует знаний учебного материала, допускает ошибки, неуверенно, с большими затруднениями выполняет задания, не демонстрирует необходимых умений, качество выполненных заданий не соответствует установленным требованиям, качество их выполнения оценено числом баллов ниже трех по оценочной системе, что соответствует допороговому уровню.

10.2. Методические указания для занятий лекционного типа

Лекционный курс предполагает систематизированное изложение основных вопросов тематического плана. В ходе лекционных занятий раскрываются базовые вопросы в рамках каждой темы дисциплины (Таблицы 4.2). Обозначаются ключевые аспекты тем, а также делаются акценты на наиболее сложные и важные положения изучаемого материала. Материалы лекций являются опорной основой для подготовки обучающихся к лабораторным работам и выполнения заданий самостоятельной работы, а также к мероприятиям текущего контроля успеваемости и промежуточной аттестации по дисциплине.

10.3. Методические указания по освоению дисциплины на лабораторных работах

Подготовку к каждой лабораторной работе обучающийся должен начать с ознакомления с планом занятия, который отражает содержание предложенной темы. Каждая выполненная работа с оформленным отчетом подлежит защите у преподавателя.

При оценивании лабораторных работ учитывается следующее:

- качество выполнения экспериментально-практической части работы и степень соответствия результатов работы заданным требованиям;
- качество оформления отчета по работе;
- качество устных ответов на контрольные вопросы при защите работы.

10.4. Методические указания по самостоятельной работе обучающихся

Самостоятельная работа обеспечивает подготовку обучающихся к аудиторным занятиям и мероприятиям текущего контроля и промежуточной аттестации по изучаемой дисциплине. Результаты этой подготовки проявляются в активности обучающихся на занятиях, в качестве выполненных заданий для самостоятельной работы и других форм текущего контроля.

При выполнении заданий для самостоятельной работы рекомендуется проработка материалов лекций по каждой пройденной теме, а также изучение рекомендуемой литературы, представленной в Разделе 6.

В процессе самостоятельной работы при изучении дисциплины обучающиеся могут работать на компьютере в специализированных аудиториях для самостоятельной работы (указано в таблице 9.1). В аудиториях имеется доступ через информационно-телекоммуникационную сеть «Интернет» к электронной информационно-образовательной среде университета (ЭИОС) и электронной библиотечной системе (ЭБС), где в электронном виде располагаются учебные и учебно-методические материалы, которые могут быть использованы для самостоятельной работы при изучении дисциплины.

Для обучающихся по заочной форме обучения самостоятельная работа является основным видом учебной деятельности.

11. ОЦЕНОЧНЫЕ СРЕДСТВА ДЛЯ КОНТРОЛЯ ОСВОЕНИЯ ДИСЦИПЛИНЫ

11.1. Типовые контрольные задания или иные материалы, необходимые для оценки знаний, умений, навыков и (или) опыта в ходе текущего контроля успеваемости

Для текущего контроля знаний обучающихся по дисциплине проводится **комплексная оценка знаний**, включающая

- тестирование на сайте преподавателя по различным разделам курса;
 - проведение лабораторных работ;
 - ответы на вопросы для самостоятельной работы для обучающихся очной формы.
- Далее для всех форм текущего контроля приведены примеры оценочных средств.

11.1.1. Типовые тестовые задания

Тестирование проводится в системе MOODLE. По приведённым в таблицах 4.2 темам проводится два теста. В разделе приведены примеры тестовых заданий для каждого теста по всем темам.

Тест 1

Раздел 1

1. Дайте определение Big Data

А) Комплексный набор инструментов обработки структурированных данных колоссальных объемов

В) Комплексный набор подходов, инструментов и методов обработки структурированных и неструктурированных данных колоссальных объемов

С) Комплексный набор методов обработки неструктурированных данных колоссальных объемов

Д) Комплексный набор методов обработки структурированных данных колоссальных объемов

2. Большие данные – это

А) Данные объемом более 1Тб

В) Данные объемом более 10Тб

С) Данные объемом более 100Тб

Д) Нет ограничений на минимальный объем

3. DATA SCIENCE -

А) это наука о методах анализа данных и извлечения из них ценной информации, знаний

В) это инструменты, подходы и методы обработки огромных объёмов данных.

С) это алгоритм, в ходе которого система обрабатывает большое число примеров, устанавливает закономерности и использует их, чтобы прогнозировать характеристики новых данных

Д) это техническая или программная система, способная решать задачи, традиционно считающиеся творческими, принадлежащие конкретной предметной области, знания о которой хранятся в памяти такой системы.

4. Machine learning -

А) это наука о методах анализа данных и извлечения из них ценной информации, знаний

В) это инструменты, подходы и методы обработки огромных объёмов данных.

С) это алгоритм, в ходе которого система обрабатывает большое число примеров, устанавливает закономерности и использует их, чтобы прогнозировать характеристики новых данных

Д) это техническая или программная система, способная решать задачи, традиционно считающиеся творческими, принадлежащие конкретной предметной области, знания о которой хранятся в памяти такой системы.

5. Artificial intelligence -

A) это наука о методах анализа данных и извлечения из них ценной информации, знаний

B) это инструменты, подходы и методы обработки огромных объемов данных.

C) это алгоритм, в ходе которого система обрабатывает большое число примеров, устанавливает закономерности и использует их, чтобы прогнозировать характеристики новых данных

D) это техническая или программная система, способная решать задачи, традиционно считающиеся творческими, принадлежащие конкретной предметной области, знания о которой хранятся в памяти такой системы.

Раздел 2

1. Транзакция - это

A) набор товаров, приобретенных покупателем в рамках одной отдельно взятой операции

B) множество событий, которые произошли одновременно

C) двумерная таблица, которая состоит из номера транзакции и перечня покупок

D) уникальный идентификатор, определяющий каждую сделку или транзакцию

2. Специалист по добыче данных - это

A) специалист, имеющий знания о окружении бизнеса, процессах, заказчиках, клиентах, потребителях, конкурентах

B) специалист, имеющий знания о том, где и каким образом хранятся данные, как получить к ним доступ и как связать между собой эти данные

C) специалист по анализу данных, который имеет, как минимум, основы статистических знаний

D) специалист, который управляет одним или большим количеством проектов

3. Поддержка набора - это

A) количество или процент транзакций, содержащих определенный набор данных

B) минимальный уровень поддержки

C) процент транзакций из всего набора, которые содержат одновременно наборы элементов A и B

D) вероятность того, что из события A следует событие B

4. Для машинного обучения подходят данные

A) Любых форматов в цифровом виде

B) Числовые типа int

C) Бинарные

D) Предварительно подготовленные, очищенные от ошибок, пропусков и выбросов, а также нормализованные и представленные в виде числовых векторов

5. Какой статистический показатель определяется как значение, приходящееся на середину ряда распределения?

A) мода

B) медиана

C) дисперсия

D) среднее арифметическое

Тест 2

Раздел 3

1. Коэффициенты регрессии - это

A) переменная, описывающая процесс, который пытаемся предсказать или понять

B) переменные, используемые для моделирования или прогнозирования значений зависимых переменных

C) параметры, которые рассчитываются в результате выполнения регрессионного анализа

D) невязки, описывающие необъяснимое количество зависимых величин, влияющих

на результат

2. Вспомогательная (искусственная) классификация -

A) производится по внешнему признаку и служит для придания множеству предметов (процессов, явлений) нужного порядка

B) производится по существенным признакам, характеризующим внутреннюю общность предметов и явлений

C) деление родового понятия только по признаку и только один раз до раскрытия всех видов

D) применяется для деления одного понятия по разным основаниям и синтеза таких простых делений в единое целое

3. При оценивании классификационных методов робастность характеризует

A) частные параметры модели

B) время, которое требуется на создание модели и ее использование

C) такие свойства, как размер дерева решений, компактность правил и т.п.

D) устойчивость к каким-либо нарушениям исходных предпосылок

4. Какое из приведённых ниже утверждений верно для кластеризации?

A) предназначена для разбиения совокупности объектов на неоднородные группы

B) поиск существующих заранее не определённых структур ("сгущение точек")

C) кластеры изучаемого набора данных заранее предопределены

D) для обучения модели необходим учитель

5. Клиент покупает билет на самолет через интернет. В момент покупки, он хочет знать насколько может упасть стоимость этого билета в ближайшем будущем и когда. К какому типу относится эта задача анализа данных?

A) прогнозирование

B) кластеризация

C) классификация

D) ассоциативные правила

Раздел 4

1. На стадии map

A) данные преобразуются и фильтруются

B) разбиваются по ключу

C) определяется финальный результат

D) параллельная сортировка данных

2. На стадии Shuffle

A) данные преобразуются и фильтруются

B) разбиваются по ключу

C) определяется финальный результат

D) выдаётся множество пар ключ-значение

3. На стадии reduce

A) данные преобразуются и фильтруются

B) разбиваются по ключу

C) определяется финальный результат обработки данных

D) считываются данные из источников

4. TF-IDF - это

A) статистическая мера, используемая для оценки важности слова в контексте документа, являющегося частью коллекции документов или корпуса.

B) отношение числа вхождений некоторого слова к общему числу слов документ

C) мера, используемая для оценки важности слова в пределах одного документа.

D) инверсия частоты, с которой некоторое слово встречается в документах коллекции

5. К стратегиям разработки MapReduce-приложений относят

A) Map only job

- B) Shuffle only job
- C) Reduce only job
- D) Shuffle Join

11.1.2. Типовые задания для лабораторных работ

Теоретические сведения и типовые задания приведены в методических указаниях (6.2).

Лабораторная работа 1

Типовое задание.

Выполнить первичный анализ данных и их очистку. Получить статистические показатели.

Лабораторная работа 2

Типовое задание.

Рассмотреть задачу прогнозирования.

Лабораторная работа 3

Типовое задание.

Рассмотреть задачи классификации.

Лабораторная работа 4

Типовое задание.

Рассмотреть задачи кластеризации.

Лабораторная работа 5

Типовое задание.

Рассмотреть задачу поиска типичных шаблонов

11.1.3. Типовые вопросы для устного и письменного опроса обучающихся очной формы обучения

По завершении лекционных занятий может быть выполнен устный или письменный опрос обучающихся для оценки работы на занятии и для оценки самостоятельной работы обучающихся.

Раздел 1.

1. Дать определение термину «большие данные».
2. Дать определение термину «Data Science».
3. Дать определение термину «Machine Learning».
4. Дать определение термину «Artificial intelligence».
5. Основные признаки BigData.
6. Жизненный цикл больших данных.
7. Проблемы анализа и обработки большого объема данных.
8. Источники больших данных.
9. Основные типы данных.
10. Принципы обработки больших данных
11. Архитектура информационных систем анализа данных
12. Теорема CAP
13. Теорема ACID
14. Организационные и человеческие факторы при внедрении ИС, работающей с большими данными.
15. Сферы применения BigData
16. Профессии, связанные с BigData
17. Стандарты в области работы с большими данными
18. СУБД для хранения больших данных. NoSQL.
19. Программное и аппаратное обеспечение, используемое для обработки и анализа больших данных

Раздел 2.

1. Источники больших данных
2. Методы сбора информации. Песочница проекта
3. Первичная очистка данных.
4. Дать определение термину «дискретизация».
5. Дать определение термину «бинаризация».
6. Трансформация временных рядов
7. Выбросы и пропуски в данных.
8. Основные понятия статистики. Статистические показатели
9. Дать определение терминам «минимум» и «максимум»
10. Дать определение термину «Среднее значение»
11. Характеристики разброса
12. Дать определение термину «Дисперсия»
13. Дать определение термину «Стандартное отклонение»
14. Дать определение термину «Медиана»
15. Матрица ковариаций и корреляций
16. Коэффициенты асимметрии, эксцесса, высшие моменты
17. Визуализация данных на разных этапах анализа больших данных

Раздел 3.

1. Методы Data Mining
2. Основные задачи Data Mining
3. Регрессионный анализ
4. Задача классификации
5. Задача кластеризации.
6. Методы кластеризации
7. Методы поиска ассоциативных правил
8. Определение дерева решений
9. Нейронные сети
10. Методика извлечения данных KDD
11. Этапы KDD
12. Методологии KDD

Раздел 4.

1. MapReduce. Состав и назначение.
2. Стадия Map
3. Стадия Shuffle
4. Стадия Reduce
5. Принципы технологии MapReduce
6. Задача Word Count
7. Мера TF-IDF
8. Приемы и стратегии разработки MapReduce-приложений
9. Фильтр Блума

11.2. Типовые контрольные задания или иные материалы, необходимые для оценки знаний, умений, навыков и (или) опыта в ходе промежуточной аттестации по дисциплине

Форма проведения промежуточной аттестации по дисциплине: зачёт с оценкой (по результатам накопительного рейтинга или в форме компьютерного тестирования).

Структура теста для итогового тестирования: в итоговом тесте 11 вопросов: 10 вопросов с выбором ответа и один вопрос открытого типа.

Вопрос открытого типа – это теоретический вопрос, взятый из раздела 5.1.

Перечень вопросов для подготовки к зачёту с оценкой (ПКС-2, ИПКС-2.1)

Перечень вопросов, выносимых на промежуточную аттестацию, приведён в разделе 5.1 настоящей рабочей программы.

Примерный тест для итогового тестирования (ПКС-2, ИПКС-2.1)

1. Основные задачи Data Mining

2. Дайте определение Big Data

A) Комплексный набор инструментов обработки структурированных данных колоссальных объемов

B) Комплексный набор подходов, инструментов и методов обработки структурированных и неструктурированных данных колоссальных объемов

C) Комплексный набор методов обработки неструктурированных данных колоссальных объемов

D) Комплексный набор методов обработки структурированных данных колоссальных объемов

3. Большие данные – это

A) Данные объемом более 1Тб

B) Данные объемом более 10Тб

C) Данные объемом более 100Тб

D) Нет ограничений на минимальный объем

4. Транзакция - это

A) набор товаров, приобретенных покупателем в рамках одной отдельно взятой операции

B) множество событий, которые произошли одновременно

C) двумерная таблица, которая состоит из номера транзакции и перечня покупок

D) уникальный идентификатор, определяющий каждую сделку или транзакцию

5. Специалист по добыче данных - это

A) специалист, имеющий знания о окружении бизнеса, процессах, заказчиках, клиентах, потребителях, конкурентах

B) специалист, имеющий знания о том, где и каким образом хранятся данные, как получить к ним доступ и как связать между собой эти данные

C) специалист по анализу данных, который имеет, как минимум, основы статистических знаний

D) специалист, который управляет одним или большим количеством проектов

6. Какой статистический показатель определяется как значение, приходящееся на середину ряда распределения?

A) мода

B) медиана

C) дисперсия

D) среднее арифметическое

7. Коэффициенты регрессии - это

A) переменная, описывающая процесс, который пытаемся предсказать или понять

B) переменные, используемые для моделирования или прогнозирования значений зависимых переменных

C) параметры, которые рассчитываются в результате выполнения регрессионного анализа

D) невязки, описывающие необъяснимое количество зависимых величин, влияющих на результат

8. Вспомогательная (искусственная) классификация -

A) производится по внешнему признаку и служит для придания множеству предметов (процессов, явлений) нужного порядка

B) производится по существенным признакам, характеризующим внутреннюю общность предметов и явлений

C) деление родового понятия только по признаку и только один раз до раскрытия всех видов

D) применяется для деления одного понятия по разным основаниям и синтеза таких простых делений в единое целое

9. Клиент покупает билет на самолет через интернет. В момент покупки, он хочет знать насколько может упасть стоимость этого билета в ближайшем будущем и когда. К какому типу относится эта задача анализа данных?

A) прогнозирование

B) кластеризация

C) классификация

D) ассоциативные правила

10. На стадии map

A) данные преобразуются и фильтруются

B) разбиваются по ключу

C) определяется финальный результат

D) параллельная сортировка данных

11. К стратегиям разработки MapReduce-приложений относят

A) Map only job

B) Shuffle only job

C) Reduce only job

D) Shuffle Join

Регламент проведения текущего контроля в форме компьютерного тестирования

Кол-во заданий в банке вопросов	Кол-во заданий, предъявляемых обучающемуся	Время на тестирование, мин.
не менее 100	11	40

Полный фонд оценочных средств для проведения промежуточной аттестации в форме компьютерного тестирования размещен в банке вопросов данного курса дисциплины в СДО MOODLE.