Министерство науки и высшего образования Российской Федерации

Муромский институт (филиал)

федерального государственного бюджетного образовательного учреждения высшего образования

«Владимирский государственный университет
имени Александра Григорьевича и Николая Григорьевича Столетовых»
(МИ ВлГУ)

Кафедра *ПИн*

«УТВЕРЖДАЮ»
Заместитель директора по УР
_____Д.Е. Андрианов
17.05.2022

РАБОЧАЯ ПРОГРАММА ДИСЦИПЛИНЫ

Анализ больших данных

Направление подготовки *09.04.04 Программная инженерия*

Профиль подготовки

Семестр	Трудоемкость, час./зач. ед.	Лекции, час.	Практические занятия, час.	Лабораторные работы, час.	Консультация, час.	Контроль, час.	Всего (контактная работа), час.	СРС, час.	Форма промежуточного контроля (экз., зач., зач. с оц.)
2	**72 / 2**	**12**		**16**	**1,2**	**0,25**	**29,45**	**42,55**	**Зач.**
Итого	**72 / 2**	**12**		**16**	**1,2**	**0,25**	**29,45**	**42,55**	

Муром, 2022 г.

1. Цель освоения дисциплины

Цель дисциплины: формирование компетенций в области использования информации, обработки и анализа ее для информационно-аналитической поддержки принятия управленческих решений. Знания, умения и навыки,

полученные в результате освоения дисциплины, помогут при сборе и анализе больших объемов структурированной и неструктурированной информации, при разработке моделей данных, и получении новых знаний.

Задачи освоения дисциплины:

− приобретение студентами знаний о технологиях подготовки, хранения, обработки и анализа данных;

− приобретение практических навыков работы с данными.

2. Место дисциплины в структуре ОПОП ВО

Изучение дисциплины "Анализ больших данных" базируется на изучении общих профессиональных дисциплин, а именно на дисциплинах "Методы оптимизации", "Современные алгоритмы обработки данных".

3. Планируемые результаты обучения по дисциплине

Планируемые результаты обучения по дисциплине, соотнесенные с планируемыми результатами освоения ОПОП (компетенциями и индикаторами достижения компетенций)

Формируемые компетенции (код, содержание компетенции)	Планируемые результаты обучения по дисциплине, в соответствии с индикатором достижения компетенции		Наименование оценочного средства
	Индикатор достижения компетенции	Результаты обучения по дисциплине	
ПК-2 Владение методами программной реализации распределенных информационных систем	ПК-2.1 Реализует методы и программные интерфейсы взаимодействия с внешними программными компонентами	Знать методы и программные интерфейсы взаимодействия с внешними программными компонентами (ПК-2.1) Умеет реализовывать методы и программные интерфейсы с целью взаимодействия с внешними программными компонентами (ПК-2.1) Владеет приемами взаимодействия с внешними программными компонентами (ПК-2.1)	тест
ПК-3 Владение навыками создания программного обеспечения для анализа, распознавания и обработки информации, систем цифровой обработки сигналов	ПК-3.2 Применяет методы машинного обучения для обработки информации	Знать методы машинного обучения, используемые для обработки информации (ПК-3.2) Умеет выбирать и применять методы машинного обучения для обработки информации (ПК-3.2) Владеет методами машинного обучения в практических задачах обработки информации (ПК-3.2)	тест

4. Структура и содержание дисциплины

Общая трудоемкость дисциплины составляет 2 зачетных единицы, 72 часа.

4.1. Форма обучения: очная

Уровень базового образования: высшее.
Срок обучения 2г.

4.1.1. Структура дисциплины

№ п\п	Раздел (тема) дисциплины	Семестр	Контактная работа обучающихся с педагогическим работником							Самостоятельная работа	Форма текущего контроля успеваемости (по неделям семестра), форма промежуточной аттестации(по семестрам)
			Лекции	Практические занятия	Лабораторные работы	Контрольные работы	КП / КР	Консультация	Контроль		
1	Технологии анализа данных	2	6		8					22	тестирование
2	Интеллектуальный анализ данных	2	6		8					20,55	тестирование
Всего за семестр		72	12		16			1,2	0,25	42,55	Зач.
Итого		72	12		16			1,2	0,25	42,55	

4.1.2. Содержание дисциплины

4.1.2.1. Перечень лекций

Семестр 2

Раздел 1. Технологии анализа данных

Лекция 1.

Большие данные (Big Data): современные подходы к обработке и хранению. Проблема множественного сравнения данных (2 часа).

Лекция 2.

Процесс анализа. Общая схема анализа. Извлечение и визуализация данных. Этапы моделирования. Процесс построения моделей. Формы представления данных, типы и виды данных. Представления наборов данных (2 часа).

Лекция 3.

Программное обеспечение в области анализа данных. Аналитические платформы: классификация и особенности применения. Языки визуального моделирования (2 часа).

Раздел 2. Интеллектуальный анализ данных

Лекция 4.

Ассоциативные правила. Аффинитивный анализ, предметный набор. Поддержка и достоверность ассоциативного правила (2 часа).

Лекция 5.

Определение кластеризации. Постановка задачи кластеризации. Цели кластеризации в Data Mining (2 часа).

Лекция 6.

Применение классификации и регрессии. Обзор методов классификации и регрессии. Статистические методы (2 часа).

4.1.2.2. Перечень практических занятий

Не планируется.

4.1.2.3. Перечень лабораторных работ

Семестр 2

Раздел 1. Технологии анализа данных

Лабораторная 1.

Понятие сценария и узла обработки. Консолидация данных (4 часа).

Лабораторная 2.

Трансформация данных. Визуализация данных (4 часа).

Раздел 2. Интеллектуальный анализ данных

Лабораторная 3.

Ассоциативные правила. Поиск ассоциативных правил (4 часа).

Лабораторная 4.

Прогнозирование с помощью линейной регрессии (4 часа).

4.1.2.4. Перечень тем и учебно-методическое обеспечение самостоятельной работы

Перечень тем, вынесенных на самостоятельное изучение:

1. Технологии KDD и Data Mining. Подготовка данных к анализу.
2. Методика извлечения знаний. Data Mining.
3. Программное обеспечение в области анализа данных.
4. Аналитические платформы: классификация и особенности применения.
5. Языки визуального моделирования.
6. Основные понятия теории нейронных сетей.
7. Основные парадигмы нейронных сетей.
8. Многослойный персептрон: класс решаемых задач, архитектура.
9. Определение дерева решений. Причины популярности и условия применимости.
10. Структура дерева решений. Выбор атрибута разбиения в узле.

Для самостоятельной работы используются методические указания по освоению дисциплины и издания из списка приведенной ниже основной и дополнительной литературы.

4.1.2.5. Перечень тем контрольных работ, рефератов, ТР, РГР, РПР

Не планируется.

4.1.2.6. Примерный перечень тем курсовых работ (проектов)

Не планируется.

4.2 Форма обучения: заочная

Уровень базового образования: высшее.

Срок обучения 2г 6м.

Семестр	Трудоемкость, час./ зач. ед.	Лекции, час.	Практические занятия, час.	Лабораторные работы, час.	Консультация, час.	Контроль,час.	Всего (контактная работа), час.	СРС, час.	Форма промежуточного контроля (экз., зач., зач. с оц.)
2	**72 / 2**	**4**		**6**	**2**	**0,5**	**12,5**	**55,75**	**Зач.(3,75)**
Итого	**72 / 2**	**4**		**6**	**2**	**0,5**	**12,5**	**55,75**	**3,75**

4.2.1. Структура дисциплины

№ п\п	Раздел (тема) дисциплины	Семестр	Контактная работа обучающихся с педагогическим работником							Самостоятельная работа	Форма текущего контроля успеваемости (по неделям семестра), форма промежуточной аттестации(по семестрам)
			Лекции	Практические занятия	Лабораторные работы	Контрольные работы	КП / КР	Консультация	Контроль		
1	Технологии анализа данных	2	4		6					55,75	тестирование
Всего за семестр		72	4		6	+		2	0,5	55,75	Зач.(3,75)
Итого		72	4		6			2	0,5	55,75	3,75

4.2.2. Содержание дисциплины

4.2.2.1. Перечень лекций

Семестр 2

Раздел 1. Технологии анализа данных

Лекция 1.

Большие данные (Big Data): современные подходы к обработке и хранению. Проблема множественного сравнения данных (2 часа).

Лекция 2.

Процесс анализа. Общая схема анализа. Извлечение и визуализация данных. Этапы моделирования. Процесс построения моделей. Формы представления данных, типы и виды данных. Представления наборов данных (2 часа).

4.2.2.2. Перечень практических занятий

Не планируется.

4.2.2.3. Перечень лабораторных работ

Семестр 2

Раздел 1. Технологии анализа данных

Лабораторная 1.

Понятие сценария и узла обработки. Консолидация данных (4 часа).

Лабораторная 2.

Трансформация данных. Визуализация данных (2 часа).

4.2.2.4. Перечень тем и учебно-методическое обеспечение самостоятельной работы

Перечень тем, вынесенных на самостоятельное изучение:

1. Программное обеспечение в области анализа данных. Аналитические платформы: классификация и особенности применения. Языки визуального моделирования.
2. Ассоциативные правила. Аффинитивный анализ, предметный набор. Поддержка и достоверность ассоциативного правила.
3. Определение кластеризации. Постановка задачи кластеризации. Цели кластеризации в Data Mining.
4. Применение классификации и регрессии. Обзор методов классификации и регрессии. Статистические методы.
5. Технологии KDD и Data Mining. Подготовка данных к анализу.
6. Методика извлечения знаний. Data Mining.
7. Программное обеспечение в области анализа данных.
8. Аналитические платформы: классификация и особенности применения.
9. Языки визуального моделирования.
10. Основные понятия теории нейронных сетей.
11. Основные парадигмы нейронных сетей.
12. Многослойный персептрон: класс решаемых задач, архитектура.
13. Определение дерева решений. Причины популярности и условия применимости.
14. Структура дерева решений. Выбор атрибута разбиения в узле.
15. Ассоциативные правила. Поиск ассоциативных правил.
16. Прогнозирование с помощью линейной регрессии.

Для самостоятельной работы используются методические указания по освоению дисциплины и издания из списка приведенной ниже основной и дополнительной литературы.

4.2.2.5. Перечень тем контрольных работ, рефератов, ТР, РГР, РПР

1. Жизненный цикл аналитики данных.
2. Hadoop Distributed File System.
3. Технология Map Reduce.
4. Архитектура Hadoop.
5. Масштабирование и многоуровневое хранение данных.
6. Интерактивный сторителлинг, дашборды.

4.2.2.6. Примерный перечень тем курсовых работ (проектов)

Не планируется.

5. Образовательные технологии

В процессе изучения дисциплины применяется контактная технология преподавания (за исключением самостоятельно изучаемых студентами вопросов). При проведении лабораторных занятий применяется имитационный или симуляционный подход. Шаги решения задач студентам демонстрируются при помощи мультимедийной техники. В дальнейшем студенты самостоятельно решают аналогичные задания.

6. Оценочные средства для текущего контроля успеваемости, промежуточной аттестации по итогам освоения дисциплины.

Фонды оценочных материалов (средств) приведены в приложении.

7. Учебно-методическое и информационное обеспечение дисциплины.

7.1. Основная учебно-методическая литература по дисциплине

1. Глебов, В. И. Практикум по математической статистике. Проверка гипотез с использованием Excel, MatCalc, R и Python : учебное пособие / В. И. Глебов, С. Я. Криволапов. — Москва : Прометей, 2019. — 86 с. — ISBN 978-5-907100-66-4. — Текст : электронный // Цифровой образовательный ресурс IPR SMART : [сайт]. — URL: https://www.iprbookshop.ru/94504.html (дата обращения: 25.11.2022). — Режим доступа: для авторизир. пользователей - https://www.iprbookshop.ru/94504.html

2. Карякин, М. И. Технологии программирования и компьютерный практикум на языке Python : учебное пособие / М. И. Карякин, К. А. Ватульян, Р. М. Мнухин. — Ростов-на-Дону, Таганрог : Издательство Южного федерального университета, 2022. — 241 с. — ISBN 978-5-9275-4108-9. — Текст : электронный // Цифровой образовательный ресурс IPR SMART : [сайт]. — URL: https://www.iprbookshop.ru/125718.html (дата обращения: 09.11.2022). — Режим доступа: для авторизир. пользователей - https://www.iprbookshop.ru/125718.html

3. Маккинли, Уэс Python и анализ данных / Уэс Маккинли ; перевод А. Слинкина. — 2-е изд. — Саратов : Профобразование, 2019. — 482 с. — ISBN 978-5-4488-0046-7. — Текст : электронный // Цифровой образовательный ресурс IPR SMART : [сайт]. — URL: https://www.iprbookshop.ru/88752.html (дата обращения: 23.08.2022). — Режим доступа: для авторизир. пользователей - https://www.iprbookshop.ru/88752.html

7.2. Дополнительная учебно-методическая литература по дисциплине

1. Протодьяконов, А. В. Алгоритмы Data Science и их практическая реализация на Python : учебное пособие / А. В. Протодьяконов, П. А. Пылов, В. Е. Садовников. — Москва, Вологда : Инфра-Инженерия, 2022. — 392 с. — ISBN 978-5-9729-1006-9. — Текст : электронный // Цифровой образовательный ресурс IPR SMART : [сайт]. — URL: https://www.iprbookshop.ru/124000.html (дата обращения: 28.09.2022). — Режим доступа: для авторизир. пользователей - https://www.iprbookshop.ru/124000.html

2. Карякин, М. И. Технологии программирования и компьютерный практикум на языке Python : учебное пособие / М. И. Карякин, К. А. Ватульян, Р. М. Мнухин. — Ростов-на-Дону, Таганрог : Издательство Южного федерального университета, 2022. — 241 с. — ISBN 978-5-9275-4108-9. — Текст : электронный // Цифровой образовательный ресурс IPR SMART : [сайт]. — URL: https://www.iprbookshop.ru/125718.html (дата обращения: 09.11.2022). — Режим доступа: для авторизир. пользователей - https://www.iprbookshop.ru/125718.html

3. Буйначев С.К. Основы программирования на языке Python : учебное пособие / Буйначев С.К., Боклаг Н.Ю.. — Екатеринбург : Уральский федеральный университет, ЭБС АСВ, 2014. — 92 с. — ISBN 978-5-7996-1198-9. — Текст : электронный // IPR SMART : [сайт]. — URL: https://www.iprbookshop.ru/66183.html (дата обращения: 24.11.2022). — Режим доступа: для авторизир. пользователей - https://www.iprbookshop.ru/66183.html

7.3. Перечень информационных технологий, используемых при осуществлении образовательного процесса по дисциплине, включая перечень программного обеспечения и информационных справочных систем

В образовательном процессе используются информационные технологии, реализованные на основе информационно-образовательного портала института (www.mivlgu.ru/iop), и инфокоммуникационной сети института:

- предоставление учебно-методических материалов в электронном виде;

- взаимодействие участников образовательного процесса через локальную сеть института и Интернет;

- предоставление сведений о результатах учебной деятельности в электронном личном кабинете обучающегося.

Информационные справочные системы:

Электронная библиотека ВлГУ - http://e.lib.vlsu.ru/

электронная библиотечная системы "IPRBooks" (http://www.iprbookshop.ru/);

Программное обеспечение:

Microsoft Visual Studio (Программа Microsoft Azure Dev Tools for Teaching (Order Number: IM126433))

Pycharm Community Edition (проприетарная лицензия и Apache License 2.0)

7.4. Перечень ресурсов информационно-телекоммуникационной сети «Интернет», необходимых для освоения дисциплины

iprbookshop.ru
e.lib.vlsu.ru
mivlgu.ru/iop

8. Материально-техническое обеспечение дисциплины

Лаборатория программного обеспечения и сопровождения компьютерных систем

Сервер «Ай Тек» на базе 2 процессоров Intel Xeon; 12 шт. компьютеров Intel Core i5-2400 3,10 GHz; 4гб, DVD-R/ Philips 19’; интерактивная доска SMART Boaerd 480 со встроенным проектором V25; маршрутизатор Gigabit Switch TEG-S16S. Маркерная доска. Доступ к сети Интернет.

9. Методические указания по освоению дисциплины

Для успешного освоения теоретического материала обучающийся: знакомится со списком рекомендуемой основной и дополнительной литературы; уточняет у преподавателя, каким дополнительным пособиям следует отдать предпочтение; ведет конспект лекций и прорабатывает лекционный материал, пользуясь как конспектом, так и учебными пособиями.

о выполнения лабораторных работ обучающийся изучает соответствующий раздел теории. Перед занятием студент знакомится с описанием заданий для выполнения работы, внимательно изучает содержание и порядок проведения лабораторной работы. Лабораторная работа проводятся в компьютерном классе. Обучающиеся выполняют индивидуальную задачу компьютерного моделирования в соответствии с заданием на лабораторную работу. Полученные результаты исследований сводятся в отчет и защищаются по традиционной методике в классе на следующем лабораторном занятии. Необходимый теоретический материал, индивидуальное задание, шаги выполнения лабораторной работы и требование к отчету приведены в методических указаниях, размещенных на информационно-образовательном портале института.

Самостоятельная работа оказывает важное влияние на формирование личности будущего специалиста, она планируется обучающимся самостоятельно. Каждый обучающийся самостоятельно определяет режим своей работы и меру труда, затрачиваемого на овладение учебным содержанием дисциплины. Он выполняет внеаудиторную работу и изучение разделов, выносимых на самостоятельную работу, по личному индивидуальному плану, в зависимости от его подготовки, времени и других условий.

Форма заключительного контроля при промежуточной аттестации – зачет. Для проведения промежуточной аттестации по дисциплине разработаны фонд оценочных средств и балльно-рейтинговая система оценки учебной деятельности студентов. Оценка по дисциплине выставляется в информационной системе и носит интегрированный характер, учитывающий результаты оценивания участия студентов в аудиторных занятиях, качества и своевременности выполнения заданий в ходе изучения дисциплины и промежуточной аттестации.

Программа составлена в соответствии с требованиями ФГОС ВО по направлению *09.04.04 Программная инженерия*
Рабочую программу составил *к.т.н., доцент Быков А.А.*_____

Программа рассмотрена и одобрена на заседании кафедры *ПИн*

протокол № 11 от 05.05.2022 года.
Заведующий кафедрой *ПИн* _____*Жизняков А.Л.*
(Подпись)

Рабочая программа рассмотрена и одобрена на заседании учебно-методической комиссии факультета

протокол №4 от 12.05.2022 года.

Председатель комиссии ФИТР _____ Рыжкова М.Н.
(Подпись) (Ф.И.О.)

Лист актуализации рабочей программы дисциплины

Программа одобрена на_____учебный год.

Протокол заседания кафедры № _____от _____20__ года.

Заведующий кафедрой_____ _____ _____
(Подпись) (Ф.И.О.)

Программа одобрена на_____учебный год.

Протокол заседания кафедры № _____от _____20__ года.

Заведующий кафедрой_____ _____ _____
(Подпись) (Ф.И.О.)

Программа одобрена на_____учебный год.

Протокол заседания кафедры № _____от _____20__ года.

Заведующий кафедрой_____ _____ _____
(Подпись) (Ф.И.О.)

Приложение

Фонд оценочных материалов (средств) по дисциплине

Анализ больших данных

1. Оценочные материалы для проведения текущего контроля успеваемости по дисциплине

1. Аналитик это …

+а)специалист в области анализа имоделирование

б) специалист в предметной области;

в)человек, решающий определенные задачи;

г) человек, который имеет опыт в программировании.

2 Эксперт это …

а) специалист в области анализа и моделирование;

+б) специалист в предметной области;

в) человек, решать определенные задачи;

г) человек, который имеет опыт в программировании.

3 Задача классификации сводится к …

а) нахождения частых зависимостей между объектами или событиями;

+б) определения класса объекта по его характеристиками;

в) определение по известным характеристиками объекта значение некоторого его параметра;

г) поиска независимых групп и их характеристик в всем множестве анализируемых данных.

4 Задача регрессии сводится к …

а) нахождения частых зависимостей между объектами или событиями;

б) определения класса объекта по его характеристиками;

+в) определение по известным характеристиками объекта значение некоторого его параметра;

г) поиска независимых групп и их характеристик в всем множестве анализируемых данных.

5 Задача кластеризации заключается в …

а) нахождения частых зависимостей между объектами или событиями;

б) определения класса объекта по его характеристиками;

в) определение по известным характеристиками объекта значение некоторого его параметра;

+г) поиска независимых групп и их характеристик в всем множестве анализируемых данных.

6 Целью поиска ассоциативных правил является …

+а) нахождения частых зависимостей между объектами или событиями;

б) определения класса объекта по его характеристиками;

в) определение по известным характеристиками объекта значение некоторого его параметра;

г) поиска независимых групп и их характеристик в всем множестве анализируемых данных.

7 До предполагаемых моделей относятся такие модели данных:

+а) модели классификации и последовательностей;

б) регрессивные, кластеризации, исключений, итоговые и ассоциации;

в) классификации, кластеризации, исключений, итоговые и ассоциации;

г) модели классификации, последовательностей и исключений.

8 В описательных моделей относятся следующие модели данных:

а) модели классификации и последовательностей;

+б) регрессивные, кластеризации, исключений, итоговые и ассоциации;

в) классификации, кластеризации, исключений, итоговые и ассоциации;

г) модели классификации, последовательностей и исключений.

9 Модели классификации описывают …

+а) правила или набор правил в соответствии с которыми можно отнести описание любого нового объекта к одному из классов;

б) функции, которые позволяют прогнозировать изменения непрерывных числовых параметров;

в) функциональные зависимости между зависимыми и независимыми показателями и переменными в понятной человеку форме;

г) группы, на которые можно разделить объекты, данные о которых подвергаются анализа.

10 Модели последовательностей описывают …

а) правила или набор правил в соответствии с которыми можно отнести описание любого нового объекта к одному из классов;

+б) функции, которые позволяют прогнозировать изменения непрерывных числовых параметров;

в) функциональные зависимости между зависимыми и независимыми показателями и переменными в понятной человеку форме;

г) группы, на которые можно разделить объекты, данные о которых подвергаются анализа.

11 Регрессивные модели описывают …

а)правила или набор правил в соответствии с которыми можно отнести описание любого нового объекта к одному из классов;

б) функции, которые позволяют прогнозировать изменения непрерывных

числовых параметров;

+ в) функциональные зависимости между зависимыми и независимыми показателями и переменными в понятной человеку форме;

г) группы, на которые можно разделить объекты, данные о которых подвергаются анализа.

12. Виды лингвистической неопределенности:

а) неточность измерений значений определенной величины, выполняемых физическими приборами;

+ б) неопределенность значений слов (Многозначность, размытость,

непонятность, нечеткость); неоднозначность смысла фраз (Синтаксическая и семантическая);

в) случайность (или наличие в внешней среде нескольких возможностей, каждая из которых случайным образом может стать действительностью); неопределенность значений слов (многозначность, размытость, неясность, нечеткость)

г) неоднозначность смысла фраз (Синтаксическая и семантическая).

13. Модели исключений описывают …

+ а) исключительные ситуации в записях, которые резко отличаются произвольной признаку от основной множества записей;

б) ограничения на данные анализируемого массива;

в) закономерности между связанными событиями;

г) группы, на которые можно разделить объекты, данные о которых подвергаются анализа.

14 Итоговые модели обнаружат …

а) исключительные ситуации в записях, которые резко отличаются произвольной признаку от основной множества записей;

+б) ограничения на данные анализируемого массива;

в) закономерности между связанными событиями;

г) группы, на которые можно разделитьобъекты, данные о которых подвергаются анализа.

15Модели ассоциации проявляют …

а)исключительные ситуации в записях, которые резко отличаются произвольной признаку от основной множества записей;

б) ограничения на данные анализируемого массива;

+в) закономерности между связанными событиями;

г) группы, на которые можно разделить объекты, данные о которых подвергаются анализа.

16 Виды физической неопределенности данных:

+ а) неточность измерений значений определенной величины, выполняемых

физическими приборами; случайность (или наличие в внешней среде нескольких возможностей, каждая из которых случайным образом может

стать действительностью)

б)неопределенность значений слов (Многозначность, размытость, непонятность, нечеткость); неоднозначность смысла фраз (Синтаксическая и семантическая);

в) случайность (или наличие в внешней среде нескольких возможностей, каждая из которых случайным образом может стать действительностью); неопределенность значений слов (многозначность, размытость, неясность, нечеткость);

г) неоднозначность смысла фраз (Синтаксическая и семантическая).

17 Очистка данных — …

+ а) комплекс методов и процедур, направленных на устранение причин, мешающих корректной обработке: аномалий, пропусков, дубликатов, противоречий, шумов и т.д.

б) процесс дополнения данных некоторой информацией, позволяющей повысить эффективность развязку аналитических задач

в) объект, содержащий структурированные данные, которые могут оказаться полезными для развязку аналитического задачи

г) комплекс методов и процедур, направленных на извлечение данных из

различных источников, обеспечение необходимого уровня их

информативности и качества, преобразования в единый формат, в котором они могут быть загружены в хранилище данных или аналитическую систему

18 Обогащение — …

а) комплекс методов и процедур, направленных на устранение

причин, мешающих корректной обработке: аномалий, пропусков, дубликатов, противоречий, шумов и т.д.

+б) процесс дополнения данных некоторой информацией, позволяющей

повысить эффективность развязку аналитических задач

в) объект, содержащий структурированные данные, которые могут оказаться полезными для развязку аналитического задачи

г) комплекс методов и процедур, направленных на извлечение данных из

различных источников, обеспечение необходимого уровня их информативности и качества, преобразования в единый формат, в котором они могут быть загружены в хранилище данных или аналитическую систему.

19 Консолидация — …

а) комплекс методов и процедур, направленных на устранение

причин, мешающих корректной обработке: аномалий, пропусков, дубликатов, противоречий, шумов и т.д.

б) процесс дополнения данных некоторой информацией, позволяющей

повысить эффективность развязку аналитических задач

в) объект, содержащий структурированные данные, которые могут оказаться

полезными для развязку аналитического задачи

+г) комплекс методов и процедур, направленных на извлечение данных из

различных источников, обеспечение необходимого уровня их

информативности и качества, преобразования в единый формат, в

котором они могут быть загружены в хранилище данных или

аналитическую систему

20 Транзакция — …

+а) некоторый набор операций над базой данных, который рассматривается как единственное завершено, с точки зрения пользователя, действие над некоторой информацией, обычно связано с обращением к базе данных

б) разновидность систем хранения, ориентирована на поддержку процесса

анализа данных целостность, обеспечивает, непротиворечивость и

хронологию данных, а также высокую скорость выполнения аналитических

запросов

в) высокоуровневые средства отражения информационной модели и описания структуры данных

г) это установление зависимости дискретной выходной переменной от входных переменных

21 Метаданные — …

а) некоторый набор операций над базой данных, который рассматривается как единственное завершено, с точки зрения пользователя, действие над некоторой информацией, обычно связано с обращением к базе данных

б) разновидность систем хранения, ориентирована на поддержку процесса

анализа данных целостность, обеспечивает, непротиворечивость и

хронологию данных, а также высокую скорость выполнения аналитических

запросов

+в) высокоуровневые средства отражения информационной модели и описания структуры данных

г) это установление зависимости дискретной выходной переменной от входных переменных

22 Классификация — …

а) некоторый набор операций над базой данных, который рассматривается как единственное завершено, с точки зрения пользователя, действие над некоторой информацией, обычно связано с обращением к базе данных

б) разновидность систем хранения, ориентирована на поддержку процесса

анализа данных целостность, обеспечивает, непротиворечивость и хронологию данных, а также высокую скорость выполнения аналитических

запросов

в) высокоуровневые средства отражения информационной модели и описания структуры данных

+г) это установление зависимости дискретной выходной переменной от входных переменных

23 Регрессия — …

+а) это установление зависимости непрерывной выходной переменной от входных переменных

б) эта группировка объектов (Наблюдений, событий) на основе данных, описывающих свойства объектов

в) выявление закономерностей между связанными событиями

г) это установление зависимости дискретной выходной переменной от входных переменных

24 Кластеризация — …

а) это установление зависимости непрерывной выходной переменной от входных переменных

+б) эта группировка объектов (Наблюдений, событий) на основе данных, описывающих свойства объектов

в) выявление закономерностей между связанными событиями

г) это установление зависимости дискретной выходной переменной от входных переменных.

25 Ассоциация — …

а)это установление зависимости непрерывной выходной переменной от входных переменных

б) эта группировка объектов (наблюдений, событий) на основе данных, описывающих свойства объектов

+в) выявление закономерностей между связанными событиями

г) это установление зависимости дискретной выходной переменной от входных переменных

26 Машинное обучение — …

а) специализированный программный решение (или набор решений), который включает в себя все инструменты для извлечения закономерностей из сырых данных

б) эта группировка объектов (Наблюдений, событий) на основе данных, описывающих свойства объектов

в) набор данных, каждая запись которого представляет собой учебный пример, содержащего заданный входной влияние, что и отвечает ему правильный выходной результат.

+г) подразделение искусственного интеллекта изучающий методы построения алгоритмов, способных обучаться на данных

27 Аналитическая платформа — …

+а) специализированный программный решение (или набор решений), который включает в себя все инструменты для извлечения закономерностей из сырых данных

б) эта группировка объектов (Наблюдений, событий) на основе данных, описывающих свойства объектов

в) набор данных, каждая запись которого представляет собой учебный пример, содержащего заданный входной влияние, что и отвечает ему правильный выходной результат.

г) подразделение искусственного интеллекта изучающий методы построения алгоритмов, способных обучаться на данных

28 Обучающая выборка — …

а) эта группировка объектов (Наблюдений, событий) на основе

данных, описывающих свойства объектов

+б) набор данных, каждая запись которого представляет собой учебный пример, содержащего заданный входной влияние, и соответствующий ему правильный выходной результат

в) набор данных, каждая запись которого представляет собой учебный пример, содержащего заданный входной влияние, что и отвечает ему

правильный выходной результат.

г) выявление в сырых данных ранее неизвестных, нетривиальных,

практически полезных и доступных интерпретации знаний, необходимых

для принятия решений в различных сферах человеческой деятельности

29 Ошибка обучения — …

+а) это ошибка, допущенная моделью на учебной множества.

б) это ошибка, полученная на тестовых примерах, то есть, что вычисляется по тем же формулам, но для тестовой множества

в) имена, типы, метки и назначения полей исходной выборки данных

г) набор данных, каждая запись которого представляет собой учебный пример, содержащего заданный входной влияние, и соответствующий ему правильный выходной результат

30 Ошибка обобщения — …

а) это ошибка, допущенная моделью на учебной множества.

+б) это ошибка, полученная на тестовых примерах, то есть, что вычисляется по тем же формулам, но для тестовой множества

в) имена, типы, метки и назначения полей исходной выборки данных

г) набор данных, каждая запись которого представляет собой учебный пример, содержащего заданный входной влияние, и соответствующий ему правильный выходной результат

Общее распределение баллов текущего контроля по видам учебных работ для студентов

Рейтинг-контроль 1	устный опрос, 1 тест	до 20 баллов
Рейтинг-контроль 2	устный опрос, 1 тест	до 20 баллов
Рейтинг-контроль 3	устный опрос, 1 тест	до 40 баллов
Посещение занятий студентом	контроль посещаемости	до 16 баллов
Дополнительные баллы (бонусы)	за своевременную защиту всех лабораторных	4
Выполнение семестрового плана самостоятельной работы	нет	0

2. Промежуточная аттестация по дисциплине

Перечень вопросов к экзамену / зачету / зачету с оценкой.
Перечень практических задач / заданий к экзамену / зачету / зачету с оценкой (при наличии)

1. Укажите фактор, способствовавший появлению тренда больших данных

Ответ:

(1) маркетинговые кампании крупных корпораций +

(2) снижение издержек на хранение данных +

(3) появление новых технологий обработки потоковых данных

(4) выпуск баз данных с обработкой данных в памяти

2. Выберите верный ответ

Ответ:

(1) большие данные – это обработка или хранение более 1 Тб информации

(2) проблема больших данных – это такая проблема, когда при существующих технологиях хранения и обработки сущностная обработка данных затруднена или невозможна +

(3) большие данные – это огромная PR-акция крупных вендоров и не более того

(4) большие данные – это явление, когда цифровые данные наиболее полно представляют изучаемый объект

3. Выберите неверный ответ:

Ответ:

(1) большие данные – это данные объёма свыше 1 Тб +

(2) проблема больших данных – это проблема, когда при существующих технологиях хранения и обработки сущностная обработка данных затруднена или невозможна

(3) большие данные – это тренд в области ИТ, подогреваемый маркетинговыми кампаниями крупных вендоров

(4) большие данные как правило не структурированы

4. Отметьте те из вариантов, в которых данные структурированы:

Ответ:

(1) данные о продажах компании, представленные в виде помесячных отчётов в формате MS Word

(2) таблица с ежедневными показаниями температуры помещения за год в файле формата csv +

(3) текст педагогической поэмы А.С. Макаренко, представленный в формате PDF

(4) библиотека фильмов, представленных в формате mpeg4 на одном жестком диске

5. Перечислите четыре основных характеристики Big Data:

Ответ:

(1) Virtualization, Volume, Variability, Vehicle

(2) Variety, Velocity, Volume, Value +

(3) Verification, Volume, Velocity, Visualization

(4) Video, Value, Variety, Volume

6. Выберите неверное высказывание:

Ответ:

(1) большие объёмы данных приводят к слабой их структуризации, поэтому появляется такое разнообразие данных +

(2) увеличившаяся производительность телекоммуникационных каналов привела к росту объёмов передаваемой информации

(3) удешевление систем хранения на единицу информации привело к росту рынка больших данных

7. Отметьте неверное понимание Variety в контексте характеристик Big Data:

Ответ:

(1) высокая скорость генерирования данных +

(2) разные типы данных в колонках таблиц реляционных СУБД +

(3) разнообразие отраслей, являющихся источниками данных +

(4) разнообразие типов данных, включающих в себя структурированные, полуструктурированные и неструктурированные

8. Принцип MapReduce состоит в том, чтобы

Ответ:

(1) производить вычисления на узлах, где информация изначально была сохранена +

(2) использовать вычислительные мощности систем хранения +

(3) использовать функциональное программирование для решения задач массивно-параллельной обработки

9. Выберите одно неверное высказывание про MapReduce:

Ответ:

(1) интерфейс для массово-параллельной обработки данных, где вычисления производятся на узлах, где информация изначально была сохранена

(2) MapReduce – это две операции: распределения и сборки данных

(3) MapReduce был придуман разработчиками Hadoop +

(4) MapReduce был анонсирован разработчиками Google

10. Во сколько раз теоретически вырастет производительность при подсчёте числа слов в тексте при работе MapReduce при переходе от одного узла к двум? (Введите число.)

Ответ:

2

11. Какие из следующих технологий СУБД не используют принцип MapReduce

Ответ:

(1) Hadoop

(2) Cassandra

(3) HDInsight

(4) Redis +

12. Какие СУБД полностью полагаются на оперативную память при хранении информации:

Ответ:

(1) Oracle Exalytics +

(2) SAP HANA +

(3) BigTable

(4) HBase

13. В чём преимущество колоночно-ориентированных СУБД?

Ответ:

(1) они позволяют выполнять более сложные SQL-запросы по сравнению с реляционными СУБД

(2) они позволяют динамически дополнять содержание записей новыми полями +

(3) они имеют более гибкие возможности аналитики

(4) они позволяют эффективно делать межколоночные сравнения

14. Для чего аналитику необходима "песочница"?

Ответ:

(1) для высокопроизводительной аналитики за счёт использования оперативной памяти и inDB операций +

(2) для хранения всех полученных от заказчика данных

(3) для построения отчётов о результатах анализа

(4) для снижения затрат, связанных с репликацией данных +

15. Какие из следующих средств разумно использовать для анализа данных, представленных единственным csv-файлом размера более 100Гб:

Ответ:

(1) Hadoop +

(2) Data Warehouse

(3) "Песочница" +

(4) Python +

16. Выберите верное утверждение:

Ответ:

(1) Data Warehouse создаются для проверки гипотез при анализе больших данных

(2) "Песочница" используется для снижения нагрузки на основной Data Warehouse +

(3) каждый Data Warehouse должен содержать "песочницу"

(4) "Песочница" необходима для любого процесса аналитики

17. Ниже приведена последовательность этапов проекта аналитики в соответствии с CRISP-DM, укажите первый этап.

Ответ:

(1) моделирование (Modeling)

(2) внедрение (Deployment)

(3) подготовка данных (Data Preparation)

(4) понимание бизнеса (Business understanding) +

(5) оценка (Evaluation)

(6) понимание данных (Data Understanding)

18. На каком из этапов процесса CRISP-DM происходит проверка гипотез?

Ответ:

(1) понимание бизнеса (Business understanding)

(2) понимание данных (Data Understanding)

(3) моделирование (Modeling) +

(4) оценка (Evaluation)

19. Вы являетесь владельцем и аналитиком в компании из 10 человек, в которой требуется проанализировать продажи за 1 год (1 млн. продаж). Какие из этапов CRISP-DM можно опустить:

Ответ:

(1) понимание бизнеса (Business understanding) +

(2) подготовка данных (Data Preparation)

(3) моделирование (Modeling)

(4) оценка (Evaluation)

20. Пример благоразумного использования Hadoop

Ответ:

(1) анализ 10 Гб данных

(2) ежедневное сохранение данных температуры, поступающих со всех городов России (по одному показанию на город, всего городов 1100 шт)

(3) посекундное сохранение данных температуры, поступающих со всех городов России (по одному показанию на город, всего городов 1100 шт)

(4) построение графика пульса пациента в реальном времени

21. Начиная с каких размеров данных обоснованно применение кластера Hadoop для хранения данных?

Ответ:

(1) 100Гб
(2) 1Тб
(3) 100Тб +
(4) 1Пб +

22. Hadoop – это:
Ответ:
(1) набор утилит, и программный каркас для выполнения распределённых программ, работающих на кластерах +
(2) распределённая СУБД, позволяющая обрабатывать большие данные
(3) язык выполнения заданий в парадигме MapReduce
(4) распределённая файловая система, предназначенная для хранения файлов большого объёма

Методические материалы, характеризующих процедуры оценивания

На основе типовых заданий программным комплексом информационно-образовательного портала МИ ВлГУ формируются в автоматическом режиме тестовые задания для студентов: три вопроса из блока 1, три вопроса из блока 2 и задача из блока 3. Программный комплекс формирует индивидуальные задания для каждого зарегистрированного в системе студента и устанавливает время прохождения тестирования. Результатом тестирования является процент правильных ответов, с учетом индивидуального семестрового рейтинга студента формируется экзаменационная оценка.

Максимальная сумма баллов, набираемая студентом по дисциплине равна 100.

Оценка в баллах	Оценка по шкале	Обоснование	***Уровень сформированности компетенций***
Более 80	«Отлично»	Содержание курса освоено полностью, без пробелов, необходимые практические навыки работы с освоенным материалом сформированы, все предусмотренные программой обучения учебные задания выполнены, качество их выполнения оценено числом баллов, близким к максимальному	***Высокий уровень***
66-80	«Хорошо»	Содержание курса освоено полностью, без пробелов, некоторые практические навыки работы с освоенным материалом сформированы недостаточно, все предусмотренные программой обучения учебные задания выполнены, качество выполнения ни одного из них не оценено минимальным числом баллов, некоторые виды заданий выполнены с ошибками	***Продвинутый уровень***

50-65	«Удовлетворительно»	Содержание курса освоено частично, но пробелы не носят существенного характера, необходимые практические навыки работы с освоенным материалом в основном сформированы, большинство предусмотренных программой обучения учебных заданий выполнено, некоторые из выполненных заданий, возможно, содержат ошибки	***Пороговый уровень***
Менее 50	«Неудовлетворительно»	Содержание курса не освоено, необходимые практические навыки работы не сформированы, выполненные учебные задания содержат грубые ошибки	***Компетенции не сформированы***

3. Задания в тестовой форме по дисциплине

Примеры заданий:

ПК-2

ПК-2.1

1. Какая из следующих СУБД подходит для организации высоко-доступного и консистентного хранилища?

Ответ:

(1) Greenplum +
(2) BigTable +
(3) CouchDB
(4) Cassandra

2. Какие характеристики объединяют следующие СУБД: Greenplum и BigTable?

Ответ:

(1) высокая-доступность +
(2) консистентность +
(3) распределённость
(4) колоночная ориентация +

3. Какие типы СУБД поддерживают одновременно высокую-доступность, консистентность и распределённость?

Ответ:

(1) NoSQL
(2) RDBMS
(3) построенные на базе HDFS
(4) никакие +

4. Какая из следующих СУБД подходит для организации высоко-доступного и распределённого хранилища?

Ответ:

(1) Cassandra +
(2) Hbase
(3) MongoDB
(4) CouchDB +

ПК-3

ПК-3.2

1. 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?

Ответ:

(1) прогнозирование +

(2) кластеризация

(3) классификация

(4) цензурирование

2. Клиент покупает билет на самолет через интернет. Как бы в данном случае формулировалась задача прогнозирования?

Ответ:

(1) в момент покупки предсказать, насколько и когда может упасть стоимость этого билета в ближайшем будущем +

(2) в момент покупки определить, к какому сегменту относится пользователь и предложить выгодные для него условия сделки

(3) выдать пользователю прогноз погоды для того места, куда он летит

(4) спрогнозировать вероятность дополнительных покупок (отель, страховка, туристический тур) и предложить наиболее вероятные пользователю +

3. Клиент покупает билет на самолет через интернет. В момент покупки стоит задача определить вероятность дополнительных покупок (отель, страховка, туристический тур) и предложить наиболее вероятные пользователю. К какому типу относится эта задача анализа данных?

Ответ:

(1) заполнение пробелов

(2) классификация

(3) прогнозирование +

(4) цензурирование

1. Инвестиционный фонд интересуется тем, почему часть финансируемых им проектов успешно переходят на второй год, а часть - нет. К какому типу относится эта задача анализа данных?

Ответ:

(1) поиск информативных признаков +

(2) построение решающего правила

(3) классификация

(4) цензурирование

Полный перечень тестовых заданий с указанием правильных ответов, размещен в банке вопросов на информационно-образовательном портале института по ссылке https://www.mivlgu.ru/iop/question/edit.php?courseid=3053

Оценка рассчитывается как процент правильно выполненных тестовых заданий из их общего числа.