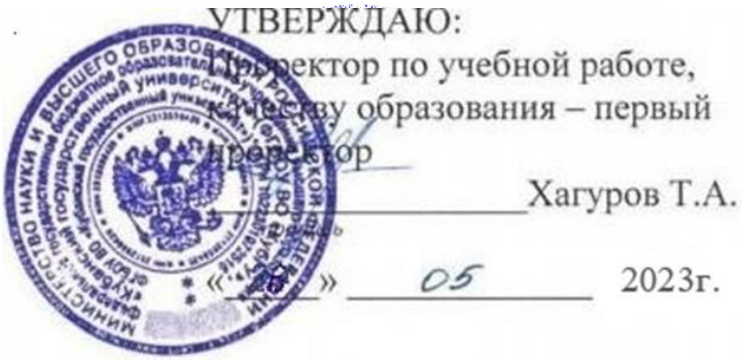
Министерство науки и высшего образования Российской Федерации
Федеральное государственное бюджетное образовательное учреждение высшего образования
«Кубанский государственный университет»

Факультет компьютерных технологий и прикладной математики
Кафедра вычислительных технологий



РАБОЧАЯ ПРОГРАММА УЧЕБНОЙ ДИСЦИПЛИНЫ
Б1.О.31 «ОБРАБОТКА БОЛЬШИХ ДАННЫХ»

Направление подготовки/специальность 02.03.02 **Фундаментальная информатика и информационные технологии**
(код и наименование направления подготовки/специальности)

Направленность (профиль) /специализация
Математическое и программное обеспечение компьютерных технологий

Программа подготовки *академический бакалавриат*

Форма обучения *очная*

Квалификация выпускника *бакалавр*

Краснодар 2023

Рабочая программа дисциплины «ОБРАБОТКИ БОЛЬШИХ ДАННЫХ» составлена в соответствии с федеральным государственным образовательным стандартом высшего образования (ФГОС ВО) по направлению подготовки 02.03.02 Фундаментальная информатика и информационные технологии

Программу составил(а):

Приходько Татьяна Александровна, доцент, к. т. н.
Ф.И.О. , должность, ученая степень, ученое звание — подпись

Рабочая программа дисциплины. «ОБРАБОТКИ БОЛЬШИХ ДАННЫХ» утверждена на заседании кафедры Вычислительных технологий
протокол № 8 «03 » мая 2023 г.

Заведующий кафедрой (разработчика) Вишняков Ю.М

(фамилия, инициалы — подпись

Утверждена на заседании учебно-методической комиссии факультета Компьютерных Технологий и Прикладной Математики
протокол № № 5 от «19» мая 2023 г

Председатель УМК факультета Коваленко А.В.

фамилия, инициалы — подпись

Рецензенты:

Гаркуша О.В., доцент кафедры информационных технологий ФБГОУ ВО «Кубанский государственный университет», кандидат физико-математических наук.

Схаляхо Ч.А., доцент КВВУ им.С.М.Штеменко, к.ф.-м.н., доцент

1. ЦЕЛИ И ЗАДАЧИ ОСВОЕНИЯ ДИСЦИПЛИНЫ

1.1 Цель освоения дисциплины

Курс «Обработка больших данных» имеет своей целью: формирование у студентов профессиональной компетенции в области разработки и использования систем обработки и анализа больших массивов данных. Данная цель соотносится с целью образовательной программой в частности с технологий разработки специализированных программных систем, отвечающих за обработку больших данных. Изучение данной дисциплины готовит выпускника к выполнению следующих профессиональных задач:

- Постановка задачи анализа данных.
- Предварительная обработка данных.
- Визуализация данных.
- Разработка, реализация и применение методов интеллектуального анализа данных к большим массивам данных.
- Представление результатов работы.

1.2 Задачи дисциплины

Основные задачи освоения дисциплины:

Студент должен **знать** методы анализа и хранения больших объемов данных, этапы жизненного цикла обработки больших данных, языки, наиболее приспособленные для обработки и аналитики больших данных, способы организации хранения и доступа к большим данным; **уметь** выполнять элементы анализа данных и интерпретировать результаты, различать характеристики SQL и NoSql БД, формулировать алгоритмы в парадигме MapReduce, выбрать подходящий инструмент анализа больших данных, выбрать подходящую технологию хранения больших данных.; **владеть** математическими методами анализа данных, языками и компьютерными методами обработки.

1.3 Место дисциплины (модуля) в структуре образовательной программы

Курс «Обработка больших данных» относится к базовой части блока Б1 дисциплин Дисциплины (модули).

Для изучения дисциплины студент должен владеть знаниями, умениями и навыками по дисциплинам:

Дискретная математика, Алгебраические структуры, Основы программирования, Алгоритмы вычислительной математики, Конструирование алгоритмов и структур данных, Теория алгоритмов и вычислительных процессов, Основы теории вероятностей и статистических методов, Алгоритмы и структуры данных, Математическая логика и теория алгоритмов, Интеллектуальный анализ данных.

Знания, получаемые при изучении дисциплины «Обработка больших данных» используются при изучении профессиональных дисциплин Распределенные задачи и алгоритмы, Программирование в компьютерных сетях, Облачные вычисления, Мультиагентные системы, а также для работ над дипломной и магистерской работой.

1.4 Перечень планируемых результатов обучения по дисциплине (модулю), соотнесенных с планируемыми результатами освоения образовательной программы.

Изучение данной учебной дисциплины направлено на формирование у обучающихся следующих **профессиональных компетенций и соотнесенных с ними индикаторов достижения компетенций:** ОПК-2; ОПК-3; ПК-5:

Код и наименование индикатора*	Результаты обучения по дисциплине *(знает, умеет, владеет (навыки и/или опыт деятельности))*
ОПК-2 Способен применять компьютерные/суперкомпьютерные методы, современное программное обеспечение, в том числе отечественного происхождения, для решения задач профессиональной деятельности	
ОПК-2.1. Знает основные положения и концепции в области программирования, архитектуру языков программирования, теории коммуникации, знает основную терминологию, знаком с содержанием Единого Реестра Российских программ	Знает стандарты обработки и анализа больших данных, и требования, связанные с созданием и использованием SQL и NoSQL систем хранения и обработки данных
ОПК-2.2. Знает особенности языков программирования, теорию алгоритмов, умеет составлять программы	Использовать современные инструментальные и вычислительные средства – язык анализа данных R, осуществлять постановку задач анализа данных, визуализацию интерпретацию результатов
ОПК-2.3. Имеет практический опыт решения задач анализа, интеграции различных типов программного обеспечения, анализа типов коммуникаций	Владеет способностью собирать, обрабатывать и интерпретировать данные современных научных исследований, необходимые для формирования выводов по соответствующим научным исследованиям
ОПК-3 Способен к разработке алгоритмических и программных решений в области системного и прикладного программирования, математических, информационных и имитационных моделей, созданию информационных ресурсов глобальных сетей, образовательного контента, прикладных баз данных, тестов и средств тестирования систем и средств на соответствие стандартам и исходным требованиям	
ОПК-3.1. Знает методы теории алгоритмов, методы системного и прикладного программирования, основные положения и концепции в области математических, информационных и имитационных моделей;	Знает математические методы анализа данных, методы и прикладные языки для разработки программных решений в области обработки больших данных, математических, информационных и имитационных моделей.
ОПК-3.2. Умеет соотносить знания в области программирования, интерпретацию прочитанного, определять и создавать информационные ресурсы глобальных сетей, образовательного контента, средств тестирования систем	Умеет корректно построить архитектуру кроссплатформенного приложения. Реализовать программу, включающую реализацию сенсорно-моторной координации и пространственного позиционирования, алгоритмы извлечения и обработки данных, включая возможности автономного принятия решений на основе ИИ.
ОПК-3.3. Имеет практический опыт применения разработки программного обеспечения.	Владеет языками системного и прикладного программирования для разработки математических, информационных и имитационных моделей, для обработки информационных ресурсов глобальных сетей и прикладных баз данных.
ПК-5 Способен применять в профессиональной деятельности современные языки программирования и методы параллельной обработки данных, операционные системы, электронные библиотеки и пакеты программ, сетевые технологии	
ПК-5.1. Знает основы разработки и реализации процессов жизненного цикла программного обеспечения	Современные языки и средства обработки дынных (язык R, Python), прикладные библиотеки для анализа данных
ПК-5.2. Умеет приобретать и использовать организационно- управленческие навыки в профессиональной и социальной деятельности	Умеет применить современные языки (R и Python) и прикладные библиотеки для анализа данных, сформулировать научную гипотезу и проверить ее достоверность
ПК-5.3. Имеет практический опыт	владеет средствами сбора, обработки и анализа

Код и наименование индикатора*	Результаты обучения по дисциплине *(знает, умеет, владеет (навыки и/или опыт деятельности))*
управления процессами жизненного цикла программных продуктов	больших данных, средствами оценки эффективности решений

В результате изучения дисциплины у студента формируются:

- представления о феномене больших данных, о научных и технических проблемах и возможностях, связанных с их появлением, о трендах в области технологий хранения и анализа больших данных;
- знания причин возникновения тренда больших данных, процессов анализа больших данных, основных подходов к обработке больших массивов данных, основ языка R;
- умения формулировать алгоритмы в парадигме MapReduce, выбрать подходящий инструмент анализа больших данных, выбрать подходящую технологию хранения больших данных.

Таблица 1. Профессиональные компетенции студента

№ п.п.	Индекс компе-тенции	Содержание компетенции (или ее части)	В результате изучения учебной дисциплины обучающие должны		
			знать	уметь	владеть
1.	ОПК-2	Способен применять компьютерные/супер-компьютерные методы, современное программное обеспечение, в том числе отечественного происхождения, для решения задач профессиональной деятельности	стандарты обработки и анализа больших данных, и требования, связанные с созданием и использованием SQL и NoSQL систем хранения и обработки данных	использовать современные инструментальные и вычислитель-ные средства (в соответствии с профилем подготовки), осуществлять постановку задач анализа данных, визуализацию интерпретацию результатов	способностью собирать, обрабатывать и интерпретиро-вать данные современных научных исследований, необходимые для формирова-ния выводов по соответствующи м научным исследованиям
	ОПК-3	Способен к разработке алгоритмических и программных решений в области системного и прикладного программирования, математических, информационных и имитационных моделей, созданию информационных ресурсов глобальных сетей, образовательного контента, прикладных баз данных, тестов и средств тестирования систем и средств на соответствие стандартам и исходным требованиям.	математические методы анализа данных, методы и прикладные языки для разработки программных решений в области обработки больших данных, математических, информационных и имитационных моделей	выполнять сбор и анализ данных, в том числе из сети Интернет, производить интерпретацию и оценку полученных результатов	языками системного и прикладного программировани я для разработки математических, информационных и имитацион-ных моделей, для обработки информационных ресурсов глобальных сетей и прикладных баз данных.
	ПК-5	Способен применять в	Современные	применить	владеет

		профессиональной деятельности современные языки программирования и методы параллельной обработки данных, операционные системы, электронные библиотеки и пакеты программ, сетевые технологии.	языки и средства обработки дынных (язык R, Python), прикладные библиотеки для анализа данных	современные языки (R и Python) и прикладные библиотеки для анализа данных, сформулировать научную гипотезу и проверить ее достоверность	средствами сбора, обработки и анализа больших данных, средствами оценки эффективности решений

2. СТРУКТУРА И СОДЕРЖАНИЕ ДИСЦИПЛИНЫ

2.1 Распределение трудоёмкости дисциплины по видам работ

Общая трудоёмкость дисциплины составляет 4 зач.ед. (144 часов), их распределение по видам работ представлено в таблице *(для студентов ОФО)*

Вид учебной работы	Всего часов	Семестры (часы)			
		6			
Контактная работа в том числе:					
Аудиторные занятия (всего):	68,3	68,3			
В том числе:					
Занятия лекционного типа	32	32			
Занятия семинарского типа (семинары, практ. занятия)					
Лабораторные занятия	32	32			
Иная контрольная работа					
Контроль самостоятельной работы	4	4			
Промежуточная аттестация (ИКР)	0,3	0,3			
Самостоятельная работа, в том числе	31	31			
В том числе:					
Курсовая работа					
Проработка учебного (теоретического) материала	12	12			
Выполнение индивидуальных заданий (подготовка сообщений, презентаций)	16	16			
Реферат					
Подготовка к текущему контролю	3	3			
Контроль:					
Подготовка к экзамену:	44,7	44,7			
Общая трудоемкость час	144	144			
в т.ч. контактная работа	68,3	68,3			
зач. ед.	4	4			

2.2 Структура дисциплины:

Распределение видов учебной работы и их трудоемкости по разделам дисциплины. Разделы дисциплины, изучаемые в _6__семестре *(очная форма)*

№	Наименование разделов	Количество часов				
		Всего	Аудиторная работа			Внеаудиторная работа
			Л	КСР	ЛР	СРС
1	2	3	4	5	6	7

1	2	3	4	5	6	7
1.	**Введение в большие данные.** Понятие Data Minig. Прикладные инструменты для работы с Big Data. Технология MapRaduce. Hadoop.	28	8	4	4	12
2.	**Технологии анализа данных:** Жизненный цикл анализа больших данных, стандарты. Алгоритмы классификации, кластеризации. Понятие корреляции и регрессионный анализ. Тестирование гипотез. Когнитивный анализ данных. Визуализация больших данных.	52	16	4	16	12
3.	**Технологии хранения больших данных**. Распределенные хранилища, NoSql хранилища, классификация и примеры.	17	8		8	1
	Итого по разделам дисциплины:	99				
	ИКР	0,3				
	Итого:	99,3	32	4	32	31
	Контроль	44,7				
	Итого по дисциплине:	144				

Примечание: Л – лекции, КСР – контрольные и самостоятельные работы, ЛР – лабораторные занятия, СРС – самостоятельная работа студента, Д-доклад, РГЗ – расчетно-графическое задание.

2.3 Содержание разделов дисциплины:

2.3.1 Занятия лекционного типа

№ раз-дела	Наименование раздела	Содержание раздела	Форма текущего контроля	Разработ. с участием представи-телей работо-дателей
1	2	3	4	5
1	**Введение в большие данные.** Понятие Data Minig. Прикладные инструменты для работы с Big Data. Технология MapRaduce. Hadoop.	**Предпосылки формирования тренда больших данных** ▪ Основные вызовы больших данных (4V) ▪ Определение термина "большие данные" ▪ Базовое представление о **Map Reduce** и **Hadoop** ▪ Представление о работе аналитика **Инструменты для обработки больших данных** ▪ Знакомство с языками и прикладными пакетами для обработки больших данных. ▪ Рассмотрение общей концепции и синтаксиса языка R (примеры).	ЛР	

2	**Технологии анализа данных:** Жизненный цикл анализа больших данных, стандарты. Алгоритмы классификации, кластеризации. Понятие корреляции и регрессионный анализ. Тестирование гипотез. Когнитивный анализ данных. Визуализация больших данных.	**Аналитика больших данных.** ▪ Процесс аналитики ▪ Стандарты жизненного цикла Big Data: ***CRISP-DM*** **Когнитивный анализ данных** ▪ Введение в Data Mining – понятие, структура, составляющие и сопутствующие науки. ▪ Задачи Data Mining и способы их решения. Классификация методов DM. ▪ Области применения DM. ▪ Классы систем DM. ▪ Процесс накопления и анализа данных: Азбука когнитивного анализа. **Аналитика больших данных.** **Математическая статистика** Основные понятия статистики и дескриптивный анализ ▪ Шкалы измерений. ▪ Генеральная совокупность и выборка. ▪ Нормальное распределение. Уровень статистической достоверности. ▪ Свойства описательных статистик (Дескриптивный анализ) ▪ Визуальное представление данных ▪ Меры изменчивости **Методы DATA MINING** ▪ Данные & знания ▪ Типовые задачи Data Mining ▪ Обучаемые и необучаемые задачи ▪ Жизненный цикл проекта DM ▪ Математический аппарат DM ▪ Стандарты DM **Задачи классификации и кластеризации** ▪ Naive Bayes ▪ Desisison Tree ▪ RandomForest ▪ K-means ▪ R и MapReduse **Методы анализа на графах** Случайные графы, безмасштабные графы, социальные сети – сети тесного мира. Закономерности, методы кластеризации на графах.	ЛР РГЗ	

		Корреляция, регрессионный анализ ○ Понятие корреляции ○ Значимость коэффициента корреляции ○ Виды связи между переменными **Тестирование гипотез** ○ Виды гипотез ○ Предварительный анализ данных ○ Параметрические и непараметрические тесты.		
3	**Технологии хранения больших данных**. Распределенные хранилища, NoSql хранилища, классификация и примеры.	**Хранилища данных** ▪ Хранилища данных ○ OLAP и OLTP системы ○ Характеристики BigData и хранилища данных ○ Почему не реляционные СУБД? ▪ Требования к хранилищам данных ▪ Регрессионный анализ **Распределенные базы данных NoSQL. Решение задач Data Mining. Задачи классификации, кластеризации** 2. Распределенные базы данных NoSQL ▪ Типы NoSQL ▪ Репликация и шардинг ▪ Пример NoSQL БД	ЛР Д	

2.3.2. Занятия семинарского типа

Занятия семинарского типа – не предусмотрены.

2.3.3. Лабораторные занятия

№ работы	№ раздела дисциплины	Наименование лабораторных работ
1	1	Ознакомление с синтаксисом языка ***R*** для анализа данных.(4ч)
2	1	Способы подготовки и отображения данных в ***R*** (4ч). Возможности ввода/вывода.
3	2	Решение задач на больших графах (2ч).
4	2	Способы анализа данных в R. Получение первичных элементарных характеристик о наборах данных (элементарные статистики). Способы импорта/экспорта данных(2ч).
5	2	Работа с диаграммами и графиками в R (2ч).
6	2	Проверка статистических гипотез (4ч)
7	2	Корреляционный анализ и регрессионный анализ данных (2ч)
8	2	Решение задач Data Mining. Задачи классификации, кластеризации: деревья решений, RandomForest, k-means. (4ч)
9	3	Изучение принципов работы распределенных баз данных
10	1-3	Круглый стол: Совместное обсуждение результатов РГЗ
11	1-3	Обсуждение итогов курса

2.3.3 Примерная тематика курсовых работ (проектов)

Учебным планом не предусмотрены.

2.3.4 Расчетно-графические задания (индивидуальное задание)

В процессе изучения дисциплины "Обработка больших данных" студентами выполняется одно расчетно-графическое (индивидуальное) задание. Темы заданий для каждого студента различны. Задача РГЗ состоит в проверке умений студентов и проверке эффективности их самостоятельной работы в плане сбора и анализа данных.

Темы заданий ежегодно обновляются. Общая тематика соответствует тематике лабораторных работ.

Примеры РГЗ – задания на анализ данных

Загрузить данные в таблицу (ниже) из указанного источника и проанализировать взаимное влияние параметров, отобразить корреляцию:

a. Роста ВВП на прирост населения
b. Прироста населения на динамику безработицы
c. Прирост людей с высшим образованием на рост промышленного производства
d. Прирост людей с высшим образованием на развитие науки
e. Прирост людей с высшим образованием на динамику доходов на душу населения
f. Динамику безработицы на динамику преступности
g. С помощью регрессионного анализа найдите зависимые переменные и поясните влияние на них независимых переменных.
h. С помощью функции predict() (см. лекции и help()) постройте прогноз по столбцу, соответствующему варианту.

Годы	Численность населения	Рост ВВП	Динамика безработицы	Динамика промышленного производства	Прирост людей, получивших очное высшее образование	Развитие науки (высокотехнологичных отраслей)	Динамика доходов на душу населения	Динамика преступности
	1	2	3	4	5	6	7	8
01.01.1990								
…								
01.01.2015								

Отчет по выполнению РГЗ должен содержать:

2. постановку задачи;
3. сформированный набор данных;
4. тексты скриптов на языке R;
5. результаты тестов на проверку гипотез о корреляции, оценка регрессии, вычисление корреляции в текстовом и графическом виде.
6. ясное и подробное пояснение каждого результата, словесную трактовку графиков;
7. выводы;
8. список использованной литературы.

2.4 Перечень учебно-методического обеспечения для самостоятельной работы обучающихся по дисциплине

№	Вид СРС	Перечень учебно-методического обеспечения дисциплины по выполнению самостоятельной работы
1	2	3
1	**Раздел 1.** Чтение публикаций по истории развития Big Data, Data Mining [3-5]. Изучение языка R и Phython [7,8] осн. список, [1-4] - дополнительный.	Приходько Т.А. Методические указания по выполнению лабораторных работ по дисциплине «Обработка больших данных», утвержденные кафедрой вычислительных технологий.
2	**Раздел2.** Изучение части курса Introduction to Data Science [6], посвященной визуализации. Визуализация стандартных наборов данных при помощи Tableau.	Приходько Т.А. Методические указания по выполнению лабораторных работ по дисциплине «Обработка больших данных», утвержденные кафедрой вычислительных технологий.
3	**Раздел 3.** Изучение парадигмы Map Reduce. Подсчет кол-ва слов, реализация алгоритма k-means в рамках парадигмы Map Reduce с использованием Hadoop. Чтение публикаций о распределенных хранилищах данных, их особенностях и принципах построения распределенных файловых систем.	Источники основной и дополнительной литературы

Учебно-методические материалы для самостоятельной работы обучающихся из числа инвалидов и лиц с ограниченными возможностями здоровья (ОВЗ) предоставляются в формах, адаптированных к ограничениям их здоровья и восприятия информации:

Для лиц с нарушениями зрения:

– в печатной форме увеличенным шрифтом, в форме электронного документа,

Для лиц с нарушениями слуха:

– в печатной форме, в форме электронного документа.

Для лиц с нарушениями опорно-двигательного аппарата:

– в печатной форме, в форме электронного документа,

Данный перечень может быть конкретизирован в зависимости от контингента обучающихся.

3. ОБРАЗОВАТЕЛЬНЫЕ ТЕХНОЛОГИИ

Семестр	Вид занятия (Л, ПР, ЛР)	Используемые интерактивные образовательные технологии	Количество часов
6	Л	Компьютерные презентации и обсуждение	32
	ЛР	Разбор конкретных ситуаций (задач), тренинги по решению задач, компьютерные симуляции (программирование алгоритмов), подготовка и обсуждение докладов.	32
	КРС	Контрольная работа	4
Итого:			68

Для лиц с ограниченными возможностями здоровья предусмотрена организация консультаций с использованием электронной почты.

4. ОЦЕНОЧНЫЕ СРЕДСТВА ДЛЯ ТЕКУЩЕГО КОНТРОЛЯ УСПЕВАЕМОСТИ И ПРОМЕЖУТОЧНОЙ АТТЕСТАЦИИ

4.1 Фонд оценочных средств для проведения текущего контроля

№ п/п	Код и наименование индикатора	Результаты обучения	Наименование оценочного средства	
			Текущий контроль	Промежуточная аттестация
1	ОПК-2.1. Знает основные положения и концепции в области программирования, архитектуру языков программирования, теории коммуникации, знает основную терминологию, знаком с содержанием Единого Реестра Российских программ	Знает стандарты обработки и анализа больших данных, и требования, связанные с созданием и использованием SQL и NoSQL систем хранения и обработки данных	Опрос по теме лабораторных работ.	Вопросы 1-43
2	ОПК-2.2. Знает особенности языков программирования, теорию алгоритмов, умеет составлять программы	Использовать современные инструментальные и вычислительные средства – язык анализа данных R, осуществлять постановку задач анализа данных, визуализацию интерпретацию результатов	Опрос по теме лабораторных работ.	Вопросы 1-43, выносимые на зачет
3	ОПК-2.3. Имеет практический опыт решения задач анализа, интеграции различных типов программного обеспечения, анализа типов коммуникаций	Владеет способностью собирать, обрабатывать и интерпретировать данные современных научных исследований, необходимые для формирования выводов по соответствующим научным исследованиям	Опрос по теме лабораторных работ.	Вопросы 1-43, выносимые на зачет
4	ОПК-3.1. Знает методы теории алгоритмов, методы системного и прикладного программирования, основные положения и концепции в области математических, информационных и имитационных моделей;	Знает математические методы анализа данных, методы и прикладные языки для разработки программных решений в области обработки больших данных, математических, информационных и имитационных моделей.	Опрос по теме лабораторных работ.	Вопросы 52-86, выносимые на зачет
5	ОПК-3.2. Умеет соотносить знания в области программирования, интерпретацию прочитанного, определять и создавать информационные ресурсы глобальных сетей, образовательного контента, средств тестирования систем	Умеет корректно построить архитектуру кроссплатформенного приложения. Реализовать программу, включающую реализацию сенсорно-моторной координации и пространственного позиционирования, алгоритмы извлечения и обработки данных,	Опрос по теме лабораторных работ.	Вопросы 52-86, выносимые на зачет

		включая возможности автономного принятия решений на основе ИИ.		
6	ОПК-3.3. Имеет практический опыт применения разработки программного обеспечения.	Владеет языками системного и прикладного программирования для разработки математических, информационных и имитационных моделей, для обработки информационных ресурсов глобальных сетей и прикладных баз данных.	Опрос по теме лабораторных работ.	Вопросы 52-86, выносимые на зачет
7	ПК-5.1. Знает основы разработки и реализации процессов жизненного цикла программного обеспечения	Современные языки и средства обработки дынных (язык R, Python), прикладные библиотеки для анализа данных	Опрос по теме лабораторных работ.	Вопросы 27-51, выносимые на зачет
8	ПК-5.2. Умеет приобретать и использовать организационно-управленческие навыки в профессиональной и социальной деятельности	Умеет применить современные языки (R и Python) и прикладные библиотеки для анализа данных, сформулировать научную гипотезу и проверить ее достоверность	Опрос по теме лабораторных работ.	Вопросы 27-51, выносимые на зачет
9	ПК-5.3. Имеет практический опыт управления процессами жизненного цикла программных продуктов	владеет средствами сбора, обработки и анализа больших данных, средствами оценки эффективности решений	Опрос по теме лабораторных работ.	Вопросы 27-51, выносимые на зачет

Фонд оценочных средств дисциплины состоит из средств текущего контроля (вопросы при защите ЛР, контрольной работы) лабораторных работ, средств итоговой аттестации (зачет в 6 семестре).

Оценка успеваемости осуществляется по результатам:

- выполнения лабораторных работ;

- выполнения лабораторных работ;

- ответа на экзамене (для выявления знания и понимания теоретического материала дисциплины).

Текущий контроль включает контрольную работу по итогам первой половины курса.

Пример экзаменационного билета:

1. Данные, информация, знания – в чем отличия? Области, где эффективно используются Big Data, примеры.
2. Свойства описательных статистик (Дескриптивный анализ). Меры изменчивости.
3. Что такое наивный байесовский алгоритм? Какую задачу обработки данных он выполняет? Приведите и поясните формулу теорема Байеса.
4. Индивидуальное задание.

Пример задания к билету:

Для чтения файла воспользуйтесь командой
chem <- read.csv(file = file.choose(), header = TRUE, sep = ",")
Дан датасет: **SmokeBan.csv**

Сокращают Ли Запреты На Курение На Рабочем Месте Курение?

Описание

Оценка влияния запрета на курение на рабочем месте на курение работников, работающих в помещениях.

Формат

Фрейм данных, содержащий 500 наблюдений по 7 переменным.

курильщик фактор. Является ли данный человек в настоящее время курильщиком?
бан фактор. Существует ли запрет на курение в рабочей зоне?
Возраст- возраст в годах.
образование, указывающий на наивысший достигнутый уровень образования: окончание средней школы (hs), выпускник средней школы, какой-либо колледж, выпускник колледжа, степень магистра (или выше).
afam фактор. Является ли индивид афроамериканцем?
hispanic (латиноамериканец) фактор. Является ли индивид латиноамериканцем?
пол фактор, указывающий на пол.

Подробности SmokeBank - это набор данных с наблюдениями за 500 работников, работающих в помещении, который является подмножеством набора данных, собранных в рамках Национального опроса по вопросам здравоохранения в 1991 году, а затем снова (с разными респондентами) в 1993 году. Набор данных содержит информацию о том, подпадали или не подпадали под действие запрета на курение на рабочем месте.

1. Найти средний возраст курильщиков (не используя цикл).
2. Подсчитать число курильщиков афроамериканцев моложе 35 лет. И отдельно курильщиков латиноамериканцев старше 30 лет.
3. Проверить влияние образования и то, является ли человек в настоящее время курильщиком, а также влияние пола на курение.
4. Построить боксплоты, сопоставляющие сообщества курильщиков афроамериканцев, латиноамериканцев и всех остальных.
5. Проверить гипотезу о среднем возрасте курильщиков всех трех групп.
6. Построить гистограммы по трем группам людей.
7. Полученные результаты показать и прокомментировать устно.

****) Все графики должны сопровождаться заголовками и подписями по осям.***

Перечень вопросов, для подготовки к экзамену

Л1

1. Назовите источники появления Больших Данных.
2. В каких областях деятельности используются большие данные, привести примеры.
3. Основные вызовы больших данных (6V).
4. Определение термина "большие данные", источники получения больших данных.
5. Перечислите и охарактеризуйте логические слои для работы с большими данными.
6. Данные, информация, знания – в чем отличия?
7. Области деятельности, где эффективно используются БД, примеры.

Л2

8. Каковы основные инструменты аналитики больших данных, провести сравнительную характеристику.
9. Сравнительная характеристика R и Python.
10. Охарактеризовать конструкции языка R
11. Перечислить типы языка R, привести примеры.
12. Структуры и типы данных в языке R, привести примеры.
13. Векторы, матрицы, фреймы и факторы в R. Сходство и различия, способы обработки.
14. Принцип массивных (векторных и матричных вычислений в R).

Л3

15. Основные понятия статистики и дескриптивный анализ
16. Генеральная совокупность и выборка.
17. Шкалы измерений.
18. Меры центральной тенденции, их сравнительный анализ.
19. Виды функций распределения. Нормальное распределение. Уровень статистической достоверности.
20. Свойства описательных статистик (Дескриптивный анализ).
21. Перечислить и охарактеризовать меры изменчивости.

Л4

22. Способы графического представления данных.
23. Примеры использования гистограммы для обработки фотографий и оценки качества изделий.
24. Виды столбчатых диаграмм и их интерпретация.
25. Boxplot и его интерпретация, связь этого графика с другими элементами анализа.
26. Для чего нужны гипотезы в анализе данных, какие существуют приемы работы с гипотезами?

Л5

27. Опишите стандарты жизненного цикла Big Data.
28. Что называется когнитивным анализом данных?
29. Назовите этапы интеллектуального анализа данных.
30. Что такое статистическое обучение?
31. В чем разница между описательными и предсказательными задачами DM? Какие методы анализа лучше приспособлены для описательных, а какие для предсказательных задач?
32. В каких случаях лучше использовать линейные, а когда нелинейные модели анализа данных?
33. Приведите математическое выражение для параметрической модели статистического обучения, для каких задач анализа данных их лучше использовать?
34. В чем состоят преимущества и недостатки непараметрических моделей анализа данных, как осуществить выбор между параметрической и непараметрической моделью?
35. Как выполняется измерение качества модели анализа данных?
36. В чем состоит фундаментальное свойство статистического обучения?

Л6

37. Основные задачи Data Mining. Какие дисциплины охватывает Data Mining?
38. Классификация методов DM.
39. Понятие кластерного анализа, Классификация алгоритмов кластеризации.
40. Зачем нужна мера близости в кластеризации? В чем достоинства алгоритмов, построенных на основе теории графов? Перечислите виды алгоритмов кластеризации на графах.
41. В чем суть алгоритмов нахождения квадратичной ошибки?
42. Плоские алгоритмы кластеризации перечислить, охарактеризовать работу.
43. Поясните суть работы алгоритмов нахождения связных компонент и алгоритмов покрывающего дерева.
44. Опишите шаги построения дендрограммы.
45. В чем состоит суть стандартизации и нормализации переменных, зачем они нужны?

Л7

46. Что такое наивный байесовский алгоритм? Какую задачу обработки данных он выполняет?
47. Приведите и поясните формулу теорема Байеса.
48. Деревья решений, опишите процесс работы, приемы остановки работы дерева.
49. Какие типы деревьев решений вы знаете, какие индексы используются при работе дерева, для чего, что такое энтропия?
50. Назовите достоинства и недостатки деревьев решения.
51. Принцип работы RandomForest.

Л8

52. Data Mining vs. Machine Learning – в чем отличия?
53. Нарисуйте схему классификации методов машинного обучения.

54. Обучение с учителем и без учителя. Приведите примеры методов для обоих вариантов.
55. Обучение с подкреплением и ансамбли – основные разновидности и принципы работы.
56. Принципы глубокого обучения в нейросетях. В чем преимущества сверточных нейронных сетей, какие задачи они решают очень хорошо?
57. Зачем нужны рекуррентные нейросети?

Л9

58. Что такое статистическая гипотеза? Какие виды гипотез вы знаете?
59. Как принято формулировать нулевую гипотезу?
60. Что такое уровень значимости, как он определяется, как влияет на решение о принятие гипотезы?
61. Каков порядок обработки данных при тестировании гипотезы о равенстве, какие еще тесты при этом должны быть пройдены, какие требования к данным выдвигаются?
62. Для чего вообще нужна гипотеза о равенстве средних?
63. Как тестируются независимые и парные выборки?

Л10

64. Понятие корреляции, коэффициент корреляции Пирсона, Спирмена, Кендела.
65. Каковы факторы, влияющие на коэффициент корреляции?
66. Назовите виды связи между переменными при корреляции.
67. Что такое регрессионный анализ, какие задачи DM можно проводить с его помощью?
68. Какие способы визуализации корреляции были изучены в курсе Big Data?

Л11

69. Перечислите основные задачи анализа сетей на графах. Приведите примеры.
70. Перечислите разновидности сложных сетей, назовите их характеристики.
71. Характерные черты безмасштабных сетей, какова их связь с сетями тесного мира?
72. Каковы закономерности динамики сложных сетей и законы распространения информации в них.
73. Свойства эластичности и надежности сложных сетей.
74. Понятие регрессии. Как используется этот вид анализа?

Л12

75. Дайте определение социального графа. Перечислите его типы и свойства. К какому семейству больших графов он относится?
76. Какие алгоритмы лежат в основе методов выделения сообществ? Дайте общее описание шагов выполнения этих алгоритмов.
77. Перечислите основные метрики больших графов.
78. Назовите критерии качества кластеризации и поясните их значение и когда они используются.
79. Приведите примеры задач, которые могут быть решены с помощью больших графов.
80. Дайте сравнительный анализ алгоритмов кластеризации.

Л13

81. Охарактеризуйте хранилища данных типа OLAP и OLTP. Назовите разницу.
82. Архитектура хранилищ данных.
83. Требования ACID. CAP-теорема, BASE архитектура: как и к каким хранилищам данных эти понятия применяются.
84. Мотивация происхождения NoSql.
85. NoSql. Классификация NoSql хранилищ (типы). Их особенности. Примеры распределенных хранилищ.
86. Понятия репликации и шардинга для хранилищ данных.

Критерии оценивания к экзамену:

- 84-100 баллов (оценка «отлично») - изложенный материал фактически верен, наличие глубоких исчерпывающих знаний в объеме пройденной программы дисциплины в соответствии с поставленными программой курса целями и задачами обучения; правильные, уверенные действия по применению полученных знаний на практике, грамотное и логически стройное изложение материала при ответе, усвоение основной и знакомство с дополнительной литературой; Практические задания выполнены в срок в полном объеме.

- 67-83 баллов (оценка «хорошо») - наличие твердых и достаточно полных знаний в объеме пройденной программы дисциплины в соответствии с целями обучения, правильные действия по применению знаний на практике, четкое изложение материала, допускаются отдельные логические и стилистические погрешности. Практические задания выполнены в срок в объеме не менее 80%.

- 50-66 баллов (оценка удовлетворительно) - наличие твердых знаний в объеме пройденного курса в соответствии с целями обучения, изложение ответов с отдельными ошибками, уверенно исправленными после дополнительных вопросов; правильные в целом действия по применению знаний на практике; Практические задания выполнены в объеме не менее 60%.

- 0-49 баллов (оценка неудовлетворительно) - ответы не связаны с вопросами, наличие грубых ошибок в ответе, непонимание сущности излагаемого вопроса, неумение применять знания на практике, неуверенность и неточность ответов на дополнительные и наводящие вопросы». Практические задания не выполнены либо предоставлены не в срок в объеме менее 50%.

Оценочные средства для инвалидов и лиц с ограниченными возможностями здоровья выбираются с учетом их индивидуальных психофизических особенностей.

– при необходимости инвалидам и лицам с ограниченными возможностями здоровья предоставляется дополнительное время для подготовки ответа на экзамене;

– при проведении процедуры оценивания результатов обучения инвалидов и лиц с ограниченными возможностями здоровья предусматривается использование технических средств, необходимых им в связи с их индивидуальными особенностями;

– при необходимости для обучающихся с ограниченными возможностями здоровья и инвалидов процедура оценивания результатов обучения по дисциплине может проводиться в несколько этапов.

Процедура оценивания результатов обучения инвалидов и лиц с ограниченными возможностями здоровья по дисциплине предусматривает предоставление информации в формах, адаптированных к ограничениям их здоровья и восприятия информации:

Для лиц с нарушениями зрения:

– в печатной форме увеличенным шрифтом, в форме электронного документа.

Для лиц с нарушениями слуха:

– в печатной форме, в форме электронного документа.

Для лиц с нарушениями опорно-двигательного аппарата:

– в печатной форме, в форме электронного документа.

Данный перечень может быть конкретизирован в зависимости от контингента обучающихся.

5. ПЕРЕЧЕНЬ ОСНОВНОЙ И ДОПОЛНИТЕЛЬНОЙ УЧЕБНОЙ ЛИТЕРАТУРЫ, НЕОБХОДИМОЙ ДЛЯ ОСВОЕНИЯ ДИСЦИПЛИНЫ

5.1 Основная литература:

1. Крутиков, В.Н. Анализ данных : учебное пособие / В.Н. Крутиков, В.В. Мешечкин ; Министерство образования и науки Российской Федерации, Федеральное государственное бюджетное образовательное учреждение высшего профессионального образования «Кемеровский государственный университет». - Кемерово : Кемеровский государственный университет, 2019. - 138 с. : ил. - Библиогр. в кн. - ISBN 978-5-8353-1770-7 ; То же [Электронный ресурс]. - URL: http://biblioclub.ru/index.php?page=book&id=278426
2. Жуковский, О.И. Информационные технологии и анализ данных : учебное пособие / О.И. Жуковский ; Министерство образования и науки Российской Федерации, Томский Государственный Университет Систем Управления и Радиоэлектроники (ТУСУР). - Томск : Эль Контент, 2020. - 130 с. : схем., ил. - Библиогр.: с. 126. - ISBN 978-5-4332-0158-3 ; То же [Электронный ресурс]. - URL: http://biblioclub.ru/index.php?page=book&id=480500
3. Базы данных в высокопроизводительных информационных системах : учебное пособие / авт.-сост. Е.И. Николаев ; Министерство образования и науки РФ, Федеральное государственное автономное образовательное учреждение высшего образования «Северо-Кавказский федеральный университет». - Ставрополь : СКФУ, 2016. - 163 с. : ил. - Библиогр.: с.161. ; То же [Электронный ресурс]. - URL: http://biblioclub.ru/index.php?page=book&id=466799

5.2. Дополнительная литература:

1. Туманов, В.Е. Проектирование хранилищ данных для систем бизнес-аналитики : учебное пособие / В.Е. Туманов. - Москва : Интернет-Университет Информационных Технологий, 2010. - 616 с. : ил., табл., схем. - (Основы информационных технологий). - ISBN 978-5-9963-0353-3 ; То же [Электронный ресурс]. - URL: http://biblioclub.ru/index.php?page=book&id=233492
2. Добронец, Б.С. Численный вероятностный анализ неопределенных данных : монография / Б.С. Добронец, О.А. Попова ; Министерство образования и науки Российской Федерации, Сибирский Федеральный университет. - Красноярск : Сибирский федеральный университет, 2014. - 168 с. : граф., ил. - Библиогр. в кн. - ISBN 978-5-7638-3093-4 ; То же [Электронный ресурс]. - URL: http://biblioclub.ru/index.php?page=book&

5.3. Интернет-ресурсы, в том числе современные профессиональные базы данных и информационные справочные системы

Электронно-библиотечные системы (ЭБС):

1. **ЭБС «ЮРАЙТ»** https://urait.ru/
2. **ЭБС «УНИВЕРСИТЕТСКАЯ БИБЛИОТЕКА ОНЛАЙН»** http://www.biblioclub.ru/
3. **ЭБС «BOOK.ru»** https://www.book.ru
4. **ЭБС «ZNANIUM.COM»** www.znanium.com
5. **ЭБС «ЛАНЬ»** https://e.lanbook.com

Профессиональные базы данных

1. **Scopus** http://www.scopus.com/
2. **ScienceDirect** https://www.sciencedirect.com/
3. **Журналы издательства Wiley** https://onlinelibrary.wiley.com/
4. **Научная электронная библиотека (НЭБ)** http://www.elibrary.ru/
5. **Полнотекстовые архивы ведущих западных научных журналов на Российской платформе научных журналов НЭИКОН** http://archive.neicon.ru
6. **Национальная электронная библиотека** (доступ к Электронной библиотеке диссертаций Российской государственной библиотеки (РГБ) https://rusneb.ru/

7. **Президентская библиотека им. Б.Н. Ельцина** https://www.prlib.ru/
8. **База данных CSD Кембриджского центра кристаллографических данных (CCDC)** https://www.ccdc.cam.ac.uk/structures/
9. **Springer Journals:** https://link.springer.com/
10. **Springer Journals Archive:** https://link.springer.com/
11. **Nature Journals:** https://www.nature.com/
12. **Springer Nature Protocols and Methods**: https://experiments.springernature.com/sources/springer-protocols
13. **Springer Materials:** http://materials.springer.com/
14. **Nano Database:** https://nano.nature.com/
15. **Springer eBooks (i.e. 2020 eBook collections):** https://link.springer.com/
16. **"Лекториум ТВ"** http://www.lektorium.tv/
17. **Университетская информационная система РОССИЯ** http://uisrussia.msu.ru

Информационные справочные системы

1. **Консультант Плюс** - справочная правовая система (доступ по локальной сети с компьютеров библиотеки)

Ресурсы свободного доступа

1. **КиберЛенинка** http://cyberleninka.ru/;
2. **Американская патентная база данных** http://www.uspto.gov/patft/
3. **Министерство науки и высшего образования Российской Федерации** https://www.minobrnauki.gov.ru/;
4. **Федеральный портал "Российское образование"** http://www.edu.ru/;
5. **Информационная система "Единое окно доступа к образовательным ресурсам"** http://window.edu.ru/;
6. **Единая коллекция цифровых образовательных ресурсов** http://school-collection.edu.ru/ .
7. **Проект Государственного института русского языка имени А.С. Пушкина "Образование на русском"** https://pushkininstitute.ru/;
8. **Справочно-информационный портал "Русский язык"** http://gramota.ru/;
9. **Служба тематических толковых словарей** http://www.glossary.ru/;
10. **Словари и энциклопедии** http://dic.academic.ru/;
11. **Образовательный портал "Учеба"** http://www.ucheba.com/;
12. **Законопроект "Об образовании в Российской Федерации". Вопросы и ответы** http://xn--273--84d1f.xn--p1ai/voprosy_i_otvety

Собственные электронные образовательные и информационные ресурсы КубГУ

1. **Электронный каталог Научной библиотеки КубГУ** http://megapro.kubsu.ru/MegaPro/Web
2. **Электронная библиотека трудов ученых КубГУ** http://megapro.kubsu.ru/MegaPro/UserEntry?Action=ToDb&idb=6
3. **Среда модульного динамического обучения** http://moodle.kubsu.ru
4. **База учебных планов, учебно-методических комплексов, публикаций и конференций** http://infoneeds.kubsu.ru/
5. **Библиотека информационных ресурсов кафедры информационных образовательных технологий** http://mschool.kubsu.ru;
6. **Электронный архив документов КубГУ** http://docspace.kubsu.ru/
7. **Электронные образовательные ресурсы кафедры информационных систем и технологий в образовании КубГУ и научно-методического журнала "ШКОЛЬНЫЕ ГОДЫ"** http://icdau.kubsu.ru/

6. Методические указания для обучающихся по освоению дисциплины

По курсу предусмотрено проведение лекционных занятий, на которых дается основной систематизированный материал, лабораторных работ, контрольной работы, зачета и экзамена.

Важнейшим этапомкурса является самостоятельная работа по дисциплине с использованием указанных литературных источников и методических указаний автора курса.

Виды и формы СР, сроки выполнения, формы контроля приведены выше в данном документе.

Для лучшего освоения дисциплины при защите ЛР студент должен ответить на несколько вопросов из лекционной части курса.

В освоении дисциплины инвалидами и лицами с ограниченными возможностями здоровья большое значение имеет индивидуальная учебная работа (консультации) – дополнительное разъяснение учебного материала.

Индивидуальные консультации по предмету являются важным фактором, способствующим индивидуализации обучения и установлению воспитательного контакта между преподавателем и обучающимся инвалидом или лицом с ограниченными возможностями здоровья.

7. Перечень информационных технологий, используемых при осуществлении образовательного процесса по дисциплине

7.1 Перечень информационных технологий

- Проверка домашних заданий и консультирование посредством электронной почты.
- Использование электронных презентаций при проведении лекций и практических занятий.

7.2 Перечень необходимого программного обеспечения

1. Phyton,
2. R, R Studio.
3. Программы для демонстрации и создания презентаций («Microsoft Power Point»).

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

Наименование специальных помещений	Оснащенность специальных помещений	Перечень лицензионного программного обеспечения
Учебные аудитории для проведения занятий лекционного типа (ауд. 129, 131, А305.)	Мебель: учебная мебель Технические средства обучения: экран, проектор, компьютер	PowerPoint.
Учебные аудитории для проведения занятий семинарского типа, групповых и индивидуальных консультаций, текущего контроля и промежуточной аттестации (ауд. 147,148)	Мебель: учебная мебель Технические средства обучения: экран, проектор, компьютер	Аудитория, (кабинет) – компьютерный класс
Учебные аудитории для проведения лабораторных работ. Лаборатория 102,105,106	Мебель: учебная мебель Технические средства обучения: компьютер	Лаборатория, укомплектованная специализированными техническими средствами обучения – компьютерный класс, с возможностью подключения к сети «Интернет», программой экранного увеличения и обеспеченный доступом в

		электронную информационно-образовательную среду университета

Для самостоятельной работы обучающихся предусмотрены помещения, укомплектованные специализированной мебелью, оснащенные компьютерной техникой с возможностью подключения к сети «Интернет» и обеспечением доступа в электронную информационно-образовательную среду университета.

Наименование помещений для самостоятельной работы обучающихся	Оснащенность помещений для самостоятельной работы обучающихся	Перечень лицензионного программного обеспечения
Помещение для самостоятельной работы обучающихся (читальный зал Научной библиотеки)	Мебель: учебная мебель Комплект специализированной мебели: компьютерные столы Оборудование: компьютерная техника с подключением к информационно-коммуникационной сети «Интернет» и доступом в электронную информационно-образовательную среду образовательной организации, веб-камеры, коммуникационное оборудование, обеспечивающее доступ к сети интернет (проводное соединение и беспроводное соединение по технологии Wi-Fi)	1. OS Windows, MS Office 2. R, R Studio. 3. Антивирус.
Помещение для самостоятельной работы обучающихся (ауд. 105, 148,150)	Мебель: учебная мебель Комплект специализированной мебели: компьютерные столы Оборудование: компьютерная техника с подключением к информационно-коммуникационной сети «Интернет» и доступом в электронную информационно-образовательную среду образовательной организации, веб-камеры, коммуникационное оборудование, обеспечивающее доступ к сети интернет (проводное соединение и беспроводное соединение по технологии Wi-Fi)	1.OS Windows, MS Office 2.R, R Studio. 3.Антивирус.