Министерство науки и высшего образования Российской Федерации

Федеральное государственное бюджетное образовательное учреждение высшего образования
«Сибирский государственный индустриальный университет»

Кафедра прикладных информационных технологий и программирования

УТВЕРЖДАЮ
Директор института
информационных технологий и
автоматизированных систем
_____ А.В. Зимин
подпись
«____» _____ 20___ г.

РАБОЧАЯ ПРОГРАММА ДИСЦИПЛИНЫ

Математические и инструментальные методы анализа данных

09.04.03 «Прикладная информатика»
(направленность (профиль): «Искусственный интеллект и машинное обучение»)

Квалификация выпускника
Магистр

Форма обучения
Очная форма

Срок обучения: 2 года

Год начала подготовки 2026

Новокузнецк
2026

СибГИУ
Подлинник электронного документа хранится в системе электронного согласования документов СибГИУ
ДОКУМЕНТ ПОДПИСАН ЭП
Сертификат: daba56d7b892efca3bf5c04ed0251c98b19b099c
Владелец: Директор ИИТиАС Зимин Алексей Валерьевич
Действителен: с 15 января 2025 г. по 10 апреля 2026 г.
Дата и время: 01-04-2026 10:05:12 ПРОВЕРИТЬ ПОДПИСЬ

1 Цели и задачи освоения учебной дисциплины

Целями учебной дисциплины являются:

– формирование целостного представления о методах и алгоритмах сбора, хранения и анализа больших объемов данных.

Задачами учебной дисциплины являются:

– знакомство с основными понятиями в области обработки «больших данных»;
– знакомство с методами и инструментами получения, хранения и анализа данных;
– получение практических навыков обработки и анализа данных с использованием специализированного программного обеспечения.

2 Место учебной дисциплины в структуре ООП по направлению подготовки (специальности)

Учебная дисциплина относится к учебным дисциплинам части, формируемой участниками образовательных отношений **Блока 1 «Дисциплины (модули)»** ООП по направлению подготовки (специальности) 09.04.03 «Прикладная информатика».

Учебная дисциплина опирается на базовые знания и компетенции, полученные в процессе получения предыдущего образования.

Учебная дисциплина дополняет знания и умения, получаемые по одновременно изучаемым и последующим учебным дисциплинам:

– Визуализация данных для принятия управленческих решений;
– Разработка и реализация проектов 1;
– Разработка и реализация проектов 2;
– Научно-исследовательская работа;
– Подготовка к процедуре защиты и защита выпускной квалификационной работы.

3 Планируемые результаты обучения по учебной дисциплине

Процесс изучения учебной дисциплины направлен на формирование следующих компетенций:

– Профессиональные компетенции

Наименование категории (группы) ПК	Код и наименование ПК	Код и наименование индикатора достижения ПК	Планируемые результаты обучения
	ПК-2: Способен к проведению работ по обработке и	ПК-2.2 Проводит анализа научных данных, результатов экспериментов и наблюдений	– знать: методы сбора и интеграции данных из различных источников, их подготовки, первичной обработки и

	анализу научно-технической информации и результатов исследований		визуализации . – уметь: осуществлять сбор, первичную обработку и визуализацию данных при решении профессиональных задач.
		ПК-2.3 Обобщает данные, результаты экспериментов и наблюдений	– знать: теоретические и прикладные основы анализа больших объемов данных и построения моделей на основе данных. – уметь: использовать типовые алгоритмы анализа и обработки больших объемов данных.

4 Объем и содержание учебной дисциплины

Учебные занятия по учебной дисциплине проводятся в форме контактной работы и в форме самостоятельной работы обучающихся.

Контактная работа включает в себя занятия лекционного типа (лекции), занятия семинарского типа (семинары, практические занятия, практикумы), промежуточную аттестацию обучающихся и иные формы взаимодействия обучающихся с педагогическими работниками и (или) лицами, привлекаемыми к реализации ООП на иных условиях, в том числе при проведении промежуточной аттестации обучающихся. Контактная работа может проводиться с применением электронного обучения, дистанционных образовательных технологий.

Объем учебной дисциплины

Семестр / курс		***ИТОГО***	***1 семестр***
Форма промежуточной аттестации			*экзамен, зачет с оценкой по КР*
Трудоёмкость	*академ. час.*	***180***	*180*
	зачетных единиц	***5***	*5*
Лекции, *академ. час.*		***16***	*16*
в форме практической подготовки		***0***	*0*
Лабораторные работы, *академ. час.*		***0***	*0*
в форме практической подготовки		***0***	*0*
Практические занятия, *академ. час.*		***32***	*32*
в форме практической подготовки		***0***	*0*
Курсовая работа, *академ. час.*		***36***	*36*
в форме практической подготовки		***0***	*0*
Консультации, *академ. час.*		***0***	*0*
в форме практической подготовки		***0***	*0*

Самостоятельная работа, *академ. час.*	***60***	*60*
в форме практической подготовки	***0***	*0*
Контроль, *академ. час.*	***36***	*36*
в форме практической подготовки	***0***	*0*

Содержание учебной дисциплины

Раздел 1 Введение в науку о данных;

Тема 1.1 Основы анализа данных (Принципы анализа данных, классический и информационный подходы. Структурированные данные. Источники данных. Сбор данных. Требования к данным. Подготовка данных к анализу);

Тема 1.2 Концепция Big Date (Большие данные) (Предпосылки формирования тренда больших данных; определение термина «большие данные». Сферы использования больших данных. Модель MapReduce; пути развития передовых аналитических инструментов; подходы к построению моделей для эффективного использования);

Тема 1.3 Консолидация, визуализация и подготовка данных (Организация хранилища данных. Хранилища данных и анализ. Визуализаторы общего назначения. Многомерная модель данных. OLAP-анализ. Качество данных. Сбор и интеграция данных. Виды источников данных. Методы интеграции данных. Преобразование и обогащение данных. Методы трансформации данных.);

Раздел 2 Интеллектуальный анализ данных и машинное обучение;

Тема 2.1 Методы и инструменты интеллектуального анализа данных (Задачи интеллектуального анализа данных (ИАД) и их классификация. Модели и методы ИАД. Цели, задачи и методы машинного обучения.);

Тема 2.2 Кластерный анализ (Введение в кластеризацию. Алгоритмы кластерного анализа.);

Тема 2.3 Классификация и регрессия (Метод опорных векторов. Метод k-ближайших соседей. Байесовский классификатор. Основы логистической регрессии. Правила классификации. Деревья решений, случайный лес. Простая линейная регрессия. Искусственные нейронные сети).

5 Перечень тем лекций

№ раздела / темы дисциплины	**Темы лекций**	**Трудоемкость, *академ.час***	
		всего	**в форме практической подготовки**
Раздел 1.	Введение в науку о данных		
Тема 1.1.	Основы анализа данных	2	
Тема 1.2.	Концепция Big Date	2	

	(Большие данные)		
Тема 1.3.	Консолидация, визуализация и подготовка данных	2	
Раздел 2.	Интеллектуальный анализ данных и машинное обучение		
Тема 2.1.	Методы и инструменты интеллектуального анализа данных	2	
Тема 2.2.	Кластерный анализ	2	
Тема 2.3.	Классификация и регрессия	6	
Итого:		**16**	**0**

6 Перечень тем практических занятий (семинаров)

№ раздела / темы дисциплины	**Темы практических занятий (семинаров)**	**Трудоемкость, *академ.час***	
		всего	**в форме практической подготовки**
Тема 1.3.	Консолидация, визуализация и подготовка данных	10	
Тема 2.2.	Кластерный анализ	8	
Тема 2.3.	Классификация и регрессия	14	
Итого:		**32**	**0**

7 Перечень тем лабораторных работ

№ раздела / темы дисциплины	**Темы лабораторных работ**	**Трудоемкость, *академ.час***	
		всего	**в форме практической подготовки**
	Отсутствуют		
Итого:		**0**	**0**

8 Перечень тем курсовых работ (проектов)

№ раздела / темы дисциплины	**Темы курсовых работ (проектов)**	**Трудоемкость, *академ.час***	
		всего	**в форме практической подготовки**
Раздел 1; Раздел 2.	1) Построение описательной модели и анализ ассоциаций по данным о клиентах организации предоставляющей услуги (мобильный оператор, интернет магазин,	36	

	розничная торговая сеть, банк, государственные органы и т.п.). 2) Построение описательной модели и сегментация клиентов организации предоставляющей услуги (мобильный оператор, интернет магазин, розничная торговая сеть, банк, страховая компания и т.п.). 3) Построение модели оценки качества конечного продукта на основе физико-химических свойств, протекания технологического процесса и т.п. 4) Построение предсказательной модели поведения клиента организации предоставляющей услуги (банк, издательство, мобильный оператор и т.д.). 5) Построение скоринговой модели для оценки риска невозврата кредита заемщиком. 6) Построение диагностической модели по данных предметной области (медицина, техническая диагностика и т.п.). 7) Построение рекомендательной системы для посетителей интернет ресурса.		
Итого:		**36**	**0**

9 Виды самостоятельной работы

№ раздела / темы дисциплины	**Виды самостоятельной работы**	**Трудоемкость, *академ.час***	
		всего	**в форме практической подготовки**
Раздел 1.	1. Изучение лекционного материала;	20	

	2. Подготовка к практическому занятию; 3. Прохождение тестирования.		
Раздел 2.	1. Изучение лекционного материала; 2. Подготовка к практическому занятию; 3. Прохождение тестирования.	40	
Курсовая работа	*Выполнение курсовой работы*	36	0
Контроль	*Подготовка к экзамену*	36	
Итого:		**132**	**0**

10 Матрица компетентностных задач по учебной дисциплине

Контролируемые разделы (темы) дисциплины	Контролируемые компетенции	Оценочные средства
Раздел 1.	ПК-2	1. Наблюдение за работой на практическом занятии; 2. Оценивание этапов выполнения курсовой работы; 3. Оценка участия в работе на лекционном занятии; 4. Тестовые задания
Раздел 2.	ПК-2	1. Наблюдение за работой на практическом занятии; 2. Оценивание этапов выполнения курсовой работы; 3. Оценка участия в работе на лекционном занятии; 4. Тестовые задания
Курсовая работа.	ПК-2	1. Вопросы к защите курсовой работы
Экзамен.	ПК-2	1. Экзамен на основе результатов текущего контроля

11 Входной контроль

Дисциплина «Математические и инструментальные методы анализа данных», изучаемая в 1 семестре, не подразумевает проведение входного контроля и рассчитана на обучающегося 1 курса, поступившего в университет в соответствии с Правилами приема в СибГИУ.

12 Текущий контроль

12.1 Текущий контроль знаний используется для оперативного и регулярного управления учебной деятельностью (в том числе

самостоятельной) обучающихся. Данный вид контроля стимулирует у обучающихся стремление к систематической самостоятельной работе по изучению учебной дисциплины и позволяет оценить степень освоения изученного материала.

Текущий контроль успеваемости осуществляется в течение семестра, в ходе повседневной учебной работы по индивидуальной инициативе педагогического работника, в том числе посредством использования СУО СибГИУ.

Формы проведения текущего контроля – наблюдение за работой на практическом занятии; оценивание этапов выполнения курсовой работы; оценка участия в работе на лекционном занятии; тестовые задания.

12.2 Описание оценочных средств

Шкалы оценивания (методики оценки)

12.2.1 Рекомендации по оцениванию этапов выполнения курсовой работы

Курсовая работа – конечный продукт, получаемый в результате планирования и выполнения комплекса учебных и исследовательских заданий. Позволяет оценить умения обучающихся самостоятельно конструировать свои знания в процессе решения практических задач и проблем, ориентироваться в информационном пространстве, а также уровень сформированности аналитических, исследовательских умений, умений практического и творческого мышления.

Оценивание этапов выполнения курсовой работы проводится в рамках текущей аттестации (три раза в семестр) в соответствии с графиком выполнения курсовой работы по учебной дисциплине. Результаты аттестации обучающихся по выполнению ими курсовой работы оцениваются руководителями в процентах от общего объема выполненной работы. Сроки выполнения курсовой работы устанавливаются кафедрой в графике выполнения курсовой работы. График выполнения курсовой работы выдается старостам групп в начале каждого семестра.

Критерии оценки % выполнения курсовой работы

Этап работы	% выполнения	Форма отчетности
Обозначена проблема и обоснована её актуальность, сделан краткий анализ различных точек зрения на рассматриваемую проблему, сформулирована цель работы, собраны исходные данные, сформирована обучающая и тестовая выборки	30	Выполненная часть курсовой работы загружается СУО СибГИУ

Выполнена очистка, преобразование, трансформация, фильтрация данных, построены и обучены модели на базе различных алгоритмов	60	Выполненная часть курсовой работы загружается СУО СибГИУ
Проведена оценка построенных моделей, проверена работоспособность моделей на тестовом множестве. Оформлена пояснительная записка к курсовой работы	100	Выполненная часть курсовой работы загружается СУО СибГИУ

Критерии оценки курсовой работы

Критерии оценки выполнения курсовой работы	Оценка
Задание выполнено полностью: цель достигнута; основные понятия выделены; имеются в наличии схемы, графическое выделение особо значимой информации; все выводы и рекомендации обоснованы, работа тщательно вычитана, отсутствуют грамматические и стилистические ошибки, работа выполнена в полном объёме; процент выполнения задания соответствует плановому.	отлично
Задание выполнено: цель выполнения достигнута; наличие правильных эталонных ответов; однако работа выполнена не в полном объёме, имеются единичные опечатки, орфографические, пунктуационные, грамматические, стилистические ошибки; график выполнения задания нарушен на 10 – 20 %.	хорошо
Задание выполнено частично: цель выполнения достигнута не полностью; многочисленные ошибки снижают качество выполненной работы; работа представлена с отклонением от запланированного срока на 50 %.	удовлетворительно
Задание не выполнено, цель работы не достигнута.	неудовлетворительно

12.2.2 Рекомендации по оцениванию результатов тестирования обучающихся

Тестирование проводится по изученным разделам (темам) учебной дисциплины для выявления уровня знаний и умений обучающегося, мотивирования обучающегося к активизации самостоятельной работы по усвоению учебного материала, выявления и устранения пробелов в знаниях, повышения дисциплины и организации учебной деятельности обучающегося.

Тестирование проводится в СУО СибГИУ в рамках часов, отведенных на самостоятельную работу обучающихся. Для тестирования используется 20 вопросов из базы типовых тестовых заданий по изученным разделам (темам) учебной дисциплины.

Обучающийся выбирает один ответ из предложенных вариантов ответа. В случае выбора неправильного ответа обучающийся получает за данный вопрос 0 баллов; в случае правильного ответа – 1 балл.

Оценка формируется в соответствии с критериями таблицы:

Максимальный балл	Проходной балл	Оценка
20	не менее 18	отлично
17	не менее 14	хорошо
13	не менее 10	удовлетворительно
9	–	неудовлетворительно

12.2.3 Рекомендации по оцениванию участия обучающихся в работе на лекционных занятиях

Формирование у обучающихся добросовестного и ответственного отношения к собственному профессиональному образованию проявляется в обязательном его присутствии на лекционном занятии.

На лекционных занятиях происходит формирование информационной компетентности обучающихся, в том числе способности фиксировать информацию с помощью конспектирования учебного материала. Преподаватель на лекционных занятиях акцентирует внимание обучающихся на том, как необходимо оформлять конспект лекций в тетради. По окончании лекционных занятий преподаватель осуществляет проверку тетрадей обучающихся и оценивает их.

Критерии оценки участия обучающихся в работе на лекционных занятиях

Наименование параметра оценки	Максимальное количество баллов за одно лекционное занятие
Наличие у обучающихся добросовестного и ответственного отношения к собственному профессиональному образованию, выраженное в посещении лекций	0,5
Умение конспектировать учебный материал в тетради, в том числе: - материал изложен последовательно, четко и аккуратно - в конспекте имеются изученные темы, основные термины, схемы	0,5
Всего	1

Если по результатам оценки участия обучающийся в лекционных занятиях в среднем за семестр обучающийся набрал не менее 0,5 баллов, то ему выставляется оценка «зачтено», менее 0,5 баллов – «не зачтено».

12.2.4 Рекомендации по оцениванию в процессе наблюдения за работой на практическом занятии

Наблюдение за работой на практическом занятии – важнейший компонент контроля деятельности обучающихся и качества их обучения.

Критерии оценки в процессе наблюдения за работой на практическом занятии:

– самостоятельность выполнения задания и работы с методическими указаниями (учитывается работа в течение занятия, быстрота и способность нахождения в методических указаниях нужной информации);

– правильность выполнения работы (учитывается последовательность выполняемых действий, умение работать с инструментами и/или средствами вычислительной техники);

– рациональность использования времени, отведенного на задание (не одобряется затянутость выполнения работы с учетом индивидуальных особенностей обучающихся);

– активность обучающегося при выполнении практической работы (учитывается индивидуальная работа и работа в группе).

Критерии оценки по результатам наблюдения за работой на практическом занятии

Наименование параметра оценки	Максимальное количество баллов
Самостоятельно выполняет задания; умеет работать с методическими указаниями.	2,5
Указанные в методических указаниях действия выполняет в правильной последовательности; умеет работать с инструментами и/или средствами вычислительной техники.	1,25
Рационально использует время, отведенное на задание; проявляет активность при выполнении практической работы.	1,25
Итого	**5**

Максимальное количество – 5 баллов соответствует оценке «отлично», 4 балла – «хорошо», 3 балла – «удовлетворительно», 2 балла – «неудовлетворительно».

Задания и вопросы для текущего контроля знаний обучающихся по разделам (темам) учебной дисциплины приведены в приложении.

13 Промежуточная аттестация

13.1 Оценочные материалы для промежуточной аттестации обучающихся по учебной дисциплине предназначены для оценки степени достижения запланированных результатов обучения по завершению изучения учебной дисциплины в установленной учебным

планом форме и позволяют определить качество усвоения изученного материала.

Итоговой формой контроля сформированности компетенций у обучающихся по учебной дисциплине является: курсовая работа; экзамен.

Оценочные материалы для промежуточной аттестации обучающихся включают: вопросы к защите курсовой работы; экзамен на основе результатов текущего контроля.

13.2 Оценивание обучающегося на защите курсовой работы

Оценка	Требования к знаниям
отлично	Оценка «отлично» выставляется обучающемуся, если выполнены все требования к выполнению и защите курсовой работы: соблюдены требования к внешнему оформлению, последовательно изложен материал; при защите обучающийся демонстрирует умения применять теоретические знания основной и дополнительной литературы, может показать и объяснить применение технологических решений, использованных в курсовой работе; даны правильные ответы на вопросы, заданные в процессе защиты курсовой работы.
хорошо	Оценка «хорошо» выставляется обучающемуся, если основные требования к выполнению и защите курсовой работы выполнены, но при этом допущены недочеты. В частности, имеются неточности в изложении материала; отсутствует логическая последовательность в решениях, имеются упущения в оформлении; на дополнительные вопросы при защите курсовой работы даны неполные ответы.
удовлетворительно	Оценка «удовлетворительно» выставляется обучающемуся, если имеются существенные отступления от требований к выполнению и защите курсовой работы. В частности: задача решена лишь частично; допущены фактические ошибки в содержании и оформлении; при ответе на дополнительные вопросы на защите курсовой работы обучающийся показывает знания только основного материала, испытывает затруднения при объяснении выбранных решений.
неудовлетворительно	Оценка «неудовлетворительно» выставляется обучающемуся, если задача решена лишь частично; допущены грубые ошибки в содержании и при оформлении или при ответе на дополнительные вопросы; во время защиты курсовой работы обучающийся обнаруживает существенное непонимание проблемы.

Если допущены существенные недостатки в оформлении курсовой работы, опущен или не написан какой-либо раздел, имеются отступления от задания на курсовую работу, то такая курсовая работа возвращается обучающемуся на доработку.

13.3 Оценивание обучающегося на экзамене

Оценка	Требования к знаниям
отлично	Оценка «отлично» выставляется обучающемуся, если он глубоко и прочно усвоил программный материал,

Оценка	Требования к знаниям
	исчерпывающе, последовательно, четко и логически его излагает, умеет тесно увязывать теорию с практикой, свободно справляется с задачами, вопросами и другими видами применения знаний, причем не затрудняется с ответом при видоизменении заданий, использует в ответе материал различной литературы, правильно обосновывает принятое нестандартное решение, обладает приемами выполнения практических задач по формированию компетенций.
хорошо	Оценка «хорошо» выставляется обучающемуся, если он твердо знает материал, грамотно и по существу излагает его, не допуская существенных неточностей в ответе на вопрос, правильно применяет теоретические положения при решении практических вопросов и задач, обладает необходимыми умениями и приемами их выполнения.
удовлетворительно	Оценка «удовлетворительно» выставляется обучающемуся, если он имеет знания только основного материала, но не усвоил его деталей, допускает неточности, недостаточно правильные формулировки, испытывает сложности при выполнении практических работ и затрудняется связать теорию вопроса с практикой.
неудовлетворительно	Оценка «неудовлетворительно» выставляется обучающемуся, который не знает значительной части программного материала, неуверенно отвечает, допускает серьезные ошибки, не имеет представлений по методике выполнения практической работы. Как правило, оценка «неудовлетворительно» ставится обучающимся, которые не могут продолжить обучение без дополнительных занятий по данной учебной дисциплине.

Задания и вопросы для промежуточной аттестации обучающихся приведены в приложении.

14 Учебно-методическое и информационное обеспечение учебной дисциплины

а) литература:

1 Болотова, Л. С. Системы поддержки принятия решений : учебник и практикум для вузов / Л. С. Болотова. — Москва : Издательство Юрайт, 2025. — 530 с. — ISBN 978-5-534-20422-3. – URL: https://urait.ru/bcode/558120 (дата обращения: 26.03.2026);

2 Станкевич, Л. А. Интеллектуальные системы и технологии : учебник и практикум для вузов / Л. А. Станкевич. — 2-е изд., перераб. и доп. — Москва : Издательство Юрайт, 2026. — 478 с. — (Высшее образование). — ISBN 978-5-534-20363-9. – URL: https://urait.ru/bcode/583592 (дата обращения: 26.03.2026);

3 Сапрыкин, О. Н. Интеллектуальный анализ данных : учебное пособие / О. Н. Сапрыкин. — Самара : Самарский университет, 2020. —

80 с. — ISBN 978-5-7883-1563-8. – URL: https://e.lanbook.com/book/188906 (дата обращения: 26.03.2026);

4 Васильев, Е. П. Интеллектуальный анализ данных в технологиях принятия решений : учебное пособие / Е. П. Васильев, В. И. Орешков. — Рязань : РГРТУ, 2023. — 180 с. — ISBN 978-5-7722-0344-6. – URL: https://e.lanbook.com/book/381533 (дата обращения: 26.03.2026);

5 Ростовцев, В. С. Искусственные нейронные сети : учебник для вузов / В. С. Ростовцев. — 5-е изд., стер. — Санкт-Петербург : Лань, 2025. — 216 с. — ISBN 978-5-507-50568-5. – URL: https://e.lanbook.com/book/447392 (дата обращения: 26.03.2026);

6 Митяков, Е. С. Искусственный интеллект и машинное обучение : учебное пособие для вузов / Е. С. Митяков, А. Г. Шмелева, А. И. Ладынин. — 2-е изд., стер. — Санкт-Петербург : Лань, 2026. — 252 с. — ISBN 978-5-507-51198-3. – URL: https://e.lanbook.com/book/507451 (дата обращения: 26.03.2026);

7 Баланов, А. Н. Искусственный интеллект. Понимание, применение и перспективы : учебник для вузов / А. Н. Баланов. — 3-е изд., стер. — Санкт-Петербург : Лань, 2026. — 312 с. — ISBN 978-5-507-55902-2. – URL: https://e.lanbook.com/book/512029 (дата обращения: 26.03.2026).

б) ресурсы информационно-телекоммуникационной сети «Интернет»:

1 Электронный каталог : сайт / Научно-техническая библиотека СибГИУ. – Новокузнецк, [199 –]. – URL: http://libr.sibsiu.ru. – URL: https://libr.sibsiu.ru;

2 Электронная библиотека // Научно-техническая библиотека СибГИУ : сайт. – Новокузнецк, [200 –]. – URL: http://library.sibsiu.ru/LibrELibraryFullText.asp. – Режим доступа: для авторизир. пользователей. – URL: https://library.sibsiu.ru/LibrELibraryFullText.asp;

3 Консультант студента : электронно-библиотечная система / ООО «КОНСУЛЬТАНТ СТУДЕНТА». – Москва, [200 –]. – URL: http://www.studentlibrary.ru. – Режим доступа: для авторизир. пользователей;

4 ЛАНЬ : электронно-библиотечная система : [коллекция «Единая профессиональная база знаний для технических вузов - Издательство Лань" ЭБС»] / ООО «Издательство ЛАНЬ». – Санкт-Петербург, [200 –]. – URL: http://e.lanbook.com. – Режим доступа: для авторизир. пользователей;

5 Образовательная платформа ЮРАЙТ / ООО «Электронное издательство ЮРАЙТ». – Москва, [200 –]. – URL: https://urait.ru. – Режим доступа: для авторизир. пользователей;

6 Университетская библиотека онлайн : электронно-библиотечная система / ООО «Директ-Медиа». – Москва, [200 –]. – URL:

https://biblioclub.ru. – Режим доступа: для авторизир. пользователей. – URL: http://www.biblioclub.ru;

7 НАУЧНАЯ ЭЛЕКТРОННАЯ БИБЛИОТЕКА eLIBRARY.RU : база данных / ООО «НЭБ». – Москва, [200 –]. – URL: http://elibrary.ru. – Режим доступа: по подписке;

8 Электронные периодические издания ИВИС : универсальная база данных / ООО «ИВИС». – Москва, [200 –]. – URL: http://eivis.ru. – Режим доступа: по подписке.

в) лицензионное и свободно распространяемое программное обеспечение:

– 7-Zip;
– Adobe Acrobat Reader;
– Astra Linux Special Edition;
– Deductor Academic;
– Kaspersky Endpoint Security;
– Loginom Community;
– Microsoft Office;
– Microsoft Windows;
– OnlyOffice;
– PyCharm;
– Р7-Офис.

г) базы данных и информационно-справочные системы:

1 ГАРАНТ : справочно-правовая система / ООО «Правовой центр «Гарант». – Кемерово, [200 –]. – Режим доступа: компьютерная сеть Сиб. гос. индустр. ун-та.;

2 КонсультантПлюс : справочно-правовая система / ООО «Информационный центр АНВИК». – Новокузнецк, [199 –]. – Режим доступа: компьютерная сеть библиотеки Сиб. гос. индустр. ун-та.;

3 Техэксперт : информационно-справочная система / ООО «Группа компаний «Кодекс». – Кемерово, [200 –]. – Режим доступа: компьютерная сеть Сиб. гос. индустр. ун-та.

15 Материально-техническое обеспечение учебной дисциплины

Материально-техническое обеспечение учебной дисциплины включает учебные аудитории, оснащенные оборудованием, компьютерной техникой, и техническими средствами обучения, в том числе:

- учебную аудиторию для проведения занятий лекционного типа, оборудованную учебной доской, экраном и мультимедийным проектором;
- учебную аудиторию для проведения занятий семинарского типа (практических занятий);
- учебную аудиторию (помещения) для групповых и индивидуальных консультаций, текущего контроля и промежуточной аттестации;

- помещения для самостоятельной работы, оснащенные компьютерной техникой с возможностью подключения к сети «Интернет» и обеспечением доступа в электронную информационно-образовательную среду, научно-техническую библиотеку СибГИУ;
- учебную аудиторию (помещения) для групповых и индивидуальных консультаций, текущего контроля и промежуточной аттестации

Рабочая программа дисциплины составлена в соответствии с требованиями ФГОС ВО по направлению подготовки (специальности) 09.04.03 «Прикладная информатика».

Составитель(и):

доцент Кораблина Татьяна Валентиновна (кафедра прикладных информационных технологий и программирования);
доцент Бабичева Надежда Борисовна (кафедра прикладных информационных технологий и программирования).

Рабочая программа дисциплины рассмотрена и утверждена на заседании кафедры.

Согласована:

Генеральный директор
Общества с ограниченной ответственностью
Научно-производственная фирма
"Радиан-М" (ООО НПФ "Радиан-М")
к.т.н., доцент  Е.Г. Руденкова

Приложение

по разделам (темам) учебной дисциплины

1 Тестирование по результатам освоения лекционного материала и самостоятельной работы

Тестирование проводится для определения степени усвоения материалов лекций, а также выполнения самостоятельной работы в рамках часов, отведенных на самостоятельную работу обучающихся, в СУО «Moodle».

Перечень вопросов для тестирования:

Раздел 1. Введение в науку о данных

1 Укажите современное определение анализа данных. Анализ данных – это … (*один ответ*)

а) обработка информации после ее сбора;

б) исследования, связанные со сбором, передачей и хранением данных;

в) средство проверки гипотез и решения задач исследователя.

2 Определите порядок шагов анализа данных в соответствии с современной концепцией:

№ шага	Описание шага
	сбор данные
	анализ (разведочный)
	постановка задачи
	анализ (подтверждающий)
	построение модели
	интерпретация
	получение результатов

3 Отправной точкой в разведочном анализе являются (*один ответ*):

а) модели;

б) данные;

в) предварительные сведения об исследуемой системе.

4 О ком идет речь? _____ – ключевая фигура в процессе анализа, он выдвигает гипотезы и для проверки их достоверности извлекает и визуализирует данные и (или) строи модели.

5 Что означает термин Knowledge Discovery in Databases (KDD)? (*один ответ*)

а) тиражирование знаний;

б) раскопка знаний;

в) извлечение знаний из баз данных;

г) разведочный анализ данных.

6 Основными факторами, которые привели к пересмотру подходов к анализу данных являются (*несколько ответов*):

а) доступность хранилищ данных;
б) большие объемы накопленных данных;
в) требование принятия решений в режиме реального времени;
г) необходимость обработки данных разнородного характера;
д) потребность в создании средств автоматического выделения и анализа скрытых данных;
е) сложный характер объектов управления;
ж) появление нового программного обеспечения.

7 Комплекс методов, используемых для создания моделей, способных самостоятельно обнаруживать в данных присущие им закономерности, называется (*один ответ*):

а) машинное обучение;
б) генетические алгоритмы;
в) нечеткая логика;
г) нейронные сети.

8 Расположите в порядке следования этапы обнаружения знаний:

№ шага	Описание шага
	применение методов Data Mining
	понимание и формулировка задачи
	подготовка данных
	интерпретация результатов
	проверка модели

9 На какие вопросы позволяют ответить описательные модели? (*несколько ответов*)

а) Какова структура клиентской базы?
б) Какой профиль хорошего для компании клиента?
в) Какие услуги будут интересны клиентам компании в следующем месяце?
г) Какой будет спрос на товар на следующий период планирования?

10 На какие вопросы позволяют ответить предсказательные модели? (*несколько ответов*)

а) Откликнется ли клиент на данную рекламную кампанию?
б) Какой размер прибыли получить компания в следующем году?
в) Какие есть связи между характеристиками клиента?
г) Какие события могут происходить совместно?

11 Отметьте неправильное высказывание по тематике Big Data

(*один ответ*).

а) Объем данных в сотни терабайт и петабайт не позволяет легко хранить и управлять ими с помощью традиционных реляционных баз данных.

б) Большинство всех Big Data являются структурированными.

в) Хранение очень больших объемов данных не всегда приводит к пониманию реальности.

г) Big Data, в контексте данных, обычно хранятся и организуются в распределенных файловых системах.

12 Предпосылки формирования тренда Больших данных являются (*несколько ответов*):

а) распространение сенсоров датчиков в различных сферах;

б) увеличение пропускной способности сетей;

в) развитие беспроводных сетей;

г) развитие и удешевление технологий хранения данных;

д) появление на рынке труда большого количества специалистов в области анализа данных;

е) появление настольных программных средств для обработки больших объемов данных.

13 Характеристики «4V» Big Data (*один ответ*):

а) объем, разнообразие, изменчивость, скорость;

б) объем, структура, стабильность, скорость;

в) пропускная способность, беспроводные технологии, облачные вычисления, структура.

14 Масштабирование по вертикали, как направление развития вычислительной инфраструктуры компании, предполагает (*один ответ*):

а) приобретение компьютеров с большим количеством процессоров, объемом оперативной памяти, дискового пространства и т.п.;

б) добавление дополнительных вычислительных узлов и распределение работы между ними;

в) приобретение компьютеров с большим количеством процессоров и добавление дополнительных вычислительных узлов.

15 MapReduce – это (*один ответ*)

а) специальная программа для обработки больших объемов данных алгоритмами Data Mining;

б) модель организации распределенных вычислений, используемая для параллельных вычислений над большими объемами данных в распределенный вычислительных сетях;

в) метод, используемый для хранения и обработки больших

объемов данных.

16 Какие действия выполняются в модели MapReduce на шагах Map и Reduce?

Действия	Шаг
Главный узлов (master node) получает входные данные задачи, разделяет их на части и передает рабочим узлам (worker) для предварительной обработки	
Главный узел получает ответы от рабочих узлов и на их основе формирует ответ	

17 Основными свойствами ХД являются (*несколько ответов*)

а) поддержание хронологии изменения данных;
б) сбор данных за большой период времени;
в) возможность интегрировать информацию, отражающую разные точки зрения на одну предметную область;
г) приведение данных к единому формату;
д) сбор данных за фиксированный промежуток времени;
е) возможность хранения данных в разных форматах;
ж) возможность интегрировать информацию, отражающую одну из точек зрения на предметную область.

18 Какие данные относятся к аддитивным, полуаддитивным, неаддитивным

Описание данных	данные
числовые данные, которые могут быть просуммированы по всем измерениям	
физические данные, которые не могут быть просуммированы ни по одному измерению	
числовые данные, которые могут быть просуммированы только по определенным измерениям	

19 Как загружаются данные в хранилище данных? (*один ответ*)

а) данные загружаются из многих источников данных регулярно;
б) данные вводятся пользователем в ручном режиме;
в) данные загружаются из одного источника однократно;
г) данные загружаются из одного источников регулярно

20 Основными достоинствами виртуального ХД являются? (*несколько ответов*)

а) упрощение процедур заполнения;
б) работа с текущими, детализированными данными;
в) минимизация объема памяти, занимаемой на носителе информацией;
г) проектирование для ответов на определенный круг вопросов.

21 Что из перечисленного является фактом, а что измерением?

Описание	Назначение
группа товаров	*факт / измерение*
количество проданного товаров из одной группы	*факт / измерение*
наименование товара	*факт / измерение*
сумма проданных товаров одной группы	*факт / измерение*
масса автомобиля марки №1	*факт / измерение*
марка автомобиля	*факт / измерение*
время разгона автомобиля марки №1 до 100 км в час	*факт / измерение*
время	*факт / измерение*
продукция	*факт / измерение*
количество полученных на складе телевизоров	*факт / измерение*
длина пройденного маршрута	*факт / измерение*

22 Какая операция обеспечивает изменение расположения измерений, представленных в отчете или на отображаемой странице? (*один ответ*)

а) конкатенция;
б) срез;
в) консолидация;
г) детализация;
д) вращение.

23 Вам необходимо проанализировать данные сразу по нескольким измерениям. Какой способ ее представления является наиболее удобным? (*один ответ*)

а) гиперкуб;
б) график;
в) гистограмма;
г) диаграмма;
д) таблица.

24 Основным назначение OLAP-систем является (*один ответ*):

а) визуализация данных;
б) сбор, хранение и анализ данных;
в) поддержка аналитической деятельности, произвольных запросов пользователей-аналитиков;

25 Для какой цели используется зоны временного хранения? (*один ответ*)

а) преобразование данных в процессе ELT;
б) хранение нормативно-справочной информации;
в) формирование оперативных отчетов;
г) хранение исторических данных.

26 Что из перечисленного относится к факторам качества данных? (*несколько ответов*)

а) непротиворечивость;
б) актуальность;
в) полнота;
г) точность;
д) тип;
е) объем;
ж) детализация

27 Какие требования предъявляются СППР по сравнению с OLTP системой?

Требование	Система
Быстрый поиск, простейшие алгоритмы обработки	*СППР/OLTP*
Аналитическая обработка с целью поиска скрытых закономерностей, построения прогнозов и моделей и т.д.	*СППР/OLTP*
Хранение только детализированных данных	*СППР/OLTP*
Хранение как детализированных, так и обобщенных данных	*СППР/OLTP*
Может содержать данные в разных форматах в зависимости от приложений	*СППР/OLTP*
Единый согласованный формат хранения данных	*СППР/OLTP*
Должна обеспечиваться максимальная нормализация	*СППР/OLTP*
Допускается контролируемая денормализация (избыточность) для эффективного извлечения данных	*СППР/OLTP*
Должна быть возможность в любое время добавлять, удалять и изменять данные	*СППР/OLTP*
Постоянная средняя загрузка процессора	*СППР/OLTP*
Основными приоритетами являются высокая производительность и доступность	*СППР/OLTP*

28 ELT-процесс включает (*один ответ*):

а) извлечение данных, преобразование данных, загрузка данных
б) обработка данных, загрузка данных, архивация данных
в) извлечения данных, передача данных, загрузка данных.

Раздел 2. Интеллектуальный анализ данных и машинное обучение

1 Число транзакций, которое содержит как условие, так и следствие – это (*один ответ*)

а) достоверность;
б) поддержка;
в) лифт;
г) улучшение;

д) левередж.

2 Мера точности правила – это (*один ответ*)

а) достоверность;
б) поддержка;
в) лифт;
г) улучшение;
д) левередж.

3 Обобщенная мера связи двух предметных наборов (*один ответ*)

а) достоверность;
б) поддержка;
в) лифт;
г) улучшение;
д) левередж.

4 Предметный набор с поддержкой больше заданного порога либо равной ему называется (*один ответ*)

а) популярный;
б) антимонопольный;
в) редкий;
г) уникальный.

5 Количество транзакций, содержащих и условие A и следствие B равно 78, общее количество транзакций 156. Определите поддержку и ввести в поле ввода _____.

6 Кластеризацию можно определить, как (*один ответ*)

а) группировка объектов по похожести их свойств; каждый кластер состоит из схожих объектов, а объекты разных кластеров существенно отличаются;

б) процедура, которая любому объекту $x \epsilon X$ ставит в соответствие метку набору $y \epsilon Y$;

в) исследование взаимной связи между событиями, которые происходят совместно;

г) установление закономерности между входными и выходными переменными.

7 Даны два объекта с признаками (1; 3) и (3; 3) соответственно. Определите между ними Евклидово расстояние и ввести в поле ввода _____.

8 При построении сети Кохонена «победителем» стал нейрон с весом 0,9 и значением признака 0,8. Определите новое значение веса для этого нейрона, если скорость обучения $\eta=0,5$ и ввести в поле ввода _____.

9 Карты Кохонена используются при решении задачи (*один ответ*)

а) кластеризации;
б) классификации;
в) поиска ассоциативных правил;
г) регрессии.

10 На карте Кохонена похожие объекты по рассматриваемому признаку будут иметь (*один ответ*)

а) одинаковый цвет;
б) различные цвета;
в) противоположные цвета по цветовой гамме;
г) уникальные цвета.

11 Если объекты на карте Кохонена расположены близко друг к другу, то (*один ответ*)

а) векторы признаков этих объектов сильно различаются;
б) векторы признаков этих объектов близки;
в) должна быть еще одна характеристика.

12 Каждой ячейке на карте Кохонена предназначен цвет (*один ответ*)

а) в соответствии со значениями признаков объектов в ней;
б) в соответствии с количеством признаков объекта в ней;
в) в соответствии с расстоянием между признаками объектов в ней.

13 На одной карте Кохонена можно использовать расцветку по ____ признаку(ам).

14 Объект имеет 4 признака. Сколько карт Кохонена необходимо построить для получения полной информации о значениях этих признаков? Ответ ввести в поле ввода _____.

15 В некоторую ячейку карты Кохонена попало два объекта со значениями признаков (45; 21) и (34; 14). Код цвета ячейки по первому признаку будет равен _____.

16 Задачу классификации и регрессии можно определить, как (*один ответ*)

а) группировка объектов по похожести их свойств; каждый кластер состоит из схожих объектов, а объекты разных кластеров существенно отличаются;
б) процедура, которая любому объекту $x \in X$ ставит в соответствие метку набору $y \in Y$;
в) исследование взаимной связи между событиями, которые происходят совместно;
г) установление закономерности между входными и выходными переменными.

17 К статистическим методам классификации и регрессии

относятся (*несколько ответов*):

а) линейная регрессия;
б) логистическая регрессия;
в) деревья решений;
г) нейронные сети;
д) метод k ближайших соседей.

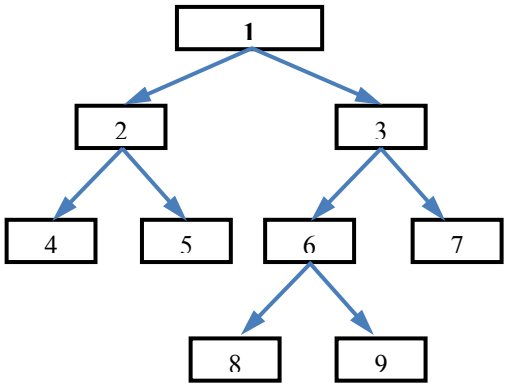
18 К методам машинного обучения для решения задачи классификации и регрессии относятся (*несколько ответов*):

а) линейная регрессия;
б) логистическая регрессия;
в) деревья решений;
г) нейронные сети;
д) метод k ближайших соседей.

19 Для набора данных, имеющих три входных атрибута (a1, a2, a3), и метку класса, принимающую два значения (к1, к2), выбирается наилучший атрибут ветвления. Прирост информации при использовании атрибута a1 для разбиения исходного подмножества составил 0,246 бит, атрибута a2 – 0,103 бит, атрибута a3 – 0,048 бит. Какой атрибут следует выбрать?

а) a2;
б) a1;
в) a3;
г) любой.

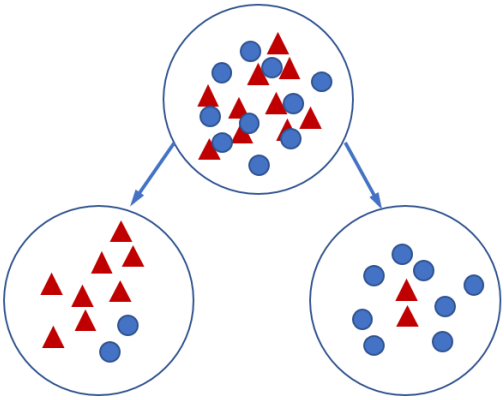
20 Какие из узлов дерева являются листьями (номера узлов записать через пробел) _____ .



20 Чистым называется такой узел, … (*один ответ*)

а) в который попали объекты, отнесенные к одному классу;
б) в который попали объекты, отнесенные к разным классам;
в) в который попало одинаковое число объектов каждого класса.

21 Для приведенного разбиения рассчитать прирост информации. Ответ округлить до двух знаков после запятой, например, 1,22. _____



22 На основании характеристик пяти объектов определить, какие два объекта находятся ближе друг к другу. При построении матрицы расстояний использовать евклидово расстояние. (Номера объектов и расстояние между ними записать через пробел, расстояние округлить до двух знаков после запятой, например, C D 1,22). _____

Номер объекта	*А*	*Б*	*В*
1	130	200	30
2	40	100	22
3	85	200	35
4	110	150	18
5	110	100	12

23 Алгоритм обратного распространения ошибки – градиентный алгоритм обучения многослойного персептрона, основанный на (*минимизации*, *максимизации*, *стабилизации*) (*среднеквадратической*, *среднемодульной*, *общей*) ошибки выходов.

Из группы в скобках выбрать по одному варианту.

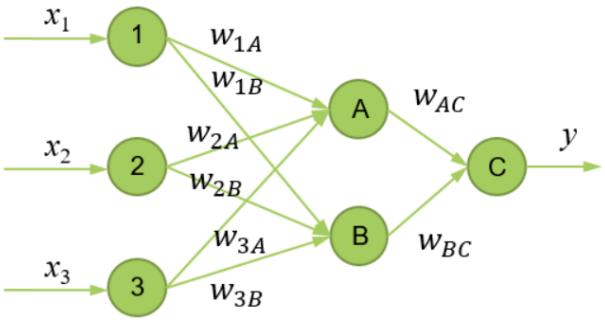
24 Определите порядок шагов процесса оптимизации при поиске оптимального состояния весов нейронной сети:

№ шага	**Описание шага**
	Вычисление градиента функции
	Достижение ближайшего минимума
	Определение точки, соответствующей новому состоянию весов
	Случайный выбор точки на рельефе
	Вычисление следующей точки

25 Какие факторы определяют работу нейронной сети (*несколько ответов*)

а) модель используемых нейронов;
б) вид активационной функции;
в) количество нейронов;
г) начальные значения весовых коэффициентов;
д) тип входных данных.

26 Для нейронной сети определить значений выходной переменной при заданных значениях входных переменных и весовых коэффициентов и логистической активационной функцией. Результат округлить до двух знаков после запятой



a=1

x_1=0,37 x_2=0,62 x_3=0,1

w_{1A}=-0,9 w_{1B}=0,82

w_{2A}=0,78 w_{2B}=0,56

w_{3A}=1,04 w_{3B}=0,55

w_{AC}=1,2 w_{BC}=0,4

27 Обучающее множество содержит 150 примеров. Какое число общих связей должно быть в сети? _____.

28 Обучающий набор имеет 150 примеров Нейронная сеть имеет 4 входных нейрона и 1 выходной. Сколько нейронов должно быть в одном скрытом слое сети? _____.

2 Выполнение практических работ

Тема 1.3 Консолидация, визуализация и подготовка данных

1 При оценке качества данных о месячных продажах имеется два выброса и одно пропущенное значение в марте месяце.

Используя Мастер *Редактирование выбросов* и Мастер *Заполнение пропусков* отредактируйте выбросы и замените пропущенное значение применяя различные методы. Сравните полученные результаты. Какие методы по вашему мнению больше подходят для исследуемых данных? Используя визуализатор, постройте графики.

2 Тренд ряда данных об объеме месячных продаж имеет сезонную составляющую. Подберите из имеющихся алгоритмов сглаживания обработки такой, который позволяет выделить полезную (медленную) составляющую сигнала с сезонностью наилучшим образом. С помощью визуализатора постройте графики.

3 Используя обработчик *Квантование*, разбить всех клиентов банка на группы в соответствии с размером взятого кредита. С помощью визуализатора *Куб* построить таблицу, используя в качестве факта количество взятых кредитов в соответствующем диапазоне, а в качестве измерений дату кредитования и метки диапазонов размера кредита.

4 В данных об абонентах сотовой компании необходимо обнаружить и заполнить пропуски, обнаружить и отредактировать выбросы. Перед заполнением пропусков и редактированием выбросов необходимо выполнить квантование данных о возрасте клиентов.

5 По имеющимся данным о показателях регионов определить степень влияния экономических показателей региона на среднедушевой

доход жителей, определите, какие параметры оказывают большее влияние на среднедушевой доход в регионе, а какие можно исключить из рассмотрения.

6 По данным о клиентах банка, выявить факторы, влияющие на возврат кредита. Представить результаты в виде таблицы и в виде матрицы факторного.

7 Аналитику необходимо оценить кредитоспособность потенциального заемщика. Предполагается, что заемщики, берущие суммы разного диапазона, ведут себя по-разному, следовательно, модели прогноза должны быть свои для каждой группы, т. е. для дальнейшего построения моделей оценки кредитоспособности различных категорий заемщиков необходимо использовать фильтрацию.

8 Произвести группировку для данных о продажах строительных материалов по сумме и количеству по городам.

9 Требуется проанализировать потребительскую корзину за четыре месяца и оценить изменение цен на продукты питания. Исходную таблицу необходимо привести к такому виду, чтобы в ней появились дополнительные поля, содержащие информация о цене рассматриваемых продуктов питания за каждый месяц в отдельности с помощью обработчика *Кросс-таблица*.

Тема 2.2 Кластерный анализ

1 Ирисы Фишера состоят из данных о 150 экземплярах ириса, по 50 экземпляров из трёх видов – Ирис щетинистый (*Iris setosa*), Ирис виргинский (*Iris virginica*) и Ирис разноцветный (*Iris versicolor*). На основании данных в соответствии с вариантом проведите кластеризацию по двум признакам ирисов, используя алгоритм k-means. Сравните результаты, полученные с помощью алгоритмов *k-means*, *g-means* и *EM кластеризации*. С помощью какого алгоритма получены более чистые кластеры?

2 Выполните кластеризацию по данным о параметрах цветков ириса в аналитической платформе Loginom , используя алгоритсы k-means, g-means, EM-кластеризации.

3 По данным абонентов телекоммуникационной компании провести сегментацию клиентов и оценить наиболее и наименее доходные сегменты. Что объединяет людей, попавших в каждый кластер? Дайте им характеристики. Результаты исследований оформите в виде таблиц. Установите принудительно 6 кластеров. Какие кластеры получены, как произведена группировка? В чем особенность кластеров? Исследуйте статистические характеристики кластеров.

Повторите исследования с 5 кластерами. Изучите статистические характеристики этой группы людей.

Используйте настроенную модель на новых данных. Проанализируйте полученные результаты. В какой кластер попал данный абонент?

4 Используя алгоритм *k-means* выполните кластеризацию по данным о состоянии банков. Проанализируйте статистически свойства кластеров, определите их мощность.

5 На основании данных в соответствии с вариантом провести кластеризацию магазинов по трем признакам: X_1 – площадь торгового зала, $м^2$, X_2 – товарооборот на одного продавца, ден. ед., X_3 – уровень рентабельности, %. Построить дендрограмму.

6 По данным об обеспеченности населения лечебными учреждениями регионов Российской Федерации провести сегментацию регионов и выяснить существуют ли заметные различия в их обеспеченности населения лечебными учреждениями с помощью алгоритмов *k-means*, *g-means* и *EM кластеризации*. Сколько и какие регионы попали в каждый кластер? Какие значения мощности амбулаторно-поликлинических учреждений, численность населения на одну больничную койку, обеспеченность врачами и средним медицинским персоналом? Для проведения анализа воспользуйтесь визуализаторами *Центы кластеров* и *Профили кластеров*.

7 По данным об обеспеченности регионов РФ высшими образовательными учреждениями выполнить кластеризацию, используя алгоритмы *k-means, g-means* и *EM кластеризации.* Сколько и какие регионы попали в каждый кластер? Выяснить имеются ли существенные различия в их обеспеченности образовательными учреждениями. В алгоритмах *k-means* и *EM кластеризации* задать *k*=5, исходя из предположения, что регионы могут быть отнесены к группам с лучшей, выше среднего, средний, ниже среднего и худшей обеспеченностью образовательными учреждениями.

8 Используя алгоритмы *k-means, g-means* и *EM кластеризации* выполнить кластеризацию по данным о состоянии банков. Проанализируйте статистически свойства кластеров, определите их мощность. Сравните результаты, полученные с помощью разных алгоритмов. Какой алгоритм по вашему мнению дает наилучшие результаты?

9 На основании данных о продажах в розничной сети (файл *Предметы бытовой химии.xlsx*) произвести анализ потребительской корзины с помощью кластеризации транзакций для выявления совместно приобретаемых товаров в супермаркете.

Сколько получилось кластеров? Какие у них характеристики: количество транзакций (*N*), ширина кластера (*W*) и мощность кластера (*S*)? Какой кластер самый мощный?

Используя визуализатор *Разбиение на кластеры* определите к какому кластеру относится транзакция.

Какие кластеры содержат клиентов, которые предпочитают только один вид товара?

10 Провести анализ потребительской корзины для выявления совместно покупаемых товаров в розничной сети, занимающейся

продажей продовольственных товаров.

Тема 2.3 Классификация и регрессия

1 На основании набора данных (в соответствии с вариантом), используя алгоритм «k ближайших соседей», определить к какому из трех классов: Ирис щетинистый (*Iris setosa*), Ирис виргинский (*Iris virginica*), Ирис разноцветный (*Iris versicolor*), относится пример последней записи в наборе данных. Задачу решить при k=1 и k=3. В качестве метрики для определения расстояния между записями использовать Евклидово расстояние и растояние Манхэттена.

2 В коммерческом банке имеется продукт «Нецелевой потребительский кредит»: займы предоставляются на любые цели с принятием решения за один день. В настоящее время решение о выдаче кредита принимается на основе скоринговой карты, построенной экспертным способом, с процентом отказа, равным 55 %, при этом объем просроченной задолженности велик. Накоплена статистическая информация о заемщиках и качестве обслуживания ими долга за несколько месяцев. Руководство банка, понимая, что высокий уровень отказов препятствует расширению розничного бизнеса в области потребительского кредитования, поставило перед отделом розничных рисков задачу разработать новую скоринговую карту, которая позволила бы значительно сократить число отказов в выдаче и снизить сумму просроченной задолженности.

3 Используя данные кредитных историй из файла с данными постройте скоринговую модель на основе логистическое регрессии. Определите оптимальный порог и порог, при котором чувствительность модели 90%.

Используйте настроенную модель на новых данных. Проанализируйте полученные результаты. В какой класс попали данные заемщики?

4 Торговая компания, осуществляющая продажу товаров, располагает информацией о своих клиентах и их покупках. Компания провела рекламную рассылку 13 504 клиентам и получила отклик в 14,5 % случаев. Необходимо построить модели отклика и проанализировать результаты, чтобы предложить способы минимизации издержек на новые почтовые рассылки.

3 Выполнение курсовой работы

В рамках учебной дисциплины предусмотрена самостоятельная работа обучающихся, результатом выполнения которой является оформленная в соответствии с требованиями пояснительная записка курсовой работы по заданной теме.

Примерный перечень тем курсовых работ представлен в рабочей программе дисциплины. Задание и методические указания к выполнению курсовой работы размещены в СУО «Moodle».

Перед выдачей темы курсовой работы обучающийся знакомится с графиком его выполнения и критериями оценки. Выполненная пояснительная записка курсовой работы загружается обучающимся в СУО «Moodle».

5 Вопросы к защите курсовой работы

1 Какая цель курсовой работы? Какие решались задачи в ходе выполнения курсовой работы?

2 В чем актуальность темы курсовой работы?

3 В чем заключается методология CRISP-DM?

4 В чем заключается проблема, для решения которой необходимо провести анализ данных в курсовой работе?

5 Какие рассмотрены точки зрения на рассматриваемую проблему?

6 Какие данные анализировались в курсовой работе?

7 Какой предварительной обработке были подвергнуты исходные данные при подготовке их к анализу?

8 Как изменилась выборка после предварительной обработки?

9 Какие этапы включает в себя интеллектуальный анализ данных?

10 Какие выбраны методы и алгоритмы для решения поставленных задач? Чем обоснован выбор?

11 Какие модели получены при решении поставленных задач?

12 Как проводилась проверка адекватности и корректности полученных моделей?

13 Как ведут себя модели на обучающей и контрольных выборках?

14 Какие инструменты использовались для визуализации полученных результатов?

15 Какие выводы могут быть сделаны на основании полученных результатов курсовой работы?

16 Какие классы моделей выделены в ИАД?

17 В каких сферах деятельности могут быть использованы большие данные?

18 В чем заключается задача поиска ассоциативных правил?

19 В чем заключается секвенциальный анализ?

20 В чем отличие поиска ассоциативных правил от секвенциального анализа?

21 Что такое поддержка и достоверность ассоциативного правила?

22 Какие этапы предполагает задача поиска ассоциативных правил?

23 В чем заключается идея алгоритма Apriori поиска ассоциативных правил?

24 В чем заключается задача кластеризации?

25 В чем отличие задачи кластеризации от задачи классификации?

26 Что используется в качестве меры «похожести» объектов при решении задачи кластеризации?

27 Какими способами можно представлять результаты кластеризации?

28 В чем заключается алгоритм k-means?

29 В чем заключаются проблемы алгоритмов кластеризации?

30 В чем заключается задача классификации?

31 В чем заключается задача регрессии?

32 В чем заключаются этапы решения задачи классификации и регрессии?

33 Какие алгоритмы используются для решения задачи классификации?

34 Из каких частей состоят классификационные правила?

35 Что из себя представляют деревья решений?

36 В чем заключается идея алгоритмов классификации (1R-алгоритм), (Naive Bayes), (ID3), (C4.5)?

37 При решении каких практических задач используются алгоритмы классификации и регрессии?

38 Структуры, свойства и принципы работы искусственных нейронных сетей.

39 Архитектуры искусственных нейронных сетей.

39 Процесс обучения нейронных сетей. Алгоритмы обучения нейронных сетей

.....