



**МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ
РОССИЙСКОЙ ФЕДЕРАЦИИ**

федеральное государственное бюджетное образовательное учреждение
высшего образования
**«ИРКУТСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ»
(ФГБОУ ВО «ИГУ»)**

Институт математики и информационных технологий
Кафедра вычислительной математики и оптимизации



Рабочая программа дисциплины (модуля)

Б1.В.ДВ.02.02 Программное обеспечение для анализа данных

Направление подготовки	01.03.02 Прикладная математика и информатика
Направленность (профиль) подготовки	Системная и бизнес-аналитика
Квалификация выпускника	бакалавр
Форма обучения	очная

Иркутск 2024 г.

1. ЦЕЛИ И ЗАДАЧИ ДИСЦИПЛИНЫ

Цели: освоение студентами современных программно-инструментальных средств для обработки и анализа больших массивов информации; приобретение методологических основ и практических навыков обработки аналитической информации.

Задачи:

- получение представления об основных понятиях, методах, моделях анализа и обработки данных и наиболее перспективных прикладных сферах их применения;
- освоение навыков работы с современными информационными технологиями анализа и обработки информации;
- получение опыта применения методов и способов построения эффективных алгоритмов и моделей для решения прикладных задач анализа и обработки данных;
- освоение основ участия в разработке и сопровождении информационных хранилищ, технологий оперативного и интеллектуального анализа данных, отражающих деятельность предприятий в различных предметных областях.

2. МЕСТО ДИСЦИПЛИНЫ В СТРУКТУРЕ ОПОП ВО

Учебная дисциплина Б1.В.ДВ.02.02 Программное обеспечение для анализа данных относится к части Блока 1 образовательной программы, формируемой участниками образовательных отношений.

Для изучения данной учебной дисциплины необходимы знания, умения и навыки, формируемые предшествующими дисциплинами: ПЕРЕЧИСЛИТЬ ДИСЦИПЛИНЫ.

Б1.О.26 Информатика и программирование

Б1.О.27 Пакеты компьютерной математики

Б1.О.28 Интернет-программирование

Перечень последующих учебных дисциплин, для которых необходимы знания, умения и навыки, формируемые данной учебной дисциплиной:

Б1.О.25 Эконометрика

Б1.В.03 Прикладная экономическая статистика

Б1.В.04 Большие данные

Б1.В.09 Управление рисками

Б2.В.01(У) Научно-исследовательская работа (получение первичных навыков научно-исследовательской работы)

3. ТРЕБОВАНИЯ К РЕЗУЛЬТАТАМ ОСВОЕНИЯ ДИСЦИПЛИНЫ

Процесс освоения дисциплины направлен на формирование следующих компетенций в соответствии с ФГОС ВО и ОП ВО по направлению подготовки 01.03.02 Прикладная математика и информатика:

ПК-2 Способен применять стандартное программное обеспечение в области управления проектами и при необходимости самостоятельно программировать нужные блоки на языках высокого уровня.

В результате освоения дисциплины обучающийся должен

- знать: возможности использования и основы работы со специализированными программным обеспечением для решения задач в сфере анализа данных;
- основные принципы работы современных средств анализа данных;

- основные принципы организации разработки проектов для анализа, трансформации, визуализации и обработки данных,
- основные различия между различными форматами данных;
- базовые принципы проектирования моделей анализа данных;

уметь:

- использовать специализированное программное обеспечение для решения задач сбора, обработки и анализа данных;
- проектировать и моделировать процессы анализа данных в различных предметных областях;

владеть:

- навыками моделирования в области эколого-экономических систем с использованием программного обеспечения для анализа данных;
- навыками фильтрации, обогащения, визуализации и статистического анализа данных с использованием специализированного программного обеспечения.
- постановкой и планированием разработки программных проектов для решения задач анализа данных.

4. СОДЕРЖАНИЕ И СТРУКТУРА ДИСЦИПЛИНЫ

Объем дисциплины составляет 2 зачетных ед., 72 час.

Форма промежуточной аттестации: зачет с оценкой, курсовая работа.

4.1. Содержание дисциплины, структурированное по темам, с указанием видов учебных занятий и отведенного на них количества академических часов

Раздел дисциплины / тема	Сем.	Виды учебной работы				Самост. работа	Формы текущего контроля; Формы промежут. аттестации
		Контактная работа преподавателя с обучающимися					
		Лекции	Лаб. занятия	Практ. занятия			
Тема 1. Анализ данных, основные понятия	4		2		4		
Тема 2. Первичная обработка данных статистического наблюдения с использованием табличных процессоров и средств программирования.	4		14		4		
Тема 3. Анализ и визуализация данных с использованием табличных процессоров и средств программирования.	4		16		4		
Тема 4. Основы работы с аналитической платформой KNIME.	4		16		4		
Итого (4 семестр):			48		16	зач.с оц., курс.раб.	

4.2. План внеаудиторной самостоятельной работы обучающихся по дисциплине

Раздел дисциплины / тема	Самостоятельная работа обучающихся			Оценочное средство	Учебно-методическое обеспечение самост. работы
	Вид самост. работы	Сроки выполнения	Затраты времени		
Тема 1. Анализ данных, основные понятия	1. Написание реферата по заданной теме.	За месяц до окончания курса	4	Реферат	ОЛ*-1, ДЛ**- 5 ИР*** –1,2,3, 4,5,6,7
Тема 2. Первичная обработка данных статистического наблюдения с использованием табличных процессоров и средств программирования.	1. Освоение базовых возможностей специализированного программного обеспечения. 2. Реализация контрольных примеров.	К окончанию выполнения практической работы по данной теме	4	Аналитический отчет (markdown)	ОЛ -1 ДЛ - 1,2 ИР – 1,2,3, 4,5,6,7

Тема 3. Анализ и визуализация данных с использованием табличных процессоров и средств программирования.	1. Освоение базовых возможностей специализированного программного обеспечения. 2. Реализация контрольных примеров.	К окончанию выполнения практической работы по данной теме	4	Аналитический отчет (markdown)	ОЛ - 1, 2 ДЛ - 1,2,3 ИР - 1,2,3, 4,5,6,7
Тема 4. Основы работы с аналитической платформой KNIME.	1. Освоение базовых возможностей специализированного программного обеспечения. 2. Реализация контрольных примеров.	К окончанию выполнения практической работы по данной теме	4	Аналитический отчет (markdown)	ОЛ - 1, 3 ДЛ - 3,4,5 ИР - 1,2,3, 4,5,6,7, 8
Общая трудоемкость самостоятельной работы (час.)			16		
Из них с использованием электронного обучения и дистанционных образовательных технологий (час.)			16		

*ОЛ – основная литература

** ДЛ – дополнительная литература

*** ИР – интернет - ресурсы

4.3. Содержание учебного материала

Тема1. Анализ данных, основные понятия.

Понятие данных. Структурированные и неструктурированные данные. Наука о данных. Этапы построения систем анализа данных. Основные источники данных. Основные понятия статистики данных.

Тема2. Первичная обработка данных статистического наблюдения с использованием табличных процессоров и средств программирования.

Абсолютные и относительные статистические величины, средние величины, вариационный анализ. Шкалы измерения. Функциональная, статистическая и корреляционная связи. Методы измерения тесноты парной корреляционной связи. Систематизация данных. Виды структурирования информации. Понятие описательных статистик. Меры среднего уровня. Анализ статистических связей. Тест хи-квадрат. Тест Стьюдента. Критерий Манна–Уитни. Дисперсионный анализ. Метод Краскела–Уоллиса. Срезы данных. Удаление и фильтрация данных по условию. Добавление Выравнивание и сортировка данных. Работа с индексами. Чистка данных. Агрегирование данных. Групповые таблицы.

Тема 3. Анализ и визуализация данных с использованием табличных процессоров и средств программирования.

Форматирование наборов и предварительная обработка данных в Microsoft Excel. Визуализация данных Описательная статистика. Сводные таблицы и сводные диаграммы в Microsoft Excel. Инструментарий в надстройке «Анализ данных» Построение графиков, статическая и интерактивная визуализации. Работа с библиотекой Matplotlib. Изменение масштаба. Нанесение рисок, меток и надписей. Добавление пояснительных надписей. Аннотации и рисование в подграфике. Использование символов и цветов. Сохранение

графиков в файле. Линейные и нелинейные графики функций, столбиковые диаграммы, гистограммы, графики плотностей распределения вероятностей, «ящички с усами» и диаграммы рассеяния.

Тема 4. Основы работы с аналитической платформой KNIME.

Базовые элементы KNIME: поток работ (workflow), рабочее пространство (workspace), узел (node), порт узла (port), статус узла (node status), соединение и настройка узла (connecting and configuring node). Получение распределенных данных с помощью REST обмена. Поиск шаблонов с помощью Knime. Деревья решений в Knime. Кластеризация в Knime. Модели прогнозирования в Knime.

4.3.1. Перечень семинарских, практических занятий и лабораторных работ

Тема занятия	Всего часов	Оценочные средства	Формируемые компетенции
Тема 1. Анализ данных, основные понятия	2	Реферат	ПК-2
Тема 2. Первичная обработка данных статистического наблюдения с использованием табличных процессоров и средств программирования.	14	Лабораторная работа	ПК-2
Тема 3. Анализ и визуализация данных с использованием табличных процессоров и средств программирования.	16	Лабораторная работа	ПК-2
Тема 4. Основы работы с аналитической платформой KNIME.	16	Лабораторная работа	ПК-2

4.3.2. Перечень тем (вопросов), выносимых на самостоятельное изучение студентами в рамках самостоятельной работы

Тема	Задание	Формируемые компетенции
Корреляционный анализ	1. Изучение основных алгоритмов. 2. Реализация контрольного примера с помощью среды Jupyter Notebook (Python).	ПК-2
Регрессионный анализ	1. Изучение основных алгоритмов. 2. Реализация контрольного примера с помощью среды RStudio.	ПК-2
Моделирование динамики популяций	1. Изучение численных алгоритмов для решения задач с дифференциальными уравнениями. 2. Реализация контрольного примера с помощью среды Jupyter Notebook (Python).	ПК-2
Методы принятия решений в условиях конфликта	1. Изучение задачи выбора одной из n возможных альтернатив. 2. Реализация контрольного примера с помощью среды Jupyter Notebook (Python).	ПК-2
Задача о назначениях	1. Изучение задачи о назначениях. 2. Реализация контрольного примера с помощью среды Jupyter Notebook (Python).	ПК-2

4.4. Методические указания по организации самостоятельной работы студентов

Самостоятельная работа студентов всех форм и видов обучения является одним из обязательных видов образовательной деятельности, обеспечивающей реализацию требований Федеральных государственных стандартов высшего образования. Согласно требованиям нормативных документов самостоятельная работа студентов является обязательным компонентом образовательного процесса, так как она обеспечивает закрепление получаемых на лекционных занятиях знаний путем приобретения навыков осмысления и расширения их содержания, навыков решения актуальных проблем формирования общекультурных и профессиональных компетенций, научно-исследовательской деятельности, подготовки к семинарам, лабораторным работам, сдаче зачетов и экзаменов. Самостоятельная работа студентов представляет собой совокупность аудиторных и внеаудиторных занятий и работ. Самостоятельная работа в рамках образовательного процесса в вузе решает следующие задачи:

- закрепление и расширение знаний, умений, полученных студентами во время аудиторных и внеаудиторных занятий, превращение их в стереотипы умственной и физической деятельности;
- приобретение дополнительных знаний и навыков по дисциплинам учебного плана;
- формирование и развитие знаний и навыков, связанных с научно-исследовательской деятельностью;
- развитие ориентации и установки на качественное освоение образовательной программы;
- развитие навыков самоорганизации;
- формирование самостоятельности мышления, способности к саморазвитию, самосовершенствованию и самореализации;
- выработка навыков эффективной самостоятельной профессиональной теоретической, практической и учебно-исследовательской деятельности.

Подготовка к лекции. Качество освоения содержания конкретной дисциплины прямо зависит от того, насколько студент сам, без внешнего принуждения формирует у себя установку на получение на лекциях новых знаний, дополняющих уже имеющиеся по данной дисциплине. Время на подготовку студентов к двухчасовой лекции по нормативам составляет не менее 0,2 часа.

Подготовка к практическому занятию. Подготовка к практическому занятию включает следующие элементы самостоятельной деятельности: четкое представление цели и задач его проведения; выделение навыков умственной, аналитической, научной деятельности, которые станут результатом предстоящей работы. Выработка навыков осуществляется с помощью получения новой информации об изучаемых процессах и с помощью знания о том, в какой степени в данное время студент владеет методами исследовательской деятельности, которыми он станет пользоваться на практическом занятии. Подготовка к практическому занятию нередко требует подбора материала, данных и специальных источников, с которыми предстоит учебная работа. Студенты должны дома подготовить к занятию 3–4 примера формулировки темы исследования, представленного в монографиях, научных статьях, отчетах. Затем они самостоятельно осуществляют поиск соответствующих источников, определяют актуальность конкретного исследования процессов и явлений, выделяют основные способы доказательства авторами научных работ ценности того, чем они занимаются. В ходе самого практического занятия студенты сначала представляют найденные ими варианты формулировки актуальности исследования, обсуждают их и обосновывают свое мнение о наилучшем варианте. Время на подготовку к практическому занятию по нормативам составляет не менее 0,2 часа.

Подготовка к семинарскому занятию. Самостоятельная подготовка к семинару направлена: на развитие способности к чтению научной и иной литературы; на поиск дополнительной информации, позволяющей глубже разобраться в некоторых вопросах; на выделение при работе с разными источниками необходимой информации, которая

требуется для полного ответа на вопросы плана семинарского занятия; на выработку умения правильно выписывать высказывания авторов из имеющихся источников информации, оформлять их по библиографическим нормам; на развитие умения осуществлять анализ выбранных источников информации; на подготовку собственного выступления по обсуждаемым вопросам; на формирование навыка оперативного реагирования на разные мнения, которые могут возникать при обсуждении тех или иных научных проблем. Время на подготовку к семинару по нормативам составляет не менее 0,2 часа.

Подготовка к коллоквиуму. Коллоквиум представляет собой коллективное обсуждение раздела дисциплины на основе самостоятельного изучения этого раздела студентами. Подготовка к данному виду учебных занятий осуществляется в следующем порядке. Преподаватель дает список вопросов, ответы на которые следует получить при изучении определенного перечня научных источников. Студентам во внеаудиторное время необходимо прочитать специальную литературу, выписать из нее ответы на вопросы, которые будут обсуждаться на коллоквиуме, мысленно сформулировать свое мнение по каждому из вопросов, которое они выскажут на занятии. Время на подготовку к коллоквиуму по нормативам составляет не менее 0,2 часа.

Подготовка к контрольной работе. Контрольная работа назначается после изучения определенного раздела (разделов) дисциплины и представляет собой совокупность развернутых письменных ответов студентов на вопросы, которые они заранее получают от преподавателя. Самостоятельная подготовка к контрольной работе включает в себя: — изучение конспектов лекций, раскрывающих материал, знание которого проверяется контрольной работой; повторение учебного материала, полученного при подготовке к семинарским, практическим занятиям и во время их проведения; изучение дополнительной литературы, в которой конкретизируется содержание проверяемых знаний; составление в мысленной форме ответов на поставленные в контрольной работе вопросы; формирование психологической установки на успешное выполнение всех заданий. Время на подготовку к контрольной работе по нормативам составляет 2 часа.

Подготовка к зачету. Самостоятельная подготовка к зачету должна осуществляться в течение всего семестра. Подготовка включает следующие действия: перечитать все лекции, а также материалы, которые готовились к семинарским и практическим занятиям в течение семестра, соотнести эту информацию с вопросами, которые даны к зачету, если информации недостаточно, ответы находят в предложенной преподавателем литературе. Рекомендуется делать краткие записи. Время на подготовку к зачету по нормативам составляет не менее 4 часов.

Подготовка к экзамену. Самостоятельная подготовка к экзамену схожа с подготовкой к зачету, особенно если он дифференцированный. Но объем учебного материала, который нужно восстановить в памяти к экзамену, вновь осмыслить и понять, значительно больше, поэтому требуется больше времени и умственных усилий. Важно сформировать целостное представление о содержании ответа на каждый вопрос, что предполагает знание разных научных трактовок сущности того или иного явления, процесса, умение раскрывать факторы, определяющие их противоречивость, знание имен ученых, изучавших обсуждаемую проблему. Необходимо также привести информацию о материалах эмпирических исследований, что указывает на всестороннюю подготовку студента к экзамену. Время на подготовку к экзамену по нормативам составляет 36 часов для бакалавров.

В ФБГОУ ВО «ИГУ» организация самостоятельной работы студентов регламентируется Положением о самостоятельной работе студентов, принятым Ученым советом ИГУ 22 июня 2012 г.

5. УЧЕБНО-МЕТОДИЧЕСКОЕ И ИНФОРМАЦИОННОЕ ОБЕСПЕЧЕНИЕ ДИСЦИПЛИНЫ

а) основная литература:

1. Бессмертный, И. А. Интеллектуальные системы : учебник и практикум для вузов / И. А. Бессмертный, А. Б. Нугуманова, А. В. Платонов. — Москва : Издательство Юрайт, 2022. — 243 с. — (Высшее образование). — ISBN 978-5-534-01042-8. — Текст : электронный // Образовательная платформа Юрайт [сайт]. — URL: <https://www.ura.it.ru/bcode/490020> (дата обращения: 22.05.2022).
2. Яковлев, В. Б. Статистика. Расчеты в Microsoft Excel : учебное пособие для вузов / В. Б. Яковлев. — 2-е изд., испр. и доп. — Москва : Издательство Юрайт, 2022. — 353 с. — (Высшее образование). — ISBN 978-5-534-01672-7. — Текст : электронный // Образовательная платформа Юрайт [сайт]. — URL: <https://www.ura.it.ru/bcode/491936> (дата обращения: 23.05.2022).
3. Моделирование систем и процессов. Практикум : учебное пособие для вузов / В. Н. Волкова [и др.] ; под редакцией В. Н. Волковой. — Москва : Издательство Юрайт, 2022. — 295 с. — (Высшее образование). — ISBN 978-5-534-01442-6. — Текст : электронный // Образовательная платформа Юрайт [сайт]. — URL: <https://www.ura.it.ru/bcode/490226> (дата обращения: 23.05.2022).

б) дополнительная литература:

1. Федоров, Д. Ю. Программирование на языке высокого уровня Python : учебное пособие для вузов / Д. Ю. Федоров. — 3-е изд., перераб. и доп. — Москва : Издательство Юрайт, 2022. — 210 с. — (Высшее образование). — ISBN 978-5-534-14638-7. — Текст : электронный // Образовательная платформа Юрайт [сайт]. — URL: <https://www.ura.it.ru/bcode/492920> (дата обращения: 30.05.2022).
2. Воронов, М. В. Системы искусственного интеллекта: учебник и практикум для вузов / М. В. Воронов, В. И. Пименов, И. А. Небаев. — Москва : Издательство Юрайт, 2022. — 256 с. — (Высшее образование). — ISBN 978-5-534-14916-6. — Текст : электронный // Образовательная платформа Юрайт [сайт]. — URL: <https://www.ura.it.ru/bcode/485440> (дата обращения: 30.05.2022).
3. Ахременко, А. С. Политический анализ и прогнозирование в 2 ч. Часть 1 : учебник и практикум для вузов / А. С. Ахременко. — 2-е изд., испр. и доп. — Москва : Издательство Юрайт, 2022. — 180 с. — (Высшее образование). — ISBN 978-5-534-07223-5. — Текст : электронный // Образовательная платформа Юрайт [сайт]. — URL: <https://www.ura.it.ru/bcode/490122> (дата обращения: 23.05.2022).
4. Гниденко, И. Г. Технологии и методы программирования : учебное пособие для вузов / И. Г. Гниденко, Ф. Ф. Павлов, Д. Ю. Федоров. — Москва : Издательство Юрайт, 2022. — 235 с. — (Высшее образование). — ISBN 978-5-534-02816-4. — Текст : электронный // Образовательная платформа Юрайт [сайт]. — URL: <https://www.ura.it.ru/bcode/489920> (дата обращения: 30.05.2022).
5. Информационные технологии в менеджменте : учебник и практикум для вузов / Е. В. Майорова [и др.] ; под редакцией Е. В. Майоровой. — Москва : Издательство Юрайт, 2022. — 368 с. — (Высшее образование). — ISBN 978-5-534-00503-5. — Текст : электронный // Образовательная платформа Юрайт [сайт]. — URL: <https://www.ura.it.ru/bcode/489927> (дата обращения: 22.05.2022).

в) базы данных, информационно-справочные и поисковые системы:

1. Научная электронная библиотека «ELIBRARY.RU» [Электронный ресурс] : сайт. — Режим доступа: <http://elibrary.ru/defaultx.asp>
2. Открытая электронная база ресурсов и исследований «Университетская информационная система РОССИЯ» [Электронный ресурс] : сайт. — Режим доступа: <http://uisrussia.msu.ru>

3. Государственная информационная система «Национальная электронная библиотека» [Электронный ресурс] : сайт. – Режим доступа: <http://нэб.рф>
4. В соответствии с п. 4.3.4. ФГОС ВО, обучающимся в течение всего периода обучения обеспечен неограниченный доступ (удаленный доступ) к электронно-библиотечным системам:
 - ЭБС «Издательство Лань». ООО «Издательство Лань». Контракт № 92 от 12.11.2018 г. Акт от 14.11.2018 г.
 - ЭБС ЭЧЗ «Библиотех». Государственный контракт № 019 от 22.02.2011 г. ООО «Библиотех». Лицензионное соглашение № 31 от 22.02.2011 г. Адрес доступа: <https://isu.bibliotech.ru/> Срок действия: с 22.11.2011 г. бессрочный.
 - ЭБС «Национальный цифровой ресурс «Рукопт». ЦКБ «Бибком». Контракт № 91 от 12.11.2018 г. Акт от 14.11.2018 г..
 - ЭБС «Айбукс.ру/ibooks.ru». ООО «Айбукс». Контракт № 90 от 12.11.2018 г. Акт № 54 от 14.11.2018 г.
 - Электронно-библиотечная система «ЭБС Юрайт». ООО «Электронное издательство Юрайт». Контракт № 70 от 04.10.2018 г.

г) интернет-ресурсы

1. <http://wombat.org.ua/AByteOfPython/AByteofPythonRussian-2.01.pdf> - это свободная книга по программированию на языке Python.
2. <https://www.python.org/> - язык программирования Python
3. www.gks.ru - Официальный сайт Государственного комитета статистики РФ
4. sophist.hse.ru - Единый архив экономических и социологических данных
5. http://library.hse.ru/e-resources/HSE_economic_journal - Экономический журнал Высшей школы экономики
6. www.cemi.rssi.ru - Центральный экономико-математический институт (ЦЭМИ) РАН
7. www.forecast.ru - Центр макроэкономического анализа и прогнозирования при ИПП РАН
8. <https://www.knime.com/> - аналитическая платформа Knime

6. МАТЕРИАЛЬНО-ТЕХНИЧЕСКОЕ ОБЕСПЕЧЕНИЕ ДИСЦИПЛИНЫ

6.1. Учебно-лабораторное оборудование

ЭТОТ РАЗДЕЛ НЕ ЗАПОЛНЯТЬ

6.2. Программное обеспечение

1. Язык программирования Python 3.7. и выше (открытое программное обеспечение).
2. Jupyter Notebook (открытое программное обеспечение).
3. Jupyter Lab (открытое программное обеспечение).
4. VisualStudioCode (Microsoft, открытое программное обеспечение)
5. Офисный пакет Microsoft Office Project Professional 2019 (лицензия ИГУ для образовательных учреждений).
6. Редакционно-издательская система MikTeX (открытое программное обеспечение).
7. Аналитическая платформа Knime.

7. ОЦЕНОЧНЫЕ МАТЕРИАЛЫ ДЛЯ ТЕКУЩЕГО КОНТРОЛЯ И ПРОМЕЖУТОЧНОЙ АТТЕСТАЦИИ

7.1. Оценочные средства текущего контроля

Вид контроля	Контролируемые темы	Контролируемые компетенции
Реферат	Тема 1. Анализ данных, основные понятия	ПК-2
Аналитический отчет (markdown)	Тема 2. Первичная обработка данных статистического наблюдения с использованием табличных процессоров и средств программирования. Тема 3. Анализ и визуализация данных с использованием табличных процессоров и средств программирования. Тема 4. Основы работы с аналитической платформой KNIME.	ПК-2

Примеры оценочных средств текущего контроля

Лабораторная работа № 1. Поиск шаблонов

Количество баллов: 5.

Цель. Построение потока работ, выполняющего решение задачи анализа рыночной корзины и поиска ассоциативных правил. Данный поток должен выполнять следующую последовательность действий: загрузить данные из текстового файла, преобразовать загруженные данные в специализированный тип данных пакета KNIME, найти частые наборы и ассоциативные правила, вывести результаты.

Создайте поток работ, приведенный на Рис. 2.

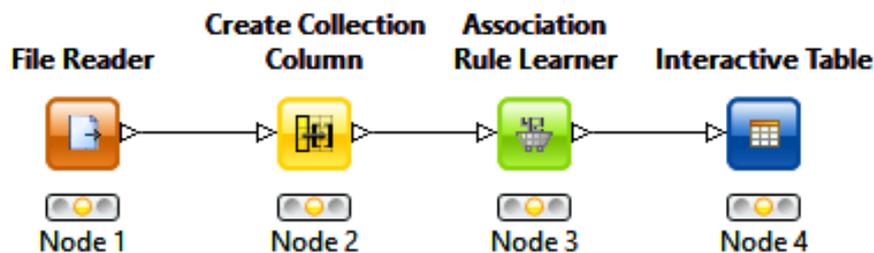


Рис. 2. Поток работ для решения задачи анализа рыночной корзины

Выполните настройку узла “File Reader” так, чтобы:

в качестве файла исходных данных фигурировал baskets.csv (см. архив baskets.zip, прилагаемый к заданию); первая строка файла трактовалась как содержащая названия столбцов (“Read column headers”); в качестве разделителя столбцов фигурировала запятая; узел обрабатывал неполные строки (кнопка “Advanced”, вкладка “Short lines”).

Выполните настройку узла “Create Collection Column” так, чтобы:

все строки исходного файла попали в выходную коллекцию данных; установите флаги “Create collection of type set”, “Ignore missing values”, “Remove aggregated columns from table”.

Выполните поток работ, предварительно настроив узел “Association Rule Learner”, указав различные значения поддержки (minimum support). Объясните полученные результаты (как данный параметр влияет на решение задачи?). Создайте скриншоты потока работ и результатов его работы для использования в качестве отчета о выполнении задания.

Лабораторная работа № 2. Деревья решений

Количество баллов: 5.

Цель. Построение потока работ, выполняющего решение задачи классификации посредством построения дерева решений. Данный поток должен выполнять следующую последовательность действий: загрузить данные из текстового файла, построить дерево

Создайте поток работ, приведенный на Рис. 2.

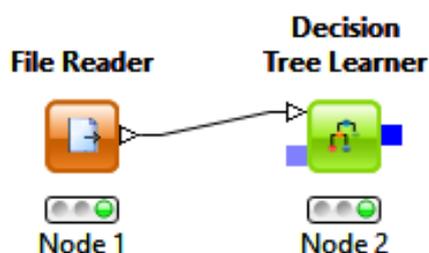


Рис. 3. Поток работ для решения задачи классификации

Выполните настройку узла “File Reader” так, чтобы:

в качестве файла исходных данных фигурировал marks.csv; первая строка файла трактовалась как содержащая названия столбцов (“Read column headers”);

Выполните настройку узла “Decision Tree Learner” так, чтобы поле FINALMARK трактовалось как признак класса. Выполните поток работ. Убедитесь, что построенное дерево решений показывает зависимость итоговой оценки только от атрибутов, отражающих мнение учителя (имеющих название вида TEACHER_xx), и ее независимость от атрибутов, отражающих персональные данные ученика (имеющих название вида PUPIL_xx). Создайте скриншоты потока работ и результатов его работы для использования в качестве отчета о выполнении задания.

Факультативное задание (1 балл)

Измените созданный поток работ, как показано на Рис. 4.

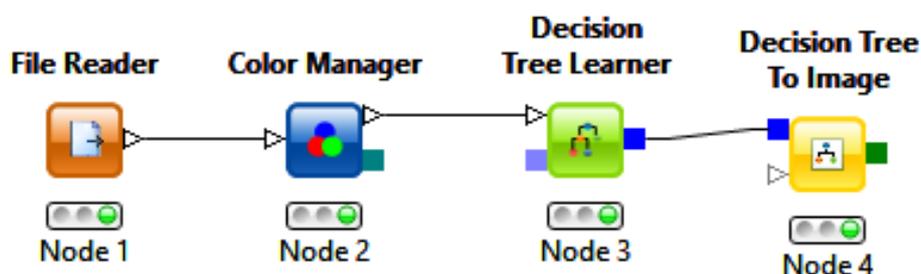


Рис. 4. Поток работ для решения задачи классификации с дополнительными узлами

Настройте узел “Color Manager”, указав столбец FINALMARK как раскрашенный. Выполните поток работ. Сравните результаты при различных настройках узла “Decision Tree Learner” (контекстное меню узла): “Decision Tree View” и “Decision Tree View (simple)”. Создайте скриншоты потока работ и результатов его работы для использования в качестве отчета о выполнении задания.

Лабораторная работа № 3. Кластеризация

Количество баллов: 5.

Цель. Построение потока работ, выполняющего решение задачи кластеризации. Данный поток должен выполнять следующую последовательность действий: загрузить данные из текстового файла, построить дерево решений, вывести результаты.

Создайте поток работ, приведенный на Рис. 5.

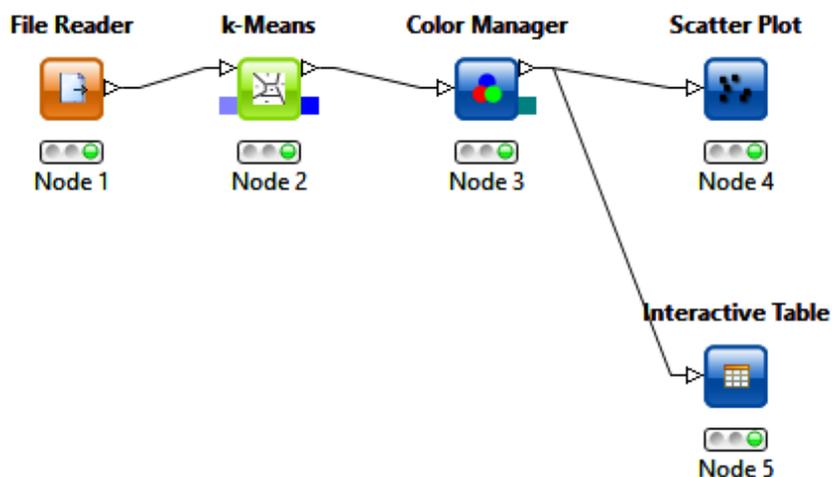


Рис. 5. Поток работ для решения задачи кластеризации

Выполните настройку узла “File Reader” так, чтобы в качестве файла исходных данных фигурировал basketball.csv. Выполните поток работ. Сравните результаты при различных параметрах: количество кластеров, цвета для отображения кластеризуемых объектов. Создайте скриншоты потока работ и результатов его работы для использования в качестве отчета о выполнении задания.

7.2. Оценочные средства для промежуточной аттестации

Список вопросов для промежуточной аттестации:

1. Структурированные и неструктурированные данные.
2. Этапы построения систем анализа данных.
3. Первичная обработка данных статистического наблюдения.
4. Методы измерения тесноты парной корреляционной связи.
5. Индексный метод исследования данных.
6. Обзор современных информационных технологий и программных средств анализа и обработки данных.
7. Языки программирования для анализа данных R, Python.
8. Системы анализа данных.
9. Аналитическая платформа Deductor.
10. Аналитические информационные системы: состав, назначение, архитектура.
11. Основные принципы построения информационных хранилищ.
12. Типы систем машинного обучения.
13. Задачи машинного обучения: классификация.
14. Задачи машинного обучения: кластеризация.
15. Задачи машинного обучения: поиск аномальных значений.

16. Задачи машинного обучения: регрессия.

Примеры оценочных средств для промежуточной аттестации:

Комплект тестовых заданий

1. Аналитик это ...
 - а) специалист в области анализа и моделирование
 - б) специалист в предметной области;
 - в) человек, решающий определенные задачи;
 - г) человек, который имеет опыт в программировании.
2. Переменная X измерена в порядковой шкале. Результаты измерений этой переменной можно представить в:
 - а) номинальной шкале измерений
 - б) количественной шкале измерений
 - г) нельзя представить в какой-либо другой шкале измерений
3. Статистика как наука изучает:
 - а) массовые явления;
 - б) единичные явления;
 - в) периодические события.
4. Закон больших чисел утверждает, что:
 - а) чем больше единиц охвачено статистическим наблюдением, тем лучше проявляется общая закономерность;
 - б) чем больше единиц охвачено статистическим наблюдением, тем хуже проявляется общая закономерность;
 - в) чем меньше единиц охвачено статистическим наблюдением, тем лучше проявляется общая закономерность.
5. Назовите основные виды ошибок регистрации: а) случайные; б) систематические; в) ошибки репрезентативности; г) расчетные:
 - а) а, б, в,
 - б) а, б
 - в) а
 - г) а, б, в, г
6. В каких единицах будет выражаться относительный показатель, если база сравнения принимается за единицу?
 - а) в коэффициентах;
 - б) в натуральных;
 - в) в процентах.
7. Значения признака, повторяющиеся с наибольшей частотой, называется
 - а) модой;
 - б) медианой.
8. Коэффициент вариации измеряет колеблемость признака в:
 - а) относительном выражении;
 - б) абсолютном выражении.
9. Абсолютный прирост исчисляется как: а) отношение уровней ряда; б) разность уровней ряда. Темп роста исчисляется как: в) отношение уровней ряда; г) разность уровней ряда:
 - а) б, в
 - б) а, в
 - в) а, г
10. К наиболее простым методам прогнозирования относят:
 - а) метод на основе среднего абсолютного прироста;
 - б) метод скользящей средней;
 - в) индексный метод.
11. Статистический индекс - это:

- а) относительная величина сравнения двух показателей;
 - б) сравнительная характеристика двух абсолютных величин;
 - в) критерий сравнения относительных величин.
12. Табличные процессоры позволяют строить диаграммы следующих типов:
- а) Гистограмма, линейчатая диаграмма, круговая диаграмма, график;
 - б) Диаграмма с областями, поверхностная диаграмма, лепестковая диаграмма;
 - и) Всё выше перечисленное;
13. Задача классификации сводится к ...
- а) определения класса объекта по его характеристиками;
 - б) нахождения частых зависимостей между объектами или событиями;
 - в) определение по известным характеристиками объекта значение некоторого его параметра;
 - г) поиска независимых групп и их характеристик в всем множестве анализируемых данных.
14. Задача регрессии сводится к ...
- а) определение по известным характеристиками объекта значение некоторого его параметра;
 - б) определения класса объекта по его характеристиками;
 - в) нахождения частых зависимостей между объектами или событиями;
 - г) поиска независимых групп и их характеристик в всем множестве анализируемых данных.
15. Задача кластеризации заключается в ...
- а) поиска независимых групп и их характеристик в всем множестве анализируемых данных;
 - б) определения класса объекта по его характеристиками;
 - в) определение по известным характеристиками объекта значение некоторого его параметра;
 - г) нахождения частых зависимостей между объектами или событиями.
16. Целью поиска ассоциативных правил является ...
- а) нахождения частых зависимостей между объектами или событиями;
 - б) определения класса объекта по его характеристиками;
 - в) определение по известным характеристиками объекта значение некоторого его параметра;
 - г) поиска независимых групп и их характеристик в всем множестве анализируемых данных.
17. Метаданные - ...
- а) высокоуровневые средства отражения информационной модели и описания структуры данных
 - б) разновидность систем хранения, ориентирована на поддержку процесса анализа данных целостность, обеспечивает, непротиворечивость и хронологию данных, а также высокую скорость выполнения аналитических запросов
 - в) некоторый набор операций над базой данных, который рассматривается как единственное завершено, с точки зрения пользователя, действие над некоторой информацией, обычно связано с обращением к базе данных
 - г) это установление зависимости дискретной выходной переменной от входных переменных
18. Классификация - ...
- а) это установление зависимости дискретной выходной переменной от входных переменных
 - б) разновидность систем хранения, ориентирована на поддержку процесса анализа данных целостность, обеспечивает, непротиворечивость и хронологию данных, а также высокую скорость выполнения аналитических запросов
 - в) высокоуровневые средства отражения информационной модели и описания структуры данных
 - г) некоторый набор операций над базой данных, который рассматривается как единственное завершено, с точки зрения пользователя, действие над некоторой информацией, обычно связано с обращением к базе данных
19. Регрессия - ...
- а) это установление зависимости непрерывной выходной переменной от входных переменных
 - б) эта группировка объектов (Наблюдений, событий) на основе данных, описывающих свойства объектов
 - в) выявление закономерностей между связанными событиями
 - г) это установление зависимости дискретной выходной переменной от входных переменных
20. Кластеризация - ...
- а) это установление зависимости непрерывной выходной переменной от входных переменных

- б) эта группировка объектов (Наблюдений, событий) на основе данных, описывающих свойства объектов
- в) выявление закономерностей между связанными событиями
- г) это установление зависимости дискретной выходной переменной от входных переменных.

21. Ассоциация - ...

- а) эта группировка объектов (наблюдений, событий) на основе данных, описывающих свойства объектов
- б) это установление зависимости непрерывной выходной переменной от входных переменных
- в) выявление закономерностей между связанными событиями
- г) это установление зависимости дискретной выходной переменной от входных переменных

22. Машинное обучение - ...

- а) подразделение искусственного интеллекта изучающий методы построения алгоритмов, способных обучаться на данных
- б) эта группировка объектов (Наблюдений, событий) на основе данных, описывающих свойства объектов
- в) набор данных, каждая запись которого представляет собой учебный пример, содержащего заданный входной влияние, что и отвечает ему правильный выходной результат.
- г) специализированный программный решение (или набор решений), который включает в себя все инструменты для извлечения закономерностей из сырых данных

23. Аналитическая платформа - ...

- а) специализированный программный решение (или набор решений), который включает в себя все инструменты для извлечения закономерностей из сырых данных
- б) эта группировка объектов (Наблюдений, событий) на основе данных, описывающих свойства объектов
- в) набор данных, каждая запись которого представляет собой учебный пример, содержащего заданный входной влияние, что и отвечает ему правильный выходной результат.
- г) подразделение искусственного интеллекта изучающий методы построения алгоритмов, способных обучаться на данных

24. Обучающая выборка - ...

- а) набор данных, каждая запись которого представляет собой учебный пример, содержащего заданный входной влияние, и соответствующий ему правильный выходной результат
- б) эта группировка объектов (Наблюдений, событий) на основе данных, описывающих свойства объектов
- в) набор данных, каждая запись которого представляет собой учебный пример, содержащего заданный входной влияние, что и отвечает ему правильный выходной результат.
- г) выявление в сырых данных ранее неизвестных, нетривиальных, практически полезных и доступных интерпретации знаний, необходимых для принятия решений в различных сферах человеческой деятельности

25. Ошибка обучения - ...

- а) это ошибка, допущенная моделью на учебной множества.
- б) это ошибка, полученная на тестовых примерах, то есть, что вычисляется по тем же формулам, но для тестовой множества
- в) имена, типы, метки и назначения полей исходной выборки данных
- г) набор данных, каждая запись которого представляет собой учебный пример, содержащего заданный входной влияние, и соответствующий ему правильный выходной результат

Примеры заданий:

Тесты и задания в ЭИОС ИГУ на сайте <https://educa.isu.ru/>

Разработчик: Кедрин В.С., к.т.н., доцент, доцент