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

РАБОЧАЯ ПРОГРАММА ДИСЦИПЛИНЫ (МОДУЛЯ)

Модели и методы интеллектуального анализа данных

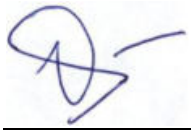
Направление подготовки *«(09.04.02) Информационные системы и технологии»*

Направленность *«Руководство разработкой программного обеспечения»*

Квалификация (степень) выпускника: *магистр*

Кострома

Рабочая программа дисциплины «Модели и методы интеллектуального анализа данных» разработана в соответствии с Федеральным государственным образовательным стандартом по направлению 09.04.02 Информационные системы и технологии (уровень магистратуры), утвержден приказом Министерства образования и науки РФ № 917 от 19.09.17.

Разработал:  Денисов А.Р., д.т.н., доцент

подпись

Рецензент:  Панин И.Г., д.т.н., доцент

подпись

УТВЕРЖДЕНО:

На заседании кафедры Информационных систем и технологий
Протокол заседания кафедры № 9 от 14.06.2019 г.
Заведующий кафедрой Информационных систем и технологий

 Киприна Л.Ю., к.т.н., доцент

Подпись

ПЕРЕУТВЕРЖДЕНО:

На заседании кафедры Информационных систем и технологий
Протокол заседания кафедры № 8 от 26.05.2020 г.
Заведующий кафедрой Информационных систем и технологий

 Киприна Л.Ю., к.т.н., доцент

Подпись

1. Цели и задачи освоения дисциплины

Цель дисциплины: – формирование представления о типах исследовательских задач, возникающих в области интеллектуального анализа данных и методах их решения, которые помогут обучающимся выявлять, формализовать и успешно решать практические задачи анализа данных, возникающие в процессе их профессиональной деятельности

Задачи дисциплины:

- Научиться формулировать задачи анализа данных, выбирать адекватные алгоритмы их решения.
- Применять полученные знания для решения нестандартных исследовательских и аналитических задач.

2. Перечень планируемых результатов обучения по дисциплине

В результате освоения дисциплины обучающийся должен:

знать:

• принципы обработки больших массивов данных, способы их представления и хранения;

• основные задачи и методы интеллектуального анализа данных

уметь:

• формулировать практические задачи анализа данных и выбирать адекватные алгоритмы их решения

• применять методы интеллектуального анализа данных для решения профессиональных задач

владеть:

• технологиями разработки алгоритмов и программными системами анализа данных;

• навыками практического применения методов интеллектуального анализа данных для решения профессиональных задач

освоить компетенции:

ОПК-4. Способен применять на практике новые научные принципы и методы исследований

Индикаторы освоенности компетенции:

ИД-4.1- знать: новые научные принципы и методы исследований

ИД-4.2- уметь: применять на практике новые научные принципы и методы исследований

ИД-4.3- иметь навыки: применения новых научных принципов и методов исследования для решения профессиональных задач

3. Место дисциплины в структуре ОП ВО

Дисциплина входит в обязательную часть Блока 1. Изучается во 2 семестре.

Дисциплина предполагает, что полученные компетенции в дальнейшем будут использованы в рамках научно-исследовательской работы и работы над ВКР

4. Объем дисциплины (модуля)

4.1. Объем дисциплины в зачетных единицах с указанием академических (астрономических) часов и виды учебной работы

Виды учебной работы,	Очная форма
Общая трудоемкость в зачетных единицах	5
Общая трудоемкость в часах	180
Аудиторные занятия в часах, в том числе:	50
Лекции	12
Практические занятия	12
Лабораторные занятия	26
Курсовой проект	4
Проведение экзамена	2,35
Самостоятельная работа в часах	87,65+36
Форма промежуточной аттестации	экзамен

4.2. Объем контактной работы на 1 обучающегося

Виды учебных занятий	Очная форма
Лекции	12
Практические занятия	12
Лабораторные занятий	26
Консультации	
Зачет/зачеты	
Экзамен/экзамены	2,35
Курсовые работы	
Курсовые проекты	4
Всего	46,35

5.Содержание дисциплины (модуля), структурированное по темам (разделам), с указанием количества часов и видов занятий

5.1 Тематический план учебной дисциплины

№	Название раздела, темы	Всего з.е./час	Аудиторные занятия			Самостоятельная работа
			Лекции	Практические	Лабораторные	
1	Введение в анализ данных. Жизненный цикл CRISP-DM	14	2	4	-	8
2	Классификация методов анализа данных	14	2	4	-	8
3	Работа с распределениями случайных величин	12	-	-	4	8
4	Задача регрессии	31,65	2	-	10	19,65
5	Задача классификации	26	2	-	6	18
6	Задача кластеризации. Многомерное сжатие данных	26	2	-	6	18
7	Задача авторегрессии	14	2	4	-	8
8	Курсовой проект	4				4
9	Экзамен	36+2,35	-	-		36+2,35
	Итого:	5/180	12	12	26	87,65+36+6,35

5.2. Содержание:

Введение в анализ данных. Жизненный цикл CRISP-DM

Назначение систем анализа данных. Причины, обусловившие актуальность данной темы: перманентный реинжиниринг, задача автоматизации интеллектуальных операций, HR, цифровая экономика, понятие BigData: 3V, 5V, 7V. Трехуровневая архитектура системы анализа данных. Сходство и различие в понятиях: Статистика, Эконометрика, Машинное обучение. Цикл машинного обучения: от постановки задачи до принятия решения. Проблема ошибок первого и второго рода. HADI и CRISP-DM. Структура проекта анализа данных: роли в команде. Стадии формирования моделей.

Классификация методов анализа данных

Задачи анализа и прогнозирования. Линейные и нелинейные методы. Основные задачи анализа данных: регрессия, классификация, кластеризация. Дополнительные методы: анализ распределений и поиск аномалий, многомерное сжатие, DEA-анализ, распознавание образов, рекомендательные системы и заполнение пропусков, ассоциативные правила. Ансамбли моделей: бэггинг, стекинг, бустинг.

Работа с распределениями случайных величин

Понятие случайной величины. Законы распределения случайных величин. Базовые законы распределения: распределение Бернулли, нормальный закон и закон Пуассона. Нормальный закон распределения, методы проверки нормальности: критерий Пирсона, критерий Шапиро-Уилка, qqplot. Вероятностный гипотико-дедуктивный подход к решению задач. Алгоритм формулирования и тестирования гипотез. Проблема множественности гипотез. Методы работы с множеством гипотез: методы Холма, Бонферрони, Шидака, Бенджамини. Задача выявления и анализа аномалий.

Задача регрессии

Общая постановка задачи регрессии. Задача линейной регрессии. Проблема корреляции входных параметров, регуляризация. Нелинейные методы: понятие дерева решений, случайный лес и градиентный бустинг, К ближайших соседей. Оценка качества регрессии. Требование статичности ошибки: гомосекдастичность и гетероскедастичность, критерии оценки статичности ошибки. Оценка значимости регрессии: R2 и критерий Фишера. Оценка параметров линейной регрессии: критерий стьюдента. Выбор лучшей модели, критерий Акаике.

Задача классификации

Общая постановка задачи классификации. Линейные методы классификации: линейная и логистическая регрессия, метод опорных векторов. Использование метода опорных векторов при решении нелинейных задач. Нелинейные методы классификации: случайный лес и градиентный бустинг, К ближайших соседей. Критерии качества результатов классификации: accuracy, precision, recall, f1 метрики. ROC-кривая. Проблема балансировки данных при решении задач классификации. Методы балансировки. Нормализация данных. Принцип GIGO и проблема качества данных. Причины низкого качества данных и методы их выявления. Задача нормализации. Нормализация количественных параметров, нормализация категориальных параметров. Устранение пропусков в данных.

Задача кластеризации. Многомерное сжатие данных

Общая постановка задачи кластеризации. Линейные методы кластеризации: k-средних, EM, MeanShift. Выбор лучшей модели по критерию Акаике. Нелинейные методы кластеризации: HDBScan. Анализ результатов кластеризации: визуализация результатов, анализ взаимного расположения множеств, использование логистической регрессии. Задача многомерного сжатия. Линейные методы многомерного сжатия: методы главных компонент и SVD-преобразований. Нелинейные методы многомерного сжатия: MDS и tSNE. Выделение и прогнозирование трендов в компонентах данных.

Задача авторегрессии

Понятие временного ряда. Задача прогнозирования временного ряда. Выделение компонент временных рядов: трендовая, сезонная и случайная компоненты. Базовые методы прогнозирования временных рядов: авторегрессия и скользящее среднее. Современные методы прогнозирования временных рядов: SARIMA и GARCH. Оценка качества авторегрессионных моделей.

6. Методические материалы для обучающихся по освоению дисциплины

6.1. Самостоятельная работа обучающихся по дисциплине (модулю)

№ п/п	**Раздел (тема) дисциплины**	**Задание**	**Часы**	**Методические рекомендации по выполнению задания**	**Форма контроля**
1.	Введение в	Выполнить	8	Сформулируйте свою позицию,	Защита

	анализ данных. Жизненный цикл CRISP-DM	лабораторные работы		отражающую ключевые моменты лекции, выполните лабораторные работы	лабораторных и практических работ и курсового проекта, экзамен
2.	Классификация методов анализа данных	Выполнить лабораторные работы	8	Сформулируйте свою позицию, отражающую ключевые моменты лекции, выполните лабораторные и практические работы	Защита лабораторных и практических работ и курсового проекта, экзамен
3	Работа с распределениями случайных величин	Выполнить лабораторные работы	8	Сформулируйте свою позицию, отражающую ключевые моменты лекции, выполните лабораторные и практические работы	Защита лабораторных и практических работ и курсового проекта, экзамен
4	Задача регрессии	Выполнить лабораторные работы	19,65	Сформулируйте свою позицию, отражающую ключевые моменты лекции, выполните лабораторные и практические работы	Защита лабораторных и практических работ и курсового проекта, экзамен
5	Задача классификации	Выполнить лабораторные работы	18	Сформулируйте свою позицию, отражающую ключевые моменты лекции, выполните лабораторные и практические работы	Защита лабораторных и практических работ и курсового проекта, экзамен
6	Задача кластеризации. Многомерное сжатие данных	Выполнить лабораторные работы	18	Сформулируйте свою позицию, отражающую ключевые моменты лекции, выполните лабораторные и практические работы	Защита лабораторных и практических работ и курсового проекта, экзамен
7	Задача авторегрессии	Выполнить лабораторные работы	8	Сформулируйте свою позицию, отражающую ключевые моменты лекции, выполните лабораторные и практические работы	Защита лабораторных и практических работ и курсового проекта, экзамен
8	Выполнение курсового проекта	В рамках темы магистерской диссертации выполнить проект по анализу данных	-	В рамках темы магистерской диссертации выделите задачу анализа данных, выполните исследование по алгоритму CRISP-DM	Защита курсового проекта
9	Подготовка к экзамену	Изучить материалы лекций, выполнить все лабораторные работы	36	Использование материалов лекций, лабораторных работ и рекомендованной литературы	экзамен

6.2. Тематика и задания для практических занятий (*при наличии*)

1. Формулирование гипотез машинного обучения.
2. Авторегрессионная задача прогнозирования финансовых трендов
3. Прогнозирование пола и возраста по фотографии.

6.3. Тематика и задания для лабораторных занятий

1. Однофакторный регрессионный анализ, линейная регрессия и регуляризаторы.
2. Классификационная задача кредитного скоринга
3. Кластеризация данных о студентах
4. Задача прогнозирования аренды велосипедов
5. Формулирование и анализ гипотез о клиентах банка.

6.4. Методические рекомендации для обучающихся по освоению дисциплины (модуля)

Рекомендуется обязательное посещение лекций и лабораторных и практических работ студентами ввиду ограниченного количества литературы и постоянного обновления теоретического и практического материала.

Самостоятельная работа студентов заключается в изучении материала лекций и рекомендованной литературы, самостоятельном изучении указанных разделов и тем дисциплины, подготовке к лабораторным работам, подготовке отчетов по лабораторным работам, выполнении индивидуальных заданий, подготовке к защите лабораторных работ, подготовке реферата. Отчет по лабораторной работе может представляться в электронной форме в виде листинга программного кода или файла в формате *.doc или *.pdf с включением изображений (скриншотов) в соответствии с заданием на лабораторную работу. Контроль самостоятельной работы студентов осуществляется в форме теоретического и практического опроса согласно перечню тем, предусмотренных в рабочей программе дисциплины.

Лекционное обучение осуществляется в аудиториях, оснащенных специализированным оборудованием, таким как: ПК, видеопроектор, оптический проектор, аудио и видеосистемы.

Лабораторные и практические задания выполняются в соответствии с тематикой лабораторных работ, приведенной в рабочей программе дисциплины, в компьютерных классах, оснащенных 7-9 ПК, объединенными в локальную сеть.

6.5. Методические рекомендации для выполнения курсовых работ (проектов)

В рамках темы магистерской диссертации выделите задачу анализа данных, выполните исследование по алгоритму CRISP-DM

7. Перечень основной и дополнительной литературы, необходимой для освоения дисциплины (модуля)

а) основная:

2. Григорьев А.А., Исаев Е.А. Методы и алгоритмы обработки данных. – М.: ИНФРА-М, 2018. – 383 с. – URL: https://znanium.com/catalog/document?id=361208

б) дополнительная:

1. Интеллектуальные информационные системы и технологии : учебное пособие / Ю.Ю. Громов, О.Г. Иванова, В.В. Алексеев и др. - Тамбов : Издательство ФГБОУ ВПО «ТГТУ», 2013. - 244 с. : ил. - - ISBN 978-5-8265-1178-7 ; То же [Электронный ресурс]. – URL: http://biblioclub.ru/index.php?page=book&id=277713

2. Серегин, М. Ю. Интеллектуальные информационные системы : учебное пособие / М.Ю. Серегин, М.А. Ивановский, А.В. - Тамбов : Издательство ФГБОУ ВПО «ТГТУ», 2012. - 205 с. : ил. - То же [Электронный ресурс]. - URL: http://biblioclub.ru/index.php?page=book&id=277790

3. Интеллектуальные системы : учебное пособие / А. Семенов, Н. Соловьев, Е. Чернопрудова, А. Цыганков. - Оренбург : ОГУ, 2013. - 236 с. ; То же [Электронный ресурс]. - URL: http://biblioclub.ru/index.php?page=book&id=259148

4. Боровская Е.В., Давыдова Н.А. Основы искусственного интеллекта. – М.: Лаборатория знаний, 2020. – 130 с. – URL: https://znanium.com/catalog/document?id=365893

5. Гуриков С.Р. Основы алгоритмизации и программирования на Python. – М.: Форум, 2020. – 343 с. – URL: https://znanium.com/catalog/document?id=366970

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

Информационно-образовательные ресурсы:

1. Федеральный портал «Российское образование», [Электронный ресурс], URL: http://www.edu.ru/

2. Официальный сайт министерства образования и науки Российской Федерации, [Электронный ресурс], URL: https://минобрнауки.рф/

3. Библиотека ГОСТов. Все ГОСТы, [Электронный ресурс], URL: http://vsegost.com/

Электронные библиотечные системы:

1. ЭБС «Лань»
2. ЭБС «Университетская библиотека online»
3. ЭБС «Znanium»

Программное обеспечение

Jupiter Anaconda for Python 3
Colab.research.google.com

9. Описание материально-технической базы, необходимой для осуществления образовательного процесса по дисциплине

Наименование специальных помещений и помещений для самостоятельной работы	Оснащенность специальных помещений и помещений для самостоятельной работы	Перечень лицензионного программного обеспечения. Реквизиты подтверждающего документа
ауд. Е-326 (занятия лекционного типа, групповые консультации, промежуточная аттестация)	Лекционная аудитория. Число посадочных мест – 80. Имеется: мультимедиа – проектор с компьютером,	Лицензионное программное обеспечение не используется

	выход в интернет; усилитель; колонки.	
ауд. Е-323 (лабораторные занятия, индивидуальные консультации, промежуточная аттестация, самостоятельная работа обучающихся)	Компьютерный класс. Число посадочных мест – 16.Число мест, оборудованных компьютерами – 8 с выходом в интернет. Имеется: мультимедиа – проектор с компьютером; интерактивная доска.	Лицензионное программное обеспечение не используется
ауд. Е-321 (лабораторные занятия, индивидуальные консультации, промежуточная аттестация, самостоятельная работа обучающихся)	Компьютерный класс. Число посадочных мест – 16.Число мест, оборудованных компьютерами – 8 с выходом в интернет. Имеется: мультимедиа – проектор с компьютером; колонки.	Лицензионное программное обеспечение не используется

Проведение занятий лекционного типа, лабораторных работ, индивидуальных и групповых консультаций, промежуточной аттестации возможно в других аудиториях КГУ, имеющих аналогичное техническое и программное оснащение.