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

РАБОЧАЯ ПРОГРАММА ДИСЦИПЛИНЫ

Машинное обучение и анализ данных

Направление подготовки *09.03.02 Информационные системы и технологии*
Направленность *«*Разработка и внедрение интеллектуальных компонентов информационных систем*»*

Квалификация выпускника: *бакалавр*

Кострома
2023

Рабочая программа дисциплины «Машинное обучение и анализ данных» разработана в соответствии с Федеральным государственным образовательным стандартом по направлению 09.03.02 Информационные системы и технологии (уровень бакалавриата), утвержден приказом Министерства образования и науки РФ № 926 от 19.09.17.

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

Рецензент: Панин И.Г., д.т.н., доцент

УТВЕРЖДЕНО:

На заседании кафедры информационных систем и технологий:
Протокол заседания кафедры №6 от 27.04.2023 г.
Заведующая кафедрой информационных систем и технологий:
Киприна Л.Ю., к.т.н., доцент

1. Цели и задачи освоения дисциплины

Цель дисциплины: Формирование систематизированного представления о концепциях, моделях и принципах технологий машинного обучения и анализа данных. Ознакомление с принципами организации информационного обмена и консолидации данных, их поиска и извлечения. Получение представления о трансформации данных и способах их визуализации.

Задачи дисциплины:

- Изучение технологий машинного обучения и анализа данных, получение представления о консолидации, трансформации данных и способах их визуализации
– Знакомство с типовыми алгоритмами решения профессиональных задач на основе анализа данных
– Развитие умений применять инструментальные средства машинного обучения и анализа данных
– Развитие компетенций в области построения и оценки качества моделей машинного обучения и анализа данных.

2. Перечень планируемых результатов обучения по дисциплине

В результате освоения дисциплины обучающийся должен:

знать:

Предметная область анализа больших данных
Типы анализа данных, виды аналитики
Стандарты проведения анализа данных: CWM (Common Warehouse Metamodel), CRISP-DM (The Cross Industrie Standard Process for Data Mining), PMML (Predicted Model Markup Language)
Источники данных, в том числе информации, необходимой для обеспечения деятельности в предметной области заказчика исследования
Типы больших данных: метаданные, полуструктурированные, структурированные, неструктурированные данные
Методы извлечения информации и знаний из гетерогенных, мульти структурированных, неструктурированных источников
Фильтрация шумовых выбросов, виды шумовых выбросов: глобальный, контекстуальный, коллективный
Методы обеспечения и оценки качества информации
Алгоритмы машинного обучения: обучение с учителем, обучение без учителя, полууправляемое обучение, обучение с подкреплением; классификация, кластеризация, обнаружение выбросов, фильтрация
Статистический анализ: A/B тестирование, корреляционный анализ, регрессионный анализ
Методы оценки моделей: оценка качества построенной модели по тестовой выборке и анализ обобщающих способностей алгоритма
Современные методы и инструментальные средства анализа больших данных
Методы интерпретации и визуализации анализа больших данных

уметь:

Использовать инструментальные средства для извлечения, преобразования, хранения и обработки данных из разнородных источников
Оценивать адекватность наборов данных
Производить очистку данных для проведения аналитических работ
Проводить интеграцию и преобразование данных
Разрабатывать и оценивать модели данных
Решать задачи кластеризации, регрессии, прогнозирования, снижения размерности и

ранжирования данных
Определять теоретические верхние оценки переобученности: сложность, разделимость, устойчивость; решать проблемы переобучения и недообучения алгоритма
Планировать и проводить аналитические работыИспользовать инструментальные средства для извлечения, преобразования, хранения и обработки
данных из разнородных источников
Оценивать адекватность наборов данных
Производить очистку данных для проведения аналитических работ
Проводить интеграцию и преобразование данных
Разрабатывать и оценивать модели данных
Решать задачи кластеризации, регрессии, прогнозирования, снижения размерности и ранжирования данных
Определять теоретические верхние оценки переобученности: сложность, разделимость, устойчивость; решать проблемы переобучения и недообучения алгоритма
Планировать и проводить аналитические работы

быть готовым к выполнению следующих трудовых действий:

Определение состава группы для проведения анализа больших данных
Разработка, обсуждение и утверждение плана и содержания аналитических работ
Определение источников данных для анализа, идентификация внешних и внутренних данных
Извлечение, проверка, очистка и агрегация данных и разработка представления данных
Оценка соответствия набора данных предметной области и задачам аналитических работ
Разработка, поверка, оценка используемых моделей больших данных

освоить компетенции:

ПК-4 Способен осуществлять сбор, обработку и анализ больших данных с использованием существующей в организации методологической и технологической инфраструктуры

Индикаторы освоенности компетенции:

ПК-4.1 Способен планировать и организовывать аналитические работы
ПК-4.2 Готов осуществлять подготовку данных для проведения аналитических работ
ПК-4.3 Способен проводить аналитические исследования в соответствии с согласованными требованиями

3. Место дисциплины в структуре ОП ВО

Дисциплина входит в часть, формируемую участниками образовательного процесса, Блока 1. Изучается в 3 семестре

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	6
Общая трудоемкость в часах	216
Аудиторные занятия в часах, в том числе:	54
Лекции	18
Практические занятия	-
Лабораторные занятия	36
Проведение экзамена	2,35
Самостоятельная работа в часах	123,65+36
Форма промежуточной аттестации	экзамен

4.2. Объем контактной работы на 1 обучающегося

Виды учебных занятий	Очная форма
Лекции	18
Практические занятия	-
Лабораторные занятий	36
Консультации	-
Зачет/зачеты	-
Экзамен/экзамены	2,35
Курсовые работы	-
Курсовые проекты	-
Всего	56,35

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	23	2	-	6	15
2	Классификация методов анализа данных	21	2	-	4	15
3	Работа с распределениями случайных величин	21	2	-	4	15
4	Задача регрессии	27	2	-	10	15
5	Задача классификации	21	2	-	4	15
6	Нормализация данных	17	2	-		15
7	Задача кластеризации. Многомерное сжатие данных	21	2	-	6	15
8	Задача авторегрессии	19	2	-	2	15
9	DEA-анализ	5,65	2	-	-	3,65
10	Экзамен	36+2,35	-	-	-	36+2,35
	Итого:	6/216	18	-	36	123,65+36+2,35

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.

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.

DEA-анализ

Понятие эффективности и методы ее оценки. Проблема оценки эффективности многокритериальных задач. Метод анализа среды функционирования (Data envelopment analysis). CRS и VRS модели. Модели ориентированные на достижение максимума эффектов (output-oriented) и минимума ресурсов (input-oriented). Метод выпуклых

оболочек (FDH). Оценка устойчивости результатов DEA-анализа.

5.3. Практическая подготовка

Код компетенции	Индикатор компетенции	Содержание задания на практическую подготовку по выбранному виду деятельности	Число часов практической подготовки			
			Всего	Лекции	Практ. занятия	Лаб.раб
ПК-4	ПК-4.1	Планирование и организация аналитических работ (лаб. работы 6, 8, 9, 10, 11)	6	-	-	6
ПК-4	ПК-4.2	Подготовка данных (лаб. работы 8, 9,10, 11)	10	-	-	10
ПК-4	ПК-4.3	Проведение аналитических исследований (лаб. работы 6, 8, 9, 10, 11)	6	-	-	6

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 анализ данных. Жизненный цикл CRISP-DM	Выполнить лабораторные работы	15	Сформулируйте свою позицию, отражающую ключевые моменты лекции, выполните лабораторные работы	Защита лабораторной работы, экзамен
2.	Классификация методов анализа данных	Выполнить лабораторные работы	15	Сформулируйте свою позицию, отражающую ключевые моменты лекции, выполните лабораторные работы	Защита лабораторной работы, экзамен
3	Работа с распределениями случайных величин	Выполнить лабораторные работы	15	Сформулируйте свою позицию, отражающую ключевые моменты лекции, выполните лабораторные работы	Защита лабораторной работы, экзамен
4	Задача регрессии	Выполнить лабораторные работы	15	Сформулируйте свою позицию, отражающую ключевые моменты лекции, выполните лабораторные работы	Защита лабораторной работы, экзамен
5	Задача классификации	Выполнить лабораторные работы	15	Сформулируйте свою позицию, отражающую ключевые моменты лекции, выполните лабораторные работы	Защита лабораторной работы, экзамен
6	Нормализация данных	Выполнить лабораторные	15	Сформулируйте свою позицию, отражающую	Защита лабораторной

		работы		ключевые моменты лекции, выполните лабораторные работы	работы, экзамен
7	Задача кластеризации. Многомерное сжатие данных	Выполнить лабораторные работы	15	Сформулируйте свою позицию, отражающую ключевые моменты лекции, выполните лабораторные работы	Защита лабораторной работы, экзамен
8	Задача авторегрессии	Выполнить лабораторные работы	15	Сформулируйте свою позицию, отражающую ключевые моменты лекции, выполните лабораторные работы	Защита лабораторной работы, экзамен
9	DEA-анализ	Сформулировать задачу для DEA-анализа	3,65	Сформулируйте свою позицию, отражающую ключевые моменты лекции, сформулируйте задачу для DEA-анализа	экзамен
12	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*)

Не предусмотрены учебным планом

6.3. Тематика и задания для лабораторных занятий

1. Основы python. List и алгоритмические структуры.
2. Основы python. Numpy.array и построение графиков
3. Моделирование центральной предельной теоремы
4. Формулирование гипотез машинного обучения.
5. Прогнозирование пола и возраста по фотографии.
6. Однофакторный регрессионный анализ, линейная регрессия и регуляризаторы.
7. Авторегрессионная задача прогнозирования финансовых трендов
8. Классификационная задача кредитного скоринга
9. Кластеризация данных о студентах
10. Задача прогнозирования аренды велосипедов
11. Формулирование и анализ гипотез о клиентах банка.

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

Рекомендуется обязательное посещение лекций и лабораторных 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 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)

Не предусмотрены учебным планом

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

а) основная:

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

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

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, выход в интернет; усилитель; колонки.	Лицензионное программное обеспечение не используется
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	Лицензионное программное обеспечение не используется

	компьютером; колонки.	

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