

Министерство науки и высшего образования Российской Федерации
Федеральное государственное автономное образовательное учреждение
высшего образования
«Национальный исследовательский ядерный университет «МИФИ»
ИНСТИТУТ ИНТЕЛЛЕКТУАЛЬНЫХ КИБЕРНЕТИЧЕСКИХ СИСТЕМ
КАФЕДРА КРИПТОЛОГИИ И ДИСКРЕТНОЙ МАТЕМАТИКИ

ОДОБРЕНО УМС ИИКС

Протокол № 8/1/2024

от 28.08.2024 г.

РАБОЧАЯ ПРОГРАММА УЧЕБНОЙ ДИСЦИПЛИНЫ
ПРЕДПОДГОТОВКА И АНАЛИЗ ДАННЫХ

Направление подготовки
(специальность)

[1] 01.04.02 Прикладная математика и информатика

Семестр	Трудоемкость, кред.	Общий объем курса, час.	Лекции, час.	Практич. занятия, час.	Лаборат. работы, час.	В форме практической подготовки/ В	СРС, час.	КСР, час.	Форма(ы) контроля, экз./зач./КР/КП
1	6	216	12	12	12		144	0	Э КП
Итого	6	216	12	12	12	0	144	0	

АННОТАЦИЯ

обучение студентов методам предподготовки и анализа данных

1. ЦЕЛИ И ЗАДАЧИ ОСВОЕНИЯ УЧЕБНОЙ ДИСЦИПЛИНЫ

обучение студентов методам предподготовки и анализа данных

2. МЕСТО УЧЕБНОЙ ДИСЦИПЛИНЫ В СТРУКТУРЕ ООП ВО

дисциплина профессионального цикла

3. ФОРМИРУЕМЫЕ КОМПЕТЕНЦИИ И ПЛАНИРУЕМЫЕ РЕЗУЛЬТАТЫ ОБУЧЕНИЯ

Универсальные и(или) общепрофессиональные компетенции:

Код и наименование компетенции	Код и наименование индикатора достижения компетенции
ОПК-2 [1] – Способен совершенствовать и реализовывать новые математические методы решения прикладных задач	З-ОПК-2 [1] – Знать основные понятия, математические методы решения прикладных задач, принципы математического моделирования и методы верификации. У-ОПК-2 [1] – Уметь применять полученную теоретическую базу для решения практических задач В-ОПК-2 [1] – Владеть основными математическими методами решения прикладных задач
ОПК-3 [1] – Способен разрабатывать математические модели и проводить их анализ при решении задач в области профессиональной деятельности	З-ОПК-3 [1] – Знать основные методы и принципы математического моделирования, методы построения математических моделей типовых профессиональных задач, способы нахождения решений математических моделей и содержательной интерпретации полученных результатов. У-ОПК-3 [1] – Уметь составлять математические модели типовых профессиональных задач, находить способы их решения и профессионально интерпретировать смысл полученного результата. В-ОПК-3 [1] – Владеть методами построения математических моделей типовых профессиональных задач, способами нахождения решений математических моделей и содержательной интерпретации полученных результатов

4. СТРУКТУРА И СОДЕРЖАНИЕ УЧЕБНОЙ ДИСЦИПЛИНЫ

Разделы учебной дисциплины, их объем, сроки изучения и формы контроля:

№ п.п	Наименование раздела учебной дисциплины	Недели	Лекции/ Практи. (семинары)/ Лабораторные работы, час.	Обязат. текущий контроль (форма*, неделя)	Максимальный балл за раздел**	Аттестация раздела (форма*, неделя)	Индикаторы освоения компетенции
	<i>1 Семестр</i>						
1	Первый раздел	1-8	16/8/24		25	КИ-8	3-ОПК-2, У-ОПК-2, В-ОПК-2, 3-ОПК-3, У-ОПК-3, В-ОПК-3
2	Второй раздел	9-16	16/8/24		25	КИ-16	3-ОПК-2, У-ОПК-2, В-ОПК-2, 3-ОПК-3, У-ОПК-3, В-ОПК-3
	<i>Итого за 1 Семестр</i>		12/12/12		50		
	Контрольные мероприятия за 1 Семестр				50	КП, Э	3-ОПК-2, У-ОПК-2, В-ОПК-2, 3-ОПК-3, У-ОПК-3, В-ОПК-3, 3-ОПК-2, У-ОПК-2, В-ОПК-2, 3-ОПК-3, У-ОПК-3, В-ОПК-3

* – сокращенное наименование формы контроля

** – сумма максимальных баллов должна быть равна 100 за семестр, включая зачет и (или) экзамен

Сокращение наименований форм текущего контроля и аттестации разделов:

Обозначение	Полное наименование
КИ	Контроль по итогам
Э	Экзамен
КП	Курсовой проект

КАЛЕНДАРНЫЙ ПЛАН

Недели	Темы занятий / Содержание	Лек., час.	Пр./сем., час.	Лаб., час.
	<i>1 Семестр</i>	12	12	12

1-8	Первый раздел	16	8	24
	Введение в предобработку данных Рассматриваются основные задачи и цели предобработки данных как ключевого этапа аналитического процесса. Анализируется роль подготовки данных в обеспечении корректности, достоверности и воспроизводимости результатов анализа. Изучаются этапы работы с данными — от сбора и интеграции до очистки, нормализации и кодирования признаков. Особое внимание уделяется инструментам и библиотекам, применяемым в предобработке данных (Python, Pandas, NumPy, SQL), а также типовым проблемам, возникающим при работе с «сырыми» данными: ошибки ввода, пропуски, дубликаты, некорректные форматы. Рассматриваются примеры влияния низкого качества исходных данных на последующие этапы анализа и построения моделей машинного обучения.	Всего аудиторных часов		
		3	1	4
		Онлайн		
		3	0	0
	Работа с пропусками Изучаются причины возникновения пропусков в данных, их природа и влияние на качество анализа. Рассматриваются методы обнаружения пропусков и недостающих значений при помощи стандартных инструментов Python и библиотеки Pandas (.isna(), .info(), .describe()). Обсуждаются стратегии обработки пропусков, включая удаление строк и столбцов, заполнение значений с помощью статистических показателей (среднего, медианы, моды) или прогнозных методов. Анализируются последствия выбора метода обработки пропусков для точности моделей и интерпретации данных. Подчеркивается важность критической оценки допустимости замещения данных и влияние выбранных подходов на достоверность результатов.	Всего аудиторных часов		
		3	1	4
		Онлайн		
		3	0	0
	Изменение типов данных Тема посвящена классификации и преобразованию типов данных в процессе подготовки к анализу. Рассматриваются основные типы данных (числовые, строковые, категориальные, логические, временные) и их особенности при обработке. Изучаются методы приведения типов, включая автоматическое и ручное преобразование, а также потенциальные ошибки и искажения, возникающие при неправильном типе переменной. Обсуждается вопрос оптимизации памяти и производительности вычислений через выбор подходящего типа данных. Особое внимание уделяется проблемам конверсии при загрузке данных из различных источников и необходимости единообразия типов для последующего анализа и визуализации.	Всего аудиторных часов		
		3	2	5
		Онлайн		
		3	0	0
	Поиск дубликатов Рассматриваются виды дубликатов данных — полные, частичные и логические, а также причины их появления при интеграции данных из разных источников. Изучаются методы обнаружения дубликатов с использованием	Всего аудиторных часов		
		3	2	5
		Онлайн		
		3	0	0

	функций Pandas (.duplicated(), .drop_duplicates()) и алгоритмические подходы к поиску похожих записей. Обсуждаются последствия наличия дублей для статистического анализа, визуализации и обучения моделей. Анализируются способы обработки и удаления дублей, включая использование ключей, индексов и специальных идентификаторов. Отдельное внимание уделяется построению процедур контроля качества данных и предотвращению повторного появления дубликатов в рабочих наборах.			
	Категоризация данных Тема направлена на изучение принципов преобразования непрерывных и дискретных признаков в категориальные переменные. Рассматриваются цели категоризации, связанные с оптимизацией, упрощением анализа и улучшением интерпретации данных. Подробно изучаются методы разбиения данных на интервалы (pd.cut(), pd.qcut()), отображение категорий с помощью функций .map(), а также подходы кодирования признаков — One-Hot Encoding и Label Encoding. Анализируются особенности работы с номинальными и порядковыми категориями, а также типичные ошибки при выборе метода кодирования. Отмечается влияние категоризации на качество анализа, скорость вычислений и поведение моделей машинного обучения.	Всего аудиторных часов		
		4	2	6
		Онлайн		
		4	0	0
9-16	Второй раздел	16	8	24
	Системное и критическое мышление в работе аналитика. Рассматривается роль системного и критического мышления в процессе обработки и анализа данных. Изучаются подходы к логике принятия решений при выборе методов очистки, преобразования и анализа данных. Обсуждается значение оценки качества исходной информации и способности аналитика выявлять потенциальные искажения, ошибки или противоречия. Особое внимание уделяется когнитивным искажениям, которые могут влиять на интерпретацию результатов и принятие решений. Формируются навыки объективной оценки источников данных, проверки гипотез и выбора оптимальных решений с учетом контекста задачи и ограничений доступных инструментов.	Всего аудиторных часов		
		3	1	4
		Онлайн		
		3	0	0
	Введение в исследовательский анализ данных (EDA). Изучаются цели, задачи и этапы проведения исследовательского анализа данных (Exploratory Data Analysis). Рассматриваются методы получения описательной статистики (.describe(), .info()) и их роль в первичном понимании структуры данных. Анализируются распределения признаков с помощью визуальных инструментов (гистограммы, boxplot, плотности вероятности) для выявления закономерностей и аномалий. Подчеркивается значение EDA в проверке гипотез, выборе моделей и формировании предположений о взаимосвязях	Всего аудиторных часов		
		3	1	4
		Онлайн		
		3	0	0

	между переменными. Рассматриваются примеры интерпретации результатов и распространенные ошибки при поверхностном анализе данных.			
	Первые графики и выводы Рассматриваются основные типы графиков и визуализаций, применяемые для первичного анализа данных: линейные, столбчатые, круговые диаграммы, а также boxplot и scatter plot. Изучаются инструменты построения графиков на языке Python с использованием библиотек Matplotlib и Seaborn. Описываются принципы выбора оптимального типа визуализации в зависимости от задачи и структуры данных. Обсуждаются правила корректной интерпретации графиков, типичные ошибки восприятия и некорректные способы визуального представления данных. Особое внимание уделяется эстетике графиков, их читаемости и способности доносить ключевые выводы до аудитории.	Всего аудиторных часов		
		3	2	5
		Онлайн		
		3	0	0
	Изучение срезов данных Тема посвящена работе с подмножествами данных и построению аналитических срезов для детального изучения закономерностей. Рассматриваются методы фильтрации данных с использованием логических условий, функции группировки (groupby) и агрегации (mean, sum, count и др.). Анализируются принципы выделения подгрупп и сравнения между ними, а также методы вычисления сводных статистик. Обсуждается роль срезов данных в выявлении структурных особенностей, сезонности и взаимосвязей. Формируются навыки систематизации результатов анализа и построения агрегированных таблиц (pivot_table) для подготовки отчетов и визуализаций.	Всего аудиторных часов		
		3	2	5
		Онлайн		
		3	0	0
	Работа с несколькими источниками данных Рассматриваются методы интеграции данных из различных форматов и источников (CSV, Excel, SQL, JSON и др.). Изучаются принципы объединения наборов данных с использованием операций merge, concat и join, а также методы согласования структуры и типов данных при слиянии. Анализируются типичные проблемы при работе с разнородными источниками: несоответствие ключей, дублирование записей, различия в кодировках и форматах хранения. Обсуждаются стратегии обеспечения целостности данных и проверка корректности объединений. Подчеркивается значение этого этапа для формирования единого аналитического пространства и обеспечения достоверности последующих расчетов.	Всего аудиторных часов		
		4	2	6
		Онлайн		
		4	0	0

Сокращенные наименования онлайн опций:

Обозначение	Полное наименование
ЭК	Электронный курс
ПМ	Полнотекстовый материал
ПЛ	Полнотекстовые лекции

ВМ	Видео-материалы
АМ	Аудио-материалы
Прз	Презентации
Т	Тесты
ЭСМ	Электронные справочные материалы
ИС	Интерактивный сайт

ТЕМЫ ЛАБОРАТОРНЫХ РАБОТ

Недели	Темы занятий / Содержание
	<i>1 Семестр</i>
	Л/Р 1 Рассматриваются методы обнаружения пропусков и недостающих значений при помощи стандартных инструментов Python и библиотеки Pandas (.isna(), .info(), .describe())
	Л/Р 2 методы обнаружения дубликатов с использованием функций Pandas (.duplicated(), .drop_duplicates())
	Л/Р 3 методы получения описательной статистики (.describe(), .info())
	Л/Р 4 методы фильтрации данных с использованием логических условий, функции группировки (groupby) и агрегации (mean, sum, count и др.)

5. ОБРАЗОВАТЕЛЬНЫЕ ТЕХНОЛОГИИ

Образовательные технологии сочетают в себе совокупность методов и средств для реализации определенного содержания обучения и воспитания в рамках дисциплины, включают решение дидактических и воспитательных задач, формируя основные понятия дисциплины, технологии проведения занятий, усвоения новых знаний, технологии повторения и контроля материала, самостоятельной работы.

6. ФОНД ОЦЕНОЧНЫХ СРЕДСТВ

Фонд оценочных средств по дисциплине обеспечивает проверку освоения планируемых результатов обучения (компетенций и их индикаторов) посредством мероприятий текущего, рубежного и промежуточного контроля по дисциплине.

Связь между формируемыми компетенциями и формами контроля их освоения представлена в следующей таблице:

Компетенция	Индикаторы освоения	Аттестационное мероприятие (КП 1)
ОПК-2	З-ОПК-2	КП, Э, КИ-8, КИ-16
	У-ОПК-2	КП, Э, КИ-8, КИ-16
	В-ОПК-2	КП, Э, КИ-8, КИ-16
ОПК-3	З-ОПК-3	КП, Э, КИ-8, КИ-16
	У-ОПК-3	КП, Э, КИ-8, КИ-16
	В-ОПК-3	КП, Э, КИ-8, КИ-16

Шкалы оценки образовательных достижений

Шкала каждого контрольного мероприятия лежит в пределах от 0 до установленного максимального балла включительно. Итоговая аттестация по дисциплине оценивается по 100-балльной шкале и представляет собой сумму баллов, заработанных студентом при выполнении заданий в рамках текущего и промежуточного контроля.

Итоговая оценка выставляется в соответствии со следующей шкалой:

Сумма баллов	Оценка по 4-х балльной шкале	Отметка о зачете	Оценка ECTS
90-100	5 – «отлично»	«Зачтено»	A
85-89	4 – «хорошо»		B
75-84			C
70-74			D
65-69	3 – «удовлетворительно»		E
60-64			
Ниже 60	2 – «неудовлетворительно»	«Не зачтено»	F

Оценка «отлично» соответствует глубокому и прочному освоению материала программы обучающимся, который последовательно, четко и логически стройно излагает свои ответы, умеет тесно увязывать теорию с практикой, использует в ответах материалы монографической литературы.

Оценка «хорошо» соответствует твердым знаниям материала обучающимся, который грамотно и, по существу, излагает свои ответы, не допуская существенных неточностей.

Оценка «удовлетворительно» соответствует базовому уровню освоения материала обучающимся, при котором освоен основной материал, но не усвоены его детали, в ответах присутствуют неточности, недостаточно правильные формулировки, нарушения логической последовательности.

Отметка «зачтено» соответствует, как минимум, базовому уровню освоения материала программы, при котором обучающийся владеет необходимыми знаниями, умениями и навыками, умеет применять теоретические положения для решения типовых практических задач.

Оценку «неудовлетворительно» / отметку «не зачтено» получает обучающийся, который не знает значительной части материала программы, допускает в ответах существенные ошибки, не выполнил все обязательные задания, предусмотренные программой. Как правило, такие обучающиеся не могут продолжить обучение без дополнительных занятий.

7. УЧЕБНО-МЕТОДИЧЕСКОЕ И ИНФОРМАЦИОННОЕ ОБЕСПЕЧЕНИЕ УЧЕБНОЙ ДИСЦИПЛИНЫ

ОСНОВНАЯ ЛИТЕРАТУРА:

ДОПОЛНИТЕЛЬНАЯ ЛИТЕРАТУРА:

ПРОГРАММНОЕ ОБЕСПЕЧЕНИЕ:

Специальное программное обеспечение не требуется

LMS И ИНТЕРНЕТ-РЕСУРСЫ:

<https://online.mephi.ru/>

<http://library.mephi.ru/>

8. МАТЕРИАЛЬНО-ТЕХНИЧЕСКОЕ ОБЕСПЕЧЕНИЕ УЧЕБНОЙ ДИСЦИПЛИНЫ

Специальное материально-техническое обеспечение не требуется

9. УЧЕБНО-МЕТОДИЧЕСКИЕ РЕКОМЕНДАЦИИ ДЛЯ СТУДЕНТОВ

Студенты должны своевременно спланировать учебное время для поэтапного и системного изучения данной учебной дисциплины в соответствии с планом лекций и семинарских занятий, графиком контроля знаний.

Успешное освоение дисциплины требует от студентов посещения лекций, активной работы во время семинарских занятий, выполнения всех домашних заданий, ознакомления с базовыми учебниками, основной и дополнительной литературой, а также предполагает творческое участие студента путем планомерной, повседневной работы.

Изучение дисциплины следует начинать с проработки учебной программы, особое внимание, уделяя целям и задачам, структуре и содержанию курса.

Во время лекций рекомендуется писать конспект. Запись лекции – одна из форм активной самостоятельной работы студентов, требующая навыков и умения кратко, схематично, последовательно и логично фиксировать основные положения, выводы, обобщения, формулировки.

При необходимости в конце лекции преподаватель оставляет время для того, чтобы студенты имели возможность задать вопросы по изучаемому материалу.

Лекции нацелены на освещение основополагающих положений теории алгоритмов и теории функций алгебры логики, наиболее трудных вопросов, как правило, связанных с доказательством необходимых утверждений и теорем, призваны способствовать формированию навыков работы с научной литературой. Предполагается также, что студенты приходят на лекции, предварительно проработав соответствующий учебный материал по источникам, рекомендуемым программой.

Конспект лекций для закрепления полученных знаний необходимо просмотреть сразу после занятий. Хорошо отметить материал конспекта лекций, который вызывает затруднения для понимания. Можно попытаться найти ответы на затруднительные вопросы, используя рекомендуемую литературу. Если самостоятельно не удалось разобраться в материале, рекомендуется сформулировать вопросы и обратиться за помощью к преподавателю на консультации или ближайшей лекции.

В процессе изучения учебной дисциплины необходимо обратить внимание на самоконтроль. Требуется регулярно отводить время для повторения пройденного материала, проверяя свои знания, умения и навыки по контрольным вопросам, а также для выполнения домашних заданий, которые выдаются после каждого семинара.

Систематическая индивидуальная работа, постоянная активность на занятиях, готовность ставить и обсуждать актуальные проблемы курса – залог успешной работы и положительной оценки.

10. УЧЕБНО-МЕТОДИЧЕСКИЕ РЕКОМЕНДАЦИИ ДЛЯ ПРЕПОДАВАТЕЛЕЙ

Учебный курс строится на интегративной основе и включает в себя как теоретические знания, так и практические навыки, получаемые студентами в ходе лекций, аудиторных практических занятий, лабораторных и самостоятельных занятий.

Данная дисциплина выполняет функции теоретической и практической подготовки студентов. Содержание дисциплины распределяется между лекционной и практической частями на основе принципа дополняемости: практические занятия, как правило, не дублируют лекции и посвящены рассмотрению практических примеров и конкретизации материала, введенного на лекции. В лекционном курсе главное место отводится общетеоретическим проблемам.

Содержание учебного курса, его объем и характер обуславливают необходимость оптимизации учебного процесса в плане отбора материала обучения и методики его организации, а также контроля текущей учебной работы. В связи с этим возрастает значимость и изменяется статус внеаудиторной (самостоятельной) работы, которая становится полноценным и обязательным видом учебно-познавательной деятельности студентов. При изучении курса самостоятельная работа включает:

- самостоятельное ознакомление студентов с теоретическим материалом, представленным в отечественных и зарубежных научно-практических публикациях;

- самостоятельное изучение тем учебной программы, достаточно хорошо обеспеченных литературой и сравнительно несложных для понимания;

- подготовку к практическим занятиям по тем разделам, которые не дублируют темы лекционной части, а потому предполагают самостоятельную проработку материала учебных пособий.

Со стороны преподавателя должен быть установлен контакт со студентами, и они должны быть информированы о порядке прохождения курса, его особенностях, учебно-методическом обеспечении по данной дисциплине. Преподаватель дает методические рекомендации обучаемым по самостоятельному изучению проблем, характеризуя пути и средства достижения поставленных перед ними задач, высказывает советы и рекомендации по изучению учебной литературы, самостоятельной работе и работе на семинарских занятиях.

Автор(ы):

Епишкина Анна Васильевна, к.т.н.