

**Федеральное государственное автономное образовательное  
учреждение высшего образования  
«Московский физико-технический институт  
(национальный исследовательский университет)»**

**УТВЕРЖДЕНО**

**Директор физтех-школы  
прикладной математики и  
информатики**

**А.М. Райгородский**

	<b>Рабочая программа дисциплины (модуля)</b>
<b>по дисциплине:</b>	Машинное обучение и анализ данных
<b>по направлению:</b>	Прикладная математика и информатика
<b>профиль подготовки:</b>	А1360: Передовые методы искусственного интеллекта Физтех-школа Прикладной Математики и Информатики кафедра дискретной математики
<b>курс:</b>	4
<b>квалификация:</b>	бакалавр

Семестр, формы промежуточной аттестации: 7 (осенний) - Дифференцированный зачет

Аудиторных часов: 60 всего, в том числе:

лекции: 30 час.

семинары: 30 час.

лабораторные занятия: 0 час.

Самостоятельная работа: 30 час.

Всего часов: 90, всего зач. ед.: 2

Программу составил: А.М. Райгородский, д-р физ.-мат. наук, профессор, ассистент

Программа обсуждена на заседании кафедры дискретной математики 12.02.2024

## Аннотация

В курсе показано, как проходит полный цикл анализа, от сбора данных до выбора оптимального решения и оценки его качества. Студенты освоят основные темы, необходимые в работе с большим массивом данных, в т.ч. современные методы классификации и регрессии, поиск структуры в данных, проведение экспериментов, построение выводов, базовая фундаментальная математика, основы программирования на Python.

### 1. Цели и задачи

#### Цель дисциплины

В курсе показано, как проходит полный цикл анализа, от сбора данных до выбора оптимального решения и оценки его качества. Студенты научатся пользоваться современными аналитическими инструментами и адаптировать их под особенности конкретных задач.

#### Задачи дисциплины

Студенты освоят основные темы, необходимые в работе с большим массивом данных, в т.ч. современные методы классификации и регрессии, поиск структуры в данных, проведение экспериментов, построение выводов, базовая фундаментальная математика, основы программирования на Python.

### 2. Перечень формируемых компетенций

Освоение дисциплины направлено на формирование следующих компетенций:

Код и наименование компетенции	Индикаторы достижения компетенции
ОПК-3 Способен составлять и оформлять научные и (или) технические (технологические, инновационные) отчеты (публикации, проекты)	ОПК-3.1 Знает основные правила оформления научных публикаций и научно-технической документации, в том числе с использованием прикладного программного обеспечения
	ОПК-3.2 Владеет на практике методологией составления научно-технических отчетов (проектов)
	ОПК-3.3 Владеет методами визуального и графического представления результатов научной (научно-технической, инновационной технологической) деятельности в виде отчетов, научных публикаций
ПК-1 Способен ставить, формализовывать и решать задачи, в том числе разрабатывать и исследовать математические модели изучаемых явлений и процессов, системно анализировать научные проблемы, получать новые научные результаты	ПК-1.1 Способен находить, анализировать и обобщать информацию об актуальных результатах исследований в рамках тематической области своей профессиональной деятельности
	ПК-1.2 Способен выдвигать гипотезы, строить математические модели для описания изучаемых явлений и процессов, оценивать качество разработанной модели
	ПК-1.3 Способен применять теоретические и (или) экспериментальные методы исследований к конкретной научной задаче и интерпретировать полученные результаты

### 3. Перечень планируемых результатов обучения по дисциплине (модулю)

В результате освоения дисциплины обучающиеся должны

знать:

- принципы построения композиций (ансамблей);
- модель случайного леса и метод градиентного бустинга;
- оценивание обобщающей способности алгоритмов;
- подбор параметров модели;
- универсальные методы оценки параметров и проверки гипотез, корреляции и причинно-следственные связи.

уметь:

- строить матричные разложения;
- строить предсказывающие алгоритмы;
- решать задачу тематического моделирования;
- понижать размерность данных;
- искать аномалии;
- визуализировать многомерные данные;
- превращать данные в выводы;
- решать задачи в области анализа текста и информационного поиска, коллаборативной фильтрации и рекомендательных системы, бизнес-аналитики, прогнозировании временных рядов;
- извлекать признаки из разнородных данных;
- сводить задачу заказчика к формальной постановке задачи машинного обучения;
- проверять качество построенной модели на исторических данных и в онлайн-эксперименте.

владеть:

- библиотеками, полезными для анализа данных, например, NumPy, SciPy, Matplotlib и Pandas;
- техникой организации эксперимента;
- техникой A/B-тестирования.

#### 4. Содержание дисциплины (модуля), структурированное по темам (разделам) с указанием отведенного на них количества академических часов и видов учебных занятий

##### 4.1. Разделы дисциплины (модуля) и трудоемкости по видам учебных занятий

№	Тема (раздел) дисциплины	Трудоемкость по видам учебных занятий, включая самостоятельную работу, час.			
		Лекции	Семинары	Лаборат. работы	Самост. работа
1	Обучение на размеченных данных	12	12		12
2	Поиск структуры в данных	8	8		8
3	Математика и Python для анализа данных	5	5		5
4	Глубокое обучение	5	5		5
Итого часов		30	30		30
Подготовка к экзамену		0 час.			
Общая трудоёмкость		90 час., 2 зач.ед.			

##### 4.2. Содержание дисциплины (модуля), структурированное по темам (разделам)

Семестр: 7 (Осенний)

###### 1. Обучение на размеченных данных

Машинное обучение и линейные моды. Борьба с переобучением и оценивание качества  
 Линейные модели: классификация и практические аспекты Решающие деревья и композиции алгоритмов  
 Нейронные сети и обзор методов

###### 2. Поиск структуры в данных

Кластеризация. Понижение размерности и матричные разложения. Визуализация и поиск аномалий. Тематическое моделирование.

###### 3. Математика и Python для анализа данных

Python и Anaconda. Основы математики для машинного обучения. Библиотеки Python и линейная алгебра. Оптимизация и матричные разложения. Случайность. Базовые концепции теории вероятностей и статистики.

#### 4. Глубокое обучение

Основы нейронных сетей и их обучение. Конволюционные нейронные сети (CNN) для обработки изображений. Рекуррентные нейронные сети (RNN) для обработки последовательностей.

#### 5. Описание материально-технической базы, необходимой для осуществления образовательного процесса по дисциплине (модулю)

Необходимое оборудование для лекций и практических занятий: учебная аудитория, оснащенная компьютером и мультимедийным оборудованием (проектор, звуковая система).

#### 6. Перечень рекомендуемой литературы

##### Основная литература

Элементы математической теории машинного обучения [Текст] / В. В. Вьюгин ; М-во образования и науки Рос. Федерации, Моск. физ.-техн. ин-т (гос. ун-т), Ин-т проблем передачи информации им. А. А. Харкевича РАН - М.МФТИ : ИППИ РАН, 2010

##### Дополнительная литература

Введение в анализ данных [Текст], учебник для бакалавриата и магистратуры /Б. Г. Миркин; НИУ "Высшая школа экономики". -М., Юрайт, 2018

#### 7. Перечень ресурсов информационно-телекоммуникационной сети "Интернет", необходимых для освоения дисциплины (модуля)

Не используются

#### 8. Перечень информационных технологий, используемых при осуществлении образовательного процесса по дисциплине (модулю), включая перечень необходимого программного обеспечения и информационных справочных систем (при необходимости)

Необходимое программное обеспечение: текстовый редактор, Anaconda.

#### 9. Методические указания для обучающихся по освоению дисциплины (модуля)

Студент, изучающий дисциплину, должен с одной стороны, овладеть общим понятийным аппаратом, а с другой стороны, должен научиться применять теоретические знания на практике.

Успешное освоение дисциплины требует:

- посещения студентом всех видов аудиторных занятий;
- ведения конспекта в ходе лекционных занятий;
- качественной самостоятельной подготовки к практическим занятиям, активной работы на них;
- активной самостоятельной и аудиторной работы студента;
- своевременной сдачи преподавателю заданий по аудиторным видам работ.

**ОЦЕНОЧНЫЕ МАТЕРИАЛЫ ПО ДИСЦИПЛИНЕ (МОДУЛЮ)**

<b>по направлению:</b>	Прикладная математика и информатика
<b>профиль подготовки:</b>	АІ360: Передовые методы искусственного интеллекта Физтех-школа Прикладной Математики и Информатики кафедра дискретной математики
<b>курс:</b>	<u>4</u>
<b>квалификация:</b>	бакалавр
Семестр, формы промежуточной аттестации: 7 (осенний) - Дифференцированный зачет	
<b>Разработчик:</b>	А.М. Райгородский, д-р физ.-мат. наук, профессор, ассистент

## 1. Компетенции, формируемые в процессе изучения дисциплины

Код и наименование компетенции	Индикаторы достижения компетенции
ОПК-3 Способен составлять и оформлять научные и (или) технические (технологические, инновационные) отчеты (публикации, проекты)	ОПК-3.1 Знает основные правила оформления научных публикаций и научно-технической документации, в том числе с использованием прикладного программного обеспечения
	ОПК-3.2 Владеет на практике методологией составления научно-технических отчетов (проектов)
	ОПК-3.3 Владеет методами визуального и графического представления результатов научной (научно-технической, инновационной технологической) деятельности в виде отчетов, научных публикаций
ПК-1 Способен ставить, формализовывать и решать задачи, в том числе разрабатывать и исследовать математические модели изучаемых явлений и процессов, системно анализировать научные проблемы, получать новые научные результаты	ПК-1.1 Способен находить, анализировать и обобщать информацию об актуальных результатах исследований в рамках тематической области своей профессиональной деятельности
	ПК-1.2 Способен выдвигать гипотезы, строить математические модели для описания изучаемых явлений и процессов, оценивать качество разработанной модели
	ПК-1.3 Способен применять теоретические и (или) экспериментальные методы исследований к конкретной научной задаче и интерпретировать полученные результаты

## 2. Показатели оценивания компетенций

В результате изучения дисциплины «Машинное обучение и анализ данных» обучающийся должен:

### знать:

- принципы построения композиций (ансамблей);
- модель случайного леса и метод градиентного бустинга;
- оценивание обобщающей способности алгоритмов;
- подбор параметров модели;
- универсальные методы оценки параметров и проверки гипотез, корреляции и причинно-следственные связи.

### уметь:

- строить матричные разложения;
- строить предсказывающие алгоритмы;
- решать задачу тематического моделирования;
- понижать размерность данных;
- искать аномалии;
- визуализировать многомерные данные;
- превращать данные в выводы;
- решать задачи в области анализа текста и информационного поиска, коллаборативной фильтрации и рекомендательных системы, бизнес-аналитики, прогнозировании временных рядов;
- извлекать признаки из разнородных данных;
- сводить задачу заказчика к формальной постановке задачи машинного обучения;
- проверять качество построенной модели на исторических данных и в онлайн-эксперименте.

### владеть:

- библиотеками, полезными для анализа данных, например, NumPy, SciPy, Matplotlib и Pandas;
- техникой организации эксперимента;
- техникой А/В-тестирования.

## 3. Перечень типовых (примерных) вопросов, заданий, тем для подготовки к текущему контролю

С целью контроля освоения обучающимися учебного материала проводится устный опрос в начале занятия по теме прошлой лекции или в конце занятия по пройденной теме.

#### **4. Перечень типовых (примерных) вопросов и тем для проведения промежуточной аттестации обучающихся**

1. Что такое машинное обучение и какие его основные виды вы знаете?
2. Опишите принципы работы алгоритмов машинного обучения, в том числе supervised, unsupervised и reinforcement learning.
3. Какие основные этапы включает в себя процесс разработки модели машинного обучения?
4. В чем разница между классификацией и регрессией? Приведите примеры задач для каждого типа.
5. Что такое переобучение и недообучение? Как можно справиться с этими проблемами?
6. Объясните концепцию кросс-валидации и ее роль в оценке модели.
7. Какие метрики используются для оценки точности моделей машинного обучения?
8. Что такое оптимизация и какие алгоритмы используются для ее реализации?
9. Каковы основные типы данных и как они влияют на выбор алгоритмов машинного обучения?
10. Опишите принципы работы нейронных сетей и их применение в машинном обучении.
11. Как можно преобразовать неструктурированные данные (например, текст) в формат, пригодный для машинного обучения?
12. Какие методы предобработки данных применяются в машинном обучении?
13. Как выбрать подходящий алгоритм машинного обучения для решения конкретной задачи?
14. Опишите процесс развертывания обученной модели машинного обучения в реальном мире.
15. Как машинное обучение может использоваться для решения задач в вашей области (например, в медицине, финансах, маркетинге)?

#### **Критерии оценивания**

Оценка "Отлично" (10) - полностью и вовремя решены все задачи без ошибок. Продemonстрирован грамотный подход к решению задач, реализованы оптимальные алгоритмы, код оформлен в едином удобочитаемом стиле.

Оценка "Отлично" (9) - полностью и вовремя решены все задачи без ошибок. Продemonстрирован грамотный подход к решению задач, реализованы оптимальные алгоритмы.

Оценка "Отлично" (8) - полностью и вовремя решены все задачи без ошибок. Продemonстрирован грамотный подход к решению задач.

Оценка "Хорошо" (7) - полностью решены все задачи. Допущены несущественные ошибки.

Оценка "Хорошо" (6) - полностью решено большинство задач. В некоторых задачах допущены и не исправлены ошибки, либо некоторые задачи решены частично.

Оценка "Хорошо" (5) - полностью решено две трети задач. В некоторых задачах допущены и не исправлены ошибки, либо некоторые задачи решены частично.

Оценка "Удовлетворительно" (4) - полностью решено более половины задач. В остальных задачах допущены и не исправлены ошибки, либо некоторые задачи решены частично.

Оценка "Удовлетворительно" (3) - полностью решено более половины задач.

Оценка "Неудовлетворительно" (2) - решено менее половины задач.

Оценка "Неудовлетворительно" (1) - не решено ни одной задачи.

#### **5. Методические материалы, определяющие процедуры оценивания знаний, умений, навыков и (или) опыта деятельности**

Дифференцированный зачет может проводиться по итогам текущей успеваемости и сдачи заданий и других видов работ, предусмотренных программой дисциплины и (или) путем организации специального опроса, проводимого в устной и (или) письменной форме.

При проведении устного дифференцированного зачета обучающемуся предоставляется 30 минут на подготовку. Опрос обучающегося не должен превышать одного астрономического часа.

Во время проведения дифференцированного зачета обучающиеся могут пользоваться программой дисциплины, а также справочной литературой, конспектами лекций или другими материалами.