

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

**УТВЕРЖДЕНО**  
**Проректор по учебной работе**

**А.А. Воронов**

	<b>Рабочая программа дисциплины (модуля)</b>
<b>по дисциплине:</b>	Язык Python и библиотеки обработки данных
<b>по направлению:</b>	Прикладные математика и физика
<b>профиль подготовки:</b>	Управление инновациями в бизнесе Физтех-школа бизнеса высоких технологий кафедра информатики и вычислительной математики
<b>курс:</b>	2
<b>квалификация:</b>	бакалавр

Семестр, формы промежуточной аттестации: 3 (осенний) - Дифференцированный зачет

Аудиторных часов: 60 всего, в том числе:

лекции: 20 час.

семинары: 0 час.

лабораторные занятия: 40 час.

Самостоятельная работа: 75 час.

Всего часов: 135, всего зач. ед.: 3

Количество контрольных работ, заданий: 2

Программу составил: Т.Ф. Хирьянов, старший преподаватель

Программа обсуждена на заседании кафедры информатики и вычислительной математики 27.04.2022

## Аннотация

Курс направлен на изучение возможностей языка Python 3 и среды Jupyter для агрегации данных и разведочного анализа данных.

В частности, происходит изучение инструментария библиотек Matplotlib, NumPy, Pandas, BeautifulSoup.

### 1. Цели и задачи

#### Цель дисциплины

Углублённое изучение языка Python 3 в среде Jupyter, стандартных модулей парсинга и агрегации данных, библиотек Matplotlib, NumPy и Pandas.

#### Задачи дисциплины

1. освоить работу на Python 3 в среде JupyterLab;
2. изучить возможности библиотеки Matplotlib по визуализации данных;
3. изучить возможности библиотеки NumPy по работе с массивами и матрицами;
4. изучить возможности библиотеки Pandas по работе с табличными данными;
5. изучить стандартные модули Python 3 по парсингу данных из веб-страниц и текстов;
6. изучить продвинутый синтаксис Python 3 как функционального и объектно-ориентированного языка.

### 2. Перечень формируемых компетенций

Освоение дисциплины направлено на формирование следующих компетенций:

Код и наименование компетенции	Индикаторы достижения компетенции
ОПК-4 Способен осуществлять сбор и обработку научно-технической и (или) технологической информации для решения фундаментальных и прикладных задач	ОПК-4.1 Владеет методами научного поиска и интеллектуального анализа информации при решении задач профессиональной деятельности
ПК-1 Способен планировать и проводить научные эксперименты (в избранной предметной области) и (или) теоретические (аналитические и имитационные) исследования	ПК-1.8 Владеет навыками работы с современными языками программирования и программными пакетами для научных расчетов
ПК-3 Способен выбирать и применять подходящее оборудование, инструменты и методы исследований для решения задач в избранной предметной области	ПК-3.2 Знает области и критерии применимости используемых теоретических подходов и умение оценивать точность приближенных аналитических методов вычислений

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

В результате освоения дисциплины обучающиеся должны

знать:

- синтаксические конструкции функционального программирования на Python 3;
- синтаксические конструкции ООП на Python 3;
- возможности научных библиотек Python по анализу и визуализации данных.

уметь:

- работать в средах Jupyter Notebook и JupyterLab;
- создавать программы на языке Python в том числе в формате Jupyter Notebook;
- использовать Pandas, Numpy и другие научные библиотеки для анализа данных;
- пользоваться разметкой Markdown для создания ячеек-пояснений в Jupyter;
- пользоваться LaTeX для написания формул;
- визуализировать данные и результаты анализа.

владеть:

- инструментарием языка Python и научных библиотек для анализа данных на практике.

### 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	Использование Jupyter Notebook и JupyterLab	2		4	5
2	Основы NumPy	4		8	20
3	Основы Pandas	4		8	20
4	Визуализация данных и зависимостей в Matplotlib и Seaborn	4		8	10
5	Парсинг данных регулярными выражениями и BeautifulSoup	4		8	10
6	Продвинутый синтаксис Python	2		4	10
Итого часов		20		40	75
Подготовка к экзамену		0 час.			
Общая трудоёмкость		135 час., 3 зач.ед.			

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

Семестр: 3 (Осенний)

##### 1. Использование Jupyter Notebook и JupyterLab

Установка и запуск Jupyter Notebook и JupyterLab.

Принципы использования Jupyter. Когда он подходит, а когда нет.

Создание ячеек и их порядок.

Синтаксис Markdown текстовых ячеек.

Управление подсветкой синтаксиса вставок кода на разных языках программирования.

Вставка изображений и графиков.

Синтаксис ввода формул LaTeX в ячейках Jupyter.

##### 2. Основы NumPy

Установка и подключение NumPy.

Массивы ndarray: отличие от списков list и стандартных массивов array.

Простые типы данных NumPy. Фиксированное число бит для чисел.

Способы создания массивов NumPy.

Векторные операции с массивами.

Срезы массивов NumPy.

Выборка элементов по логическому критерию.

Матричные операции в NumPy.

Линейная алгебра в NumPy.

##### 3. Основы Pandas

Установка и подключение Pandas.

Типы Series и DataFrame для работы с сериями и таблицами данных.

Индексация серий и фреймов. Локаторы loc и iloc. Срезы по индексам.  
Векторные операции с сериями. Логические операции &, | и особенности их приоритета.  
Выборка строк по логическому условию. Метод query.  
Статистика данных в таблице. Перцентили, медиана, среднее, отклонение. Гистограммы.  
Функции агрегации данных. Группировка по категориальным параметрам.

#### 4. Визуализация данных и зависимостей в Matplotlib и Seaborn

Установка и подключение Matplotlib и Seaborn.  
Типы графиков, диаграмм, гистограмм. Адекватность их применения для визуализации данных.  
Управление цветами, видами линий, подписями на графиках.  
Трёхмерные графики.  
Анимация графиков.

#### 5. Парсинг данных регулярными выражениями и BeautifulSoup

Основы разметки веб-страниц HTML и описание структуры гипертекстовых документов.  
Установка и подключение библиотеки BeautifulSoup.  
Основы парсинга страниц HTML при помощи BeautifulSoup.  
Регулярные выражения в Python. Поиск необходимых подстрок по шаблону в сыром тексте.  
Формирование листа Pandas с данными на основе данных на веб-страницах.

#### 6. Продвинутый синтаксис Python

Итерируемые объекты. Генераторы и итераторы. Ключевое слово yield.  
Библиотека itertools. Сопроцессы.  
Декораторы функций.  
Объекты и классы. Атрибуты и методы. Конструктор.  
Обработка исключений. Инструкции try, except, finally.

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

Большая лекционная аудитория, подходящая для учебного потока (факультет, оснащённая мультимедиа проектором и экраном для чтения лекций.  
Учебные аудитории — сетевые компьютерные классы с установленным необходимым программным обеспечением.

### 6. Перечень рекомендуемой литературы

#### Основная литература

1. Python 3. Самое необходимое / Н. А. Прохоренко, В. А. Дронов, Санкт-Петербург, БХВ, 2021
2. Алгоритмы. Руководство по разработке [Текст], [учеб. пособие для вузов] /С. Скиена ; [пер. с англ. С. Таранушенко], The Algorithm, esign Manual. -СПб., БХВ-Петербург, 2018

#### Дополнительная литература

1. Программирование на Python 3, подробное руководство/М. Саммерфилд,-СПб, Символ-Плюс, 2020
2. Дискретная математика для программистов, учебное пособие / Р. Хаггарт . — Москва, Техносфера, 2012.— URL: <https://ibooks.ru/bookshelf/337430/reading> (дата обращения: 26.11.2020). - Полный текст (Режим доступа : из сети МФТИ / Удаленный доступ)
3. Алгоритмы и программы на языках С и PYTHON. Сортировка. Поиск. Строки, Электронная версия печатной публикации / В. В. Прут. — Москва, МФТИ, 2020

## **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)**

На ПК в компьютерных классах должно быть установлено следующее ПО:

1. Операционная система GNU/Linux;
2. Интерпретатор Python версии не ниже 3.9;
3. Среда разработки IDLE;
4. Среды JupyterLab, Jupyter Notebook, Ipython;
5. Библиотеки Numpy, Pandas, xlrd, NetworkX, Matplotlib, Seaborn и PyGame для Python 3;
6. Среда разработки JetBrains Python Charm community edition;

На лекциях используются мультимедийные технологии, включая демонстрацию презентаций.

Для контроля и коррекции знаний обучающихся используются автоматизированное компьютерное тестирование на основе Ejudge или CMS Moodle.

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

Изложение материала происходит преимущественно на лекциях, сопровождается мультимедиа-презентацией с примерами кода и блок-схемами алгоритмов. На лабораторных занятиях также происходит изложение нового материала: в начале каждой лабораторной работы и далее по мере необходимости. На контрольных работах изложение нового материала исключено, преподаватель оказывает только консультации по условиям задач.

Учёт, контроль и оценка знаний студентов

В течение лабораторной работы успеваемость отслеживается по результатам и своевременности сдачи лабораторных работ. Таким образом достигается раннее выявление отстающих студентов с передачей докладных в деканат.

Посещаемость лекций не отмечается, но каждая лабораторная работа завязана на материал прошедшей лекции, что делает посещение лекций насущной необходимостью в течение семестра.

Дифференцированный зачёт принимается в устной форме, при этом учитываются оценки по контрольным и оценки по практическим лабораторным работам. Устный ответ практически исключает списывание, показывает владение базовой терминологией предмета, а также позволяет проверить знание концепций и подходов к анализу данных.

Самостоятельная домашняя работа предполагается после каждой лабораторной работы.

**ОЦЕНОЧНЫЕ МАТЕРИАЛЫ ПО ДИСЦИПЛИНЕ (МОДУЛЮ)**

**по направлению:** Прикладные математика и физика  
**профиль подготовки:** Управление инновациями в бизнесе  
Физтех-школа бизнеса высоких технологий  
кафедра информатики и вычислительной математики  
**курс:** 2  
**квалификация:** бакалавр

Семестр, формы промежуточной аттестации: 3 (осенний) - Дифференцированный зачет

**Разработчик:** Т.Ф. Хирьянов, старший преподаватель

## 1. Компетенции, формируемые в процессе изучения дисциплины

Код и наименование компетенции	Индикаторы достижения компетенции
ОПК-4 Способен осуществлять сбор и обработку научно-технической и (или) технологической информации для решения фундаментальных и прикладных задач	ОПК-4.1 Владеет методами научного поиска и интеллектуального анализа информации при решении задач профессиональной деятельности
ПК-1 Способен планировать и проводить научные эксперименты (в избранной предметной области) и (или) теоретические (аналитические и имитационные) исследования	ПК-1.8 Владеет навыками работы с современными языками программирования и программными пакетами для научных расчетов
ПК-3 Способен выбирать и применять подходящее оборудование, инструменты и методы исследований для решения задач в избранной предметной области	ПК-3.2 Знает области и критерии применимости используемых теоретических подходов и умение оценивать точность приближенных аналитических методов вычислений

## 2. Показатели оценивания компетенций

В результате изучения дисциплины «Язык Python и библиотеки обработки данных» обучающийся должен:

### знать:

- синтаксические конструкции функционального программирования на Python 3;
- синтаксические конструкции ООП на Python 3;
- возможности научных библиотек Python по анализу и визуализации данных.

### уметь:

- работать в средах Jupyter Notebook и JupyterLab;
- создавать программы на языке Python в том числе в формате Jupyter Notebook;
- использовать Pandas, Numpy и другие научные библиотеки для анализа данных;
- пользоваться разметкой Markdown для создания ячеек-пояснений в Jupyter;
- пользоваться LaTeX для написания формул;
- визуализировать данные и результаты анализа.

### владеть:

- инструментарием языка Python и научных библиотек для анализа данных на практике.

## 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 прошлого занятия.

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

1. Рекурсивные сортировки. Быстрая сортировка. Сортировка слиянием.
2. Пирамида (куча). Пирамидальная сортировка.
3. Устойчивость сортировок.
4. Списки: односвязный, двусвязный, кольцо.
5. стек. Дек.
6. Очередь.
7. Очередь с приоритетами. Пирамида (куча).
8. Хеш-функция. Хеширование. Открытая и закрытая хеш-таблица.
9. Графы и способы их представления: список рёбер, матрица смежности, списки смежности
10. Определение дерева. Поиск в глубину.
11. Связность неориентированных графов: выделение компонент связности.

12. Поиск в ширину. Алгоритм Дейкстры.
13. Эйлеров цикл. Эйлеров путь.
14. Взвешенный граф. Кратчайшее расстояние между двумя вершинами.
15. Алгоритм Флойда-Уоршелла.
16. Минимальное остовное дерево. Алгоритм Прима.
17. Проверка изоморфизма графов.
18. Построение гамильтонова цикла.
19. Задача о коммивояжере
20. Орграфы. Топологическая сортировка.
21. Проверка равенства строк. Простой и вероятностный алгоритмы.
22. Поиск подстроки в строке. Алгоритм Рабина-Карпа.
23. Алгоритм Кнута-Морриса-Пракса.
24. Z-алгоритм.
25. Конечный автомат для поиска подстрок и регулярных выражений.

#### Критерии оценивания

Оценка по десятибалльной шкале за работу на лабораторном практикуме выставляется преподавателем практикума исходя из количества и качества выполненных практических работ за семестр.

Оценка за выполнение контестов выставляется автоматически исходя из суммарного рейтинга обучающегося в системе Ejudge и также нормируется к десятибалльной шкале.

Итоговая оценка за зачёт не должна отличаться от среднего арифметического оценок по контестам и по практическим лабораторным работам более чем на три балла.

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

Дифференцированный зачёт принимается в устной форме с учётом оценки по контестам и оценки по лабораторному практикуму. Устный ответ практически исключает списывание, показывает владение базовой терминологией предмета, умение говорить на языке информатики, а также позволяет проверить знание сложных алгоритмов, которые долго программируются, но могут быть относительно легко устно объяснены.

На дифференцированном зачёте предлагается ответить на два-три вопроса по теории и решить одну короткую алгоритмическую задачу на бумаге без использования компьютера.

Пример задания на устном зачёте:

1. Очередь.
2. Эйлеров цикл. Эйлеров путь.
3. Задача: реализовать алгоритм Рабина-Карпа с полиномиальной хеш-функцией.



### 3. Перечень типовых контрольных заданий, используемых для оценки знаний, умений, навыков

Итоговая аттестация по дисциплине «Язык Python и библиотеки обработки данных» осуществляется в форме дифференцированного зачета. Оценка за зачёт выставляется как взвешенная сумма оценок лабораторных работ, выполняемых в течение семестра.

#### Пример лабораторной работы по анализу данных

Загрузите датасет `titanic.csv` и, используя описанные выше способы работы с данными, найдите ответы на вопросы

1. Какое количество мужчин и женщин ехало на корабле? В качестве ответа приведите два числа через пробел.
2. Какой части пассажиров удалось выжить? Посчитайте долю выживших пассажиров. Ответ приведите в процентах (число в интервале от 0 до 100, знак процента не нужен), округлив до двух знаков.
3. Какую долю пассажиры первого класса составляли среди всех пассажиров? Ответ приведите в процентах (число в интервале от 0 до 100, знак процента не нужен), округлив до двух знаков.
4. Какого возраста были пассажиры? Посчитайте среднее и медиану возраста пассажиров. В качестве ответа приведите два числа через пробел.
5. Коррелируют ли число братьев/сестер/супругов с числом родителей/детей? Посчитайте корреляцию Пирсона между признаками `SibSp` и `Parch`.
6. Какое самое популярное женское имя на корабле? Извлеките из полного имени пассажира (колонка `Name`) его личное имя (`First Name`).

Это задание — типичный пример того, с чем сталкивается специалист по анализу данных. Данные очень разнородные и шумные, но из них требуется извлечь необходимую информацию. Попробуйте вручную разобрать несколько значений столбца `Name` и выработать правило для извлечения имен, а также разделения их на женские и мужские.

Если ответом является нецелое число, то целую и дробную часть необходимо разграничивать точкой, например, 0.42. При необходимости округляйте дробную часть до двух знаков.

## Пример дополнительных вопросов при сдаче лабораторной работы

1. Выберите верные утверждения:

- Объекты описываются с помощью признаков
- Одна из задач машинного обучения — научиться делать прогнозы для объектов
- Одна из задач машинного обучения — научиться делать прогнозы для признаков
- Признаки описываются с помощью объектов

2. Что из этого — корректные названия типов признаков?

- Устойчивые признаки
- Нетривиальные признаки
- Номинальные (категориальные) признаки
- Числовые (количественные) признаки
- Бинарные признаки

3. Какие из этих задач являются задачами классификации?

- Прогноз оценки студента по пятибалльной шкале на экзамене по машинному обучению в следующей сессии
- Поиск групп похожих пользователей интернет-магазина
- Разделение книг, хранящихся в электронной библиотеке, на научные и художественные
- Прогноз температуры на следующий день

4. Какая из этих фраз наиболее точно описывает переобучение?

- Переобучение — это ситуация, в которой алгоритм выдает недетерминированные ответы на новых данных (то есть при разных запусках на одном и том же объекте можно получить разные предсказания)
- Переобучение — это ситуация, в которой алгоритм показывает одинаково плохое качество и на обучающей выборке, и на новых данных
- Переобучение — это ситуация, в которой алгоритм часто отказывается от построения прогноза на новых данных.
- Переобучение — это ситуация, в которой алгоритм показывает хорошее качество на обучающей выборке, но при этом плохо работает на новых данных

#### 4. Критерии оценивания

Оценка	Баллы	Критерии
отлично	10	Обучающийся ответил на все вопросы, но не с первой попытки.
	9	Обучающийся допустил не более одной ошибки или воспользовался помощью преподавателя.
	8	Обучающийся работая самостоятельно, допустил не более двух численных ошибок в лабораторной работе.
хорошо	7	Обучающийся если он ответил на подавляющее большинство вопросов в лабораторной работе, может быть с помощью преподавателя или товарищей.
	6	Обучающийся ответил на подавляющее большинство вопросов в лабораторной работе, может быть с помощью преподавателя или товарищей.
	5	Обучающийся ответил на основные вопросы в лабораторной работе, может быть с помощью преподавателя или товарищей.
удовлетворительно	4	Обучающийся ответил на основные вопросы в лабораторной работе;
	3	Обучающийся ответил на некоторые вопросы в лабораторной работе.
неудовлетворительно	2	Обучающийся не справился с работой.
	1	Обучающийся демонстрирует полное отсутствие знаний по предмету или пытался выдать чужую работу за свою.