

МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ

**Федеральное государственное автономное  
образовательное учреждение высшего образования  
«Национальный исследовательский Нижегородский государственный университет  
им. Н.И. Лобачевского»**

Институт биологии и биомедицины

(факультет / институт / филиал)

УТВЕРЖДЕНО  
решением ученого совета ННГУ  
протокол от  
«25» января 2023 г. № 1

**Рабочая программа дисциплины**

Базы данных и основы научного  
программирования

(наименование дисциплины (модуля))

Уровень высшего образования  
магистратура

(бакалавриат / магистратура / специалитет)

Направление подготовки / специальность  
19.04.01 Биотехнология

(указывается код и наименование направления подготовки / специальности)

Направленность образовательной программы  
Общая биотехнология

(указывается профиль / магистерская программа / специализация)

Форма обучения

очная

(очная / очно-заочная / заочная)

Нижний Новгород

2023 год

## 1. Место дисциплины в структуре ООП

Дисциплина Б1.О.07 Базы данных и основы научного программирования относится к обязательной части, Блока 1 ООП направления подготовки 19.04.01 «Биотехнология».

## 2. Планируемые результаты обучения по дисциплине, соотнесенные с планируемыми результатами освоения образовательной программы (компетенциями и индикаторами достижения компетенций)

Формируемые компетенции (код, содержание компетенции)	Планируемые результаты обучения по дисциплине (модулю), в соответствии с индикатором достижения компетенции		Наименование оценочного средства
	Индикатор достижения компетенции (код, содержание индикатора)	Результаты обучения по дисциплине	
<b>ОПК-3.</b> Способен разрабатывать алгоритмы и участвовать в разработке программ в сфере своей профессиональной деятельности	<b>ОПК-3.1</b> Знаком с типичными алгоритмами разработки программ в сфере своей профессиональной деятельности.	<i>Знать</i> основы научного программирования <i>Уметь</i> составлять алгоритмы разработки программ в сфере своей профессиональной деятельности <i>Владеть</i> навыками составления алгоритмов при разработке программ в сфере своей профессиональной деятельности	Вопросы к устному опросу
	<b>ОПК-3.2.</b> Использует информационные и телекоммуникационные технологии сбора, размещения, хранения, накопления, преобразования и передачи данных в профессионально-ориентированных информационных системах.	<i>Знать</i> информационные и телекоммуникационные технологии сбора, размещения, хранения, накопления, преобразования и передачи данных в профессионально-ориентированных информационных системах. <i>Уметь</i> проводить поиск и анализ информации в профессионально-ориентированных информационных системах <i>Владеть</i> навыками сбора, размещения, хранения, накопления, преобразования и передачи данных в профессионально-ориентированных информационных системах.	Практические задания
	<b>ОПК-3.3.</b> Имеет опыт разработки алгоритмов, программ в сфере своей профессиональной	<i>Знать</i> алгоритмы составления программ в сфере своей профессиональной деятельности <i>Уметь</i> разрабатывать алгоритмы программ в сфере своей профессиональной деятельности	Практические задания

	деятельности	<b>Владеть</b> навыками разработки алгоритмов программ в сфере своей профессиональной деятельности	
<b>ОПК-4.</b> Способен выбирать и использовать современные инструментальные методы и технологии, осваивать новые методы и технику исследований для решения конкретных задач профессиональной деятельности	<b>ОПК-4.1.</b> Понимает, может описать принципы и перечислить современные инструментальные методы и технологии, необходимые для решения задач профессиональной деятельности.	<b>Знать</b> современные методические подходы для решения задач профессиональной деятельности. <b>Уметь</b> использовать современные инструментальные методы и технологии, необходимые для решения задач профессиональной деятельности. <b>Владеть</b> навыками использования современных инструментальных методов и технологий, необходимых для решения задач профессиональной деятельности	Доклад с презентацией  Тест
	<b>ОПК-4.2.</b> Применяет современные инструментальные методы и технологии, необходимые для решения задач профессиональной деятельности.	<b>Знать</b> современные инструментальные методы и технологии, необходимые для решения задач профессиональной деятельности <b>Уметь</b> использовать современные инструментальные методы и технологии, необходимые для решения задач профессиональной деятельности. <b>Владеть</b> навыкам использования современных инструментальных методов и технологий, необходимых для решения задач профессиональной деятельности.	Практические задания
	<b>ОПК-4.3.</b> Может осваивать новые методы и техники исследований, необходимые для решения задач профессиональной деятельности.	<b>Знать</b> основные достижения и проблемы новых методов и техник исследований <b>Уметь</b> проводить поиск и анализ информации о новых методах и техник исследований <b>Владеть</b> навыками освоения методов и техник исследований для решения задач профессиональной деятельности	Практические задания

### 3. Структура и содержание дисциплины

#### 3.1 Трудоемкость дисциплины

	<b>очная форма обучения</b>
<b>Общая трудоемкость</b>	<b>3 ЗЕТ</b>
<b>Часов по учебному плану</b>	<b>108</b>
<b>в том числе</b>	
<b>аудиторные занятия (контактная работа):</b>	<b>42</b>

- занятия лекционного типа	14
- лабораторные работы	28
- практические занятия	65
самостоятельная работа	1
КСР	зачет
Промежуточная аттестация – Зачет/экзамен	

### 3.2. Содержание дисциплины

Наименование и краткое содержание разделов и тем дисциплины (модуля),	Всего (часы)	в том числе				
		контактная работа (работа во взаимодействии с преподавателем), часы				Самостоятельная работа обучающегося, часы
		из них				
		Занятия лекционного типа	Занятия лабораторного типа	Занятия семинарского типа	Всего	
	Очная	Очная	Очная	Очная	Очная	Очная
Тема 1. Основы статистики в Python. Описательная статистика. Построение диаграмм размаха, гистограмм, функций плотности вероятности, планок погрешностей.	12	2		4	6	6
Тема 2. Основы статистики в Python. Корреляционный анализ. Проверка статистических гипотез.	10	2		2	4	6
Тема 3. Численное решение дифференциальных уравнений, решение систем линейных уравнений с помощью библиотек SciPy и NumPy. Методы Монте-Карло.	10	2		2	4	6
Тема 4. Базы данных Введение в язык баз данных SQL. Элементы проектирования баз данных. Системы управления базами данных. Физическая организация данных и методы доступа. Обеспечение защиты данных в БД. Перспективные направления развития БД	25	2		6	8	17
Тема 5. Методы предобработки данных. Метод k ближайших соседей в задачах классификации и восстановления регрессии. Метрики качества алгоритмов машинного обучения.	14	2		6	8	6
Тема 6. Экспериментальная оценка качества обучения. Метод перекрестного контроля. Методы снижения размерности: метод главных компонент.	10	2		2	4	6
Тема 7. Понятие регрессии. Линейная регрессия. Метод наименьших квадратов.	10	2		2	4	6

Проверка статистической значимости модели. Борьба с переобучением в задаче восстановления регрессии: сокращение числа признаков, регуляризация.						
<i>Тема 8. Логистическая регрессия. Линейный и квадратичный дискриминантный анализ. Машина опорных векторов.</i>	8			2	2	6
<i>Тема 9. Деревья решений. Алгоритм CART. Ансамбли решающих правил. Баггинг, бустинг.</i>	8			2	2	6
Текущий контроль	1				1	
<b><i>Промежуточная аттестация - зачет</i></b>						
<b><i>Итого</i></b>	108	14		28	42	65

Текущий контроль успеваемости реализуется в рамках занятий семинарского типа.

#### **4. Учебно-методическое обеспечение самостоятельной работы обучающихся**

Цель самостоятельной работы - подготовка современного компетентного специалиста и формирование способностей и навыков к непрерывному самообразованию и профессиональному совершенствованию.

Основные виды самостоятельной работы студентов в рамках освоения дисциплины:

- изучение понятийного аппарата дисциплины;
- работа над основной и дополнительной литературой
- изучение материала по темам дисциплины в сети Интернет;
- подготовка к практическим занятиям;
- изучение тем самостоятельной подготовки и подготовка доклада с презентацией;
- подготовка к устным опросам;
- подготовка к зачету.

Самостоятельная работа является наиболее деятельным и творческим процессом, который выполняет ряд дидактических функций: способствует формированию диалектического мышления, вырабатывает высокую культуру умственного труда, совершенствует способы организации познавательной деятельности, воспитывает ответственность, целеустремленность, систематичность и последовательность в работе студентов, развивает у них бережное отношение к своему времени, способность доводить до конца начатое дело.

##### **• Изучение понятийного аппарата дисциплины.**

Вся система индивидуальной самостоятельной работы должна быть подчинена усвоению понятийного аппарата, поскольку одной из важнейших задач подготовки современного грамотного специалиста является овладение и грамотное применение профессиональной терминологии. Лучшему усвоению и пониманию дисциплины помогут учебники, монографии, справочники и интернет ресурсы, указанные в списке литературы.

##### **• Работа над основной и дополнительной литературой**

Изучение рекомендованной литературы следует начинать с учебников и учебных пособий, затем переходить к научным монографиям и материалам периодических изданий.

Студент должен уметь самостоятельно подбирать необходимую для учебной и научной работы литературу. При этом следует обращаться к предметным каталогам и библиографическим справочникам, которые имеются в библиотеках.

Для аккумуляции информации по изучаемым темам рекомендуется формировать личный архив, а также каталог используемых источников, что может использоваться не только в рамках данного курса, но и для последующей подготовке к итоговой аттестации на выпускном курсе.

- ***Самоподготовка к практическим занятиям***

При подготовке к практическому занятию необходимо помнить, что данная дисциплина тесно связана с ранее изучаемыми дисциплинами.

На практических занятиях студент должен уметь последовательно излагать свои мысли и аргументировано их отстаивать.

Для достижения этой цели необходимо:

- 1) ознакомиться с соответствующей темой программы изучаемой дисциплины;
- 2) осмыслить круг изучаемых вопросов и логику их рассмотрения;
- 3) изучить рекомендованную учебно-методическим комплексом литературу по данной теме, составить конспект; ознакомиться с нормативными документами;
- 4) тщательно изучить лекционный материал;
- 5) ознакомиться с вопросами очередного практического занятия;
- 6) подготовить сообщение по каждому из вынесенных на практическое занятие вопросу.

Изучение вопросов очередной темы требует глубокого усвоения теоретических основ дисциплины, раскрытия сущности основных положений, проблемных аспектов темы и анализа фактического материала.

- ***Самостоятельная работа студента при подготовке к промежуточной аттестации:***

Промежуточной формой контроля успеваемости студентов является зачет.

Для успешного прохождения промежуточной аттестации рекомендуется в начале семестра изучить перечень вопросов к зачету по данной дисциплине, а также использовать в процессе обучения материалы, разработанные в ходе подготовки к практическим занятиям. Это позволит в процессе изучения тем сформировать более правильное и обобщенное видение существа того или иного вопроса за счет:

- 1) уточняющих вопросов преподавателю;
- 2) самостоятельного уточнения вопросов на смежных дисциплинах;
- 3) углубленного изучения вопросов темы по учебным пособиям.

- ***Изучение сайтов по темам дисциплины в сети Интернет***

Ресурсы Интернет являются одним из альтернативных источников быстрого поиска требуемой информации. Их использование возможно для получения основных и дополнительных сведений по изучаемым материалам.

Самостоятельная работа по освоению материала проводится к практическим занятиям семинарского типа (лабораторные занятия не предусмотрены) с привлечением конспектов лекций, знаний, полученных на предыдущих практических занятиях, основной и дополнительной литературы по всем темам курса. Кроме того, самостоятельная работа студентов по разделам включает подготовку к устным опросам и семинарским занятиям.

- ***Изучение тем самостоятельной подготовки и подготовка доклада с презентацией.***

Особое место отводится самостоятельной проработке студентами отдельных разделов и тем по изучаемой дисциплине. При докладе с презентацией на практическом занятии можно воспользоваться следующим алгоритмом изложения темы: название, актуальность исследования, цели и задачи предмета исследования, оценка современного состояния

вопроса, используемые материалы и методы исследования, выводы, перспективы развития и возможности внедрения. Время доклада – 7-10 минут. Презентация должна быть выполнена в программе PowerPoint. Презентация должна быть хорошо иллюстрирована (рисунками, схемами, таблицами), логически согласована с рефератом. Желательно свободное изложение материала без зачитывания печатного текста.

Контрольные вопросы и задания для проведения текущего контроля и промежуточной аттестации по итогам освоения дисциплины приведены в п. 5.2.

## 5. Фонд оценочных средств для промежуточной аттестации по дисциплине (модулю), включающий:

### 5.1 Описание шкал оценивания результатов обучения по дисциплине

Уровень сформированности компетенций (индикатора достижения компетенций)	Шкала оценивания сформированности компетенций						
	плохо	неудовлетворительно	удовлетворительно	хорошо	очень хорошо	отлично	превосходно
	не зачтено		зачтено				
<u>Знания</u>	Отсутствие знаний теоретического материала.  Невозможность оценить полноту знаний вследствие отказа обучающегося от ответа	Уровень знаний ниже минимальных требований. Имели место грубые ошибки.	Минимально допустимый уровень знаний. Допущено много негрубых ошибок.	Уровень знаний в объеме, соответствующем программе подготовки. Допущено несколько негрубых ошибок	Уровень знаний в объеме, соответствующем программе подготовки. Допущено несколько несущественных ошибок	Уровень знаний в объеме, соответствующем программе подготовки, без ошибок.	Уровень знаний в объеме, превышающем программу подготовки.
<u>Умения</u>	Отсутствие минимальных умений . Невозможность оценить наличие умений вследствие отказа обучающегося от ответа	При решении стандартных задач не продемонстрированы основные умения.  Имели место грубые ошибки.	Продemonstrированы основные умения. Решены типовые задачи с негрубыми ошибками. Выполнены все задания, но не в полном объеме.	Продemonstrированы все основные умения. Решены все основные задачи с негрубыми ошибками. Выполнены все задания, в полном объеме, но некоторые с недочетами.	Продemonстрированы все основные умения. Решены все основные задачи . Выполнены все задания, в полном объеме, но некоторые с недочетами.	Продemonстрированы все основные умения, решены все основные задачи с отдельными несущественными недочетами, выполнены все задания в полном объеме.	Продemonстрированы все основные умения, решены все основные задачи. Выполнены все задания, в полном объеме без недочетов
<u>Навыки</u>	Отсутствие владения материалом. Невозможность оценить	При решении стандартных задач не продемонстрированы	Имеется минимальный набор	Продemonстрированы базовые навыки	Продemonстрированы базовые навыки	Продemonстрированы навыки при решении	Продemonстрирован творческий подход к решению

	наличие навыков вследствие отказа обучающегося от ответа	базовые навыки. Имели место грубые ошибки.	навыков для решения стандартных задач с некоторыми недочетами	при решении стандартных задач с некоторыми недочетами	при решении стандартных задач без ошибок и недочетов.	нестандартных задач без ошибок и недочетов.	нестандартных задач
--	--	--	---	---	---	---	---------------------

### Шкала оценки при промежуточной аттестации

Оценка		Уровень подготовки
	<b>превосходно</b>	Все компетенции (части компетенций), на формирование которых направлена дисциплина, сформированы на уровне не ниже «превосходно», продемонстрированы знания, умения, владения по соответствующим компетенциям на уровне, выше предусмотренного программой
<b>зачтено</b>	<b>отлично</b>	Все компетенции (части компетенций), на формирование которых направлена дисциплина, сформированы на уровне не ниже «отлично», при этом хотя бы одна компетенция сформирована на уровне «отлично»
	<b>очень хорошо</b>	Все компетенции (части компетенций), на формирование которых направлена дисциплина, сформированы на уровне не ниже «очень хорошо», при этом хотя бы одна компетенция сформирована на уровне «очень хорошо»
	<b>хорошо</b>	Все компетенции (части компетенций), на формирование которых направлена дисциплина, сформированы на уровне не ниже «хорошо», при этом хотя бы одна компетенция сформирована на уровне «хорошо»
	<b>удовлетворительно</b>	Все компетенции (части компетенций), на формирование которых направлена дисциплина, сформированы на уровне не ниже «удовлетворительно», при этом хотя бы одна компетенция сформирована на уровне «удовлетворительно»
<b>не зачтено</b>	<b>неудовлетворительно</b>	Хотя бы одна компетенция сформирована на уровне «неудовлетворительно», ни одна из компетенций не сформирована на уровне «плохо»
	<b>плохо</b>	Хотя бы одна компетенция сформирована на уровне «плохо»

## 5.2 Типовые контрольные задания или иные материалы, необходимые для оценки результатов обучения.

### 5.2.1 Контрольные вопросы к зачету

#### для оценки сформированности компетенции ОПК-3:

1. Основы статистики в Python. Типы данных. Описательные статистики. Построение диаграмм размаха, гистограмм, функций плотности вероятности, планок погрешностей.
2. Основы статистики в Python. Корреляционный анализ. Проверка статистических гипотез. FDR-коррекция.
3. Численное решение дифференциальных уравнений и их систем с помощью библиотек SciPy и NumPy. Пример биологической системы, моделируемой дифференциальными уравнениями.



4. Решение систем линейных алгебраических уравнений с помощью библиотек SciPy и NumPy. Метод Монте-Карло.
5. Постановка задачи машинного обучения. Основные классы задач в машинном обучении. Основные проблемы машинного обучения. Примеры практических задач.
6. Вероятностная постановка задачи обучения с учителем. Наивный байесовский классификатор. Сглаживание Лапласа.
7. Методы предобработки данных. Метод k ближайших соседей в задачах классификации и восстановления регрессии.
8. Экспериментальная оценка качества обучения и выбор параметров модели. Обучающая, проверочная и тестовая выборки. Метод перекрестного (скользящего) контроля. Метрики качества алгоритмов машинного обучения.
9. Метод сокращения размерности: метод главных компонент.
10. Линейная регрессия. Метод наименьших квадратов. Проверка статистической значимости модели. Коэффициент детерминации.
11. Методы борьбы с переобучением в задаче восстановления регрессии. Отбор признаков. Регуляризация.
12. Логистическая регрессия. Логистическая функция и softmax.
13. Линейный дискриминантный анализ. Квадратичный дискриминантный анализ.
14. Машина опорных векторов. Оптимальная разделяющая гиперплоскость. Опорные векторы. Случай линейно-разделимых и неразделимых классов. Ядра и спрямляющие пространства. Случай с более чем двумя классами.
15. Деревья решений. Алгоритм CART. Борьба с переобучением.
16. Ансамбли решающих правил. Баггинг. Случайный лес. Экстремально случайные деревья.
17. Ансамбли решающих правил. Бустинг. AdaBoost. Градиентный бустинг деревьев решений.
18. Нейронные сети. Персептрон Розенблатта. Алгоритм обратного распространения ошибки. Борьба с переобучением.
19. Основные понятия доверенного искусственного интеллекта.
20. Обучение без учителя. Задача кластеризации. Метод центров тяжести. Метод медоидов. Алгоритмы иерархической кластеризации.

**для оценки сформированности компетенции ОПК-4:**

21. Автоматизированные информационные системы (АИС), основанные на данных. Предметная область АИС. Классификация АИС.
22. Компоненты системы баз данных. Уровни представления данных. Физическая и логическая независимость данных.
23. Реляционная модель данных (РМД). Структуризация данных в РМД. Основные операции. Ограничения целостности. Достоинства и недостатки РМД.
24. Системы управления базами данных (СУБД). Назначение СУБД. Классификация СУБД. Основные функции СУБД.
25. Системы управления базами данных (СУБД). Требования к реляционным СУБД (по Кодду).
26. Структура памяти и структура хранимых данных. Управление свободным пространством памяти.
27. Способы доступа к данным. Индексирование данных. Способы организации индексов.
28. Защита данных от сбоев. Защита данных от несанкционированного доступа.

29. Требования к проекту базы данных. Этапы проектирования базы данных.
30. Инфологическое и логическое проектирование базы данных.

### **5.2.2. Типовые вопросы для устного опроса для оценки сформированности компетенции ОПК-3:**

1. Какие типы переменных встречаются при анализе научных данных?
2. Какая выборка называется репрезентативной?
3. Что такое статистические выбросы?
4. Какие существуют основные описательные статистики?
5. Что представляет собой диаграмма размаха («ящик с усами», boxplot)?
6. Что такое гистограмма и функция плотности вероятности?
7. Для чего на графиках строятся планки погрешностей (error bar)?
8. Что такое корреляция и какие типы корреляции бывают?
9. Какой график используется для графического представления корреляционной связи?
10. Что такое ковариация?
11. В чём заключается разница между коэффициентами корреляции Пирсона и Спирмена?
12. Какие критерии нормальности вам известны?
13. В чём заключается суть дисперсионного анализа (ANOVA)?
14. Что такое FDR-коррекция?
15. Какие задачи машинного обучения являются задачами обучения с учителем?
16. Что такое нормализация данных?
17. Что такое переобучение?
18. В чём заключается суть метода k ближайших соседей?
19. Для чего используются обучающие и тестовые выборки?
20. В чём заключается суть метода перекрёстного контроля?
21. Какие существуют эвристические подходы для выбора количества главных компонент?
22. Для чего используется метод наименьших квадратов?
23. Что такое коэффициент детерминации?
24. Какие существуют причины переобучения в задаче восстановления регрессии?
25. Какие предположения задаются в линейном дискриминантном анализе?
26. Что такое логистическая регрессия?
27. Что такое оптимальная разделяющая гиперплоскость?
28. В чём заключается суть использования дерева решений?
29. Какие существуют алгоритмы построения деревьев решений и в чём их отличие?
30. В чём заключается суть баггинга?
31. Чем отличаются экстремально случайные деревья от случайного леса?
32. В чём заключается суть бустинга?
33. Что такое нейронная сеть?
34. Что такое глубокое обучение?
35. Какая задача является задачей обучения без учителя?
36. В чём заключается суть метода центров тяжести?
37. Чем метод медоидов отличается от метода центров тяжести?
38. Что такое иерархическая кластеризация?

### **5.2.3 Типовые темы докладов с презентацией для оценки сформированности компетенции ОПК-4:**

- Модели фотосинтетических процессов лежащие в основе автоматизированного анализа данных РАМ-флуориметрии с использованием программного обеспечения Dual-RAM-100 (Heinz Walz GmbH, Германия).
- Модели фотосинтетических процессов лежащие в основе автоматизированного анализа данных ЛР-теста с использованием программного обеспечения M-PEA-2 (Hansatech Instruments Ltd, Великобритания).

- Модели фотосинтетических и транспирационных процессов лежащие в основе автоматизированного анализа данных с использованием программного обеспечения инфракрасного газоанализатора GFS-3000 (Heinz Walz GmbH, Германия).
- Современные подходы для автоматизированного выявления параметров ответа мембранного потенциала в электрофизиологических измерениях на клеточном уровне.
- уровне.

### 3.2.4. Тесты для оценки компетенции «ОПК-4»

Что такое переменная в программировании?

- A) Инструкция, которая выполняет действие
- B) Способ хранения данных, которые могут изменяться\*
- C) Фиксированное значение данных

Какой тип данных используется для хранения текста в Python?

- A) int
- B) float
- C) string\*

Что делает цикл for в Python?

- A) Повторяет блок кода определённое количество раз\*
- B) Выполняет условие, если оно истинно
- C) Создаёт новую функцию

Какой оператор используется для сравнения двух значений на равенство?

- A) =
- B) ==\*
- C) ===

### 5.2.5. Типовые практические задания для оценки сформированности компетенции ОПК-3

Практические задания предполагают решение задач в области научного программирования и машинного обучения (написание соответствующих программ с использованием необходимых библиотек языка Python в приложении Jupyter Notebook).

*Задание 1.*

Загрузите из файла heart.csv данные о сердечных заболеваниях. Выполните следующие подзадания:

- 1) Сколько образцов (объектов) содержит данный датасет?
- 2) Сколько атрибутов (признаков) содержит данный датасет? Подробно опишите значение каждого признака.
- 3) Опишите тип каждого признака (числовой / дискретный / непрерывный / категориальный / номинальный / бинарный / ординальный)?
- 4) Вычислите, сколько мужчин/женщин в датасете?
- 5) Вычислите описательные статистики для количественных признаков (среднее значение, медиана, мода, размах, дисперсия, среднеквадратичное отклонение, 1й/2й/3й квартили, межквартильный размах).
- 6) Постройте гистограммы для признаков age, trestbps, chol, thalach, oldpeak. Расположите гистограммы на одном графике в одну линию. Подпишите оси каждой гистограммы.
- 7) Постройте диаграммы размаха для признаков age, trestbps, chol, thalach, oldpeak.
- 8) Постройте на одном графике две кривые PDF (probability density function) для признака chol. Одна PDF для мужчин, другая - для женщин. Подпишите оси, добавьте легенду.
- 9) Для признаков, которые не были указаны в п.6-7, постройте полигоны частот (подграфики расположите в нескольких строках и столбцах). Подпишите оси.

10) Сгруппируйте данные по полу и вычислите для каждой группы среднее значение признака chol, применив функцию агрегации. Изобразите результаты в виде столбчатой диаграммы, где столбцы должны соответствовать полу, а высота столбцов - соответствующим средним значениям признака chol. Добавьте к каждому столбцу планку погрешности, отражающую среднее квадратичное отклонение.

11) Постройте следующие диаграммы рассеяния:

- trestbps от age
- chol от age
- thalach от age
- oldpeak от age

Изобразите точки на диаграммах разными цветами в зависимости от пола. Подпишите оси и добавьте легенду. Попробуйте визуально определить, коррелируют ли рассматриваемые переменные с возрастом. Проверьте свои предположения, вычислив коэффициенты корреляции Спирмена. Сделайте выводы.

12) Проверьте признаки age, trestbps, chol, thalach, oldpeak на нормальность с помощью критерия Шапиро-Уилка.

## Задание 2.

Загрузите датасет ISOLET (<https://archive.ics.uci.edu/ml/datasets/ISOLET>). Выполните следующие подзадания:

- 1) Опишите рассматриваемый датасет.
- 2) Выполните нормализацию признаков.
- 3) Разбейте данные на обучающую и тестовую выборку.
- 4) Вызовите метод  $k$  ближайших соседей. Постройте графики зависимости ошибки этого метода на обучающей и тестовой выборках от  $k$ . Сделайте вывод.
- 5) Выполните процедуру перекрестного контроля (5-fold, 10-fold, LOO) с обучающей выборкой. Постройте графики зависимости CV-ошибки от числа используемых соседей в методе  $k$  ближайших соседей. Выберите наилучшую модель и проверьте ее качество на тестовой выборке.
- 6) Примените к рассматриваемым данным
  - Линейный дискриминантный анализ
  - Квадратичный дискриминантный анализ
  - Логистическую регрессию

Для каждого метода вычислите ошибки на обучающей и тестовой выборках.

- 7) На рассматриваемых данных обучите следующие классификаторы:
  - Random Forest
  - Extremely Random Trees

Постройте графики зависимости ошибки на обучающей и тестовой выборке от количества используемых деревьев.

- 8) Натренируйте на рассматриваемых данных нейронную сеть с одним скрытым слоем ("vanilla") с 200 нейронами в нем (`hidden_layer_sizes = (200,)`). В качестве функции активации используйте положительную срезку (`activation = 'relu'`). Постройте графики зависимости ошибки на обучающей и тестовой выборке от параметра  $\alpha$  (weight decay).

## 5.2.6. Типовые практические задания для оценки сформированности компетенции ОПК-4

**Задание 1.** Вручную сделать глобальное выравнивание нуклеотидных последовательностей ATGAGTCTCT и CTGTCTCCTG, используя матрицу штрафов DNAfull и линейный штраф за гэп равный 10. Проверить результат с помощью EMBOSS Needle.

**Задание 2.** Построить глобальное выравнивание нуклеотидных последовательностей человеческого и мышинного гена, кодирующего бета-актин. Построить выравнивание аминокислотных последовательностей соответствующих белков. Объяснить получившуюся разницу.

**Задание 3.** Написать функцию, принимающую на вход нуклеотидную последовательность и возвращающую ее обратно-комплементарную.

**Задание 4.** Используя Python, построить хеш-таблицу (dict) для позиций  $k$ -мер'ов в геноме SARS-CoV-2 ( $k$  – переменная).

**Задание 5.** Посчитать экзонные длины генов GNG4, SPRR4.

**Задание 6.** Провести вычисления метода median-of-ratios для матрицы TCGA (без использования DESeq2), убедиться, что результат совпал с DESeq2.

**Задание 7.** Посчитать нормировочные множители для матрицы TCGA с помощью библиотеки edgeR; выяснить, сильно ли отличаются коэффициенты от DESeq2.

**Задание 8.** Вывести топ-50 наиболее дифференциально экспрессированных генов по  $t$ -критерию Стьюдента (после нормализации FPKM на sizeFactors и логарифмирования).

**Задание 9.** Главный комплекс гистосовместимости человека кодируется генами HLA-A, HLA-B и HLA-C (класс I) и HLA-DRB1, HLA-DQB1 и HLA-DPB1 (класс II). С использованием корреляционного анализа убедиться, что фактор транскрипции SP1 активирует данные гены.

**Задание 10.** Даны гены, значимо изменившие свою экспрессию в клеточной линии HT-29 при некотором воздействии. Проанализировав функциональную принадлежность данных генов, предположить, что это было за воздействие.

## **6. Учебно-методическое и информационное обеспечение дисциплины**

### **а) основная литература:**

1. Анализ данных : учебник / В. С. Мхитарян [и др.] ; под редакцией В. С. Мхитаряна. - Москва : Юрайт, 2023. - 490 с. - (Высшее образование). - ISBN 978-5-534-00616-2. - Текст : электронный // <https://urait.ru/bcode/511020>
2. Маккинли У. Python и анализ данных. – М.: ДМК Пресс, 2015. - 481 с. – ISBN: 978-5-9706031-5-4. Текст : электронный // <https://www.studentlibrary.ru/book/ISBN9785970603154.html>

### **б) дополнительная литература:**

1. Богданов, Е. П. Интеллектуальный анализ данных : практикум для магистрантов направления 09.04.03 «Прикладная информатика» профиль подготовки «Информационные системы и технологии корпоративного управления» / Е. П. Богданов. - Волгоград : ФГБОУ ВО Волгоградский ГАУ, 2019. - 112 с. - Текст : электронный. - URL: <https://znanium.com/catalog/product/1087885>

### **в) Интернет ресурсы**

1. Курс лекций по машинному обучению Н.Ю. Золотых. Режим доступа: <http://www.uic.unn.ru:8103/~zny/ml/>.
2. Информационно-аналитический ресурс, посвященный машинному обучению, распознаванию образов и интеллектуальному анализу данных. Режим доступа: <http://www.machinelearning.ru>.
3. Документация библиотеки SciPy. Режим доступа: <https://docs.scipy.org/doc/scipy/>.
4. Документация библиотеки scikit-learn. Режим доступа: [https://scikit-learn.org/stable/user\\_guide.html](https://scikit-learn.org/stable/user_guide.html).

### **г) программное обеспечение**

1. Anaconda3 (дистрибутив Python)
2. Jupyter Notebook (интерактивная оболочка Python)

## **7. Материально-техническое обеспечение дисциплины**

Учебная аудитория для проведения занятий лекционного типа, занятий семинарского типа, групповых и индивидуальных консультаций, текущего контроля и промежуточной аттестации. Учебная мебель, доска, компьютерные терминалы, экран, проектор, плоттер, принтер, беспроводной Интернет, лицензионное программное обеспечение.

Помещения для самостоятельной работы обучающихся оснащены компьютерной техникой с возможностью подключения к сети "Интернет" и обеспечением доступа в электронную информационно-образовательную среду организации.

Программа составлена в соответствии с требованиями ФГОС ВО по направлению 19.04.01 Биотехнология.

Автор к.ф.-м.н., М.Ю. Кириллин

Рецензент к.б.н. Синицына Ю.В.

Заведующий кафедрой прикладной математики ИТММ М.Ю. Иванченко

Программа одобрена на заседании методической комиссии ИББМ от «6» сентября 2022 года, протокол № 1.