

# РАБОЧАЯ ПРОГРАММА

модуля(курса)

«Основы технологии машинного обучения и искусственного интеллекта»

## 1. АННОТАЦИЯ

Цель - формирование компетенций, необходимых для выполнения нового вида профессиональной деятельности в области создания алгоритмов и компьютерных программ, пригодных для практического применения.

В ходе изучения дисциплины решаются следующие задачи:

- овладение навыками программирования на Python;
- овладение основами технологии машинного обучения и искусственного интеллекта.

## 2.СОДЕРЖАНИЕ

Учебная программа курса

№ п/п	Наименование модуля, разделов и тем	Содержание обучения (по темам в дидактических единицах), наименование и тематика лабораторных работ, практических занятий (семинаров), самостоятельной работы с указанием кол-ва часов, используемых образовательных технологий и рекомендуемой литературы
1.	2.	3.
	<b>Раздел 1.</b>	<b>Программирование на Python</b>
1	Тема 1. Введение в Python и Jupyter Notebook.	Язык программирования Python. Среда Jupyter Notebook. Установка дистрибутива Anaconda (или другой эквивалентной системы). Работа в Google Colab (или другой эквивалентной среде). (0,5 часа)
2	Тема 2. Контейнеры в языке Python.	Списки (list), кортежи (tuple) и словари (dict). Использование контейнеров для обработки текстовых данных. (0,5 часа)
3	Тема 3. Основные конструкции языка Python.	Синтаксические конструкции языка питон для организации ветвлений и циклов и привить навыки работы с ними. Булевский (логический) тип данных. Логические выражения. Ветвления (конструкции if, if else, if elif else). Цикл for. Цикл while. Команды break и continue. (0,5 часа)
4	Тема 4. Функции, модули и библиотеки в языке Python.	Концепция функции, модуля и библиотеки, формальных параметров, область видимости переменных. (0,5 часа)
5	Тема 5. Работа с текстовыми данными.	Простейшие функции работы с текстовыми строками и текстовыми файлами. (1 час)
6	Тема 6. Введение в объектно-ориентированное программирование.	Основные понятия объектно-ориентированного программирования (инкапсуляция, наследование, полиморфизм). Классы, объекты, методы. Создание своих классов. (1 час)
	Тема 7. Работа с табличными данными.	Особенности работы с табличными данными, работы с функциями библиотеки Pandas. Библиотека Pandas и работа с табличными данными. Обзор библиотеки Pandas. Возможности библиотеки для решения задач обработки данных.(1 час)
	<b>Раздел 2.</b>	<b>Основы технологии машинного обучения и искусственного интеллекта</b>
	Тема 1. Введение в искусственный	Задачи и методы искусственного интеллекта, история его развития. Концепции: сильный (общий) и слабый (прикладной) искусственный интеллект; ИИ, основанный на правилах

интеллект и машинное обучение.	(дедуктивный подход) и ИИ, основанный на машинном обучении (дедуктивный подход); обучение с учителем и без учителя; задачи классификации, регрессии, предсказания временного ряда, кластеризации. Перспективы развития ИИ; обзор основных достижений ИИ, в том числе в решении задач обработки естественного языка. Примеры задач машинного обучения; связь между машинным обучением обработкой данных. Концепции недообучения и переобучения, их возможные причины. Специфические вопросы, относящиеся к методам машинного обучения для решения задач обработки естественного языка (классификация задач, метод “мешок слов”, векторные представления и др.) (1 час)
Тема 2. Задача обучения с учителем.	Методы решения задач обучения с учителем. Линейная регрессия и метод наименьших квадратов; метод ближайшего соседа. Методы, основанные на построении деревьев решений и их ансамблей. Логистическая регрессия и нейронные сети. Метод обратного распространения ошибки (backpropagation). (1 час)
Тема 3. Введение в библиотеку scikit-learn и решение с помощью нее задач обучения с учителем.	Обзор библиотеки. Основные методы. Пример решения задач классификации и регрессии (в том числе из области обработки естественного языка и лингвистики) для обучения (настраивания) модели, тестирования и верификации. (0,5 часа)
Тема 4. Задача обучения без учителя.	основных методах решения задач обучения без учителя. Дается представление о методе k-средних и DBSCAN для решения задач кластеризации. Ставится задача понижения размерности и дается представление о методе главных компонент. (0,5 часа)
Тема 5. Решения задач обучения без учителя с помощью библиотеки scikit-learn.	Работа в библиотеке scikit-learn для решения задач кластеризации и понижения размерности (0,5 часа)
Тема 6. Глубокое обучение.	Принципы глубокого обучения, в том числе для решения задач обработки естественного языка и обработки неструктурированных данных. (0,5 часа)
Тема 7. Методы глубокого обучения в библиотеке TensorFlow.	Решение задач ИИ в библиотеке TensorFlow (или другой эквивалентной), включая задачи компьютерной лингвистики и обработки естественного языка, методами глубокого обучения. (1 часа)
Практические занятия (семинары)	Практические занятия по темам лекций (24 часа)
Самостоятельная работа	Выполнение домашних заданий по теме занятия (24 часа)
Зачет	Собеседование (2 часа)

### 3. ОЦЕНКА КАЧЕСТВА ОСВОЕНИЯ ПРОГРАММЫ

*(формы аттестации, оценочные и методические материалы)*

Практические задания с использованием необходимых библиотек языка Python в приложении Jupyter Notebook).

1. Загрузите из файла heart.csv (<https://www.kaggle.com/ronitf/heart-diseaseuci?select=heart.csv>) данные о сердечных заболеваниях.

Выполните следующие задания:

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, среднееквадратичное 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 среднееквадратичное 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

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

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).

Кейсы по машинному обучению для проведения практических занятий

1.	Отток	пользователей	Telecom
<a href="https://github.com/Ninjalemur/telecom_users/blob/main/telecom_users.csv">https://github.com/Ninjalemur/telecom_users/blob/main/telecom_users.csv</a>			

2. Оценка стоимости квартиры  
<https://www.kaggle.com/datasets/hugoncosta/price-of-flats-in-moscow>

3. Распознавание фейковых новостей в социальных медиа  
<https://www.kaggle.com/datasets/mdepak/fakenewsnet>

Устный опрос

1. Какие основные типы графиков используются для визуализации научных данных?
2. Какие типы переменных встречаются при анализе научных данных?
3. Какая выборка называется репрезентативной?
4. Что такое статистические выбросы?
5. Какие существуют основные описательные статистики?
6. Что представляет собой диаграмма размаха («ящик с усами», `boxplot`)?
7. Что такое гистограмма и функция плотности вероятности?
8. Для чего на графиках строятся планки погрешностей (`error bar`)?
9. Что такое корреляция и какие типы корреляции бывают?
10. Какой график используется для графического представления корреляционной связи?
11. Что такое ковариация?
12. В чём заключается разница между коэффициентами корреляции Пирсона и Спирмена?
13. Какие критерии нормальности вам известны?
14. В чём заключается суть дисперсионного анализа (ANOVA)?
15. Что такое FDR-коррекция?
16. Какие задачи машинного обучения являются задачами обучения с учителем?
17. Что такое нормализация данных?
18. Что такое переобучение?
19. В чём заключается суть метода  $k$  ближайших соседей?

Примеры вопросов к устным опросам для оценки знаний компетенции ПК-3

1. Для чего используются обучающие и тестовые выборки?
2. В чём заключается суть метода перекрёстного контроля?
3. В чём заключается суть метода главных компонент?
4. Какие существуют эвристические подходы для выбора количества главных компонент?
5. Для чего используется метод наименьших квадратов?
6. Что такое коэффициент детерминации?
7. Какие существуют причины переобучения в задаче восстановления регрессии?
8. Какие предположения задаются в линейном дискриминантном анализе?
9. Что такое логистическая регрессия?
10. Что такое оптимальная разделяющая гиперплоскость?
11. В чём заключается суть использования дерева решений?
12. Какие существуют алгоритмы построения деревьев решений и в чём их отличие?
13. Что такое энтропия?
14. В чём заключается суть баггинга?
15. Чем отличаются экстремально случайные деревья от случайного леса?
16. В чём заключается суть бустинга?
17. Что такое нейронная сеть?
18. Что такое метод обратного распространения ошибки?
19. Что такое глубокое обучение?

20. Какая задача является задачей обучения без учителя?
21. В чём заключается суть метода центров тяжести?
22. Чем метод медоидов отличается от метода центров тяжести?
23. Что такое иерархическая кластеризация?

Промежуточная аттестация. Перечень вопросов для подготовки дифференцированному зачету

1. Способы визуализации научных данных с использованием Matplotlib. Иерархическая структура рисунка в Matplotlib. Основные элементы графика. Основные типы графиков.
2. Основы статистики в Python. Описательные статистики. Построение диаграмм размаха, гистограмм, функций плотности вероятности, планок погрешностей.
3. Основы статистики в Python. Корреляционный анализ. Проверка статистических гипотез. FDR-коррекция.
4. Численное решение дифференциальных уравнений, решение систем линейных уравнений с помощью библиотек SciPy и NumPy. Методы Монте-Карло.
5. Постановка задачи машинного обучения. Основные классы задач в машинном обучении. Примеры практических задач.
6. Вероятностная постановка задачи обучения с учителем. Функция потерь. Средний риск. Эмпирический риск.
7. Экспериментальная оценка качества обучения и выбор параметров модели. Обучающая, проверочная и тестовая выборки. Метод перекрестного (скользящего) контроля. Переобучение и недообучение.
8. Наивный байесовский классификатор. Сглаживание Лапласа. Ошибки 1-го и 2-го рода. Чувствительность, специфичность, точность, полнота. ROC-кривая. Использование наивного байесовского классификатора для количественных признаков.
9. Метод k ближайших соседей в задачах классификации и восстановления регрессии.
10. Методы предобработки данных. Методы понижения размерности. Метод главных компонент. Сингулярное разложение.
11. Линейная регрессия. Метод наименьших квадратов. Проверка статистической значимости модели. Коэффициент детерминации.
12. Методы борьбы с переобучением в задаче восстановления регрессии. Отбор признаков. Регуляризация.
13. Линейный дискриминантный анализ. Квадратичный дискриминантный анализ.
14. Логистическая регрессия. Логистическая функция и softmax.
15. Машина опорных векторов. Оптимальная разделяющая гиперплоскость. Опорные векторы. Случаи линейно-разделимых и неразделимых классов. Ядра и спрямляющие пространства.
16. Деревья решений. Алгоритм CART.
17. Ансамбли решающих правил. Баггинг. Случайный лес. Экстремально случайные деревья.
18. Ансамбли решающих правил. Бустинг. AdaBoost. Градиентный бустинг деревьев решений.
19. Нейронные сети. Персептрон Розенблатта. Сети прямого распространения. Алгоритм обратного распространения ошибки.
20. Глубокое обучение. Сверточные нейронные сети.
21. Обучение без учителя. Задача кластеризации. Метод центров тяжести. Метод медоидов.
22. Алгоритмы кластеризации: EM-алгоритм, алгоритм DBSCAN. Алгоритмы иерархической кластеризации.

Формы и методы контроля и оценки результатов освоения модуля

№ п/п	Наименование процедуры	Основные показатели оценки	Формы и методы контроля и оценки
1	Промежуточная аттестация.	Владеет конструкциями программирования на Python и Jupyter Notebook . Владеет основами	Дифференцированный зачет/лабораторная работа

	4. Анализ данных и машинное обучение	технологии машинного обучения и искусственного интеллекта	
--	--------------------------------------	---	--

#### Критерии оценки

№ п/п	Наименование процедуры	Основные показатели оценки	Формы и методы контроля и оценки
1	Основы программирования в системе «1С:Предприятие 8	<p>Зачтено. Уровень знаний в объеме, соответствующем программе подготовки. Фрагментарные, либо сформированные, но содержащие отдельные пробелы знания о программировании на Python и Jupyter Notebook . Владеет основами технологии машинного обучения и искусственного интеллекта</p> <p>Незачтено. Уровень знаний ниже минимальных требований. Имели место грубые ошибки.</p> <p>Отсутствие знаний о программировании на Python и Jupyter Notebook . Владеет основами технологии машинного обучения и искусственного интеллекта</p>	Дифференцированный зачет/Лабораторная работа

#### 4. УСЛОВИЯ РЕАЛИЗАЦИИ ПРОГРАММЫ МОДУЛЯ

4.1. Помещения представляют собой учебные аудитории для проведения учебных занятий, предусмотренных программой (лекционного и семинарского типа), оснащенные оборудованием и техническими средствами обучения.

Помещения для самостоятельной работы обучающихся оснащены компьютерной техникой с возможностью подключения к сети "Интернет" и обеспечены доступом в электронную информационно-образовательную среду

Обучение данной дисциплине организовано следующим образом:

1. Предусмотрены 2 вида занятий: лекционные и занятия семинарского типа (практические занятия). Лекции читаются для всего потока одновременно. Для организации аудиторной практики, а также контроля самостоятельной работы слушателя занятия ведутся в учебных группах.

2. В рамках лекционных занятий основное внимание уделяется формированию у студентов целостного понимания истории развития, текущего состояния и перспективных направлений в области базового программирования, развитию кругозора. Лекционное время преимущественно расходуется не на разбор синтаксиса конкретных конструкций языков программирования, а на изучение основных принципов и концепций, лежащих в основе языков и технологий программирования, прагматике использования тех или иных алгоритмических и технических решений (в пределах рамок курса).

3. В рамках аудиторных практических занятий основное внимание уделяется изучению синтаксиса языков, их основных элементов, со студентами обсуждаются возможные способы решения типовых учебных задач, проводится их сравнительный анализ, прототипируются наиболее важные решения, ставятся задачи для самостоятельной работы.

4. Самостоятельная работа студентов в ходе всего учебного года предполагает выполнение ряда лабораторных работ. При этом в каждой лабораторной работе студенты проходят весь путь, начиная от постановки учебной задачи до сдачи преподавателю работающей программы с краткой документацией.

5. Основная сложность, стоящая перед преподавателем, заключается в, как правило, весьма сильном разбросе в начальном уровне подготовки студентов в области информатики и программирования. Для того чтобы преодолеть указанную проблему, в рамках лекционных занятий дается материал разного уровня. Значительная часть материала ориентирована на тех, кто изучает данный предмет впервые. Вместе с тем, во многих лекциях встречаются блоки повышенного уровня сложности, рассчитанные на наиболее подготовленных студентов. На практике для работы со студентами, начальный уровень подготовки которых значительно превышает средний, предусмотрены дополнительные задания повышенной сложности.

#### **4.2. Перечень основной и дополнительной литературы:**

а) основная литература:

1. МакГрат М. (2021) Программирование на Python для начинающих Эксмо
2. Géron Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems. 2nd Edition. O'Reilly Media. 2019. Рус. пер. 1-го издания: О. Жерон Прикладное машинное обучение с помощью Scikit-Learn и TensorFlow. ООО Диалектика. 2018
3. A.C. Müller, S. Guido Introduction to Machine Learning with Python: A Guide for Data Scientists. Рус. пер.: А. Мюллер, С. Гвидо Введение в машинное обучение с помощью Python. Вильямс, 2017.

б) дополнительная литература:

4. Bhasin H. (2019). Python Basics : A Self-Teaching Introduction. Dulles, Virginia: Mercury Learning & Information. <http://search.ebscohost.com/login.aspx?direct=true&site=eds-live&db=edsebk&AN=1991381>
5. G. James, D. Witten, T. Hastie, R. Tibshirani An Introduction to Statistical Learning with Applications in R. Springer, 2013. <https://faculty.marshall.usc.edu/gareth-james/ISL/>  
Рус. пер.: Г. Джеймс, Д. Уиттон, Т. Хасте, Р. Тибишрани Введение в статистическое обучение с примерами на языке R. ДМК Пресс, 2016.
6. I. Goodfellow, Y. Bengio, A. Courville Deep Learning The MIT Press 2016. Рус. пер.: И. Бенджио, Я. Гудфеллоу, А. Курвилль Глубокое обучение. ДМК Пресс, 2017
7. Материалы курса лекций «Основы программирования»: НОУ ИНТУИТ:
8. <http://www.intuit.ru/studies/courses/2193/67/info>, режим доступа – свободный.

в) программное обеспечение и Интернет-ресурсы

9. Anaconda: the Most World's Most Popular Data Science Platform <https://www.anaconda.com/>
10. scikit-learn: Machine Learning in Python <https://scikit-learn.org/>
11. TensorFlow: Комплексная платформа машинного обучения с открытым исходным кодом <https://www.tensorflow.org/>
12. pyTorch: An open source machine learning framework that accelerates the path from research prototyping to production deployment. <https://pytorch.org/>

#### **4.3. МАТЕРИАЛЬНО-ТЕХНИЧЕСКОЕ ОБЕСПЕЧЕНИЕ ДИСЦИПЛИНЫ (МОДУЛЯ)**

Помещения представляют собой учебные аудитории для проведения учебных занятий, предусмотренных программой, оснащенные оборудованием и техническими средствами обучения: компьютер преподавателя с возможностью подключения к сети Интернет, экран для демонстрации и проектор, компьютеры для студентов с возможностью подключения к сети Интернет.

Помещения для самостоятельной работы обучающихся оснащены компьютерной техникой с возможностью подключения к сети «Интернет» и обеспечением доступа в электронную

информационно-образовательную среду организации.

Каждый обучающийся в течение всего периода обучения обеспечен индивидуальным неограниченным доступом к электронно-библиотечным системам (электронным библиотекам) («Консультант студента», «Лань», «Znanium», «Юрайт») и к электронной информационно-образовательной среде организации (portal.unn.ru), в системе электронного обучения ННГУ <https://e-learning.unn.ru/>. Данные электронно-библиотечные системы (электронные библиотеки) и электронная информационно-образовательная среда обеспечивают возможность доступа обучающегося из любой точки, в которой имеется доступ к информационно-телекоммуникационной сети «Интернет», как на территории организации (в библиотеке ИЭП ННГУ), так и вне ее.