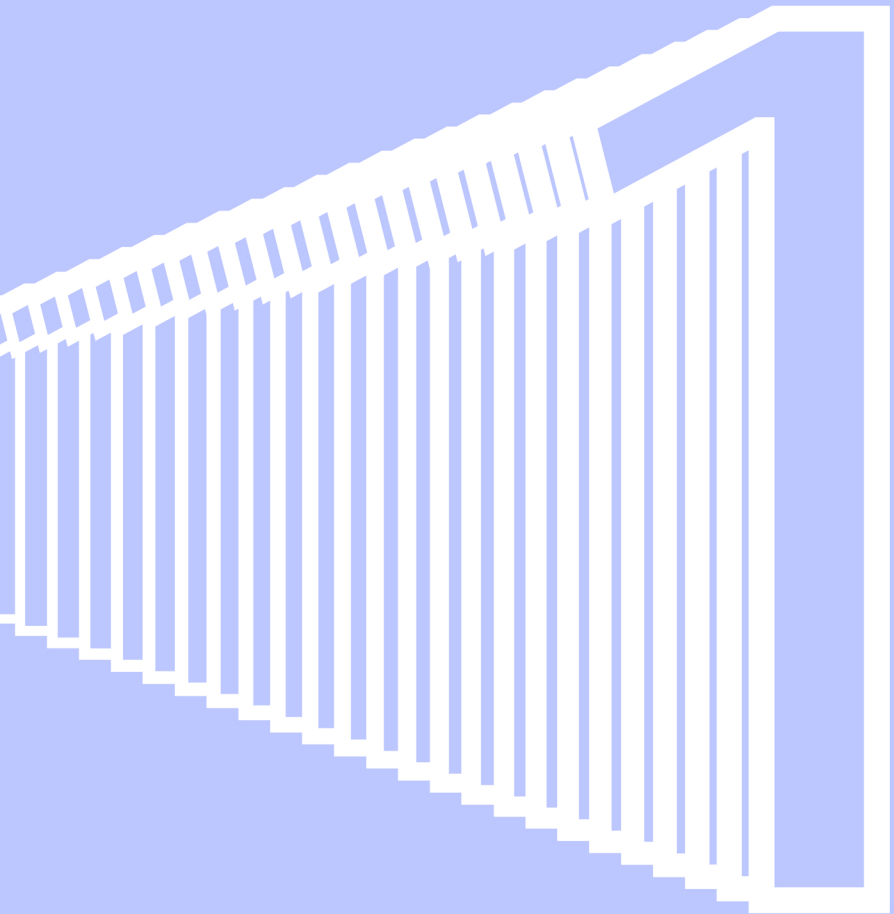
ТУЛКИТ DATA SCIENCE

«БОЛЬШИЕ ДАННЫЕ И МАШИННОЕ ОБУЧЕНИЕ»



Data Science

2022 | ФЦДО

Методическая разработка
«Тулкит Data Science»

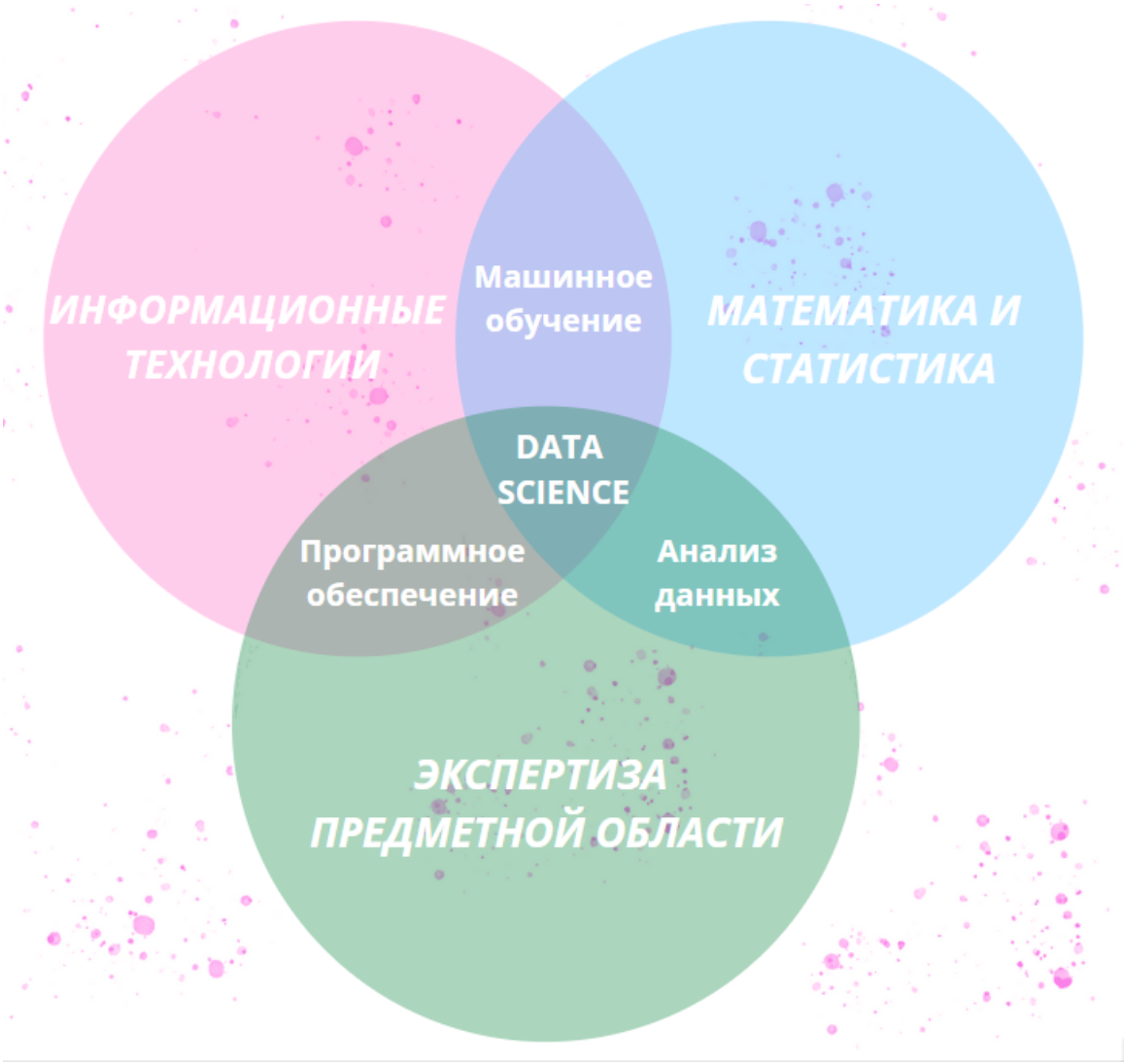
Направление: «Большие данные и машинное обучение»

Автор: старший методист Черная Анна Александровна

Организация: ФГБОУ ДО «Федеральный центр дополнительного образования и организации отдыха и оздоровления детей»

Что такое Data Science?

Data Science (наука о данных) — это набор конкретных дисциплин из разных направлений, отвечающих за анализ данных и поиск оптимальных решений на их основе.



Процесс Data Science:

- назначение цели исследования;
- сбор данных;
- подготовка данных;
- исследование данных;
- моделирование данных;
- отображение и автоматизация.



Анализ данных — процесс исследования, фильтрации, преобразования и моделирования данных с целью извлечения полезной информации и принятия решений.

Машинное обучение — процесс автоматизированной обработки структурированных данных для нахождения полезных закономерностей. Основная идея заключается в том, чтобы научить машину «обучаться» и принимать решения без участия человека. То есть не быть явно запрограммированной, а оптимизировать (минимизировать) некоторую целевую переменную на основе выявленных закономерностей.

Данный процесс содержит несколько стадий:

- подготовка данных;
- построение модели;
- оценка модели;
- оптимизация;
- предсказание.



Услуги специалистов по машинному обучению актуальны в настоящее время. Уже сейчас есть автомобили с беспилотным управлением, и специалисты по машинному обучению все время их дообучают. В медицине применяется машинное обучение для обследования пациента, которое основывается на статистике со всего мира.

Еще одной востребованной профессией является разработчик моделей Big Data. Специалист фиксирует события, выявляет закономерности, чтобы использовать их на практике. Все это помогает понять, какой товар или услуга заинтересуют клиента, кому одобрить кредит, что нужно предпринять, чтобы клиент не ушел. Эта профессия тесно связана с машинным обучением, так как для него необходим сбор и обработка данных.

В перечне «сквозных» цифровых технологий (утверждены распоряжением Правительства Российской Федерации от 28 июля 2018 года № 1632-р) и субтехнологий указаны следующие разделы: машинное

обучение; методы и способы визуализации, представления и использования больших данных человеком; интеллектуальный анализ данных.

Подводя итоги, машинное обучение (МО) замечательно подходит для:

- задач, существующие решения которых требуют большого объема ручной настройки или длинных списков правил — один алгоритм МО часто способен упростить код и выполняться лучше;
- сложных задач, для которых традиционный подход вообще не предлагает хороших решений — лучшие приемы МО могут найти решение;
- изменяющихся сред — система МО способна адаптироваться к новым данным;
- получения сведений о сложных задачах и крупных объемах данных.

Как учим?

Для того, чтобы начать обучение по данному направлению, необходимо иметь навыки программирования, решения алгоритмических задач; математические и статистические знания; навык межотраслевой коммуникации.

Углубленный модуль направлен на изучение анализа данных и машинного обучения. Возможны разные направления: сбор и обработка данных, работа с Big Data, визуализация данных, программирование на разных языках (Python, R), оптимизация моделей машинного обучения.

Командные проекты — реальные заказы от партнеров. А также межквантумные проекты.

Дети могут участвовать в мероприятиях регионального и федерального уровней (хакатоны, Олимпиада НТИ, партнерские конкурсы, WorldSkills и др.)

Возраст — от 13 лет.

Чему учим?

МОДУЛЬ на 72 часа

Ключевые темы

- Библиотеки для работы с данными на языке программирования Python.
- Сбор, подготовка и очистка данных.
- Визуализация данных: мозаичные диаграммы, диаграммы размаха, графики плотности, диаграммы рассеяния, «ящик с усами».
- Применение моделей. Модель с учителем и без учителя. Модели классификации и регрессии.
- Оценка производительности модели.
- Оптимизация производительности модели.

Возможные проекты

Создание робота, который учится передвигаться на определенное расстояние с помощью метода «обучение с подкреплением».

Разработка алгоритма машинного обучения, определяющего диагноз по рентгеновским снимкам (искривление позвоночника, переломы).

Разработка алгоритма машинного обучения классификации клиентских заявок.

Разработка алгоритма машинного обучения, определяющего ожидаемый доход магазина в следующем месяце.

Итоги углубленного модуля

Количественные:

- Не менее четырех нормализованных наборов данных.
- Не менее двух решенных задач машинного обучения.

Качественные:

- Умение работать с данными (загрузка из файла и выгрузка в файл; добавление, удаление и изменение данных) с помощью языка программирования Python.
- Знание моделей машинного обучения.
- Умение визуализировать данные и находить взаимосвязи в них.
- Навыки нормализации данных.
- Навыки применения моделей машинного обучения.
- Навык оценки и оптимизации модели.

Ограничения

1 уровень

1. Как работает обучение с учителем и без учителя?
2. Назовите типы задач машинного обучения.
3. В чем разница между искусственным интеллектом и машинным обучением?
4. Какие этапы рабочего процесса в машинном обучении?
5. Что такое переобучение и недообучение?
6. В чем отличие машинного обучения от классического программирования?

2 уровень

1. Подберите датасет с сайта https://www.kaggle.com/ и визуализируйте данные, используя известные диаграммы.
2. Возьмите известную задачу машинного обучения, примените к ней разные модели машинного обучения. Какая работает лучше?
3. Возьмите известную задачу машинного обучения с готовым решением, запустите ее на своем компьютере, определите работоспособность модели, при необходимости оптимизируйте.

3 уровень

1. Придумайте задачу машинного обучения и соберите для нее обучающую выборку.
2. Нормализуйте набор данных (заполните пропуски, уберите лишние данные, приведите к нужному типу данных).
3. Решите задачу кредитного скоринга.

Образовательная программа углубленного модуля 72 часа

Пояснительная записка

В настоящее время непрерывно растет поток различной информации и объем генерируемых данных. Если обработать эту информацию, то с ее помощью можно делать прогнозы и выводы в областях бизнеса, робототехники, медицины и промышленности: металлургия, нефтехимия, пищевая отрасль и др. Но чем больше данных, тем сложнее с ними работать и их анализировать. Необходимо иметь инструмент работы с большими данными. Таким инструментом может выступать машинное обучение.

Машинное обучение — это методики анализа данных, которые позволяют аналитической системе обучаться в ходе решения множества сходных задач.

Для задач машинного обучения необходимо иметь навыки программирования, решения алгоритмических задач; математические и статистические знания; навык межотраслевой коммуникации.

В ходе занятий по программе обучающиеся познакомятся с понятием анализа данных и машинного обучения, поймут их особенности, возможности и взаимосвязь, получат навыки в данной области для применения на более сложных задачах.

Цель модуля

Формирование компетенций по работе с данными и применение их на реальных задачах анализа данных и машинного обучения.

Задачи модуля

- Погрузить участников в проектную деятельность для формирования навыков ведения проекта.
- На протяжении всех занятий формировать 4К-компетенци (критическое мышление, креативное мышление, коммуникация, кооперация).
- Познакомить с понятиями анализа данных и машинного обучения.
- Научить собирать и нормализовать данные.
- Сформировать навык визуализации данных (подбор необходимых инструментов для визуализации, определение взаимосвязи данных по графику).
- Дать основные навыки применения моделей машинного обучения.

- Сформировать способность оценки и оптимизации модели.

Методы

- Кейс-метод.
- Проектная деятельность.
- Датаскаутинг.
- Демонстрация наглядного материала.
- Мозговой штурм.
- Метод проектов.

Формы работы

Программой предусмотрена фронтальная, групповая и индивидуальная формы обучения, в том числе:

- интерактивные лекции;
- самостоятельная работа обучающихся;
- защита проектных работ;
- игровые технологии.

Приветствуются встречи с приглашенными спикерами, совместные конференции, видеоконференции или вебинары с другими квантумами и экспертами, индивидуальные и групповые консультации.

Требования к результатам освоения программ модуля

Профессиональные и знаниевые компетенции (Hard Skills):

- знание и понимание работы алгоритмов машинного обучения;
- умение собирать и анализировать данные;
- умение визуализировать данные и делать на основе этого выводы;
- навык использования библиотек NumPy, Pandas, Matplotlib, Scikit-learn;
- написание алгоритма машинного обучения;
- навык нормализации данных;
- умение оценивать и оптимизировать модель.

Личностные и межличностные компетенции (Soft Skills):

- работа в команде: эффективное распределение задач и др.;
- развитие познавательных интересов обучающихся;
- умение ориентироваться в информационном пространстве, продуктивно использовать техническую литературу для поиска сложных решений;
- навыки ведения проекта, проявление компетенции в вопросах, связанных с темой проекта, выбор наиболее эффективных решений задач в зависимости от конкретных условий;
- развитие критического мышления;
- проявление технического мышления, познавательной деятельности, творческой инициативы, самостоятельности;

- способность творчески решать технические задачи;
- готовность и способность применения теоретических знаний;
- способность правильно организовывать рабочее место и время для достижения поставленных целей.

Артефакты:

- Не менее четырех нормализованных наборов данных.
- Не менее двух применений моделей машинного обучения.

Рекомендации наставникам

В рамках углубленного модуля педагогу следует сконцентрироваться на формировании основных профессиональных компетенций (Hard Skills), а также научить генерировать идеи по применению машинного обучения.

На помощь приходят статьи, новости и сайты про машинное обучение и его применение на реальных задачах. Важно спрашивать у обучающихся, какие примеры они могут привести. Дать понимание, что машинное обучение применимо к таким областям, как медицина, робототехника и т. д.

Стоит также учитывать, что предлагаемые материалы носят характер рекомендаций, а не жестких планов и правил. Тьютор должен проявлять гибкость и пересматривать план работы в соответствии с уровнем и интересами группы.

Учебно-тематическое планирование

Режим занятий: 2 раза по 2 часа в неделю.
Продолжительность одного занятия: 2 академических часа.
Структура двухчасового занятия:

- 45 минут — рабочая часть;
- 15 минут — перерыв (отдых);
- 45 минут — рабочая часть.

Программа рассчитана на 18 недель обучения, общее количество академических часов — 72. Основной формой являются групповые занятия. В основе образовательного процесса лежит проектный подход.

Основная форма работы подачи теории — занятия в группах по 14 человек. Практические задания планируется выполнять как индивидуально и в парах, так и в малых группах. Занятия проводятся в виде бесед, семинаров, лекций; для наглядности подаваемого материала используется различный мультимедийный материал — презентации, видеоролики и пр.

Кейс 1. «Распределяющая шляпа»

Описание

Коля выбирает, в какой квантум ему и его друзьям Лене и Роме пойти. Ребята из IT-квантума предложили использовать методы машинного обучения для определения, какой квантум подойдет Коле, Лене и Роме, но им нужны данные для обучения (возраст, любимый предмет в школе, хобби и другие увлечения). Можно ли применить к этой задаче методы машинного обучения? Что будет целевой переменной? Какие входные данные нужны? Как собрать обучающую выборку? Как подобрать данные по форме представления?

Категория кейса: углубленный.

Место в структуре модуля: в рамках данного кейса обучающиеся знакомятся основными понятиями машинного обучения и реализуют задачу сбора и анализа данных.

Количество учебных часов: 8.

Занятие 1

Цель: познакомиться с понятием «машинное обучение».

Что делаем: знакомимся с понятиями «машинное обучение», «целевая переменная», «обучающая выборка»; знакомимся с типами задач машинного обучения и определяем, к какому типу относится данная задача.

Компетенции: умение находить, анализировать и использовать информацию, умение ставить вопросы (инициативное сотрудничество в поиске и сборе информации), знание и понимание основных понятий: «машинное обучение», «целевая переменная», «обучающая выборка».

Количество часов: 2.

Занятие 2

Цель: создать форму для сбора обучающей выборки.

Что делаем: подбираем входные признаки и их тип; выбираем способ сбора данных; создаем анкету в Яндекс.Формы или Google Forms; рассылаем анкету на заполнение; знакомимся с наборами данных и сайтом kaggle.com.

Компетенции: умение находить, анализировать и использовать информацию; навыки самостоятельного решения проблем творческого и поискового характера; умение создавать анкеты в Яндекс.Формы или Google Forms.

Количество часов: 2.

Занятие 3

Цель: провести анализ собранных данных.

Что делаем: определяем связь между целевой переменной и входными признаками; ищем пропуски в данных, изучаем их возникновение; проводим нормализацию данных.

Компетенции: умение находить, анализировать и использовать информацию; навыки самостоятельного решения проблем творческого и поискового характера; навыки формулирования проблемы, выдвижения гипотезы; умение ставить вопросы.

Количество часов: 2.

Занятие 4

Цель: презентовать свои наработки**.**

Что делаем: демонстрируем свои разработки; обсуждаем, задаем вопросы; вносим доработки по необходимости.

Компетенции: навык качественной презентации, умение отвечать на вопросы.

Количество часов: 2.

Метод работы с кейсом: кейс-метод.

Минимально необходимый уровень входных компетенций: для прохождения кейса не требуется специальных знаний.

Предполагаемые результаты обучающихся, формируемые навыки

Универсальные навыки (Soft Skills):

- умение находить, анализировать и использовать информацию;
- умение ставить вопросы (инициативное сотрудничество в поиске и сборе информации);
- навыки самостоятельного решения проблем творческого и поискового характера;
- навыки формулирования проблемы, выдвижения гипотезы;
- умение ставить вопросы;
- навык качественной презентации;
- умение отвечать на вопросы.

Предметные навыки (Hard Skills):

- знание и понимание основных понятий: «машинное обучение», «целевая переменная», «обучающая выборка»;
- умение создавать анкеты в Яндекс.Формы или Google Forms.

Процедуры и формы выявления образовательного результата

Универсальные навыки оцениваются с помощью групповой рефлексии. Предметные навыки преподаватель оценивает в процессе работы над проектом и по результатам итоговой презентации.

Необходимые материалы и оборудование

- Компьютер/ноутбук с доступом к интернету.
- Оборудование для презентации: проектор, экран.

Кейс 2. «Визуализируй»

Описание

Компания, которая делает видеоигры, хочет создать алгоритм для определения успеха продаж. Данные для алгоритма уже собраны (https://www.kaggle.com/rush4ratio/video-game-sales-with-ratings), но перед построением модели машинного обучения необходимо проверить взаимосвязь обучающих признаков и целевой переменной. Этот этап называется визуализацией данных. Какие способы визуализации данных вы знаете? Может ли она повлиять на выбор входных признаков? Какая библиотека языка Python позволяет визуализировать данные (строить графики)? Какие составляющие должны быть у графика? Как определить, какой вид диаграммы необходимо применить?

Категория кейса: углубленный.

Место в структуре модуля: в рамках данного кейса изучается один из этапов рабочего процесса машинного обучения — визуализация данных. Она дает наглядное представление о взаимосвязи данных, помогает лучше понять модель и генерируемые прогнозы.

Количество учебных часов: 14.

Занятие 1

Цель: познакомиться с понятием визуализации данных.

Что делаем: знакомимся с понятием визуализации данных и видами графиков: диаграммы размаха, мозаичные, рассеяния, графики плотности; устанавливаем необходимое ПО.

Компетенции: умение искать информацию в свободных источниках; навыки самостоятельного решения проблем творческого и поискового характера; знание и понимание основных понятий визуализации данных.

Количество часов: 4.

Занятие 2

Цель: научиться работать с библиотеками Pandas и NumPy.

Что делаем: знакомимся с библиотеками Pandas и NumPy, подгружаем данные из файла, «чистим» данные, определяем отсутствующие данные и их важность.

Компетенции: умение искать информацию в свободных источниках; навыки самостоятельного решения проблем творческого и поискового характера; умение работать с библиотеками Pandas и NumPy; умение анализировать информацию.

Количество часов: 4.

Занятие 3

Цель: научиться строить диаграммы.

Что делаем: формируем навык работы с библиотекой Matplotlib; определяем основные составляющие графиков; строим диаграммы размаха, мозаичные, рассеяния, графики плотности.

Компетенции: умение искать информацию в свободных источниках; навыки самостоятельного решения проблем творческого и поискового характера; умение работать с библиотекой Matplotlib.

Количество часов: 2.

Занятие 4

Цель: научиться анализировать диаграммы.

Что делаем: строим диаграммы для определения взаимосвязей входных данных и целевой переменной, проводим анализ получившихся результатов, определяем шум по графику; делаем вывод о взаимосвязи входных данных и целевой переменной.

Компетенции: умение искать информацию в свободных источниках; навыки самостоятельного решения проблем творческого и поискового характера; умение работать с библиотеками Pandas, Matplotlib и NumPy; умение анализировать информацию.

Количество часов: 2.

Занятие 5

Цель: презентовать свои наработки**.**

Что делаем: демонстрируем свои разработки, обсуждаем, задаем вопросы.

Компетенции: навык качественной презентации, умение отвечать на вопросы.

Количество часов: 2.

Метод работы с кейсом: кейс-метод.

Минимально необходимый уровень входных компетенций: основные понятия языка Python, датаскаутинг.

Предполагаемые результаты обучающихся, формируемые навыки

Универсальные навыки (Soft Skills):

- умение искать информацию в свободных источниках;
- навыки самостоятельного решения проблем творческого и поискового характера;
- умение анализировать информацию;
- навык качественной презентации;
- умение отвечать на вопросы.

Предметные навыки (Hard Skills):

- умение работать с библиотекой Matplotlib;
- умение работать с библиотеками Pandas и NumPy;

- знание и понимание основных методов и инструментов визуализации данных.

Процедуры и формы выявления образовательного результата

Универсальные навыки оцениваются с помощью групповой рефлексии. Предметные навыки преподаватель оценивает в процессе работы над кейсом и по результатам итоговой презентации.

Необходимые материалы и оборудование

- Компьютер/ноутбук с доступом к интернету и установленным ПО для работы с языком Python.
- Оборудование для презентации: проектор, экран.

Кейс 3. «Классификация»

Описание

Стекольный завод производит семь типов стекла. Прежде чем поставить стекло на мебельную фабрику, завод сборки автомобилей и другие предприятия, стекло проверяют и сортируют. Необходимо построить модель машинного обучения для определения типа стекла (бытовое, фары и т. д.). Дан набор данных для обучения (https://www.kaggle.com/uciml/glass), который содержит 214 экземпляров и искаженные данные. Что будет являться целевой переменной? Все ли данные подходят по форме представления? Почему? Необходимо ли нормализовать данные? Каким образом? Какой метод лучше подобрать для небольшого количества экземпляров?

Категория кейса: углубленный.

Место в структуре модуля: в рамках данного кейса обучающиеся знакомятся с задачей классификации и работы с небольшим набором данных. Проводят рабочий процесс от анализа и обработки данных до построения модели.

Количество учебных часов: 10.

Занятие 1

Цель: определить обучающую выборку.

Что делаем: определяем целевую переменную, знакомимся с набором данных, выявляем подходящую форму входных признаков, визуализируем данные.

Компетенции: умение искать информацию в свободных источниках; навыки самостоятельного решения проблем творческого и поискового характера; умение работать с библиотеками Matplotlib, Pandas и NumPy; умение делать вывод на основе графиков.

Количество часов: 2.

Занятие 2

Цель: написать программы для нормализации признаков.

Что делаем: изучаем способы нормализации данных, пишем код, реализующий нормализацию данных, нормализуем набор данных из кейса.

Компетенции: умение искать информацию в свободных источниках; навыки самостоятельного решения проблем творческого и поискового характера; знание и понимание основных понятий нормализации данных.

Количество часов: 2.

Занятие 3

Цель: изучить модели, применяемые к задачам классификации.

Что делаем: изучаем алгоритм дерева решений, метод k-ближайших соседей, логистическую регрессию и применяем их для классификации типов стекла; знакомимся с библиотекой Scikit-learn.

Компетенции: умение искать информацию в свободных источниках; навыки самостоятельного решения проблем творческого и поискового характера; знание и понимание основных алгоритмов классификации; умение работать с библиотекой Scikit-learn.

Количество часов: 2.

Занятие 4

Цель: обучить модель на данных.

Что делаем: делим выборку на обучающую и тестовую; используя библиотеку Scikit-learn, обучаем модель на обучающей выборке и тестируем на оставшихся данных; знакомимся с понятиями переобучения и недообучения; делаем выводы.

Компетенции: умение искать информацию в свободных источниках; навыки самостоятельного решения проблем творческого и поискового характера; знание и понимание переобучения и недообучения; умение объективно оценивать результаты своей работы; умение работать с библиотекой Scikit-learn; навык обучения и тестирования модели на данных.

Количество часов: 2.

Занятие 5

Цель: презентовать свои наработки**.**

Что делаем: демонстрируем свои разработки, обсуждаем, задаем вопросы. Вносим доработки по необходимости.

Компетенции: навык качественной презентации, умение отвечать на вопросы, умение объективно оценивать результаты своей работы.

Количество часов: 2.

Метод работы с кейсом: кейс-метод.

Минимально необходимый уровень входных компетенций: для прохождения кейса требуются знания, полученные в предыдущем кейсе «Визуализация» (№ 2); основные понятия языка Python, датаскаутинг.

Предполагаемые результаты обучающихся, формируемые навыки

Универсальные навыки (Soft Skills):

- умение искать информацию в свободных источниках;
- навыки самостоятельного решения проблем творческого и поискового характера;
- навык качественной презентации;
- умение отвечать на вопросы;
- умение объективно оценивать результаты своей работы.

Предметные навыки (Hard Skills):

- умение работать с библиотеками Matplotlib, Pandas и NumPy;
- умение делать вывод на основе графиков;
- знание и понимание основных понятий нормализации данных;

- знание и понимание основных алгоритмов классификации;
- умение работать с библиотекой Scikit-learn;
- знание и понимание переобучения и недообучения;
- навык обучения и тестирования модели на данных.

Процедуры и формы выявления образовательного результата

Универсальные навыки оцениваются с помощью групповой рефлексии. Предметные навыки преподаватель оценивает в процессе работы над кейсом и по результатам итоговой презентации.

Необходимые материалы и оборудование

- Компьютер/ноутбук с доступом к интернету и установленным ПО для работы с языком Python.
- Оборудование для презентации: проектор, экран.

Кейс 4. «Чаевые»

Описание

Водитель такси Николай Иванович хочет получать больше чаевых от клиентов. Он брал заказы с разным местом высадки и посадки, длительными поездками и разными тарифами. Николай Иванович мечтает о системе, которая будет давать подсказки, какой заказ выгоднее. Можно ли помочь ему? Что будет являться целевой переменной? Необходимо ли нормализовать данные? Какой метод применить к данной задаче?

Набор данных для обучения: https://www.kaggle.com/chicago/chicago-taxi-trips-bq.

Категория кейса: углубленный.

Место в структуре модуля: в рамках данного кейса обучающиеся знакомятся с задачей регрессии для предсказания размера чаевых от каждой поездки. Проводят рабочий процесс от анализа и обработки данных до оценки модели. Изучение оценки задач регрессии.

Количество учебных часов: 14.

Занятие 1

Цель: определить обучающую выборку.

Что делаем: определяем целевую переменную, отбрасываем «шум» из входных признаков, выявляем подходящую форму входных признаков, визуализируем данные, делаем выводы.

Компетенции: умение искать информацию в свободных источниках; навыки самостоятельного решения проблем творческого и поискового характера.

Количество часов: 2.

Занятие 2

Цель: подготовить данные.

Что делаем: формулируем задачу МО, подготавливаем данные, устраняем недостоверные данные (денежные чаевые обычно не регистрируются).

Компетенции: умение искать информацию в свободных источниках; навыки самостоятельного решения проблем творческого и поискового характера.

Количество часов: 2.

Занятие 3

Цель: изучить модели, применяемые к задачам регрессии.

Что делаем: изучаем алгоритмы регрессионной модели, линейной регрессии, метод наименьших квадратов, нелинейную регрессию.

Компетенции: умение искать информацию в свободных источниках; навыки самостоятельного решения проблем творческого и поискового характера; знание и понимание основных алгоритмов регрессии; умение

работать с библиотекой Scikit-learn; знание и понимание алгоритмов задач регрессии и области их применения.

Количество часов: 2.

Занятие 4

Цель: обучить модель на данных.

Что делаем: делим выборку на обучающую и тестовую; используя библиотеку Scikit-learn, обучаем модель на обучающей выборке и тестируем на оставшихся данных; делаем выводы.

Компетенции: умение искать информацию в свободных источниках; навыки самостоятельного решения проблем творческого и поискового характера; умение объективно оценивать результаты своей работы; умение работать с библиотекой Scikit-learn; навык обучения и тестирования модели на данных.

Количество часов: 2.

Занятие 5

Цель: сделать оценку модели.

Что делаем: изучаем методы RMSE (среднеквадратическая ошибка), исследование остатков, сеточный поиск, делаем выводы.

Компетенции: умение искать информацию в свободных источниках; навыки самостоятельного решения проблем творческого и поискового характера; умение объективно оценивать результаты своей работы; умение работать с библиотекой Scikit-learn; знание и понимание оценки модели.

Количество часов: 2.

Занятие 6

Цель: оптимизация модели.

Что делаем: используя выводы прошлого занятия, оптимизируем модель (удаляем/добавляем данные, изменяем типы данных) и переобучаем ее.

Компетенции: умение искать информацию в свободных источниках; навыки самостоятельного решения проблем творческого и поискового характера; умение объективно оценивать результаты своей работы; умение работать с библиотекой Scikit-learn; навык исправления модели.

Количество часов: 2.

Занятие 7

Цель: презентовать свои наработки.

Что делаем: демонстрируем свои разработки, обсуждаем, задаем вопросы. Вносим доработки по необходимости.

Компетенции: навык качественной презентации, умение отвечать на вопросы

Количество часов: 2.

Метод работы с кейсом: кейс-метод.

Минимально необходимый уровень входных компетенций: основные понятия языка Python, датаскаутинг.

Предполагаемые результаты обучающихся, формируемые навыки

Универсальные навыки (Soft Skills):

- умение объективно оценивать результаты своей работы;
- умение искать информацию в свободных источниках;
- навыки самостоятельного решения проблем творческого и поискового характера.

Предметные навыки (Hard Skills):

- знание и понимание основных алгоритмов регрессии;
- умение работать с библиотекой Scikit-learn;
- знание и понимание алгоритмов задач регрессии;
- навык обучения и тестирования модели на данных.

Процедуры и формы выявления образовательного результата

Универсальные навыки оцениваются с помощью групповой рефлексии. Предметные навыки преподаватель оценивает в процессе работы над кейсом и по результатам итоговой презентации.

Необходимые материалы и оборудование

- Компьютер/ноутбук с доступом к интернету и установленным ПО для работы с языком Python.
- Оборудование для презентации: проектор, экран.

Кейс 5. «Диагноз»

Описание

Уровень заболеваемости диабетом растет с каждым годом. Это связано с изменением образа жизни. Люди могут годами жить, не зная о своем диагнозе, в то время как организму наносится вред. Поэтому важно вовремя диагностировать наличие диабета у пациента.

Набор данных содержит диагностические измерения пациентов. Какую модель машинного обучения нужно применить, чтобы предсказать, есть у пациентов диабет или нет?

Набор данных для обучения: https://www.kaggle.com/uciml/pima-indians-diabetes-database.

Категория кейса: углубленный.

Место в структуре модуля: в рамках данного кейса обучающиеся знакомятся с задачей регрессии. Проводят рабочий процесс от анализа и обработки данных до оценки модели. Изучение оценки задач классификации.

Количество учебных часов: 14.

Занятие 1

Цель: определить обучающую выборку.

Что делаем: определяем целевую переменную, отбрасываем «шум» из входных признаков, выявляем подходящую форму входных признаков, визуализируем данные, делаем выводы.

Компетенции: умение искать информацию в свободных источниках; навыки самостоятельного решения проблем творческого и поискового характера.

Количество часов: 2.

Занятие 2

Цель: подготовить данные.

Что делаем: формулируем задачу МО, подготавливаем данные, устраняем пустые значения.

Компетенции: умение искать информацию в свободных источниках; навыки самостоятельного решения проблем творческого и поискового характера.

Количество часов: 2.

Занятие 3

Цель: обучить модель на данных.

Что делаем: делим выборку на обучающую и тестовую; используя библиотеку Scikit-learn, обучаем модель на обучающей выборке и тестируем на оставшихся данных; делаем выводы.

Компетенции: умение искать информацию в свободных источниках; навыки самостоятельного решения проблем творческого и поискового

характера; умение объективно оценивать результаты своей работы; умение работать с библиотекой Scikit-learn; навык обучения и тестирования модели на данных.

Количество часов: 2.

Занятие 4

Цель: сделать оценку модели.

Что делаем: изучаем методы RMSE (среднеквадратическая ошибка), исследование остатков, сеточный поиск, делаем выводы.

Компетенции: умение искать информацию в свободных источниках; навыки самостоятельного решения проблем творческого и поискового характера; умение объективно оценивать результаты своей работы; умение работать с библиотекой Scikit-learn; знание и понимание оценки модели.

Количество часов: 2.

Занятие 6

Цель: оптимизация модели.

Что делаем: используя выводы прошлого занятия, оптимизируем модель (удаляем/добавляем данные, изменяем типы данных) и переобучаем ее.

Компетенции: умение искать информацию в свободных источниках; навыки самостоятельного решения проблем творческого и поискового характера; умение объективно оценивать результаты своей работы; умение работать с библиотекой Scikit-learn; навык исправления модели.

Количество часов: 2.

Занятие 7

Цель: презентовать свои наработки.

Что делаем: демонстрируем свои разработки, обсуждаем, задаем вопросы. Вносим доработки по необходимости.

Компетенции: навык качественной презентации, умение отвечать на вопросы.

Количество часов: 2.

Метод работы с кейсом: кейс-метод.

Минимально необходимый уровень входных компетенций: основные понятия языка Python, датаскаутинг.

Предполагаемые результаты обучающихся, формируемые навыки

Универсальные навыки (Soft Skills):

- умение объективно оценивать результаты своей работы;
- умение искать информацию в свободных источниках;
- навыки самостоятельного решения проблем творческого и поискового характера.

Предметные навыки (Hard Skills):

- знание и понимание основных алгоритмов регрессии;
- умение работать с библиотекой Scikit-learn;
- знание и понимание алгоритмов задач регрессии;
- навык обучения и тестирования модели на данных.

Процедуры и формы выявления образовательного результата

Универсальные навыки оцениваются с помощью групповой рефлексии. Предметные навыки преподаватель оценивает в процессе работы над кейсом и по результатам итоговой презентации.

Необходимые материалы и оборудование

- Компьютер/ноутбук с доступом к интернету и установленным ПО для работы с языком Python.
- Оборудование для презентации: проектор, экран.

Кейс 6. «Искусственный доктор»

Описание

Искусственный интеллект может изменить диагностику и лечение заболеваний, выполняя классификацию, трудную для специалистов. Данные содержат записи жалоб пациентов. Как преобразовать записи в признаки? Как «чистить» данные? Какую модель машинного обучения нужно применить для определения заболевания?

Набор данных для обучения (https://www.kaggle.com/paultimothymooney/medical-speech-transcription-and-intent/data#).

Категория кейса: углубленный.

Место в структуре модуля: в рамках данного кейса обучающиеся учатся работать с текстом и записями. Проводят рабочий процесс от анализа и обработки данных до оценки модели. Изучение оценки задач классификации.

Количество учебных часов: 10.

Занятие 1

Цель: определить обучающую выборку.

Что делаем: определяем целевую переменную, отбрасываем «шум» из входных признаков, выявляем подходящую форму входных признаков, визуализируем данные, делаем выводы.

Компетенции: умение искать информацию в свободных источниках; навыки самостоятельного решения проблем творческого и поискового характера.

Количество часов: 2.

Занятие 2

Цель: подготовить данные.

Что делаем: изучаем техники обработки естественного языка: модель «мешок слов» (токенизация и преобразования, векторизация), тематическое моделирование (частота слова, латентно-семантический анализ, вероятностные методы).

Компетенции: умение искать информацию в свободных источниках; навыки самостоятельного решения проблем творческого и поискового характера.

Количество часов: 2.

Занятие 3

Цель: обучить модель на данных.

Что делаем: делим выборку на обучающую и тестовую; используя библиотеку Scikit-learn, обучаем модель на обучающей выборке; знакомимся с моделью на базе наивного байесовского классификатора.

Компетенции: умение искать информацию в свободных источниках; навыки самостоятельного решения проблем творческого и поискового характера; умение объективно оценивать результаты своей работы; умение работать с библиотекой Scikit-learn; навык обучения модели на данных.

Количество часов: 2.

Занятие 4

Цель: оптимизация модели.

Что делаем: используя выводы прошлого занятия, оптимизируем модель (удаляем/добавляем данные, изменяем типы данных) и переобучаем ее.

Компетенции: умение искать информацию в свободных источниках; навыки самостоятельного решения проблем творческого и поискового характера; умение объективно оценивать результаты своей работы; умение работать с библиотекой Scikit-learn; навык исправления модели.

Количество часов: 2.

Занятие 5

Цель: презентовать свои наработки.

Что делаем: демонстрируем свои разработки, обсуждаем, задаем вопросы. Вносим доработки по необходимости.

Компетенции: навык качественной презентации, умение отвечать на вопросы.

Количество часов: 2.

Метод работы с кейсом: кейс-метод.

Минимально необходимый уровень входных компетенций: основные понятия языка Python, датаскаутинг.

Предполагаемые результаты обучающихся, формируемые навыки

Универсальные навыки (Soft Skills):

- умение объективно оценивать результаты своей работы;
- умение искать информацию в свободных источниках;
- навыки самостоятельного решения проблем творческого и поискового характера.

Предметные навыки (Hard Skills):

- знание и понимание основных алгоритмов регрессии;
- умение работать с библиотекой Scikit-learn;
- знание и понимание алгоритмов задач регрессии;
- навык обучения и тестирования модели на данных.

Процедуры и формы выявления образовательного результата

Универсальные навыки оцениваются с помощью групповой рефлексии. Предметные навыки преподаватель оценивает в процессе работы над кейсом и по результатам итоговой презентации.

Необходимые материалы и оборудование

- Компьютер/ноутбук с доступом к интернету и установленным ПО для работы с языком Python.
- Оборудование для презентации: проектор, экран.

Мастер-класс №1. «Ромашка»

Тема: «Сбор данных для задачи машинного обучения».
Длительность: от 40 минут до 2 часов.
Целевая аудитория: школьники 14–18 лет, их родители.
Цели и задачи:

- познакомить обучающихся с понятиями «машинное обучение», «анализ данных»;
- определить критерии отбора данных;
- научить создавать Яндекс.Формы и подбирать вопросы для анкетирования;
- дать основные навыки анализа данных и их нормализации.

Требования к входным навыкам участников: нет.
Краткое описание:

Супермаркет «Ромашка» предлагает своим покупателям несколько категорий скидок: товары для дачи, товары для дома, товары для детей, товары для учебы, инструменты, бытовая техника.

Необходимо собрать обучающую выборку для задачи выбора категории скидок. Как собрать обучающую выборку? Какие признаки она должна содержать? Можно ли использовать данные в исходном виде?

Варианты входных признаков: возраст, есть ли дети, семейное положение, есть ли машина и т. д.

План проведения/алгоритм действий:

1. Погружение в тему, сбор и анализ данных в машинном обучении. Приводим и обсуждаем примеры задач машинного обучения, придумываем свои идеи.
2. Обсуждаем: как собрать данные (Яндекс.Формы, Google Forms); что такое «сырые» данные; что такое нормализация данных и зачем ее делать.
3. Знакомим с задачей «Ромашка». Придумываем признаки.
4. Создаем анкету, основываясь на признаках. Формулируем вопросы, чтобы ответ был однозначным. Важно! Анкета должна содержать целевую переменную (какую категорию скидок выбирает пользователь).
5. Заполняем анкету данными.
6. Анализируем собранные данные. Скорее всего, данные будут «сырые». Их необходимо будет обрабатывать. Например, все слова (строки) заменять на числовые и категориальные переменные.
7. Рефлексия.

Необходимое оборудование и расходные материалы (для проведения МК):

- компьютер наставника с подключением к интернету, компьютеры для участников с подключением к интернету;
- проектор;
- распечатанные задания.

Предполагаемые результаты обучающихся

Умения и навыки:

- знание и понимание основных понятий: «сбор и анализ данных», «машинное обучение», «обучающая выборка»;
- навыки создания форм (Яндекс.Формы, Google Forms);
- умение находить, анализировать и использовать информацию;
- умение ставить вопросы;
- навыки самостоятельного решения проблем творческого и поискового характера.

Артефакт: набор данных (датасет).

Мастер-класс №2. «Работа с данными»

Тема: «Работа с данными с помощью языка Python».
Длительность: от 40 минут до 2 часов.
Целевая аудитория: школьники 14–18 лет, их родители.
Цели и задачи:

- познакомить обучающихся с понятием «нормализация данных»;
- определить, как необходимо преобразовывать данные;
- научить пользоваться Jupyter Notebook;
- дать основные навыки создания собственных тетрадей в Jupyter Notebook;
- дать основные навыки работы с данными с помощью библиотек Pandas и NumPy.

Требования к входным навыкам участников: нет.
Краткое описание:

Часто данные анкетирования содержат пропуски, ошибки и неправильно введенные данные. Прежде чем использовать, их необходимо привести к «нужному» виду: изменить тип данных, привести в бинарный или категориальный вид, заполнить пропуски или удалить строки.

Трудность заключается в том, что иногда этих данных очень много и обработать их вручную невозможно.

На мастер-классе участники узнают о нормализации данных и научатся пользоваться библиотеками Pandas и NumPy.

План проведения/алгоритм действий:

1. Погружение в тему нормализации данных. Обсуждаем, зачем нужно нормализовать данные.
2. Знакомим с набором данных (заранее подготовленным). Обсуждаем, что нужно изменить.
3. Открываем на компьютерах в браузере Jupyter Notebook (заранее подготовленный с частично написанным кодом). Пробуем запустить строчку кода, написать свою и т. д.
4. Запускаем и поясняем строчки кода: загрузка набора данных, получение информации о данных (пустые значения, количество строк и т. д.), вывод первых пяти строк набора данных.
5. Обсуждаем, на что влияют пропуски в данных (их стоит удалить или заполнить); какие данные избыточны, где нужно изменить типы данных.
6. Запускаем и поясняем строчки кода: удаление, изменение и добавление данных (строк, столбцов).
7. Участники самостоятельно пишут строчки кода: удаление, изменение и добавление данных (строк, столбцов).
8. Запускаем и поясняем строчки кода нормализации данных.
9. Участники самостоятельно пишут строчки кода нормализации данных.
10. Рефлексия.

Необходимое оборудование и расходные материалы (для проведения МК):

- компьютер наставника с подключением к интернету, компьютеры для участников с подключением к интернету;
- проектор;
- набор данных для обработки;
- страница Jupyter Notebook с кодом и комментариями, подразумевающая дополнения строчек кода от участников.

Предполагаемые результаты обучающихся

Умения и навыки:

- знание и понимание понятия «нормализация данных»;
- навыки создания страниц Jupyter Notebook;
- умение анализировать информацию;
- основные навыки работы с данными с помощью библиотек Pandas и NumPy.

Артефакт: нормализованный набор данных.

Список литературы

Для преподавателей

1. Kaggle: Your Machine Learning and Data Science Community [Электронный ресурс]. — Режим доступа: https://www.kaggle.com/.
2. Хенрик Бринк, Джозеф Ричардс, Марк Феверфол. Машинное обучение. — 2018 г.
3. Франсуа Шолле. Глубокое обучение на Python. — 2018 г.
4. Джоэл Грас. Data Science. Наука о данных с нуля. — 2017 г.
5. NumPy — Режим доступа: https://numpy.org/.
6. Pandas — Режим доступа: https://pandas.pydata.org/.
7. Matplotlib: Visualization with Python [Электронный ресурс]. — Режим доступа: https://matplotlib.org/.
8. Scikit-learn. Machine Learning in Python [Электронный ресурс]. — Режим доступа: https://scikit-learn.org/stable/.
9. TensorFlow [Электронный ресурс]. — Режим доступа: https://www.tensorflow.org/.
10. Jupyter Notebook для начинающих: учебник [Электронный ресурс]. — Режим доступа: https://webdevblog.ru/jupyter-notebook-dlya-nachinajushhih-uchebnik/.
11. Галерея интересных ноутбуков Jupyter [Электронный ресурс]. — Режим доступа: https://github.com/jupyter/jupyter/wiki/A-gallery-of-interesting-Jupyter-Notebooks.
12. Обзор библиотек для машинного обучения на Python [Электронный ресурс]. — Режим доступа: https://neerc.ifmo.ru/wiki/index.php?title=%D0%9E%D0%B1%D0%B7%D0%BE%D1%80_%D0%B1%D0%B8%D0%B1%D0%BB%D0%B8%D0%BE%D1%82%D0%B5%D0%BA_%D0%B4%D0%BB%D1%8F_%D0%BC%D0%B0%D1%88%D0%B8%D0%BD%D0%BD%D0%BE%D0%B3%D0%BE_%D0%BE%D0%B1%D1%83%D1%87%D0%B5%D0%BD%D0%B8%D1%8F_%D0%BD%D0%B0_Python.
13. Профессиональный информационно-аналитический ресурс, посвященный машинному обучению, распознаванию образов и интеллектуальному анализу данных [Электронный ресурс]. — Режим доступа: http://www.machinelearning.ru/wiki/index.php?title=%D0%97%D0%B0%D0%B3%D0%BB%D0%B0%D0%B2%D0%BD%D0%B0%D1%8F_%D1%81%D1%82%D1%80%D0%B0%D0%BD%D0%B8%D1%86%D0%B0.
14. Курс «Введение в анализ данных» (2019) [Электронный ресурс]. — Режим доступа: https://www.youtube.com/watch?v=5Yl6_2d0CN0&=&list=PLrCZzMib1e9p6lpNv-yt6uvHGyBxQncEh.
15. Введение в машинное обучение [Электронный ресурс]. — Режим доступа: https://www.coursera.org/learn/vvedenie-mashinnoe-obuchenie.
16. Курс «Алгоритмы интеллектуальной обработки больших объемов данных», или Data Mining [Электронный ресурс]. — Режим доступа: https://www.youtube.com/playlist?list=PLrCZzMib1e9qq4fABTTgA5XPB-N1aahIV.

17. Learn Python — Free Interactive Python Tutorial [Электронный ресурс]. — Режим доступа: https://www.learnpython.org/.
18. Орельен Жерон. Прикладное машинное обучение с помощью Scikit-Learn и TensorFlow: концепции, инструменты и техники для создания интеллектуальных систем. — 2018 г. — 688 с.
19. Силен Дэви, Майсман Арно, Али Мохамед. Основы Data Science и Big Data. Python и наука о данных — СПб: Питер, 2020 г. — 336 с.
20. Medical Speech, Transcription, and Intent (English) [Электронный ресурс]. — Режим доступа: https://www.figure-eight.com/dataset/audio-recording-and-transcription-for-medical-scenarios/
21. Smith, J.W., Everhart, J.E., Dickson, W.C., Knowler, W.C., & Johannes, R.S. (1988). Using the ADAP learning algorithm to forecast the onset of diabetes mellitus. In Proceedings of the Symposium on Computer Applications and Medical Care (pp. 261--265). IEEE Computer Society Press

Для детей

1. Kaggle: Your Machine Learning and Data Science Community [Электронный ресурс]. — Режим доступа: https://www.kaggle.com/.
2. Jupyter Notebook для начинающих: учебник [Электронный ресурс]. — Режим доступа: https://webdevblog.ru/jupyter-notebook-dlya-nachinajushhih-uchebnik/.
3. Обзор библиотек для машинного обучения на Python [Электронный ресурс]. — Режим доступа: https://neerc.ifmo.ru/wiki/index.php?title=%D0%9E%D0%B1%D0%B7%D0%BE%D1%80_%D0%B1%D0%B8%D0%B1%D0%BB%D0%B8%D0%BE%D1%82%D0%B5%D0%BA_%D0%B4%D0%BB%D1%8F_%D0%BC%D0%B0%D1%88%D0%B8%D0%BD%D0%BD%D0%BE%D0%B3%D0%BE_%D0%BE%D0%B1%D1%83%D1%87%D0%B5%D0%BD%D0%B8%D1%8F_%D0%BD%D0%B0_Python.
4. Профессиональный информационно-аналитический ресурс, посвященный машинному обучению, распознаванию образов и интеллектуальному анализу данных [Электронный ресурс]. — Режим доступа: http://www.machinelearning.ru/wiki/index.php?title=%D0%97%D0%B0%D0%B3%D0%BB%D0%B0%D0%B2%D0%BD%D0%B0%D1%8F_%D1%81%D1%82%D1%80%D0%B0%D0%BD%D0%B8%D1%86%D0%B0.
5. Курс «Введение в анализ данных» (2019) [Электронный ресурс]. — Режим доступа: https://www.youtube.com/watch?v=5Yl6_2d0CN0&=&list=PLrCZzMib1e9p6lpNv-yt6uvHGyBxQncEh.
6. Введение в машинное обучение [Электронный ресурс]. — Режим доступа: https://www.coursera.org/learn/vvedenie-mashinnoe-obuchenie.
7. Курс «Алгоритмы интеллектуальной обработки больших объемов данных», или Data Mining [Электронный ресурс]. — Режим доступа: https://www.youtube.com/playlist?list=PLrCZzMib1e9qq4fABTTgA5XPB-N1aahIV.
8. NumPy — Режим доступа: https://numpy.org/.
9. Pandas — Режим доступа: https://pandas.pydata.org/.

10. Matplotlib: Visualization with Python [Электронный ресурс]. — Режим доступа: https://matplotlib.org/.
11. Scikit-learn. Machine Learning in Python [Электронный ресурс]. — Режим доступа: https://scikit-learn.org/stable/.