МИНИСТЕРСТВО ВЫСШЕГО ОБРАЗОВАНИЯ, НАУКИ И ИННОВАЦИЙ РЕСПУБЛИКИ УЗБЕКИСТАН

ТАШКЕНТСКИЙ ГОСУДАРСТВЕННЫЙ ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ ИМЕНИ ИСЛАМА КАРИМОВА

ИНТЕЛЛЕКТУАЛЬНЫЙ АНАЛИЗ ДАННЫХ

МЕТОДИЧЕСКОЕ ПОСОБИЕ

Ташкент 2023

Интеллектуальный анализ данных. Методическое пособие. Юсупбеков А.Н., Гулямов Ш.М., Мухамедханов У.Т., Адилов Ф.Т. Ешматова Б.И., Иваньян А.И. – Ташкент: ТашГТУ, 2023. 34 с.

Методическое пособие разработано совместно творчески сотрудничающими организациями – Ташкентский государственный технический университет и СП ООО «Химавтоматика» для широкого использования современного научно-методического и лабораторного оборудования в совместных учебно-тренинговых центрах ТашГТУ и «Химавтоматика» для подготовки, переподготовки и повышения квалификации специалистов в сфере систем искусственного интеллекта и индустриальной автоматизации.

Методическое пособие разработано для проведения лабораторных работ по предмету «Интеллектуальный анализ данных и машинное обучение». Работа предназначена для магистрантов по специальности 70610701 – Искусственный интеллект.

В данной работе представлены методические рекомендации к выполнению и оформлению лабораторных работ с развитыми конструкторско-расчётными, организационно-техническими частями.

Печатается по решению научно-методического совета ТашГТУ.
Протокол №7 от 26 апреля 2023 г.

Рецензенты: д.т.н., проф. Мусурманов Р.К. (НУУз) ;
д.т.н., проф. Севинов Ж. У. (ТашГТУ)

ВВЕДЕНИЕ

В последние годы искусственный интеллект развивается очень быстрыми темпами, даже сама технология нейронных сетей стала более доступной обычным пользователям. Каждый день разрабатываются новые нейронные сети для биометрических измерений, маркетинговых и научных исследований. Также появилось множество технологий, использующих нейронные сети для ускорения или упрощения процессов. Так, например, корпорация Intel планируют выпускать все новые поколения процессоров со встроенным модулем искусственного интеллекта для ускорения сложных расчетов. Нейросети нашли свое применение и в сфере медицины. Разработано множество интеллектуальных систем, которые способны выявлять заболевания по различным прямым или косвенным признакам. Сфера развлечений также не осталась без внимания. Множество голосовых «ассистентов», поисковых систем, компьютерных игр и бортовых компьютеров автомобилей используют технологии нейронных сетей.

Машинное обучение являет собой самый простой вариант искусственного интеллекта. Оно предполагает, что с помощью различных методов на основе большого количества «тренировочных» данных можно классифицировать или предсказать любой объект, явление или событие.

Актуальность выбранных тем состоит в том, что отсутствуют аналоги, а также электронный вариант поможет улучшить восприятие информации и упростить к ней доступ. Цель работы - разработать лабораторный практикум «Интеллектуальный анализ данных и машинное обучение». В соответствии с поставленной целью в работе определены следующие задачи: проанализировать литературу и интернет-источники по теме «Интеллектуальный анализ данных и машинное обучение» с целью сбора информации о методах машинного обучения и особенностях их применения. Спроектировать структуру лабораторного практикума «Интеллектуальный анализ данных и машинное обучение». Разработать лабораторные работы, видеоуроки и презентационный материал.

Лабораторная работа № 1
МАШИННОЕ ОБУЧЕНИЕ В ИНФОРМАЦИОННОЙ БЕЗОПАСНОСТИ

Цель: Несмотря на всеобщее мнение, что машинное обучение было изобретено в двадцать первом веке — это распространенное заблуждение. В последние годы машинное обучение получило широкое распространение только потому, что появились платформы с достаточной вычислительной производительностью.

Идея сделать машины, которые будут заниматься безопасностью людей уже давно рассматривается как один из самых перспективных вариантов использования машинного обучения. По данным агентства «CB insights» (рис.1) насчитывается около девяноста перспективных проектов, которые пытаются автоматизировать некоторые процессы и задачи безопасности.

Рис.1- Перспективные разработки в области машинного обучения и информационной безопасности

На данный момент главными проблемами использования машинного обучения в информационной безопасности — это избыточное внимание со стороны маркетологов и дезинформация относительно возможностей данных систем. Обычно инвесторов заманивают красивым словом «Искусственный интеллект».

«CB insights» выделяют около 10 направлений использования машинного обучения, но несмотря на это оно не стало «спасением» в области кибербезопасности и на это есть несколько причин.

Первая причина — очень узкая направленность каждой модели машинного обучения. Одна нейросеть неспособна выполнять множество отличающихся друг от друга задач. Если её задачей является распознавание лиц на фотографиях, то распознавать голоса на аудиозаписях она не сможет.

Вторая причина — малый объем данных для качественного обучения модели. Все модели обучаются на данных, полученных заранее. В дальнейшем такая модель может совершать большое количество ошибок и требовать доработки обучающей выборки, что занимает очень много времени и ресурсов.

Третья причина — продукт машинного обучения не может объяснить и ответить за свои решения. В некоторых случаях разработчики такого программного обеспечения ссылаются на то, что невозможно узнать, что именно происходит внутри модели и как именно она пришла к конкретному решению. И именно поэтому на текущий момент инциденты, связанные с информационной безопасностью подтверждают люди, они же несут и ответственность, а машины лишь помогают им в этом.

Однажды машинное обучение сможет эффективно защищать информацию, но оно же может стать и лучшим оружием для совершения атак на информационные системы и ресурсы.

В последнее время больше развивается идея использования машинного обучения не для самих атак, а для проведения операций, связанных с социальной инженерией. Нейросети используют для подделки голоса человека, сбора информации о ком-либо или для генерации фальшивых личностей и страниц в социальных сетях.

Также можно использовать нейросеть для анализа зашифрованного трафика. Данный анализ направлен не на извлечение данных, а на определение трафика через информацию о получателе, отправителе, числе переданных пакетов и их размере, временных параметрах.

Из плюсов данного метода можно выделить:

- возможность обучить нейросеть на виртуальных машинах;
- отсутствие необходимости пропускать через модель весь трафик, а использовать лишь метаданные.

Также перспективой развития машинного обучения в

информационной безопасности является разработка систем поиска уязвимостей. Одним из его направлений можно назвать автоматизацию фаззинга. Фаззинг — это метод тестирования приложений, который подразумевает отправку на вход неправильных, испорченных или неожиданных данных и чаще всего является автоматическим или полуавтоматическим. Машинное обучение хорошо справляется с задачей, где нужно искать закономерности в структурированных или слабоструктурированных данных. Также развивает направление анализа статического кода и динамического анализа исполняемых файлов и процессов с помощью машинного обучения. Глубокий анализ кода позволяет нейросети находить не только уязвимый код, но и похожий на уязвимый.

На данный момент автоматизация процессов защиты с помощью машинного обучения невозможна по причине того, что данных для первоначального обучения всегда очень мало (рис.2) и создавать их достаточно сложно и долго. Из чего следует вывод, что необходимо модифицировать сами методы машинного обучения или тратить большое количество ресурсов и времени на сбор или генерацию данных для первоначального обучения модели.

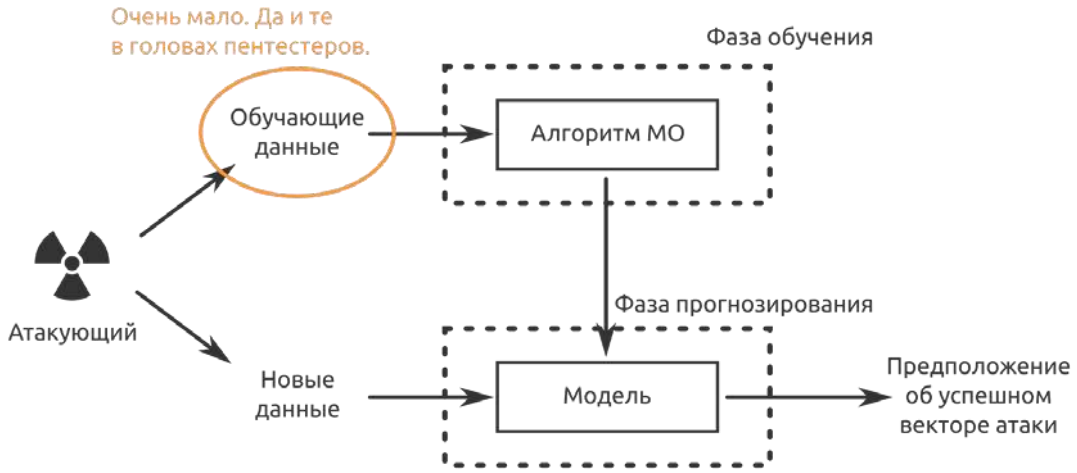

Рис. 2- Схема обучения модели во время реальной атаки.

На данный момент один из самых интересных проектов — это «Deep Exploit» (рис.3).

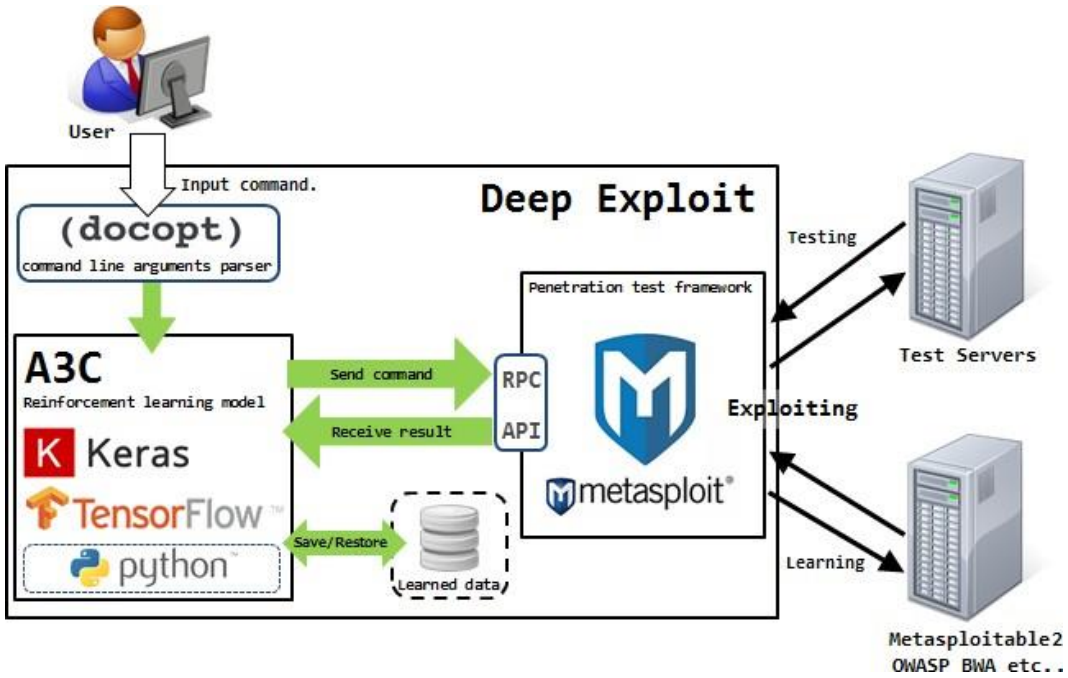

Рис.3- Схема работы системы «Deep Exploit»

Данная система может работать как в режиме сбора данных, так и в режиме брут-форс атаки.

Первый режим позволяет сканировать все возможные открытые порты приложения и тестировать их на уязвимости, которые ранее срабатывали для них в других приложениях.

Второй режим используется как направленная атака на конкретный порт конкретного приложения и применяет к нему все возможные уязвимости и их комбинации.

«Deep Exploit» способна самообучаться и находить новые методы и возможности эксплуатации уязвимостей.

На текущий момент развития искусственный интеллект не может полностью заменить человека, поскольку у обученных машин существуют большие проблемы с выстраиванием логического мышления, чего нет у человека, а ведь это может напрямую повлиять на достижения цели по проникновению в систему. Также машина не может самостоятельно оценивать уровень важности конкретной уязвимости, а значит и не может оценить уровень наносимого урона системе.

1.1. Обзор литературных источников

Для создания лабораторного практикума «Интеллектуальный анализ данных и машинное обучение» были отобраны источники литературы и интернет-источники, которые наиболее ясно, чётко и доступно раскрывают понятие и суть данной темы.

Особенно стоит отметить учебник Флах П. «Машинное обучение. Наука и искусство построения алгоритмов, которые извлекают знания из данных» [28], в котором рассматриваются задачи и проблемы, решаемые методами машинного обучения; этапы и применения разнообразных моделей обучения. Эти данные позволяют глубже понять смысл данной темы и разобраться в ней более конкретно.

В книге Вьюгина В. В. «Математические основы теории машинного обучения и прогнозирования» [7] дано конкретное определение понятия машинное обучение, помогает уяснить некоторые современные математические проблемы данной области и их решения. Первая часть книги — статистическая теория машинного обучения — использует методы теории вероятностей и математической статистики. В основе данного подхода лежит предположение о том, что наблюдаемые исходы генерируются вероятностным источником, возможно, с неизвестными параметрами.

Книга «Машинное обучение» [28] рассчитана на тех, кто хочет решать самые разнообразные задачи при помощи машинного обучения. Как правило, для этого нужен Python.

В курсе лекций Загинайлова Ю. Н. [14] изложены теоретические осно- вы информационной безопасности технической системы. Приведены объекты обеспечения информационной безопасности, угрозы объектам, политики и структуры систем обеспечения информационной безопасности. Рассмотрены понятия и классификации защищаемой информации, угроз безопасности информации, объектов, способов, средств и систем защиты информации.

В книге «Введение в машинное обучение с помощью Python. Руководство для специалистов по работе с данными» [12] рассказывается о том, что машинное обучение стало неотъемлемой частью различных коммерческих и исследовательских проектов, однако эта область не является прерогативой больших компаний с мощными аналитическими командами. Эта книга научит

практическим способам построения систем машинного обучения, даже если пользователь еще новичок в этой области.

В учебном пособии Васильева В. И. «Интеллектуальные системы защиты информации» [5] рассмотрены основы построения интеллектуальных систем защиты информации в корпоративных информационных системах. Автор сделал акцент на построении биометрических систем идентификации личности, систем обнаружения и предотвращения вторжений, анализа и управления информационными рисками. Изложены современные подходы к созданию данного класса систем с использованием методов теории нейронных сетей, искусственных иммунных систем, нечетких когнитивных моделей.

В течение последнего десятилетия произошел взрыв в вычислительных и информационных технологиях. С его помощью поступают огромные объемы данных в различных областях, таких как медицина, биология, финансы и маркетинг. Задача понимания этих данных привела к разработке новых инструментов в области статистики и породила новые области, такие как интеллектуальный анализ данных, машинное обучение и биоинформатика. Многие из этих инструментов имеют общие основы, но часто выражаются с другой терминологией. В книге «The Elements of Statistical Learning: Data Mining, Inference, and Prediction» описываются важные идеи в этих областях в общих концептуальных рамках. Хотя этот подход является статистическим, акцент делается на концепциях, а не на математике. Приводится много примеров с использованием цветной графики. Это ценный ресурс для статистиков и всех, кто интересуется разработкой данных в науке или промышленности. Охват книги обширен: от контролируемого обучения (прогнозирования) до неконтролируемого обучения. Многие темы включают в себя нейронные сети, вспомогательные векторные машины, деревья классификации — первое всестороннее рассмотрение этой темы в любой книге. В этом крупном новом выпуске представлены многие темы, не затронутые в оригинале, включая графические модели, случайные леса, ансамблевые методы, алгоритмы наименьшей регрессии и пути для лассо, неотрицательной матричной факторизации и спектральной кластеризации.

В книге «Распознавание образов. Построение и обучение вероятностных моделей» [21] рассматриваются несколько практически важных примеров решения задач статистического

обучения, в которых пространства признаков и ответов и обучающие наборы устроены слишком сложно и нерегулярно, так что стандартные методы статистического обучения в них нельзя применить буквально, но можно применять после построения адекватных вероятностных моделей. Значительная часть описываемых методов строго обоснована: простые технические детали доказательств сформулированы и предложены в качестве упражнений, более сложные, но не слишком громоздкие доказательства предъявлены.

Книга «Машинное обучение: новый искусственный интеллект» Э. Алпайдина [2] представляет собой краткое введение в машинное обучение. Книга дает общее представление о машинном обучении, описывает суть основных алгоритмов обучения без погружения в технические подробности и обсуждает некоторые примеры их применения на уровне, достаточном для понимания основ.

1.2. Обзор интернет-источников

При анализе интернет-источников необходимо выделить работу «Машинное обучение для чайников» [20] размещенную на интернет-ресурсе
«Newtonew.com». В данной статье рассказывается для чего собственно нужна эта технология и приведены грамотные примеры использования, которые помогут разобраться с нуля в данной теме.
Приведены классы задач машинного обучения:

- задача регрессии: на основании различных признаков предсказать вещественный ответ;
- задача классификации: на основании различных признаков предсказать категориальный ответ;
- задача кластеризации: разбиение данных на похожие категории;
- задача уменьшения размерности: научиться описывать данные не N признаками, а меньшим числом;
- задача выявления аномалий: на основании признаков научиться различать аномалии от «не-аномалий».

Несмотря на множество преимуществ предыдущей статьи удалось найти видеолекцию с полным курсом про машинное обучение, которую ведет Воронцов К. В. [12]. Он рассказывает про основные понятия машинного обучения: объект, ответ, признак, предсказательная модель, метод обучения, эмпирический риск,

переобучение. И показывает всё на наглядных примерах, что очень помогает в понимании материала.

В работе «Машинное обучение» [12] рассказывается о том, что обширный подраздел искусственного интеллекта, изучающий методы построения алгоритмов, способен обучаться. Машинное обучение — не только математическая, но и практическая, инженерная дисциплина. Чистая теория, как правило, не приводит сразу к методам и алгоритмам, применимым на практике. Чтобы заставить их хорошо работать, приходится изобретать дополнительные эвристики, компенсирующие несоответствие сделанных в теории предположений условиям реальных задач. Практически ни одно исследование в машинном обучении не обходится без эксперимента на модельных или реальных данных, подтверждающего практическую работоспособность метода.

В статье Генрихова И. В. одной из центральных задач распознавания образов является задача распознавания по прецедентам. Известным инструментом решения данной задачи являются деревья решений. Синтез классического решающего дерева представляет собой итерационный процесс. Как правило, для построения очередной внутренней вершины дерева выбирается признак, который наилучшим образом удовлетворяет некоторому критерию ветвления, т.е. наилучшим образом разделяет текущее множество обучающих объектов.

Лабораторная работа № 2
АНАЛИЗ ДАННЫХ И МАШИННОЕ ОБУЧЕНИЕ

Цель: Описание интерфейса программного продукта. В качестве интерфейса программного продукта был выбран веб-сайт. Он был разработан на основе фреймворка Django версии 2.0 для Python. В качестве базы данных используется PostgreSQL 11, которая распространяется на бесплатной основе. Python является интерпретируемым языком программирования и в последние годы стал очень популярным.

Все модели базы данных также создаются Django самостоятельно через «конектор». Необходимо лишь создать описание модели в файле «models.py» и произвести миграции.

Одним из главных преимуществ Django это многофункциональная и гибкая «админ-панель» (рис.4). Она очень легко настраивается под нужды администратора и поддерживает подключение различных модулей. Также через нее можно создавать и редактировать различные записи на сайте и в базе данных.

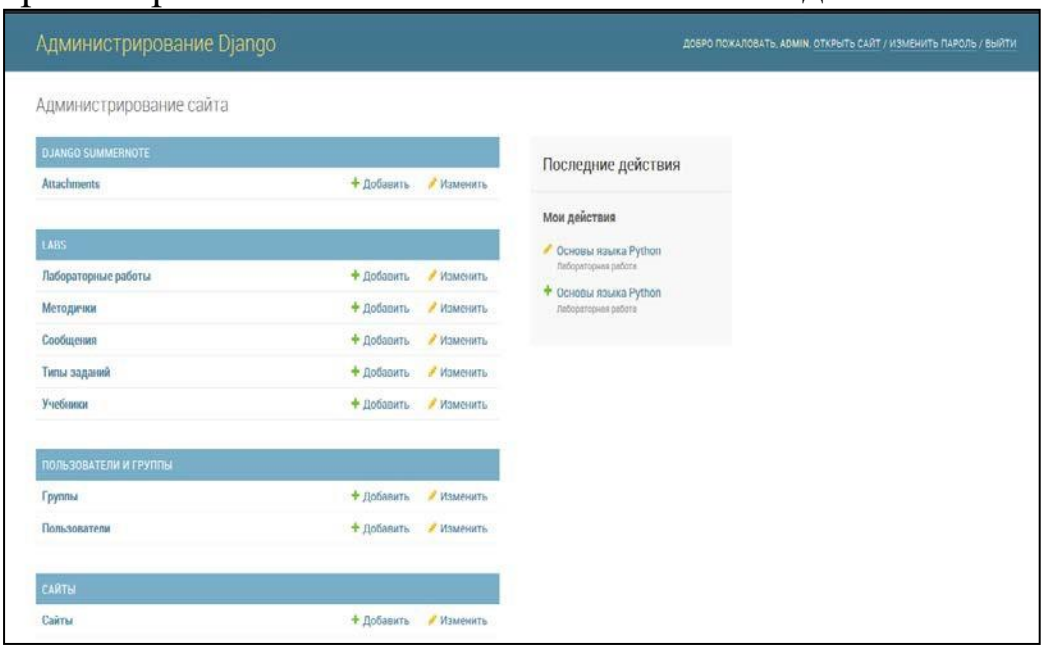

Рис.4- Внешний вид «админ-панели»

«Админ-панель» является инструментом для добавления лабораторных работ, проектных заданий и методического материала на страницы сайта (рис.5). Также модуль «DjangoSummerNote» позволяет удобно редактировать содержание лабораторной работы: изменять шрифты, добавлять к нему декорации, прикреплять изображения и видеофайлы (рис.6).

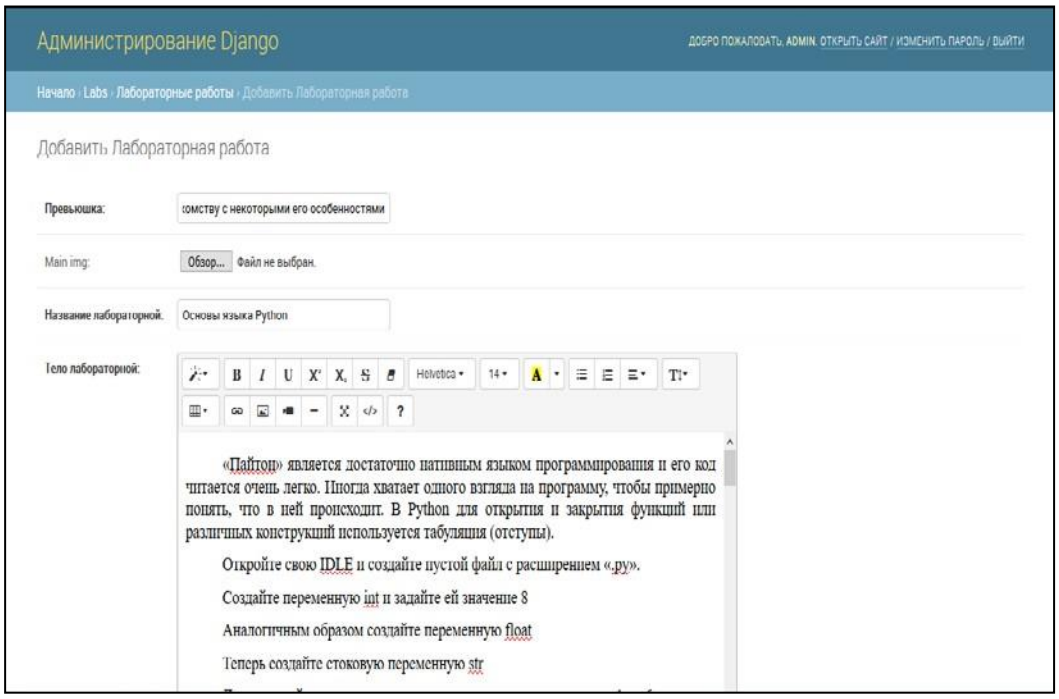

Рис.5- Добавление лабораторных работ через «админ-панель»

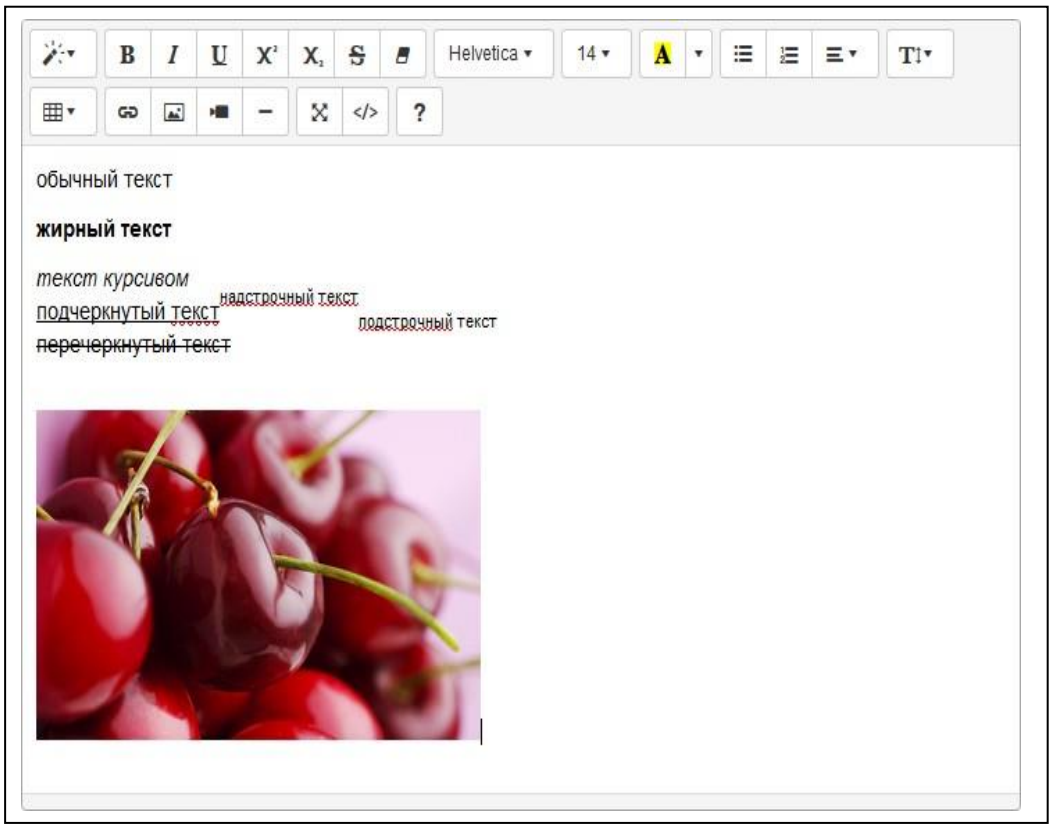

Рис.6- Модуль «DjangoSummerNote».

При переходе на сайт пользователь попадает на страницу приветствия (рис. 7). В дальнейшем на ней появится не только текстовое, но и видеоприветствие с описанием.

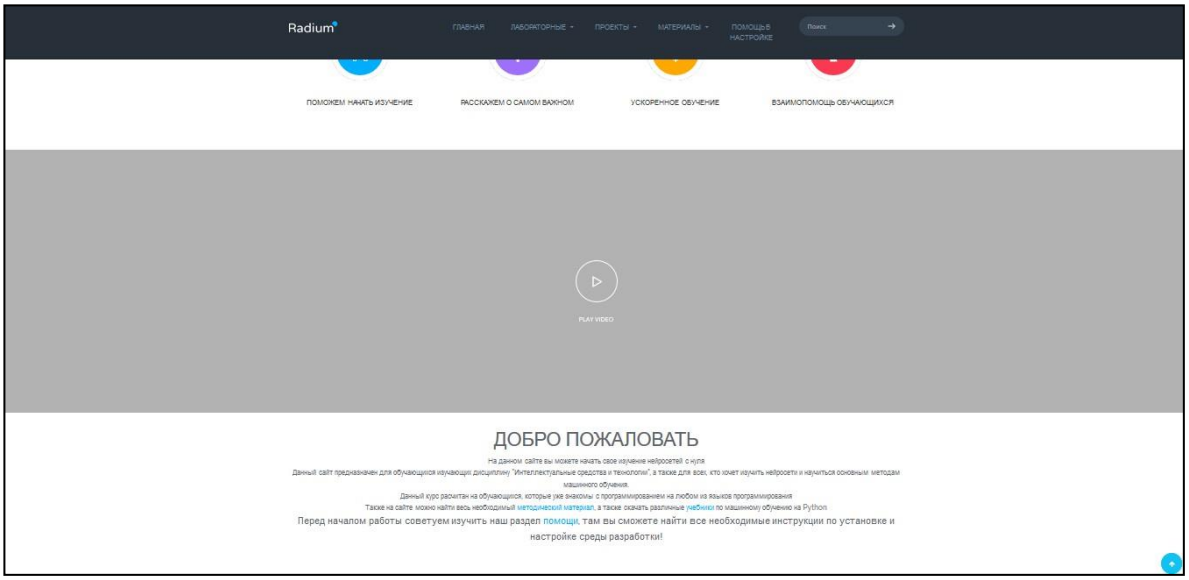

Рис.7- Главная страница сайта

Пользуясь меню сайта, пользователь может перейти на страницы с лабораторными работами (рис.8), проектными заданиями (рис.9), различными полезными ресурсами, ссылками на книги, статьями и другими полезными источниками.

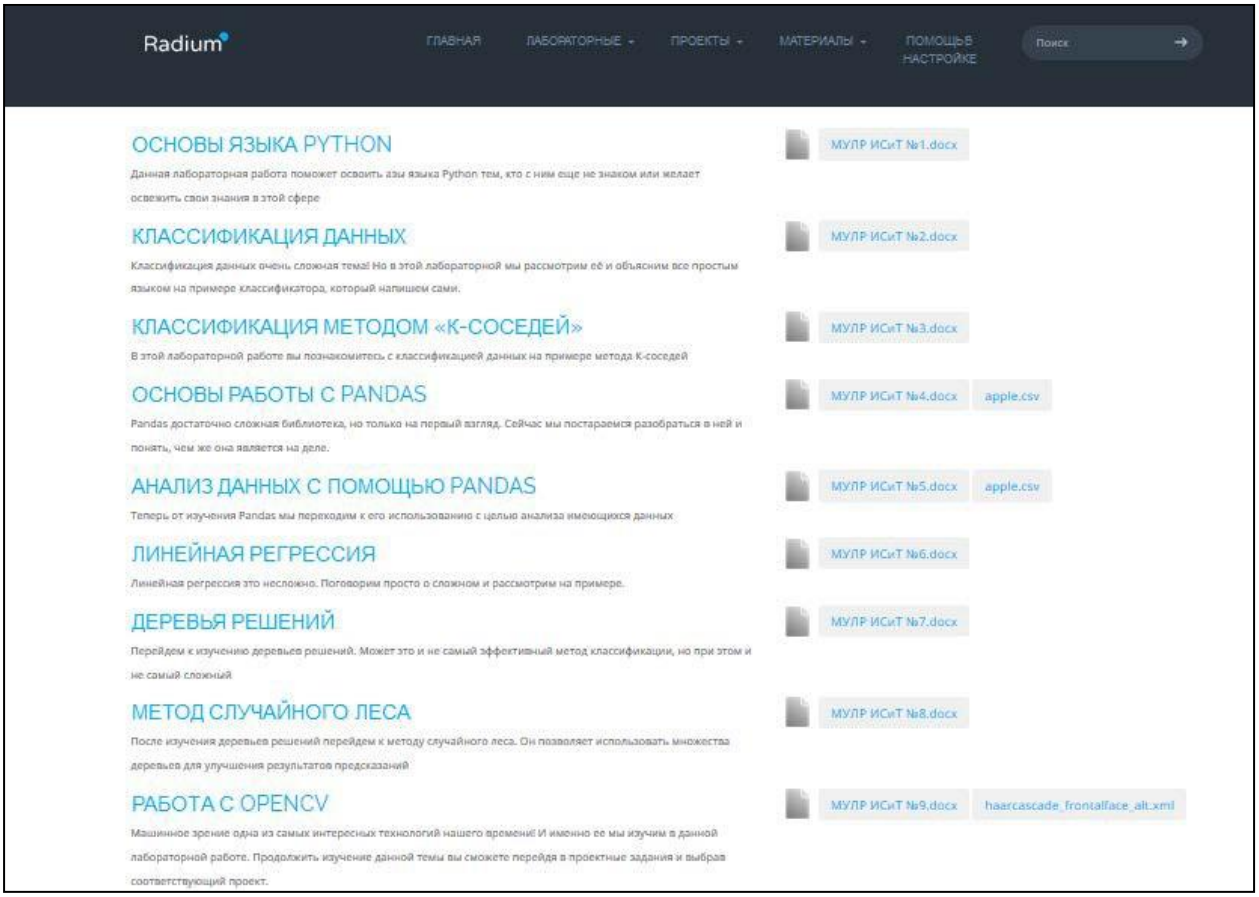

Рис.8 - Страница со списком лабораторных работ

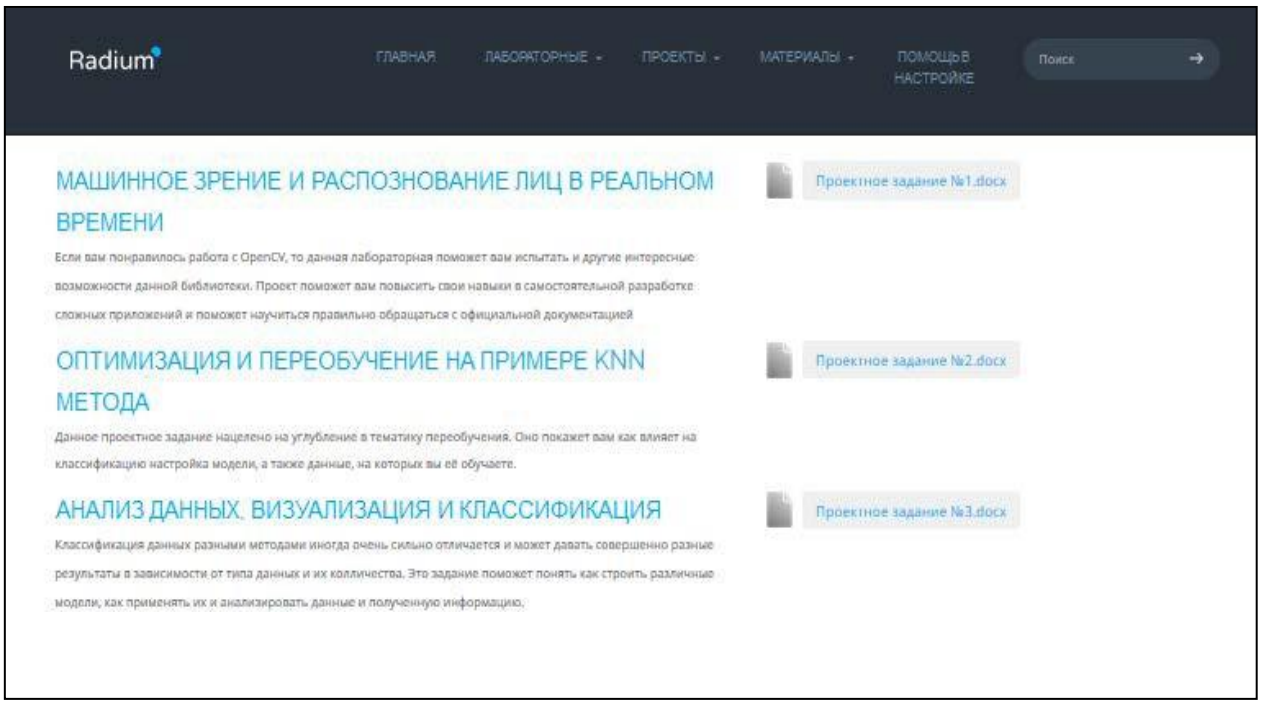

Рис. 9- Страница со списком проектных заданий

На странице с лабораторной работой (рис.10) находится сама лабораторная работа с целью, задачами, описанием хода работы и вопросами для самоконтроля. Также вначале каждой страницы есть видеоплеер, в котором можно посмотреть видеоурок с выполнением лабораторной работы с подробными объяснениями. На странице со списками лабораторных и проектных работ прикреплены файлы для выполнения лабораторных работ, а также документ с методическими указаниями по выполнению лабораторных работ, для тех, кто хочет работать офлайн.

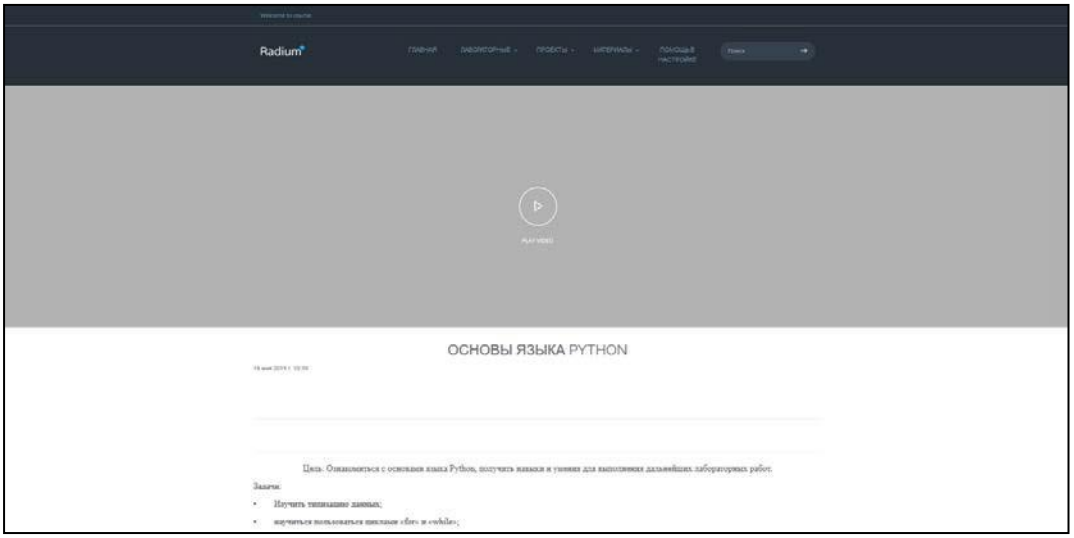

Рис.10- Страница лабораторной работы

В разделе «Учебники» (рис.11) пользователь может найти ссылки на различные учебники и методички по машинному обучению и анализу данных.

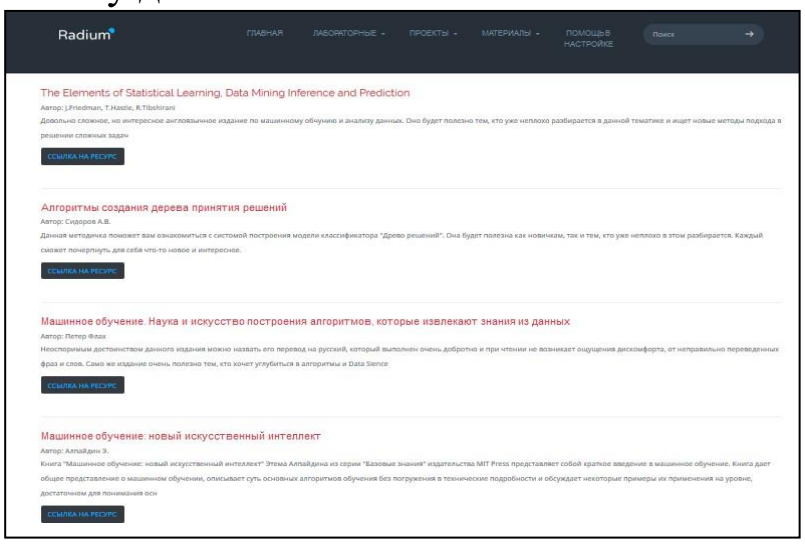

Рис.11- Раздел «Учебники»

Раздел «Полезные ссылки» (рис.12) содержит ссылки на различные онлайн курсы, форумы и статьи на тему машинного обучения. В некоторых случаях они имеют большое преимущество над учебниками в том плане, что объяснение там идет более «простым» языком, понятным тем, кто только начал изучать данную тематику.

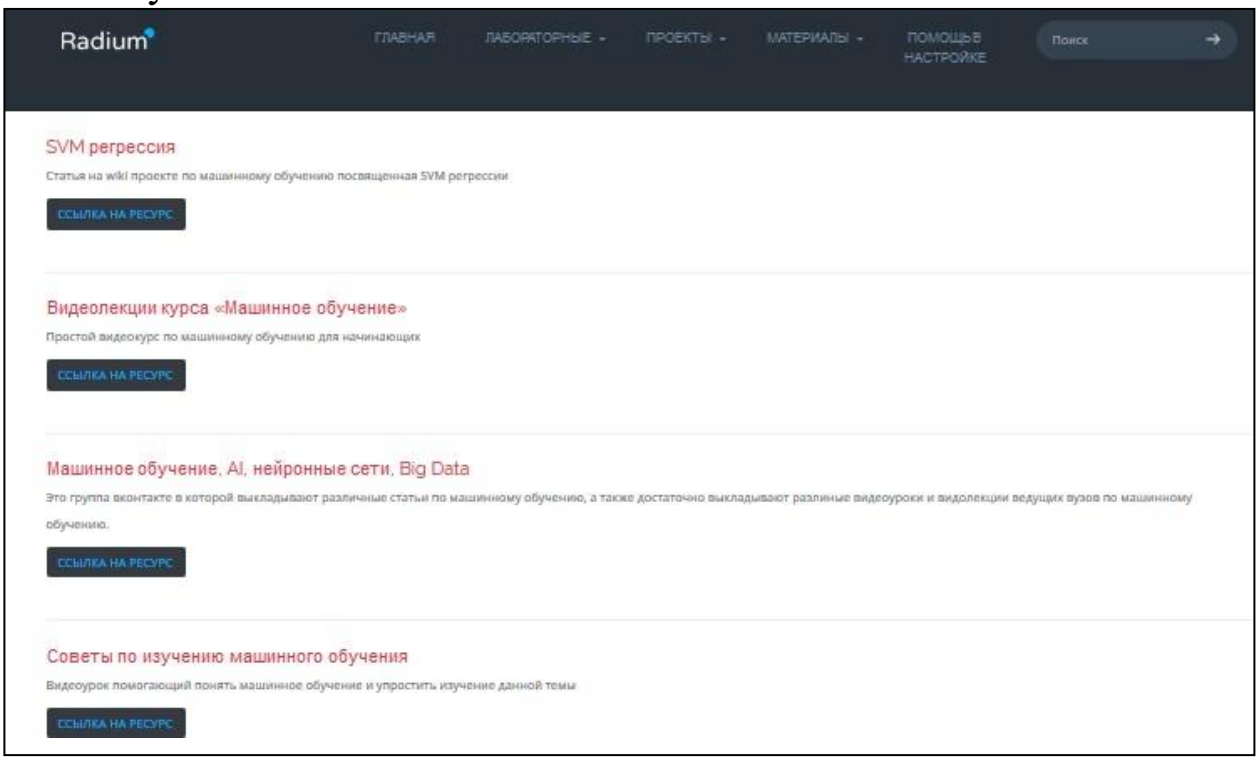

Рис.12- Раздел «Полезные ссылки»

После того, как у обучающихся возникли проблемы с настройкой среды разработки, был добавлен раздел «помощь в настройке». Он предназначен для того, чтобы помочь обучающимся выбрать редактор кода, установить и настроить интерпретатор языка Python и установить все необходимые для выполнения лабораторных работ библиотек.

Поскольку лабораторные работы должны выполняться последовательно, то стрелочками был указан порядок выполнения (рис.13).

Рис.13- Структура заданий и материалов

В отличие от лабораторных работ, проектные задания выполняются по желанию и не имеют четкой последовательности.

Помощь по настройке предназначена для того, чтобы помочь обучающимся настроить интерпретатор, выбрать себе подходящую среду разработки, а также помогает с установкой всех необходимых библиотек Python.

Лабораторная работа №3
ОСНОВЫ ЯЗЫКА PYTHON

Цель: ознакомиться с основами языка Python, получить умения для вы- полнения дальнейших лабораторных работ.

Задачи:

- изучить типизацию данных;
- научиться пользоваться циклами «for» и «while»;
- рассмотреть «ветвление» в Python;
- отработать задачи с использованием конструкции «try-except»;
- разобрать функции и пространства имён.

Данная лабораторная работа предназначена для ознакомления обучающихся с языком Python и помощи в выполнении лабораторных и проектных заданий. В ней обучающиеся поэтапно проходят все основные аспекты языка такие, как:

1. Типизация данных.
2. Пространство имён.
3. Функции.
4. Циклы.
5. Ветвления.
6. Исключения.

Сформировать у обучающихся понятие о переобучении и научить оптимизации модели машинного обучения: закрепить умения по использованию метода, сформировать понятие переобучения, научиться оптимизировать модель.

Любая модель обучения имеет свои особенности и черты, но есть вещь, которая объединяет их — это переобучение. В данной работе обучающиеся будут тестировать свою модель на устойчивость и рассматривать на графиках последствия переобучения, пытаться избегать подобных эффектов и правильно настраивать свою модель.

Лабораторная работа №4
КЛАССИФИКАЦИЯ ДАННЫХ

Цель: научиться работать с данными при помощи визуальных инструментов и разобрать азы классификации при помощи построения простейшего классификатора со статичными параметрами.

Задачи:

- научиться анализировать данные;
- сформировать понятие математических срезов;
- получить умения в работе с визуальными инструментами;
- построить классификатор на основе данных, полученных при анализе;
- научиться калибровать нейросеть для получения более точных ответов.

Первая лабораторная работа, в которой начинается изучение принципов машинного обучения. В ней обучающиеся будут учиться работать с графической информацией и производить анализ полученных данных и самостоятельно строить простейший классификатор. Данная работа поможет понять простейшие понятия и принципы работы более сложных классификаторов, научиться использовать различные методы классификации и визуализировать данные разными средствами. Кроме того, научиться использовать различные методы классификации, научиться работать с имеющимися данными и закрепить умения по использованию инструментов визуализации.

Работа с данными — это один из самых сложных этапов машинного обучения и данная работа поможет обучающимся еще глубже погрузиться в неё. Обучающиеся должны рассмотреть, как будут вести себя различные модели машинного обучения при работе с различными наборами данных. Так как все модели по-разному работают с разными типами данных, а некоторые и вовсе поддерживают только определенные типы данных, то и результат может очень сильно отличаться.

Лабораторная работа № 5
Классификация методом «К-ближайших соседей»

Цель: изучить простейший метод классификации данных

«К-ближайших соседей» и научиться производить оценку данных с помощью визуальных инструментов Python.

Задачи:

- детально разобрать метод машинного обучения «К-ближайших со-седей»;
- научиться работать с информацией;
- сформировать понятие математических срезов;
- получить умения в работе с визуальными инструментами;
- научиться калибровать нейросеть для получения более точных от- ветов.

В этой лабораторной работе будет рассматриваться метод машинного обучения «К-ближайших соседей», а также его положительные и отрицательные стороны; рассмотрим библиотеку для визуализации данных и математического анализа «matplotlib»; будем производить калибровку нейронной сети для получения максимального процента правильных ответов, а также обсудим проблему переобучения в нейросетях.

Описание презентации на тему «Метод "К-ближайших соседей"»

Презентационный материал (рис. 3.1) по теме «Линейной регрессии» был подобран и скомпонован таким образом, чтобы объяснить материал тем обучающимся, которые еще не были знакомы с машинным обучением и имеют слабое представление о том, что же это такое.

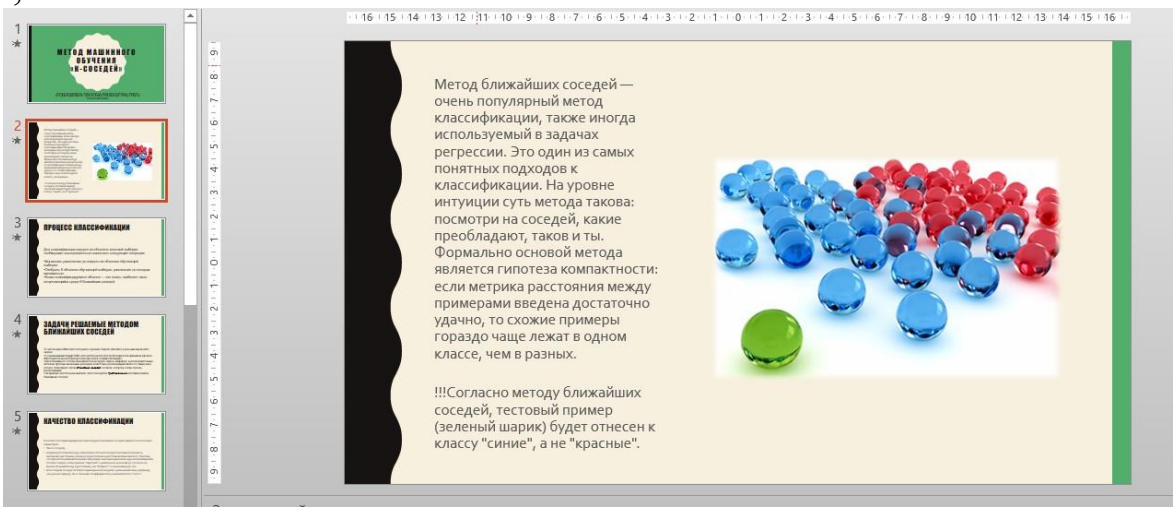

Рис.3.1- Презентационный материал по теме «Метод «К-ближайших соседей»

В подобранном материале даются простые определения и понятные примеры, а также некоторая часть информации о строении алгоритмов данного метода машинного обучения.

Лабораторная работа № 6
ОСНОВЫ РАБОТЫ С PANDAS

Цель: научиться пользоваться библиотекой Pandas и её встроенными объектами для визуализации данных в данных сетях.

Задачи:

- получить умения по использованию библиотеки Pandas;
- сформировать понятия о DataFrame и Series;
- научиться строить графики с помощью scatter matrix (матрица рассеивания) и matplotlib.

Pandas — это библиотека языка Python, которая позволяет создавать объекты, в которых данные хранятся в табличной форме. Это основной структурный объект, необходимый для анализа и очистки данных перед их использованием в обучении модели. Это первая лабораторная работа, связанная с Pandas. В ней обучающиеся разберут структуру и основные функции данной библиотеки.

Лабораторная работа №7
АНАЛИЗ ДАННЫХ С ПОМОЩЬЮ PANDAS

Цель: научиться пользоваться библиотекой Pandas и её встроенными объектами для анализа данных в данных сетях.

Задачи:

- получить умения по использованию библиотеки Pandas;
- научиться анализировать и обрабатывать данные с помощью Pandas;
- закрепить умения визуализации в Pandas.

В продолжении к лабораторной работе № 4 обучающиеся переходят к анализу данных с помощью библиотеки Pandas. В данной работе обучающиеся научатся делать срезы, группировки, индексацию данных по разным параметрам, а также строить графики по созданным срезам. Знания и умения полученные в ходе выполнения работы, помогут понять, некоторые механизмы машинного обучения.

Лабораторная работа №8
ЛИНЕЙНАЯ РЕГРЕССИЯ

Цель: понять и научиться применять метод линейной регрессии в машинном обучении.

Задачи:

- изучить модель линейного регрессора;
- произвести обучение модели;
- рассмотреть особенности данного метода машинного обучения;
- произвести предсказание на основе созданной модели.

Линейная регрессия является одним из самых универсальных методов машинного обучения, т.к. может принимать на вход данные различных типов и структур. Также данная модель достаточно устойчива к переобучению и имеет большую скорость обработки данных. В этой лабораторной работе будут рассматриваться все нюансы и принципы работы этого метода классификации.

Описание презентации на тему «Метод линейной регрессии»

Линейная регрессия является одним из сложных для понимания методов машинного обучения, так как имеет достаточно сложные алгоритмы работы. Материал (рис.6.1) для презентации подобран таким образом, чтобы очень доступно объяснить сложные вещи обучающимся.

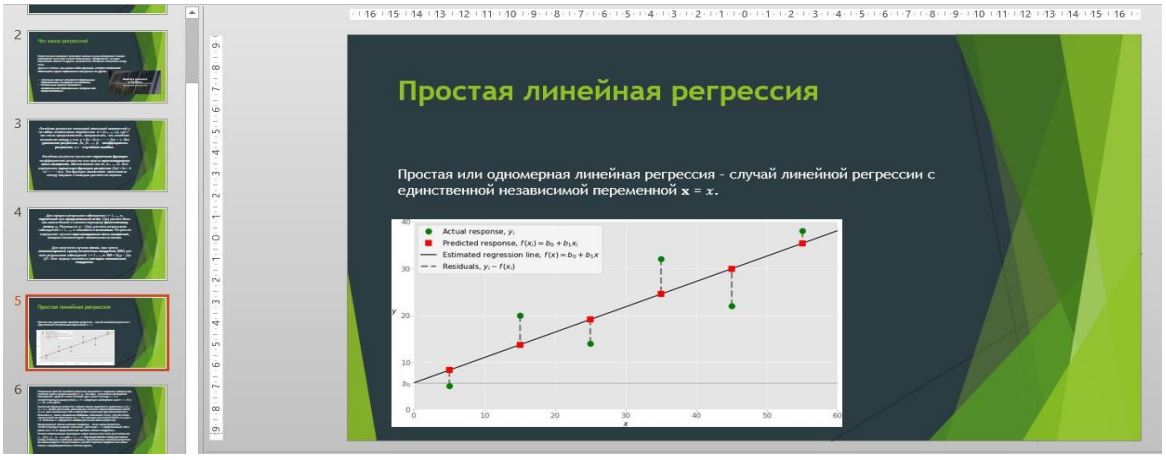

Рис.6.1 — Презентационный материал по теме «Метод линейной регрессии»

Несмотря на то, что данный метод очень тяжелый для понимания, он очень прост в использовании и позволяет получать достаточно хорошие предсказания.

Лабораторная работа №9
ДЕРЕВЬЯ РЕШЕНИЙ

Цель: познакомить обучающихся с методом машинного обучения, построенном на деревьях решений, а также научить строить сами деревья.

Задачи:

- рассмотреть понятие дерева решений;
- рассмотреть варианты применения данной классификации;
- обучить модель на основе классов;
- отобразить дополнительный класс на модели и посмотреть результат;
- рассмотреть плюсы и минусы данной модели.

Одним из самых простых для понимания является классификатор «Древо решений». Эта модель строит «дерево» классов, у каждого из которых есть свои параметры. Попадающие под эти параметры данные причисляются к тому или иному классу. Модель очень чувствительна к переобучению и данным, которые получает на вход. Лабораторная работа поможет нам разобраться во всех тонкостях данного метода и покажет все его нюансы.

Описание презентации на тему «Метод "Деревья решений"»

Среди методов машинного обучения «Древо решений» является одним из самых наглядных и простых для понимания. В презентации данный метод объясняется на примере классической задачи (рис.9.1) из теории вероятности. Подробно описывается алгоритмический ход работы метода машинного обучения с приложением математических расчётов и графиков.

Рис.9.1 — Пример работы древа решений. (часть 1)

После окончания всех расчетов предоставляется конечный граф (рис. 9.2), на примере которого можно легко понять причины решений модели. Данный метод машинного обучения является одним из самых простых для интерпретации и позволяет отслеживать причины принятия каждого решения и уступает в этом плане только методу «Случайного леса».

Рис.9.2 — Пример работы древа решений. (часть 2)

Но несмотря на все плюсы данного метода он очень неустойчив к переобучению и при плохой обучающей выборке будет выдавать очень плохие результаты предсказаний. По этой причине он используется только в достаточно простых задачах классификации.

Лабораторная работа №10
МЕТОД СЛУЧАЙНОГО ЛЕСА

Цель: сформировать понятие случайного леса, а также научить обучающихся использовать данную модель для решения задач.

Задачи:

- рассмотреть понятие случайного леса;
- рассмотреть пример кода для решения простых задач;
- научить подбирать параметры модели для улучшения качества прогнозов модели.

Случайный лес является модифицированным вариантом «Древа решений» и не имеет всех его отрицательных особенной, например, он не чувствителен к переобучению. В лабораторной работе будет рассмотрено строение данного метода классификации и поговорим об его устройстве.

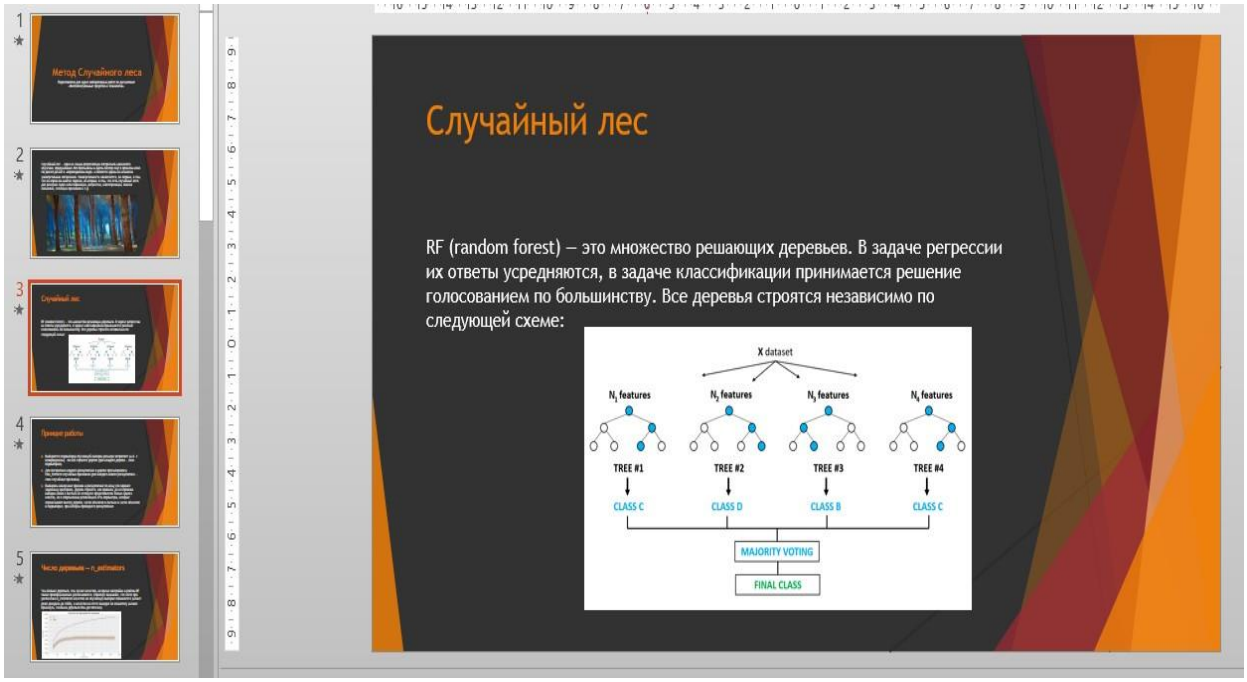

Рис. 10.1 — Презентационный материал по теме «Случайный лес»

Описание презентации на тему «Метод Случайного леса»

Метод «Случайного леса» является самым популярным методом машинного обучения для решения большинства классических задач регрессии, классификации и предсказания. По отдельным данным, данный метод используется в 70 % решаемых задач. Он устойчив к переобучению и позволяет очень быстро решать сложные задачи с точностью предсказаний приближенными к 95 % и более правильных ответов.

В материале (рис.10.1) для данной презентации рассказывает о самом «Случайном лесе», его математическом устройстве и о всех возможных настройках модели, которые помогают оптимизировать точность предсказаний.

Для каждого параметра настройки выделены отдельные слайды с описанием самого параметра, возможных его значений и наглядными представлениями того, как он влияет на процесс классификации или регрессии.

Лабораторная работа №11
РАБОТА С OPENCV

Цель: научить обучающихся основам работы с машинным зрением и показать основные алгоритмы работы с ним.

Задачи:

- разобрать импорт и просмотр изображения;
- разобрать кадрирование;
- научиться изменять размер изображения;
- научиться переворачивать изображение;
- рассмотреть способ преобразование изображения в черно-белое;
- научиться работать со сглаживанием и размытием;
- изучить метод распознавания лиц.

В последние годы машинное зрение получило большое распространение, вызвало интерес со стороны не только ученых, но и различных инженеров и разработчиков «интеллектуальных» приложений. В лабораторной работе обучающиеся рассмотрят вариант библиотеки OpenCV, написанной на языке «C++» для Python, опробуем некоторый её функционал и протестируем классификатор для распознавания лиц. Кроме того, сформируется умение по использованию библиотеки OpenCV самостоятельному поиску информации, связанной с решением поставленной задачи:

- закрепить умение работы с машинным зрением;
- сформировать умение поиска информации;
- сформировать умения по разработке полноценного приложения;
- научиться отрабатывать ошибки при разработке приложения.

Самостоятельная работа - неотъемлемая задача каждого обучающегося и именно поэтому обучающимся предоставляется возможность самим развиваться в данной области. Проектное задание «Машинное зрение и распознавание лиц в реальном времени» позволит обучающимся больше углубиться в изучение библиотеки OpenCV, а также научит их использовать официальную документацию к программным продуктам.

Больше всего трудностей у обучающихся может возникнуть лабораторная работа под названием «Работа с OpenCV». Данная лабораторная работа является одной из самых сложных частей, связанных с изучением искусственного интеллекта, и знакомит обучающихся с машинным зрением. Чтобы помочь обучающимся в изучении и понимании материала видеоурок по данной теме был

переработан и дополнен краткими справочными вставками, а также было более детально рассмотрено последнее задание по распознаванию лиц на изображениях.

Так как видео лекции являются одним из самых эффективных наглядных методов обучения, было решено записать видеоурок к каждой лабораторной работе.

Каждый видеоурок прикрепляется к лабораторной работе и отображается на ее странице в разработанном интерфейсе в виде онлайн видеоплеера (рис.11.1).

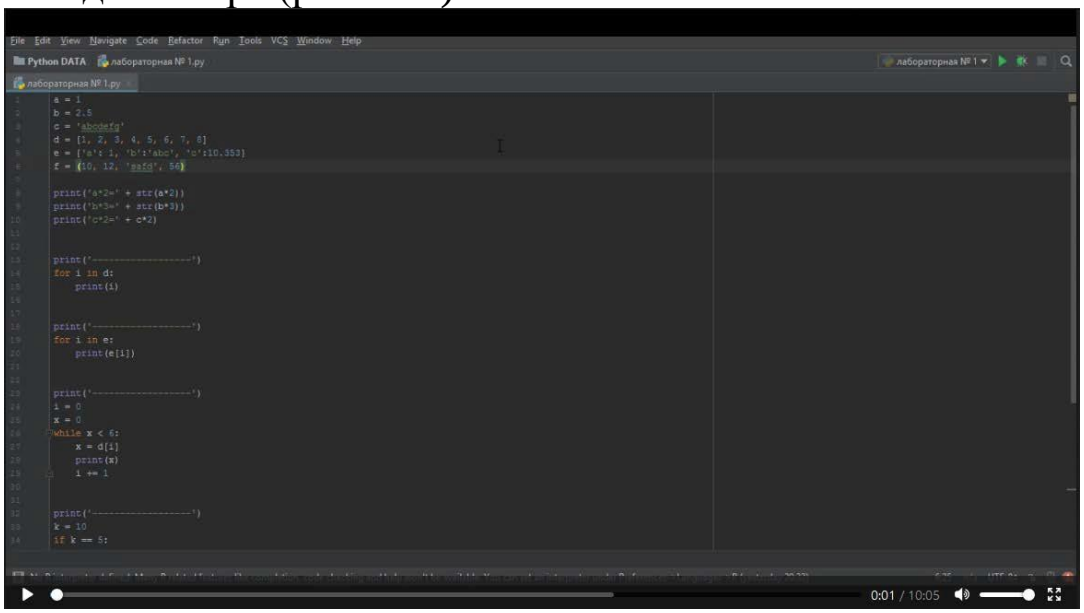

Рис. 11.1. Видеоплеер

В каждом видеоуроке даётся краткое изложение теоретического материала. Далее на примере написанного кода объясняется принцип работы различных методов машинного обучения.

Лабораторные работы и видеоуроки по темам «Основы работы с Pandas» и «Анализ данных с помощью Pandas» направлены на изучение библиотеки Pandas. В видеоуроке подробно объясняется применение библиотеки и её элементов в машинном обучении, а также приведены примеры работы с данными элементами.

Самый первый видеоурок посвящён ознакомлению с языком программирования Python и в нём рассказывается о различных особенностях и возможностях этого языка программирования, рассматриваются основные структурные компоненты, типизация данных, циклы, ветвления, исключения, математические функции с разными типами данных и преобразование типов данных.

Одним из самых важных является видеоурок к лабораторной работе

«Работа с OpenCV». Данная лабораторная работа является одной из самых сложных в курсе и может вызвать у обучающихся затруднения при её выполнении. Наличие видеоурока облегчает выполнение задачи для обучающегося, а также поможет улучшить усвоение как теоретического, так и практического материала лабораторной работы.

Презентационное сопровождение лабораторных работ

Машинное обучение является сложной для понимания темой именно на уровне алгоритмов работы методов машинного обучения. Чтобы помочь обучающимся разобраться в алгоритмах машинного обучения были разработаны следующие презентационные материалы:

1. Метод «К-ближайших соседей».
2. Метод «Линейной регрессии».
3. Метод «Дерева решений».
4. Метод «Случайного леса».

Каждый презентационный материал направлен на углубленное изучение каждого отдельно взятого алгоритма машинного обучения и позволяет лучше понять принцип его работы, так как понимание принципа позволяет более рационально и осознано применять каждый из алгоритмов.

ЗАКЛЮЧЕНИЕ

Машинное обучение и нейросети стали новой отраслью науки, технологий и бизнеса. Они позволяют решать самые разнообразные задачи, будь то распознавание лиц и поиск людей, медицинский анализ, анализ рынка продаж, предсказание катастроф или подбор музыки, которая вам может понравиться. Машинное обучение — это не обучение в общепринятом понимании. Даже глубокое обучение не позволяет машине стать по-настоящему интеллектуальной. Её решения складываются исключительно из ранее изученных ситуаций и не могут порождать другие нестандартные или нелогичные на первый взгляд решения или ответы, как это может делать человек. Ведь человек может выстраивать свои логические цепочки, основываясь не только на полученном ранее опыте, но и на основе своих догадок и предчувствий, а также на основе неявных связей между вещами и явлениями. В последнее время потребность в таком мышлении сильно возросла и машинное обучение стало применяться повсеместно и находит все новые и новые сферы применения. Если раньше оно применялось только в IT-сфере и алгоритмизации, то

сейчас машинное обучение начинает просачиваться во все сферы деятельности и всё более ориентируется на потребителей.

Большая потребность в сфере разработки нейросетевых технологий породило большое количество различных форумов, книг, журналов и онлайн курсов, связанных с изучением данной тематики. Именно данный скачок интереса позволил развиться этой технологии в краткие сроки и вызвать интерес у людей со всего мира. У большинства материалов, связанных с искусственным интеллектом, есть проблема с доступностью излагаемой информации, они написаны сложным техническим языком и это вызывает у обучающихся трудности с усвоением материала. Переработка материала является первоочередной задачей при создании любого курса по изучению искусственного интеллекта, так как доступность этого материала напрямую повлияет на качество знаний и умений, полученных обучающимися.

Именно по этим причинам данная тема очень актуальна в наше время. Также данная тема оказалась очень интересной, потому что «искусственный интеллект» вызвал много интереса как со стороны простых людей, так со стороны учёных и бизнесменов. Машинное обучение уже проникло в такие сферы деятельности как:

1. IT-сфере: приложения на основе нейросетей, машинное зрение.
2. Кибербезопасность.
3. Маркетинг.
4. Медицина.
5. Диагностика техники.
6. Биоинформатика.

ЛИТЕРАТУРА

1. Алгоритмы создания дерева принятия решений [Электронный ресурс]. — Режим доступа: http://econf.rae.ru/pdf/2014/03/3245.pdf (дата обращения: 09.04.2019).
2. Алпайдин Э. Машинное обучение: новый искусственный интеллект [Текст] / Э. Алпайдин. — Москва: Точка, 2017. — 208 с.
3. Баргесян А. А. Технологии анализа данных: Data Mining, Visual Mining, Text Mining, OLAP [Текст] / А. А. Баргесян, М. С. Куприянов, В. В. Степаненко, и т. д. — 2-е изд., перераб. и доп. — Санкт-Петербург: БХВ-Петербург, 2007. — 384 с.
4. Бринк Х. Машинное обучение [Текст] / Х. Бринк, Дж. Ричардс, М. Феверолф. — пер. с англ. Рузмайкина И. — Санкт-Петербург: Питер, 2017. — 336 с.
5. Васильев В. И. Интеллектуальные системы защиты информации [Текст]: учебное пособие / В. И. Васильев. — изд. 2-е, испр. — Москва: Ма- шиностроение, 2012. — 171 с.
6. Видеолекции курса «Машинное обучение» [Электронный ресурс]. —Режим доступа: https://yandexdataschool.ru/edu-process/courses/machine-learning#item-1 (дата обращения: 09.04.2019).
7. Вьюгин В. В. Математические основы теории машинного обучения и прогнозирования [Текст] / В. В. Вьюгин. — Москва: МЦНМО 2013. — 305 с.
8. Деревья решений — общие принципы работы [Электронный ресурс]. - Режим доступа: https://basegroup.ru/community/articles /description (дата обращения: 10.04.2019).
9. Загинайлов Ю. Н. Теория информационной безопасности и методо- логия защиты информации [Текст] : курс лекций / Ю. Н. Загинайлов. — Бар- наул: АлтГТУ им. И. И. Ползунова, 2010. -104 с.
10. Знакомство с машинным обучением [Электронный ресурс]. — Ре- жим доступа: https://www.google.ru/about/main/machine-learning-qa/ (дата об- ращения: 08.04.2019).
11. Искусство анализа данных: взгляд изнутри [Электронный ре-сурс]. — Режим доступа: https://www.osp.ru/cio/2018/02/13054071/ (дата об- ращения: 08.04.2019).
12. курс «Введение в машинное обучение» [Электронный ресурс].—Режим доступа: https://www.coursera.org/learn/vvedenie-mashinnoe-obuchenie /home/welcome (дата обращения: 19.05.2019).
13. Курс «Нейронные сети и машинное зрение» [Электронный ре-

сурс]. — Режим доступа: https://stepik.org/course/50352/promo (дата обраще- ния: 19.05.2019).

14. Логистическая регрессия и ROC-анализ — математический аппарат [Электронный ресурс]. — Режим доступа: https://basegroup.ru/commu- nity/articles/logistic (дата обращения: 01.04.2019).

15. Машинное обучение в Offensive Security [Электронный ресурс]. — Режим доступа: https://habr.com/ru/company/pm/blog/419617/ (дата обращения: 14.05.2019).

16. Машинное обучение — это легко [Электронный ресурс]. — Режим доступа: https://habr.com/post/319288/ (дата обращения: 08.03.2019).

17. Машинное обучение [Электронный ресурс]. — Режим доступа: http://www.machinelearning.ru/wiki/index.php?title=Машинное_обучени е (дата обращения: 24.03.2019).

18. Машинное обучение для чайников [Электронный ресурс]. — Режим доступа: https://newtonew.com/tech/machine-learning-novice (дата обращения: 20.03.2019).

19. Машинное обучение и анализ данных [Электронный ресурс]. — Режим доступа: http://www.uic.unn.ru/~zny/ml/ (дата обращения: 04.04.2019).

20. Машинное обучение и анализ данных [Электронный ресурс]. — Режим доступа: https://elibrary.ru/title_about.asp?id=32828 (дата обращения: 09.03.2019).

21. Мерков А. Б. Распознавание образов. Построение и обучение веро- ятностных моделей [Текст] / А. Б. Мерков. — Москва: Ленанд, 2014. — 240 с.

22. Методы построения деревьев решений в задачах классификации в Data Mining [Электронный ресурс]. — Режим доступа: https://ami.nstu.ru/ ~vms/lecture/data_mining/trees.htm (дата обращения: 03.04.2019).

23. Мюллер А. Введение в машинное обучение с помощью Python. Ру- ководство для специалистов по работе с данными [Текст] / А. Мюллер, С. Гвидо. — пер. Груздев А. — Москва: Альфа-книга, 2017. — 480 с.

24. Основные принципы подготовки презентаций [Электронный ресурс]. — Режим доступа: https://studme.org /50391/menedzhment /osno-vnye _printsipy_podgotovki_prezentatsiy (дата обращения: 22.02.2018).

25. Простыми словами: как работает машинное обучение [Электронный ресурс]. — Режим доступа: https://www.kaspersky.ru /blog/machine-learning- explained/13605/ (дата обращения: 14.03.2019).

26. Советы по изучению машинного обучения [Электронный ресурс]. — Режим доступа: https://www.youtube.com/watch?v=IiDmp-O2Yok (да- та обращения: 12.03.2019).

27. Тархов Д. А. Нейросетевые модели и алгоритмы [Текст] / Д. А. Тархов. — Москва: Радиотехника, 2014. — 352 с.

28. Флах П. Машинное обучение. Наука и искусство построения алго- ритмов, которые извлекают знания из данных [Текст] / П. Флах. — пер. с англ. Слинкина А. А. — Москва: ДМКПресс, 2015. — 400 с.

СОДЕРЖАНИЕ

Редакторы: Ахметжанова Г.М.
Сидикова К.А.