МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ РОССИЙСКОЙ ФЕДЕРАЦИИ
ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ ОБРАЗОВАТЕЛЬНОЕ
УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ
«РОССИЙСКИЙ ГОСУДАРСТВЕННЫЙ ПЕДАГОГИЧЕСКИЙ УНИВЕРСИТЕТ
им. А. И. ГЕРЦЕНА»

44.04.01 – ПЕДАГОГИЧЕСКОЕ ОБРАЗОВАНИЕ

МАГИСТЕРСКАЯ ПРОГРАММА: «ТЕХНОЛОГИИ И МЕНЕДЖМЕНТ ЭЛЕКТРОННОГО ОБУЧЕНИЯ»

Выпускная квалификационная работа

Методика подготовки будущих инженеров к образовательному дата-майнингу

Обучающегося 2 курса
очной формы обучения
Жукова Николая Николаевича

Научный руководитель:
к.пед.н., доцент
Государев Илья Борисович

Рецензент:
к.пед.н., доцент,
Авксентьева Елена Юрьевна

Санкт-Петербург
2016

ОГЛАВЛЕНИЕ

ВВЕДЕНИЕ

Современный этап развития информатики и информационных систем, а также программных платформ и языков программирования предполагает качественную подготовку инженеров различных направлений, в том числе и направления 09.03.01 «Информатика и вычислительная техника». Подготовка по данной специальности позволяет сформировать у выпускника множество профессиональных компетенций [3].

Требования стандарта уточняются и реализуются в виде образовательных программ, включая конкретные дисциплины, такие, например, как «Программирование», преподаваемые на базе кафедры компьютерных технологий и электронного обучения института компьютерных наук и технологического образования (ИКНиТО) РГПУ им. А.И. Герцена.

Для такой дисциплины, как «Программирование», следующие профессиональные компетенции (ПК) характерные для проектно-конструкторской деятельности являются наиболее актуальными на сегодняшний день:

- освоение методик использования программных средств для решения практических задач (ПК-2);
- разработка моделей компонент информационных систем, включая модели баз данных (ПК-4);
- разработка компонент программных комплексов и баз данных, использование современных инструментальных средств и технологий программирования (ПК-5).

Следует принять во внимание утверждение стратегии развития информационных технологий в Российской Федерации на 2014-2020 годы и на перспективу до 2025 года, принятую Правительством РФ от 1 ноября 2013 г.

2036-р [1]. В постановлении уточняются, что ключевыми направлениями в этот период, среди прочих, становятся обработка больших данных и машинное обучение, а в прикладных исследованиях приоритетными считаются проектирование и разработка систем поиска и распознавания данных, извлечение информации, анализ больших массивов данных и извлечение знаний и новые методы и программное обеспечение для предсказательного моделирования сложных инженерных решений.

Аналогичные прогнозы были опубликованы Консорциумом Новых Медиа по заказу Европейской комиссии в докладе 2014 года о перспективах внедрения новых образовательных технологий в школы и в докладе 2016 года — в университеты, а также Открытым университетом Великобритании в докладе 2015 года Innovating Pedagogy 2015.

Выполнение требований стандарта и формирования у студентов требуемых компетенций, подготовки к решению профессиональных задач и реализации стратегии развития информационных технологий в РФ возможно, при условии корректировки содержания их квалификационной подготовки в сторону, обозначенную в постановлении Правительства РФ.

Актуальность темы обосновывается противоречием между потребностью в электронных образовательных ресурсах (ЭОР), электронной информационно-образовательной среде (ЭИОС) и методике подготовки специалистов по ключевым направлениям, обозначенным выше (машинное обучение, анализ больших массивов данных), и их фактическим отсутствием.

Таким образом, предметом исследования является методика подготовки студентов к извлечению и анализу данных, а целью — разработка электронного образовательного ресурса для подготовки студентов к извлечению и анализу данных.

Для достижения поставленной цели необходимо решить следующие задачи:

1. Проанализировать существующие ЭОР, ЭИОС, посвященные тематике извлечения и анализа данных, а также машинного обучения.
2. Выявить требования к ЭОР и ЭИОС для подготовки к дата-майнингу на основе проведенного анализа.
3. Спроектировать модель ЭУМК.
4. Реализовать разработанную модель с использованием оптимальных технических средств.
5. Провести апробацию на площадке кафедры компьютерных технологий и электронного обучения (КТиЭО).

Новизна работы заключается в создании нового подхода к подготовке будущих программистов и формировании необходимых компетенций для решения задач ДМ.

Практическая значимость определяется тем, что:

- разработан ЭУМК в виде электронного образовательного ресурса, готовый к использованию на базе любой профильной кафедры;
- разработаны набор практических, лабораторных работ, тестовых заданий, а также тем для выпускных квалификационных работ и т.д.

Апробация результатов работы осуществлялась на площадке кафедры компьютерных технологий и электронного обучения института компьютерных наук и технологического образования РГПУ им. А.И. Герцена в рамках курсов «Программирование», «E-learning-решения управления знаниями в образовательных учреждениях».

Результатом магистерской диссертации является готовый к использованию ЭОР, позволяющий осуществлять подготовку бакалавров по дисциплине «Программирование».

Структура магистерской диссертации. Магистерская диссертация состоит из введения, двух глав, заключения, библиографии и приложений. Она содержит 65 страниц, 11 рисунков и 2 таблицы.

ГЛАВА 1. АНАЛИЗ НАУЧНЫХ ПУБЛИКАЦИЙ В ОБЛАСТИ ИЗВЛЕЧЕНИЯ ДАННЫХ

1.1. Подходы к трактовке термина «data mining»

Термины «извлечение данных» (data mining) и «обнаружение знаний в базах данных» (knowledge discovery in databases, KDD) впервые появились в англоязычных научных публикациях [10]. На сегодняшний день, их, как правило, считают синонимами как в англоязычном, так и в русскоязычном научном сообществах. По словам одного из основоположников этого научного направления — Григория Пятецкого-Шапиро, «обнаружение знаний в базах данных» является более точным. В последние годы появилось существенное количество близких по значению к придуманному Пятецким-Шапиро определению или являющихся синонимами. Приведем наиболее часто упоминаемые термины:

1) анализ данных (data analysis);
2) добыча данных;
3) просев информации;
4) интеллектуальный анализ данных (intellectual data analysis);
5) глубинный анализ данных;
6) машинное обучение (machine learning).

Причем, пункты 2, 3, 4 являются вариациями перевода англоязычного понятия “data mining”.

В настоящее время не сформировалось окончательного взгляда какой термин является более точным для описания этой области. Существуют несколько подходов для описания взаимосвязи «извлечения данных» и «обнаружения знаний в базах данных».

В книге «Анализ данных. Часть 1» под авторством Федина Ф. О. термин “извлечение данных” (data mining, DM) трактуется как один из этапов извлечения данных из баз данных (knowledge discovery in databases, KDD). Данный этап включает в себя построение решающей модели и осуществляется практически в конце всего процесса после этапов: (1) выборки, (2) очистки, (3) трансформации данных и перед этапом (5) интерпретации [22].

Другой подход [31] заключается в том, чтобы считать извлечение данных прикладной деятельностью, выполняемой исследователем, которому требуется решить конкретную задачу при наличии соответствующего набора данных, а машинное обучение — частью сферы искусственного интеллекта. Классическое и формальное определение термина «машинное обучение» определяет его как компьютерную программу, обучаемую на некоторых опытных данных E, относительно некоторого класса задач T с измерением производительности P, если производительность на задачах T, измеренная с помощью P, улучшается с опытными данными E [30].

Поскольку анализ взаимосвязей различных терминов в этой области не являлся темой диссертации, было принято решение использовать в диссертации первоначальный термин для обозначения этой области науки — «извлечение данных» (data mining).

Однако, с учетом многообразия, продемонстрированного выше, на этапе анализа публикационной активности исследовались всевозможные трактовки термина «извлечение данных» («обнаружение знаний в базах данных») в различном контексте. Здесь и далее автор, употребляя термин «извлечение данных» будет иметь в виду все возможные вариации (data mining, data analysis, knowledge discovery in databases etc).

1.2. Анализ публикационной активности в области извлечения и анализа данных

В результате работы на этом этапе был произведен отбор и систематизация информации, посвященной извлечению и анализу данных. Было проанализировано множество источников, всё множество которых можно представить с помощью трех направлений:

1. Публикации, подтверждающие актуальность анализа данных и его подобластей. Например, текст-майнинга («text mining») — извлечение данных из текстов [17];
2. Публикации, относящиеся к области интеллектуального анализа данных образовательного процесса [26, 5, 13]. К этой категории относится разработка информационных систем, решающих различные задачи в области дистанционного и электронного обучения с использованием различных алгоритмов и моделей извлечения и анализа данных.
3. Публикации, раскрывающие различные аспекты, связанные с обучением или подготовкой к извлечению данных.

Необходимо отметить значительный объем публикаций, касающихся анализа и извлечения данных. Например, только по ключевому запросу “data mining” (дата обращения: 25.04.2016) в российской научной электронной библиотеке, интегрированной с Российским индексом научного цитирования (РИНЦ) «Elibrary», нашлось 15247 источников.

Ниже представлены краткие выводы по каждому из выделенных направлений.

1.2.1. Актуальность анализа данных

В работе Пиотровской выделяется значимость текст-майнинга (интеллектуального анализа текстов) — области извлечения данных, целью которой является получение информации из коллекций текстовых документов с

использованием различных алгоритмов машинного обучения и обработки естественного языка.

Автор подчеркивает, что, используя программные продукты для статистической обработки текстовых данных, реализуются: (1) классификация, (2) кластеризация и (3) предварительная обработка данных и другие [17]. Среди ПО для текст-майнинга с помощью сравнительной характеристики выделяются такие среды как: Rapid miner, Weka, R, представляющие собой открытое ПО, активно использующееся и применяющееся крупными компаниями, поддерживаемое сообществом и расширяемое с помощью дополнительных плагинов.

Область извлечения данных тесно связана с другими науками (например, математической статистикой, дискретной математикой, теорией множеств и т.д.). В ходе анкетирования соискателей и выпускников, описанного в статье Константиновской Н. В., Мухаматзановой, подтверждается актуальность внедрения методов математической статистики в процесс обучения студентов. Лишь около 5% респондентов отмечают, что не имеют никаких затруднений в ходе выполнения исследовательской работы, а основные трудности у целевой аудитории возникают с построением математических моделей и выбором метода статистической обработки данных [12].

Говоря про практические навыки, можно отметить, что использование программного обеспечения, позволяющего проводить анализ данных студентами педагогического направления, повышает общий уровень профессиональной подготовки и мотивации проведения педагогических экспериментов в учебно-исследовательской и научно-исследовательской работе в вузе и в будущей профессиональной деятельности [13].

1.2.2. Интеллектуальный анализ данных для образования

С активным развитием электронного образования в России [2] и мире стало возможным извлечение и анализ данных для исследования образовательного процесса.

В научной публикации Анохина П. В. подчеркивается, что в рамках дистанционного обучения генерируются значительные объемы данных, большая часть которых, по словам автора, плохо или вообще не структурирована, большинство полезных знаний — не является тривиальными [5]. Для работы в такой предметной области и используется анализ данных образовательного процесса (EDM - educational data mining) [26]. EDM представляет собой область науки, связанную с (1) разработкой методов для изучения данных, поступающих из образовательной сферы и (2) использования множества различных уже существующих алгоритмов анализа данных для лучшего понимания студентов, оценки условий, в которых они учатся.

Основные методы, применяемые в EDM учитывают специфику данных. К наиболее часто использующимся типам задач, решаемым в этой области автор относит следующие: (1) кластеризация, (2) прогнозирование, (3) выявление зависимостей, (4) использование нейронных сетей для анализа потока изменяющихся данных.

В статье отмечается, что для удобства и эффективности обработки данных могут использоваться системы интеллектуального анализа такие как RapidMiner, Weka [5]. Эти среды предоставляют широкие возможности как для проведения экспериментов для решения задач, так и для визуализации, анализа полученных результатов.

Некоторые авторы для анализа данных контроля знаний, накапливаемых в процессе обучения, предлагают использовать архитектуру системы обработки знаний на основе онтологий.

1.2.3. Обучение или подготовка к извлечению данных

Подготовка специалистов к дату майнингу может быть организована в рамках различных дисциплин, направлений и уровней высшего образования.

Задача подготовки магистров в рамках дисциплины «Модели и методы интеллектуального анализа данных», проводимой на базе кафедры автоматизированных систем и обработки информации и управления при

Казанском национальном исследовательском техническом университете (КНИТУ-КАИ) по направлению «Информационные системы и технологии», решается с использованием аналитической платформы Deductor [18].

В статье рассматривается применение этой аналитической платформы для решения задач анализа данных, демонстрируется, что для успешной деятельности в этом направлении необходимо использовать методы OLAP (Online Transaction Processing) и дата майнинг, а также на примере решения задачи классификации показывается применение аналитической платформы.

Одной из целей изучения дисциплины является освоение и применение различных моделей и методов интеллектуального анализа данных для принятия решений в сфере информационных технологий на базе средств дата майнинга.

В процессе решения задачи классификации наиболее оптимально использование модели деревьев принятия решений (decision trees) как наиболее простого и удобного инструмента. Аналитическая платформа Deductor позволяет использовать большие возможности этого ПО по настройке процесса построения дерева принятия решений. В данном программном продукте реализовано множество вспомогательных процедур, позволяющих решить данную задачу эффективнее: (1) отсечение несущественных факторов (pruning), (2) ранжирование факторов по степени влияния на результат; (3) описание способа классификации с помощью формальных правил; (4) определение достоверности и поддержке того или иного правила.

При описании используемой аналитической платформы Deductor, следует отметить следующие факты:

- не является кроссплатформенным и может быть запущен только на компьютерах под управлением операционной системы Windows (Windows XP, Vista, 7, 8);
- для использования в образовательных целях доступна специальная версия платформы, включающая множество алгоритмов анализа и

визуализации данных (стоимость коммерческой версии для персонального применения на момент 18 мая 2016 года составляет от 35 000 рублей.).

В статье Васильевой Т. В. даются методологические рекомендации, с учетом которых могут быть организованы курсы “Методы и представления и анализа данных” и “Методы прогнозирования”. Первый из этих двух курсов, по словам автора, должен включать в себя анализ входных данных - наблюдений и процесс их преобразования и визуализации. По результатам работы в рамках этих курсов, студенты должны осознавать существование следующих задач: (1) описание исходной ситуации, (2) обнаружение (усиление) закономерностей (предположений и гипотез) и (3) использования этих закономерностей [7].

Авторы статей, посвященных подготовке студентов в области ДМ, уверены в том, что эту деятельность следует осуществлять в рамках “computer science”-курсов [29]. Это могут быть, например, курсы программирования для веба, разработки программных продуктов и т. д. Данный подход широко используется множеством университетов, но в свою очередь порождает и ряд сложностей одной из которых является то, что студентам, для успешного завершения такого курса, требуются базовые знания и навыки в множестве различных областей, которых у них может не быть.

Согласно [29] все возможные варианты организации курсов подготовки в области извлечения данных можно условно разделить на 4 направления или подхода:

- алгоритмический (математический);
- подход, основанный на использовании учебников;
- предметно-ориентированный подход;
- прикладной подход.

Кратко опишем каждый из них.

При использовании алгоритмического (математического) подхода упор делается на отдельные области ДМ (например, алгоритмы и модели высшей математики, статистику, дифференцирование, интегрирование, теорию вероятностей и дискретную математику) в приложении к использованию различных специальных методов извлечения данных таких как: кластеризация, линейная регрессия, классификация (в том числе деревья принятия решений) и т.д.

Каждый такой алгоритм описывается с помощью набора математических формул и рассматривается с теоретической точки зрения. Основной упор в таких курсах делается именно на теорию и меньше — на практическое применение таких алгоритмов. В качестве вспомогательных материалов для таких курсов как правило не используются специальные учебные пособия, предпочтение отдается использованию научных статей. В качестве инструментов реализации алгоритмов ДМ на практике в рамках таких курсов часто используется язык программирования R, а также программный пакет для анализа Weka (Университет Вермонта).

Несколько отличается, от описанного выше, подход (2), при котором преподавание курса полностью строится на использовании специального учебника, упор в котором делается последовательное описание алгоритмов. Предполагается, что преподаватель четко следует структуре, предложенной в нём. В качестве методов извлечения данных аналогично предыдущему способу выбраны кластеризация, линейная регрессия и классификация (например, деревья принятия решений), идущие в том же порядке, что и в тексте учебника.

При предметно-ориентированном подходе (3) для преподавания не используется какой-либо специальный учебник и не делается упор на изложении математической основы моделей, используемых в алгоритмах извлечения данных. Напротив, здесь преподаватель сочетает исследование различных тем и алгоритмов в области кластеризации, линейной регрессии и классификации с применением их в реальных ситуациях. Вместо использование единственного

учебника, предполагается задействование множества научных статей, материалов с профильных веб-сайтов и фрагментов книг, в которых описывается теоретическая база, необходимая для полного понимания конкретной темы или алгоритма.

Используя прикладной подход (4), преподаватель комбинирует теорию с активным применением алгоритмов на практике. Этот подход близок к предметно-ориентированному, однако, в отличие от него не предполагает использование каких-либо учебников и опирается на презентации и раздаточные материалы, подготовленные преподавателем. Основное отличие от подходов, описанных выше, в том, что алгоритмы и модели извлечения данных исследуются в контексте решения актуальных задач в этой области (например, решаются актуальные задачи реальных предприятий). Теоретический материал преподается в объеме, необходимом для понимания области применения конкретного алгоритма.

Авторы статьи [29] не выделяют наиболее оптимальный метод среди описанных в статье, однако, предлагают гибридную модель курса, содержащую в себе элементы прикладного и предметно-ориентированного подхода. В качестве прикладного инструмента для исследования алгоритмов извлечения данных предлагается использовать платформу анализа данных Weka (Университет Вермонта) [30].

В заключение преподавателю предлагается учесть следующие рекомендации:

- студентам может потребоваться дополнительное время для изучения некоторых базовых областей науки (таких как статистика или теория вероятностей) и это следует учесть при проектировании курса;
- преподавателю следует изначально запланировать определенный объем материала, который может быть пропущен по мере необходимости, если студентам требуется дополнительное время получения базовых знаний и навыков.

- если в рамках курса запланирована работа над индивидуальным проектом, авторы статьи рекомендуют, чтобы проект был групповой.

Последняя рекомендация позволит студентам обучать друг друга (внутри группы) и находить нестандартные решения для поставленных преподавателем задач.

Гибридная модель курса, предлагаемая в статье выше, используется в рамках курса «Data mining», проводящегося для студентов направления «Information Systems» и описанного (Sakha'a Al Manaseer) в статье Improve Teaching Method of Data Mining Course [28]. Автор выделяет следующие этапы процесса обнаружения знаний в данных (knowledge discovery process):

- предобработка данных (data preprocessing), включающую нормализацию данных, выбор наиболее репрезентативных характеристик, уменьшение размерности и т.д;
- извлечение данных (data mining) включает обнаружение закономерностей, применение алгоритмов классификации, кластеризации, выявление зависимостей и анализ "выбросов" данных;
- постобработка данных (оценка работы, интерпретация модели, визуализация результатов).

В начале курса, автор знакомит студентов с основными этапами анализа данных и прикладными областями, где конкретные алгоритмы могут быть применены, кратко описывается каждый из этапов; далее студенты делятся на группы, выбирая прикладную область, в которой будут они будут работать и конечную цель анализа данных. Далее идет обсуждение сделанного выбора, целей исследования, задач, которые необходимо решить в процессе.

После этого для каждого этапа в процессе обнаружения знаний в данных группы решают подзадачи, входящие в эти этапы. Например, предобработка данных может включать заполнение пропущенных значений, выбраковку "выбросов" данных и т.д.

После самостоятельной работы студентов преподаватель обсуждает с представителями каждой группы результаты их деятельности.

Этот процесс повторяется для каждого этапа.

В конце подготовки по этому курсу помимо представления полученных в результате анализа данных результатов, проводится письменный экзамен.

В ходе педагогического эксперимента, описанного автором в статье, было обнаружено, что для повышения эффективности преподавания данного курса студенты на каждом этапе процесса обнаружения знаний могли выбирать новую предметную область. Таким образом, обучающиеся овладевали большим набором знаний (теоретических и практических), умений и навыков решения задач из различных областей.

Как уже было сказано выше, некоторые исследователи и ученые определяют область извлечения данных как один из этапов интеллектуального анализа данных. К инструментам, использующимся в этой области помимо алгоритмов кластеризации, классификации и других, относятся нейронные сети (neural networks). В статье [8] отмечается, что обучение студентов использованию инструмента нейронных сетей позволит решать множество трудноформализуемых и неформализуемых задач. К таким задачам относится, например, анализ финансовой и банковской деятельности биржевых, фондовых и валютных рынков, связанных с высокими рисками моделей поведения клиентов. Точность решения реальных задач в этой области достигает 95%. Автор, далее, отмечает основные моменты, на которые важно обратить внимание при обучении инструментарию искусственных сетей. Кратко обозначим их ниже:

- модель искусственного нейрона можно описать с помощью простых функций (например, модель нейрона Маккалока-Питтса возможно представить средствами табличного процессора);

- решение реальной задачи возможно с помощью искусственной нейронной сети, состоящей из нескольких слоёв (содержащей не более десятка нейронов);
- задачи, предлагаемые к решению, средствами НС должны быть тщательно отобраны таким образом, чтобы инструмент позволял построить эффективное решение;
- в наборе задач должны быть и такие, которые нельзя решить с помощью НС, но возможно решить с помощью других алгоритмов извлечения данных (это позволит сравнить несколько алгоритмов, выбрать наиболее оптимальный и.д.), что в конечном итоге поможет сформировать у студентов верное представление о возможностях и ограничениях при решении задач с помощью этого инструмента;
- НС чувствительных к входным данных, а значит необходимо обратить внимание на этап предобработки данных (очистки от выбросов), выработка навыков отбора, оценки и подготовки их к анализу должна стать неотъемлемой часть обучения;
- подбор задач должен быть выполнен таким образом, чтобы раскрыть максимально предметные области, в которых могут быть использованы НС и не ограничиваться областью экономики.

В другой работе этого же автора внимание уделяется технологиям кластеризации данных для подготовки студентов [9]. Автор обращает внимание, что множество задач анализа данных эффективно можно решить именно с использованием алгоритмов кластеризации. Для успешного их применения необходимо сформировать у студентов четкое представление о сути понятия «кластер», которое лежит в основе различных моделей концепции интеллектуального анализа данных. Здесь может возникнуть путаница из-за разницы между понятием «кластер» в различных областях науки. Например, в случае подготовки студентов экономического направления, может возникнуть ситуация, при которой обучающиеся этой категории могут соотносить понятие

«экономического кластера» с результатами работы программного средства, анализирующего экономические данные.

В процессе обучения для решения практических задач используются множество различных инструментов. Одним из наиболее используемых решений является язык программирования R. В статье [20] для решения задач в области извлечения данных и касающихся таких областей как теория вероятностей, математическая статистика и компьютерное моделирование в различных предметных областях предпочтение отдается именно этому языку.

Приведем ниже основные преимущества такого подхода, отмеченные автором статьи:

- свободно распространяемый язык для всего спектра операционных систем и платформ;
- возможность расширить стандартный набор возможностей с помощью большого количества пакетов расширений (охватывающих множество предметных областей)
- большое количество вспомогательных материалов, созданных в помощь изучающему этот язык.

Использование R как основного инструмента в рамках курса «Прикладная статистика», организованного в институте информационных наук и технологий безопасности РГГУ, предполагает теоретические, лабораторные работы, индивидуальные компьютерные тесты для контроля знаний студентов. Обучение студентов проводится в рамках смешанной модели.

Для успешной подготовки в области ДМ необходимо тщательно продумать и организовать среду, в которой будет происходить взаимодействие студентов и преподавателя. При проведении такого курса может быть использован специально разработанный электронный образовательный ресурс (ЭОР). В статье Есаулова [11] дается описание электронного учебного пособия «Эволюционные алгоритмы

интеллектуального анализа данных» по дисциплине “Системы искусственного интеллекта”. Для данного курса автор предлагает следующую структуру:

- постановка задачи поисковой оптимизации;
- эволюционные алгоритмы;
- алгоритмы роя.

Автор обращает внимание читателя на следующие особенности:

- основной материал должен быть оформлен в виде структурированной контекстно-ориентированной гипертекстовой среды;
- содержание должно иметь прямые ссылки на все разделы учебника;
- разрабатываемые элементы дидактических единиц требуется делать интерактивными;
- контекстно-связанные ссылки, имеющиеся в тексте должны указывать на все разделы учебного пособия;
- должна существовать возможность продемонстрировать лабораторные работы, предложенные студентам для выполнения;
- должна быть возможность с помощью встроенной тестовой системы организовать контроль усвояемости знаний.

Для реализации такого продукта автором предлагается использовать технологии гипертекстовой разметки HTML, язык каскадных стилей CSS и язык программирования JavaScript.

В качестве более привычного и распространенного инструмента для решения задач ДМ возможно использование ПП с установленными расширениями (например, с помощью продуктов компании Microsoft). Одно из таких решений описывается в статье “A Tools-based Approach to Teaching Data Mining Methods” [27]. В качестве инструментов обучения автором используется:

- программный продукт Microsoft Excel с установленными расширениями, позволяющими решать определенные задачи в области дата майнинга — на стороне клиента;
- программное обеспечение Microsoft Cloud Computing и SQL Server 2008 — на стороне сервера.

Описанный в статье подход позволяет сделать вывод, о том, что набор интегрированных друг с другом инструментов, указанных выше, позволяет сфокусировать внимание студентов в процессе их подготовки на аналитических аспектах извлечения данных и использовании алгоритмов для решения практических задач. Для подготовки активно используются практические демонстрации механизма работы алгоритмов, домашние задания и проектная деятельность. Обучающиеся, в результате прохождения курса, получают целостное представление о процессах извлечения и анализа данных, о примерах применения конкретных алгоритмов для поддержки принятия решений.

Таким образом, студент выступает в роли, скорее, аналитика, а не разработчика программного обеспечения, решающего задачи извлечения и анализа данных. В рамках такого курса предполагается изучить:

- основы анализа данных;
- построение, конфигурирование и проведение оценку работы различных вычислительных моделей извлечения данных, их тестирование и сравнение;
- способы использования моделей извлечения данных для анализа и предсказания для принятия решений.

Подобное построение курса позволяет уместить его в один семестр, в течение которого обеспечить студентов необходимым объемом теоретических и практических знаний, умений и навыков.

Структура такого курса может быть представлена тремя основными компонентами:

1) аналитическая компонента;
2) инструменто-ориентированная прикладная компонента;
3) множество различных наборов данных.

В аналитической компоненте раскрывается теоретические и практические аспекты жизненного цикла проекта, связанного с извлечением и анализом данных, демонстрируются основные алгоритмы классификации, кластеризации, предсказания (деревья принятия решений, нейронные сети, логистическая регрессия и т.д.), объясняется как протестировать и проверить работу построенной модели анализа данных и, наконец, демонстрируется как применить модели для решения задач принятия решений и организации предсказаний.

В рамках инструменто-ориентированной (прикладной) компоненты демонстрируется как использовать инструменты для использования алгоритмов, описанных на предыдущем этапе. Автор статьи уверен, что ПП Microsoft Excel (с использованием специальных расширений) - оптимальное средство для организации несложного анализа и визуализации данных. Эти расширения сконфигурированы таким образом, чтобы отправить данные для анализа на сервер (SQL Server 2008) с платформой Microsoft Cloud Computing. В свою очередь на сервере выполняются необходимые операции по анализу данных и результаты отправляются обратно на клиент в Excel. Расширения предоставляют режим мастера и графический интерфейс, позволяющий конфигурировать выполнение процессов, управлять данными, и визуализировать результаты.

Множество различных наборов данных позволяют продемонстрировать возможности алгоритмов и обеспечить получение прикладного опыта в исследовании.

Далее, в статье приводится описание нескольких тем, которые автор статьи включил в курс по машинному обучению: (1) основы анализа данных, (2) задача анализа ассоциативных правил, (3) классификация и предсказание, (4) кластеризация.

В заключение, отмечается, что используя описанную выше структуру и предлагаемый инструментарий, курс может быть проведен в течение одного семестра, обеспечивая студентов необходимым объемом знаний и навыками их применения в области извлечения данных без необходимости программировать алгоритмы вручную.

Другой подход к практической подготовке студентов обозначен в статье [25]. Он предполагает использование проприетарного платного продукта SAS Enterprise Miner, а само обучение было построено вокруг руководства по использованию этого программного продукта. В рамках практических занятий, которые проводил автор статьи, обсуждалось использование множества алгоритмов для анализа данных, включая следующие:

- деревья принятия решений;
- искусственные нейронные сети;
- логистическая регрессия;
- кластеризация;
- поиск ближайших соседей.

В рамках одного занятия преподаватель сравнивал работу разных алгоритмов с одним и тем же набором данных. Это позволило показать их достоинства и недостатки, область применения, смоделировать ситуацию, когда исследователь последовательно осуществляет поиск наиболее оптимального решения.

Курс завершался групповой проектной работой по решению задачи предсказания поступления студентов.

1.3. Анализ существующих электронных образовательных ресурсов по подготовке к анализу данных

Помимо анализа публикационной активности, необходимо было также проанализировать существующие электронные образовательные ресурсы по

подготовке к извлечению и анализу данных. Специфика этой области требует анализа не только разработок, созданных отечественными исследователями, но и реализованных за рубежом и предназначенных для англоязычной аудитории.

Так как рассматриваемая область входит в число перспективных научных отраслей стран всего мира и учитывая тот факт, что следующее десятилетие будет характеризоваться процессами замещения аудиторных занятий различными формами онлайн-взаимодействия и появлением большого количества массовых открытых онлайн-курсов (англ. Massive Open Online Course, MOOC) [6], поиск осуществлялся среди наиболее популярных MOOC площадках.

MOOC разрабатываются ведущими университетами всего мира, такими как Гарвард, Стенфорд, МТИ (Массачусетский технологический институт), Калифорнийский университет Беркли и так далее. Эти и многие другие учебные заведения инвестируют значительные средства в инновационные проекты по разработке учебных платформ таких как EDx, Udacity, Coursera и т. д.

В России наиболее популярными и востребованными сервисами получения образования онлайн являются: (1) Национальная платформа открытого образования (2) Интернет-университет информационных технологий «ИНТУИТ», (3) межвузовская площадка электронного образования «Универсариум», (4) Лекториум.

Образовательные программы, представленные на этих площадках разрабатываются крупнейшими рейтинговыми российскими вузами, среди которых СПБГУ, ИТМО, МГУ, ВГИК, МГТУ им. Баумана, МАИ, РГГУ и др. [4]

РГПУ им. А. И. Герцена активно реализует поддержку образовательного процесса средствами дистанционных технологий: развернута собственная LMS площадка (анг. Learning Management System - система управления обучением) на базе системы управления курсами Moodle, преподавателями различных филиалов, факультетов и институтов вуза постоянно разрабатываются и дополняются множество дистанционных курсов.

На этапе анализа было принято решение проанализировать наиболее крупные и положительно зарекомендовавшие себя отечественные и зарубежные платформы МООС. На предмет наличия курсов соответствующей тематики осуществлялся мониторинг следующих наиболее популярных отечественных и зарубежных платформ МООС:

- Coursera (coursera.org)
- EdX (сноска)
- Lectorium (сноска)
- Udacity (сноска)

Кроме них анализировался Интернет-университет информационных технологий «ИНТУИТ» и курсы, размещенные там и, в последнюю очередь были просмотрены опубликованные в свободном доступе материалы по анализу данных (теоретической и практической направленности), разработанные крупнейшми вузами мира (например, Стэнфордским университетом).

Опишем кратко наиболее востребованные площадки для онлайн-обучения и курсы, посвященные теме исследования и размещенные на этих площадках.

Coursera

Платформа Coursera предлагает множество как отдельных курсов, так и специализаций (когда подготовка производится по нескольким смежным направлениям), посвященных анализу данных и опубликованных несколькими крупными иностранными и отечественными вузами. Приведем небольшой список наиболее популярных курсов и специализаций:

- курс «Машинное обучение» («Machine learning»), университет Стенфорда;
- специализация «Анализ текстов, обнаружение шаблонов в данных, визуализация данных» («Analyze Text, Discover Patterns, Visualize Data»), университет Иллинойса;

- специализация «Изучение фундаментальных основ науки о данных» («Learn Data Science Fundamentals»), Уэслианский университет;
- курс «Введение в машинное обучение» Высшая школа экономики (факультет компьютерных наук совместно с школой анализа данных компании Яндекс);
- курс «Машинное обучение на основе больших данных» («Machine Learning With Big Data»), Калифорнийский университет в Сан-Диего;
- специализация «Наука о данных» («Data science»), университет Джона Хопкинса;
- курс «Овладение анализом данных в Excel» («Mastering Data Analysis in Excel»), Дюкский университет;
- курс «Численные методы анализа данных» («Computational Methods for Data Analysis»), Вашингтонский университет;
- специализация «Машинное обучение» («Machine Learning»), Вашингтонский университет;
- курс «Нейронные сети для машинного обучения» («Neural Networks for Machine Learning»), университет Торонто.

Курсы, представленные на этой платформе, обладают некоторыми общими чертами, опишем их подробнее. Каждый курс разбит на определенное количество недель. Одна неделя, как правило, посвящена одной или нескольким смежным темам. После ознакомления с теорией по конкретной теме, представленной в виде набора видеолекций (или текста), студенту предлагается пройти компьютерный тест на овладение материалом. Тест может содержать вопросы как закрытого, так и открытого типа. Необходимо заметить, что часто внутри видеолекции встроены вопросы для самопроверки. Эти вопросы, однако, могут быть пропущены, а ответ на них может быть подобран методом перебора всех возможных вариантов. Как

правило сразу после теста в конце темы располагается практическое задание (“Assignment”), которое в большинстве случаев предполагает решение практической задачи или нескольких задач, ответ которой необходимо предоставить через специальную форму (ответ может загружаться внутрь курса в виде текстового файла). Практическим заданием может быть и эссе. В этом случае, проверка проводится по системе перекрестного оценивания («peer review»), когда каждый участник курса оценивает одну или несколько работ других студентов, выставляя определенный балл. Выполненным считается задание, которое получило средний балл больше или равный проходному баллу.

Помимо выполнения практических заданий и сдачи тестов по материалом каждой темы, курсы могут предполагать выполнение курсового проекта (“capstone project”) как в группе, так и самостоятельно. Задание, использующееся в проекте, учитывает весь объем теоретического материала, предоставленного студенту в рамках курса, и информацию, найденную им самостоятельно.

Обратная связь с преподавателями и кураторами курса организована с использованием форумов “прикрепленных” к видеозаписям, темам, практическим заданиям и т. д. Это позволяет, например, задавать вопросы непосредственно по тому материалу, который был непонятен.

Проанализируем 2 курса, размещенных на площадке Coursera из списка, приведенного выше.

«Машинное обучение» «Machine learning»

Курс «Машинное обучение», предлагаемый профессором Andrew Ng (Стенфорд), являющимся также основателем этой платформы, состоит из 11 разделов (недель). Раздел, в свою очередь, состоит из одной или нескольких тем, каждая тема раскрывает какую-то область машинного обучения (например, алгоритм, модель или гипотезу).

Важной особенностью данного курса является автоматизированная система для проверки правильности выполнения заданий. Она реализована в виде отдельного модуля (предоставляемого разработчиками курса), написанного с

использованием системы математических вычислений, в которой задействован совместимый с MATLAB язык высокого уровня и не требует проверки правильности выполнения преподавателем вручную задания каждого студента. Обучающийся запускает на выполнение данный модуль и в интерактивном режиме самостоятельно осуществляет проверку реализованного им решения. Проверяющий модуль, сравнивая ответы, которые дал пользователь с эталонными, предоставляет комментарии по загруженному решению. Для того чтобы решение было зачтено, перед отправкой решения обучающийся должен пройти процесс аутентификации, т.е. ввести электронную почту и специальный проверочный код, предоставляемый платформой Coursera. Выполнение практических заданий предусматривает создание программ на платформе GNU/Octave. Не смотря на то, что в курсе есть отдельная тема, посвященная работе в среде GNU/Octave, обучающемуся потребуются базовые навыки программирования.

Ниже приведем структуру содержания этого курса с кратким описанием каждой темы по неделям:

Неделя 1

1. Введение (здесь даются определения основным терминам в области машинного обучения, рассказывается о классификации алгоритмов).
2. Рассматривается один из алгоритмов обучения с учителем - линейная регрессия с одной переменной (для решения выбрана задача предсказания стоимости жилья).
3. Элементы линейной алгебры (даются основы необходимые и достаточные для понимания тех алгоритмов, которые планируется изучить).

Неделя 2

1. Алгоритм линейной регрессии с использованием нескольких переменных (обобщение и усложнение задачи из предыдущего раздела).
2. Введение в среду GNU/Octave (описываются основные понятия этой среды, языковые конструкции: циклы, условные операторы и т.д., организация вычислений наиболее оптимальным способом, создание функций, работа с данными и т.д.).

Неделя 3

1. Алгоритм логистической регрессии (описывается один из самых простых алгоритмом классификации данных).
2. Регуляризация (методика, при которой к условию задачи добавляются дополнительные требования, позволяющие предотвратить переобучение - ситуацию, когда задача может быть решена только с уже имеющимися данными)

Неделя 4. Введение в описание нейронных сетей (область применения, особенности, устройство).

Неделя 5. Обучение нейронных сетей (рассматривается алгоритм обратного распространения ошибки «backpropagation»).

Неделя 6. Оценка качества работы алгоритмов машинного обучения. Лучшие практики при разработке решения задач средствами алгоритмов машинного обучения. Отладка работы. Дизайн систем машинного обучения.

Неделя 7. Алгоритм SVM (“Support vector machine”). Теория и практика применение алгоритма.

Неделя 8. Алгоритмы, работающие по принципу “обучение без учителя”. Снижение размерности в машинном обучении:

- алгоритм k-means;
- метод главных компонент.

Неделя 9. Обнаружение аномалий. Рекомендательные системы. Рассматриваются прикладные задачи (1) построение системы для обнаружения аномалий (подозрительные транзакции в банковской деятельности), (2) построение рекомендательной системы (с использованием метода совместной фильтрации).

Неделя 10. Машинное обучение с большими данными (“big data”). Рассматриваются эффективные алгоритмы обработки (например, “map-reduce” или параллельная обработка данных).

Неделя 11. Задача распознавания образов. Описывается задача распознавания образов и способы её решения.

Приведем пример практического задания, которое необходимо выполнить в первую неделю обучения по данному курсу.

Задача — построить алгоритм, который будет, используя тренировочный набор данных, строить модель для предсказания цены на жилье, основываясь на его площади, количестве комнат.

Данные описывают цены на недвижимость, их можно представить в виде таблицы, в которой первый столбец — площадь дома в квадратных футах, второй столбец — количество спален, третий столбец - обозначающих стоимость жилья (значения этого столбца для домов, которых в выборке нет, нам и предстоит предсказать).

Для решения пользователю даются два набора данных:

- тренировочный (который используется на этапе построения модели для предсказания);
- тестовый (для объектов из этого набора данных будет необходимо определить стоимость жилья).

В результате решения задачи будет сформирована модель, используя которую можно предсказать (с определенной точностью) стоимость жилья, параметры которого содержатся в файле с тестовыми данными.

После этого необходимо запустить проверочный модуль (описанный выше), который не только сравнит ответы, но и осуществит проверку того, что ответ не был подобран вручную (или найден методом перебора).

Курс «Введение в машинное обучение», представляемый Высшей школой экономики (факультет компьютерных наук совместно со школой анализа данных компании Яндекс) представлен набором из 7 разделов (недель). Внутри каждого раздела может находится одна или несколько тем.

В качестве инструмента для выполнения практических заданий используется современный язык программирования Python, а также большое число специальных библиотек для этого языка, позволяющих решать различного рода задачи (например, Pandas, NumPy и др.). Для изучающего этот курс также будут необходимы базовые навыки программирования. Теоретическая часть представляет собой набор видеолекций различной длины, для использования доступны текстовые варианты лекций, презентации с материалами, показываемыми в них. Внутри лекций встроены вопросы для самопроверки, в конце тем — содержатся тесты, где могут присутствовать вопросы открытого и закрытого типов.

Опишем краткую программу этого курса по неделям:

Неделя 1. Введение в машинное обучение. Примеры решаемых задач. Описание логических методов: решающие деревья и решающие леса.

Неделя 2. Метрические методы классификации. Линейные методы, стохастический градиент.

Неделя 3. Метод опорных векторов (SVM). Логистическая регрессия. Метрики качества классификации.

Неделя 4. Линейная регрессия. Понижение размерности, метод главных компонент.

Неделя 5. Композиции алгоритмов, алгоритм градиентного бустинга. Нейронные сети.

Неделя 6. Кластеризация и визуализация. Частичное обучение.

Неделя 7. Прикладные задачи анализа данных: постановки и методы решения.

Опишем пример практического задания, которое необходимо решить на первой неделе обучения.

Задача — для данного набора данных (список пассажиров парохода «Титаник») построить модель решающего дерева и с помощью него определить параметр в данных, оказавшийся наиболее важным при спасении.

Данные описывают пассажиров парохода «Титаник», их можно представить в виде таблицы со множество столбцов (например, имя пассажира, класс, пол, возраст и.т.д.).

Алгоритм выполнения практического задания будет следующим: (1) обучающийся пишет программный код, позволяющий решить задачу и находит ответ, (2) полученный ответ или несколько ответов (если задание представляет собой несколько связанных между собой вопросов) публикуется в специальную форму на странице задания курса (3) если даны правильные ответы, задание считается выполненным, в противном случае, дается комментарий с рекомендациями по исправлению ошибок.

Udacity

Одним из наиболее популярных курсов, который следует упомянуть говоря про машинное обучение, является курс “Введение в искусственный интеллект” (“Intro to Artificial Intelligence”), разработанный Себастьяном Траном (Sebastian Thrun) и представленный на площадке Udacity. Опубликованный МООС привлек внимание около 160000 тысяч студентов из 190 стран мира.

Однако, наиболее соответствует теме подготовке к ДМ другой курс — “Введение в машинное обучение” (“Intro to Machine Learning”), который также подготовлен при участии Себастьяна Трана. Его структура представляет собой 14 уроков (тем), приведенных в списке ниже:

Изучение алгоритмов машинного обучения с учителем (наивный байесовский классификатор, SVM, деревья принятия решений).

Подготовка данных к анализу (преобразование данных, выявление выбросов, нормирование, отбор).

Использование моделей регрессии для решения задач.

Обучение без учителя (алгоритм кластеризации K-means).

Работа с ключевыми параметрами (“фичами”): создание, отбор, извлечение.

Основные трудности при запуске и тестировании работы алгоритмов извлечения данных (распределение данных на тестовый и тренировочный наборы, перекрестная проверка данных, точность и т. д.).

В конце каждой темы обучающемуся предлагается выполнить проект, связанный с изученным материалом. В конце курса также для исследования предоставляется реальный набор данных и формулируется задача анализа.

Помимо этого курса, на площадке Udacity организована возможность получения так называемого “nanodegree” (“наностепень”) — помимо стандартного функционала этой площадки, при получении “наностепени” предлагается больший объем консультаций, в том числе и в индивидуальном формате. Обучающийся, с помощью сервиса календарей компании Google может назначить время для индивидуальной консультации с ассистентом курса. Кроме того после успешного освоения специальности выпускнику предлагается помощь в трудоустройстве. Сами курсы организуются при участии крупнейших компаний, функционирующих в ИТ отрасли. В настоящий момент уровень “nanodegree” может быть получен по следующим направлениям:

1. инженер по разработке алгоритмов машинного обучения (“Machine Learning Engineer”) (создан при участии Google);
2. аналитик данных (“Data Analyst”) (создан при участии Facebook, MongoDB).

Проанализируем структуру курса для специализации аналитика данных “Data Analyst”. Весь объем материала представлен в 8 частях, при этом в структуре выделяется теоретическая и практическая составляющие. Авторы предполагают, что задания, выполненные в рамках практической части, будут составлять основу для портфолио студента. Теория представлена набором видеолекций различной продолжительности. Видеолекции чередуются вопросами для самопроверки.

Практическая подготовка по этому курсу предполагает выполнение всех этапов решения типовой задачи средствами машинного обучения:

- постановка задачи анализа;
- формирование модели, выбор алгоритма решения;
- настройка алгоритма, запуск, оценка результатов;
- визуализация результатов.

Модуль практической работы предполагает использования программного продукта Anaconda, которая сочетает в себе несколько языков программирования (Python, R), средства визуализации и обработки больших объемов данных, а также средства их анализа.

В качестве проектов, которые необходимо будет выполнить, авторы предлагают решить следующие задачи:

- анализ результатов эксперимента, использующего эффект Струпа (требуется сформировать гипотезу или набор гипотез, позволяющих объяснить результаты эксперимента, определить значения некоторых статистических характеристик, визуализировать распределение объектов исследования, опровергнуть или подтвердить нулевую гипотезу);
- проанализировать данные об играх в бейсбол и предоставить ответ на следующие вопросы: (1) существует ли взаимосвязь между различными метриками качества в этом наборе данных, оценить

степень взаимосвязи (сильная или слабая) (2) выделить черты игроков с наибольшим уровнем заработной платы;

- оценить качество, достоверность, точность, полноту и другие показатели, очистить данные от выбросов, в качестве источника данных может быть выбрана БД, в которой используется SQL или документоориентированная БД.
- исследовать данные с использованием языка программирования R на предмет выявления зависимостей, выбросов, аномалий, определить существует ли распределение в данных; реализовать средствами языка R визуализацию данных; выявить ключевые характеристики (“фичи”) данных для использования в предсказательных моделях;
- решить задачу фильтрации спама, используя алгоритмы машинного обучения;
- использовать различные библиотеки для визуализации и при помощи них наглядно отобразить результаты исследования (в качестве библиотек используются D3.js и dimple.js);
- провести практический эксперимент с помощью методики A/B, проанализировать результаты и предложить дальнейшие шаги для следующего эксперимента.

На основании качественного анализа каждого из представленных ресурсов по обучению ДМ необходимо отметить, что представленные выше ресурсы имеют достаточно много отличий между собой.

Общие характеристики исследованных ресурсов, предположительно, могут быть рассмотрены как критерии оптимальности для проектируемого ресурса. Отзывы, оставленные обучающимися на этих курсах и успешно их завершившими, также были учтены при проектировании структуры, контента, дизайна ЭОР.

Критерии, отмеченные большинством респондентов-экспертов, сведены в таблицу, приведенную ниже.

Таблица 1.1. – Сравнение онлайн-курсов по подготовке к ДМ

Критерий сравнения	Machine Learning (Coursera)	Введение в машинное обучение (Coursera)	Data Analyst (Udacity)
Всего оценок	17 275	662	280
Средний балл по всем оценкам	4,9 / 5	4,5 / 5	4,4 / 5
Количество отзывов	3830	105	-
Количество отзывов от успешно завершивших обучение	1508	34	-
Распределение оценок за курс по пятибалльной шкале от всех пользователей	91,5% — «5» 8,00% — «4» 0,50% — «3»	71% — «5» 17% — «4» 6% — «3» 3% — «2» 3% — «1»	57,9% — «5» 33,2% — «4» 4,6% — «3» 2,9% — «2» 1,4% — «1»
Распределение оценок за курс по	92% — «5» 8% — «4»	65% — «5» 31% — «4» 4% — «3»	-

пятибалльной шкале от завершивших курс			
Замечания к курсу	Недостаточно теоретических сведений об алгоритмах. Несбалансированное по сложности содержание. GNU/Octave недостаточно адекватный инструмент для выполнения практических заданий. Недостаточное количество практических заданий с реальными данными.	Недостаточное количество сведений для выполнения практических заданий. Достаточно высокий порог вхождения (сложность математической теории, наличие навыков программирования на Python). Рассогласованность теории и практики. Наличие ошибок в системе проверки заданий.	Достаточно высокий порог вхождения: требуются навыки программирования (Python, R), знания в области математической статистики. В самом курсе эта информация не предоставляется. Несбалансированное по сложности содержание курса.

Выводы по главе 1

Анализ статей отечественных и зарубежных авторов и электронных образовательных ресурсов позволяет сформулировать вывод о недостаточно

разработанности вопроса подготовки к извлечению и анализу данных в вузе. Кроме этого, по характеру содержания подготовка к извлечению и анализу данных актуальна как сама по себе, так и в качестве средства подготовки в области веб-программирования.

ГЛАВА 2. ПРОЕКТИРОВАНИЕ И РАЗРАБОТКА ЭЛЕКТРОННОГО ОБРАЗОВАТЕЛЬНОГО РЕСУРСА ДЛЯ ПОДГОТОВКИ К ИЗВЛЕЧЕНИЮ И АНАЛИЗУ ДАННЫХ

2.1. Описание содержания подготовки и решаемых подготовкой задач

2.1.1. Общие характеристики содержания подготовки

Обучение веб-программированию проводится на кафедре КТиЭО РГПУ им. А.И. Герцена в рамках курса дисциплины «Программирование», начиная со второго семестра. В качестве языка программирования, использующегося для подготовки студентов, выбран JavaScript. Реализовать подготовку студентов к ДМ наиболее оптимально именно в рамках этой дисциплины.

Элементы подготовки к ДМ могут быть внедрены и в структуру других дисциплин, включенных в содержимое образовательной программы магистратуры по направлениям «Технологии и менеджмент электронного обучения» и «Корпоративное электронное обучение». Это могут быть, например, «E-learning - решения управления знаниями в образовательных учреждениях» и «Анализ данных в педагогических исследованиях» также преподаваемых на площадке кафедры (во втором и третьем семестрах соответственно).

На этапе анализа существующих методик было выявлено, что большинство их разработчиков рекомендуют проводить обозначенный курс в течение одного семестра. Поскольку в образовательном стандарте направления 09.03.01 «Информатика и вычислительная техника» дисциплина «Программирование» продолжается на протяжении нескольких семестров, курс может быть более детализированным и охватывать больший объем материала. Основные требования к разрабатываемому ЭОР были продиктованы ФГОС, а также стратегией развития информационных технологий в РФ.

Приведем профессиональные компетенции, реализованные в рамках используемой на кафедре образовательной программы:

- освоение методик использования программных средств для решения практических задач (ПК-2);
- разработка моделей компонент информационных систем, включая модели баз данных (ПК-4);
- разработка компонент программных комплексов и баз данных, использование современных инструментальных средств и технологий программирования (ПК-5).

Для успешного формирования указанных выше компетенций необходимо, чтобы студенты

1. **Знали:**
   - основные синтаксические конструкции и понятия JavaScript (циклы, условные операторы, функции и т.д.);
   - методику работы с объектной моделью документов (DOM) и использовали наиболее оптимальные техники для доступа к html-элементам;
   - основные модели данных JavaScript, а также дискретные структуры, используемые в веб-приложениях (массив, ассоциативный массив), методы перебора и сортировки массивов (в том числе и ассоциативных).
2. **Умели:**

- анализировать и обоснованно использовать программное обеспечения для отладки JavaScript-сценариев;
- анализировать и обоснованно использовать библиотеки и фреймворкидля реализации клиентского компонента (AngularJS, KnockoutJS, jquery, AmplifyJS);

- отслеживать перспективные технологии и тренды в JavaScript-разработке, а также анализировать и обоснованно использовать для решения поставленных задач современные, актуальные на данный момент технологические решения (в том числе кроссплатформенные и кроссбраузерные решения) и технологии (AJAX, REST, Comet, Web Sockets, HTML5, oAuth, OpenID, RSS, ATOM и др.).

3. **Владели:**

- базовым набором стандартных языковых средств разработки веб-страниц (HTML5, CSS3, SVG);
- навыками реализации кроссбраузерных решений на языке JavaScript;
- актуальными моделями/паттернами структуры/проектирования клиентский веб-сценариев;
- инструментарием проектирования/разработки/менеджмента веб-проекта;
- инструментарием размещения и поддержки веб-ресурсов (хостинг, CMS).

Кратко обозначим содержание дисциплины «Программирование», которое позволяет реализовать обозначенные выше компетенции.

Таблица 2.1. – Содержимое дисциплины с указанием разделов.

№ п/п	Название и краткое содержание темы
1.	**Введение в JavaScript и основы синтаксиса** История языка, описание его возможностей, области применения. Создание скрипта на JavaScript. Установка и настройка программного обеспечения, необходимого для работы с JS. Переменные, константы и типы данных, операторы, создание комментариев, форматы

	данных, приведение форматов. Выражения: арифметические, отношений, логические, тернарный оператор. Условные инструкции (if, switch), работа с циклами (while, for, foreach). Понятие функции, функции, определяемые пользователем, встроенные, аргументы функций, передача аргументов, значения, возвращаемые функцией.
2.	**Классы и объекты в JavaScript.** ООП. Реализация ООП в JS. Определение и использование классов, типов, подклассы. Методы и свойства, изменение свойств, понятие «прототип» (prototype, class). Объектная модель документа (DOM), взаимодействие с элементами веб-страницы Query selector.
3.	**Паттерны программирования.** Примеры. MVC, MVVM. Использование дополнительных библиотек и фреймворков: JQuery, Bootstrap, Angular JS, Knockout JS. Разработка веб-приложения: DHTML, Single page application, AJAX. Клиент-серверное взаимодействие.
4.	**Взаимодействие JavaScript с CSS3 и HTML5.** Хранение данных на стороне клиента: sessionStorage, localStorage, библиотека AmplifyJS, работа с cookies, геопозицирование, фоновые потоки выполнения (worker). Веб-сокеты. Node.js

С учетом специфики решаемых в области ИД задач, язык программирования JavaScript не может быть использован как основной. В качестве альтернативы JavaScript может быть использован язык программирования Python. Это позволит, с одной стороны, организовать подготовку студентов наиболее эффективно (язык содержит большое количество специализированных библиотек, расширений и на его основе строятся наиболее эффективные инструменты для анализа данных), а с другой стороны, не потребует значительного изменения содержимого курса, поскольку Python активно развивается и используется разработчиками в том числе и для решения задач веб-программирования.

При использовании Python уровень вхождения для студентов остается аналогичным тому, что необходим для освоения дисциплины при использовании языка JavaScript.

С учетом данных текущего учебного плана, трудоемкость дисциплины «Программирование» оценивается в 4 кредита, что составляет 144 часа на теоретическое обучение. При этом, аудиторная нагрузка составляет 72 часа, 72 часа отводится студентам на самостоятельную работу. В аудиторную нагрузку входят 27 часов лекций и 45 часов лабораторных работ.

2.1.2.Подготовка к применению ДМ

Подготовка к ДМ включает работу по следующим направлениям:

1. Описание полного цикла решения задачи с помощью ДМ.
2. Назначение, цель и детализированное представление каждого из этапов решения.
3. Исследование использования различных алгоритмов и методов ИД.
4. Изучение практических аспектов реализации алгоритмов и методов ИД.
5. Опишем содержимое каждого из направлений.

Исследователи в области извлечения и анализа данных [23, 14], как правило, описывают (1) полный цикл решения задачи как последовательность следующих этапов:

- анализ предметной области
- постановка задачи;
- сбор данных;
- подготовка данных (обработка);
- выбор модели для решения;
- подбор параметров модели или метода и алгоритма обучения модели;

- обучение модели;
- анализ качества обучения;
- анализ выявленных закономерностей.

На этапе анализа предметной области исследуются свойства её объектов, целью является выявление закономерностей в отношениях между значениями этих свойств. Под предметной областью, как правило, понимают область реальной действительности, которую возможно описать и провести её моделирование. На этом этапе важно попытаться разделить свойства предметной области на существенные и не значащие.

В результате этого этапа исследователь должен получить модель предметной области.

На следующем этапе, которым является постановка задачи, специалисту требуется сформулировать задачу и формализовать её, причем, в формулировке может требоваться включить описание поведения объектов исследования (как статического, так и динамического).

После постановки задачи исследователь переходит к этапам сбора и подготовки данных к анализу. Здесь следует отметить, что при решении реальных задач этот этап может повторяться несколько раз и может выполняться параллельно с выполнением других этапов при необходимости. На этом этапе должна быть спроектирована и создана база данных. Некоторые авторы [15] отмечают, что его продолжительность может достигать 80% от общего времени, отведенного на решение задачи. Подготовка данных включает проверку данных на дублирование, целостность, выбросы и многое другое. Как правило для различных типов данных, используются различные процедуры. Например, для текстовых данных проводится операция стемминга — нахождения основы слова [21], что является частью процесса нормализации текста.

Этапы выбора модели для решения и подбора её параметров - является ключевым моментом в процессе осмысления устройства и тенденций

анализируемого объекта. Модель является вручную созданным аналогом изучаемого объекта. Выделяют несколько типов моделей, позволяющих решать различные типы задач. Например, задача классификации множества объектов на подкатегории может быть решена при использовании прогнозирующей (классификационной модели), дескриптивная модель, в свою очередь, позволяет решать задачи кластеризации, когда требуется обобщить «похожие» друг на друга объекты.

На этапе обучения модели ей на вход подается выборка данных (или множество различных по объему выборок), для которых известны результирующие значения. В результате этого процесса модель «настраивается» на предсказание «правильных» результатов.

Некоторые исследователи при использовании иного подхода к структуризации процесса решения задачи обозначают этап анализа качества обучения как проверку и оценку модели и включают в него тестирование решения с использованием множества различных по объему выборок данных.

В завершение процесса решения задачи проводится анализ выявленных закономерностей, который включает описание полученных данных, их визуализацию, рекомендации по коррекции и обновлению модели и т.д.

Исследование использования различных алгоритмов и методов ИД (3) заключается в:

- анализе их работы;
- исследовании области применения (для решения каких задач тот или иной алгоритм может быть использован);
- выявлении достоинств и недостатков алгоритмов (и обзор решений для их устранения);
- анализе сложностей при использовании (например, настройка алгоритма или его «переобучение») и их устранение; «переобучение»

— одна из наиболее распространенных ситуаций на этапе построения решения;

- подготовке данных к использованию;
- программировании алгоритмов.

Изучение практических аспектов реализации алгоритмов и методов ИД (4) включает решение специалистом множества смежных задач в числе которых находятся:

- выбор наиболее оптимального решения для хранения данных (часто извлечение данных связывают с термином “big data”, суть которого сводится к решению множества технических задач, выбору технологий обработки и хранения сверх крупных объемов данных и т.д.).
- программная реализация оптимальных и эффективных алгоритмов ДМ на всех этапах: обучение, сбор данных, подготовка данных и т.д.;
- визуализация результатов анализа (использование прикладного ПО для визуализации, библиотек для построения графиков, платформ для анализа);
- обзор, описание и анализ различных инструментов для решения задачи ДМ (сюда входит ряд вопросов, связанных с целесообразностью программирования извлечения и анализа данных, необходимости использования библиотек, расширений для языков программирования для решения отдельных задач, а также вопросы, связанные с использованием платформ для решения задач ИД).

2.1.3. Методика проведения занятий

Перед описанием непосредственно самой методики проведения занятий, требуется сформулировать критерий готовности студента к извлечению данных.

Будем говорить, что сформирована готовность студентов к реализации извлечения данных, если в результате подготовки они способны:

1) формулировать задачу извлечения данных в рамках конкретной предметной области;
2) определить тип решаемой задачи и применить соответствующие действия для решения;
3) успешно исследовать предметную область, выделять значащие и незначащие параметры исследуемых объектов;
4) использовать терминологический аппарат (для описания решения, в дискуссиях и т.д.) и наиболее распространенные подходы к решению задач, касающиеся основных моделей, использующихся в ИД;
5) использовать адекватные инструменты и подходы для преодоления возникших препятствий на всех этапах решения задачи ИД;
6) анализировать полученные результаты, правильно их оценивать, руководствуясь ими для принятия решений;
7) отслеживать перспективные технологии в области ИД и применять их;
8) реализовать необходимый алгоритм различными способами, с учетом его особенностей.

В процессе подготовки студенты должны овладеть решением задач ИД, представляющих два основных типа:

- предсказательные — необходимо дать предварительный прогноз для ситуаций, о которых данных нет;
- описательные — требуется охарактеризовать имеющиеся скрытые закономерности.

К описательным задачам можно отнести [15, 16]:

1. Задачу поиска ассоциативных правил или паттернов (например, алгоритм Априори).
2. Задача группировки объектов (кластеризация), кластерный анализ (например, алгоритм k-средних).

К предсказательным задачам относятся:

1. Задача классификации объектов (например, деревья решений).
2. Регрессионный анализ (например, алгоритм линейной регрессии).

Приведем примеры задач, опытом решения которых должны овладеть студенты для успешного формирования готовности к извлечению данных:

1. Анализ рыночной корзины (market basket analysis) — на основании содержимого корзины покупателя делаются рекомендации о том, какой товар наиболее вероятно необходимо рекомендовать для покупки.
2. Анализ данных об автомобилях и их владельцах - на основании нескольких параметров автомобилей (марка, стоимость, возраст и т.д.) и водителя (стаж, возраст и т.д.) определить кластеры, соответствующие определенной рисковой группе.
3. Классификация ирисов Фишера — на основании нескольких характеристик, содержащихся в наборе данных о экземплярах ириса, необходимо определить к какому виду тот или иной экземпляр относится.
4. Прогнозирование стоимости жилья — на основании данных о стоимости жилья (площадь квартиры, количество комнат, стоимость) предсказать целевое значение (стоимость) для объектов, не содержащихся в выборке.

Опишем процесс решения одной задачи каждого типа.

Анализ данных об автомобилях и их владельцах построен на использовании метода решения ситуационных задач и включает как информационную деятельность, так и интерактивное взаимодействие с преподавателем и группой.

Студенты, ознакомившись с формулировкой задания и данными для анализа (текст с заданием предоставляется в электронном виде или как раздаточный материал) осуществляют подготовку и преобразование данных для анализа. Сохранение данных возможно или в специально созданную базу данных в виде таблицы или в “плоский файл” (например, CSV) (содержание файла см. в Приложение №1). Фрагмент структуры этого файла представлен ниже (см. Рисунок 2.1.), а полное содержание представлено в приложениях к работе (см. Приложение 1).

№	Car	CarPrice	CarAge	DriversAge	DriversExp
1	Acura	0.521	10	25	3
2	Audi	0.866	1	24	3
3	BMW	0.496	4	29	3

Рисунок 2.1. – Фрагмент файла с данными для анализа

На этом этапе они определяют и обсуждают с преподавателем с помощью какого инструмента данная задача может быть решена (к моменту решения задачи в качестве инструмента может быть использовано несколько различных решений):

- платформа анализа данных Knime;
- язык программирования Python с расширением SciPy.

От выбора их решения зависит необходимо ли будет реализовывать модель для извлечения данных вручную или можно воспользоваться готовыми компонентами платформы или расширениями языка.

Импорт данных в Knime реализуется компонентом CSV Reader, в его настройках указывается тип разделителя и некоторые другие настройки (см. Рисунок 2.2.).

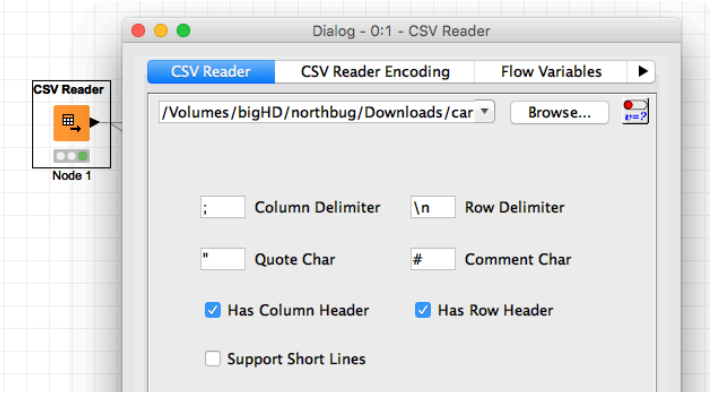

Рисунок 2.2. – Импорт CSV файла в Knime

В Python для чтения данных необходимо импортировать специальный пакет и после этого данные могут быть загружены в виде ассоциативного массива в переменную (см. Рисунок 2.3.).

```
>>> import csv
>>> with open('eggs.csv', newline='') as csvfile:
...     spamreader = csv.reader(csvfile, delimiter=' ', quotechar='|')
```

Рисунок 2.3. – Импорт библиотеки и данных файла в CSV

Дальнейший ход решения задачи зависит от выбранного инструмента и может включать этап нормирования данных.

Реализация этого действия для приведенных выше инструментов различается.

При использовании Knime нормировать данные можно с помощью специализированного узла “Normalizer”, позволяющего выбирать способ нормализации. Для текущей версии Knime их три: (1) с помощью линейно вычисляемого среднего значения по заданным максимальному и минимальному значению; (2) приведение к распределению по методу Гаусса со средним значением равным 0 и отклонением равным 1; (3) с помощью десятичного масштабирования (когда максимальное значение в столбце по модулю делится j-раз на 10 до того как его абсолютное значение не станет меньше или равным 1, все остальные значения в столбце затем делятся на 10 в степени j) (см. Рисунок 2.4.).

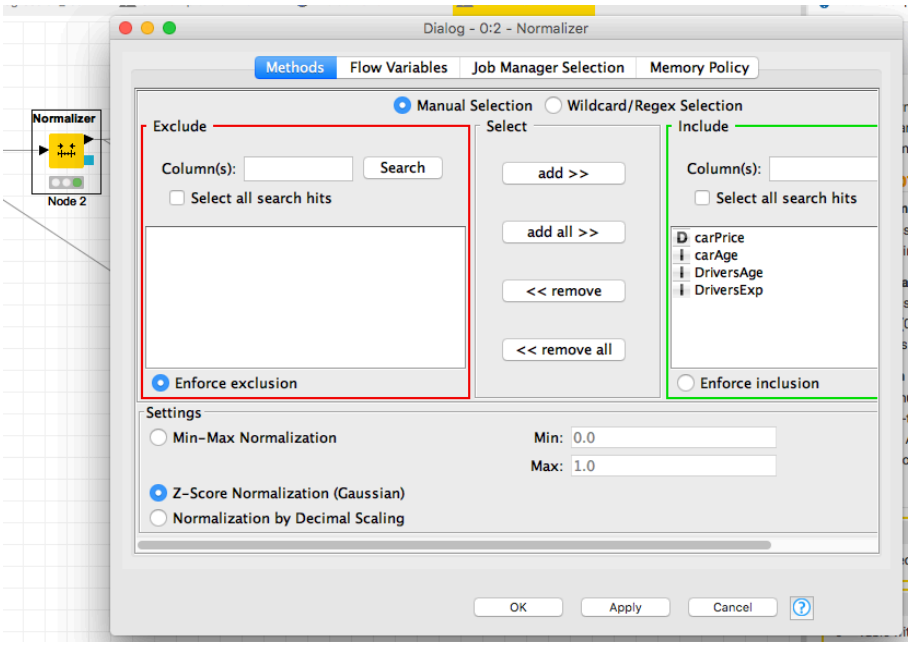

Рисунок 2.4. – Нормализация данных с помощью узла “Normalizer”

Приведем фрагмент нормализованных данных (см. Рисунок 2.5.).

Row ID	Car	carPrice	carAge	DriversAge	DriversExp
1	Acura	-0.316	1.104	-1.129	-0.916
2	Audi	0.204	-1.299	-1.219	-0.916
3	BMW	-0.353	-0.498	-0.768	-0.916
4	Buick	-0.176	0.837	1.129	1.335

Рисунок 2.5. – Фрагмент файла с нормализованными данными

В языке Python аналогичные операции обработки данных могут быть выполнены с помощью пакета sklearn.preprocessing, находящегося в пакете расширения scikit-learn. Пример нормирования значений матрицы размерности 3 приведен ниже. (см. Рисунок 2.6.).

```
>>> from sklearn import preprocessing
>>> import numpy as np
>>> X = np.array([[ 1., -1.,  2.],
...               [ 2.,  0.,  0.],
...               [ 0.,  1., -1.]])
>>> X_scaled = preprocessing.scale(X)

>>> X_scaled
array([[ 0.  ..., -1.22...,  1.33...],
       [ 1.22...,  0.  ..., -0.26...],
       [-1.22...,  1.22..., -1.06...]])
```

Рисунок 2.6. – Пример нормирования данных

В качестве моделей для решения задачи может быть использован алгоритм k-средних или иерархическая кластеризация, описанные ранее на лекции.

В результате решения задачи должны быть сформированы множества классов автомобилей, содержащие метку кластера.

Для решения задачи средствами Knime может быть использована компонента “K-means” или “Hierarhical Clustering” соответственно. В Python это реализуется с помощью пакета scipy.cluster.hierarchy или sklearn.cluster.KMeans соответственно.

Требуемый результат может быть представлен в виде таблицы или в виде дендрограммы (см. Рисунок 2.7.).

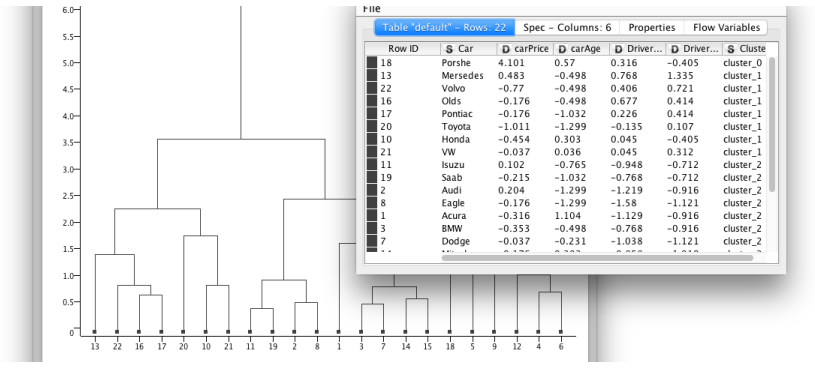

Рисунок 2.7. – Дендрограмма и таблица с метками кластеров

Групповое обсуждение, организованное после решения задачи предполагает, сравнительный анализ решений студентов, обсуждение результатов полученных с использованием нормализованных и «сырых» данных.

Задание тренировочно-проверочного типа, предполагающее исследовательскую деятельность студента и предлагающее решение задачи прогнозирования стоимости жилья, основано на анализе кода, предоставленного преподавателем (см. Рисунок 2.8.).

```
1
2   function [J, grad] = costFunction(theta, X, y)
3
4   % Initialize some useful values
5   m = length(y); % number of training examples
6
7   J = 0;
8   grad = zeros(size(theta));
9
10  for i=1:m
11      x_row = X(i,:);
12      J += (-y(i)*log(hypothesis(theta,x_row)) - ((1-y(i))*log(1-hypothesis(theta,x_row))));
13  end
14  J *= 1/m;
15  % ============================================================
16  for j=1:size(theta,1)
17      temp = 0;
18      for i=1:m
19          x_row = X(i,:);
20          temp+= (hypothesis(theta,x_row) - y(i))*x_row(j);
21          grad(j) = 1/m*temp;
22      end
23  end
24  % ============================================================
25  end
26
27
```

Рисунок 2.8. – Функция потерь для решения задачи классификации

В коде допущены ошибки, в результате которой необходимое значение вычисляется неверно. Ошибки допущены в строках 21 и 23. Накапливаемая переменная temp должна использоваться для вычисления после цикла for, переменную temp после цикла for и после произведения следует обнулить.

Студент должен определить тип задачи и использующуюся модель, самостоятельно найти ошибку. Поскольку речь идет о вычислении функции потерь, то здесь решается задача классификации. Анализируя код, представленный выше, можно обнаружить вычисление логарифма (в том числе и во вспомогательных файлах, где реализована функция hypothesis()). Таким образом, можно определить, что эта функция используется для логистической регрессии.

После этого подтвердить правильность своего решения (например, сократив набор данных и вычислив функцию потерь вручную).

Обучающемуся доступно несколько векторов-путей решений:

- анализ кода и каждого этапа, который он решает, сопоставление его с алгоритмом работы использованной модели;
- решение задачи без опоры на существующий код, с использованием альтернативных средств, а после этого реверс-инжиниринг решения преподавателя).

Второй путь предполагает использование платформы Knime и осуществление всего процесса решения задачи: (1) импорта множества данных для обучения в формате CSV, (2) визуализации их на графике данных (см. Рисунок 2.9.).

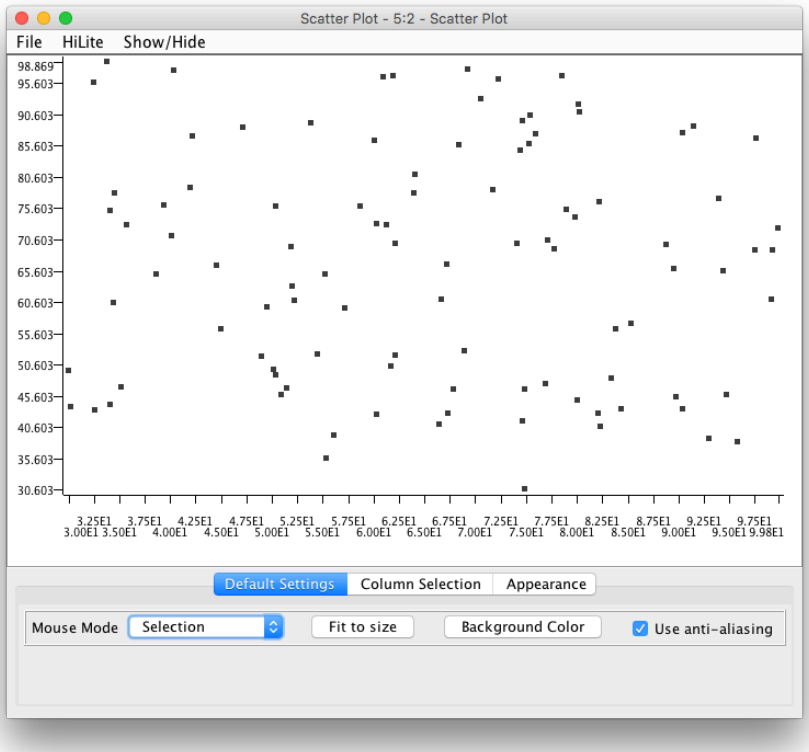

Рисунок 2.9. – Визуализация данных

Обучение классификатора решению этой задачи (3). Для этого в Knime используется узел “Polynomial Regression Lerner”, позволяющий построить модель для предсказания (см. Рисунок 2.10.).

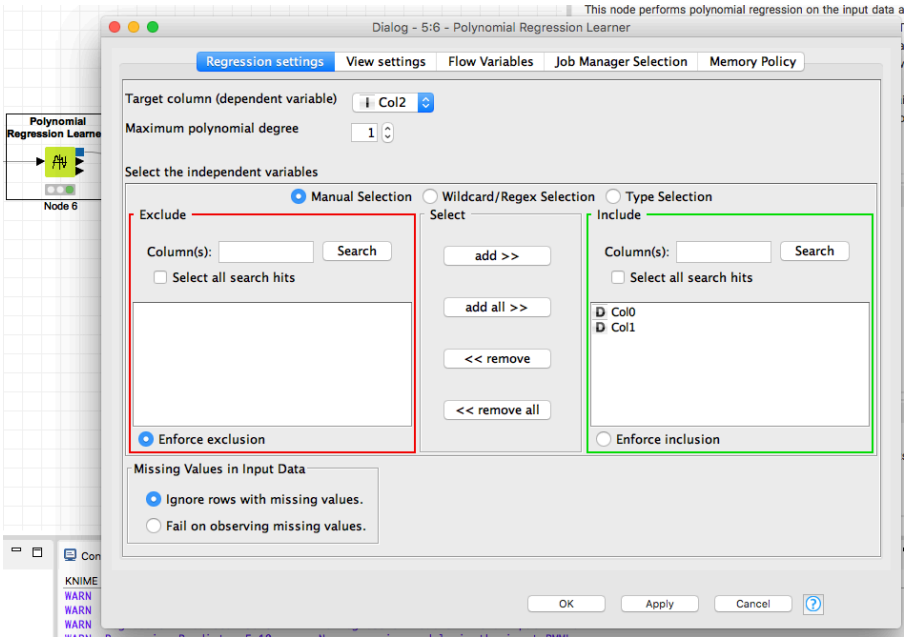

Рисунок 2.10. – Построение модели для предсказания

Модель настраивают и после её настройки значение целевого параметра может быть предсказано (см. Рисунок 2.11.).

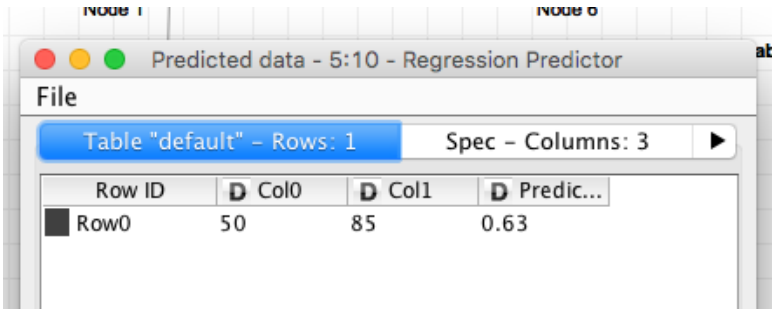

Рисунок 2.11. – Предсказанная вероятность

Для объекта со значениями 50 и 85 соответственно предсказываемое значение будет составлять 0,63.

2.2. Описание средств и форм подготовки

В качестве дополнительных средств и форм подготовки студентов к ИД предполагается предотвращать возникновение устоявшегося негативного отношения к ошибкам у студентов. Этот метод, условно обозначенный как «мотивированное тестами обучение» (“test driven instruction”) представляет собой один из видов электронного обучения (в том числе и извлечению данных), при которых изначально проектируется набор тестов, а затем формулируются задания. Задания, в свою очередь, готовятся путем анализа реальных ситуаций. Здесь следует провести параллель с одной методологией разработки, в которую заложена идея разработки через тестирование (TDD - test driven development). Процесс создания кода при использовании этой методологии начинается с написания теста, определяющего требования или спецификации к модулю (или любому другому “готовому” фрагменту кода например, функции, методу, классу и т.д.). Этот процесс итеративно повторяется при добавлении нового функционала.

Поскольку процесс организации проверки выполнения студентами лабораторных и практических работ вручную не представляется возможным, в силу их большого числа и разнообразия, его автоматизация даст возможность ускорить взаимодействие между обучающим и обучающимся и облегчить процесс корректировки и исправления кода.

Автоматизация тестирования привела к появлению большого количества фреймворков и библиотек для большинства современных языков (в том числе и для Python), которые могут быть использованы для решения этой задачи.

С учетом описанного выше, деятельность на занятии можно представить с помощью следующих шагов:

- преподаватель формулирует задачу (изначально построенную с учетом принципа TDI), которую студенты должны решить;

- тщательно описываются входные и выходные значения (наборы данных);
- при необходимости предоставляется шаблон решения (частично или полностью правильный);
- студент должен реализовать решение своей задачи и запустить проверяющий модуль, который отправит данные на сервер, развернутый преподавателем, где ответ будет автоматически проверен;
- в случае, если ответ студента и преподавателя совпадают, задание считается выполненным;
- в случае, если ответы не совпадают, преподаватель может вручную осуществить анализ кода (code review) или сформировать рекомендации (в автоматическом, используя эвристики, или ручном режиме) с учетом возвращаемых ответов.

Комбинация данных инструментов позволяет с одной стороны снизить нагрузку преподавателя, поскольку он не вынужден проверять множество решенных студентами лабораторных работ вручную, а с другой стороны, при выявлении неправильного или сомнительного решения, провести анализ кода.

Выводы по главе 2

На основе анализа содержимого существующих ЭОР, ЭИОС, посвященных извлечению, анализу данных и машинному обучению были сформулированы требования к методике подготовки инженеров, а также к ЭУМК, в рамках которого эта методика будет опубликована. Осуществлено проектирование ресурса, отбор и анализ инструментов для его реализации, проведена разработка с применением итеративной модели жизненного цикла.

Разработанный контент может быть адаптирован к использованию в системе управления обучением (LMS) Moodle, развернутой на базе РГПУ им. А.И. Герцена или аналогичной (при применении стандарта SCORM).

ЗАКЛЮЧЕНИЕ

Автором данной диссертации спроектирован и реализован электронный образовательный ресурс на тему «Методика подготовки будущих инженеров к образовательному дата-майнингу»:

- проанализированы существующие ЭОР, ЭИОС, посвященные тематике извлечения и анализа данных, а также машинного обучения;
- выявлены требования к ЭОР и ЭИОС для подготовки к дата-майнингу на основе проведенного анализа;
- спроектирована модель ЭУМК;
- реализована разработанная модель с использованием оптимальных технических средств;
- проведена апробация на площадке кафедры компьютерных технологий и электронного обучения (КТиЭО).

Таким образом, поставленные задачи полностью решены и цель диссертации достигнута. С содержимым ресурса можно познакомиться по адресу http://mlcourse.ru.

К числу основных направлений дальнейшего совершенствования созданного продукта можно отнести добавление функционала для реализации подхода, обозначенного в пункте «2.3. Описание средств и форм подготовки», а также обновление и расширение содержимого (например, подготовки к использованию алгоритмов распознавания образов).

СПИСОК СОКРАЩЕНИЙ И УСЛОВНЫХ ОБОЗНАЧЕНИЙ

Условное сокращение	**Полное наименование**
DM	Data mining
KDD	Knowledge discovery in databases
ML	Machine Learning
MOOC	Massive online open courses
БД	База данных
ДМ	Дата майнинг
ИД	Извлечение данных
Кафедра КТиЭО	Кафедра компьютерных технологий и электронного обучения
ИКНиТО	Институт компьютерных наук и технологического образования
ПК	Персональный компьютер
ПО	Программное обеспечение
ПП	Программный продукт
РГПУ им. А. И. Герцена	Российский государственный педагогический университет имени А. И. Герцена
ЭИОС	Электронная информационно-образовательная среда
ЭОР	Электронный образовательный ресурс

СПИСОК ЛИТЕРАТУРЫ

1. Акт правительства Российской Федерации "Стратегия развития отрасли информационных технологий в Российской Федерации на 2014 - 2020 годы и на перспективу до 2025 года" от 1 ноября 2013 года № 2036-р // официальный сайт Правительства России. 2013 г. — [Электронный ресурс] URL: http://government.ru/media/files/41d49f3cb61f7b636df2.pdf
2. Закон Российской Федерации "Об образовании в Российской Федерации" от 21 декабря 2012 года № 273-ФЗ // Российская газета. 2012 г. № 5976 (303).
3. Приказ Минобрнауки России от 12 января 2016 г. № 5 «Об утверждении федерального государственного образовательного стандарта высшего образования по направлению подготовки 09.03.01 информатика и вычислительная техника (уровень бакалавриата)». — [Электронный ресурс] URL: http://минобрнауки.рф/документы/7980
4. Андреев А.А. Российские открытые образовательные ресурсы и массовые открытые дистанционные курсы [Текст] / А.А. Андреев // Высшее образование в России. – 2014. – No 6. – С. 150-155.
5. Анохин П. В. Использование интеллектуального анализа данных образовательного процесса при дистанционном обучении // Автоматизация, мехатроника, информационные технологии. Материалы III Международной научно-технической интернет-конференции молодых ученых. Омск, 15-16 мая 2013 года.: Материалы конференции. — Омск, 2013. — с. 204-207
6. Бадарч Д., Токарева Н.Г., Цветкова М.С. МООК: реконструкция высшего образования. С. 135-146.

7. Васильева Т. В. Методологическая подготовка студентов специальности «Прикладная математика» на примере курсов «Методы представления и анализа данных» и «Методы прогнозирования» // Труды Дальневосточного государственного технического университета. - Владивосток: ДФУ (Владивосток), 2005. - С. 48-50.
8. Демин И.С. Нейросетевые технологии в подготовке экономистов // Социально-экономические явления и процессы. - 2014. - №1 (59). - С. 11-13.
9. Демин И.С. Технологии кластеризации данных в подготовке экономистов // Человеческий капитал. — 2012. — №12 (48). — С. 55-59.
10. Дюк В. А., Флегонтов А. В., Фомина И. К. Применение технологий интеллектуального анализа данных в естественнонаучных, технических и гуманитарных областях // Известия РГПУ им. А.И. Герцена. 2011. №138. URL: http://cyberleninka.ru/article/n/primenenie-tehnologiy-intellektualnogo-analiza-dannyh-v-estestvennonauchnyh-tehnicheskih-i-gumanitarnyh-oblastyah (дата обращения: 09.06.2016).
11. Есаулов В.А. Разработка электронного учебного пособия «Эволюционные алгоритмы интеллектуального анализа данных» // Традиции русской инженерной школы: вчера, сегодня, завтра. - Новочеркасск: Южно-Российский государственный политехнический университет (НПИ) имени М.И. Платова, 2015. - С. 139-142.
12. Константиновская Н.В., Мухаматзанова М.Ш. О перспективах обучения статистическому анализу медико-социологических данных // Волгоградский научно-медицинский журнал. - 2010. - №1 (25). - С. 13-14.
13. Никитин П.В., Горохова Р. И. Компьютерные системы анализа данных в подготовке будущих учителей // Вестник Марийского государственного университета. - 2015. - №1 (16). - С. 44-47.

14. НОУ ИНТУИТ | Лекция | Методы поиска ассоциативных правил // НОУ «ИНТУИТ» URL: http://www.intuit.ru/studies/courses/6/6/lecture/186?page=5 (дата обращения: 06.06.2016).

15. НОУ ИНТУИТ | Лекция | Процесс Data Mining. Начальные этапы // НОУ «ИНТУИТ» URL: http://www.intuit.ru/studies/courses/6/6/lecture/192 (дата обращения: 06.06.2016).

16. НОУ ИНТУИТ | Лекция | Процесс Data Mining. Начальные этапы // НОУ «ИНТУИТ» URL: http://www.intuit.ru/studies/courses/6/6/lecture/192?page=2 (дата обращения: 06.06.2016).

17. Пиотровская К. Р. ТЕКСТ-МАЙНИНГ: ПЕРСПЕКТИВЫ РАЗВИТИЯ // Известия Российского государственного педагогического университета им. АИ Герцена. – 2014. – №. 168.

18. Ризаев И.С., Яхина З.Т., Мифтахутдинов Д.И. Компьютерные технологии обучения методам data mining обработки данных // Образовательные технологии и общество. - 2015. - №2. - С. 514-526.

19. Сидорова Н. П. Применение средств интеллектуального анализа данных для оценки качества подготовки специалистов [текст] // Международная научно-практическая конференция “Перспективы, организационные формы и эффективность развития сотрудничества ВУЗов стран Таможенного союза и СНГ” Королёв, 23-24 мая 2013 г.: Материалы конференции. — Финансово-технологическая академия (Королев), 2013. — С. 399-403

20. Синицын В. Ю. Практикум по теории вероятностей и математической статистике в вычислительной среде R // Проблемы и перспективы развития образования в России. - 2013. - №24. - С. 42-46.

21. Стемминг – Википедия [Электронный ресурс]. // Википедия URL: https://ru.wikipedia.org/wiki/Стемминг (Дата обращения: 06.06.2016)

22. Федин Ф.О. Анализ данных. Часть 1. Подготовка данных к анализу [Электронный ресурс]: учебное пособие/ Федин Ф.О., Федин Ф.Ф.— Электрон. текстовые данные.— М.: Московский городской педагогический университет, 2012.— 204 c.— Режим доступа: http://www.iprbookshop.ru/26444.— ЭБС «IPRbooks», по паролю

23. Data mining – Википедия [Электронный ресурс]. // Википедия URL: https://ru.wikipedia.org/wiki/Data_mining (Дата обращения: 06.06.2016)

24. Data mining with WEKA, Part 1: Introduction and regression // Data mining with WEKA, Part 1: Introduction and regression URL: http://www.ibm.com/developerworks/library/os-weka1/ (дата обращения: 06.06.16).

25. Dickey D. A. Teaching Data Mining in a University Environment //Proceedings of the SUGI 30 Conference. – 2005. – С. 1-10.

26. Educational data mining - Wikipedia, the free encyclopedia // Wikipedia URL: https://en.wikipedia.org/wiki/Educational_data_mining (дата обращения: 07.06.2016).

27. Jafar M. J., Anderson R. A tools-based approach to teaching data mining methods //Journal of Information Technology Education: Innovations in Practice. – 2010. – Т. 9.

28. Malibari A. et al. Improve Teaching Method of Data Mining Course //International Journal of Modern Education and Computer Science. – 2012. – Т. 4. – №. 2. – С. 15. URL: http://www.mecs-press.org/ijmecs/ijmecs-v4-n2/IJMECS-V4-N2-3.pdf

29. Sanati-Mehrizy R., Sanati-Mehrizy P., Minaie A., Dean C. How a data mining course should be taught in an undergraduate computer science

curriculum // ASEE ANNUAL CONFERENCE AND EXPOSITION, CONFERENCE PROCEEDINGS. - Louisville, KY:

30. What is Machine Learning: A Tour of Authoritative Definitions and a Handy One-Liner You Can Use // Machine Learning Mastery URL: http://machinelearningmastery.com/what-is-machine-learning/ (дата обращения: 07.06.2016).
31. What is the difference between data mining, statistics, machine learning and AI? // Cross Validated URL: http://stats.stackexchange.com/questions/5026/what-is-the-difference-between-data-mining-statistics-machine-learning-and-ai (дата обращения: 07.06.2016).

ПРИЛОЖЕНИЕ 1

Numb	Car	CarPrice	CarAge	DriversAge	DriversExp
1	Acura	0,521	10	25	3
2	Audi	0,866	1	24	3
3	BMW	0,496	4	29	3
4	Buick	0,614	9	50	25
5	Corvetter	1,235	15	62	38
6	Chrysler	0,614	9	43	21
7	Dodge	0,706	5	26	1
8	Eagle	0,614	1	20	1
9	Ford	0,706	11	54	10
10	Honda	0,429	7	38	8
11	Isuzu	0,798	3	27	5
12	Mazda	0,126	10	51	20
13	Mersedes	1,051	4	46	25
14	Mitsub.	0,614	7	28	2
15	Nissan	0,429	6	31	6
16	Olds	0,614	4	45	16
17	Pontiac	0,614	2	40	16
18	Porshe	3,454	8	41	8
19	Saab	0,588	2	29	5
20	Toyota	0,059	1	36	13
21	VW	0,706	6	38	15
22	Volvo	0,219	4	42	19

ПРИЛОЖЕНИЕ 2

Интеллект-карта понятия «Подготовка к дата-майнингу», http://bit.ly/dm-imap