УДК 378

АНАЛИЗ ДАННЫХ: МЕТОДОЛОГИЯ, МЕТОДЫ И ИХ РЕАЛИЗАЦИЯ В ПРОЦЕССЕ ОБУЧЕНИЯ

Брит Анна Александровна, кандидат физико-математических наук,
доцент кафедры «Информационные технологии и математическое обеспечение информационных систем», ИЭиУ АПК
Красноярский государственный аграрный университет, Красноярск, Россия
e-mail: anna.a.brit@gmail.com
Миндалев Игорь Викторович, доцент кафедры «Информационные технологии и математическое обеспечение информационных систем», ИЭиУ АПК
Красноярский государственный аграрный университет, Красноярск, Россия
e-mail: mindalev@rambler.ru
Болдарук Ирина Ивановна, старший преподаватель кафедры «Информационные технологии и математическое обеспечение информационных систем», ИЭиУ АПК
Красноярский государственный аграрный университет, Красноярск, Россия
e-mail: boldaruk1@mail.ru
Кузнецова Александра Сергеевна, кандидат физико-математических наук,
доцент кафедры «Информационные технологии и математическое обеспечение информационных систем», ИЭиУ АПК
Красноярский государственный аграрный университет, Красноярск, Россия
e-mail: alexakuznetsova85@gmail.com

Аннотация. В связи с непрерывно растущим потоком данных во всех отраслях, методология их поиска и отбора, методы анализа данных становятся важной частью подготовки современного специалиста. В статье описывается возможность применения анализа данных, как междисциплинарной области знаний, при осуществлении образовательного процесса.

Ключевые слова: анализ данных, Data mining, методология, методы, данные, научно-исследовательская работа, практика.

DATA ANALYSIS: METHODOLOGY, METHODS AND THEIR IMPLEMENTATION IN THE LEARNING PROCESS

Brit Anna Alexandrovna, Candidate of Physical and Mathematical Sciences,
associate professor of the Department of «Information technology and mathematical support of information systems», Institute of Economics and Management in AIC
Krasnoyarsk state agrarian university, Krasnoyarsk, Russia
e-mail: anna.a.brit@gmail.com
Mindalev Igor Viktorovich,
associate professor of the Department of «Information technology and mathematical support of information systems», Institute of Economics and Management in AIC
Krasnoyarsk state agrarian university, Krasnoyarsk, Russia
e-mail: mindalev@rambler.ru
Boldaruk Irina Ivanovna, Senior Lecturer of the Department of "Information technology and mathematical support of information systems", Institute of Economics and Management in AIC
Krasnoyarsk state agrarian university, Krasnoyarsk, Russia
e-mail: boldaruk1@mail.ru
Kuznetsova Alexandra Sergeevna, Candidate of Physical and Mathematical Sciences,
associate professor of the Department of «Information technology and mathematical support of information systems», Institute of Economics and Management in AIC
Krasnoyarsk state agrarian university, Krasnoyarsk, Russia
e-mail: alexakuznetsova85@gmail.com

Abstract. In connection with the continuously growing flow of data in all industries, the methodology for their search and selection, methods of data analysis are becoming an important part of the

training of a modern specialist. The article describes the possibility of using data analysis as an interdisciplinary field of knowledge in the implementation of the educational process.

Key words: data analysis, data mining, methodology, methods, data, research work, practice.

Ежегодно объем данных в мире увеличивается с огромной скоростью. По прогнозам IDC к 2025 году будет создано порядка 175 зетабайт данных. [10]

Термин «большие данные» был впервые введен редактором журнала Nature Линчем Клиффордом в 2008 году. [9] Это понятие было посвящено большому росту объемов информации по всему миру. Большие данные – это метаданные; массив данных огромных размеров, которые обладают свойствами: объем, разнообразие, скорость, изменчивость, достоверность. Большими данными в основном обладают корпорации и в реальности аналитики очень редко работают с большими данными. [5]

Понятие «анализ данных» будем рассматривать, как междисциплинарную область знаний, и определим, как «область математики и информатики, занимающуюся построением и исследованием наиболее общих математических методов и вычислительных алгоритмов извлечения знаний из экспериментальных данных». [4] Таким образом, Data-аналитик – специалист, который получает знания из данных. Цикл аналитической деятельности (рисунок 1) состоит из четырех уровней: данные (data), информация (information), знание (knowledge), мудрость (wisdom).

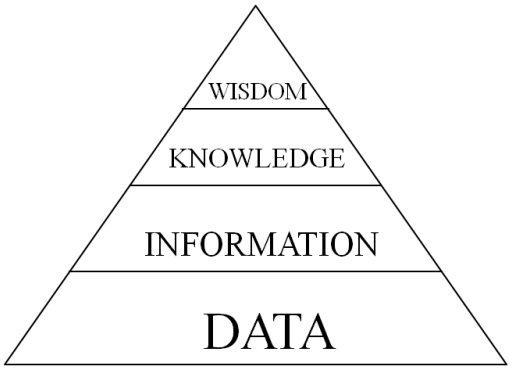

Рисунок 1 – DIKW модель по Р.Акоффу

Сначала происходит получение данных из внешнего мира, далее осуществляется анализ отношений и связей данных, в результате которых получается информация, после рассмотрения информации в ситуационном контексте будут получены знания и только после взвешенных рекомендаций и обоснований появляется глубокое понимание – мудрость. Целью аналитической работы является получение оснований для принятия управленческого решения, которые могут быть разработаны при применении математических методов и моделей с помощью технических и программных средств (Data mining).

Выделяют следующие методологии Data mining: Cross-Industry Standard Process for Data Mining (CRISP-DM) - межотраслевой стандартный процесс исследования данных; Knowledge Discovery in Databases (KDD) - обнаружение знаний в базах данных; Sample, Explore, Modify, Model, Assess (SEMMA) - пробуйте, исследуйте, модифицируйте, моделируйте, оценивайте; Team Data Science Process (MS TDSP) - процесс сбора данных в команде. [5] Ни одна из методологий не содержит готовых алгоритмов для решения задачи и не содержит готовых решений.

Одной из самых распространенных методологий исследования данных является CRISP-DM. Эта методология опирается на понимание бизнес-процессов предприятия и состоит из выполнения шести этапов и соответствующих им задач. (рисунок 2)

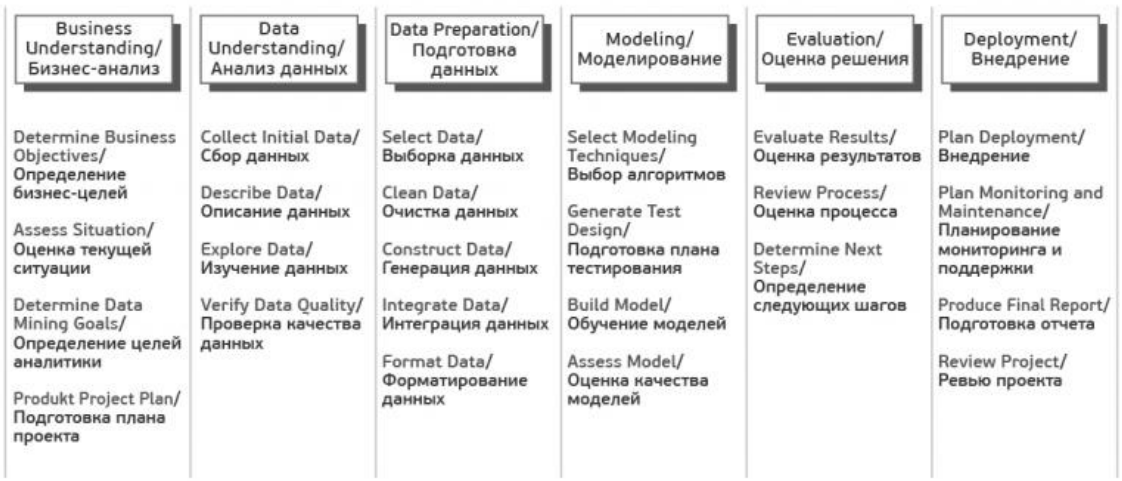

Рисунок 2 – Этапы CRISP-DM [7]

Отметим, что самыми важными и трудозатратными этапами, к которым чаще всего возвращаются, являются этапы анализа данных и подготовки данных. Они составляют большую часть времени работы аналитика данных.

В Data mining выделяют статистические и кибернетические методы: методы описательной статистики, статистические коэффициенты, эконометрические методы и модели, методы классификации и кластеризации, сетевой анализ и теория графов, методы нечеткой логики, генетические алгоритмы, искусственные нейронные сети, системы обработки экспертных знаний и др.

В основном применяются инструменты для анализа данных - Excel, SAS, SPSS, Deductor, Python, R и др., внедрение осуществляется на языках программирования – Python, Java, C++ и др. [5]

На примере студентов, обучавшихся по направлению 01.03.02 «Прикладная математика и информатика», рассмотрим применение методологии Data mining в образовательном процессе в рамках построения цифровой образовательной среды, которая включает в себя совокупность информационно-образовательных ресурсов, средства информационных и коммуникационных технологий, систему современных педагогических технологий и комплекс разносторонних коммуникативных отношений в образовательной деятельности. [1,6,8]

Для проведения анализа данных необходимо изучение следующих дисциплин «Статистика», «Теория вероятностей и математическая статистика», «Эконометрика», «Математическое моделирование микроэкономических и макроэкономических процессов», «Теория систем и системный анализ», «Базы данных», «Языки и методы программирования», «Программирование на C++».

Для того, чтобы пройти весь процесс от определения бизнес-целей до внедрения проекта, одной дисциплины не достаточно.

Это возможно осуществить в учебное время при прохождении производственных практик и написании выпускной квалификационной работы (ВКР). Во время первой производственной практики «Практика по получению профессиональных умений и опыта профессиональной деятельности» на 3 курсе студенту необходимо получить техническое задание от руководителя практики для анализа бизнес-процессов предприятия или организации, последующего анализа данных и подготовки данных в соответствии с задачами этапов CRISP-DM, что определит вторую главу ВКР. Далее до следующей практики, которая проходит на 4 курсе, необходимо описать методы и/или модели для реализации этапа моделирования CRISP-DM, что составит первую главу ВКР. Этапы проведения моделирования, оценки решения и внедрения CRISP-DM пройдут во время второй производственной практики «Преддипломная практика», описание этих этапов составит третью главу ВКР.

Также возможно осуществить работу во внеучебное время в рамках научно-исследовательской работы для выполнения проектных работ, участия в различных конкурсах, грантах и конференциях, написания научных статей. [2,3] Этот вариант работы обязывает преподавателя владеть приемами передачи знаний и организацией учебного процесса таким образом, чтобы заинтересовать студента и мотивировать на дальнейшее участие в научно-исследовательской

работе. И определяет качества, которыми должен обладать студент – инициативность, самостоятельность, коммуникативность, для успешного осуществления этого вида работ.

Проведенный педагогический эксперимент показал, что первый вариант является наиболее предпочтительным, так как обязывает всех участников соблюдать сроки выполнения отчетов по практикам и сроки написания ВКР, установленные во время учебного процесса. Полученные результаты могут быть полезными при обучении студентов других направлений.

Публикация данной статьи и участие в стажировке «Интеллектуальный анализ больших данных: решение социально-значимых задач и методика преподавания» осуществлено при поддержке Краевого государственного автономного учреждения «Красноярский краевой фонд поддержки научной и научно-технической деятельности»

Список литературы

1. Амбросенко Н.Д. Концепция формирования электронной информационной образовательной среды университета / Н.Д. Амбросенко // В сборнике: Проблемы современной аграрной науки. Материалы международной заочной научной конференции. 2017. С. 195-198.
2. Болдарук И.И. Научно-исследовательская работа студентов как форма организации учебного процесса / И.И. Болдарук, А.А. Брит, Л.Н. Шевцова // В сборнике: Наука и образование: опыт, проблемы, перспективы развития. Материалы международной научно-практической конференции. Ответственные за выпуск Е.И. Сорокатая, В.Л. Бопп. 2020. С. 115-117
3. Пушкарева Т.П. О формировании математической компетентности студентов на основе проектно-целевого подхода / Т.П. Пушкарева, В.В. Калитина // Современные проблемы науки и образования. 2019. № 4. С. 113.
4. Словари и энциклопедии на академике. [Электронный ресурс]. https://dic.academic.ru (дата обращения: 25.09.2021).
5. Стажировка Центра прикладного анализа больших данных. [Электронный ресурс]. https://moodle.ido.tsu.ru (дата обращения: 25.09.2021).
6. Титовская Н.В. Использование LMS Moodle в Красноярском ГАУ / Н.В. Титовская, С.Н. Титовский // В сборнике: Наука и образование: опыт, проблемы, перспективы развития. материалы международной научно-практической конференции. 2018. С. 268-271.
7. Хабр: сообщество IT-специалистов. CRISP-DM: проверенная методология для Data Scientist-ов. [Электронный ресурс]. https://habr.com/ru (дата обращения: 25.09.2021)./
8. Шилова О.Н. Цифровая образовательная среда: педагогический взгляд / О.Н. Шилова // ЧиО. - 2020. - №2 (63). - URL: https://cyberleninka.ru/article/n/tsifrovaya-obrazovatelnaya-sreda-pedagogicheskiy-vzglyad (дата обращения: 25.09.2021).
9. Lynch C. Bigdata: How do your data grow? / C. Lynch // Nature. 2008. V.455. № 7209. P.28-29.
10. Reinsel D. The Digitization of the World From Edge to Core / D. Reinsel, J. Gantz, J. Rydning // An IDC White Paper. November, 2018. P.28