Федеральное государственное бюджетное образовательное учреждение
высшего образования
Московский государственный университет имени М.В. Ломоносова
Высшая школа управления и инноваций

УТВЕРЖДАЮ
и.о.декана
/В.В.Печковская /
«12» февраля 2019 г.

РАБОЧАЯ ПРОГРАММА ДИСЦИПЛИНЫ

МЕТОДЫ АНАЛИЗА ДАННЫХ

МАГИСТРАТУРА

27.04.05 "ИННОВАТИКА"

Форма обучения:

очная

Рабочая программа рассмотрена и одобрена
Советом факультета

(протокол № 2, 12 февраля 2019 г.)

Москва 2019

Рабочая программа дисциплины (модуля) разработана в соответствии с самостоятельно установленным МГУ образовательным стандартом (ОС МГУ) для реализуемых основных профессиональных образовательных программ высшего образования по направлению подготовки / специальности 27.04.05 «Инноватика» (программа магистратуры), утвержденным приказом МГУ от 22 июля 2011 года № 729 (в редакции приказов МГУ от 22 ноября 2011 года № 1066, от 21 декабря 2011 года № 1228, от 30 декабря 2011 года № 1289, от 22 мая 2015 года № 490, от 30 июня 2016 года № 746).

Год (годы) приема на обучение: 2019, 2020.

I. Цели и задачи учебной дисциплины

Целью изучения дисциплины «Методы анализа данных» является изучить технологии анализа данных: OLAP, KDD, Data Mining и подготовки данных; дать представление об автоматизированных моделях анализа данных, применить методы анализа данных на примере решения задач сегментации, классификации, прогнозирования.

Задачами дисциплины являются:

- изучение понятийно-категориального аппарата в области углубленного анализа данных;
- формирование представлений об общей методологии консолидации, подготовки и анализа данных;
- обеспечение освоения современных методов OLAP, KDD, Data Mining;
- формирование навыков и умений, необходимых для создания и развития корпоративных аналитических систем.

В результате изучения данного курса обучающиеся получат знания об общей методологии и конкретных методах углубленного анализа данных, приобретут навыки и умения построения автоматизированных аналитических моделей.

II. Место дисциплины в структуре ОПОП ВО

Дисциплина «Методы анализа данных» относится к профессиональному блоку вариативной части (дисциплинам по выбору студента) учебного плана программы магистратуры 27.04.05. «Инноватика».

Изучение дисциплины базируется на знаниях и умениях, полученных обучающимися в процессе изучения естественных, гуманитарных, социальных и экономических дисциплин бакалавриата, таких как «Математика», «Системный анализ и теория принятия решений», «Методы исследования в менеджменте», а также дисциплины программы магистратуры «Моделирование и количественные методы в бизнесе».

Для успешного освоения дисциплины обучающийся должен:

Знать:

- фундаментальные положения дискретной математики;
- теоретические основы теории оптимизации и регрессионного анализа;
- основные проблемы современной философии и подходов к их решению;

Уметь:

- использовать междисциплинарные системные связи наук;
- анализировать и оценивать философские проблемы при решении социальных и профессиональных задач;
- применять математический инструментарий к решению социальных и профессиональных проблем.

Владеть:

- навыками работы в среде Excel;
- навыками выбора наиболее актуальных направлений научных исследований, ставить задачи исследования и определять способы решения поставленных задач;
- самостоятельно приобретать и использовать в практической деятельности новые знания и умения в различных сферах деятельности.

Знания, навыки и умения, полученные при изучении дисциплины «Методы анализа данных» обеспечивают успешное освоение дисциплины «Сенсорные сети и

нейрокоммуникации» и необходимы для прохождения преддипломной практики, осуществления научно-исследовательской работы и написания выпускной квалификационной работы (магистерской диссертации). Изучается на 2 курсе (3 семестр).

III. Требования к результатам освоения дисциплины

В результате освоения дисциплины должны быть сформированы следующие компетенции:

УК-1. Способен формулировать научно обоснованные гипотезы, создавать теоретические модели явлений и процессов, применять методологию научного познания в профессиональной деятельности.

УК-2. Готовность к саморазвитию, самореализации, использованию творческого потенциала

УК-3. Готовность действовать в нестандартных ситуациях, нести социальную и этическую ответственность за принятые решения.

ОПК-3. Способность решать профессиональные задачи на основе философии, математических методов и моделей для управления инновациями, компьютерных технологий в инновационной сфере

ОПК-4. Способность к абстрактному мышлению, анализу, синтезу

ПК-1. Способность разработать план и программу организации инновационной деятельности научно-производственного подразделения, осуществлять технико-экономическое обоснование инновационных проектов и программ

ПК-3. Способностью произвести оценку экономического потенциала инновации, затрат на инновационный проект и осуществление инновационной деятельности в организации.

ПК-4. Способность найти (выбрать) оптимальные решения при создании новой наукоемкой продукции с учетом требований качества, стоимости, сроков исполнения, конкурентоспособности и экологической безопасности

ПК-6. Способность применять теории и методы теоретической и прикладной инноватики, систем и стратегий управления, управления качеством инновационных проектов, выбирать соответствующие методы решения экспериментальных и теоретических задач.

ПК-8. Способностью выполнить анализ результатов научного эксперимента (исследования) с использованием соответствующих методов и инструментов обработки, интерпретировать, представлять и применять полученные результаты в практической деятельности.

ПК-10. Способностью критически анализировать современные проблемы инноватики с учётом экономического, социального, экологического и технологического аспектов жизнедеятельности человека.

В результате изучения дисциплины студент должен:

Знать: основные технологии анализа данных: OLAP, KDD и Data Mining.

Уметь: строить автоматизированные модели анализа данных.

Владеть: навыками анализа данных на примере решения задач сегментации, классификации, прогнозирования.

Иметь опыт построения автоматизированных аналитических моделей

Форма обучения: очная.

IV. Формы контроля

Контроль за освоением дисциплины осуществляется в каждом дисциплинарном разделе отдельно.

Рубежный контроль: контрольная работа, тестирование по отдельным разделам дисциплины.

Итоговая аттестация в 3 семестре – зачет в письменной форме.

Результаты текущего контроля и итоговой аттестации формируют рейтинговую оценку работы обучающегося. Распределение баллов по отдельным видам работ в процессе освоения дисциплины «Методы анализа данных» осуществляется в соответствии с Приложением 1.

V. Объём дисциплины и виды учебной работы

Объем курса – 72 часов, 2 зачетные единицы, в том числе 24 часов – аудиторная нагрузка, из которых 5 часов – лекции, 19 часов – семинары, 48 часов – самостоятельная работа студентов. Изучается на 2 курсе (3 семестр), итоговая форма отчетности – **зачет.**

Вид учебной работы	Всего часов
Контактные занятия (всего)	24
В том числе:	-
Лекции	5
Практические занятия (ПЗ)	-
Семинары (С)	19
Лабораторные работы (ЛР)	-
Самостоятельная работа (всего)	48
В том числе:	-
Домашние задания	12
Реферат	6
Подготовка к опросу	8
Подготовка к тестированию	8
Подготовка к контрольной работе	10
Вид промежуточной аттестации Зачет	 4
Общая трудоемкость (часы) Зачетные единицы	72 2

VI. Структура и содержание дисциплины

п/п	Раздел	Содержание (темы)
1	Введение	Основы анализа данных. Методология построения моделей сложных систем. Модель черного ящика. Основные этапы моделирования. Методика анализа данных.
2	Методы интеллектуального анализа данных	Определения OLAP, Data Mining, KDD и взаимосвязи между ними. OLAP. Аналитическая отчетность и многомерное

		представление данных. Хранилище данных. Измерения и факты. Основные операции над кубом данных. Типы задач, решаемые методами Data Mining: классификация, кластеризация, регрессия, ассоциация, поиск последовательных шаблонов. Алгоритмы, получившие наибольшее распространение для каждого типа задач: самоорганизующиеся карты, деревья решений, линейная регрессия, нейронные сети, ассоциативные правила. Практикум: – построение аналитической отчетности; – построение регрессионной прогнозной модели спроса.
3	Примеры практических приложений в экономике и бизнесе	Задача сегментации клиентов фирмы. Оценка кредитоспособности физических лиц. Задача прогнозирования потребности в продукции. Задачи прогнозирования продаж, поступления финансовых средств и др. Примеры комбинации методов Data Mining. Практикум: – построение скоринговой модели кредитования (деревья решений); – построение нейросетевой прогнозной модели спроса.
4	Подготовка данных и интерпретация результатов	Этапы подготовки данных. Выдвижение гипотез. Методы сбора и систематизации фактов. Методы проведения экспертиз для выявления наиболее значимых факторов. Понятия парциальной и комплексной обработки. Анализ качества полученных моделей. Практикум: – построение сценария предобработки данных в программе Deductor.
5	Практические аспекты	Критерии выбора аналитических платформ и пакетов Data Mining. Основные этапы внедрения систем анализа данных. Категории пользователей аналитических систем; требования, предъявляемые к каждой группе пользователей. Способы снижения рисков проектов Data Mining.

Разделы дисциплин и виды занятий (ак. часы)

п/п	Наименование раздела дисциплины	Лекция	Практи-ческие занятия	Лабора-торные занятия	Семи нар	СРС	Формы текущего контроля
1	**Введение**	1	-	-	3	4	**Домашнее задание Реферат Опрос Тест КР**
2	**Методы интеллектуального анализа данных**	1	-	-	4	8	**Домашнее задание Реферат Опрос Тест КР**
3	**Примеры практических приложений в экономике и бизнесе**	1	-	-	4	8	**Домашнее задание Реферат Опрос Тест КР**
4	**Подготовка данных и интерпретация результатов**	1	-	-	4	12	**Домашнее задание Реферат Опрос Тест КР**
5	**Практические аспекты**	1	-	-	4	12	**Домашнее задание Реферат Опрос Тест КР**
	Промежуточная аттестация (зачет)					4	
	Итого	**5**	-	-	**19**	**48**	

Разделы дисциплины и междисциплинарные связи

№ п/п	Наименование обеспечиваемых (последующих) дисциплин	№ № разделов данной дисциплины, необходимых для изучения обеспечиваемых (последующих) дисциплин				
		1	2	3	4	5
1.	Сенсорные сети и нейрокоммуникации	-	+	+	-	+

VII. Образовательные технологии

В процессе освоения дисциплины «Методы анализа данных» используются следующие образовательные технологии:

1. Стандартные методы обучения:

- лекции;
- семинары;
- письменные или устные домашние задания;
- консультации преподавателей;
- самостоятельная работа студентов, в которую входит освоение теоретического материала, подготовка к семинарам, выполнение указанных выше письменных работ.

2. Методы обучения с применением интерактивных форм образовательных технологий:

- интерактивные лекции;
- анализ деловых ситуаций на основе кейс-метода и имитационных моделей;
- круглые столы;
- обсуждение подготовленных студентами рефератов;
- групповые дискуссии и проекты;
- обсуждение результатов работы студенческих исследовательских групп.

VIII. Учебно-методическое, информационное и материально-техническое обеспечение дисциплины

Учебно-методическое и информационное обеспечение дисциплины

а) Основная литература:

1. Айзек, М.П. Вычисления, графики и анализ данных в Excel 2013. [Текст] / М.П. Айзек. – СПб.: Наука и техника, 2015. – 416 с.
2. Горяинова, Е.Р. Прикладные методы анализа статистических данных: учебное пособие [Текст] / Е.Р. Горяинова, А.Р. Панков, Е.Н. Платонов. – М.: ИД ГУ ВШЭ, 2018. – 310 с.
3. Кибзун, А.И. Теория вероятностей и математическая статистика. Базовый курс с примерами и задачами [Текст] / А.И. Кибзун, Е.Р. Горяинова, А.В. Наумов. – М.: Огни, 2014. – 232 с.
4. Мхитарян, В.С., Шишов В.Ф., Козлов А.Ю. Анализ данных в MS Excel: учебное пособие. [Текст] / В.С. Мхитарян. – М.: КУРС, 2018. – 368 с.
5. Паклин, Н. Орешков, В. Бизнес аналитика. От данных к знаниям [Текст] / Н. Палкин, В. Орешков. – М.: Питер, 2013. – 704 с.

б) Дополнительная литература:

6. Волкова, В.Н., Горелова Г.В., Козлов В.Н. и др. Моделирование систем и процессов: учебник для акад. Бакалавриата [Текст] / В.Н. Волкова, Г.В. Горелова, В.Н. Козлов. – М.: Юрайт, 2016, – 596 с.
7. Корячко, В., Бакулева, М., Орешков В. Интеллектуальные системы и нечеткая логика [Текст] / В. Корячко, М. Бакулева. – М.: КУРС Инфра-М, 2017, 352 с.

Перечень лицензионного программного обеспечения

MS Office

Перечень профессиональных баз данных и информационных справочных систем

1. ЭБС «Юрайт» [раздел «ВАША ПОДПИСКА: учебники и учебные пособия издательства «Юрайт»]: сайт. – URL: https://www.biblio-online.ru/catalog/
2. ЭБС издательства «Лань» [учебные, научные издания, первоисточники, художественные произведения различных издательств; журналы] : сайт. – URL: http://e.lanbook.com
3. https://www.econ.msu.ru/elibrary – электронная библиотека Экономического факультета МГУ.
4. http://nbmgu.ru/ – Научная библиотека МГУ имени М.В. Ломоносова

Перечень ресурсов информационно-телекоммуникационной сети «Интернет»

1. http://www.olap.ru – информационный портал, посвященный технологиям интерактивной аналитической обработки
2. http://neiroset.ru – Информационный портал «Нейросеть.ру»
3. https://neuralnet.info – информационный портал о нейросетях
4. https://proglib.io/p/neural-nets-guide – информационный портал «Библиотека программиста»

Рекомендуемые обучающие, справочно-информационные, контролирующие и прочие компьютерные программы, используемые при изучении дисциплины

№ п/п	Название рекомендуемых по разделам и темам программы технических и компьютерных средств обучения	Номера тем
1.	MS PowerPoint	1-5
2.	MS Excel	1-5
3.	Deductor 5.3	1-5

Методические указания для обучающихся по освоению дисциплины

В процессе изучения курса обучающиеся обязаны соблюдать дисциплину, вовремя приходить на занятия, делать домашние задания, осуществлять подготовку к семинарам и контрольным работам, проявлять активность на занятиях.

При этом важное значение имеет самостоятельная работа, которая направлена на формирование у учащегося умений и навыков правильного оформления конспекта и работы с ним, работы с литературой и электронными источниками информации, её анализа, синтеза и обобщения. Для проведения самостоятельной работы обучающимся предоставляется список учебно-методической литературы.

Материально-техническое обеспечение дисциплины

Для проведения образовательного процесса необходима аудитория, оборудованная компьютером и проектором, необходимыми для демонстрации презентаций. Обязательное программное обеспечение – MS Office.

IX. ФОНД ОЦЕНОЧНЫХ СРЕДСТВ

Темы курсовых работ

Курсовая работа по дисциплине «Методы анализа данных» не предусмотрена.

Темы рефератов

1. Методология анализа сложных систем.
2. Основные концепции построения хранилищ данных.
3. Построение автоматизированных систем предобработки данных.
4. Алгоритмы и технологии DataMining.
5. Построение корпоративных автоматизированных аналитических систем на основе методология KDD.
6. Обзор алгоритмов построения деревьев решений.
7. Математические основы нейросетевых технологий.
8. Методология построения регрессионных моделей.
9. Применение технологии деревьев решений для оценки кредитоспособности физических лиц.
10. Применение нейросетевых технологий для оценки кредитоспособности физических лиц.
11. Ассоциативные правила, как инструмент повышения прибыльности в розничной и оптовой торговле.
12. Задачи классификации, как инструмент повышения эффективности бизнеса.
13. Методология построения и верификации прогнозных моделей.
14. Основные методы прогнозирования.
15. Методы оценки качества прогнозных моделей.

Вопросы для текущего контроля и самостоятельной работы студентов

1. Методыпостроения моделей сложных систем.
2. Модель черного ящика.
3. Основные этапы моделирования.
4. Методика анализа данных.
5. Определения OLAP, Data Mining, KDD и взаимосвязи между ними.
6. Особенности OLAP.
7. Аналитическая отчетность и многомерное представление данных.
8. Хранилище данных.
9. Измерения и факты.
10. Основные операции над кубом данных.
11. Типы задач, решаемые методами Data Mining.
12. Алгоритмы, получившие наибольшее распространение для каждого типа задач: самоорганизующиеся карты, деревья решений, линейная регрессия, нейронные сети, ассоциативные правила.
13. Этапы подготовки данных.
14. Выдвижение гипотез.
15. Методы сбора и систематизации фактов.
16. Методы проведения экспертиз для выявления наиболее значимых факторов.
17. Понятия парциальной и комплексной обработки.
18. Анализ качества полученных моделей.
19. Критерии выбора аналитических платформ и пакетов Data Mining.

20. Основные этапы внедрения систем анализа данных.
21. Категории пользователей аналитических систем.
22. Способы снижения рисков проектов Data Mining.

Пример теста для контроля знаний обучающихся

1. Какие из нижеперечисленных признаков являются количественными:
а) плотность населения
б) уровень занятости населения
в) среднедушевой доход
г) пол человека
д) возраст
е) уровень образования (начальное, среднее, высшее)

2. Какие из нижеперечисленных признаков являются дискретными:
а) денежные доходы населения
б) число детей в семье
в) прибыль предприятия
г) пол человека
д) тарифный разряд

3. Типологические группировки применяются для:
а) характеристики структурных сдвигов
б) характеристики взаимосвязей между отдельными признаками
в) разделения совокупности на качественно однородные типы
г) характеристики структуры совокупности

4. Структурные группировки применяются для:
а) разделения совокупности на качественно однородные типы
б) характеристики взаимосвязей между отдельными признаками
в) характеристики структуры совокупности

5. Часть зависимой переменной в регрессионной модели, которая может быть объяснена значением регрессора:
а) случайное возмущение;
б) отклик;
в) уравнение регрессии;
г) остаток.

6. Гипотеза является сложной, если:
а) она состоит из конечного числа простых гипотез;
б) она состоит из бесконечного числа простых гипотез;
в) Содержит только одно предположение.

7. Коррелированность возмущений с различными номерами называется:
а) гомоскедастичностью;
б) гетероскедастичностью;
в) автокорреляцией.

8. Критической областью называют:
а) совокупность значений критерия, при которых нулевую гипотезу принимают;

б) совокупность значений критерия, при которых нулевую гипотезу отвергают.

9. Причины гетероскедастичности (множественный выбор):
а) исследование неоднородных объектов;
б) характер наблюдений;
в) ошибки спецификации;
г) ошибки измерений.

10. Под мультиколлинеарностью понимается линейная зависимость (единичный выбор):
а) зависимой переменной с одним или несколькими регрессорами;
б) двух, или нескольких регрессоров;
в) зависимой переменной с возмущением;
г) регрессоров с возмущением.

11. С увеличением объема выборки длина доверительного интервала прогнозируемого значения зависимой переменной (единичный выбор):
а) увеличивается;
б) уменьшается;
в) не меняется.

12. Как переводится DATA MINING?
a) “добыча” или “раскопка данных”.
b) “значение”.
c) “хранение”.
d) “перечисление данных”.

13. Какое требование к переработке информации не верно?
a) Данные имеют неограниченный объем.
b) Данные являются разнородными.
c) Результаты должны быть конкретны и понятны.
d) Инструменты для обработки сырых данных должны быть сложны в использовании.

14. Какая концепция положена в основу современной технологии Data Mining?
a) Концепция естествознания.
b) Концепция управления.
c) Концепция шаблонов (паттернов).
d) Концепция становления.

15. Сколько выделяют стандартных типов закономерностей?
a. 4.
b. 5.
c. 6.
d. 7.

16. Если несколько событий связаны друг с другом, то это…
a) Ассоциация.
b) Последовательность.
c) Классификация.
d) Кластеризация.

17. Основой для каких систем служит историческая информация, хранящаяся в БД в виде временных рядов?

a) Классификации.
b) Последовательности.
c) Прогнозирования.
d) Ассоциации.

18. Какую иерархическую структуру создают деревья решений?

a) "ЕСЛИ... ТО...".
b) "НИ... НИ...".
c) "КОГДА... ТО...".
d) "... НИКОГДА...".

19. С чем связано направление эволюционного программирования?

a) Постановка вопроса вида "значение параметра А больше х?".
b) Использование метода "ближайшего соседа".
c) Подача значений входных параметров, на основе которых нужно принимать какие-то решения, прогнозировать развитие ситуации.
d) Поиск зависимости целевых переменных от остальных в форме функций какого-то определенного вида.

20. Что называют хромосомами в генетических алгоритмах?

a) Кодировку исходных логических закономерностей в базе данных.
b) Направление эволюционного программирования.
c) Большой класс систем.
d) Набор закономерностей.

Вопросы к зачету

1. Какие значения множества данных могут рассматриваться как аномальные?
2. Каково ожидаемое влияние аномальных значений на результаты анализа?
3. Как применяется визуальный анализ для выявления аномалий в одномерных и двумерных множествах данных?
4. Всегда ли аномальные значения являются нежелательными в данных?
5. Какие методы корректировки аномальных значений вам известны?
6. Что такое ETL-системы?
7. Чем вызвана необходимость использования ETL?
8. Каковы основные этапы процесса ETL и решаемые им задачи?
9. Зачем необходимо выполнять очистку данных?
10. Что такое поток данных с точки зрения ETL?
11. В чем заключается процесс снижения размерности исходных данных?
12. По каким причинам при подготовке данных может потребоваться сокращение их размерности?
13. Каких преимуществ можно добиться путем сокращения размерности данных?
14. По каким направлениям может производится сокращение размерности?
15. В каких режимах может производится сокращение размерности данных?
16. Какими свойствами должны обладать алгоритмы сокращения размерности данных для эффективной работы?
17. В чем отличие трансформации данных от предобработки и очистки?
18. С какими проблемами связана необходимость трансформации данных?

19. Каковы цели трансформации данных в аналитическом приложении?
20. Почему, несмотря на то, что трансформация данных производится на этапе консолидации данных, её необходимо применять и в аналитическом приложении?
21. Каково назначение систем OLTP и СППР?
22. Отличия OLTP-систем и СППР?
23. Какие факторы стимулировали появление концепции ХД?
24. В чем заключаются основные различия между ХД и обычными базами данных?
25. Какие функциональные требования предъявляются к ХД?
26. Каковы предпосылки появления концепции ВХД?
27. Каковы принципы функционирования ВХД?
28. В чем состоят преимущества и недостатки ВХД по сравнению с многомерными и реляционными хранилищами?
29. Сохранится ли информация в ВХД после перезагрузки компьютера?
30. Обеспечивают ли ВХД поддержку исторических данных?
31. Что понимается в данных под пропущенным значением?
32. Почему пропущенные значения в анализируемых данных необходимо восстанавливать?
33. Каково происхождение пропусков в данных?
34. Каким требованиям должны удовлетворять алгоритмы восстановления пропущенных значений?
35. В каком случае пропущенные значения можно восстановить вручную?
36. Какие данные с точки зрения восстановления пропущенных значений являются упорядоченными, а какие – неупорядоченными?
37. В чем различие подходов между восстановлением пропусков в упорядоченных и неупорядоченных данных?
38. В чем преимущества и недостатки методики подстановки констант вместо пропущенных значений?
39. Как определяются наиболее вероятные значения для подстановки вместо пропусков.
40. Каковы цели этапа загрузки данных в ETL процессе?
41. Каковы основные причины неполной загрузки данных?
42. Что следует делать с записями, которые не попали в ХД в процессе загрузки?
43. Что следует предпринять, если все данные загрузить не удалось?
44. Почему для переноса данных в ХД организуется несколько потоков?
45. Как можно выполнить проверку результатов загрузки данных?
46. Когда возникает необходимость загрузки данных непосредственно из источников (минуя ХД)?
47. Каковы причины отказа от использования ХД и применения непосредственной загрузки анализируемых данных из источников?
48. Какие проблемы порождает непосредственная загрузка данных из источников?
49. Какие основные типы источников данных, из которых чаще всего приходится загружать данные непосредственно, вам известны?
50. Почему загрузка данных из текстовых фалов с разделителями наиболее проблематична?
51. Какие типы источников данных наиболее удобны для непосредственного доступа и почему?
52. Каковы место и роль извлечения данных в общей структуре ETL процесса?
53. Из каких соображений выбираются используемые источники данных?
54. В чем сложность извлечения данных из отдельных структурированных источников (файлов MS Excel, TXT и т.д.).
55. Почему извлечение данных из СУБД является наименее проблематичным?

56. Все ли данные нужно извлекать из источника при пополнении ХД?
57. Как вы видите роль аналитика в процессе организации извлечения данных в рамках ETL.
58. В чем заключается процедура консолидации данных, каковы ее цели?
59. Какие основные виды источников данных вы знаете?
60. Какие задачи решаются при консолидации?
61. Какие причины мешают корректной аналитической обработке и требуют использования методов очистки данных?
62. В чем заключается цель процедуры обогащения данных?
63. Как вы понимаете термин «многомерный куб», «гиперкуб»?
64. Какова роль измерений и фактов в многомерной модели данных?
65. В чем принципиальное отличие многомерной модели от реляционной?
66. Каковы преимущества и недостатки многомерной модели данных.
67. Какие действия применяются к измерениям для извлечения нужной информации из многомерного куба.
68. Чем отличаются понятия «информация» и «данные»? Всегда ли в данных присутствует информация?
69. В чем выражается субъективность информации и объективность данных.
70. В чем заключается процесс обогащения данных и какова его цель?
71. Основные отличия между внутренним и внешним обогащением данных?
72. Каких преимуществ в бизнесе позволяет добиться обогащение анализируемых данных?
73. Какие источники данных во внешнем окружении предприятия могут использоваться для обогащения данных?
74. Что представляют собой дубликаты и противоречия?
75. Какие проблемы при анализе данных могут вызвать дубликаты и противоречия?
76. Всегда ли дубликаты и противоречия являются следствием ошибок и их необходимо удалять?
77. В каких случаях обработку дубликатов и противоречий не производят совсем?
78. Когда требуется объединение дублирующихся и противоречивых записей?
79. Какие принципы лежат в основе построения ХД?
80. Каковы цели использования концепции ХД в процессе поддержки принятия решений и интеллектуального анализа данных?
81. Зачем выполняется агрегирование данных?
82. Что такое метаданные и какова их роль в процессе функционирования ХД?
83. Какие виды метаданных вам известны?
84. Какие ХД называются кросс-платформенными?
85. Какие архитектуры ХД вам известны?
86. Как вы понимаете термин «качество данных»?
87. Почему оценке качества данных уделяют большое внимание на всех этапах подготовки данных к анализу?
88. Каковы основные цели оценки качества данных?
89. Какие способы реализации процесса оценки качества данных вам известны?
90. Какие выводы о качестве данных могут быть сделаны по результатам его оценки?
91. Какие предположения можно сделать о качестве данных, зная их происхождение и методику сбора?
92. Почему при реализации аналитических проектов выполняется два этапа очистки данных – перед загрузкой их в ХД и в аналитической системе?
93. Какие значения данных называются фиктивными, каково их происхождение?
94. Какие нарушения и ошибки в данных называются критичными?

95. Какие ошибки являются наиболее типичными для отдельных ячеек таблиц?
96. Какие записи в таблице являются противоречивыми?
97. Почему предобработка данных, загруженных в аналитическое приложение, необходима независимо от уровня очистки данных на предыдущих этапах их жизненного цикла (в OLTP-системе, ETL-процессе или ХД)?
98. Почему в данных, поступающих для анализа в аналитическое приложение, все еще присутствуют проблемы, связанные с качеством данных?
99. Почему мониторинг качества данных и борьба за качество проводятся на всех этапах процесса сбора, консолидации и анализа данных, а не на каком-то одном этапе?
100. Почему выявление и устранение проблем в данных удобнее производить в местах их появления?
101. Почему некоторые задачи повышения качества данных целесообразно решать только в процессе их предобработки в аналитическом приложении? Приведите примеры таких задач.
102. С какой целью в аналитическом приложении производится снижение размерности входных данных и устранение незначащих признаков?
103. Какие преимущества дает непосредственное участие аналитика в процессе подготовки данных к анализу?
104. Почему окончательная предобработка данных может быть выполнена только в аналитическом приложении и только с учетом требований конкретной задачи анализа?
105. Какие виды закономерностей в рядах данных вам известны?
106. Почему ряды данных разделяют на детерминированную и случайную составляющую?
107. Как визуально проявляется шум в рядах данных, как он влияет на распознаваемость закономерностей?
108. Почему шум в данных носит случайный характер?
109. В чем отличие природы шумов в рядах анализируемых данных и технических системах?
110. Как может повлиять на качество данных непродуманная процедура подавления шума в них?
111. Каковы место и цель этапа преобразования данных в ETL-процессе?
112. Какие типичные операции выполняются при преобразовании данных в ETL?
113. Какие данные называют детализированными, а какие – агрегированными, в чем заключается процедура агрегирования данных, какова её цель?
114. Нужно ли агрегировать все данные, загружаемые в ХД?
115. Зачем может потребоваться перевод значений при преобразовании данных в ETL?
116. С какой целью в процессе преобразования данных могут создаваться новые поля?
117. Каковы цели очистки данных в ETL?
118. В какой форме хранятся данные в РХД?
119. Как соотносятся таблицы фактов и таблицы данных в РХД?
120. Чем отличаются схемы «звезда» и «снежинка»?
121. Какие преимущества дает схема «снежинка» при анализе данных с иерархией измерений?
122. С какой целью производится сокращение числа значений признаков и записей в исходной выборке данных?
123. По какому принципу производится сокращение числа значений признаков?
124. Что такое точки среза и как они выбираются?
125. Какова цель сокращения количества записей исходного множества данных, и из каких соображений оно выбирается?
126. В чем заключается принцип обучения окнами и в чем его преимущество?
127. Какое преобразование лежит в основе спектральной обработки данных?

128. Из чего состоит частотный спектр?
129. Почему при подавлении высокочастотных спектральных составляющих восстановленные данные оказываются сглаженными, а шум в них уменьшается?
130. Что такое полоса пропускания фильтра, на что она влияет, из каких соображений выбирается?
131. Почему если полоса пропускания равна 0, ряд данных после фильтрации представляет собой прямую линию?
132. Каковы основные этапы процесса частотной фильтрации?
133. На чем основан принцип разделения шумовой и детерминированных составляющих ряда данных?
134. Почему при неправильном выборе порога для удаления шума может быть потеряна часть полезных данных?
135. Что такое пространственный фильтр (маска)?
136. Почему количество коэффициентов пространственного фильтра должно быть нечетным?
137. Какие эффекты возникают на границах ряда в процессе пространственной фильтрации?
138. Каков принцип работы пространственного фильтра?
139. Что такое отклик пространственного фильтра?
140. Какие аспекты качества данных можно оценить с помощью профайлинга?
141. Какие приемы можно использовать для визуальной оценки качества данных с помощью таблиц?
142. Какие проблемы в данных можно выявить с помощью графиков и диаграмм?
143. Какие ошибки в данных являются трудноформализуемыми?
144. Какие данные называются временными рядами? Приведите примеры.
145. Какие временные ряды называются одномерными и многомерными?
146. Для чего ряды данных преобразуют в табличную форму?
147. В чем заключается механизм преобразования данных скользящим окном и для чего оно используется?
148. Что такое глубина погружения и горизонт прогноза? Как они выбираются?
149. В чем заключается трансформация даты, и каковы её цели?
150. Какие виды преобразования дат вам известны?
151. Почему при подготовке данных к анализу требуется фильтрация и в чем она заключается?
152. Какую роль в фильтрации играют значения и условия?
153. Какие условия фильтрации для числовых данных вам известны?
154. Почему сокращение признаков является самым эффективным направлением снижения размерности исходных данных?
155. В чем вы видите отличие незначащих признаков от избыточных?
156. В чем заключаются задачи отбора и композиции признаков?
157. Как производится отбор признаков на основе их статистических характеристик?
158. Что такое энтропия и как она используется для отбора признаков?
159. В чем заключается основной принцип метода главных компонент?
160. Какую роль играет порог значимости в методе главных компонент?
161. Какую роль играют отношения между условиями при фильтрации данных?

Зачет проводится путем ответа на вопросы и выполнения заданий на компьютере (2 вопроса и 1 задание).

Примерные задачи к зачету

1) В файле **P1_1.XLS** содержатся данные финансовых расчетов с потребителями компании за последние 4 месяца. Каждая строка в приведенной базе данных содержит информацию об одной операции отгрузки товара, а именно, имя потребителя, месяц, категорию отгрузки, сумма отгрузки, сумма поступившей оплаты.

А) создайте сводную таблицу для вычисления количества операций отгрузки по каждому потребителю и по каждой категории за все 4 месяца.

Б) создайте сводную таблицу для вычисления общих сумм поставок по каждому потребителю за каждый месяц. Используя полученные данные, постройте соответствующие временные ряды для каждого потребителя.

В) постройте гистограмму (одну) для поступивших оплат только для двух категорий отгрузки «Оборудование» и «Материалы».

2) Основываясь на данных о продажах из файла **«Продажи»** и других сопутствующих справочниках сформировать сценарии ежедневных отчетов по долевому объему продаж в каждом из отделов по группам товаров за последние 10 дней. В каждом из отчетов должен присутствовать полный перечень товарных групп, упорядоченных по возрастанию кода.

Примеры контрольной работы

В 1

Необходимо построить сценарий в аналитической платформе Deductor, который ежедневно формирует отчет, показывающий 10 лидеров продаж по сумме продаж по итогам последних 10 дней. Результат необходимо визуализировать в виде упорядоченной по убыванию столбчатой диаграммы. Необходимо произвести визуализацию в двух видах и сформировать соответствующие отчеты, а именно, с отображением сумм продаж, с отображением долей продаж в общей сумме продаж этих товаров. Исходные данные находятся в файле «Продажи», характеристики товаров в файле «Товары».

В 2

В файле представлены некоторые исходные данные, а именно столбцы «Вход1», «Вход2», «Выход». В данных присутствуют дубликаты и противоречия. Необходимо построить сценария в аналитической платформе Deductor, который исключает дубликаты и противоречия.

Данные представлены в файле «Дубликаты и противоречия».

В3

Коммерческий директор хочет иметь информацию о последних тенденциях в изменении суммовых объемов продаж по товарным группам. Для этого предлагается вычислить относительное изменение объемов продаж за последние 10 дней по отношению к объемам продаж за предыдущие 10 дней. Если это изменение менее -0,3, то товарной группе присваиваем категорию «Провал», от -0,3 до -0,1 – «Падение», от -0,1 до 0 – «Уменьшение», от 0 до 0,1 – «Увеличение», от 0,1 до 0,3 – «Подъем», свыше 0,3 – «Взлет». Необходимо построить соответствующий сценарий в аналитической платформе Deductor. Необходимые данные находятся в файлах «Продажи», «Товары», «Товарные группы».

Пример итоговой контрольной работы

1. В чем состоит цель поиска ассоциативных правил. Дайте определения категориям «транзакция», «поддержка набора», «поддержка правила», «достоверность правила».

2. Алгоритм поиска ассоциативных правил Apriori: этапы работы. В чем состоит отличие алгоритма Apriori от алгоритмов AIS и SETM.
3. Организационные факторы при внедрении Data Mining в деятельность компании.
4. Человеческие факторы при внедрении Data Mining в деятельность компании. Основные роли специалистов в процессах Data Mining.
5. Перечислите стандарты методологии Data Mining. В чем состоят их особенности.

Примеры домашнего задания

1) Вы работаете в небольшой туристической фирме и планируете массовую рассылку рекламного буклета. Ваши средства ограничены, поэтому вы хотите послать ее тем, кто готов тратить на путешествия и отдых в большей степени. В файле **P5_1.XLS** содержатся данные о случайной выборке клиентов размером 925 (пол, возраст, суммы, затраченные на путешествия и отдых в предыдущем году). Используйте данные, чтобы понять, насколько пол и возраст влияют на объем затрат. Сформулируйте обоснованные рекомендации относительно контингента для рассылки рекламной брошюры.
2) Основываясь на данных о продажах из файла **«Продажи»** и других сопутствующих справочниках сформировать сценарий ежедневного отчета по 5 лидерам товарных групп по суммам продаж за последние 15 дней. Отчет визуализировать с помощью столбчатых диаграмм с информацией о сумме продаж и названии товарных групп.

Приложение 1

СИСТЕМА РЕЙТИНГОВОЙ ОЦЕНКИ И КОНТРОЛЯ ЗНАНИЙ СТУДЕНТОВ

№ п/п	СТРУКТУРА	Баллы по каждому модулю
1.	Оценка за активное участие в учебном процессе и посещение занятий: Всех занятий Не менее 75% Не менее 50% Не менее 25% Итого:	 5 4 3 2 до 5
2.	устный опрос в форме собеседования (УО-1) письменный опрос в виде теста (ПР-1) письменная контрольная работа (ПР-2) письменная работа в форме реферата (ПР-4) Итого:	15 10 10 10 45
3.	Зачет	50
	ВСЕГО:	100

Пересчет на 5 балльную систему

2 (неудовлетворительно)	3 (удовлетворительно)	4 (хорошо)	5 (отлично)
< 50	**50-64**	**65-84**	**85-100**

Язык преподавания: русский.

Автор программы: Косоруков Олег Анатольевич, д.т.н., профессор, профессор Высшей школы управления и инноваций МГУ имени М.В. Ломоносова.

Преподаватель (преподаватели) программы: Косоруков Олег Анатольевич, д.т.н., профессор, профессор Высшей школы управления и инноваций МГУ имени М.В. Ломоносова.