Автономная некоммерческая организация высшего образования
«Институт социальных наук»

УТВЕРЖДАЮ

Ректор Бианкина А.О.

« 01 » июня 2022 г.

РАБОЧАЯ ПРОГРАММА ДИСЦИПЛИНЫ

Анализ данных

для студентов направления подготовки

38.03.05 Бизнес-информатика

Профиль

«Бизнес-аналитика»

Квалификация (степень) выпускника – бакалавр

Форма обучения: очная

Москва

Рабочая программа дисциплины «Анализ данных»

Направление подготовки 38.03.05 Бизнес –информатика

Составитель

Программа рассмотрена и согласована на заседании кафедры экономики и управления (протокол № от « » _____20 г.)

Заведующий кафедрой _____

(подпись)

СОДЕРЖАНИЕ

1. Перечень планируемых результатов обучения по дисциплине, соотнесенных с планируемыми результатами освоения программы

1.1. Дисциплина «Анализ данных» обеспечивает овладение следующими компетенциями:

Таблица 1.1

Код компетенции	Наименование компетенции	Код этапа освоения компетенции	Наименование этапа освоения компетенции
ДПК-31	Сбор, обработка и анализ больших данных с использованием существующей в организации методологической и технологической инфраструктуры	ДПК-31.1	Способность планировать и проводить аналитические работы, использовать математический аппарат, информационные технологии, современные языки статистической обработки и программные средства решения эконометрических задач и задач анализа данных.

В результате освоения дисциплины у студентов должны быть сформированы:

Таблица 1.2

ОТФ/ТФ (при наличии профстандарта)/ профессиональные действия	Код этапа освоения компетенции	Результаты обучения
Выполнение работ и управление работами по созданию (модификации) и сопровождению ИС, автоматизирующих задачи организационного управления и бизнес-процессы/ Разработка модели бизнес-процессов заказчика	ДПК -31.1	на уровне знаний: – Теоретические и прикладные вопросы теории нечетких множеств, анализа данных; – основные понятия и основные методы, многомерной математической статистики; – современные ИКТ и ИС, их возможности; – средства бизнес-аналитики, современные языки статистической обработки (R, Python) и графические платформы; – основные понятия и основные методы теории анализа данных, интеллектуальной обработки данных, теории нечетких множеств, теории прогнозирования, эконометрики, многомерной математической статистики – технологии анализа данных: статистический анализ, семантический анализ, анализ изображений, машинное обучение, методы сравнения средних, частотный анализ, анализ соответствий, кластерный анализ, дискриминантный анализ, факторный анализ, деревья классификации, моделирование структурными уравнениями.
		на уровне умений: - обрабатывать эмпирические и экспериментальные данные, осуществлять предобработку и очистку данных, выполнять разведывательный анализ; - использовать математические и инструментальные средства для анализа данных в процессе эконометрического моделирования, предикативной аналитики, сбора, обработки и анализа больших данных; - Программировать на языках статистической обработки, ориентированных на работу с большими данными: для статистической обработки данных и работы с графикой, для работы с разрозненными фрагментами данных в больших массивах, для работы с базами структурированных и неструктурированных данных;

		- оценивать качество решения задач сбора, обработки и анализа больших данных с использованием существующей в организации методологической и технологической инфраструктуры; - Проводить сравнительный анализ методов и инструментальных средств анализа данных.

2. Объем и место дисциплины в структуре ОП ВО

Объем дисциплины

Общая трудоемкость дисциплины составляет 4 зачетных единиц 144 академических часов.

Таблица 2

Вид работы	Трудоемкость (акад/астр.часы)
Общая трудоемкость	144/108
Контактная работа с преподавателем	52/39
Лекции	24/18
Практические занятия	28/21
Самостоятельная работа	56/42
Контроль	36/27
Формы текущего контроля	Задания, Практическое контрольное задание, выполнение расчетного задания
Форма промежуточной аттестации	Экзамен

Место дисциплины в структуре ОП ВО

Дисциплина реализуется с применением дистанционных образовательных технологий (*далее - ДОТ*).

Дисциплина Б1.В.11 «Анализ данных» относится к вариативной части учебного плана по направлению «Бизнес-информатика» 38.03.05. Преподавание дисциплины «Анализ данных» основано на дисциплинах – Б1.Б.07.03 «Теория вероятностей и математическая статистика», Б1.Б.07.01 - «Математический анализ», Б1.В.21 «Дифференциальные и разностные уравнения». В свою очередь она создаёт необходимые предпосылки для освоения программ таких дисциплин, как Б1.В.03 «Моделирование бизнес-процессов», Б1.В.10 «Архитектура предприятия», Б1.В.01 «Нечеткая логика и нейронные сети», изучаемой с дисциплиной одновременно.

Дисциплина изучается в 5-м семестре 3-го курса.

Формой промежуточной аттестации в соответствии с учебным планом является экзамен.

3. Содержание и структура дисциплины

Таблица 3

№ п/п	Наименование тем	Объем дисциплины, час.						Форма текущего контроля успеваемости, промежуточной аттестации
		Всего	Контактная работа обучающихся с преподавателем по видам учебных занятий				СР	
			Л	ЛР	ПЗ	КСР		
Тема 1	Основы анализа	14	4		4		6	РГЗ

	данных							
Тема 2	Предобработка и очистка данных	22	4		6		12	ДЗ
Тема 3	Кластерный анализ	22	4		4		14	ДЗ, РГЗ
Тема 4.	Анализ взаимосвязей между переменными. Ассоциативные правила	18	4		4		10	ДЗ
Тема 5	Задачи классификации. Деревья решений	32	8		10		14	ДЗ, ПКЗ
Контроль		36/27						
Промежуточная аттестация						2*		Экзамен
Всего:		144/108	24/18		28/19,5		56/45	

2* - консультация (не входит в общий объем дисциплины)
ДЗ – Кейс-задание,
РГЗ- расчетно- графическая работа
ПКЗ - контрольные работы,
Э – экзамен

3.Содержание дисциплины

Тема 1. Основы анализа данных

Введение. Понятие анализа данных. Задачи систем поддержки принятия решений. OLTP и OLAP-системы. Принципы построения информационных хранилищ. Модели информационных хранилищ. Многомерная модель данных. Правила Кодда. Размерностные модели. MOLAP, ROLAP, HOLAP- системы. Витрины данных. ETL (Extracting Transformatingand Loading) – средство извлечения, обработки и загрузки данных. Добыча данных. Добыча данных в управлении качеством. Data Mining. Стандарты Data Mining. Стандарт CWM, CRISP, PMML. Жизненный цикл процесса анализа данных. Классификация методов Data Mining. Модели Data Mining. Понятие данные и знания. Процесс обнаружения знаний. Классификация задач Data Mining. Методы анализа данных. Разведочный анализ данных. Очистка и фильтрация данных. Статистические диаграммы. «Ящичные» диаграммы. Диаграммы «ствол-листья». Задачи классификации и регрессии. Использование статистических пакетов для интеллектуального анализа данных. Понятие бизнес-аналитики. Средства бизнес-аналитики. Средства легкой бизнес-аналитики. Qlik View, Qlik Sence, Power BI.

Общая характеристика языка R. Графические средства языка.

Тема 2. Предобработка и очистка данных

Методология KDD. Задачи предобработки данных. Технология ETL. Просмотр данных. Очистка данных. Оценка качества данных. Заполнение пропущенных данных. Аномальные и предельные данные. Использование ящичной диаграммы. Выявление дубликатов и противоречий. Корреляционный анализ. Использование факторного анализа при предобработке данных. Трансформация данных. Квантование. Сэмплинг. Группировка данных. Решение задач предобработки и очистки данных в R (Python).

Тема 3. Кластерный анализ

Постановка задач кластерного анализа. Определение кластера. Параметры кластера. Меры близости. Метрики кластерного анализа. Базовые алгоритмы кластеризации. Иерархическая кластеризация. Дендограммы. Метод К-средних. Профили кластеров. Взаимосвязь кластерного и регрессионного анализа. Использование пакета Deductor для решения задач кластерного анализа. Кластерный анализ в средствах интеллектуального анализа Microsoft Office (на R, SPSS, Python).

Тема 4. Анализ взаимосвязей между переменными. Ассоциативные правила

Основные положения непараметрической и нечисловой статистики. Таблицы сопряженности. Таблица сопряженности 2x2. Таблицы флагов и заголовков. Непараметрические и нечисловые критерии. Канонический анализ. Корреляционная матрица. Коэффициенты канонической корреляции. Меры избыточности переменных. Задачи ассоциации. Ассоциативные правила. Поддержка и достоверность ассоциативных правил. Лифт. Алгоритмы построения ассоциативных правил. Рекомендации по генерации правил. Алгоритм apriori. Использование пакета Deductor для построения ассоциативных правил.

Тема 5. Задачи классификации. Деревья решений

Формулировка задачи классификации. Классификационный анализ с обучением. Деревья решений. Алгоритмы построения деревьев решений. Классификация критериев разбиений. Критерий Gini. Деревья классификации и их свойства. Типы ветвления. Методы и алгоритмы построения деревьев. Алгоритм CART. Определение прекращения построения дерева классификации. Использование нейронных сетей для решения задач классификации. Карты Кохонена. Логистическая регрессия. Сравнение результатов классификации различными методами.

Примеры алгоритмов построения деревьев решений. Использование статистических пакетов Deductor, SPSS, Excel (R, Python) для построения деревьев решений.

4. Материалы текущего контроля успеваемости обучающихся и фонд оценочных средств промежуточной аттестации по дисциплине

Промежуточная аттестация может проводиться с использованием ДОТ.

4.1. Формы и методы текущего контроля успеваемости обучающихся и промежуточной аттестации.

В ходе реализации дисциплины «Анализ данных» используются следующие методы текущего контроля успеваемости обучающихся:

Таблица 4.1

Тема (раздел)	Формы (методы) текущего контроля успеваемости
Тема 1. Основы анализа данных	Защита расчетно-графического задания
Тема 2. Предобработка и очистка данных	Защита задания
Тема 3. Кластерный анализ	Защита расчетно-графического задания
Тема 4. Анализ взаимосвязей между переменными. Ассоциативные правила	Защита задания, Тестирование
Тема 5. Задачи классификации. Деревья решений	Защита задания, Практическое контрольное задание,

В дисциплине используются следующие активные и интерактивные методы обучения:

-дискуссии в период обсуждения предложенных оценочных материалов;

-выполнение и защита задания и контрольной работы;

-интерактивная работа по решению практических задач на компьютерах в компьютерном классе с текущим обсуждением хода и результатов решения задачи, использованию современных программных средств аналитики, data mining;

-выполнение тестирования;

- методы коллективных обсуждений на занятиях семинарского типа;

- тренинги в решении практических задач, направленных на формирование универсальных и общепрофессиональных компетенций;

Признаками данных методов являются:

- активизация мышления студентов, причем учащийся вынужден быть активным;

- длительное время активности — учащийся работает не эпизодически, а в течение всего учебного процесса. Поэтому данные методы в основном реализуются на занятиях семинарского типа;
- самостоятельность в выработке и поиске решений поставленных задач;
- мотивированность к обучению путем использовать балльно-рейтинговой системы оценивания.

4.1.2. Экзамен проводится с применением следующих методов (средств):

Экзамен проводится в компьютерном классе в устной форме. Во время экзамена проверяется уровень знаний по «Анализу данных», а также уровень умений решать учебные задачи анализа данных с использованием программных приложений. К экзамену студенты должны решить задания по всем темам учебной дисциплины. Результаты решения задач могут быть использованы при решении практической задачи в соответствии с имеемым перечнем задач. Пример задач приведен в программе. При ответе на вопросы студент показывает умение решать практические задачи на примере приложения Deductor.

4. 2. Материалы текущего контроля успеваемости обучающихся.

4.2.1 Кейс-задания

Все задания оформлены в виде excel-файлов.

Кейс-задание 1. Кластерный анализ. Кейс-задание оформлено в файле Excel. Состоит из нескольких задач.

Пример задачи.

Кейс-задание 2. Ассоциативные правила

Построить ассоциативные правила по имеемым транзакциям. Рассчитать характеристики для каждого правила.

Транзакционная база данных	
TID	**Приобретенные покупки**
100	ремень, женская сумка, портмоне
200	женская сумка, косметичка
300	женская сумка, ремень, ключница, портмоне
400	дамский зонт, ключница, косметичка
500	ремень, женская сумка, портмоне, ключница
600	косметичка, портмоне
700	ремень, портфель

Кейс-задание 3. Деревья решений.

Построить дерево решений по данным, приведенным в таблице.

Рейтинг	Возраст	УровеньДохода	Образование
0	35	3000	0
0	25	5000	1
0	31	7000	1
1	56	1000	0
1	62	1100	1
1	49	1500	0

Решить задачу логистической регрессии. Определить качество построенной модели классификации. Решить данную задачу другим методами классификации, реализованными в Deductor Academic. Сравнить результаты решения задачи классификации с помощью таблицы сопряженности.

Рейтинг	Образование, A1	Доход, A2	Возраст, A3
низкий	высшее	малый	35
низкий	среднее	большой	40
высокий	высшее	большой	30

высокий	высшее	большой	30
низкий	среднее	малый	30
высокий	высшее	малый	35
высокий	высшее	большой	45
высокий	высшее	большой	35

4.2.2. Практическое контрольное задание

Практическое контрольное задание включает пять задач. Шаблоны контрольной работы размещены в файле Excel. К тематике задач относятся: задача очистки данных, иерархическая задача кластерного анализа, решение задачи кластерного анализа методов к-средних, построение ассоциативных правил, построение дерева решений.

Пример задачи. Построить дендрограмму, используя Евклидово расстояние и метод "дальнего соседа". Перед построением кластеров выполнить стандартизацию значений атрибутов

Номер объекта	x1	x2
1	3,00	10,00
2	4,00	11,00
3	6,00	10,00
4	10,00	9,00
5	11,00	9,00
6	10,00	7,00

Найти ассоциативные правила, если множества транзакций имеют вид

TID	Предметные наборы			
TID1	зубная паста	крем для бритья	шампунь	
TID2	мыло	дезодорант	шампунь	
TID3	шампунь	дезодорант	лосьен после бритья	шампунь
TID4	крем для бритья	шампунь	дезодорант	лосьен после бритья
TID5	лосьен после бритья	мыло	зубная паста	
TID6	дезодорант	мыло	лосьен после бритья	дезодорант
TID7	дезодорант	шампунь		
TID8	зубная паста	дезодорант	крем для бритья	
TID9	дезодорант	мыло	лосьен после бритья	
TID10	лосьен после бритья	шампунь		

4.2.3. Расчетно-графическое задание

Расчетно-графическое задание 1.

Использование пакета QlikView и Power BI для решения задач анализа данных о демографической ситуации в России. Для каждого варианта приведены таблицы с указанием вида исходных данных, которые будут анализироваться средствами бизнес-аналитики.

Вариант	год	область	регион	городское население	зарплата	миграция	Млад. Смертность	Рождаемость	Смертность	население	осн. Фонды	преступления	сельскеХоз	Трудоспособное Население	безработные
1	+	+	+	+	-	+	-	+	+	+	-	+	-	+	+
2	+	+	+	-	+	+	-	+	-	+	+	-	-	+	+
3	+	+	+	+	-	-	+	+	+	+	-	+	+	+	+
4	+	+	+	-	+	-	+	-	-	+	+	-	+	+	-
5	+	+	+	+	-	+	-	-	+	+	-	+	-	+	-

6	+	+	+	-	+	+	-	-	-	+	+	-	-	+	-
7	+	+	+	+	-	-	+	+	+	+	-	+	+	+	+
8	+	+	+	-	+	-	+	+	-	+	+	-	+	+	+
9	+	+	+	+	-	+	-	+	+	+	-	+	-	+	+
10	+	+	+	-	+	+	-	-	-	+	+	-	-	+	-
11	+	+	+	+	-	-	+	-	+	+	-	+	+	+	-
12	+	+	+	-	+	-	+	-	-	+	+	-	+	+	-
13	+	+	+	+	-	+	-	+	+	+	-	+	-	+	+
14	+	+	+	-	+	+	-	+	-	+	+	-	-	+	+
15	+	+	+	+	-	-	+	+	+	+	-	+	+	+	+
16	+	+	+	-	+	-	+	-	-	+	+	-	+	+	-
17	+	+	+	+	-	+	-	-	+	+	-	+	-	+	-
18	+	+	+	-	+	+	-	-	-	+	+	-	-	+	-
19	+	+	+	+	-	-	+	+	+	+	-	+	+	+	+
20	+	+	+	-	+	-	+	+	-	+	+	-	+	+	+
21	+	+	+	+	-	+	-	+	+	+	-	+	-	+	+
22	+	+	+	-	+	+	-	-	-	+	+	-	-	+	-
23	+	+	+	+	-	-	+	-	+	+	-	+	+	+	-
24	+	+	+	-	+	-	+	-	-	+	+	-	+	+	-
25	+	+	+	+	-	+	-	+	+	+	-	+	-	+	+

Расчетно-графическое задание 2. Задание на тему «Кластерный анализ» выполняется в соответствии с разработанным учебно-методическим пособием. При выполнении задания выполняется кластерный анализ варианта исходных данных, которые содержат учебные наборы данных, находящиеся в Kaggle на других ресурсах по data science.

4.2.4. Тесты

ЗАДАНИЕ № 1 (- выберите один вариант ответа)

Коэффициент парной корреляции характеризует тесноту _ связи между _ переменными.

ВАРИАНТЫ ОТВЕТОВ:

1) линейной … несколькими

2) нелинейной … несколькими

3) линейной … двумя

4) нелинейной … двумя

ЗАДАНИЕ № 2 (- выберите варианты согласно тексту задания)

Установите соответствие между наименованиями элементов уравнения $Y=b_0+b_1X+e$ и их буквенными обозначениями:

1. параметры регрессии
2. объясняющая переменная
3. объясняемая переменная
4. случайные отклонения

ВАРИАНТЫ ОТВЕТОВ:

A) Y **B)** b_0, b_1

C) X **D)** e

ЗАДАНИЕ № 3 (- выберите несколько вариантов ответа)

Для линейного уравнения регрессии $y = a + bx + \varepsilon$ метод наименьших квадратов используется при оценивании параметров…

ВАРИАНТЫ ОТВЕТОВ:

1) B **2)** y

3) X **4)** a

ЗАДАНИЕ № 4 (- выберите один вариант ответа**)**
Сколько параметров содержит парное линейное уравнение регрессии?

ВАРИАНТЫ ОТВЕТОВ:
А) 1
Б) 2
В) 3
Г) 4

ЗАДАНИЕ № 5 (-выберите несколько вариантов ответа)
При выполнении предпосылок МНК оценки параметров регрессии обладают свойствами:

ВАРИАНТЫ ОТВЕТОВ:

1) достоверность **2)** эффективность

3) несмещенность **4)** несостоятельность

ЗАДАНИЕ № 6 *(- выберите один вариант ответа*)
Как влияет увеличение объема выборки на величину остаточной дисперсии случайной величины?

ВАРИАНТЫ ОТВЕТОВ:

А) Никак.
Б) Остаточная дисперсия увеличивается.
В) Остаточная дисперсия уменьшается.
Г) Результат зависит от конкретного вида случайной величины.

ЗАДАНИЕ № 7 *(- выберите один вариант ответа*)
При каком значении параметра x оценка случайной величины y, полученная в рамках парной линейной регрессионной модели, будет наиболее точной?

ВАРИАНТЫ ОТВЕТОВ:

А) При $x = (x_{\min} + x_{\max})/2$, где $x_{\min}$, $x_{\max}$ - минимальное и максимальное значения параметра x из обследованного интервала.

Б) При $x = \sqrt{x_{\min} x_{\max}}$

В) При $x = \bar{x}$, где $\bar{x}$ - среднее значение параметра x из обследованного интервала.

Г) Точность одинакова при всех x.

ЗАДАНИЕ № 8 *(- выберите один вариант ответа*)

Рассматривается парная линейная регрессионная модель. Как изменится ширина доверительного интервала для условного математического ожидания случайной величины $\bar{y}(x)$ при увеличении объема выборки в 4 раза?

ВАРИАНТЫ ОТВЕТОВ:

А) Увеличится в 4 раза.
Б) Уменьшится в 4 раза.
В) Увеличится в 2 раза.
Г) Уменьшится в 2 раза.

ЗАДАНИЕ № 9 (- выберите несколько вариантов ответа)

Гомоскедастичность остатков подразумевает …

ВАРИАНТЫ ОТВЕТОВ:

1) рост дисперсии остатков с увеличением значения фактора

2) одинаковую дисперсию остатков при каждом значении фактора

3) уменьшение дисперсии остатка с уменьшением значения фактора

4) максимальную дисперсию остатков при средних значениях фактора

ЗАДАНИЕ № 10 (- выберите несколько вариантов ответа)

В кластерном анализе используются методы объединения …

ВАРИАНТЫ ОТВЕТОВ:

1) Ближнего соседа

2) Дальнего соседа

3) Среднего соседа

4) центроидный метод

ЗАДАНИЕ № 11 (- выберите несколько вариантов ответа)

В кластерном анализе для определения близости между кластерами используются метрики …

ВАРИАНТЫ ОТВЕТОВ:

1) Эвклидово расстояние

2) Куб Эвклидова расстояния

3) Взвешенное эвклидово расстояние

4) Квадрат Эвклидова расстояния

ЗАДАНИЕ № 12 (- выберите один вариант ответа)

В дискриминантном анализе обучающая выборка используется для …

ВАРИАНТЫ ОТВЕТОВ:

1) Выявления значимых признаков

2) Выявления аномального измерения

3) Разделения объектов на классы

4) Выбора вида модели

ЗАДАНИЕ № 13 (- выберите один вариант ответа)

В факторном анализе при *n* измерениях и *k* факторах матрица факторных нагрузок имеет размерность …

ВАРИАНТЫ ОТВЕТОВ:

1) *n*x*n*

2) *k*x*k*

3) *n*x*k*

4) *k*x*n*

ЗАДАНИЕ № 14 (- выберите несколько вариантов ответа)

Метод главных компонент …

ВАРИАНТЫ ОТВЕТОВ:

1) Является частным случаем метода факторного анализа

2) Предназначен для снижения размерности задачи

3) Устраняет проблему коррелированности факторов

4) Предназначен для классификации

ЗАДАНИЕ № 15 (- выберите один вариант ответа)

Сигмоидальная активизационная функция искусственного нейрона имеет вид…

ВАРИАНТЫ ОТВЕТОВ:

1) $y = \begin{cases} 1, \text{ если } S \geq T \\ 0, \text{ если } S < T \end{cases}$

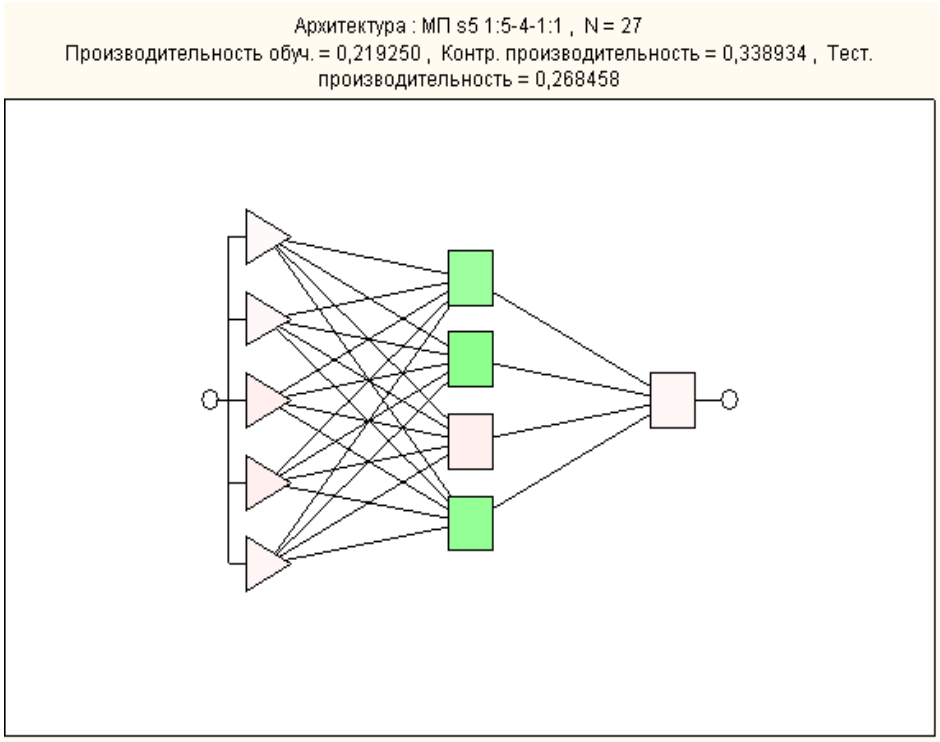
2) $y = \dfrac{1}{1+e^{-S}}$

3) $y = \begin{cases} 1, \text{ если } S > 0 \\ -1, \text{ если } S \leq 0 \end{cases}$

4) $y = \begin{cases} S, \text{ если } S \geq 0 \\ 0, \text{ если } S < 0 \end{cases}$

ЗАДАНИЕ № 16 (- выберите один вариант ответа)

На рисунке приведена архитектура многослойного персептрона. Определить число рецепторных, реагирующих и ассоциативных элементов



ВАРИАНТЫ ОТВЕТОВ:

1) Рецепторных 4, ассоциативных 5, реагирующих -1

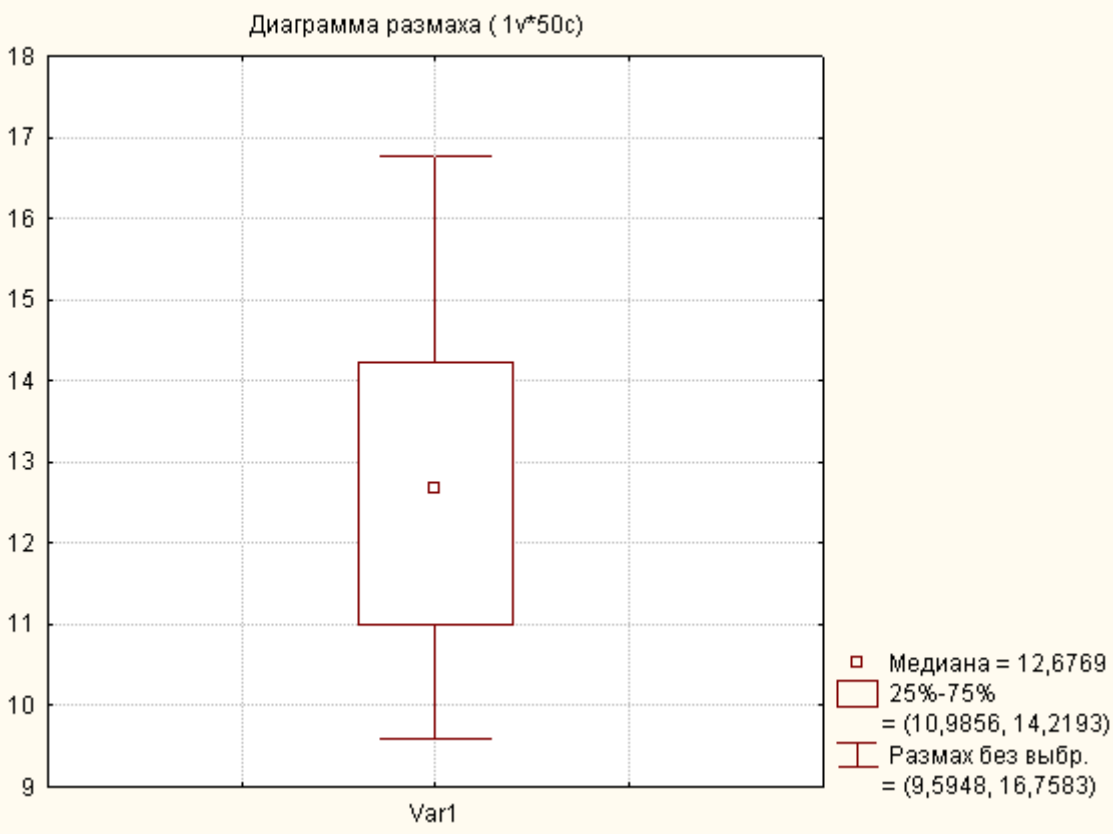
2) Рецепторных 4, ассоциативных 1, реагирующих -5

3) Рецепторных 5, ассоциативных 4, реагирующих -1

4) Рецепторных 1, ассоциативных 5, реагирующих -4

ЗАДАНИЕ № 16 (- выберите один вариант ответа)

На рисунке приведена диаграмма размаха («ящик с усами». Определить интерквартильный размах с точностью до второго знака



ВАРИАНТЫ ОТВЕТОВ:

1) 3,23 2) 7,16

3) 12,68 4) 16,76

ЗАДАНИЕ № 18 (- свободный ответ)

После перепроверки партии электронных ламп из 110 штук, изготовленных по двум технологиям, получены результаты, приведенные в таблице. Необходимо оценить связь качества ламп с технологией из изготовления с помощью мер связи.

Технология	Число дефектных ламп	Число годных ламп	Всего ламп
1	95	15	110
2	70	40	110
Итого	165	55	220

Вычислить значение критерия χ^2 с точностью до двух знаков

ЗАДАНИЕ № 19 (- свободный ответ)

40 пациентов начали применять два препарата. При этом 8 пациентов признали сильным действие обоих препаратов. 11 –слабым действие обоих; 16 – сильным действие первого и слабым действие второго; 5 –наоборот. Таблица сопряженности имеет вид:

Препарат 1(Опыт 1)	Препарат 2 (Опыт 2)	
	Сильное (+)	Слабое (-)
Сильное (+)	*a* (8)	*b* (16)
Слабое (-)	*c* (5)	*d* (11)

a, d соответствуют неизменным условиям опыта, когда опыт 1 и 2 одинаково действуют на объект или не действуют. Значения *b, c* соответствуют разным условиям опыта, когда действует только один. Таким образом, при условии $b \approx c$ опыт не влияет на объект. Определить значение коэффициента сопряженности с точностью до двух знаков после запятой

ЗАДАНИЕ № 20 (- выберите один вариант ответа)

Два студента расставили значимость предметов учебного плана по рангам. Более важному предмету соответствует меньший ранг. Студенты провели ранжирования без использования равных рангов.

	1	2	3	4	5	6	7	8	9	10
Предмет1,ri	2	1	3	4	6	8	5	10	7	9
Предмет2,si	1	3	4	2	7	10	8	5	6	9

Оценить коэффициент корреляции Спирмена с точностью до двух знаков, используя непараметрическую статистику.

1. 0,6
2. -0,2
3. 0,7
4. 0,55

Ключи к заданиям

1) 1
2) A-3, B-1,C-2, D-4
3) 1,4
4) 2
5) 2,3
6) B
7) A
8) Г
9) 1,3
10) 1,2,4
11) 1,3,4
12) 3
13) 2
14) 1,2,3
15) 2
16) 3
17) 1
18) 15,15
19) 0,05
20) 0,7

4.3. Оценочные средства для промежуточной аттестации.

Таблица 3

Код компетенции	Наименование компетенции	Код этапа освоения компетенции	Наименование этапа освоения компетенции
ДПК-31	Сбор, обработка и анализ больших данных с использованием существующей в организации методологической и технологической инфраструктуры	ДПК-31.1	Способность планировать и проводить аналитические работы, использовать математический аппарат, информационные технологии, современные языки статистической обработки и программные средства решения эконометрических задач и задач анализа данных.

Таблица 4

Этап освоения компетенции	Показатель оценивания	Критерий оценивания
ДПК -31.1	1. Самостоятельно решает простейшие задачи планирования и выполнения аналитических работ, использования математического аппарата, информационных технологий, современных языков и средств статистической обработки. 2. Демонстрирует умение планировать, проводить и управлять аналитическими работами, использовать современные ИТ. 3. Показывает знания и умения использовать технологии анализа данных, решать задачи эконометрического моделирования	1. Представлены результаты выполнения учебных кейсов по решению задач аналитики данных, эконометрического моделирования. 2. Приведены скрипты, результаты решения задач разведывательного анализа, интеллектуального анализа, многомерной статистики с использованием статистических пакетов, языков статистической обработки (R, Python). 3. Правильно выполнения интерпретация результатов моделирования, их валидация 4. Сделаны правильные ответы на поставленные вопросы или тесты

Для оценки сформированности компетенций, знаний и умений, соответствующих данным компетенциям, используются контрольные вопросы, а также задачи, при решении которых необходимо построить имитационные модели, спланировать и провести эксперименты с ними.

Типовые вопросы, выносимые на экзамен:

1. Дать характеристику систем поддержки принятия решений, хранилищ данных.
2. Сформулировать свойства OLAP и OLTP-систем, найти их отличия.
3. Дать определение технологии KDD. Охарактеризовать этапы анализа данных KDD.
4. Объяснить содержание основных элементов математической статистики, используемых для анализа данных. Дать характеристику операций агрегирования данных.
5. Охарактеризовать содержание начальных этапов KDD, предобработки, очистка и трансформации данных.
6. Сделать обзор основного содержания разведывательного анализа, содержания модели Тьюки. Указать основные задачи разведывательного анализа.
7. Дать определение понятия аномалии. Выполнить характеристику методов борьбы с аномалиями. Дать характеристику ящичных диаграмм. Привести примеры.
8. Дать определение повторной выборки. Указать методы повторной выборки и организации их использования.
9. Назвать основные графические средства анализа. Характеризовать организацию построения гистограмм и вероятностных графиков, а также их использования при разведывательном анализе.
10. Описать организацию проверки гипотез о законах распределения. Привести примеры проверки гипотез в R.

11. Описать организацию проверки гипотез с использованием Т-критерия. Привести примеры проверки гипотез в R.
12. Определить гистограмму распределения и диаграмма «ящик с усами». Описать их использование при проверке гипотезы о законе распределения.
13. Охарактеризовать язык R. Выполнить обзор его основных возможностей.
14. Охарактеризовать графическую среду RStudio. Привести примеры решения простейших задач с помощью данной графической среды
15. Дать общую характеристика SPSS. Привести примеры решения задач описательной статистики.
16. Определить задачу кластерного анализа. Дать общую характеристику методов кластерного анализа.
17. Привести примеры метрик кластерного анализа.
18. Дать характеристику методов определения близости между кластерами. Привести примеры решения задач определения близости.
19. Объяснить содержание иерархической кластеризации. Охарактеризовать агломеративный и дивизимный методы. Привести примеры в R.
20. Характеризовать метод к-средних. Привести примеры решения задач в SPSS и в R.
21. Дать характеристику метода k-средних, методов определения числа кластеров, метода локтя.
22. Определить структуру ассоциативных правил, понятия антимонотонности.
23. Определить метрики построения ассоциативных правил.
24. Характеризовать алгоритм построения ассоциативных правил a'priori, указать на параметры, используемые при построении правил.
25. Дать определение деревьев решений. Дать общую характеристика деревьев решений.
26. Сделать обзор алгоритмов построения деревьев решений. Характеризовать алгоритм CARD, C4.5.
27. Характеризовать задачи классификации. Дать определение ROC-кривой. Описать организацию оценки качества классификации с помощью AUC. Объяснить организацию решения задачи классификации в Deductor. Привести пример построения ROC-кривой.
28. Привести характеристику построения деревьев решения в R.
29. Характеризовать метод random forest. Дать характеристику, привести примеры решения задач классификации с помощью метода случайного леса.
30. Охарактеризовать таблицу сопряженности (conclusion). Описать организацию построения таблиц сопряженности в R. Уточнить содержание ошибок первого и второго рода.
31. Дать определение логистической регрессии. Привести примеры решения задач бинарной классификации различными методами. Определить понятие ансамбля методов.
32. Дать определение нейронной сети. Классифицировать нейронные сети.
33. Характеризовать активизационные функции нейрона.
34. Привести примеры архитектура нейронной сети. Построить нейронные сети в Deductor.
35. Дать характеристику основ синтаксиса языка R, структуры данных языка.
36. Характеризовать средства импорта и экспорта данных. Привести примеры.
37. Классифицировать графику в R. Привести примеры.
38. Характеризовать средства анализ выборки в R. Привести примеры.
39. Продемонстрировать организацию проверки статистических гипотез в R, в SPSS. Описать содержание и организацию использования критерия Стьюдента и критерия Манна-Уитни.

40. Продемонстрировать решение задач корреляционного анализа в R. Сделать обзор средств корреляционного анализа.
41. Характеризовать корреляции Пирсона, Спирмена, Кендалла, частной корреляции. Показать примеры их использования

Типовые контрольные задания на экзамен:

Задача 1. Проверить гипотезу о значимом отличии среднего балла за экзамены в десятом и одиннадцатом классах, используя критерий Стьюдента и критерий Манна-Уитни. Построить диаграммы «ящик с усами» для школьников, имеющих разные хобби. Построить диаграмму «дерево-листья». Данные находятся в файле тестыШкола.txt. Задачу решить в R и в SPSS.

Построить задачу классификации хобби в зависимости от результатов тестирования. Задачу классификации решить с помощью деревьев решений в R.

Задача 2. Создать случайную последовательность размером в 500 наблюдений с использованием генератора равномерно распределенных чисел в диапазоне от 0 до 10. Проверить статистическую гипотезу о числовых значениях параметров:

1 $H_0 : a = 0{,}5;\ H_1 : a \neq 0{,}5$.

2 $H_0 : a = 5;\ H_1 : a > 5$.

Построить гистограмму распределения в R. Построить гистограмму частот и гистограмму относительных частот. При построении гистограммы оценить и задать число интервалов. Указать название осей и название гистограммы, а также заливку синего цвета. На диаграмму поместить кривую ядерной плотности, а также аппроксимацию равномерным законом распределения. При построении кривой регулировать ее гладкость.

- Оценить статистические характеристики.

- При проверке гипотезы: использовать одновыборочный Т-критерий. Задать уровень значимости 0,05. Использовать одностороннюю и двухстороннюю проверки гипотезы.

- Проверить гипотезу о равномерном законе распределения с помощью критерия Колмогорова-Смирнова.

В R использовать функцию t.test

Задача 3. В файле ГосСлужба.txt приведены данные по стажу работы, стажу в должности и возрасту в государственной службе.

–Построить гистограммы распределения случайных величин.

–Оценить выборочные характеристики.

–Проверить статистические гипотезы о значимом отличии стажа в должности, стажа работы на гос. службе и возраста для мужчин и женщин с использованием t-критерия и критерия Манна-Уитни.

–Построить диаграммы размаха для случайных величин: возраст, стаж службы.

Задачу решить в SPSS.

Задача 4. Таксомоторную компанию интересует зависимость между средним пробегом автомашины в расчете на 1 л топлива и возрастом машины. Были взяты 12 автомашин одной марки. Поскольку водителями были мужчины и женщины, предполагалось, что какая-то часть изменчивости пробега определяется разной техникой вождения у мужчин и женщин. Значения среднего пробега были рассчитаны на основе сведений о расходе горючего после прохождения машиной расстояния 100 км. Данные приведены в таблице.

Пол (мужчины, женщины)	Возраст машины, лет	Расход горючего, км.
мужчина	3	8,92
женщина	4	8,8
женщина	3	9,48
мужчина	2	9,68
женщина	1	10,2

мужчина	5	8,44
мужчина	4	8,24
мужчина	1	9,6
женщина	1	10,4
мужчина	2	9,24
женщина	2	9,92
мужчина	3	8,08

- Определить, значимы ли различия между пробегом для водителей-мужчин и водителей женщин, используя Т-тест для независимых групп (двухсторонний и односторонний). Для проверки гипотезы проверить гипотезу о постоянстве дисперсии. Сравнить результаты проверки гипотезы с результатами проверки по критерию Манна-Уитни. Построить диаграммы размаха.

- Построить ящичные диаграммы для водителей мужчин и водителей-женщин.

- Решить задачу построения описательной статистики в SPSS.

Для проверки гипотезы по критерию Манна-Уитни в R использовать функцию wilcox.test($y \sim x$, data)

Задача 5. Создать две случайные последовательности двух случайных величинах, размером в 200 наблюдений, полученных с помощью генераторов нормально распределенных случайных чисел, имеющих одинаковое математическое ожидание, равное 5 и ско, соответственно 1 и 2.

- Проверить гипотезу о равенстве математических ожиданий и дисперсий данных величин.

- Изменить генератор, добавив в первый генератор смещение математического ожидания. Вновь проверить статистическую гипотезу.

- Проверить гипотезы о нормальном законе распределения.

- Найти сумму пяти случайных величин, равномерно распределенных на интервале 0, 2. Проверить гипотезу о нормальном законе распределения суммы.

Задачу решить с помощью статистических критериев в R. Построить вероятностные и квантиль-квантиль графики.

Задача 6. Решить задачу кластерного анализа для файла Семейное положение.txt. при решении задачи кластерного анализа:

- определить склонность к кластеризации;
- определить лучшую метрику иерархической кластеризации;
- выполнить иерархическую кластеризацию;
- определить состав и центроиды кластеров;
- Решить задачу кластеризации методом k-средних;
- Выполнить интерпретацию полученных кластеров;
- Визуализировать полученную кластеризацию;
- Задачу решить в RStudio и в SPSS.

Задача 7. В наборе Animals библиотеки cluster имеются данные о 20 животных. Заданы 6 бинарных признаков: теплокровные/нетеплокровные; летают/не летают; позвоночный/беспозвоночный; находящихся под угрозой вымирания; живущих в группах. Решить задачу кластерного анализа наблюдений в SPSS и в R. Использовать иерархическую кластеризацию и кластеризацию методом k-средних.

Шкала оценивания.

Оценка результатов производится на основе балльно-рейтинговой системы (БРС). Использование БРС осуществляется в соответствии с приказом от 28 августа 2014 г. №168 «О применении балльно-рейтинговой системы оценки знаний студентов». БРС по дисциплине отражена в схеме расчетов рейтинговых баллов (далее – схема расчетов).

Схема расчетов сформирована в соответствии с учебным планом направления, согласована с руководителем научно-образовательного направления, утверждена деканом факультета. Схема расчетов доводится до сведения студентов на первом занятии по данной дисциплине и является составной частью рабочей программы дисциплины и содержит информацию по изучению дисциплины, указанную в Положении о балльно-рейтинговой системе оценки знаний обучающихся в РАНХиГС.

На основании п. 14 Положения о балльно-рейтинговой системе оценки знаний обучающихся в РАНХиГС в институте принята следующая шкала перевода оценки из многобалльной системы в пятибалльную:

Таблица 4.2

Количество баллов	**Оценка**	
	прописью	**буквой**
96-100	отлично	A
86-95	отлично	B
71-85	хорошо	C
61-70	хорошо	D
51-60	удовлетворительно	E

5. Методические указания для обучающихся по освоению дисциплины

Рабочей программой дисциплины предусмотрены следующие виды аудиторных занятий: лекции, практические занятия, контрольные работы. На лекциях рассматриваются наиболее сложный материал дисциплины. Лекция сопровождается презентациями, компьютерными текстами лекции, что позволяет студенту самостоятельно работать над повторением и закреплением лекционного материала. Для этого студенту должно быть предоставлено право самостоятельно работать в компьютерных классах в сети Интернет.

Практические занятия предназначены для самостоятельной работы студентов по решении конкретных задач дискретно математики. Ряд практических занятий проводится в компьютерных классах с использованием Excel. Каждое практическое занятие сопровождается домашними заданиями, выдаваемыми студентам для решения внеаудиторное время. Для оказания помощи в решении задач имеются тексты практических заданий с условиями задач и вариантами их решения.

Большинство тем основано на использовании приложения Deductor, R. Каждый студент может скачать бесплатную версию приложения (академическая версия), получить доступ к порталу данного приложения для получения актуальной информации о нем. Академическая версия имеет ограниченный функционал. В частности, нет возможности использовать современные хранилища данных. Встроенная бесплатная база данных позволяет построить хранилища. Однако все возможности современных хранилищ данных не реализованы.

Расчетно-графическое задание выполняется в средах бизнес-аналитики Qlik View, Qlik Sense, MS BI. Отчет представляется в распечатанной виде. В нем должны быть скрины основных окон разработанных платформ.

С целью контроля сформированности компетенций разработан фонд контрольных заданий. Его использование позволяет реализовать балльно-рейтинговую оценку, определенную приказом от 28 августа 2014 г. №168 «О применении балльно-рейтинговой системы оценки знаний студентов».

С целью активизации самостоятельной работы студентов в системе дистанционного

обучения Moodle разработан учебный курс «Анализ данных», включающий набор файлов с текстами лекций, практикума, примерами задач, а также набором тестов для организации электронного обучения студентов.

Для активизации работы студентов во время контактной работы с преподавателем отдельные занятия проводятся в интерактивной форме. В основном, интерактивная форма занятий обеспечивается при проведении занятий в компьютерном классе. Интерактивная форма обеспечивается наличием разработанных файлом с заданиями, наличием контрольных вопросов, возможностью доступа к системе дистанционного обучения, а также к тестеру.

Для работы с печатными и электронными ресурсами СЗИУ имеется возможность доступа к электронным ресурсам. Организация работы студентов с электронной библиотекой указана на сайте института (странице сайта – «Научная библиотека»).

Контрольные вопросы для подготовки к занятиям

Таблица 4.3

№ п/п	Наименование темы дисциплины	Контрольные вопросы для самопроверки
1	Тема 1. Основы анализа данных. Системы	1. Дайте сравнительный анализ OLAP и OLTP систем. Сферы их применения. 2. В чем отличие информационного хранилища от баз данных? 3. Принципы построения информационных хранилищ. Классификация информационных хранилищ. 4. Модели информационных хранилищ. Многомерная модель данных. Нормальная форма. Денормализация моделей данных. 5. Правила Кодда. Зачем применяется денормализация моделей? 6. Размерностные модели. В чем отличие таблицы фактов от размерностной таблицы? 7. Дайте характеристику стандартам Data Mining.
2	Тема 2. Предобработка и очистка данных	1. Дайте характеристику этапа ETL (Extracting Transformating and Loading). 2. Какие задачи решаются Data Mining? 3. Каково предназначение и средства разведочный анализ данных? Дайте характеристику диаграммы «ящик с усами» 4. Назовите какие операции выполняются при агрегировании данных. 5. Приведите примеры использования статистических пакетов для разведочного анализа. 6. Назовите и выполните сравнительный анализ графических средств анализа. Дайте характеристику биржевых диаграмм. 7. Для чего используются диаграммы рассеяния?
3	Тема 3. Кластерный анализ	1. Зачем используются ассоциативные правила? Приведите примеры задач использования ассоциативных правил. 2. Дайте определение ассоциативного правила. Зачем используются обобщенные правила? Что такое транзакция. Приведите примеры транзакций. 3. Какие показатели используются для построения правила? 4. Алгоритмы построения ассоциативных правил. Алгоритм apriori. 5. Общая характеристика пакета Deductor. 6. Использование пакета Deductor для решения задач интеллектуального анализа данных.
4	Тема 4. Анализ взаимосвязей между переменными, ассоциативные правила	1. Дайте определение задачи классификации. Какие методы решения задачи классификации Вы знаете? 2. Особенности решения задач классификации с обучением. 3. Деревья классификации и их свойства. 4. Приведите примеры алгоритмы построения деревьев. 5. Как определяется правило остановки построения дерева? 6. Алгоритм CART? Приведите пример его использования.

5	Тема 5. Задачи классификации. Деревья решений	1. Что понимается под кластером? Назовите характеристики кластера. Что такое «центроид» кластера? 2. Дайте классификацию методов кластерного анализа. Приведите примеры их применения в практической жизни. 3. Зачем используются меры близости? Назовите методы определения близости между кластерами. 4. Когда применяется метод ближнего соседа, дальнего соседа? Сравните их. 5. Дайте характеристику метрик кластерного анализа. 6. Поясните содержание «дендограммы» и организацию ее применения. 7. Что понимается под профилем кластера. 8. Использование статистических пакетов для решения задач кластерного анализа. 9. Дайте характеристику метода к-средних.

6. Учебная литература и ресурсы информационно-телекоммуникационной сети "Интернет", включая перечень учебно-методического обеспечения для самостоятельной работы обучающихся по дисциплине

6.1. Основная литература

1. Барсегян А.А, Куприянов М.С., Степаненко В.В., Холод И.И. Анализ данных и процессов. 3-е изд. [Электронный ресурс]- СПб. : БХВ-Петербург, 2010, 512 с.-
2. Методы и модели прогнозирования социально- экономических процессов : [учеб. пособие] / Т. С. Клебанова [и др.] ; Федер. гос. бюджетное образовательное учреждение высш. проф. образования, Рос. акад. нар. хоз-ва и гос. службы при Президенте Рос. Федерации, Сев.-Зап. ин-т упр. - СПб. : Изд-во СЗИУ РАНХиГС, 2012. - 564 с.
3. Миркин, Борис Григорьевич. Введение в анализ данных [Электронный ресурс] : учебник и практикум / Б. Г. Миркин ; Нац. исслед. ун-т Высш. шк. экономики. - Электрон. дан. - М. : Юрайт, 2016. - 174 с.
4. Наследов, Андрей Дмитриевич. Математические методы психологического исследования : анализ и интерпретация данных : [учебное пособие] / А.Д. Наследова. - СПб. : Речь, 2007. - 390 с.
5. Паклин, Николай Борисович. Бизнес-аналитика: от данных к знаниям : [хранилища данных и OLAP, очистка и предобработка данных, основные алгоритмы Data Mining, сравнение и ансамбли моделей, решение бизнес задач на аналитической платформе Deductor] : учеб. пособие / Н. Паклин, В. Орешков. - 2-е изд., испр. - СПб.[и др.] : Питер, 2013. - 701 с.

Все источники основной литературы взаимозаменяемы.

6.2.Дополнительная литература

1. Барсегян А.А, Куприянов М.С., Степаненко В.В., Холод И.И. Технология анализа данных: Data Mining, Visual Mining, Text Mining, OLAP. – СПб.: БХВ-Петербург, 2004.
2. Боровиков В.П., Ивченко Г.И. Прогнозирование в системе STATISTICA в среде Windows. – М.: Финансы и статистика, 2000.
3. Винстон, Уэйн Л. Excel 2007 : Анализ данных и бизнес- моделирование = Excel 2007: Data Analysisi and Business Modeling : [пер. с англ.] / Уэйн Л. Винстон. - М. : Рус. Редакция ; СПб. : БХВ-Петербург, 2008. - 594 с.
4. Кацко И.А., Паклин Н.Б. Практикум по анализу данных на компьютере. – М.: КолосС, 2009. -278 с.
5. Ларсон Б. Разработка Бизнес-аналитики в Microsoft SQL Server 2005. – Москва: «Питер», 2008.
6. Наследов А. SPSS 19. Профессиональный статистический анализ данных. – СПб. : Питер, 2011.

7. Наумов В.Н. Средства бизнес-аналитики. – СПб.: СЗИУ, 2016. .
8. Тихомиров Н.П. Методы эконометрики и многомерного статистического анализа. – М.: Экономика, 2011.
9. Халафян А.А. STATISTICA 6. Статистический анализ данных. – М.: ООО «Бином-Пресс», 2007.

6.3. Учебно-методическое обеспечение самостоятельной работы.

1. Положение об организации самостоятельной работы студентов АНОВО «Институт социальных наук»
2. Положение о курсовой работе (проекте) выполняемой студентами «АНОВО «Институт социальных наук»

6.4. Нормативные правовые документы.

Не используются

6.5. Интернет-ресурсы.

Русскоязычные ресурсы

Электронные учебники электронно - библиотечной системы (ЭБС) «Айбукс»
Электронные учебники электронно – библиотечной системы (ЭБС) «Лань»
Рекомендуется использовать следующий интернет-ресурсы
http://serg.fedosin.ru/ts.htm
http://window.edu.ru/resource/188/64188/files/chernyshov.pdf

6.6. Иные источники.

Не используются.

7. Материально-техническая база, информационные технологии, программное обеспечение и информационные справочные системы

Курс включает использование программного обеспечения Microsoft Excel, Microsoft Word, Microsoft Power Point для подготовки текстового и табличного материала, графических иллюстраций. При проведении занятий используются средства бизнес-аналитики.

Методы обучения с использованием информационных технологий (компьютерное тестирование, демонстрация мультимедийных материалов).

Интернет-сервисы и электронные ресурсы (поисковые системы, электронная почта, профессиональные тематические чаты и форумы, системы аудио и видео конференций, онлайн энциклопедии, справочники, библиотеки, электронные учебные и учебно-методические материалы).

Для организации дистанционного обучения используется система Moodle.

№ п/п	Наименование
1.	Компьютерные классы с персональными ЭВМ, объединенными в локальные сети с выходом в Интернет
2.	Пакет Excel -2013, 2016, professional plus
3.	Аналитическая платформа Qlik View, MS BI
4.	Система бизнес-аналитики Deductor Academic
5.	Средства интеллектуального анализа SQ Lserver. Надстройка Analysis services, data mining ad-insforOffice.
6.	SPSS
7.	Язык R (Python)

8.	Мультимедийные средства в каждом компьютерном классе и в лекционной аудитории
9.	Браузер, сетевые коммуникационные средства для выхода в Интернет
10.	Система дистанционного обучения Moodle

Компьютерные классы из расчета 1 ПЭВМ для одного обучаемого. Каждому обучающемуся должна быть предоставлена возможность доступа к сетям типа Интернет в течение не менее 20% времени, отведенного на самостоятельную подготовку.