

Документ подписан простой электронной подписью
Информация о владельце:
ФИО: Андрей Драгомирович Хлутков
Должность: директор
Дата подписания: 17.09.2024 18:04:30
Уникальный программный ключ:
880f7c07c583b07b775f6604a630281b13ca9fd2

**Федеральное государственное бюджетное образовательное
учреждение высшего образования
«РОССИЙСКАЯ АКАДЕМИЯ НАРОДНОГО ХОЗЯЙСТВА
И ГОСУДАРСТВЕННОЙ СЛУЖБЫ
ПРИ ПРЕЗИДЕНТЕ РОССИЙСКОЙ ФЕДЕРАЦИИ»**

Северо-Западный институт управления – филиал РАНХиГС

Кафедра бизнес-информатики
(наименование кафедры)

УТВЕРЖДЕНО
Директор СЗИУ РАНХиГС
А.Д.Хлутков

ПРОГРАММА МАГИСТРАТУРЫ
Аналитическое обеспечение информационной безопасности
(наименование образовательной программы)

РАБОЧАЯ ПРОГРАММА ДИСЦИПЛИНЫ,
реализуемой без применения электронного (онлайн) курса
Б1.В.ДВ.01.01 «Методы анализа данных»
(индекс, наименование дисциплины, в соответствии с учебным планом)
Методы анализа данных
(краткое наименование дисциплины)

38.04.05 Бизнес-информатика
(код, наименование направления подготовки)

«Аналитическое обеспечение информационной безопасности»
(профиль)

очная
(форма обучения)

Год набора – 2024

Санкт-Петербург, 2024 г.

Автор–составитель:

Доктор военных наук, кандидат технических наук, профессор, заведующий кафедрой бизнес-информатики Наумов Владимир Николаевич

Заведующий кафедрой бизнес-информатики

д.в.н., профессор

Наумов Владимир Николаевич

РПД одобрена протоколом заседания кафедры бизнес-информатики № 6 от 06.03.2023 г.

СОДЕРЖАНИЕ

1. Перечень планируемых результатов обучения по дисциплине, соотнесенных с планируемыми результатами освоения образовательной программы.....	4
2. Объем и место дисциплины в структуре образовательной программы.....	5
3. Содержание и структура дисциплины	6
4. Материалы текущего контроля успеваемости обучающихся и фонд оценочных средств промежуточной аттестации по дисциплине	8
5. Оценочные материалы промежуточной аттестации по дисциплине.....	31
6. Методические указания для обучающихся по освоению дисциплины.....	38
7. Учебная литература и ресурсы информационно-телекоммуникационной сети "Интернет", учебно-методическое обеспечение самостоятельной работы обучающихся по дисциплине.....	38
7.1. Основная литература.....	38
7.2. Дополнительная литература.....	40
7.3. Учебно-методическое обеспечение самостоятельной работы.....	40
7.4. Нормативные правовые документы.....	41
7.5. Интернет-ресурсы.....	41
7.6. Иные источники.....	41
8. Материально-техническая база, информационные технологии, программное обеспечение и информационные справочные системы	41

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

1.1. Дисциплина «Методы анализа данных» обеспечивает овладение следующими компетенциями:

Таблица 1.1

Код компетенции	Наименование компетенции	Код этапа освоения компетенции	Наименование этапа освоения компетенции
ПКс-2	Способен обосновывать подходы и требования к системе обеспечения информационной безопасности, оценивать уровни безопасности компьютерных систем и сетей	ПКс-2.3	Оценивает уровни безопасности компьютерных систем и сетей

1.2.В результате освоения дисциплины у магистрантов должны быть сформированы:

Таблица 1.2

ОТФ/ТФ (при наличии профстандарта)/ профессиональные действия	Код этапа освоения компетенции	Результаты обучения
06.033 – СПЕЦИАЛИСТ ПО ЗАЩИТЕ ИНФОРМАЦИИ В АВТОМАТИЗИРОВАННЫХ СИСТЕМАХ Е - Формирование требований к защите информации в автоматизированных системах	ПКс-2.3	<p>на уровне знаний:</p> <ul style="list-style-type: none"> – теоретические и прикладные вопросы анализа данных с целью анализа, обоснования и выбора решений в оценке и управлении информационной безопасностью; – современные ИКТ и ИС в сфере анализа данных и машинного обучения, их возможности; – средства бизнес-аналитики, современные языки статистической обработки (R, Python) и графические платформы; – основные понятия и основные методы анализа данных, интеллектуальной обработки данных, эконометрики, многомерной математической статистики; – технологии анализа данных
08.037 - БИЗНЕС-АНАЛИТИК Ф - Аналитическое обеспечение разработки стратегии изменений организации		

2. Объем и место дисциплины в структуре ОП ВО

Объем дисциплины

Общая трудоемкость дисциплины составляет 3 зачетных единиц 108 академических часов.

Дисциплина реализуется частично с применением дистанционных образовательных технологий (далее – ДОТ). Доступ к системе дистанционных образовательных технологий осуществляется каждым обучающимся самостоятельно с любого устройства на портале: <https://lms.ranepa.ru/>. Пароль и логин к личному кабинету / профилю предоставляется магистранту в деканате.

Таблица 2

Вид работы	Трудоемкость в акад. часах ауд./ЭО, ДОТ	Трудоемкость в астрон. часах ауд./ЭО, ДОТ
Общая трудоемкость	108	81
Контактная работа с преподавателем	48	36
Лекции	20	18
Практические занятия	28	21
Самостоятельная работа	22	16,5
Консультация	2	1,5
Контроль	36	27
Формы текущего контроля	Задания, контрольная работа, тест, опрос	
Форма промежуточной аттестации	Экзамен	

Место дисциплины в структуре ОП ВО

Дисциплина Б1.В.07 «Методы анализа данных» относится к вариативной части учебного плана по направлению магистратуры «Бизнес-информатика». Преподавание дисциплины «Методы анализа данных» основано на дисциплинах «Теория вероятностей и математическая статистика», «Математический анализ», «Программирование», изучаемых в период обучения магистрантов по программам специалитета и бакалавриата, а также дисциплин Б1.В.02 – «Математические методы статистической обработки и анализа данных», ФТД.02 – «Предсказательная аналитика», Б1.В.ДВ.02.01 «Менеджмент данных» (Б1.В.ДВ.02.02 – «Большие данные в бизнес-аналитике»). В свою очередь она создаёт необходимые предпосылки для освоения программ таких дисциплин, как Б1.В.09 – «Интеллектуальный анализ текстов и изображений», Б1.О.07 «Аналитическая поддержка принятия решений», а также при выполнении научно-исследовательской и выпускной квалификационной работы.

Дисциплина осваивается с применением электронного (онлайн) курса (далее – ЭК). 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/ДОТ	ПЗ/ДОТ	КС Р/ДОТ	СРО		СП

Тема 1	Основы анализа и визуализации данных	12	4	4		4		О/Т
Тема 2	Предобработка и очистка данных	12	4	4		4		О/Зад/Т
Тема 3	Кластерный анализ	20	6	8		6		О/Зад/Т
Тема 4.	Задачи классификации	26	6	12		8		О/Зад/Т
Промежуточная аттестация		36			2*	36		Экзамен
Всего (акад./астр. часы):		144/108	24/18	28/21	2	54(54)/40,5		

Примечание:

2* - консультация, не входящая в общий объем дисциплины

Используемые сокращения:

Л – занятия лекционного типа (лекции и иные учебные занятия, предусматривающие преимущественную передачу учебной информации педагогическими работниками организации и (или) лицами, привлекаемыми организацией к реализации образовательных программ на иных условиях, обучающимся) ;

ПЗ – практические занятия (виды занятия семинарского типа за исключением лабораторных работ) ;

КСР – индивидуальная работа обучающихся с педагогическими работниками организации и (или) лицами, привлекаемыми организацией к реализации образовательных программ на иных условиях (в том числе индивидуальные консультации) ;

СР – самостоятельная работа, осуществляемая без участия педагогических работников организации и (или) лиц, привлекаемых организацией к реализации образовательных программ на иных условиях;

СП – самопроверка;

СРО – самостоятельная работа обучающегося

контрольные работы (К), опрос (О), тестирование (Т). Выполнение задания (Зад)

3.Содержание дисциплины

Тема 1. Основы анализа и визуализации данных

Введение. Понятие анализа данных. Задачи систем поддержки принятия решений. Понятия данные, информация, знания. Стандарты Data Mining. Стандарт CWM, CRISP, PMML. Жизненный цикл процесса анализа данных. Анализ данных и машинное обучение. Data mining. Добыча данных. Классификация методов Data Mining. Модели Data Mining. Языки аналитики данных.

Интегрированные средства разработки Rstudio, Anaconda navigator (Jupyter Notebook). Платформа Loginom. Парадигма Low code. Графические средства анализа данных. Графические библиотеки. Статистические диаграммы. «Ящичные» диаграммы. Диаграммы «ствол-листья». Задачи классификации и регрессии.

Бизнес-аналитика. Ракурсы бизнес-аналитики. Средства BI. Qlik View, Qlik Sence, Power BI.

Тема 2. Предобработка и очистка данных

Методология KDD. Задачи предобработки данных. ETL (Extracting Transforming and Loading) – средство извлечения, обработки и загрузки данных. 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).

Основные положения непараметрической и нечисловой статистики. Частотный

анализ. Таблицы сопряженности. Таблица сопряженности 2x2. Непараметрические и нечисловые критерии. Канонический анализ. Корреляционная матрица. Коэффициенты канонической корреляции. Меры избыточности переменных. Разведочный факторный анализ.

Тема 3. Кластерный анализ

Постановка задач кластерного анализа. Определение кластера. Параметры кластера. Меры близости. Метрики кластерного анализа. Базовые алгоритмы кластеризации. Иерархическая кластеризация. Дендрограммы. Метод К-средних. Метод DBscan Профили кластеров. Оценка качества кластерного анализа. Использование пакета Logitom для решения задач кластерного анализа. Кластерный анализ в средствах интеллектуального анализа Microsoft Office (на R, JASP, Python).

Тема 4. Задачи классификации

Формулировка задачи классификации. Классификационный анализ с обучением. Метод k-ближайших соседей. Наивный байесовский классификатор. Логистическая регрессия. Деревья решений. Алгоритмы построения деревьев решений. Классификация критериев разбиений. Информационный критерий. Критерий Gini. Деревья классификации и их свойства. Типы ветвления. Методы и алгоритмы построения деревьев. Алгоритм CART. Использование нейронных сетей для решения задач классификации. Карты Кохонена. Ансамблевые методы классификации. Сравнение результатов классификации различными методами. Верификация. Оценка качества классификации. ROC-кривая. Показатель AUC. Таблица сопряженности (матрица путаницы). Показатели точности: accuracy, recall, precision, F1

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

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

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

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

Таблица 4.1

Тема (раздел)	Формы (методы) текущего контроля успеваемости
Тема 1. Основы анализа данных	О/Т
Тема 2. Предобработка и очистка данных	О/Зад/Т
Тема 3. Кластерный анализ	О/Зад/Т
Тема 4. Задачи классификации	О/Зад/Т

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 магистрантов, 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.

4.2. Материалы для текущего контроля успеваемости

4.2.1 Кейс-задания Типовые оценочные материалы по теме 1

Типовые вопросы для опроса по теме 1

1. Дайте характеристику стандартов анализа данных.
2. В чем отличие информационного хранилища от баз данных?
3. Сравните понятия анализа и интеллектуального анализа данных.
4. Дайте характеристику ракурсов бизнес-аналитики.
5. Назовите задачи, решаемые средствами BI.
6. Назовите библиотеки графического анализа данных.
7. Какие диаграммы распределения вы знаете. Дайте их сравнительный анализ.

Задание. Решить задачу графического анализа набора данных Ирисы. Решить задачу графического анализа в `logiplot`

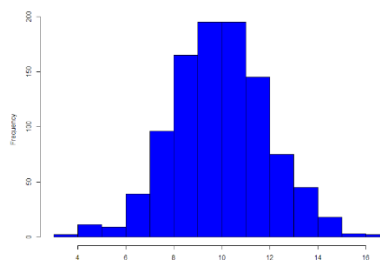
1. Импортировать данные из excel. Рассмотреть три файла, содержащих данные о цветах. Каждый файл содержит 50 наблюдений для цветка одного класса из классического набора данных Iris.
2. Выполнить операцию объединения данных.
3. Сравнить файл, полученный после объединения с данными файла Ирис.
4. Построить гистограммы с помощью панели выбора диаграмм.
5. Импортировать текстовый файл `iris.txt`.
6. С помощью визуализаторов найти описательную статистику.
7. Выполнить демопример «Ирисы Фишера»

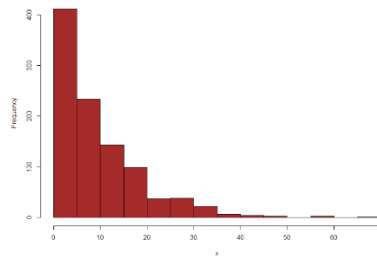
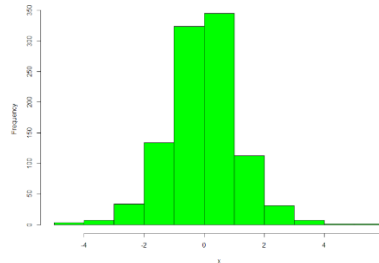
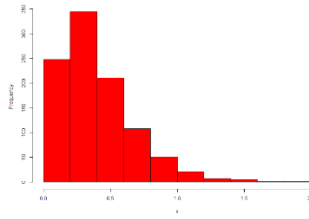
Тест

1. Гистограммы распределения

Поставить соответствие между заданными последовательностями и гистограммами распределения, если последовательности представляют собой случайные числа, полученные с помощью различных генераторов с разными законами распределения:

```
set.seed(1234)
x1<-rnorm(1000,10,2)
x2<-rgamma(1000,2,5)
x3<-rt(1000,8)
x4<-rexp(1000,0.1)
```





Варианты ответов:

1. X1
2. X2
3. X3
4. X4

2. Магический квадрант Гартнера. Выберите квадранты, которые входят в магический квадрант Гартнера.

1. Лидеры
2. Претенденты
3. участники (активисты)
4. провидцы (визуализаторы)
5. нишевые игроки
6. конкуренты (активисты)
7. вендоры
8. кандидаты (инноваторы)

3. Модели данных OLAP

Какие схемы используются при построении многомерной модели данных OLAP?

1. Иерархическая
2. индексных списков
3. звезда
4. нормализованная схема "сущность -отношение"
5. снежинка
6. цепочка
7. сетевая

5. Хранилище данных. Укажите свойства хранилищ данных

1. минимизация избыточности
2. платформенезависимость (кроссплатформенность)
3. зависимость от времени
4. интегрированность данных

5. неразрушаемость информации
 6. интероперабельность
 7. наследование свойств
 8. нормализация
6. Кривая гиперцикла Гартнера. Сколько стадий содержит кривая гиперцикла Гартнера?

7. Методология KDD. Сколько этапов содержит технология KDD?

Типовые оценочные материалы по теме 2

Типовые вопросы для опроса по теме 2

1. Дайте характеристику этапа ETL (Extracting Transforming 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?

Задания по теме 2

Задание 1. Решить задачу разведочного анализа для набора данных Boston в dataset Orange данный набор данных называется Housing.

Для решения задачи в excel, R, JASP имеются данные Boston. В python использовать `sklearn.datasets.load_boston`

Задание 2. Набор данных имеет пропуски и аномальные значения

X	Y
1	1
-1	-3
3	4
2	2
5	7
2	2
3	4
5	7
1	1
3	4
4	25
4	5
5	7
1	1
3	4
3	4
2	2
4	6
0	-1
3	4
4	5

7	10
1	1
7	10
3	4
3	4
3	3
4	5
4	19
4	6
1	0
-4	-23
4	6
1	1
2	2
5	7
1	39
4	5
5	7
7	11
4	6
6	8
4	5
2	3
4	6
1	1
5	7
5	7
5	7
3	4
-1	-3
4	6
8	12
3	4
6	9
6	9
4	5
5	7
8	12
3	4
4	6
3	4
3	4
6	9
1	1
4	
4	6
4	5
5	7

5	7
1	1
2	2
2	2
6	9
4	5
2	2
5	
1	1
2	2
1	1
5	7
5	7
2	2
-1	
3	3
3	4
4	5
7	10
6	8
5	7
3	4
4	6
4	5
6	9
5	7
5	
4	5
4	5
1	0
1	1

5. Проверить гипотезу о нормальном законе распределения для переменных X,Y
6. Решить задачу восстановления пропущенных данных;
7. Найти аномальные значения.
8. Проанализировать корреляцию между переменными. Проверить гипотезу о значимости коэффициента корреляции

Тест

2. Критерий проверки гипотезы о нормальном законе распределения

Выберите из предложенного списка статистических критериев критерий, который позволяет проверить гипотезу о нормальном законе распределения

- Колмогорова-Смирнова
- Пирсона (Chi квадрат)
- Стьюдента
- Фишера
- Шапиро
- Шапиро-Пятецкого
- Манна-Уитни

- Андерсона-Дарлинга

3. Асимметрия

Последовательность задана с помощью предложения

$x <- c(2,3,4,NA,3,4,5,NA,5,7,13,NA,26,33, 17, NA, NA, NA, 14,28,36)$

Чему равна асимметрия данной последовательности, если все пропуски заменены средним значением имеемых данных. Ответ дать с точностью до двух знаков после запятой

4. Значимость коэффициента корреляции

Чему равен уровень значимости при принятии гипотезы о наличии корреляции Пирсона для двух случайных величин/ Ответ дать с точностью до двух знаков после запятой

Таблица наблюдений

x	y
0	3
2	4
3	5
1	3
4	4
5	3

5. Интерквартильный размах

Последовательность задана с помощью предложения

$x <- c(2,3,4,NA,3,4,5,NA,5,7,13,NA,26,33, 17, NA, NA, NA, 14,28,36)$

Чему равен интерквартильный размах для данной последовательности, если все пропуски заменены средним значением имеемых данных. Ответ дать с точностью до целых

6. Корреляционный анализ

Чему равен коэффициент корреляции Спирмена для случайных величин X, Y

Таблица наблюдений

x	y
0	3
2	4
3	5
1	3
4	4
5	3

Ответ дать с точностью до двух знаков после запятой

7. коэффициент корреляции

Чему равен коэффициент корреляции Пирсона для случайных величин

Таблица наблюдений

x	y
0	3
2	4
3	5
1	3

4	4
5	3

8. Критерий Шапиро-Уилка

Сгенерировать случайную выборку, размером в 500 наблюдений, используя генератор нормально распределенных случайных чисел с математическим ожиданием 10, дисперсией 16. При этом начальное значение генератора случайных чисел равно 4321. Чему равен уровень значимости при проверке гипотезы о нормальном законе распределения с помощью критерия Шапиро-Уилка. Ответ дать с точностью до двух знаков после запятой. В качестве альтернативной использовать двустороннюю гипотезу.

9. Проверка гипотезы о законе распределения

Сгенерировать случайную выборку, размером в 500 наблюдений, используя генератор нормально распределенных случайных чисел с математическим ожиданием 10, дисперсией 16. При этом начальное значение генератора случайных чисел равно 4321. Чему равно наблюдаемое значение критерия Колмогорова-Смирнова. Ответ дать с точностью до двух знаков после запятой. В качестве альтернативной использовать двустороннюю гипотезу.

10. Проверка гипотезы о значимости отличий

Имеется выборка значений двух случайных величин

x	y
0	3
2	4
3	5
1	3
4	4
5	3

Проверить гипотезу о равенстве математических ожиданий при допущении о неравных дисперсии при допущении о независимости данных величин.

Чему равно наблюдаемое значение критерия Стьюдента. Ответ дать с точностью до одного знака после запятой.

11. Проверка гипотезы о равенстве дисперсии

Данные заданы двумя последовательностями

$x \leftarrow (0, 2, 3, 1, 4, 5)$

$y \leftarrow (3, 4, 5, 3, 4, 3)$

Чему равно наблюдаемое значение критерия Фишера при проверке гипотезы о равенстве дисперсий? Ответ дать с точностью до двух знаков после запятой

12. Пропуски данных

Последовательность задана с помощью предложения

$x \leftarrow (2, 3, 4, NA, 3, 4, 5, NA, 5, 7, 13, NA, 26, 33, 17, NA, NA, NA, 14, 28, 36)$

Чему равно среднее данной последовательности, если все пропуски заменены медианным значением. Ответ дать с точностью до двух знаков после запятой

13. Пропуски данных

Последовательность задана с помощью предложения

$x \leftarrow (2, 3, 4, NA, 3, 4, 5, NA, 5, 7, 13, NA, 26, 33, 17, NA, NA, NA, 14, 28, 36)$

Чему равна медиана данной последовательности, если все пропуски заменены средним значением имеемых данных. Ответ дать с точностью до двух знаков после запятой
Ключи:

1	2	3	4	5	6	7	8	9	10	11	12	13
1- 3,2- 1,3- 4, 4- 2	1,4.5,7	0,86	0,62	9	0,28	0,26	0,57	0,96	- 1,4	0,66	11,52	13,33

Типовые оценочные материалы по теме 3

Типовые вопросы для опроса по теме 3

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 k-средних.

Задания по теме 3

Задание 1.

Имеется выборка данных о 6 предприятиях. Найти расстояние между объектами с помощью различных метрик

	X1	X2	X3
1	120	9100	11000
2	180	8400	16000
3	840	13000	20000
4	410	11300	16000
5	460	12000	15000
6	560	11500	13000

С помощью принципа "ближайшего соседа" для метрики Евклида построить дендограмму. Задачу решить в различных приложениях. Построить workflow-diagram Orange

Задание 2.

На предприятии существуют 5 отделов. Поскольку в них имеется разное число сотрудников, разные виды деятельности и др. решено сгруппировать отделы. Решить задачу группирования иерархическим и методом k-средних.

	Стоимость производственных фондов, X1	Среднемесячный объем работ, X2
1	699	190
2	510	210
3	340	110
4	290	95
5	310	130

Задачу решить в R, JASP, Python

Задание 3. Расчетно-графическое задание.

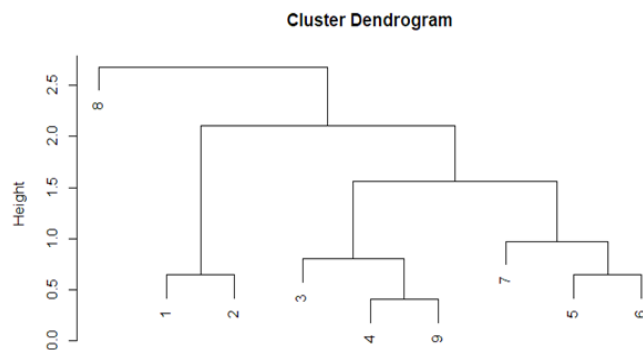
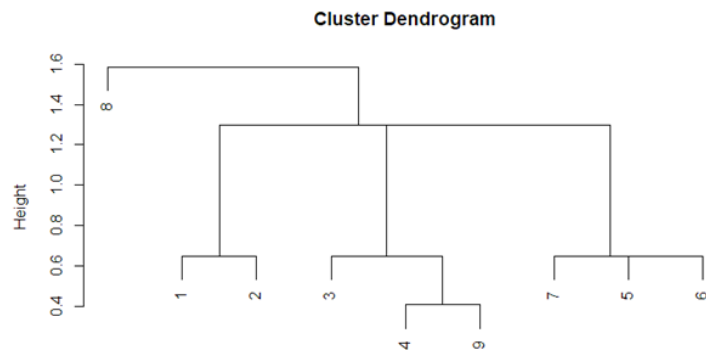
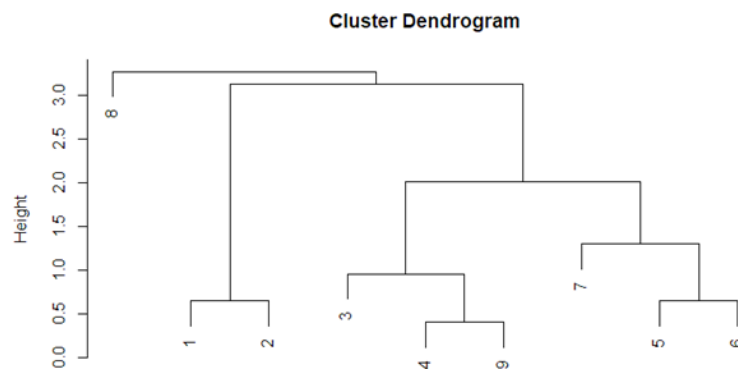
Решить задачу кластерного анализа для предложенного набора данных. Имеется 30 наборов данных, размещенных в Moodle. Каждый магистрант решает задачу кластерного анализа для своего варианта исходных данных. При решении задачи кластерного анализа следует решать задачи иерархического кластерного анализа и кластерного анализа, решенного с помощью метода k-средних. В Moodle находится пример решения задачи и оформления отчета. В примере приведены скрипты, позволяющие решить данные задачи на языке R.

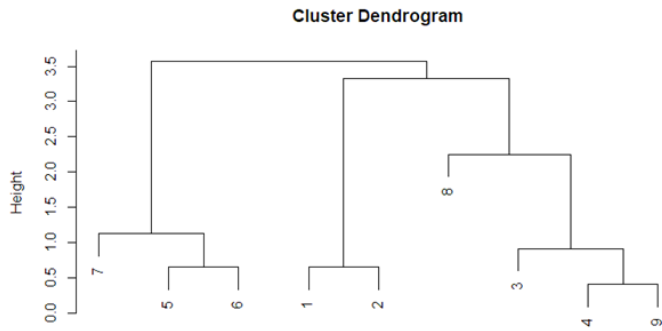
Тест

1. Дендрограммы. Поставить соответствие между дендрограммами и используемыми методами для наблюдений, признаки которых заданы последовательностями

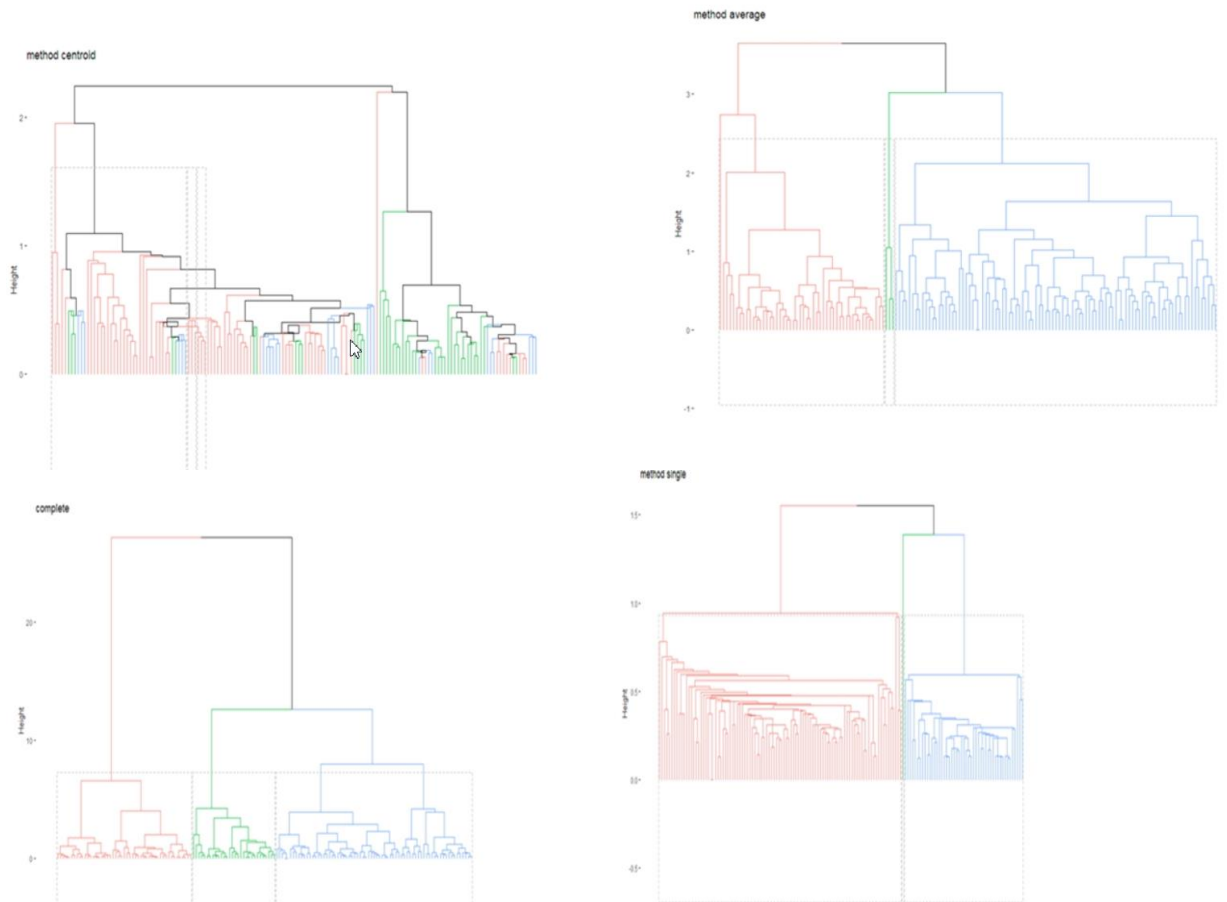
$x \leftarrow c(2,3,4,5,6,7,8,3,5)$

$y \leftarrow c(3,2,5,6,3,4,5,10,7)$.





2. Иерархический кластерный анализ. При решении задачи кластерного анализа для набора данных iris различными методами построены дендрограммы с выделением кластеров прямоугольниками для заданного числа кластеров равного трем. Выберите лучший из методов, если дендрограммы имеют следующий вид



- ближнего соседа
- дальнего соседа
- центроидный
- средней связи
- взвешенной средней связи

3. Методы иерархической кластеризации. Какой из методов кластерного анализа при анализе кандидатов на включения в кластер на текущем шаге использует результаты оценки дисперсий?

1. метод Варда
2. метод ближайшего соседа
3. метод полной связи
4. центроидный метод
5. метод невзвешенного попарного среднего

4. Внутрикластерная дисперсия. Используя RStudio решить задачу кластерного анализа методом k-средних для набора данных iris. При решении задачи использовать начальное значение генератора случайных чисел 1234. Задать максимальное число итераций равное 10. Число кластеров задать равным трем. Перед решением задачи выполнить стандартизацию значений параметров с помощью функции scale. Кластеризацию выполнять по четырем переменным набора данных (длине и ширине чашелистика и лепестка). Чему равно суммарное значение внутрикластерной дисперсии для всех трех кластеров?

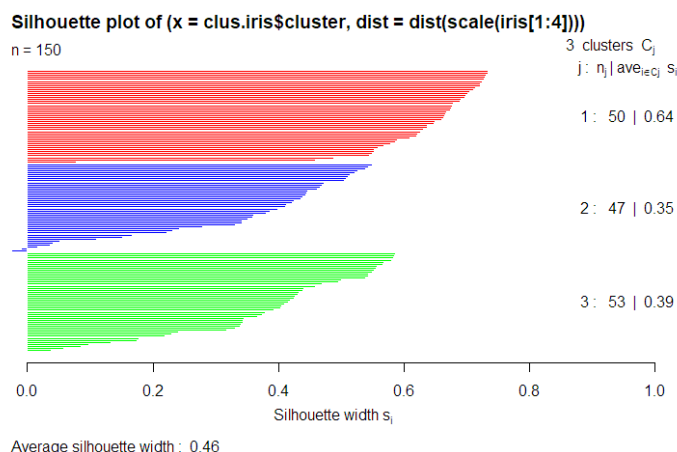
5. Дендрограммы. Набор данных задан двумя последовательностями значений признаков.

`x<-c(2,3,4,5,6,7,8,3,5)`

`y<-c(3,2,5,6,3,4,5,10,7)`.

Чему равно значение расстояния между четвертым и девятым наблюдениями, если все данными были стандартизованы с помощью функции scale, а также при построении дендрограммы использовать метод полной связи (дальнего соседа)? Ответ дать с точностью до двух знаков после запятой

6. Диаграмма силуэтов. После решения задачи кластерного анализа для набора данных iris получена диаграмма силуэтов, которая имеет вид:



Укажите номер кластера, качество формирования которого наилучшее при ее оценке методом силуэтов? Номер лучшего кластера указать числом

7. метод k-средних. Используя RStudio решить задачу кластерного анализа методом k-средних для набора данных iris. При решении задачи использовать начальное значение генератора случайных чисел 1234. Задать максимальное число итераций равное 10. Число кластеров задать равным трем. Перед решением задачи выполнить стандартизацию значений параметров с помощью функции scale. Кластеризацию выполнять по четырем переменным набора данных (длине и ширине чашелистика и лепестка). Чему равно число наблюдений, попавших во второй кластер?

8. Метрики расстояния. Чему равно Евклидово расстояние между объектами, характеризуемыми двумя признаками, если известны значения признаков

2	4
3	5

Ответ дать с точностью до двух знаков после запятой

9. Профили кластеров. Задача кластерного анализа для набора данных iris решена в

JASP. Задача решалась методом двухэтапного кластерного анализа. В результате ее решения для заданного числа кластеров равного трем получена таблица профилей кластеров, имеющая вид:

Проанализировав данную таблицу, укажите номер кластера, качество которого

		Центроиды							
		SEPALL		SEPALW		PETALL		PETALW	
		Среднее	Станд. отклонения	Среднее	Станд. отклонения	Среднее	Станд. отклонения	Среднее	Станд. отклонения
Кластер	1	5,0060	,35249	3,4280	,37906	1,4620	,17366	,2460	,10539
	2	6,1742	,59177	2,8333	,29940	4,8065	,76298	1,6366	,40908
	3	7,4286	,41519	3,3857	,34847	6,2286	,35456	2,2000	,25820
	Объединенный	5,8433	,82807	3,0573	,43587	3,7580	1,76530	1,1993	,76224

наихудшее. В ответе номер указать числом

10. Расстояние Манхеттена. Чему равно расстояние Манхеттена между объектами, характеризуемыми двумя признаками, если известны значения признаков

2	4
3	5

Ответ дать с точностью до целых

11. Расстояние Минковского. Используя набор данных ирисы, с помощью JASP определить расстояние между переменными SEPALL, SEPALW с помощью метрики Минковского второго порядка. При решении задачи выполнить стандартизацию значений переменных с помощью Z-оценки. Задачу решить с помощью иерархической кластеризации, используя метод Варда. Результаты взять из матрицы близости. Ответ дать с точностью до двух знаков после запятой

Типовые оценочные материалы по теме 4

Типовые вопросы для опроса по теме 4

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 алгоритмов деревьев.
5. Как определяется правило остановки построения дерева?
6. Алгоритм CART? Приведите пример его использования.

Типовые задания по теме 4

Задание 1. Построить дерево решений по данным, приведенным в таблице.

Рейтинг	Возраст	Уровень Дохода	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

Задание 2. 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

Задание 3. Расчетно-графическое задание.

Решить задачу классификации для предложенного набора данных. Имеется 15 наборов данных, размещенных в Moodle. Каждый магистрант решает задачу классификации для своего варианта исходных данных. При решении задачи классификации следует решать задачи классификации методами логистической регрессии, деревьев решений, k-ближайших соседей, случайного леса. В Moodle находится пример решения задачи и оформления отчета. В примере приведены скрипты, позволяющие решить данные задачи на языке R, а также с помощью платформы Anaconda Navigator и приложения Orange.

Тест по теме 4

1. Деревья решений. В результате построения дерева решений при решении задачи классификации ирисов получена такая характеристика узлов:

```
library(tree)
iris.tree<-tree(iris[,5]~.,iris[,-5])
iris.tree
1) root 150 329.600 setosa ( 0.33333 0.33333 0.33333 )
2) Petal.Length < 2.45 50 0.000 setosa ( 1.00000 0.00000 0.00000 ) *
3) Petal.Length > 2.45 100 138.600 versicolor ( 0.00000 0.50000 0.50000 )
6) Petal.Width < 1.75 54 33.320 versicolor ( 0.00000 0.90741 0.09259 )
12) Petal.Length < 4.95 48 9.721 versicolor ( 0.00000 0.97917 0.02083 )
24) Sepal.Length < 5.15 5 5.004 versicolor ( 0.00000 0.80000 0.20000 ) *
25) Sepal.Length > 5.15 43 0.000 versicolor ( 0.00000 1.00000 0.00000 ) *
13) Petal.Length > 4.95 6 7.638 virginica ( 0.00000 0.33333 0.66667 ) *
7) Petal.Width > 1.75 46 9.635 virginica ( 0.00000 0.02174 0.97826 )
14) Petal.Length < 4.95 6 5.407 virginica ( 0.00000 0.16667 0.83333 ) *
15) Petal.Length > 4.95 40 0.000 virginica ( 0.00000 0.00000 1.00000 ) *
```

Для шестого узла поставьте соответствие значений параметров узла и их названий

- 54
- 33.320
- 0.00000
- 0.90741
- 0.09259

Варианты ответов;

- Вероятность третьего класса в узле;
- Вероятность первого класса в узле;
- Количество элементов в узле;
- Вероятность второго класса в узле;
- Доля правильно классифицированных в узле;
- Доля неверной классификации в узле

2. Выбор классификатора. Использование различных методов классификации позволяет получить следующие таблицы сопряженности.

Метод линейного дискриминантного анализа

69535	442
4546	477

Метод k ближайших соседей при k=9

69938	39
4973	50

Логистическая регрессия

69770	207
5005	18

Метод опорных векторов

69882	103
4890	125

Выбрать лучший метод по показателю precision

- метод опорных векторов
- метод k-ближайших соседей
- логистическая регрессия
- линейный дискриминантный анализ

3.Использование различных методов классификации позволяет получить следующие таблицы сопряженности.

Метод линейного дискриминантного анализа

69535	442
.	477

Метод k ближайших соседей при k=9

69938	39
4973	50

Логистическая регрессия

69770	207
5005	18

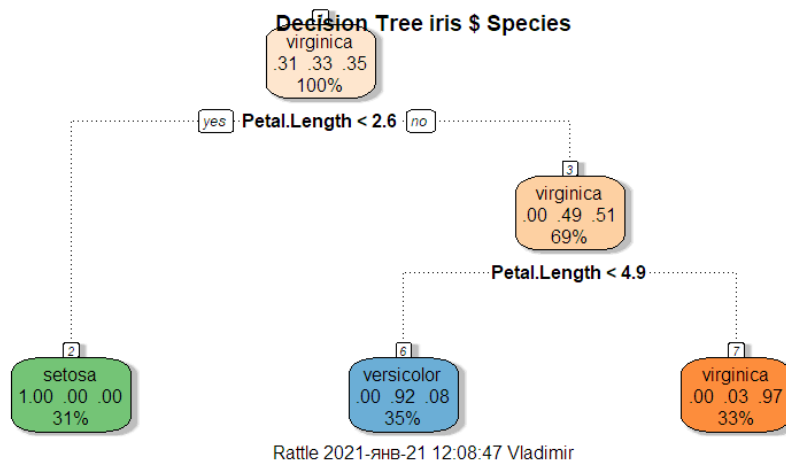
Метод опорных векторов

69882	103
4890	125

Выбрать лучший метод по показателю accuracy

- метод опорных векторов
- метод k-ближайших соседей
- логистическая регрессия
- линейный дискриминантный анализ
- Деревья решений.

4. При решении задачи классификации на наборе данных iris получено дерево решений



Какова вероятность ошибки при классификации цветка, если, длина лепестка больше 2,6 см и меньше 4,9 см. Ответ дать с точностью до двух знаков после запятой. В качестве разделителя использовать запятую

5. Качество классификатора

Таблица сопряженности (матрица путаницы, confusion matrix), полученная при проверке качества бинарного классификатора, имеет вид

25	10
8	14

Строками матрицы являются истинные значения тестируемых объектов, а столбца - результаты тестирования. Чему равно значение показателя точности классификатора (precision)? Ответ дать с точностью до двух знаков после запятой

6. Качество классификатора

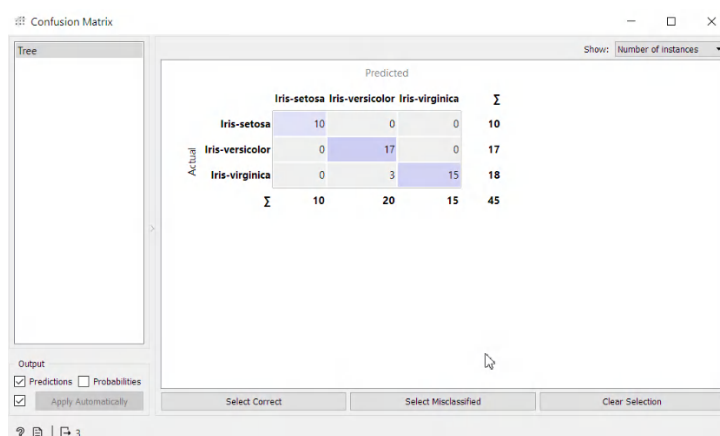
Таблица сопряженности (матрица путаницы, confusion matrix), полученная при проверке качества бинарного классификатора, имеет вид

25	10
8	14

Строками матрицы являются истинные значения тестируемых объектов, а столбца - результаты тестирования. Чему равна точность классификатора (recall)? Ответ дать с точностью до двух знаков после запятой

7. Качество классификации

В результате решения задачи классификации с помощью Orange методом деревьев решений получена матрица



Данная матрица получена при проверке качества классификатора с помощью тестовой выборки. Определить значение показателя AC. Ответ дать с точностью до двух знаков после запятой

8. Таблица сопряженности (confusion matrix). Таблица сопряженности (матрица путаницы, confusion matrix), полученная при проверке качества бинарного классификатора, имеет вид

25	10
8	14

Строками матрицы являются истинные значения тестируемых объектов, а столбца - результаты тестирования. Чему равна точность классификатора (accuracy)? Ответ дать с точностью до двух знаков после запятой

Задание 4. Выполнить анализ данных из набора Kaggle одного из следующих наборов данных, связанных с безопасностью информации.

1. Обнаружение вредоносных программ для Android
<https://www.kaggle.com/datasets/subhajournal/android-malware-detection>

Обнаружение вредоносных программ для Android с помощью машинного обучения — это подход к обнаружению и классификации вредоносных приложений для устройств Android. Одним из точных способов выявления подозрительности приложений является мониторинг сети, к которой подключено андроид-устройство. Машинное обучение — это разновидность искусственного интеллекта, которая фокусируется на разработке компьютерных программ, которые могут получать доступ к данным и использовать их для самостоятельного обучения. Эту технологию можно использовать для построения моделей, которые оценивают поступающие данные, чтобы делать прогнозы и обнаруживать аномалии. Это можно применить для обнаружения вредоносных приложений Android, создав модель, которая ищет закономерности, связанные с поведением вредоносных программ. Модель может использовать различные функции, такие как запрашиваемые разрешения, сделанные вызовы API, сетевая активность и т. д.

Набор данных содержит четыре метки, а именно Android_Adware, Android_Scareware, Android_SMS_Malware и Benign. Набор данных включает 355 630 записей или экземпляров (строк) с 85 столбцами. Данные были удалены из репозитория CIC.

2. Набор данных N-BaIoT для обнаружения атак ботнетов IoT.
<https://www.kaggle.com/datasets/mkashifn/nbaiot-dataset>

Этот набор данных восполняет недостаток общедоступных наборов данных ботнетов, особенно для Интернета вещей. Он предлагает *реальные* данные о трафике, собранные с 9 коммерческих IoT-устройств, достоверно зараженных Mirai и BASHLITE.

Источник:

URL-адрес:

http://archive.ics.uci.edu/ml/datasets/detection_of_IoT_botnet_attacks_N_BaIoT

Характеристики набора данных:

- Многовариантный, последовательный
- Количество экземпляров: 7062606
- Область: Характеристики атрибутов компьютера:
- Реальное количество атрибутов: 115
- Дата сдачи: 2018-03-19
- Связанные задачи: классификация, кластеризация
- Отсутствующие значения: Н/Д

Информация о наборе данных:

Прогнозируемый атрибут:

• Первоначально стремились различать безопасные и вредоносные данные трафика с помощью методов обнаружения аномалий.

• Однако, поскольку вредоносные данные можно разделить на 10 атак, осуществляемых двумя ботнетами, набор данных также можно использовать для многоклассовой классификации: 10 классов атак плюс 1 класс «безвредных».

3.Обнаружение мошенничества

<https://www.kaggle.com/datasets/whenamancodes/fraud-detection/code>

О наборе данных

Важно, чтобы компании, выпускающие кредитные карты, могли распознавать мошеннические транзакции по кредитным картам, чтобы с клиентов не взималась плата за товары, которые они не покупали.

О данных

Набор данных содержит транзакции, совершенные по кредитным картам в сентябре 2013 года держателями карт из Европы. Этот набор данных представляет транзакции, которые произошли за два дня, где у нас есть 492 мошенничества из 284 807 транзакций. Набор данных сильно несбалансирован, на положительный класс (мошенничество) приходится 0,172% всех транзакций.

Он содержит только числовые входные переменные, которые являются результатом преобразования PCA. К сожалению, из соображений конфиденциальности мы не можем предоставить исходные функции и дополнительную справочную информацию о данных. Признаки V1, V2, ... V28 являются основными компонентами, полученными с помощью PCA, единственные признаки, которые не были преобразованы с помощью PCA, — это «Время» и «Количество». Функция «Время» содержит секунды, прошедшие между каждой транзакцией и первой транзакцией в наборе данных. Функция «Сумма» — это сумма транзакции, эту функцию можно использовать, например, для зависимого от затрат обучения. Функция «Класс» — это переменная ответа, которая принимает значение 1 в случае мошенничества и 0 в противном случае.

Учитывая коэффициент дисбаланса классов, мы рекомендуем измерять точность с помощью площади под кривой точности-отзыва (AUPRC). Точность матрицы путаницы не имеет значения для несбалансированной классификации.

4.Набор данных для обнаружения мошенничества с Ethereum

Набор данных, содержащий мошенничество и действительные транзакции Ethereum.

<https://www.kaggle.com/datasets/vagifa/ethereum-frauddetection-dataset>

О наборе данных

Контекст

Этот набор данных содержит строки известных случаев мошенничества и действительных транзакций, совершенных с использованием криптовалюты Ethereum. Этот набор данных **несбалансирован**, поэтому имейте это в виду при моделировании.

Содержание

Вот описание строк набора данных:

- Индекс: порядковый номер строки
- Address: адрес учетной записи ethereum
- ФЛАГ: является ли транзакция мошенничеством или нет
- Среднее время между отправленными транзакциями: среднее время между отправленными транзакциями для аккаунта в минутах.
 - Avg_min_between_received_tnx: Среднее время между полученными транзакциями для аккаунта в минутах.
 - Time_Diff_between_first_and_last(Mins): разница во времени между первой и последней транзакцией.
 - Sent_tnx: общее количество отправленных обычных транзакций.
 - Received_tnx: общее количество полученных обычных транзакций.
 - Number_of_Created_Contracts: общее количество созданных контрактных транзакций.
 - Unique_Received_From_Addresses: Всего уникальных адресов, с которых были получены транзакции.
 - Unique_Sent_To_Addresses20: Всего уникальных адресов, с которых были отправлены транзакции.
 - Min_Value_Received: минимальное значение в эфире, когда-либо полученное
 - Max_Value_Received: максимальное значение в эфире, когда-либо полученное
 - Avg_Value_Received5Среднее значение в эфире, когда-либо полученное
 - Min_Val_Sent: минимальное значение эфира, когда-либо отправленное
 - Max_Val_Sent: максимальное количество эфира, когда-либо отправленное

- Avg_Val_Sent: Среднее значение эфира, когда-либо отправленного
- Min_Value_Sent_To_Contract: Минимальная стоимость эфира, отправленная на контракт.
- Max_Value_Sent_To_Contract: максимальное количество эфира, отправленного на контракт.
- Avg_Value_Sent_To_Contract: средняя стоимость эфира, отправленного на контракты.
- Total_Transactions(Including_Tnx_to_Create_Contract): общее количество транзакций
- Total_Ether_Sent: Общее количество эфира, отправленного на адрес аккаунта
- Total_Ether_Received: Общее количество эфира, полученное для адреса аккаунта.
- Total_Ether_Sent_Contracts: общее количество эфира, отправленного на адреса контрактов.
- Total_Ether_Balance: общий баланс эфира после проведенных транзакций.
- Total_ERC20_Tnxs: общее количество транзакций по передаче токенов ERC20.
- ERC20_Total_Ether_Received: общее количество транзакций, полученных токеном ERC20 в эфире.
- ERC20_Total_Ether_Sent: общее количество транзакций, отправленных токеном ERC20 в эфире.
- ERC20_Total_Ether_Sent_Contract: общий перевод токена ERC20 на другие контракты в эфире.
- ERC20_Uniq_Sent_Addr: количество транзакций с токенами ERC20, отправленных на уникальные адреса учетных записей.
- ERC20_Uniq_Rec_Addr: количество транзакций с токенами ERC20, полученных с уникальных адресов.
- ERC20_Uniq_Rec_Contract_Addr: количество транзакций с токенами ERC20, полученных с уникальных адресов контрактов.
- ERC20_Avg_Time_Between_Sent_Tnx: среднее время между отправленными транзакциями с токеном ERC20 в минутах.
- ERC20_Avg_Time_Between_Rec_Tnx: среднее время между транзакциями, полученными токеном ERC20, в минутах.
- ERC20_Avg_Time_Between_Contract_Tnx: Среднее время токена ERC20 между отправленными транзакциями токена.
- ERC20_Min_Val_Rec: минимальное значение в эфире, полученное от транзакций с токенами ERC20 для учетной записи.
- ERC20_Max_Val_Rec: максимальное значение в эфире, полученное от транзакций с токенами ERC20 для учетной записи.
- ERC20_Avg_Val_Rec: среднее значение в эфире, полученное от транзакций с токенами ERC20 для учетной записи.
- ERC20_Min_Val_Sent: минимальное значение в эфире, отправленное из транзакций с токенами ERC20 для учетной записи.
- ERC20_Max_Val_Sent: максимальное значение в эфире, отправленное из транзакций с токенами ERC20 для учетной записи.
- ERC20_Avg_Val_Sent: среднее значение в эфире, отправленное из транзакций с токенами ERC20 для учетной записи.
- ERC20_Uniq_Sent-Token_Name: количество переданных уникальных токенов ERC20.
- ERC20_Uniq_Rec-Token_Name: количество полученных уникальных токенов ERC20.
- ERC20_Most_Sent-Token_Type: наиболее часто отправляемый токен для учетной записи через транзакцию ERC20.
- ERC20_Most_Rec-Token_Type: Самый популярный токен для учетной записи через транзакции ERC20.

Набор данных классификатора надежности пароля

<https://www.kaggle.com/datasets/bhavikbb/password-strength-classifier-dataset>

Пароли, используемые в анализе, взяты из утечки 000webhost, которая доступна в Интернете. Файлы содержали пароли с еще одной колонкой, т.е. их надежность, основанная на коммерческих измерителях надежности паролей.

5. Анализ памяти вредоносных программ / CIC-MalMem-2022

Набор данных, содержащий разбивку безопасных и вредоносных дампов памяти.

<https://www.kaggle.com/datasets/luccagodoy/obfuscated-malware-memory-2022-cic>

«Обфусцированное вредоносное ПО — это вредоносное ПО, которое прячется, чтобы избежать обнаружения и уничтожения. Набор данных обфусцированного вредоносного ПО предназначен для тестирования методов обнаружения запутанного вредоносного ПО в памяти. Он состоит из программ-шпионов, программ-вымогателей и вредоносных программ-троянов и предоставляет сбалансированный набор данных, который можно использовать для тестирования запутанных систем обнаружения вредоносных программ.

Этот набор данных использует режим отладки для процесса дампа памяти, чтобы процесс дампа не отображался в дампах памяти. Это работает, чтобы представить более точный пример того, что средний пользователь будет работать во время атаки вредоносного ПО».

Источник: <https://www.unb.ca/cic/datasets/malmem-2022.html> .

6.Обнаружение вредоносных программ

<https://www.kaggle.com/datasets/piyushrumao/malware-executable-detection>

Набор данных содержит функции, извлеченные из вредоносных и невредоносных исполняемых файлов Windows. Создан обучающий файл из исполняемых файлов Windows. Файл состоит из 373 образцов, из которых 301 — вредоносные файлы, а остальные 72 — не вредоносные. Набор данных несбалансирован с образцами вредоносных программ больше, чем с обычными образцами. Существует 531 функция, представленная от F_1 до F_531, и столбец меток, указывающий, является ли файл вредоносным или нет. Поскольку представление двоичных шестнадцатеричных имен функций было затруднено, они были представлены как F_1, F_2,... и т. д. То же самое относится к вызовам DLL, которые включены в него. Некоторые функции из 531 функции могут быть исключены, поскольку они имеют очень мало значения, и их необходимо изучить в рамках разработки функций. Метка столбца будет содержать истинное значение исполняемого файла, независимо от того, является ли он вредоносным ПО или нет

7.Вредоносное ПО для сетевого трафика Android <https://www.kaggle.com/datasets/xwolf12/network-traffic-android-malware>

О наборе данных

Введение

Android — одна из самых популярных мобильных операционных систем в мире. Из-за своего технологического влияния, открытого исходного кода и возможности установки сторонних приложений без какого-либо централизованного контроля, Android в последнее время стал мишенью для вредоносных программ. Даже если он включает механизмы безопасности, последние новости о вредоносных действиях и уязвимостях Android указывают на важность продолжения разработки методов и платформ для повышения его безопасности.

5. Оценочные материалы промежуточной аттестации по дисциплине

5.1 Экзамен проводится с применением следующих методов (средств):

Экзамен проводится в компьютерном классе в устной форме. Во время экзамена проверяется уровень знаний, а также уровень умений решать учебные задачи анализа данных с использованием программных приложений. К экзамену магистранты 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 магистрант показывает умение решать практические задачи с использованием интегрированных средств разработки IDE Rstudio, Anaconda navigator (Jupyter Notebook)

Промежуточная аттестация может проводиться устно в ДОТ/письменно / тестирование. Для успешного освоения курса учащемуся рекомендуется ознакомиться с литературой, размещенной в разделе 6, и материалами, выложенными в ДОТ.

5.2.Перечень компетенций с указанием этапов их формирования в процессе освоения образовательной программы. Показатели и критерии оценивания компетенций с учетом этапа их формирования

Таблица 4.2

Код компетенции	Наименование компетенции	Код этапа освоения компетенции	Наименование этапа освоения компетенции
ПКс-2	Способен обосновывать подходы и требования к системе обеспечения информационной безопасности, оценивать уровни безопасности компьютерных систем и сетей	ПКс-2.3	Оценивает уровни безопасности компьютерных систем и сетей

Показатели и критерии оценивания компетенций на различных этапах их формирования

Таблица 4.3

Код компонента компетенции	Показатель оценивания	Критерий оценивания
ПКс-2.3	Приводит оценку уровней безопасности компьютерных систем и сетей	<p>1. Представлены результаты выполнения учебных кейсов по решению задач аналитики данных.</p> <p>2. Приведены скрипты, результаты решения задач разведывательного анализа, интеллектуального анализа, многомерной статистики с использованием статистических пакетов, языков статистической обработки (R, Python).</p> <p>3. Правильно выполнения интерпретация результатов моделирования, их валидация</p> <p>4. Сделаны правильные ответы на поставленные вопросы или тесты</p>

Типовые оценочные материалы промежуточной аттестации.

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 T-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 JASP. 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 k-средних. Привести примеры решения задач в JASP 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 JASP. 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 JASP.

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 T-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 JASP.

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 JASP.

Для проверки гипотезы по критерию Манна-Уитни в R использовать функцию `wilcox.test(y ~ 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 JASP.

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 JASP и в R. Использовать иерархическую кластеризацию и кластеризацию методом k-средних.

Задача 8. Решить задачу кластерного анализа для набора данных USArrests.txt

Задача 9. Решить задачу классификации для набора данных Заемщик. Разделить выборку на обучающую и тестовую. Проверить качество решения задачи с помощью

таблицы сопряженности. Задачу решить в R и в Orange. При решении задачи в Orange выбрать лучший классификатор.

Задача 10. Решить задачу классификации с помощью деревьев решений для файла Ирис в R, Orange, jupyter. Оценить качество классификации. С этой целью всю выборку разделить на две: обучающую и контролирующую, например, с помощью скрипта

```
•# Обучающая выборка
•train <- iris[c(1:25, 51:75, 101:125), 1:4]
•# Тестовая выборка
•test <- iris[c(1:25, 51:75, 101:125)+25, 1:4]
•# Код класса для обучающей выборки
cl <- iris[c(1:25, 51:75, 101:125), 5].
```

Задача 11. Решить задачу классификации для файла Продукт.txt. Задачу решить различными способами. Выбрать лучший классификатор с помощью таблицы сопряженности. Проверку качества классификации выполнить на обучающей выборке. Задачу решить в R и в Orange.

Задача 12. Решить задачу анализа данных (Kwan K.C. et al., 1976) по скорости выведения из организма человека индометацина – одного из наиболее активных противовоспалительных препаратов. Данные находятся в наборе данных Indometh.

- Построить диаграмму зависимости концентрации в организме от времени вывода.
- Построить ящичную диаграмму для каждого испытуемого;
- Выявить аномальные наблюдения;
- Найти средние значения для каждой группы с помощью функции
- Построить гистограммы распределения.
- Построить ящичные диаграммы для каждого
- Найти описательные статистики.
- Проверить гипотезы о законах распределения.

Задача 13. В наборе данных InsectSprays, хранятся результаты эксперимента по изучению эффективности шести видов инсектицидных средств.

-Построить ящички с усами для каждого спрея.

-Построить описательную статистику.

-Проверить гипотезу о значимости отличий между результатами использования инсектицидов A, C; A, F; E, D с помощью ящичной диаграммы.

Задача 14. Решить задачу проверки статистической гипотезы о суточном расходе энергии у худощавых женщин и женщин с избыточным весом, если данные находятся в наборе energy библиотеки ISwR. При проверке гипотезы использовать t-тест и тест Манна-Уитни. Построить ящичные диаграммы, гистограммы распределения. Выявить аномальные наблюдения. Проверить гипотезу о постоянстве дисперсии.

Задача 15. Решить задачу кластерного анализа для набора данных Florida. В наборе данных хранятся результаты голосования на выборах Президента США в 2000 году. Атрибутами являются кандидаты в президенты (Гор, Буш и др. Всего 10 кандидатов. Adams, G. D. and Fastnow, C. F. (2000) *A note on the voting irregularities in Palm Beach, FL. Formerly at <http://madison.hss.cmu.edu/>, but no longer available there.* 67 строк представляют собой 67 населенных пунктов (Набор данных Florida {carData})

Задача 16. Решить задачу кластерного анализа для штатов США. Данные о штатах находятся в файле Штаты.txt.

- -определить склонность к кластеризации;
- -выполнить иерархическую кластеризацию;
- Решить задачу кластеризации методом k-средних;
- Выполнить интерпретацию полученных кластеров

Задачу решить иерархическим методом, а также методом k-средних в JASP и в R.

Задача 17. Решить задачу классификации для набора данных Оператор.xlsx Задачу решить с помощью деревьев решений.

При решении задачи классификации необходимо создать обучающую и контролирующую выборку. При создании контролирующей и обучающей выборки использовать функцию `sample`. Тестовую выборку получить случайно, выбрав примерно 30 процентов от всей выборки. Например, с помощью функции

```
test.num <- sample(1:nrow(sales), 50, replace = FALSE),
```

где `sales` – имя фрейма с данными;

50 – размер тестовой выборки.

Шкала оценки

Оценка результатов производится на основе Положения о текущем контроле успеваемости обучающихся и промежуточной аттестации обучающихся по образовательным программам среднего профессионального и высшего образования в федеральном государственном бюджетном образовательном учреждении высшего образования «Российская академия народного хозяйства и государственной службы при Президенте Российской Федерации», утвержденного Приказом Ректора РАНХиГС при Президенте РФ от 30.01.2018 г. № 02-66 (п.10 раздела 3 (первый абзац) и п.11), а также Решения Ученого совета Северо-западного института управления РАНХиГС при Президенте РФ от 19.06.2018, протокол № 11.

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

Перевод балльных оценок в академические отметки «отлично», «хорошо», «удовлетворительно»

- «Отлично» (A) - от 96 по 100 баллов – теоретическое содержание курса освоено полностью, без пробелов необходимые практические навыки работы с освоенным материалом сформированы, все предусмотренные программой обучения учебные задания выполнены, качество их выполнения оценено максимальным числом баллов.

- «Отлично» (B) - от 86 по 95 баллов – теоретическое содержание курса освоено полностью, без пробелов необходимые практические навыки работы с освоенным материалом сформированы, все предусмотренные программой обучения учебные задания выполнены, качество их выполнения оценено числом баллов, близким к максимальному.

- «Хорошо» (C) - от 71 по 85 баллов – теоретическое содержание курса освоено полностью, без пробелов, некоторые практические навыки работы с освоенным материалом сформированы недостаточно, все предусмотренные программой обучения учебные задания выполнены, качество выполнения ни одного из них не оценено минимальным числом

баллов, некоторые виды заданий выполнены с ошибками.

- «Хорошо» (D) - от 61 по 70 баллов – теоретическое содержание курса освоено полностью, без пробелов, некоторые практические навыки работы с освоенным материалом сформированы недостаточно, большинство предусмотренных программой обучения учебных заданий выполнены, качество выполнения ни одного из них не оценено минимальным числом баллов, некоторые виды заданий выполнены с ошибками.

- «Удовлетворительно» (E) - от 51 по 60 баллов – теоретическое содержание курса освоено частично, но пробелы не носят существенного характера, необходимые практические навыки работы с освоенным материалом в основном сформированы, большинство предусмотренных программой обучения учебных заданий выполнено, некоторые из выполненных заданий выполнены с ошибками.

6. Методические указания для обучающихся по освоению дисциплины

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

Практические занятия предназначены для самостоятельной работы магистрантов по решению конкретных задач анализа данных. Каждое практическое занятие сопровождается домашними заданиями, выдаваемыми магистрантам для решения внеаудиторное время. Для оказания помощи в решении задач имеются тексты практических заданий с условиями задач и вариантами их решения.

Подготовка к лекции заключается в следующем:

- внимательно прочитайте материал предыдущей лекции;
- узнайте тему предстоящей лекции (по тематическому плану, по информации лектора);
- ознакомьтесь с учебным материалом по учебнику и учебным пособиям;
- постарайтесь уяснить место изучаемой темы в своей профессиональной подготовке;
- запишите возможные вопросы, которые вы зададите лектору на лекции.

Подготовка к практическим занятиям:

- внимательно прочитайте материал лекций, относящихся к данному семинарскому занятию, ознакомьтесь с учебным материалом по учебнику и учебным пособиям;
- выпишите основные термины;
- ответьте на контрольные вопросы по семинарским занятиям, готовьтесь дать развернутый ответ на каждый из вопросов;
- уясните, какие учебные элементы остались для вас неясными и постарайтесь получить на них ответ заранее (до семинарского занятия) во время текущих консультаций преподавателя;
- готовиться можно индивидуально, парами или в составе малой группы, последние являются эффективными формами работы;
- рабочая программа дисциплины в части целей, перечню знаний, умений, терминов и учебных вопросов может быть использована вами в качестве ориентира в организации обучения.

Подготовка к экзамену.

К экзамену необходимо готовится целенаправленно, регулярно, систематически и с первых дней обучения по данной дисциплине. Попытки освоить дисциплину в период

зачётно-экзаменационной сессии, как правило, показывают не слишком удовлетворительные результаты. В самом начале учебного курса познакомьтесь со следующей учебно-методической документацией:

- программой дисциплины;
- перечнем знаний и умений, которыми магистрант должен владеть;
- тематическими планами лекций, семинарских занятий;
- контрольными мероприятиями;
- учебником, учебными пособиями по дисциплине, а также электронными ресурсами;
- перечнем вопросов к экзамену.

После этого у вас должно сформироваться четкое представление об объеме и характере знаний и умений, которыми надо будет овладеть по дисциплине. Систематическое выполнение учебной работы на лекциях и семинарских занятиях позволит успешно освоить дисциплину и создать хорошую базу для сдачи экзамена.

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

7.1. Основная литература

1. Афанасьев, Владимир Николаевич. Анализ временных рядов и прогнозирование. - Саратов: Ай Пи Ар Медиа. – 310 с. Текст : электронный. - URL: <http://www.iprbookshop.ru/90196.html> (дата обращения: 12.11.2020). - Режим доступа: для авторизир. пользователей
2. Голоктионова Ю.Г., Ильминская С.А., Илюхина И.Б., Луговской А.М., Лисичкина Н.В. и др. Прогнозирование и планирование в экономике. - Москва: Прометей – 544 с. Текст : электронный. - URL: <http://www.iprbookshop.ru/94511.html> (дата обращения: 01.10.2020). - Режим доступа: для авторизир. пользователей
3. Мастицкий С. Э. (2020) Анализ временных рядов с помощью R. — Электронная книга, адрес доступа: <https://ranalytics.github.io/tsa-with-r>
4. Миркин, Борис Григорьевич. Введение в анализ данных – М.: Юрайт, 2020 – 174 с. Текст : электронный // ЭБС Юрайт [сайт]. — URL: <https://urait.ru/bcode/450262> (дата обращения: 01.10.2020)
5. Мхитарян В. С., Архипова М. Ю., Дуброва Т. А., Миронкина Ю. Н., Сиротин В. П. Анализ данных. – М.: Юрайт, 2020 – 490 с. Текст : электронный // ЭБС Юрайт [сайт]. — URL: <https://urait.ru/bcode/450166> (дата обращения: 29.09.2020)
6. О'Нил, Кэти. Data Science : Инсайдерская информация для новичков. Включая язык R : [пер. с англ.] – СПб. Питер. – 368 с. Текст: электронный. - URL: <http://new.ibooks.ru/bookshelf/359209/reading> (дата обращения: 25.01.2021)
7. Хайндман Р. Дж, Атанасопулос Дж. Прогнозирование: принципы и практика. [Электронный ресурс] –URL: <https://otexts.com/fpp3/>
Все источники основной литературы взаимозаменяемы.

7.2 Дополнительная литература

1. Гусарова Н.Ф, Анализ социальных сетей. Основные понятия и метрики. – СПб.: Университет ИТМО, 2016 – 67 с.
2. Жерон, Орельен. Прикладное машинное обучение с помощью Scikit-Learn и TensorFlow : Концепции, инструменты и техники для создания интеллектуальных систем : полноцветное издание : перевод с английского - ПрМ.: Диалектика. -684 с.
3. Ланц Б. – Машинное обучение на R/ - СПб. : Питер. – 2020 – 464 с.

4. Principles of Econometrics with R [Электронный ресурс] – URL: <https://bookdown.org/ccolonescu/RPoE4/>
5. Люк Д. Анализ сетей (графов) в среде R. Руководство пользователя- М.: ДМК Пресс, 2017 – 250 с.
6. Маккинни У. Python и анализ данных. – М.: ДМК, 2015 – 482 с.
7. Нильсен Эйлин. Практический анализ временных рядов: прогнозирование со статистикой и машинное обучение. –М.: ООО Диалектика – 2021 – 544 с.
8. Наумов Владимир Николаевич. Средства бизнес- аналитики: учеб. пособие / В. Н. Наумов ; Федер. гос. бюджет. образоват. учреждение высш. образования "Рос. акад. нар. хоз-ва и гос. службы при Президенте Рос. Федерации", Сев.-Зап. ин-т упр. - СПб. : СЗИУ - фил. РАНХиГС, 2016. - 107 с.
9. Наумов В.Н. Анализ данных и машинное обучение: методы и инструментальные средства. Федер. гос. бюджет. образоват. учреждение высш. образования "Рос. акад. нар. хоз-ва и гос. службы при Президенте Рос. Федерации", Сев.-Зап. ин-т упр. - СПб. : СЗИУ - фил. РАНХиГС, 2020. - 260 с.
10. Шолле Ф. Глубокое обучение на Python. – СПб.: Питер. 2018. -400 с.
11. Шолле Ф. Глубокое обучение на R. – СПб.: Питер. 2018. -400 с.

7.3. Учебно-методическое обеспечение самостоятельной работы.

1. Положение об организации самостоятельной работы магистрантов федерального государственного бюджетного образовательного учреждения высшего образования «Российская академия народного хозяйства и государственной службы при Президенте Российской Федерации» (в ред. приказа РАНХиГС от 11.05.2016 г. № 01-2211);
2. Положение о курсовой работе (проекте) выполняемой магистрантами федерального государственного бюджетного образовательного учреждения высшего образования «Российская академия народного хозяйства и государственной службы при Президенте Российской Федерации» (в ред. приказа РАНХиГС от 11.05.2016 г. № 01-2211)

7.4. Нормативные правовые документы.

Не используются

7.5. Интернет-ресурсы.

СЗИУ располагает доступом через сайт научной библиотеки <http://nwapa.spb.ru/> к следующим подписным электронным ресурсам:

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>

7.6. Иные источники.

Не используются.

8. 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, Loginom
5.	Средства интеллектуального анализа SQ Lserver. Настройка Analysis services, data mining ad-insfor Office.
6.	Облачные технологии Loginom –server, Google colab
7.	Язык R, Python, Anaconda navigator, Rstudio
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.