Министерство науки и высшего образования Российской Федерации
Федеральное государственное бюджетное образовательное учреждение
высшего образования
**«Владимирский государственный университет
имени Александра Григорьевича и Николая Григорьевича Столетовых»
(ВлГУ)**

Институт экономики и туризма
(Наименование института)

УТВЕРЖДАЮ:
Директор института
Козлов Д.А.
«11» сентября 2023 года

РАБОЧАЯ ПРОГРАММА ДИСЦИПЛИНЫ

Интеллектуальный анализ данных

(НАИМЕНОВАНИЕ ДИСЦИПЛИНЫ)

направление подготовки / специальность

01.03.05 СТАТИСТИКА
(код и наименование направления подготовки (специальности)

направленность (профиль) подготовки

«БИЗНЕС-АНАЛИТИКА**»**
(направленность (профиль) подготовки))

г. Владимир

2023

1. ЦЕЛИ ОСВОЕНИЯ ДИСЦИПЛИНЫ

Цель освоения дисциплины (модуля) «Интеллектуальный анализ данных» являются овладение студентами моделями и методами интеллектуального анализа данных и машинного обучения в задачах поиска информации, обработки и анализа данных, а также приобретение навыков исследователя данных (data scientist) и разработчика математических моделей, методов и алгоритмов анализа данных.

Задачи:

1. Знакомство с основными моделями и методами машинного обучения и разработки данных.

2. Умение адекватно применять указанные модели и методы, а также программные средства, в которых они реализованы.

3. Приобретение опыта анализа реальных данных с помощью изученных методов.

2. МЕСТО ДИСЦИПЛИНЫ В СТРУКТУРЕ ОПОП

Дисциплина «Интеллектуальный анализ данных» относится к обязательной части ОПОП бакалавриата по направлению 01.03.05 Статистика.

3. ПЛАНИРУЕМЫЕ РЕЗУЛЬТАТЫ ОБУЧЕНИЯ ПО ДИСЦИПЛИНЕ

Планируемые результаты обучения по дисциплине, соотнесенные с планируемыми результатами освоения ОПОП (компетенциями и индикаторами достижения компетенций)

Формируемые компетенции (код, содержание компетенции)	Планируемые результаты обучения по дисциплине, в соответствии с индикатором достижения компетенции		Наименование оценочного средства
	Индикатор достижения компетенции *(код, содержание индикатора)*	Результаты обучения по дисциплине	
ОПК – 3 Способен осознанно применять методы математической и дескриптивной статистики для анализа количественных данных, в том числе с применением необходимой вычислительной техники и стандартных компьютерных	ОПК-3.1 Знает современный статистический и математический инструментарий для решения профессиональных задач	Знает современный статистический и математический инструментарий для решения профессиональных задач, используемы во многомерных статистических методах Умеет применять на практике статистический инструментарий Владеет методами многомерного статистического анализа для решения профессиональных задач	Тестовые вопросы Ситуационные задачи Практико-ориентированное задание Эссе
	ОПК-3.2 Умеет использовать информационно-коммуникационные технологии и программные средства для	Знает основные программные средства, необходимые для применения методов многомерного статистического анализа Умеет пользоваться информационными технологиями для анализа данных Владеет основными программными	

программ, содержательно интерпретировать полученные результаты, готовить статистические материалы для докладов, публикаций и других аналитических материалов	анализа количественных данных	средствами для анализа количественных данных при использовании многомерных статистических методов	
	ОПК-3.3 Владеет навыками интерпретации полученных результатов анализа количественных данных и подготовки материалов для докладов, публикаций и других аналитических материалов	Знает способы интерпретации и представления полученных данных Умеет анализировать имеющиеся данные и интерпретировать их Владеет навыками интерпретации полученных результатов анализа количественных данных и подготовки материалов для докладов, публикаций и других аналитических материалов	
ОПК-4. Способен понимать принципы работы современных информационных технологий и использовать их для решения задач профессиональной деятельности	ОПК-4.1 Знает принципы работы современных информационных технологий	Знает основные информационные технологии для осуществления интеллектуального анализа данных Умеет использовать информационные технологии для интеллектуального анализа данных Владеет информационными технологиями для решения профессиональных задач	Тестовые вопросы Ситуационные задачи Практико-ориентированное задание Эссе
	ОПК-4.2 Умеет выбирать информационные технологии	Знает критерии выбора информационных технологий Умеет выбирать информационные технологии Владеет критериями отбора информационных технологи для проведения интеллектуального анализа данных	
	ОПК-4.3 Владеет навыками использования современных информационных технологий при решении задач профессиональной деятельности	Знает специальные и прикладные программы умеет применять информационные технологии на практике Владеет навыками использования современных информационных технологий при решении задач профессиональной деятельности	

4. ОБЪЕМ И СТРУКТУРА ДИСЦИПЛИНЫ

Трудоемкость дисциплины составляет 4 зачетные единицы, 144 часа.

Тематический план
форма обучения – очная

№ п/п	Наименование тем и/или разделов/тем дисциплины	Семестр	Неделя семестра	Контактная работа обучающихся с педагогическим работником				Самостоятельная Работа	Формы текущего контроля успеваемости, форма промежуточной аттестации *(по семестрам)*
				Лекции	Практические занятия	Лабораторные работы	*в форме практической подготовки*		
1	Введение, основные понятия анализа данных	7	1	2	2			5	
2	Математические объекты и методы в анализе данных	7	2,3	2	4			5	
3	Линейная регрессия и классификация	7	4		2			5	
4	Оценивание качества алгоритмов	7	5,6	2	4			5	Рейтинг-контроль №1
5	Логические методы	7	7,8	2	4			5	
6	Композиции алгоритмов	7	9,10	2	4			5	
7	Особенности реальных данных	7	11,12	2	4			5	Рейтинг-контроль №2
8	Анализ частых множеств признаков и ассоциативных правил	7	13,14	2	4			5	
9	Кластеризация данных	7	15,16	2	4			7	
10	Нейронные сети	7	17,18	2	4			7	Рейтинг-контроль №3
Всего за 7 семестр:				18	36			54	Экзамен (36)
Наличие в дисциплине КП/КР									
Итого по дисциплине				18	36			54	Экзамен (36)

Содержание лекционных занятий по дисциплине

Тема 1. Введение, основные понятия анализа данных.

Введение в машинное обучение и анализ данных. Анализ данных в различных прикладных областях. Основные определения. Этапы анализа данных. Постановки задач машинного обучения. Примеры прикладных задач и их типы: классификация, регрессия, ранжирование, кластеризация, поиск структуры в данных.

Тема 3. Линейная регрессия и классификация.

Линейная регрессия. Квадратичная функция потерь и предположение о нормальном распределении шума. Метод наименьших квадратов: аналитическое решение и оптимизационный подход. Стохастический градиентный спуск. Тонкости градиентного спуска: размер шага, начальное приближение, нормировка признаков. Проблема переобучения. Регуляризация. Линейная классификация. Аппроксимация дискретной функции потерь. Отступ. Примеры аппроксимаций, их особенности. Градиентный спуск, регуляризация. Классификация и оценки принадлежности классам. Кредитный скоринг. Логистическая регрессия: откуда берется такая функция потерь и почему она позволяет предсказывать вероятности. Максимизация зазора как пример регуляризации и устранения неоднозначности решения.

Тема 4. Оценивание качества алгоритмов.

Регрессия: квадратичные и абсолютные потери, абсолютные логарифмические отклонения. Примеры использования. Классификация: доля верных ответов, ее недостатки. Точность и полнота, их объединение: арифметическое среднее, минимум, гармоническое среднее (F-мера). Оценки принадлежности классам: площади под кривыми. AUC-ROC, AUC-PRC, их свойства. Оценивание качества алгоритмов. Отложенная выборка, ее недостатки. Оценка полного скользящего контроля. Кросс-валидация. Leave-one-out. Практические особенности кросс-валидации. Стратификация. Потенциальные проблемы с разбиением зависимой или динамической выборки.

Тема 5. Логические методы.

Логические методы и их интерпретируемость. Простейший пример: список решений. Пример решающего списка для задачи фильтрации нежелательных сообщений. Деревья решений. Проблема построения оптимального дерева решений. Жадный алгоритм, основные его параметры. Построение деревьев решений. Критерий ветвления. Выбор оптимального разбиения в задачах регрессии. Сложности выбора разбиения в задаче классификации. Примеры критериев: энтропийный (прирост информации), Джини и их модификации. Критерии завершения построения. Регуляризация и стрижка деревьев.

Тема 6. Композиции алгоритмов.

Простейший пример: уменьшение дисперсии при усреднении алгоритмов методом бутстреп. Блендинг алгоритмов. Понятие смещения и разброса (иллюстрация на примере линейных методов и решающих деревьев). Уменьшение разброса с помощью усреднения. Случайный лес. Оценка out-of-bag.

Тема 7. Особенности реальных данных.

Неполнота и противоречивость. Шумы и выбросы в данных. Методы поиска выбросов. Пропуски в данных, методы их восстановления. Несбалансированные выборки: проблемы и методы борьбы. Задача отбора признаков, примеры подходов.

Тема 8. Анализ частых множеств признаков и ассоциативных правил.

Задача анализа потребительской корзины. Поддержка и достоверность. Частые, замкнутые и максимальные частые множества. Алгоритм Априори. Меры “интересности правил”.

Тема 9. Кластеризация данных.

Простые эвристические подходы. Алгоритм K-Means. Проблема устойчивости результатов и важность грамотной инициализации, алгоритм K-Means++. Выбор числа кластеров. Оценка качества кластеризации.

Тема 10. Нейронные сети.

Типичные задачи. Алгоритм обратного распространения ошибки. Блоки нейронной сети. Архитектуры современных нейронных сетей. Типы нейронных сетей для различных видов данных. Нейронные сети для анализа изображений и видео

Содержание практических занятий по дисциплине

Тема 1. Введение, основные понятия анализа данных.

Введение в машинное обучение и анализ данных. Анализ данных в различных прикладных областях. Основные определения. Этапы анализа данных. Постановки задач машинного обучения. Примеры прикладных задач и их типы: классификация, регрессия, ранжирование, кластеризация, поиск структуры в данных.

Тема 2. Математические объекты и методы в анализе данных.

Линейная алгебра и анализ данных. Линейные пространства, их примеры из машинного обучения (признаки в кредитом скоринге, векторные представления текстов). Коллинеарность и линейная независимость. Скалярное произведение, косинус угла, примеры их применения. Векторы и матрицы, операции над ними. Матричное умножение. Системы линейных уравнений. Обратная матрица. Математический анализ и анализ данных (на примере парной линейной регрессии и МНК). Производная и градиент, их свойства и интерпретации. Типы функций: непрерывные, разрывные, гладкие. Градиентный спуск. Выпуклые функции и их особое место в оптимизации. Теория вероятностей и анализ

данных. Случайные величины. Дискретные и непрерывные распределения, их свойства. Примеры распределений и их важность в анализе данных: биномиальное, пуассоновское, нормальное, экспоненциальное. Характеристики распределений: среднее, медиана, дисперсия, квантили. Пример их использования при генерации признаков. Центральная предельная теорема. Математическая статистика и анализ данных. Оценивание параметров распределений. Метод максимального правдоподобия. Пример использования: анализ текстов и наивный байесовский классификатор. Доверительные интервалы и бутстрэппинг.

Тема 3. Линейная регрессия и классификация.

Линейная регрессия. Квадратичная функция потерь и предположение о нормальном распределении шума. Метод наименьших квадратов: аналитическое решение и оптимизационный подход. Стохастический градиентный спуск. Тонкости градиентного спуска: размер шага, начальное приближение, нормировка признаков. Проблема переобучения. Регуляризация. Линейная классификация. Аппроксимация дискретной функции потерь. Отступ. Примеры аппроксимаций, их особенности. Градиентный спуск, регуляризация. Классификация и оценки принадлежности классам. Кредитный скоринг. Логистическая регрессия: откуда берется такая функция потерь и почему она позволяет предсказывать вероятности. Максимизация зазора как пример регуляризации и устранения неоднозначности решения.

Тема 4. Оценивание качества алгоритмов.

Регрессия: квадратичные и абсолютные потери, абсолютные логарифмические отклонения. Примеры использования. Классификация: доля верных ответов, ее недостатки. Точность и полнота, их объединение: арифметическое среднее, минимум, гармоническое среднее (F-мера). Оценки принадлежности классам: площади под кривыми. AUC-ROC, AUC-PRC, их свойства. Оценивание качества алгоритмов. Отложенная выборка, ее недостатки. Оценка полного скользящего контроля. Кросс-валидация. Leave-one-out. Практические особенности кросс-валидации. Стратификация. Потенциальные проблемы с разбиением зависимой или динамической выборки.

Тема 5. Логические методы.

Логические методы и их интерпретируемость. Простейший пример: список решений. Пример решающего списка для задачи фильтрации нежелательных сообщений. Деревья решений. Проблема построения оптимального дерева решений. Жадный алгоритм, основные его параметры. Построение деревьев решений. Критерий ветвления. Выбор оптимального разбиения в задачах регрессии. Сложности выбора разбиения в задаче

классификации. Примеры критериев: энтропийный (прирост информации), Джини и их модификации. Критерии завершения построения. Регуляризация и стрижка деревьев.

Тема 6. Композиции алгоритмов.

Простейший пример: уменьшение дисперсии при усреднении алгоритмов методом бутстреп. Блендинг алгоритмов. Понятие смещения и разброса (иллюстрация на примере линейных методов и решающих деревьев). Уменьшение разброса с помощью усреднения. Случайный лес. Оценка out-of-bag.

Тема 7. Особенности реальных данных.

Неполнота и противоречивость. Шумы и выбросы в данных. Методы поиска выбросов. Пропуски в данных, методы их восстановления. Несбалансированные выборки: проблемы и методы борьбы. Задача отбора признаков, примеры подходов.

Тема 8. Анализ частых множеств признаков и ассоциативных правил.

Задача анализа потребительской корзины. Поддержка и достоверность. Частые, замкнутые и максимальные частые множества. Алгоритм Априори. Меры “интересности правил”.

Тема 9. Кластеризация данных.

Простые эвристические подходы. Алгоритм K-Means. Проблема устойчивости результатов и важность грамотной инициализации, алгоритм K-Means++. Выбор числа кластеров. Оценка качества кластеризации.

Тема 10. Нейронные сети.

Типичные задачи. Алгоритм обратного распространения ошибки. Блоки нейронной сети. Архитектуры современных нейронных сетей. Типы нейронных сетей для различных видов данных. Нейронные сети для анализа изображений и видео.

5. ОЦЕНОЧНЫЕ СРЕДСТВА ДЛЯ ТЕКУЩЕГО КОНТРОЛЯ УСПЕВАЕМОСТИ, ПРОМЕЖУТОЧНОЙ АТТЕСТАЦИИ ПО ИТОГАМ ОСВОЕНИЯ ДИСЦИПЛИНЫ И УЧЕБНО-МЕТОДИЧЕСКОЕ ОБЕСПЕЧЕНИЕ САМОСТОЯТЕЛЬНОЙ РАБОТЫ СТУДЕНТОВ

5.1. Текущий контроль успеваемости *(рейтинг-контроль 1, рейтинг-контроль 2, рейтинг-контроль 3).*

Рейтинг-контроль №1

1. Какую часть мирового рынка Data Mining занимают услуги или консультации по эффективному внедрению этой технологии для решения актуальных бизнес-задач:

а) более 75% ;

б) около половины;

в) менее 10% рынка.

2. В основу программного продукта Cognos 4Thought положена технология:

а) множественной регрессии;

б) нейронных сетей;

в) деревьев решений.

3. Достаточно высокая стоимость, невозможность добавлять свои функции, сложность подготовки данных, практическое отсутствие в интерфейсе терминов предметной области – это слабые стороны:

а) адаптация программного обеспечения под конкретную задачу;

б) заказ готового решения у фирмы-разработчика;

в) готового программного обеспечения .

4. MAP-сплайны в системе STATISTICA – это:

а) непараметрическая процедура, в работе которой не используется никаких предположений об общем виде функциональных связей между зависимыми и независимыми переменными;

б) процедура, опирающаяся на предположения о типе и накладывающая ограничения на класс зависимостей;

в) параметрическая процедура, основанная на предположениях о виде функциональных связей между зависимыми и независимыми переменными.

5. Готовые алгоритмы, полная конфиденциальность информации, техническая поддержка производителя, общение с другими пользователями пакета — это преимущества использования:

а) адаптация программного обеспечения под конкретную задачу;

б) заказ готового решения у фирмы-разработчика

в) готового программного обеспечения.

6. Cognos 4Thought предназначен для:

а) прогнозирования;

б) моделирования;

в) оба варианта верны;

г) нет правильного ответа.

7. Какое решение в большей мере требует наличия высококвалифицированных специалистов при внедрении и использования инструмента Data Mining:

а) адаптация программного обеспечения под конкретную задачу;

б) заказ готового решения у фирмы-разработчика;

в) использование готового программного обеспечения.

8. Рабочее пространство STATISTICA Data Miner не включает такого элемента:

а) тестирование;

б) анализ данных, моделирование;

в) подготовка, преобразования и очистка данных.

9. Вариант использования адаптированного программного обеспечения Data Mining:

а) имеет неоспоримые преимущества перед использованием готового программного обеспечения;

б) всегда проигрывает перед использованием готового программного обеспечения;

в) имеет как сильные, так и слабые стороны.

10. Инструментальное средство для оперативного анализа данных и формирования отчетов по OLAP-технологии:

а) Cognos Impromptu;

б) Cognos PowerPlay;

в) Cognos Scenario.

11. На этапе подготовки данных:

а) специалисты компании Разработчика подготавливают данные для их дальнейшего анализа;

б) специалисты компании Разработчика и Заказчика подготавливают данные для их дальнейшего анализа;

в) специалисты компании Заказчика подготавливают данные для их дальнейшего анализа.

12. На каких этапах 4Thought поддерживает анализ данных:

а) сбор данных;

б) преобразование данных;

в) интерпретация модели;

г) на всех этапах.

13. Преимуществом использования адаптированного программного обеспечения Data Mining по сравнению с готовыми программными продуктами и их самостоятельным использованием является:

а) наличие терминов предметной области;

б) сложность подготовки данных;

в) полная конфиденциальность информации.

14. Охарактеризуйте систему Cognos Scenario:

а) является средством оперативного анализа данных;

б) формирует отчеты по OLAP–технологии;

в) интеллектуальное инструментальное средство поиска данных.

15. Преимуществом использования адаптированного программного обеспечения Data Mining по сравнению с готовыми программными продуктами и их самостоятельным использованием является:

а) адаптированность;

б) не требуется дописывать программный код;

в) сложность подготовки данных.

16. Охарактеризуйте систему Cognos Scenario:

а) формирует отчеты по OLAP–технологии;

б) позволяет руководителям выявлять скрытые тенденции и модели бизнеса;

в) является средством оперативного анализа данных.

17. На этапе первичного исследования данных:

а) всю работу осуществляет заказчик;

б) со стороны заказчика может потребоваться максимальное участие;

в) со стороны заказчика может потребоваться лишь минимальное участие.

18. Модуль Oracle Data Mining доступен из таких редакций:

а) Personal Edition;

б) Enterprise Edition;

в) OneStandard Edition.

19. Постановка бизнес-задачи – это этап, который:

а) формулирует конкретные бизнес-задачи, и они уже не могут быть изменены;

б) формулирует конкретные бизнес-задачи, и они могут быть изменены в ходе прохождения именно этого цикла;

в) формулирует конкретные бизнес-задачи, и они не могут быть изменены в ходе прохождения именно этого цикла.

20. Архитектура хранилища типа «звезда» в Deductor называется:

а) сценарием;

б) процессом;

в) проектом.

Рейтинг-контроль №2

1. «Извлечение полезных сведений невозможно без хорошего понимания сути данных», верно ли утверждение:

а) верно;

б) неверно. Технологии не нужно понимание данных.

в) неверно. Технологии Data Mining не нужен аналитик, поэтому понимание кем либо данных — излишне.

2. Большинство аналитических методов, используемые в технологии Data mining – это:

а) новейшие математические алгоритмы и методы;

б) известные математические алгоритмы и методы;

в) классические статистические методы.

3. Выберите характеристику, наиболее подходящую для Data Mining:

а) подходит для понимания ретроспективных данных;

б) подходит для обобщения ретроспективных данных;

в) опирается на ретроспективные данные для получения ответов на вопросы о будущем.

4. Частью какой из перечисленных стадий является валидация закономерностей:

а) свободный поиск;

б) анализ исключений;

в) прогностическое моделирование.

5. Какая из перечисленных ниже групп методов достаточно часто использует для выявления взаимосвязей в данных концепцию усреднения по выборке:

а) Data Mining;

б) OLAP;

в) статистические методы.

6. На стадии свободного поиска осуществляется:

а) использование выявленных закономерностей для предсказания неизвестных значений;

б) выявление закономерностей;

в) анализ исключений.

7. В результате использования инструментов Data Mining пользователь может:

а) получить гипотезы о взаимосвязях в данных, самостоятельно выдвинутые инструментом Data Mining;

б) получить подтверждение или опровержение гипотез, выдвинутых пользователем;

в) оба варианта верны;

г) нет верного ответа.

8. Нейронные сети относятся к группам:

а) методов на основе уравнений;

б) статистических методов;

в) методов кросс-табуляции.

9. Закономерности, найденные в процессе использования технологии Data Mining должны обладать такими свойствами:

а) быть очевидными;

б) чем больше найдено закономерностей, тем лучше;

в) быть неочевидными.

10. Какой из перечисленных ниже пунктов не является названием стадии Data Mining:

а) свободный поиск;

б) индукция правил;

в) анализ исключений.

11. Закономерности, найденные в процессе использования технологии Data Mining должны обладать такими свойствами:

а) быть объективными;

б) быть очевидными;

в) чем больше найдено закономерностей, тем лучше.

12. Какие из перечисленных ниже пунктов являются названиями стадий Data Mining:

а) прогностическое моделирование;

б) свободный поиск;

в) оба варианта верны;

г) нет верного ответа.

13. Инструменты Data Mining:

а) могут самостоятельно строить гипотезы о взаимосвязях в данных;

б) могут самостоятельно строить гипотезы о взаимосвязях в данных, которые обязательно подтверждаются;

в) не могут самостоятельно строить гипотезы о взаимосвязях в данных.

14. Для какой шкалы применимы только такие операции как равно, не равно, больше, меньше:

а) номинальная шкала;

б) интервальная шкала;

в) порядковая шкала.

15. Если сравнивать Data Mining, машинное обучение и статистику, какая из дисциплин сконцентрирована на едином процессе анализа данных, включает очистку данных, обучение, интеграцию и визуализацию результатов:

а) Data Mining;

б) статистика;

в) машинное обучение.

Рейтинг-контроль №3

1. К какой категории данных относится вес измеряемых объектов:

а) дискретным данным;

б) непрерывным данным;

в) оба варианта верны;

г) нет верного ответа.

2. Назовите фактор, обусловивший возникновение и развитие Data Mining:

а) совершенствование аппаратного и программного обеспечения;

б) совершенствование технологий хранения и записи данных;

в) оба варианта верны;

г) нет верного ответа.

3. Для какой шкалы применимы только такие операции как равно и не равно:

а) порядковая шкала;

б) номинальная шкала;

в) интервальная шкала.

4. В процессе работы Data Mining программы пользователь может получить такие результаты:

а) только статистически достоверные результаты;

б) только верные результаты, ложные выводы исключены;

в) большой процент ложных, недостоверных или бессмысленных результатов.

5. Такие данные как температура воздуха относятся к:

а) дискретным данным;

б) непрерывным данным;

в) оба варианта верны;

г) нет верного ответа.

6. Data Mining — это процесс обнаружения в сырых данных знаний, необходимых для:

а) принятия решений в различных сферах человеческой деятельности;

б) увеличения стоимости анализа данных;

в) замены аналитика в процессе принятия решений.

7. Объект описывается как:

а) свойство, характеризующее объект;

б) набор атрибутов;

в) поле таблицы.

8. Подготовка данных в процессе Data Mining является:

а) необязательным этапом работы;

б) может вообще отсутствовать;

в) существенным этапом работы.

9. Свойство, характеризующее объект:

а) Данные;

б) Атрибут;

в) Инструменты Data Mining.

10. Какая из перечисленных дисциплин более сосредоточена на теории проверки гипотез:

а) Data Mining;

б) визуализация;

в) статистика.

11. Процесс обнаружения в сырых данных знаний, необходимых для принятия решений в различных сферах человеческой деятельности:

а) Данные;

б) Data Mining;

в) Атрибут.

12. Data Mining — это процесс обнаружения в сырых данных:

а) практических закономерностей;

б) большого количества закономерностей;

в) ранее сформулированных гипотез.

13. Номинальная шкала – это шкала:

а) содержащая категории, которые могут упорядочиваться;

б) содержащая только две категории;

в) содержащая только категории, которые не могут упорядочиваться.

14. Data Mining — это процесс обнаружения в сырых данных:

а) объективных закономерностей;

б) ранее сформулированных гипотез;

в) большого количества закономерностей.

15. Объектом не является:

а) строка таблицы;

б) переменная;

в) запись.

5.2. Промежуточная аттестация по итогам освоения дисциплины производится в виде экзамена, который включает в себя ответы на теоретические вопросы.

1. Введение в машинное обучение и анализ данных.

2. Анализ данных в различных прикладных областях.

3. Основные определения.

4. Этапы анализа данных.

5. Постановки задач машинного обучения.

6. Примеры прикладных задач и их типы.

7. Линейная алгебра и анализ данных.

8. Линейные пространства, их примеры из машинного обучения.

9. Коллинеарность и линейная независимость.

10. Скалярное произведение, косинус угла, примеры их применения.

11. Векторы и матрицы, операции над ними.

12. Матричное умножение.

13. Системы линейных уравнений.

14. Обратная матрица.

15. Математический анализ и анализ данных (на примере парной линейной регрессии и МНК).

16. Производная и градиент, их свойства и интерпретации.

16. Типы функций: непрерывные, разрывные, гладкие.

17. Градиентный спуск.

18. Выпуклые функции и их особое место в оптимизации.

19. Теория вероятностей и анализ данных.

20. Случайные величины.

21. Дискретные и непрерывные распределения, их свойства.

22. Примеры распределений и их важность в анализе данных: биномиальное, пуассоновское, нормальное, экспоненциальное.

23. Характеристики распределений: среднее, медиана, дисперсия, квантили.

24. Пример их использования при генерации признаков.

25. Центральная предельная теорема.

26. Математическая статистика и анализ данных.

27. Оценивание параметров распределений.

28. Метод максимального правдоподобия.

29. Доверительные интервалы и бутстрэппинг.

30. Линейная регрессия.

31. Квадратичная функция потерь и предположение о нормальном распределении шума.

32. Метод наименьших квадратов: аналитическое решение и оптимизационный подход.

33. Стохастический градиентный спуск.

34. Тонкости градиентного спуска: размер шага, начальное приближение, нормировка признаков.

35. Проблема переобучения.

36. Регуляризация.

37. Линейная классификация.

38. Аппроксимация дискретной функции потерь.

39. Логистическая регрессия.

40. Регрессия: квадратичные и абсолютные потери, абсолютные логарифмические отклонения.

41. Примеры использования.

42. Классификация: доля верных ответов, ее недостатки.

43. Точность и полнота, их объединение: арифметическое среднее, минимум, гармоническое среднее (F-мера).

44. Оценки принадлежности классам: площади под кривыми. AUC-ROC, AUC-PRC, их свойства.

45. Оценивание качества алгоритмов.

46. Отложенная выборка, ее недостатки.

47. Оценка полного скользящего контроля.

48. Кросс-валидация. Leave-one-out.

49. Практические особенности кросс-валидации.

50. Стратификация.

51. Потенциальные проблемы с разбиением зависимой или динамической выборки.

52. Логические методы и их интерпретируемость.

53. Деревья решений. Проблема построения оптимального дерева решений.

54. Жадный алгоритм, основные его параметры.

55. Построение деревьев решений. Критерий ветвления.

56. Выбор оптимального разбиения в задачах регрессии. Сложности выбора разбиения в задаче классификации.

57. Примеры критериев: энтропийный (прирост информации),

58. Джини и их модификации.

59. Критерии завершения построения.

60. Регуляризация и стрижка деревьев.

61. Уменьшение дисперсии при усреднении алгоритмов методом бутстреп.

62. Блендинг алгоритмов.

63. Понятие смещения и разброса (иллюстрация на примере линейных методов и решающих деревьев).

64. Уменьшение разброса с помощью усреднения.

65. Случайный лес.

66. Оценка out-of-bag.

67. Шумы и выбросы в данных.

68. Методы поиска выбросов.

69. Пропуски в данных, методы их восстановления.

70. Несбалансированные выборки: проблемы и методы борьбы.

71. Задача отбора признаков, примеры подходов.

72. Задача анализа потребительской корзины.

73. Поддержка и достоверность.

74. Частые, замкнутые и максимальные частые множества.

75. Алгоритм Априори.

76. Меры “интересности правил”.

77. Алгоритм K-Means.

78. Проблема устойчивости результатов и важность грамотной инициализации, алгоритм K-Means++.

79. Выбор числа кластеров.

80. Оценка качества кластеризации.

81. Нейронные сети. Типичные задачи.

82. Алгоритм обратного распространения ошибки.

83. Блоки нейронной сети.

84. Архитектуры современных нейронных сетей.

85. Типы нейронных сетей для различных видов данных.

86. Нейронные сети для анализа изображений и видео

5.3. Самостоятельная работа обучающегося производится в виде решения задач, докладов (эссе), презентаций.

Тематика задач для самостоятельной работы.

Задача 1. Изучение опыта применения методов кластеризации данных.

Задача 2. Программирование методов кластеризации данных.

Задача 3. Лингвистическое резюмирование результатов кластеризации данных.

Задача 4. Прогнозирование на основе статистического подхода.

Задача 5. Прогнозирование на основе нечеткого подхода.

Требования по подготовке доклада (эссе).

Эссе - это самостоятельная письменная работа на тему, предложенную преподавателем. Оно должен содержать:

- введение, содержащее постановку проблемы;
- основную часть, содержащую логически выдержанное изложение темы (предпосылок и путей решения поставленной проблемы);
- краткие выводы, обобщающие позицию автора по проблеме;
- список использованной литературы (указывается только та литература, которой фактически пользовался автор; все случаи использования источников - цитаты,

сведения, оценки и т.д. - отмечаются ссылками в виде сносок или примечаний с указанием страниц источника).

Объем эссе должен составлять 7-10 страниц (до 4 тыс. слов) печатного текста (шрифт Times, размер 12, полуторный интервал). Включение в эссе материалов, не имеющих прямого отношения к теме, а также источников, не указанных в базовом списке литературы (в частности, текстов из Интернета), служит основанием для признания работы не соответствующей требованиям или существенного снижения общей оценки.

Эссе оценивается по следующим критериям:

- самостоятельность выполнения работы, способность аргументировано защищать основные положения и выводы. Эссе, выполненное несамостоятельно, по другим критериям не оценивается;
- соответствие формальным требованиям: структура, наличие списка литературы, сносок, грамотность изложения;
- способность сформулировать проблему;
- уровень освоения темы и изложения материала: обоснованность отбора материала, использование первичных источников, способность самостоятельно осмысливать выявленные факты, логика изложения**;**
- четкость и содержательность выводов.

Тематика эссе

1. Дескриптивный анализ данных.
2. Полный статистический дескрипт вероятностной структуры и параметров данных.
3. Полиномиальные и стохастические модели. Оценка параметров движения.
4. Моделирование динамических стохастических процессов в среде Матлаб (R).
5. Регрессионная оценка параметров движения.
6. Метод статистических испытаний.
7. Имитационное моделирование случайных событий и процессов. Метод Монте-Карло.
8. Многомерные задачи классификации и распознавания.
9. Основные технологии статистической классификации многомерных случайных объектов средствами ИАД.
10. Прогнозирование на основе фильтра Калмана.
11. Прогнозирование состояния сложных динамических систем статистическими и другими средствами ИАД.

12. ИНС с обратным распространением ошибки.

13. Моделирование двухслойной ИНС с обратным распространением ошибки и применение в задачах распознавания.

14. Модели и прогнозирование хаотических процессов.

15. Моделирование хаотических процессов средствами ИАД.

16. Анализ возможности построения прогноза в хаотических средах.

Требования по подготовке презентации

Общие требования к презентации:

- Презентация не должна быть меньше 10 слайдов.
- Первый лист – это титульный лист, на котором обязательно должны быть представлены: название проекта; название выпускающей организации; фамилия, имя, отчество автора; вуз, где учится автор проекта и его группа.
- Следующим слайдом должно быть содержание, где представлены основные части (моменты) презентации. Желательно, чтобы из содержания по гиперссылке можно перейти на необходимую страницу и вернуться вновь на содержание.
- Дизайн-эргономические требования: сочетаемость цветов, ограниченное количество объектов на слайде, цвет текста.
- Презентация не может состоять из сплошного не структурированного текста.
- Последними слайдами урока-презентации должны быть глоссарий и список литературы.

Создание презентации состоит из трех этапов:

I. Планирование презентации – это многошаговая процедура, включающая определение целей, формирование структуры и логики подачи материала. Планирование презентации включает в себя:

1. Определение целей.
2. Определение основной идеи презентации.
3. Подбор дополнительной информации.
4. Планирование выступления.
5. Создание структуры презентации.
6. Проверка логики подачи материала.
7. Подготовка заключения.

II. Разработка презентации – методологические особенности подготовки слайдов презентации, включая вертикальную и горизонтальную логику, содержание и соотношение текстовой и графической информации.

III. Репетиция презентации – это проверка и отладка созданной презентации.

В оформлении презентаций выделяют два блока: оформление слайдов и представление информации на них. Для создания качественной презентации необходимо соблюдать ряд требований, предъявляемых к оформлению данных блоков.

Оформление слайдов:

Стиль	· Соблюдайте единый стиль оформления · Избегайте стилей, которые будут отвлекать от самой презентации. ·Вспомогательная информация (управляющие кнопки) не должны преобладать над основной информацией (текстом, иллюстрациями).
Фон	Для фона предпочтительны холодные тона
Использование цвета	· На одном слайде рекомендуется использовать не более трех цветов: один для фона, один для заголовка, один для текста. · Для фона и текста используйте контрастные цвета. · Обратите внимание на цвет гиперссылок (до и после использования).
Анимационные эффекты	· Используйте возможности компьютерной анимации для представления информации на слайде. · Не стоит злоупотреблять различными анимационными эффектами, они не должны отвлекать внимание от содержания информации на слайде.

Представление информации:

Содержание информации	· Используйте короткие слова и предложения. · Минимизируйте количество предлогов, наречий, прилагательных. · Заголовки должны привлекать внимание аудитории.
Расположение информации на странице	· Предпочтительно горизонтальное расположение информации. · Наиболее важная информация должна располагаться в центре экрана. · Если на слайде располагается картинка, надпись должна располагаться под ней.
Шрифты	· Для заголовков – не менее 24. · Для информации не менее 18. · Шрифты без засечек легче читать с большого расстояния. · Нельзя смешивать разные типы шрифтов в одной презентации. · Для выделения информации следует использовать жирный шрифт, курсив или подчеркивание. · Нельзя злоупотреблять прописными буквами (они читаются хуже строчных).
Способы выделения информации	Следует использовать: · рамки; границы, заливку; · штриховку, стрелки; · рисунки, диаграммы, схемы для иллюстрации наиболее важных фактов.
Объем информации	· Не стоит заполнять один слайд слишком большим объемом информации: люди могут единовременно запомнить не более трех фактов, выводов, определений. · Наибольшая эффективность достигается тогда, когда ключевые пункты отображаются по одному на каждом отдельном слайде.
Виды слайдов	Для обеспечения разнообразия следует использовать разные виды слайдов: · с текстом; · с таблицами; · с диаграммами.

Тематика презентаций

1. Пакеты Numpy, Scipy, математические операции в них.
2. Пакет Pandas, работа с данными в нем.
3. Линейные методы классификации и регрессии.
4. Метрики качества алгоритмов машинного обучения, кросс-валидация.
5. Деревья решений, их построение.
6. Композиции алгоритмов. Случайные леса.
7. Работа с реальными данными. Предобработка признаков.
8. Кластеризация реальных данных.
9. Поиск частых множеств и ассоциативных правил.

Фонд оценочных материалов (ФОМ) для проведения аттестации уровня сформированности компетенций обучающихся по дисциплине оформляется отдельным документом.

6. УЧЕБНО-МЕТОДИЧЕСКОЕ И ИНФОРМАЦИОННОЕ ОБЕСПЕЧЕНИЕ ДИСЦИПЛИНЫ

6.1. Книгообеспеченность

Наименование литературы: автор, название, вид издания, издательство	Год изда ния	КНИГООБЕСПЕЧЕННОСТЬ
		Наличие в электронной библиотеке ВлГУ
1	2	3
Основная литература		
1. **Анализ данных при помощи Microsoft Power BI и Power Pivot для Excel : практическое руководство** / А. Феррари, М. Руссо ; пер. с англ. А. Ю. Гинько. - Москва : ДМК Пресс, 2020. - 288 с. - ISBN 978-5-97060-858-6	2020	https://znanium.com/catalog/document?id=367157
2. **Статистические методы анализа данных : учебное пособие** / В. Н. Клячкин, Ю. Е. Кувайскова, В. А. Алексеева. - Москва : Финансы и Статистика, 2021. - 240 с. - ISBN 978-5-00184-057-2.	2021	https://znanium.com/catalog/product/1831431
3. **Данные: хранение и обработка : учебник** / Э. Г. Дадян. — Москва : ИНФРА-М, 2021. — 205 с. — (Высшее образование: Бакалавриат). - ISBN 978-5-16-016447-2	2021	https://znanium.com/catalog/product/1149101
Дополнительная литература		
1 **Практический подход к проектированию баз данных : учебное пособие** / М. В. Махмутова. - 2-е изд., стер. - Москва : ФЛИНТА, 2023. - 159 с. - ISBN 978-5-9765-3694-4.	2023	https://znanium.com/catalog/product/2091322
2. **Базы данных : учебник** / Л.И. Шустова, О.В. Тараканов. — Москва : ИНФРА-М, 2023. — 304 с. + Доп. материалы [Электронный ресурс]. — (Высшее образование: Бакалавриат). — DOI 10.12737/11549. - ISBN 978-5-16-010485-0..	2023	https://znanium.com/catalog/product/1986697

3. **Базы данных: практикум : учебно-практическое пособие** / А. С. Копырин. - Москва : ФЛИНТА, 2021. - 106 с. - ISBN 978-5-9765-4752-0.	2021	https://znanium.com/catalog/product/1851992

6.2. Периодические издания

1. Журнал «КомпьютерПресс» http://www.compress.ru
2. Журнал «ComputerWorld Россия» http://www.osp.ru/cw
3. Журнал «PC Week / RE (Компьютерная неделя)» http://www.pcweek.ru
4. Журнал «Информационное общество» http://www.infosoc.iis.ru
5. Журнал «CRN / RE (ИТ-бизнес)» http://www.crn.ru
6. Журнал «Вопросы статистики». Входит в список ВАК.
7. Журнал «Учет и статистика».

6.3. Интернет-ресурсы

1. http://www.spssbase.com/ Иллюстрированный самоучитель по SPSS
2. http://www.spss.ru Официальный сайт российского офиса компании SPSS

7. МАТЕРИАЛЬНО-ТЕХНИЧЕСКОЕ ОБЕСПЕЧЕНИЕ ДИСЦИПЛИНЫ

Учебная аудитория, компьютерный класс с выходом в Internet для проведения занятий лекционного и семинарского типа, текущего и промежуточного контроля, групповых и индивидуальных консультаций.

Демонстрационное оборудование: 13 компьютеров kraftwey с мышками и клавиатурой, проектор Panasonic, экран, ноутбук Asus X58Le, 12 станций ThinkCentre M70, звуковые колонки Genius SW-HF5.1, доска настенная.

Количество посадочных мест: 18.

Расположена по адресу: 600005, Российская Федерация, Владимирская область, г.о. город Владимир, г. Владимир, ул. Горького, д. 79, 2 этаж учебного корпуса № 6, 52,5 м^2, № 2.

Перечень используемого лицензионного программного обеспечения: пакет MS-Office, Microsoft Windows, 7-Zip, AcrobatReader; СПС «Консультант Плюс» (инсталированный ресурс ВлГУ).

Примечание

В соответствии с нормативно-правовыми актами для инвалидов и лиц с ограниченными возможностями здоровья при необходимости тестирование может быть

проведено только в письменной или устной форме, а также могут быть использованы другие материалы контроля качества знаний, предусмотренные рабочей программой дисциплины.

Рабочую программу составил к.ф.м.н. доцент Крылов В.Е.

Рецензент (представитель работодателя):
Председатель счетной палаты Владимирской обл., кандидат экономических наук Тулякова И.В.

Программа рассмотрена и одобрена на заседании кафедры БИиЭ
протокол № 1 от 30 августа 2023 года.
Заведующий кафедрой: д.э.н., профессор Тесленко И.Б.

Рабочая программа рассмотрена и одобрена на заседании учебно-методической комиссии направления 01.03.05 Статистика
протокол № 1 от «05» сентября 2023 года.
Председатель комиссии: к.э.н., доцент Яресь О.Б.