МИНОБРНАУКИ РОССИИ
ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ
«ВОРОНЕЖСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ»
(ФГБОУ ВО «ВГУ»)

УТВЕРЖДАЮ
Заведующий кафедрой
международной экономики и
внешнеэкономической деятельности


Ендовицкая Е.В.
20.03.2024 г.

РАБОЧАЯ ПРОГРАММА УЧЕБНОЙ ДИСЦИПЛИНЫ
Б1.В.08 Аналитические системы Big Data

1. Шифр и наименование направления подготовки/специальности:

38.03.01 «Экономика»

2. Профиль подготовки/специализация: Мировая экономика

3. Квалификация (степень) выпускника: бакалавр

4. Форма обучения: очная

5. Кафедра, отвечающая за реализацию дисциплины:

Международной экономики и внешнеэкономической деятельности

6. Составители программы:
Гайворонская Светлана Анатольевна, кандидат технических наук, доцент

7. Рекомендована:
НМС факультета международных отношений протокол № 3 от 20.03.2024 г.

8. Учебный год: 2026 – 2027 **Семестр:** 6

9. **Цели и задачи учебной дисциплины:**
Цель учебной дисциплины: формирование у студентов профессиональной компетенции в области разработки и использования систем обработки и анализа больших массивов данных. Приобретенные знания позволят сформировать у студентов практические навыки по: постановке задач анализа данных; предварительной обработке данных; визуализации данных; применению методов интеллектуального анализа данных к большим массивам данных.
Задачи учебной дисциплины:
- формирование у обучающихся знаний о технологиях подготовки, хранения, обработки и анализа больших данных;

- формирование у обучающихся навыков применения статистических и математических методов для анализа больших объемов информации;
- формирование у обучающихся представления о базовых принципах сбора, обработки и анализа маркетинговой информации для принятия управленческих решений.

10. Место учебной дисциплины в структуре ООП: дисциплина относится к блоку Б1 учебного плана, включена в его вариативную часть, является обязательной

11. Планируемые результаты обучения по дисциплине/модулю (знания, умения, навыки), соотнесенные с планируемыми результатами освоения образовательной программы (компетенциями выпускников):

Код	Название компетенции	Код	Индикатор(ы)	Планируемые результаты обучения
ПК-4	Способен применять современные информационные технологии для решения профессиональных задач	ПК-4.1	Отбирает информацию для формирования данных для экономических процессов и явлений с использованием IT-технологий.	Знать: принципы осуществления экономического анализа с применением современных программных продуктов. Уметь: осуществлять поиск информации с применением специализированных программных продуктов. Владеть: навыками отбора информации с использованием IT- технологий.
		ПК- 4.3	Применяет IT - технологии для систематизации и анализа массива данных в профессиональной деятельности.	Знать: стандарты, модели и алгоритмы обработки и анализа больших данных. Уметь: формально описывать задачи, возникающие в бизнес-аналитике, и сводить их к математическим или технологическим задачам. Владеть: навыками обработки экономической информации с применением IT- технологий.
		ПК - 4.4	Интерпретирует результаты обработки экономических данных для принятия управленческих решений	Знать: основные принципы интерпретации результатов анализа больших данных Уметь: осуществлять выбор процедур обработки информации в зависимости от природы используемых данных и интерпретировать результаты анализа больших данных. Владеть: навыками применения соответствующих методов обработки данных и интерпретации полученных результатов

12. Объем дисциплины в зачетных единицах/час. 3/108.

Форма промежуточной аттестации: зачет.

13. Виды учебной работы

Вид учебной работы		Трудоемкость		
		Всего	По семестрам	
			5 семестр	6 семестр
Контактная работа		32	-	32
в том числе:	лекции	-	-	-
	практические	-	-	-
	лабораторные	32	-	32
	курсовая работа	-	-	-
Самостоятельная работа		76	-	76
Промежуточная аттестация		-	-	-
Итого:		108	-	108

13.1. Содержание дисциплины

№ п/п	Наименование раздела дисциплины	Содержание раздела дисциплины	Реализация раздела дисциплины с помощью онлайн-курса, ЭУМК *
Лабораторные занятия			
1	Введение в анализ больших данных. Обзор источников информации.	Основные определения, термины, задачи анализа больших данных. Понятие Data Mining. Когнитивный анализ данных. Обзор источников информации для Big Data (открытые источники информации: статистические сборники, опубликованные отчеты и результаты исследований). Методики сбора данных.	https://edu.vsu.ru
2	Технологии хранения и обработки больших данных.	Обзор технологий хранения больших данных. Базы данных. Системы управления базами данных. Модели данных. Подготовка исходных данных для анализа: первичная обработка и визуализация имеющихся данных	https://edu.vsu.ru
3	Статистические методы анализа данных.	Основные понятия математической статистики. Методы анализа данных: дескриптивная статистика, параметрические, непараметрические, номинальные методы	https://edu.vsu.ru
4	Современные программные средства анализа больших данных.	Обзор современных популярных программных средства анализа данных: Statistica, SPSS, Excel, R-Studio и другие; их преимущества и недостатки.	https://edu.vsu.ru
5	Сбор и хранение больших данных.	Поиск источников информации в сети Интернет: открытые и закрытые источники данных. Портал открытых данных РФ.	https://edu.vsu.ru
6	Методы обработки и анализа данных	Группировка данных, обнаружение значимых корреляций, зависимостей и тенденций в ре-	https://edu.vsu.ru

		зультате анализа имеющейся информации, выявления отношений между данными различного типа. Применение различных методов выделения, извлечения и группировки данных, которые позволяют выявить систематизированные структуры данных и вывести из них правила для принятия решений и прогнозирования их последствий	
7	Визуализация исходной информации и аналитических данных.	Возможности графического представления информации: графические функции отображения одномерных и многомерных данных, графический вывод с использованием графических параметров	https://edu.vsu.ru

13.2. Темы (разделы) дисциплины и виды занятий

№ п/п	Наименование темы (раздела) дисциплины	Виды занятий (часов)				
		Лекции	Практические	Лабораторные	Самостоятельная работа	Всего
1	Введение в анализ больших данных. Обзор источников информации.			2	10	12
2	Технологии хранения и обработки больших данных.			4	10	14
3	Статистические методы анализа данных.			4	10	14
4	Современные программные средства анализа больших данных.			4	10	14
5	Сбор и хранение больших данных.			6	12	18
6	Методы обработки и анализа данных			6	12	18
7	Визуализация исходной информации и аналитических данных.			6	12	18
	Итого:			32	76	108

14. Методические указания для обучающихся по освоению дисциплины

Для освоения дисциплины обучающимся необходимо работать с лекционными материалами (конспектами лекций) и практическими заданиями, размещенными на образовательном портале https://edu.vsu.ru/, основной и дополнительной литературой, выполнять задания на практических занятиях и в процессе самостоятельной работы, пройти текущие аттестации.

Дополнительные методические рекомендации по выполнению практических заданий, а также замечания по результатам их выполнения могут размещаться на портале https://edu.vsu.ru/ в виде индивидуальных комментариев и файлов обратной связи, сообщений форума и других элементов электронного курса.

Виды самостоятельной работы: проработка учебного материала, разобранного на лабораторном занятии с использованием учебной и научной литературы; выполнение домашних заданий (практических и теоретических); подготовка к лабораторным занятиям, контрольным работам, тестированию.

15. Перечень основной и дополнительной литературы, ресурсов интернет, необходимых для освоения дисциплины

а) основная литература:

№ п/п	Источник
1	Адлер, Ю. Практическое руководство по статистическому управлению процессами : [16+] / Ю. Адлер, В. Л. Шпер ; ред. В. Ионов. – Москва : Альпина Паблишер, 2019. – 234 с. : ил. – Режим доступа: по подписке. – URL: https://biblioclub.ru/index.php?page=book&id=570307 (дата обращения: 16.06.2024). – Библиогр. в кн. – ISBN 978-5-9614-2053-1. – Текст : электронный.
2	Информационный менеджмент : учебное пособие для бакалавров очной и заочной формы обучения : [16+] / А. С. Сенин, Е. А. Бубенок, М. Н. Дудин [и др.] ; Российская академия народного хозяйства и государственной службы при Президенте Российской Федерации. – Москва : Дело, 2018. – 297 с. : ил., табл. – Режим доступа: по подписке. – URL: https://biblioclub.ru/index.php?page=book&id=577554 (дата обращения: 16.06.2024). – Библиогр. в кн. – ISBN 978-5-7749-1402-9. – Текст : электронный.

б) дополнительная литература:

№ п/п	Источник
	Москалев, С. М. Интернет-технологии и реклама в бизнесе : учебное пособие / С. М. Москалев ; Министерство сельского хозяйства Российской Федерации, Санкт-Петербургский государственный аграрный университет (СПбГАУ). – Санкт-Петербург : Санкт-Петербургский государственный аграрный университет (СПбГАУ), 2018. – 101 с. : ил. – Режим доступа: по подписке. – URL: https://biblioclub.ru/index.php?page=book&id=491717 (дата обращения: 16.06.2024). – Библиогр. в кн. – Текст : электронный.
1.	Березовская, Е. А. Работа с сервисом бизнес-аналитики Yandex DataLens : учебное пособие : [16+] / Е. А. Березовская, С. В. Крюков ; Южный федеральный университет, Экономический факультет. – Ростов-на-Дону ; Таганрог : Южный федеральный университет, 2022. – 94 с. : ил., табл. – Режим доступа: по подписке. – URL: https://biblioclub.ru/index.php?page=book&id=698669 (дата обращения: 16.06.2024). – Библиогр. в кн. – ISBN 978-5-9275-4119-5. – Текст : электронный.
2.	Березовская, Е. А. Экономическая аналитика : учебное пособие : [16+] / Е. А. Березовская, С. В. Крюков ; Южный федеральный университет. – Ростов-на-Дону ; Таганрог : Южный федеральный университет, 2021. – 106 с. : ил. – Режим доступа: по подписке. – URL: https://biblioclub.ru/index.php?page=book&id=691189 (дата обращения: 16.06.2024). – Библиогр. в кн. – ISBN 978-5-9275-3905-5. – Текст : электронный.
3.	Гладких, Т. В. Информационные системы учета и контроля ресурсов предприятия : учебное пособие : [16+] / Т. В. Гладких, Л. А. Коробова, М. Н. Ивлиев ; науч. ред. Д. С. Сайко ; Воронежский государственный университет инженерных технологий. – Воронеж : Воронежский государственный университет инженерных технологий, 2020. – 89 с. : ил., табл., схем., граф. – Режим доступа: по подписке. – URL: https://biblioclub.ru/index.php?page=book&id=612378 (дата обращения: 04.06.2022).
4.	Реброва, Н. П. Маркетинговые исследования : теоретические и практические аспекты : учебное пособие : [16+] / Н. П. Реброва, Е. А. Лунева. – Москва : Прометей, 2020. – 159 с. : схем., ил., табл. – Режим доступа: по подписке. – URL: https://biblioclub.ru/index.php?page=book&id=612089 (дата обращения: 04.06.2024). – Библиогр. в кн. – ISBN 978-5-907244-41-2. – Текст : электронный.
5.	Внешнеторговая деятельность : инфраструктурное обеспечение цифровизации экономики : учебное пособие : [16+] / О. П. Кузнецова, С. Н. Кошкина, Е. Н. Гусарская, А. Н. Силаенков ; Омский государственный технический университет. – Омск : Омский государственный технический университет (ОмГТУ), 2020. – 128 с. : ил. – Режим доступа: по подписке. – URL: https://biblioclub.ru/index.php?page=book&id=683238 (дата обращения:

	02.04.2024). – Библиогр. в кн. – ISBN 978-5-8149-3148-1. – Текст : электронный.

в) информационные электронно-образовательные ресурсы**:**

№ п/п	Источник
1	*Каталог ЗНБ ВГУ. – URL: https://lib.vsu.ru/*
2	*ЭБС Издательства «Лань» – <URL:http://www.e.lanbook.com/*
3	*ЭБС "Университетская библиотека Online» – <URL:http://www.biblioclub.ru/*
4	*Евростат. – URL: https://ec.europa.eu/eurostat*
5	*Электронный университет https://edu.vsu.ru/*

16. Перечень учебно-методического обеспечения для самостоятельной работы

№ п/п	Источник
1	Финансы устойчивого развития : учебник : в 2 книгах / Е. В. Алтухова, А. А. Алиев, Э. А. Асяева [и др.] ; под общ. ред. К. В. Ордова ; Российский экономический университет имени Г. В. Плеханова. – Москва : Юнити-Дана, 2023. – Книга 1. – 256 с. : табл., схем. – Режим доступа: по подписке. – URL: https://biblioclub.ru/index.php?page=book&id=712624 (дата обращения: 03.06.2024). – Библиогр.: с. 219-224. – ISBN 978-5-238-03682-3 (кн. 1). – ISBN 978-5-238-03681-6. – Текст : электронный.
2	Задания для практических занятий, размещенные на https://edu.vsu.ru/

17. Информационные технологии, используемые для реализации учебной дисциплины, включая программное обеспечение и информационно-справочные системы (при необходимости)

Дисциплина реализуется с применением элементов электронного обучения и дистанционных образовательных технологий (ЭОиДОТ) («Электронный университет» https://edu.vsu.ru/).

Используются такие средства информационно-коммуникационных технологий. как текстовые редакторы, электронные таблицы, средства подготовки презентаций, облачный сервис Яндекс.

18. Материально-техническое обеспечение дисциплины:

Компьютерный класс: 25 персональных компьютеров HP ProDesk 400 G5 DM/SATA 1Tb/Монитор ЖК 21,5" BenQ BL2283, 1920*1080 LED, 16:9, 250кд, 1000:1, DC 20000000:1, 5мс, IPS, 178/178, HDMI, колонки мультимедийный проектор NEC, экран настенный 153×200.

Программное обеспечение:

Office Standard 2019 Single OLV NL Each AcademicEdition Additional Product,
Win Pro 10 32-bit/64-bit All Lng PK Lic Online DwnLd NR
"Microsoft Access 2019
(Single OLV NL Each AcademicEdition Addi-tional Product)"
Неисключительные права на ПО Dr. Web Enterprise Security Suite Комплексная защита Dr. Web Desktop Security Suite
IBM SPSS® Statistics Base

19. Оценочные средства для проведения текущей и промежуточной аттестаций

Порядок оценки освоения обучающимися учебного материала определяется содержанием следующих разделов дисциплины:

№ п/п	Наименование раздела дисциплины	Компетенция(и)	Индикатор(ы) достижения компетенции	Оценочные средства
1	Введение в анализ больших данных. Обзор источников информации.	ПК 4 Способен применять современные информационные технологии для решения профессиональных задач	ПК 4.1 Отбирает информацию для формирования данных для экономических процессов и явлений с использованием IT- технологий	Доклад Тест
2	Технологии хранения и обработки больших данных.		ПК 4.3 Применяет IT - технологии для систематизации и анализа массива данных в профессиональной деятельности	Лабораторные задания Тест
3	Статистические методы анализа данных.		ПК 4.3	Лабораторные задания Тест
4	Современные программные средства анализа больших данных.		ПК 4.1	Лабораторные задания
5	Сбор и хранение больших данных.		ПК 4.3	Доклад
6	Методы обработки и анализа данных		ПК 4.3	Контрольная работа
7	Визуализация исходной информации и аналитических данных.		ПК 4.4 Интерпретирует результаты обработки экономических данных для принятия управленческих решений	Доклад
8	Применение и кластерного анализа в специализированных программах		ПК 4.3	Лабораторные задания
9	Применение факторного анализа в специализированных программах		ПК 4.3	Лабораторные задания
Промежуточная аттестация, форма контроля – зачет				Перечень вопросов, пример КИМ приведены в п. 20.2

20. Типовые оценочные средства и методические материалы, определяющие процедуры оценивания

20.1. Текущий контроль успеваемости

Контроль успеваемости по дисциплине осуществляется с помощью следующих оценочных средств:

20.1.1 Перечень тем для докладов

1. Бизнес-проблемы и наука о данных. Формулировка бизнес-проблем. Решения, основанные на данных.
2. Проектная и процессная организация аналитики.
3. Business Intelligence. Business Analytics. Определение и соотношение понятий.
4. Enterprise Decision Management. Суть концепции. Смысл управления решениями в организации.
5. Data Science и Big Data. Определение и соотношение понятий с точки зрения бизнеса и инженерии.
6. Базы данных и хранилища данных.

7. Функциональные классы аналитических систем.
8. Системы оптимизации. Экспертные системы. Системы машинного обучения.
9. Операционная бизнес-аналитика. Аналитическая отчетность. ERP-системы.
10. Облачные решения в области бизнес-аналитики.
11. Инфраструктура анализа данных.
12. Приложения и сервисы, основанные на данных.
13. A/B тестирование и оптимизационные алгоритмы.

Критерии оценивания	Шкала оценок
Полное соответствие всем перечисленным критериям: полнота охвата выбранной темы; привлечение дополнительных источников литературы, в том числе из статей в ведущих журналах по теме, докладов ведущих конференций; структурирование презентации и управление временем доклада; уместное применение иллюстративного материала; корректность оформления ссылок; управление обсуждением доклада	*Зачтено*
Доклад не соответствует любым трем из перечисленных показателей: полнота охвата выбранной темы; структурирование презентации и управление временем доклада; уместное применение иллюстративного материала; корректность оформления ссылок; управление обсуждением доклада. Обучающийся демонстрирует отрывочные, фрагментарные знания, допускает грубые ошибки, не владеет материалом	*Не зачтено*

20.1.2 Перечень лабораторных заданий

Подобрать данные для таблицы, приведенной ниже и проанализировать их взаимное влияние, отобразить корреляцию:
1. Роста ВВП на прирост населения
2. Прироста населения на динамику безработицы
3. Прирост людей с высшим образованием на рост промышленного производства
4. Прирост людей с высшим образованием на развитие науки
5. Прирост людей с высшим образованием на динамику доходов на душу населения
6. Динамику безработицы на динамику преступности
7. С помощью регрессионного анализа найдите зависимые переменные и поясните влияние на них независимых переменных.

Период	Численность населения	Рост ВВП	Рост ВВП на душу населения	Динамика промышленного производства	Развитие науки (высокотехнологичных отраслей)	Динамика доходов на душу населения

Критерии оценивания	Шкала оценок
Полное соответствие ответа обучающегося всем перечисленным критериям. Продемонстрировано знание принципов применения методов одномерного и многомерного статистического анализа, умения выбирать процедуры обработки информации в зависимости от природы используемых данных, навыками обработки экономической информации	*Зачтено*
Ответ на контрольно-измерительный материал не соответствует любым трем из перечисленных показателей, обучающийся дает неполные ответы на дополнительные вопросы. Обучающийся демонстрирует отрывочные, фрагментарные знания, допускает грубые ошибки, не владеет основными методами статистического анализа, не демонстрирует умения обработки экономической информации	*Не зачтено*

20.1.3 Перечень заданий для контрольных работ

1. Найти потоковые данные, формирующие базу больших данных.
2. Провести процедуру структуризации и записи данных.
3. Написать функцию обращения к данным.
4. Провести визуализацию и первичный статистический анализ больших данных.

Критерии оценивания	Шкала оценок
Полное соответствие ответа обучающегося всем перечисленным критериям. Продемонстрировано знание принципов применения методов одномерного и многомерного статистического анализа, умения выбирать процедуры обработки информации в зависимости от природы используемых данных, навыками обработки экономической информации с применением специализированных программ.	*Зачтено*
Ответ на контрольно-измерительный материал не соответствует любым трем из перечисленных показателей, обучающийся дает неполные ответы на дополнительные вопросы. Обучающийся демонстрирует отрывочные, фрагментарные знания, допускает грубые ошибки, не владеет основными методами статистического анализа, не демонстрирует умения обработки экономической информации.	*Не зачтено*

20.1.4 Тест

Данные задания рекомендуются к использованию при проведении диагностических работ с целью оценки остаточных знаний по результатам освоения данной дисциплины

1) закрытые задания (тестовые, средний уровень. сложности):

1. Выберите правильный вариант ответа:

Для машинного обучения подходят данные числовые типа int;

а) предварительно подготовленные, очищенные от ошибок, пропусков и выбросов, а также нормализованные и представленные в виде числовых векторов;

б) любых форматов в цифровом виде;

в) бинарные.

2. Выберите одно неверное высказывание про MapReduce:

а) интерфейс для массово-параллельной обработки данных, где вычисления производятся на узлах, где информация изначально была сохранена

б) MapReduce – это две операции: распределения и сборки данных

в) MapReduce был придуман разработчиками Hadoop

г) MapReduce был анонсирован разработчиками Google

3. Выберите правильный вариант ответа:

В чём преимущество колоночно-ориентированных СУБД?

а) они позволяют выполнять более сложные SQL-запросы по сравнению с реляционными СУБД

б) они позволяют динамически дополнять содержание записей новыми полями

в) они имеют более гибкие возможности аналитики

г) они позволяют эффективно делать межколоночные сравнения

4. Перечислите четыре основных характеристики Big Data:

а) Virtualization, Volume, Variability, Vehicle

б) Variety, Velocity, Volume, Value

в) Verification, Volume, Velocity, Visualization

г) Video, Value, Variety, Volume

5. Выберите правильный вариант ответа:
Hadoop – это:

а) набор утилит, и программный каркас для выполнения распределённых программ, работающих на кластерах
б) распределённая СУБД, позволяющая обрабатывать большие данные
в) язык выполнения заданий в парадигме MapReduce
г) распределённая файловая система, предназначенная для хранения файлов большого объёма

6. Выберите правильный вариант ответа:
Формат Parquet считается

а) строковым
б) неструктурированным
в) полуструктурированным
г) колоночным (столбцовым)

7. Укажите 2 верных способа подключить библиотеку pandas в python:

а) import pd
б) import pandas as pd
в) import pandas
г) as pd import pandas

8. Базы данных (БД) - это…

а) объектно-реляционная система управления базами данных;
б) программа, с помощью которой осуществляется хранение, обработка и поиск информации в базе данных;
в) структурная совокупность взаимосвязанных данных определенной предметной области (реальных объектов, процессов, явлений и т.д.).

9. Выберите правильный вариант ответа:
В каком случае применение Tableau наиболее оправдано необходимо реализовать гибкое интерактивное визуальное представление данных;
а) проведено исследование, результатом которого стала таблица объект-свойства, необходимо предоставить отчетность;
б) имеются данные, необходимо более получить ясное понимание этих данных;
в) не оправдано

10. Выберите правильный вариант ответа:
Какие из следующих технологий СУБД не используют принцип MapReduce:

а) Hadoop
б) Cassandra
в) Redis
г) HDInsight

11. Выберите правильный вариант ответа:
Статистическая связь - это:

а) когда зависимость между факторным и результирующим показателями неизвестна;
б) когда каждому факторному соответствует свой результирующий показатель;
в) когда каждому факторному соответствует несколько разных значений результирующего показателя

12. Определите 3 основных свойства хорошего дашборда:
а) определена целевая аудитория
б) автономен, отсутствует необходимость поддерживать и дорабатывать
в) отвечает на задачу в целом, но не на конкретно заданные вопросы
г) решает конкретную проблему
д) определены показатели эффективности

13. Выберите правильный вариант ответа:
Основной функцией базы данных является:
а) автоматизация вычислений
б) предоставление единого хранилища для всей информации, относящейся к определенной теме
в) построение и модифицирование графических объектов

14. Выберите правильный вариант ответа:
При каком значении линейного коэффициента корреляции связь между Y и X можно признать более существенной:
а) 0,25;
б) 0,14;
в) - 0,57.

15. Выберите правильный вариант ответа:
Имеет ли Python аналог Data Frame из R:
а) да, библиотека Pandas
б) да, библиотека SciPy
в) нет
г) да, библиотека NumPy

16. Выберите правильный вариант ответа:
Какие вероятные разочарования тренда больших данных?
а) из-за угрозы безопасности личной жизни (privacy) граждан будут упрощены процедуры сбора данных, что приведёт к падению ценности больших данных
б) из-за угрозы безопасности личной жизни (privacy) граждан будут усложнены процедуры сбора данных, что приведёт к падению ценности больших данных
в) нет

17. Выберите правильный вариант ответа:
Отметьте неверное понимание Variety в контексте характеристик Big Data:
а) высокая скорость генерирования данных
б) разные типы данных в колонках таблиц реляционных СУБД
в) разнообразие отраслей, являющихся источниками данных
г) разнообразие типов данных, включающих в себя структурированные, полуструктурированные и неструктурированные

2) открытые задания (тестовые, повышенный уровень сложности):

1. Определите, какие товары наиболее часто покупаются вместе
Ответ Для достижения более точных результатов, необходимо провести анализ ассоциативных правил. Это позволит выявить связи между товарами, которые часто покупаются вместе, а также определить, какие товары являются наиболее важными для удовлетворения потребностей клиентов. Для этого можно использовать алгоритмы машинного обу-

чения, такие как Apriori или FP-Growth. После этого можно определить, какие товары стоит размещать вместе на полках магазина или какие товары можно предложить в качестве дополнительных при покупке, чтобы увеличить средний чек и удовлетворенность клиентов.

2. Необходимо определить, какие факторы влияют на уровень лояльности клиентов в компании

Ответ: Для этого можно провести анализ данных, используя методы статистического анализа, машинного обучения и искусственного интеллекта. С помощью этих методов можно определить, какие факторы наиболее важны для клиентов и как они влияют на уровень лояльности. Например, можно использовать методы классификации и регрессии, чтобы выявить зависимость между уровнем лояльности и различными факторами, такими как возраст, пол, доход, образование и т.д. Также можно использовать методы кластеризации, чтобы выделить группы клиентов с разным уровнем лояльности и определить, какие факторы наиболее важны для каждой группы.

3. Необходимо предсказать вероятность оттока клиентов из компании (ПК 4.1)

Ответ: Сначала необходимо собрать данные о клиентах, такие как история покупок, демографические данные, информация об использовании продуктов или услуг компании и т.д. Затем можно провести анализ данных, используя методы статистического анализа и машинного обучения, чтобы выявить факторы, которые наиболее сильно влияют на вероятность оттока клиентов.

Например, можно использовать методы классификации и регрессии для выявления зависимости между вероятностью оттока и различными факторами, такими как длительность пользования продуктами или услугами компании, частота покупок, уровень удовлетворенности клиента и т.д. Также можно использовать методы кластеризации, чтобы выделить группы клиентов с разным уровнем вероятности оттока и определить, какие факторы наиболее важны для каждой группы.

4. Определите, какие пользователи наиболее активны в социальной сети

Ответ: Для оценки количества публикаций, лайков, комментариев и репостов, которые сделал каждый пользователь, можно использовать следующие методы: API социальной сети. С помощью API можно получить доступ к данным о пользовательской активности, таким как количество публикаций, лайков, комментариев и репостов. Специальные инструменты для сбора данных. Некоторые инструменты, такие как Socialbakers, Hootsuite и Sprout Social, позволяют собирать данные о пользовательской активности в социальных сетях.

5. Как можно автоматизировать запуск пакетных задач в рамках конвейера обработки больших данных по расписанию?

Ответ: Для автоматизации запуска пакетных задач в рамках конвейера обработки больших данных по расписанию можно использовать специальные инструменты для планирования задач, такие как cron или Airflow.

Cron - это стандартный инструмент в Unix-подобных операционных системах, который позволяет запускать задачи по расписанию. Для использования cron необходимо создать файл crontab, в котором указать расписание запуска задач и команды, которые нужно выполнить.

Airflow - это более сложный инструмент, который позволяет создавать и управлять конвейерами обработки данных. Airflow позволяет создавать DAG (Directed Acyclic Graph), который определяет порядок выполнения задач и их зависимости. Каждая задача в DAG представляется в виде оператора, который выполняет определенную функцию.

В обоих случаях необходимо настроить параметры запуска задач, указать пути к файлам и скриптам, а также определить зависимости между задачами, если это необходимо. При правильной настройке эти инструменты могут значительно упростить и автоматизировать процесс обработки больших данных.

6. Какие задачи решают графовые БД?
Ответ: хранение информации о графах, распределенное хранение с учетом минимизации передачи информации.

7. В каких случаях требуется СУБД со свойством расширяемости записей?
Ответ: требуется добавлять оценки пользователей музыкальным композициям для целей дальнейшей выдачи рекомендаций; в проекте требуется индексировать веб-страницы интернета. Каждый месяц аналитики анализируют и добавляют новые признаки, которые вычисляются по проиндексированной веб-странице.

8. Укажите фактор, способствовавший появлению тренда больших данных.
Ответ: маркетинговые кампании крупных корпораций, снижение издержек на хранение данных.

9. Базовые принципы обработки больших данных:
Ответ: горизонтальная адаптивность, стабильность в работе при отказах, концентрация данных.

10. Какие существуют технологии потоковой обработки событий в режиме реального времени.
Ответ: Существует множество технологий потоковой обработки событий в режиме реального времени. Некоторые из наиболее популярных включают в себя Apache Kafka, Apache Flink, Apache Storm, Amazon Kinesis, Google Cloud Pub/Sub, Microsoft Azure Event Hubs и многие другие. Эти технологии позволяют обрабатывать большие объемы данных в режиме реального времени и использовать их для различных целей, включая мониторинг и анализ производительности систем, обнаружение аномалий, анализ поведения пользователей и многое другое.

11. Какое СУБД подходит для полнотекстового интеллектуального поиска и аналитики по полуструктурированным данным в формате JSON
Ответ: Для полнотекстового интеллектуального поиска и аналитики по полуструктурированным данным в формате JSON подходят различные СУБД, такие как Elasticsearch, MongoDB, Couchbase, Apache Cassandra и др. Однако, Elasticsearch является одним из наиболее популярных и мощных инструментов для полнотекстового поиска и аналитики данных в формате JSON. Он предоставляет широкий набор функций для поиска, фильтрации, агрегации и визуализации данных, а также может интегрироваться с другими инструментами и системами.

12. Какой фреймворк больше подходит для распределенного глубокого машинного обучения (Deep Learning)
Ответ: TensorFlow и PyTorch являются наиболее популярными фреймворками для распределенного глубокого машинного обучения. Они оба поддерживают распределенное обучение на нескольких устройствах и могут работать с кластерами серверов. Однако, TensorFlow имеет более широкую экосистему инструментов и библиотек, в то время как PyTorch предоставляет более простой и интуитивный интерфейс для создания моделей. Выбор фреймворка зависит от конкретных потребностей и задач.

13. Назовите несколько реальных приложений алгоритмов машинного обучения:

Ответ: Биоинформатика
Робототехника, автоматизация процессов
Обработка естественного языка
Анализ настроений
Обнаружение мошенничества
Системы распознавания лица и голоса
Борьба с обмыванием денег

14. Какими способами можно уменьшить размерность набора данных?

Ответ:Факторный анализ
Анализ главных компонентов
Isomap
Автокодирование
Полуопределенное вложение

15. Чем интеллектуальный анализ данных отличается от машинного обучения?

Ответ: Интеллектуальный анализ данных – это дисциплина, которая занимается извлечением данных из не уточненных источников, чтобы их можно было проанализировать и изучить для получения значимых закономерностей. Машинное обучение фокусируется на разработке алгоритмов и методологий, которые могут помочь машинам учиться и развиваться самостоятельно.

16. В какой ситуации наиболее эффективны NoSQL решения типа ключ-значение?

Ответ: Потоковая обработка логов кластера серверов и быстрого сохранения без требования оперативной аналитики

17. Объясните разницу между KNN и кластеризацией k-средних.

Ответ: KNN расшифровывается как K-Nearest Neighbours, который представляет собой контролируемый метод обучения, требующий помеченных данных, которые затем используются для классификации точек на основе их расстояния от ближайшей точки.
Кластеризация K-средних – это алгоритм машинного обучения без учителя, в котором предоставляется модель с немаркированными данными, а затем алгоритм группирует точки наблюдения / данных на основе сходства, измеренного с использованием среднего значения расстояний между разными точками.

18. Объясните недостатки линейной модели?

Ответ: Линейная модель основана на слишком большом количестве теоретических предположений, которые в большинстве случаев не соответствуют действительности.
Дискретные или бинарные результаты нельзя получить с помощью линейной модели.
Высокая негибкость.

19. Отметьте причины создания NoSQL баз данных

Ответ: высокая стоимость горизонтальной масштабируемости RDBMS при сохранении требования высокой доступности.

20. Определите понятие Data Mining:

Ответ: анализ данных с помощью статистических и математических методов предназначенный для поиска ранее неизвестных закономерностей в больших массивах информации.

21. Этапы процесса Data Mining:

Ответ: очистка данных, интеграция данных, выборка данных, преобразование данных, интеллектуальный анализ данных, оценка модели, представление знаний/визуализация.

22. С некоторой периодичностью персонал предприятия списывает группы расходных материалов на различных участках учета. Для выявления ошибок, акты списания выборочно проверяются аудитором. Как бы в данном случае формулировалась задача классификации?
Ответ: научиться автоматически выявлять ошибочные списания с ожидаемой ошибкой не ниже 97%; определить три категории: «ошибочные», «под сомнением», «безошибочные» и найти правило отнесения к этим категориям.

23. Технологии обработки больших данных:
Ответ: NoSQL, MapReduce, Hadoop, R.

24. Как бы вы поступили с отсутствующими данными в наборе данных?
Ответ: Можно заменить отсутствующее значение другим значением, используя меру центральной тенденции, такую как среднее значение, медиана или мода, используя следующий подход:
Непрерывные переменные: заменить отсутствующие на среднее значение
Порядковые переменные: замените отсутствующие на медиану
Категориальные переменные: заменить отсутствующие на режим
В случае, если у нас очень небольшая доля отсутствующих значений в большом наборе данных, мы также можем удалить их. dropna () из библиотеки Pandas.

25. Определите разницу между примесью Джини и энтропией в дереве решений.
Ответ: Примесь Джини и Энтропия – это метрики, которые могут помочь разделить дерево решений. Первый измеряет вероятность правильной классификации случайной выборки, если вы случайным образом выбираете метку в ветке.
Энтропия – это мера неопределенности вашей модели. Энтропия самая низкая по направлению к листовому узлу. Прирост информации – это разница энтропий, наблюдаемая между набором данных до и после разделения атрибута. Он имеет максимальное значение около листового узла. Разница между энтропиями может помочь понять уровень неопределенности в дереве решений.

26. Что такое выбросы и как их обнаружить? (ПК 4.4)
Ответ: Выбросы – это те точки данных, значение которых значительно отличается от среднего значения набора данных. Коробчатая диаграмма, линейные модели и модели на основе близости часто используются для отбора выбросов в наборе данных. Для большинства моделей настоятельно рекомендуется обрабатывать выбросы путем их ограничения или исключения из набора данных.

27. Что такое A / B-тестирование? (ПК 4.4)
Ответ: A / B-тестирование – это тестирование с двумя переменными, выполняемое в рандомизированных экспериментах для определения того, какая из двух выбранных моделей лучше подходит для данного набора данных.

28. Объясните кластерную выборку:
Ответ: Кластерная выборка – это метод группировки, используемый для совокупности, в которой есть отдельные подмножества однородных элементов. Кластерная выборка, обычно используемая для маркетинговых исследований, делит данный набор данных на более мелкие группы и случайным образом выбирает выборку из групп.

29. Несколько маленьких деревьев решений лучше, чем одно большое? Обоснуйте.
Ответ: Наличие нескольких небольших деревьев решений – это то же самое, что использование модели случайного леса, которая, как известно, является более точной (низкий

уровень смещения) и менее подвержена проблеме переобучения (высокая дисперсия). Итак, да, иметь несколько маленьких деревьев решений было бы предпочтительнее, чем иметь одно большое.

30. Что делает среднеквадратическую ошибку плохим показателем производительности модели?

Ответ: MSE или среднеквадратическая ошибка основана на связывании значительно более высокого веса с большими ошибками, что делает больший акцент на более широких отклонениях. Однако это хорошо работает в большинстве алгоритмов, чтобы минимизировать ошибку модели и стоимость.

Иногда лучшим вариантом для MSE является MAE (средняя абсолютная ошибка) или MAPE (средняя абсолютная ошибка в процентах), что устраняет вышеуказанный недостаток и легко интерпретируется.

Критерии и шкалы оценивания:

Для оценивания выполнения заданий используется балльная шкала:

1) закрытые задания (тестовые, средний уровень сложности):

- 1 балл – указан верный ответ;
- 0 баллов – указан неверный ответ, в том числе частично.

2) открытые задания (тестовые, повышенный уровень сложности):

- 2 балла – указан верный ответ;
- 0 баллов – указан неверный ответ, в том числе частично.

3) открытые задания (мини-кейсы, средний уровень сложности):

- 5 баллов – задание выполнено верно (получен правильный ответ, обоснован (аргументирован) ход выполнения (при необходимости));
- 2 балла – выполнение задания содержит незначительные ошибки, но приведен правильный ход рассуждений, или получен верный ответ, но отсутствует обоснование хода его выполнения (если оно было необходимым), или задание выполнено не полностью, но получены промежуточные (частичные) результаты, отражающие правильность хода выполнения задания, или, в случае если задание состоит из выполнения нескольких подзаданий, 50% которых выполнено верно;
- 0 баллов – задание не выполнено или выполнено неверно (ход выполнения ошибочен или содержит грубые ошибки, значительно влияющие на дальнейшее его изучение).

20.2. Промежуточная аттестация

Промежуточная аттестация по дисциплине осуществляется с помощью следующих оценочных средств:

Перечень вопросов к зачету:

1. Охарактеризуйте возможности применения программы для экономического анализа.
2. Интерпретация результатов факторного анализа
3. Какие существуют основные характеристики распределения, алгоритм вычисления
4. Модель факторного анализа, принципы выбора числа факторов
5. Алгоритм вычислений факторного анализа.
6. Определите сущность понятия «большие данные».
7. Основные вызовы больших данных.
8. Процесс аналитики анализа больших данных.
9. Дайте характеристику Big Data на мировом рынке.
10. Охарактеризуйте Big Data в России.
11. Определите понятие Data Mining.
12. В чем состоит когнитивный анализ данных.
13. Какие модели данных вы знаете?
14. Основные описательные статистики.

15. Определите различия между параметрическими, непараметрическими и номинальными методами.
16. Опишите основную идею корреляционного анализа.
17. Регрессионный анализ.
18. Основная идея дисперсионного анализа.
19. Сущность кластерного анализа.
20. Дискриминантный анализ: модель и общая процедура выполнения.
21. Цели факторного анализа.
22. Программные средства анализа данных: Statistica, SPSS, Excel и другие; их преимущества и недостатки.
23. Преимущества работа с данными в программе R.
24. Представление исходных данных в программе R.
25. Выполнение анализа данных в R.
26. Цели применения кластерного анализа
27. Условия применения параметрических критериев при проверке гипотез
28. Методы объединения кластеров (иерархический кластерный анализ)
29. Условия применения непараметрических критериев при проверке гипотез
30. Методы объединения кластеров (метод К-средних)
31. Основные типы шкал и соответствующие им меры средней тенденции и меры разброса
32. Алгоритм вычислений кластерного анализа
33. Способы проверки гипотез о соответствии эмпирического распределения одному из теоретических законов
34. Какие показатели характеризуют форму и тесноту корреляционной связи
35. Сущность дисперсионного анализа
36. Общее описание регрессионной модели. Особенности использования регрессионных моделей при анализе данных выборочных исследований
37. Задачи, решаемые с помощью дисперсионного анализа
38. Множественный регрессионный анализ. Проверка качества полученной модели, требования к исходным данным.
39. Алгоритм вычислений одномерного дисперсионного анализа
40. Цели и задачи корреляционного анализа при анализе экономических данных

Пример контрольно-измерительного материала

УТВЕРЖДАЮ
заведующая кафедрой международной экономики и внешнеэкономической деятельности
_____ Е.В. Ендовицкая
подпись __ ___ 202_ г.

Направление подготовки 38.03.01 «Экономика»
Дисциплина Аналитические системы Big Data Курс 3
Форма обучения _очная_ Вид аттестации _промежуточная__ Вид контроля ___зачет___

Контрольно-измерительный материал №2

1. Алгоритм вычислений факторного анализа.
2. Общее описание регрессионной модели. Особенности использования регрессионных моделей при анализе данных выборочных исследований

Описание технологии проведения
Контрольно-измерительные материалы промежуточной аттестации включают теоретические вопросы, позволяющие оценить уровень полученных знаний и практические задания, которые позволяют оценить степень сформированности умений и навыков. При оценивании используются количественные шкалы оценок.

Промежуточная аттестация по дисциплинам с применением электронного обучения, дистанционных образовательных технологий (далее – ЭО, ДОТ) проводится в рамках электронного курса, размещенного в ЭИОС (образовательный портал «Электронный университет ВГУ» (LMS Moodle, https://edu.vsu.ru/)).

Обучающиеся, проходящие промежуточную аттестацию с применением ДОТ, должны располагать техническими средствами и программным обеспечением, позволяющим обеспечить процедуры аттестации. Обучающийся самостоятельно обеспечивает выполнение необходимых технических требований для проведения промежуточной аттестации с применением дистанционных образовательных технологий.

Идентификация личности обучающегося при прохождении промежуточной аттестации обеспечивается посредством использования каждым обучающимся индивидуального логина и пароля при входе в личный кабинет, размещенный в ЭИОС образовательной организации. Требования к выполнению заданий, шкалы и критерии оценивания

Критерии оценивания	Шкала оценок
Полное соответствие ответа обучающегося всем перечисленным критериям. Продемонстрировано знание принципов применения методов одномерного и многомерного статистического анализа, умения выбирать процедуры обработки информации в зависимости от природы используемых данных, навыками обработки экономической информации с применением специализированных программ	*Зачтено*
Ответ на контрольно-измерительный материал не соответствует любым трем из перечисленных показателей, обучающийся дает неполные ответы на дополнительные вопросы. Обучающийся демонстрирует отрывочные, фрагментарные знания, допускает грубые ошибки, не владеет основными методами статистического анализа, не демонстрирует умения обработки экономической информации	*Не зачтено*