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

ФОНД ОЦЕНОЧНЫХ МАТЕРИАЛОВ (СРЕДСТВ) ПО ДИСЦИПЛИНЕ

ВВЕДЕНИЕ В BIG DATA

(наименование дисциплины)

направление подготовки / специальность

01.03.05. СТАТИСТИКА

(код и наименование направления подготовки (специальности)

направленность (профиль) подготовки

«Бизнес-аналитика»

(наименование направленности (профиля) подготовки))

Владимир, 2023

1. ПЕРЕЧЕНЬ КОМПЕТЕНЦИЙ И ПЛАНИРУЕМЫЕ РЕЗУЛЬТАТЫ ОБУЧЕНИЯ ПО ДИСЦИПЛИНЕ

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-2. Способен формировать упорядоченные сводные массивы статистической информации и осуществлять расчет сводных и производных показателей в соответствии с утвержденными методиками, в том числе с применением необходимой вычислительной техники и стандартных компьютерных программ	ОПК-2.1. Знает методики формирования упорядоченных массивов статистической информации для решения профессиональных задач ОПК-2.2. Умеет применять современные информационные технологии и программные средства, для формирования массивов статистической информации ОПК-2.3. Владеет навыками расчета сводных и производных показателей для решения практических задач профессиональной деятельности	Знает методики работы с большими массивами информации для решения профессиональных задач Умеет применять современные информационные технологии и программные средства, для работы с большими массивами информации для решения профессиональных задач Владеет навыками расчета сводных и производных показателей при работе с большими массивами информации для решения профессиональных задач	Рейтинг-контроль, тесты, семинары, эссе.
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 ОПК-4.2. Умеет выбирать информационные технологии ОПК-4.3. Владеет навыками использования современных информационных технологий при решении задач профессиональной деятельности	Знает принципы работы с большими данными Умеет выбирать информационные технологии для работы с большими данными Владеет навыками использования технологий работы с большими даннвми	Рейтинг-контроль, тесты, семинары, эссе.

2. ОЦЕНОЧНЫЕ МАТЕРИАЛЫ ДЛЯ ПРОВЕДЕНИЯ ТЕКУЩЕГО КОНТРОЛЯ УСПЕВАЕМОСТИ ПО ДИСЦИПЛИНЕ

Рейтинг-контроль №1

Тестовые задания:

1. Принятый способ представления данных: показатели должны быть:

1) ПО СТРОКАМ;

2) ПО ЯЧЕЙКАМ;

3) ПО СТОЛБЦАМ;

4) ПО ДИАГОНАЛИ.

2. Интервальные данные – это (подчеркните правильные ответы):

1) данные с интервалом;
2) данные об интервалах;
3) количество измерений в каждом интервале;
4) количество интервалов в каждом измерении.

3. Среди ниже приведённых нечисловые данные следующие:

1) баллы;
2) дихотомические;
3) ранги;
4) рейтинги.

4. Простейшие статистические характеристики – это:

1) математическое ожидание;
2) среднее;
3) с.к.о.;
4) дисперсия.

5. Приведение к нормальной форме - это:

1) деление на с.к.о.;
2) округление;
3) деление на среднее;
4) деление на константу интегрирования.

6. Какие функции Excel имеют отношение к оцифровке:

1) РАНГ;
2) СЧЁТЕСЛИ;
3) КОРРЕЛ;
4) СУММЕСЛИ.

7. Многомерность в статистике - это:

1) ПЕРЕМЕННЫХ БОЛЬШЕ ОДНОЙ;
2) ИЗМЕРЕНИЙ БОЛЬШЕ 10;
3) ПЕРЕМЕННЫХ БОЛЬШЕ ДВУХ;

4) ИЗМЕРЕНИЙ БОЛЬШЕ 5

8. Следующие программы являются специализированными статистическими паке- тами:

1) EXCEL;

2) GRAPHER;

3) SPSS;

4) STATISTICA.

9. Проверка статистической гипотезы включает в себя:

1) ранжирование;

2) принятие уровня значимости;

3) вычисление эмпирического значения;

4) вычисление критического значения.

Рейтинг-контроль №2

Тестовые задания:

1. Кластерный анализ предназначен для:

1) ГРУППИРОВКИ ОБЪЕКТОВ;

2) РАНЖИРОВАНИЯ ОБЪЕКТОВ;

3) ГРУППИРОВКИ ПОКАЗАТЕЛЕЙ;

4) РАНЖИРОВАНИЯ ПОКАЗАТЕЛЕЙ.

2. Опции кластерного анализа:

1) расстояние межу группами;

2) РАССТОЯНИЕ МЕЖУ ОБЪЕКТАМИ;

3) РАССТОЯНИЕ МЕЖУ ПОКАЗАТЕЛЯМИ;

4) РАССТОЯНИЕ МЕЖДУ ТЕЛАМИ.

3. Кластерный анализ реализован в программах:

1) EXCEL;

2) SPSS;

3) AGRAPHER;

4) STATISTICA.

4. Снижение размерности это:

1) уменьшение числа измерений;

2) УМЕНЬШЕНИЕ ЧИСЛА ПОКАЗАТЕЛЕЙ;

3) УМЕНЬШЕНИЕ ЧИСЛА ОБЪЕКТОВ;

4) УМЕНЬШЕНИЕ ЧИСЛА ЗНАКОВ.

5. Компонентный анализ реализован в программах:

1) EXCEL;
2) SPSS;
3) AGRAPHER;
4) STATISTICA.

6. Методы, относящиеся к снижению размерности:

1) факторный анализ;
2) регрессия;
3) компонентный анализ;
4) корреляция.

7. Компонентный анализ позволяет:

1) СОРТИРОВАТЬ;

2) РАНЖИРОВАТЬ;

3) ГРУППИРОВАТЬ;

4) УПОРЯДОЧИВАТЬ.

8. Дихотомическая шкала это:

1) СОСТОЯЩАЯ ИЗ “ДА” И “НЕТ”;

2) СОСТОЯЩАЯ ИЗ ДВУХ ЧИСЕЛ;

3) СОСТОЯЩАЯ ИЗ “ИСТИНА” И “ЛОЖЬ”;

4) СОСТОЯЩАЯ ИЗ ДВУХ РАНГОВ.

9. К нечисловым шкалам относятся:

1) НОМИНАЛЬНАЯ;

2) АБСОЛЮТНАЯ;

3) ИНТЕРВАЛОВ;

4) РАНГОВАЯ.

10. Существует шкал для описания данных:

1) 4;

2) 6;

3) 5;

4) 7.

Рейтинг-контроль №3

Тестовые задания

1. Количество наблюдений - это:

1) РАЗМЕРНОСТЬ;

2) ШИРИНА;

3) ОБЪЁМ ВЫБОРКИ;

4) ПОВЕРХНОСТЬ ВЫБОРКИ.

2.Элементы таблицы сопряжённости называются:

1) КООРДИНАТЫ;

2) СКОРОСТИ;

3) ДЛИНЫ;

4) ЧАСТОТЫ.

3.Методы анализа таблиц сопряжённости:

1) КРИТЕРИЙ РОЗЕНБАУМА;

2) ХИ-КВАДРАТ;

3) КРИТЕРИЙ КОЛМОГОРОВА-СМИРНОВА;

4) КРИТЕРИЙ ФИШЕРА.

4. В ходе анализа таблицы сопряжённости выполняется:

1) проверка на соответствие;
2) проверка на непротиворечивость;
3) проверка на монотонность;
4) проверка на значимость.

5. Максимальная размерность таблицы сопряжённости может быть:

1) 3;
2) 5;
3) 10;
4) КАКАЯ УГОДНО.

6. Вычисляемое значение критерия хи-квадрат называется:

1) Численное значение;
2) реальное значение;
3) экспериментальное значение;
4) эмпирическое значение.

7. Вычисляемое значение хи-квадрат сравнивается с:

1) КРИТИЧЕСКИМ ЗНАЧЕНИЕМ;
2) ПРЕДЕЛЬНЫМ ЗНАЧЕНИЕМ;
3) ЭТАЛОННЫМ ЗНАЧЕНИЕМ;
4) ГРАНИЧНЫМ ЗНАЧЕНИЕМ.

8. То, с чем сравнивается вычисляемое значение хи-квадрат, вычисляется в EXCEL функцией:

1) ХИ2РАСП;
2) ХИ2ТЕСТ;
3) ХИ2ОБР;
4) ХИ2.

9. К коэффициентам связи относятся:

1) КОЭФФИЦИЕНТ КОНТИНГЕНЦИИ;

2) КОЭФФИЦИЕНТ АССОЦИАЦИИ;

3) КОЭФФИЦИЕНТ ЧУПРОВА-КРАМЕРА;

4) КОЭФФИЦИЕНТ КОЛЛИГАЦИИ.

10. К разновидности критерия хи-квадрат относятся:

1) критерий Вилкоксона;
2) информационный критерий;
3) критерий Джонкира;
4) критерий максимального правдоподобия.

11. Выявление вкладов, вносимых каждой клеткой таблицы, называется:

1) разбиение хи-квадрат;
2) локализация хи-квадрат;
3) анализ хи-квадрат;
4) сортировка хи-квадрат.

12 Лог-линейный анализ - это:

1) анализ синтеза таблиц;
2) анализ достоверности таблиц;
3) статистический анализ связи таблиц;
4) анализ разброса таблиц.

Иные оценочные материалы для проведения текущего контроля успеваемости
Практические задачи

1. Дан набор данных заданной структуры и программа SAS Data step, производящая определенную обработку и вычисления с использованием данного набора. Перепишитеэту программу на SAS DS2 с использованием параллельных нитей и созданием пользовательского пакета, чтобы результат обработки сохранился тем же, но код мог выполняться в параллельной среде.

2. Дан набор заданной структуры, постройте модель прогнозирования отклика с использованием процедуры impstat с алгоритмом random forest с заданным числом деревьев. Примените полученную модель к тестовому набору данных той же структуры, визуализируйте полученный график Lift. Постройте на том же наборе модель с использование высокопроизводительной версии метода GLM. Примените к тестовому набору. Сравните

результаты GLM и Random Forest по AUC.

3. Дан текстовый корпус документов, лежащих в указанной директории. Создайте в SAS Tex Miner проект, который: выберет файлы с расширением pdf; осуществит парсинг набора с определением частей речи и сохранением в признаковом пространстве только существительных и глаголов; осуществит фильтрацию документов и признаков с использованием заданной схемой определения весов лексем (например, на основе tf-idf); выделитзаданное количество ключевых тематик по методу SVD. В ответе укажите топ 5 ключевыхслов во второй выявленной тематике. Какой документ имеет наибольший вес в это тема- тике?

Тематика докладов

1. Факторный анализ.
2. Дискриминантный анализ.
3. Кластерный анализ.
4. Основные направления развития методов обработки и хранения данных.
5. Проблема хранения неструктурированных данных.
6. Проблема преобразования данных.
7. Семантические анализаторы.
8. Самообучающиеся автоматы.
9. Языки для Big Data.
10. Фреймворки.
11. Базы данных.
12. Аналитические платформы.
13. Аналитика Big Data — реалии и перспективы в России и мире.
14. Big data: применение и возможности.
15. Рынок Big data в России.

Тематика эссе

1. Аналитическая деятельность агентства Moody’s.
2. Аналитическая продукция Lexis-Nexis.
3. Принцип работы системы Watson.
4. Прогнозы развития систем аналитики по данным ведущих мировых агентств.
5. Направления исследований в области аналитики компании IBM.
6. Технология анализа данных Text Mining.
7. Когнитивная аналитика в продуктах компании IBM

8. Аналитические решения от компании SAS.

9. Современные технологии обработки больших объемов данных.

10. Основные принципы работы OLAP-систем.

11. Перспективы развития машинного перевода

12. Применение систем Business Intelligence.

13. Сравнительный анализ возможностей систем Business Intelligence ведущих производителей (SAP, Oracle, SAS, IBM).

14. Технологии визуализации данных.

15. Геоинформационные системы.

Тематика презентаций

1. Анализ данных социальных сетей.

2. Возможности использования Big Data в различных сферах деятельности.

3. Направления исследований в области аналитики компании IBM.

4. Технология анализа данных Text Mining.

5. Когнитивная аналитика в продуктах компании IBM.

6. Современные технологии обработки больших объемов данных.

7. Основные принципы работы OLAP-систем.

8. Перспективы развития машинного перевода.

9. Краткий обзор российского рынка информационно-аналитических систем.

10. Деятельность ведущих российских информационно-аналитических агентств.

11. Хэш-функции и показатели

12. Методы табличной группировки и консолидации данных

13. Разработка пользовательского приложения на основе базы данных.

14.Применение кластерных алгоритмов в линейных и параллельных вычислениях. 15. Средства реализации методов интеллектуального анализа.

3. ОЦЕНОЧНЫЕ МАТЕРИАЛЫ ДЛЯ ПРОВЕДЕНИЯ ПРОМЕЖУТОЧНОЙ АТТЕСТАЦИИ ПО ДИСЦИПЛИНЕ

Вопросы к экзамену

1. Факторный анализ.
2. Дискриминантный анализ.
3. Кластерный анализ.
4. Многомерное шкалирование.

5. Методы контроля качества.

1. Основные направления развития методов обработки и хранения данных.
2. Volume.
3. Закон Мура.
4. Velocity. Variety.
5. Фреймворк Hadoop.
6. Проблема хранения неструктурированных данных.
7. Проблема преобразования данных.
8. Семантические анализаторы.
9. Самообучающиеся автоматы.
10. Языки для Big Data: R.
11. Языки для Big Data: Python.
12. Языки для Big Data: Julia.
13. Языки для Big Data: Java.
14. Языки для Big Data: Scala.
15. Языки для Big Data: MATLAB.
16. Языки для Big Data: Kafka,.
17. Языки для Big Data: Hadoop.
18. Языки для Big Data: Go.
19. Фреймворки для Big Data: Hadoop.
20. Фреймворки для Big Data: Spark.
21. Фреймворки для Big Data: Storm.
22. Базы данных для Big Data: Hive.
23. Базы данных для Big Data: Impala.
24. Базы данных для Big Data: Presto.
25. Базы данных для Big Data: Drill.
26. Аналитические платформы для Big Data: Rapid Miner.
27. Аналитические платформы для Big Data: IBM SPSS Modeler.
28. Аналитические платформы для Big Data: KNIME.
29. Аналитические платформы для Big Data: Qlik Analytics Platform.
30. Аналитические платформы для Big Data: STATISTICA Data Miner.
31. Аналитические платформы для Big Data: Informatica Intelligent Data Platform.
32. Аналитические платформы для Big Data: World Programming System.
33. Аналитические платформы для Big Data: Deductor.
34. Аналитические платформы для Big Data: SAS Enterprise Miner.

35. Zookeeper.
36. Flume.
37. IBM Watson Analytics.
38. Dell EMC Analytic Insights Module.
39. Windows Azure HDInsight.
40. Microsoft Azure Machine Learning.
41. Pentaho Data Integration.
42. Teradata Aster Analytics.
43. SAP BusinessObjects Predictive Analytics.
44. Oracle Big Data Preparation.
45. Аналитика Big Data — реалии и перспективы в России и мире.
46. Data Mining.
47. Краудсорсинг.
48. Смешение и интеграция данных.
49. Машинное обучение.
50. Искусственные нейронные сети.
51. Распознавание образов.
52. Прогнозная аналитика.
53. Имитационное моделирование.
54. Пространственный анализ.
55. Статистический анализ.
56. Визуализация аналитических данных.
57. Big data: применение и возможности.
58. Решения на основе Big data.
59. Рынок Big data в России.
60. Big data в банках.
61. Big data в бизнесе.
62. Big data в маркетинге.

Критерии оценки тестирования студентов

Оценка выполнения тестов	Критерий оценки
0,5 балла за правильный ответ на 1 вопрос	Правильно выбранный вариант ответа (в случае закрытого теста), правильно вписанный ответ (в случае открытого теста)

Регламент проведения тестирования и оценивания

№	Вид работы	Продолжительность
1.	Предел длительности тестирования (20 вопросов)	35-40 мин.
2.	Внесение исправлений	до 5 мин.
	Итого (в расчете на тест)	до 45 мин.

Критерии оценки выполнения заданий студентами

Регламент выполнения заданий

№	Вид работы	Продолжительность
1.	Предел длительности защиты задания	до 5-7 мин.
2.	Внесение исправлений в представленное решение	до 2 мин.
3.	Комментарии преподавателя	до 1 мин.
	Итого (в расчете на одно задание)	до 10 мин.

Оценка в баллах	Критерии оценивания задания
15 баллов	Задание выполнены полностью, все элементы и взаимосвязи модели (проекта) обоснованы.
10 баллов	Задание выполнены полностью, но нет достаточного обоснования взаимосвязей, элементов модели (проекта)
5 баллов	Модели (проекты) имеют незаконченную структуру. Обоснование модели (проекта) дано частично.
0 баллов	Задание не выполнено.

Критерии оценки устных ответов студентов

Регламент проведения устного опроса

№	Вид работы	Продолжительность
1.	Предел длительности ответа на каждый вопрос	до 3 мин.
2.	Внесение студентами уточнений и дополнений	до 1 мин.
3.	Дискуссия с участием учебной группы по ответу на вопрос	до 2 мин.
4.	Комментарии преподавателя	до 1 мин.
	Итого продолжительность устного ответа (на один) вопрос)	до 7 мин.

Оценка в баллах	Критерии оценивания ответа
5	Ответ отличается последовательностью, полнотой, логикой изложения. Легко воспринимается аудиторией. При ответе на вопросы выступающий демонстрирует глубину владения материалом. Ответы формулируются аргументировано, обосновывается собственная позиция в проблемных ситуациях.
4	Ответ отличается последовательностью, логикой изложения. Но обоснование сделанных выводов не достаточно аргументировано. Неполно раскрыто содержание проблемы.

3	Ответ направлен на пересказ содержания проблемы, но не демонстрирует умение выделять главное, существенное. Выступающий не владеет пониманием сути излагаемой проблемы

Критерии оценки участия в дискуссии

В целях закрепления практического материала и углубления теоретических знаний по разделам дисциплины предполагается проведение обсуждений в форме дискуссий по актуальным темам, вопросам, что позволяет углубить процесс познания, раскрыть понимание прикладной значимости осваиваемой дисциплины.

Критерии	**Оценка в баллах**
Демонстрирует полное понимание обсуждаемой проблемы, высказывает собственное суждение по вопросу, аргументировано отвечает на вопросы участников дискуссии, соблюдает регламент выступления.	1
Понимает суть рассматриваемой проблемы, может высказать типовое суждение по вопросу, отвечает на вопросы участников семинара, однако выступление носит затянутый или не аргументированный характер.	0,5
Принимает участие в обсуждении, однако собственного мнения по вопросу не высказывает, либо высказывает мнение, не отличающееся от мнения других докладчиков.	0,2
Не принимает участия в обсуждении	0

Показатели, критерии и шкала оценивания компетенций промежуточной аттестации знаний по учебной дисциплине «Введение в Big Data» на экзамене.

Промежуточная аттестация по итогам освоения дисциплины – экзамен.

Оценка в баллах	**Оценка за ответ на зачете**	**Критерии оценивания компетенций**	**Уровень освоения компетенций**
91 -100 Баллов	Отлично	Выставляется обучающемуся, если он глубоко и прочно усвоил теоретический и практический материал, может продемонстрировать это на занятиях и в ходе промежуточной аттестации. Обучающийся исчерпывающе и логически стройно излагает учебный материал, умеет увязывать теорию с практикой, справляется с решением задач профессиональной направленности высокого уровня сложности, правильно обосновывает принятые решения. Свободно ориентируется в учебной и профессиональной литературе.	Высокий
76 – 90 баллов	Хорошо	Выставляется обучающемуся, если он знает теоретический и практический материал, грамотно и по существу излагает его на занятиях и в ходе промежуточной аттестации,	Хороший

		не допуская существенных неточностей. Обучающийся правильно применяет теоретические положения при решении практических задач профессиональной направленности разного уровня сложности, владеет необходимыми для этого навыками и приёмами. Достаточно хорошо ориентируется в учебной и профессиональной литературе.	
61 – 75 баллов	Удовлетворительно	Выставляется обучающемуся, если он знает на базовом уровне теоретический и практический материал, допускает отдельные ошибки при его изложении на занятиях и в ходе промежуточной аттестации. Обучающийся испытывает определённые затруднения в применении теоретических положений при решении практических задач профессиональной направленности стандартного уровня сложности, владеет необходимыми для этого базовыми навыками и приёмами. Демонстрирует достаточный уровень знания учебной литературы по дисциплине.	Достаточный
0 – 60 баллов	Неудовлетворительно	Выставляется обучающемуся, если он не знает на базовом уровне теоретический и практический материал, допускает грубые ошибки при его изложении на занятиях и в ходе промежуточной аттестации. Обучающийся испытывает серьёзные затруднения в применении теоретических положений при решении практических задач профессиональной направленности стандартного уровня сложности, не владеет необходимыми для этого базовыми навыками и приёмами. Демонстрирует фрагментарные знания учебной литературы по дисциплине.	Компетенции не сформированы

4. ИТОГОВЫЕ ТЕСТОВЫЕ ЗАДАНИЯ ПО ДИСЦИПЛИНЕ

№ п/п	Контролируемые разделы (темы)	Тестовые задания	Код контролируемой компетенции (или ее части)
1	Методы многомерного статистического анализа и анализа нечисловой информации	**1. Перечислите основные критерии качества данных при использовании технологий интеллектуального анализа** **2. В чем суть метода стохастического градиентного спуска?** **3. Укажите несколько правильных ответов. Регрессионный анализ – одна из важнейших областей машинного обучения, представляющий собой набор статистических методов исследования регрессии, позволяющих:**	ОПК-2 ОПК-4

		а) определить степень детерминированности вариации критериальной переменной предикторами; б) предсказать значение критериальной переменной с помощью предикторов; в) определить вклад отдельных предикторов в вариацию критериальной переменной; г) предсказать, к какой категории принадлежит тот или иной объект, на основе его характеристик. **4. Сопоставьте названия методов и их описание.** Методы: 1) методы опорных векторов; 2) метод дерева решений; 3) метод случайного леса; 4) наивный байесовский метод. Описание методов: а) набор схожих алгоритмов для автоматической классификации объектов, при котором осуществляется перевод исходных векторов в пространство более высокой размерности и осуществляется поиск разделяющей гиперплоскости с максимальным зазором в этом пространстве б) алгоритм классификации объектов, при котором осуществляется построение древовидной структуры, состоящий из решающих правил вида «если ..., то ...», которые генерируются автоматически в процессе обучения на основании обобщения множества отдельных наблюдений. в) алгоритм автоматической классификации, который предполагает использование большого ансамбля решающих деревьев, каждое из которых само по себе даёт результат очень невысокого качества, но за счёт их большого количества качество результата повышается до приемлемого уровня. г) алгоритм классификации, который предполагает использование теоремы Байеса со строгими (наивными) предположениями о статистической независимости признаков **5. Расположите в правильной последовательности. Наивный байесовский метод – алгоритм классификации, который предполагает использование теоремы Байеса со строгими (наивными) предположениями о статистической независимости признаков. Классификация с применением наивного байесовского метода проводится в следующем порядке:** **1) для классифицируемого объекта вычисляются функции правдоподобия каждого из классов;** **2) по функциям правдоподобия каждого из классов вычисляются апостериорные вероятности классов;** **3) объект относится к тому классу, для которого апостериорная вероятность максимальна.** **6. Какие методы используются для обработки противоречий в данных?**	
2	Технологии хранения и обработки больших данных	**7. Что включает очистка данных при использовании технологий интеллектуального анализа** **8. Какие методы используются для обработки шумов в данных?** **9. Перечислите методы выбора признаков, используемые для уменьшения размерности данных при проведении интеллектуального анализа** **10. Перечислите методы выделения признаков, используемые для уменьшения размерности данных при проведении интеллектуального анализа**	ОПК-2 ОПК-4

		11. Вставьте пропущенное слово. Метод k-ближайших соседей – метрический алгоритм для автоматической … объектов, при котором объект присваивается тому классу, который является наиболее распространённым среди k соседей данного элемента, классы которых уже известны. **12. Укажите несколько правильных ответов. К особенностям метода k-ближайших соседей относят:** а) не требуется осуществления обучения перед выполнением прогнозов в реальном времени (в результате метод является легкореализуемым, позволяет добавлять новые данные, имеет высокую скорость выполнения); б) может быть применим к выборкам с большим количеством атрибутов (многомерным); в) перед применением необходимо определить функцию расстояния (метрику). г) строить разделяющую поверхность с использованием только небольшого подмножества точек, лежащих в зоне, критической для разделения, тогда как остальные, верно, классифицируемые наблюдения обучающей выборки вне этой зоны игнорируются.	
3	Программирование обработки и загрузки больших данных	**13. Перечислите способы машинного обучения** **14. Перечислите виды задач, решаемые с помощью машинного обучения** **15. Перечислите основные области классификации как задачи машинного обучения**	ОПК-2 ОПК-4
4	Аналитика в больших данных	**16. Вставьте пропущенное слово. Целями … в зависимости от решаемой задачи являются:** **1) расширение знаний о предметной области;** **2) сжатие данных (замена множества данных, входящим в кластер, одним представителем);** **3) обнаружение новизны (данных, которые невозможно отнести ни к одному из кластеров).** **17. Укажите несколько правильных ответов. Одним из основных методов обучения персептрона является метод коррекции ошибки. Существует следующие модификации метода коррекции ошибок, которые отличаются между собой в зависимости способом выбора величины и знака подкрепления:** а) метод коррекции ошибок без квантования; б) метод коррекции ошибок с квантованием; в) метод коррекции ошибок со случайным знаком подкрепления; г) метод коррекции ошибок со случайными возмущениями. **18. Вставьте пропущенное слово. Пассивно-агрессивные алгоритмы – это семейство алгоритмов, используемых для крупномасштабного обучения. Их особенность заключается в том, что обучение осуществляется не на основе одновременно поступающего на вход пакета данных, а на основе входных данных, которые поступают в … порядке (модель машинного обучения обновляется шаг за шагом по мере поступления новых данных).** **19. Укажите один неверный ответ. Персептрон является простейшим видом искусственных нейронных сетей. В основе персептрона лежит математическая модель восприятия информации мозгом. Персептрон в своем виде представляет систему из элементов следующих разных типов:** а) сенсоров; б) нейронов; в) ассоциативных элементов; г) реагирующих элементов.	ОПК-2 ОПК-4

		20. Вставьте пропущенное слово. Одним из основных методов обучения персептрона является метод коррекции …, который представляет собой такой метод обучения, при котором вес связи не изменяется до тех пор, пока текущая реакция персептрона остается правильной (при появлении неправильной реакции вес (величина подкрепления) изменяется на единицу, а знак подкрепления (+/-) определяется противоположным от знака ошибки). **21. На основе какого языка был создан R?** **22. Приведите округленный результат прогнозирования пустого значения в таблице с помощью функции ТЕНДЕНЦИЯ:** **Год** | **1** | **2** | **3** | **4** **Тыс. руб** | **67** | **120** | **160** | **23. Какая из следующих СУБД подходит для организации высоко-доступного и консистентного хранилища?** **24. Укажите факторы, способствовавшие появлению тренда больших данных**	

Ключ

1	2	3	4	5	6
Качество данных – обобщенное понятие, отражающее степень их пригодности к решению определенной задачи. Основными критериями качества данных являются: доступность, точность, взаимосвязанность, полнота, непротиворечивость, однозначность, релевантность, надежность и своевременность.	Метод стохастического градиентного спуска – методом оптимизации в машинном обучении. Стохастический градиентный спуск – это метод итерации для оптимизации целевой функции с подходящими свойствами гладкости (например, дифференцируемости или субдифференцируемости). В данном методе градиент оптимизируемой функции считается на каждом шаге как градиент от одного, случайно выбранного элемента.	а) б) в)	1а) 2б) 3в) 4г)	1-2-3	Для обработки противоречий используют следующие методы: исключение противоречивых значений (применяется в том случае, если причина противоречия вызвано ошибкой) и объединение записей с агрегированием числовых значений выходных атрибутов (применяется в том случае, если данные отражают реальные события).
7	**8**	**9**	**10**	**11**	**12**
Очистка данных включает обработку пропущенных значений, дубликатов, противоречий, аномальных	Для обработки шумов (обусловленных воздействием случайных факторов флуктуаций значений признаков) используют метод их	Методы выбора признаков оставляют некоторое подмножество исходного набора	Методы выделения признаков составляют из уже исходных признаков новые, все также	Классификации	а), б), в)

значений и выбросов, шумов, фиктивных значений и ошибок ввода данных	описания с помощью различных математических моделей в соответствии с их временной, спектральной и пространственной структурой с последующим исключением из основного набора данных.	признаков, избавляясь от признаков избыточных и слабо информативных. К методам выбора признаков относят: 1) методы фильтров (измеряют релевантность признаков на основе функции μ, и затем решают по правилу κ, какие признаки оставить в результирующем множестве); 2) оберточные методы (находят подмножество искомых признаков последовательно, используя некоторый классификатор в качестве источника оценки качества выбранных признаков); 3) встроенные методы (для выбора признаков используется непосредственно структуру некоторого классификатора) и другие.	полностью описывающие пространство набора данных, но уменьшая его размерность и теряя в репрезентативности данных, т.к. становится непонятно, за что отвечают новые признаки. К методам выделения признаков относят: 1) метод главных компонент; 2) метод разложения по усеченным сингулярным значениям; 3) метод неотрицательного матричного разложения; 4) метода линейного дискриминантного анализа.		
13	**14**	**15**	**16**	**17**	**18**
Способы машинного обучения делятся на несколько категорий: 1)	Все задачи, решаемые с помощью машинного обучения, относятся к одной из	Как правило, классификацию можно разбить на две области: а) бинарная	кластеризации	**а), б), в), г)**	Последовательном

обучение с учителем (обучение осуществляется принудительно с помощью примеров «стимул-реакция»); 2) обучение без учителя (обучение осуществляется спонтанно без вмешательства извне); 3) обучение с подкреплением (обучение осуществляется при взаимодействии с некоторой средой).	следующих категорий: 1) регрессия (является частным случаем задач прогнозирования, выполняется с помощью обучения с учителем на этапе тестирования); 2) классификация (выполняется с помощью обучения с учителем на этапе собственно обучения); 3) кластеризация (выполняется с помощью обучения без учителя); 4) понижение размерности данных (выполняется с помощью обучения без учителя); 5) выявление аномалий (выполняется с помощью обучения без учителя).	классификация, в которой требуется сгруппировать результат в одну из двух групп (0 или 1, истинное или ложное, положительное или отрицательное); б) мультиклассовая классификация, где требуется сгруппировать результат в одну из нескольких (более двух) групп.			
19	**20**	**21**	**22**	**23**	**24**
б)	Ошибки	S	**209**	Greenplum, BigTable	маркетинговые кампании крупных корпораций, снижение издержек на хранение данных

Критерии оценки

Оценка в баллах	**Оценка за итоговый тест**
65-80 баллов	«Отлично»
50-64 баллов	«Хорошо»
40-49 баллов	«Удовлетворительно»
Менее 40 баллов	«Неудовлетворительно»

Разработчик: ст. преподаватель Виноградов Д.В.

Фонд оценочных материалов (средств) рассмотрен и одобрен на заседании кафедры «Бизнес-информатика и экономика»

Протокол № 1 от 30.08.2023 года

Заведующий кафедрой д.э.н., профессор Тесленко И.Б.

Фонд оценочных материалов (средств) рассмотрен и одобрен на заседании учебно-методической комиссии направления 01.03.05 Статистика

Протокол № 1 от 05.09.2023 года

Председатель комиссии к.э.н., доцент Яресь О.Б.