

МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ
Федеральное государственное автономное образовательное учреждение
высшего образования
«СЕВЕРО-КАВКАЗСКИЙ ФЕДЕРАЛЬНЫЙ УНИВЕРСИТЕТ»

УТВЕРЖДАЮ

Декан факультета математики
и компьютерных наук имени
профессора. Н.И. Червякова,
канд. физ.-мат. наук, доцент
Грובה Т.А.

ФОНД ОЦЕНОЧНЫХ СРЕДСТВ ПО ДИСЦИПЛИНЕ
«Big Data Analysis and Mining / Анализ и обработка больших данных»

Направление подготовки	02.04.01 Математика и компьютерные науки
Направленность (профиль)	Training Program for Electronic Information
Год начала обучения	2026 г.
Форма обучения	очная
Реализуется в семестре	2

Введение

1. Назначение

Фонд оценочных средств по дисциплине «Big Data Analysis and Mining / Анализ и обработка больших данных» предназначен для объективной оценки уровня сформированности компетенций.

2. ФОС является приложением к программе дисциплины «Big Data Analysis and Mining / Анализ и обработка больших данных».

3. Разработчик Р. А. Воронкин, доцент департамента цифровых, робототехнических систем и электроники института перспективной инженерии.

4. Проведена экспертиза ФОС.

Члены экспертной группы:

Члены экспертной группы:

Председатель:

Поддубная Н.А. – председатель УМК факультета математики и компьютерных наук имени профессора Н.И. Червякова

Члены комиссии:

Гладков А.В. – член УМК кафедры вычислительной математики и кибернетики;

Андрухив Л.В. – член УМК кафедры математического моделирования;

Копыткова Л.Б. – член УМК кафедры математического анализа, алгебры и геометрии.

Представитель организации-работодателя: Лапин В.Г. – начальник отдела АСУ Ставропольского краевого клинического консультативно-диагностического центра.

Экспертное заключение: ФОС по дисциплине позволяет оценить уровень сформированности компетенций. Приступить к апробации.

5.Срок действия ФОС определяется сроком реализации образовательной программы.

1. Описание критериев оценивания компетенции на различных этапах их формирования, описание шкал оценивания

Компетенция (ии), индикатор (ы)	Уровни сформированности компетенци(ий),			
	Минимальный уровень не достигнут (Неудовлетвори тельно) 2 балла	Минимальный уровень (удовлетворите льно) 3 балла	Средний уровень (хорошо) 4 балла	Высокий уровень (отлично) 5 баллов
ОПК-2				
Результаты обучения по дисциплине (модулю): <i>Индикатор:</i> ИД-1 _{ОПК-2} Проводит исследования новых математических моделей в естественных науках	Не способен обоснованно выбрать или описать математическую модель для исследования данных в рамках дисциплины Big Data Analysis and Mining. Допускает существенные ошибки при постановке исследовательской задачи, не понимает связи между свойствами данных, выбранной моделью и методами анализа больших данных.	Способен описать отдельные математические модели, применяемые при анализе больших данных, но испытывает затруднения при их обоснованном выборе и адаптации к исследовательской задаче. Может применять базовые методы классификации, кластеризации, регрессии или поиска закономерностей, однако допускает ошибки в интерпретации результатов и оценке ограничений модели.	Уверенно выбирает и применяет математические модели для исследования больших данных с учетом структуры, объема и качества исходных данных. Способен обосновать выбор модели, выполнить анализ результатов, оценить качество решения и выявить основные ограничения применяемого подхода.	Полностью овладеп подходами к исследованию и применению математических моделей в задачах анализа больших данных. Способен самостоятельно формулировать исследовательскую задачу, выбирать, модифицировать и обосновывать математические модели, применять их в распределенной среде обработки данных, интерпретировать результаты и делать аргументированные выводы о применимости модели к задачам естественных наук.

Результаты обучения по дисциплине (модулю): Индикатор: ИД-2_{опк-2} Разрабатывает новые и совершенствует существующие концепции, теории и методы	Не способен предложить обоснованные варианты разработки или совершенствования методов анализа больших данных. Не понимает различий между существующими подходами к обработке, анализу и интеллектуальному извлечению закономерностей из больших данных, допускает существенные ошибки при объяснении их назначения и ограничений..	Способен описать отдельные существующие методы анализа больших данных, включая методы распределенной обработки, классификации, кластеризации, регрессии или поиска ассоциативных правил, но испытывает затруднения при их сравнении, адаптации и совершенствовании. Предлагаемые улучшения носят частичный характер и не всегда достаточно обоснованы с точки зрения качества, масштабируемости и/или применимости к исследуемым данным.	Уверенно анализирует существующие концепции, методы и подходы, применяемые в Big Data Analysis and Mining. Способен предложить обоснованные способы их совершенствования с учетом структуры данных, объема выборки, вычислительных ограничений, требований к качеству результатов и особенностей распределенной обработки.	Полностью овладевает подходами к разработке и совершенствованию концепций, теорий и методов анализа больших данных. Способен самостоятельно формулировать новые методические решения, адаптировать и модифицировать существующие алгоритмы и модели, обосновывать их преимущества, оценивать ограничения, проводить сравнительный анализ альтернатив и аргументированно выбирать наиболее эффективный подход для решения исследовательских и прикладных задач в области Big Data Analysis and Mining.
--	---	---	---	--

ПК-6

Результаты обучения по дисциплине (модулю): Индикатор: ИД-1_{ПК-6} С помощью современных пакетов прикладных программ моделирования, исследует вновь создаваемые математические модели;	Не способен использовать современные пакеты прикладных программ и программные среды для моделирования и исследования математических моделей в задачах анализа больших данных. Испытывает	Способен использовать отдельные современные программные средства для моделирования и анализа больших данных, включая Hadoop, Apache Spark, PySpark, Spark SQL или Spark MLlib, но выполняет работу преимущественно	Уверенно применяет современные пакеты прикладных программ и программные среды для исследования математических моделей в рамках дисциплины Big Data Analysis and Mining. Способен подготовить	Полностью овладевает применением современных программных средств моделирования и анализа больших данных. Способен самостоятельно выбирать и использовать Hadoop, Apache Spark, PySpark, Spark SQL, Spark MLlib и другие
--	---	---	---	--

	существенные затруднения при работе с инструментами обработки данных, не может корректно подготовить данные, запустить модель и интерпретировать полученные результаты.	по образцу. Допускает ошибки при подготовке данных, настройке модели, выборе параметров и анализе результатов исследования.	данные, реализовать модель или вычислительный эксперимент, выполнить обработку результатов, оценить качество модели и обосновать применимость выбранного инструментария.	инструменты для исследования вновь создаваемых математических моделей, проводить вычислительные эксперименты, сравнивать результаты, оценивать масштабируемость, устойчивость и качество модели, а также делать аргументированные выводы о её применимости к исследовательским и прикладным задачам.
Результаты обучения по дисциплине (модулю): <i>Индикатор:</i> ИД-2 _{ПК-6} Использует современные методы разработки алгоритмов математических моделей	Не способен использовать современные методы разработки алгоритмов математических моделей в задачах анализа больших данных. Допускает существенные ошибки при описании алгоритма, не учитывает особенности распределенной обработки, объем данных, вычислительную сложность и ограничения применяемых моделей.	Способен применять отдельные современные методы разработки алгоритмов математических моделей в рамках дисциплины Big Data Analysis and Mining, но выполняет работу преимущественно по образцу. Может описать базовый алгоритм обработки или анализа данных, однако испытывает затруднения при его адаптации к большим данным, оценке вычислительной эффективности, масштабируемости и качества получаемых результатов.	Уверенно использует современные методы разработки алгоритмов математических моделей для решения задач анализа больших данных. Способен выбрать подходящий алгоритмический подход, реализовать его с учетом особенностей распределенной обработки, выполнить оценку качества, производительности и применимости модели к исследуемым данным.	Полностью овладел современными методами разработки алгоритмов математических моделей в области Big Data Analysis and Mining. Способен самостоятельно проектировать, модифицировать и обосновывать алгоритмы для обработки и интеллектуального анализа больших данных, учитывать требования масштабируемости, устойчивости, воспроизводимости и вычислительной эффективности, а также реализовывать и исследовать алгоритмы с использованием современных инструментов распределенной обработки данных.
Результаты обучения по дисциплине (модулю):	Не способен использовать	Способен реализовывать	Уверенно использует	Полностью овладел применением

<i>Индикатор:</i> ИД-З_{ПК-6} Использует современные языки программирования для реализации конкретных алгоритмов и математических моделей	современные языки программирования для реализации алгоритмов и математических моделей в рамках дисциплины Big Data Analysis and Mining. Допускает существенные ошибки в программной реализации, не понимает структуры обрабатываемых данных, не может корректно запустить алгоритм, получить и интерпретировать результат.	отдельные алгоритмы обработки и анализа больших данных на современных языках программирования, включая Python, PySpark, Spark SQL или другие используемые в курсе средства, но выполняет работу преимущественно по образцу. Допускает ошибки в структуре программы, обработке данных, настройке параметров модели и проверке корректности полученных результатов.	современные языки программирования и программные интерфейсы для реализации конкретных алгоритмов и математических моделей в задачах анализа больших данных. Способен подготовить данные, реализовать алгоритм обработки или интеллектуального анализа, выполнить запуск в распределенной среде, проверить корректность работы программы и оценить качество полученных результатов.	современных языков программирования для реализации алгоритмов и математических моделей в области Big Data Analysis and Mining. Способен самостоятельно разрабатывать программную реализацию алгоритмов обработки, анализа и интеллектуального извлечения закономерностей из больших данных, использовать Python, PySpark, Spark SQL, Spark MLlib и другие современные средства, учитывать требования масштабируемости и производительности и вычислительной эффективности, а также аргументированно интерпретировать результаты работы реализованных моделей.
--	---	--	---	--

Оценивание уровня сформированности компетенции по дисциплине осуществляется на основе «Положения о проведении текущего контроля успеваемости и промежуточной аттестации обучающихся по образовательным программам высшего образования - программам бакалавриата, программам специалитета, программам магистратуры - в федеральном государственном автономном образовательном учреждении высшего образования «Северо-Кавказский федеральный университет» в актуальной редакции.

ОЦЕНОЧНЫЕ СРЕДСТВА ДЛЯ ПРОВЕРКИ УРОВНЯ СФОРМИРОВАННОСТИ КОМПЕТЕНЦИЙ

Номер задания	Правильный ответ	Содержание вопроса	Компетенция
		Форма обучения очная, Семестр 2 (3)	
1	Большие данные характеризуются значительным объемом, высокой скоростью поступления, разнообразием форматов, достоверностью и ценностью извлекаемой информации	Дайте определение понятию «большие данные» с учетом признаков 5V	ОПК-2
2	HDFS	Какая технология используется в экосистеме Hadoop для распределенного хранения больших файлов? А) Spark SQL Б) HDFS В) MLlib Г) Kafka Streams	ПК-6
3	MapReduce представляет вычисления в виде этапов Map, Shuffle и Reduce	Опишите основную идею модели распределенной обработки данных MapReduce	ПК-6
4	Apache Spark выполняет вычисления в памяти и поддерживает построение DAG вычислений, что делает его более гибким и быстрым по сравнению с классическим MapReduce	В чем состоит основное преимущество Apache Spark по сравнению с классическим Hadoop MapReduce?	ПК-6
5	RDD, DataFrame, Dataset	Назовите основные абстракции данных, используемые в Apache Spark для распределенной обработки данных	ПК-6
6	Parquet	Какой формат хранения данных является колоночным и часто применяется при обработке больших данных в Hadoop и Spark? А) TXT Б) CSV В) Parquet Г) XML	ПК-6
7	Классификация используется для отнесения объектов к заранее заданным классам на основе обучающей выборки	Что такое классификация как задача интеллектуального анализа больших данных?	ОПК-2
8	Кластеризация используется для группировки объектов по сходству без	Что такое кластеризация в задачах Data Mining на больших данных?	ОПК-2

	заранее заданных меток классов		
9	Support показывает частоту появления набора элементов, confidence показывает надежность правила, lift показывает силу связи между элементами относительно случайного совпадения	Объясните смысл метрик support, confidence и lift при поиске ассоциативных правил	ОПК-2
10	Качество модели оценивается с помощью метрик, соответствующих типу задачи: accuracy, precision, recall, F1 для классификации; RMSE, MAE, R ² для регрессии; silhouette score для кластеризации	Какие метрики могут использоваться для оценки качества моделей интеллектуального анализа больших данных?	ОПК-2
11	Репликация в HDFS повышает отказоустойчивость хранения данных за счет размещения копий блоков на разных узлах кластера	Для чего используется механизм репликации блоков в HDFS?	ПК-6
12	NameNode хранит метаданные файловой системы, а DataNode хранит непосредственно блоки данных	Укажите различие между NameNode и DataNode в архитектуре HDFS	ПК-6
13	Shuffle	Как называется этап распределенной обработки, при котором данные перераспределяются между узлами по ключам перед выполнением агрегации? А) Cache Б) Shuffle В) Checkpoint Г) Broadcast	ПК-6
14	Lazy evaluation означает, что преобразования данных в Spark не выполняются немедленно, а откладываются до вызова действия	Что означает принцип ленивого выполнения вычислений в Apache Spark?	ПК-6
15	Transformation создает новое распределенное представление данных, а action запускает вычисления и возвращает результат или сохраняет данные	В чем различие между transformation и action в Apache Spark?	ПК-6

16	Spark SQL позволяет выполнять SQL-запросы к распределенным данным и использовать оптимизатор выполнения запросов	Для чего применяется Spark SQL в задачах анализа больших данных?	ПК-6
17	Catalyst Optimizer	Как называется встроенный оптимизатор запросов в Apache Spark SQL? А) YARN Scheduler Б) Catalyst Optimizer В) HDFS Manager Г) ML Pipeline	ПК-6
18	Broadcast join целесообразен, когда одна из соединяемых таблиц достаточно мала и может быть передана на все узлы кластера	В каком случае целесообразно использовать broadcast join в Spark?	ПК-6
19	Партиционирование позволяет разделить данные на части и повысить эффективность параллельной обработки	Какую роль играет партиционирование данных в системах Big Data?	ПК-6
20	Cache используется для сохранения промежуточных данных в памяти при их многократном использовании	Для чего применяется кэширование данных в Apache Spark?	ПК-6
21	DataFrame представляет распределенную коллекцию данных со схемой и именованными столбцами	Что такое DataFrame в Apache Spark?	ПК-6
22	Structured Streaming используется для обработки потоковых данных на основе модели DataFrame и SQL	Какой компонент Apache Spark применяется для потоковой обработки данных?	ПК-6
23	Checkpoint сохраняет состояние потоковой обработки и позволяет восстановить вычисления после сбоя	Для чего используется checkpointing в Spark Structured Streaming?	ПК-6
24	Временное окно используется для группировки потоковых событий по интервалам времени	Что такое оконная обработка событий в потоковой аналитике?	ОПК-2
25	Data Mining представляет собой процесс выявления скрытых закономерностей, зависимостей и	Дайте определение понятию Data Mining	ОПК-2

	полезных знаний в данных		
26	Предобработка данных включает очистку, преобразование типов, обработку пропусков, кодирование признаков и нормализацию	Какие основные операции включает предобработка данных перед применением моделей Data Mining?	ОПК-2
27	Утечка данных возникает, когда в обучающую выборку попадает информация, недоступная на момент реального применения модели	Что такое утечка данных при построении моделей машинного обучения?	ОПК-2
28	Переобучение означает, что модель хорошо работает на обучающих данных, но плохо обобщает закономерности на новых данных	Что такое переобучение модели в задачах анализа данных?	ОПК-2
29	Drift означает изменение статистических свойств данных или зависимости между признаками и целевой переменной во времени	Что такое дрейф данных или модели в задачах Big Data Analysis and Mining?	ОПК-2
30	Spark MLlib	Какая библиотека Apache Spark предназначена для реализации алгоритмов машинного обучения и интеллектуального анализа данных? А) HDFS Б) Spark MLlib В) YARN Г) Hive Metastore	ПК-6
31	YARN управляет ресурсами кластера и распределяет вычислительные задачи между узлами	Для чего используется YARN в экосистеме Hadoop?	ПК-6
32	Hive позволяет выполнять SQL-подобные запросы к данным, хранящимся в распределенной файловой системе	Какую роль выполняет Apache Hive в экосистеме больших данных?	ПК-6
33	DAG представляет вычисления в виде направленного ациклического графа зависимостей между операциями	Что такое DAG в Apache Spark?	ПК-6
34	Driver управляет выполнением Spark-приложения, а executors выполняют	Укажите различие между driver и executor в архитектуре Apache Spark	ПК-6

	задачи обработки данных на рабочих узлах		
35	Stage представляет набор задач, которые могут быть выполнены без дополнительного shuffle	Что такое stage в модели выполнения Apache Spark?	ПК-6
36	Task представляет минимальную единицу выполнения работы над отдельной партицией данных	Что такое task в Apache Spark?	ПК-6
37	reduceByKey обычно эффективнее groupByKey, так как выполняет частичную агрегацию данных до передачи по сети	Почему при агрегации данных в Spark часто рекомендуется использовать reduceByKey вместо groupByKey?	ПК-6
38	repartition увеличивает или изменяет число партиций с полным перераспределением данных, а coalesce обычно уменьшает число партиций с меньшими затратами	В чем различие между операциями repartition и coalesce в Apache Spark?	ПК-6
39	Перекося данных возникает, когда отдельные партиции или ключи содержат значительно больше данных, чем остальные, что замедляет выполнение задачи	Что такое data skew в распределенной обработке данных?	ПК-6
40	Spark SQL, DataFrame API и Spark MLlib	Какие компоненты Apache Spark могут использоваться для реализации полного цикла анализа больших данных: от обработки таблиц до построения моделей машинного обучения?	ПК-6
41	ETL включает извлечение, преобразование и загрузку данных	Раскройте содержание процесса ETL в задачах подготовки больших данных	ОПК-2
42	Слой raw хранит исходные данные, слой cleaned содержит очищенные данные, слой curated содержит подготовленные аналитические наборы или витрины данных	Объясните назначение слоев raw, cleaned и curated в batch-пайплайне обработки данных	ОПК-2

43	Идемпотентность означает, что повторный запуск обработки с теми же входными данными приводит к тому же результату без нежелательных побочных эффектов	Что означает идемпотентность пайплайна обработки данных?	ОПК-2
44	Схема данных определяет структуру набора данных, типы столбцов и правила интерпретации значений	Для чего используется схема данных при обработке больших данных?	ОПК-2
45	Schema evolution позволяет изменять структуру данных во времени при сохранении совместимости с ранее записанными данными	Что такое эволюция схемы данных в системах Big Data?	ОПК-2
46	Аномалия представляет собой объект или событие, существенно отличающееся от типичного поведения данных	Что такое аномалия в задачах интеллектуального анализа больших данных?	ОПК-2
47	Регрессия используется для предсказания числового значения на основе входных признаков	Что такое регрессия как задача Data Mining?	ОПК-2
48	Обучающая выборка используется для настройки модели, тестовая выборка используется для независимой оценки качества модели	В чем различие между обучающей и тестовой выборками при построении моделей анализа данных?	ОПК-2
49	Feature engineering представляет собой процесс создания, отбора и преобразования признаков для повышения качества модели	Что такое feature engineering в задачах анализа больших данных?	ОПК-2
50	Pipeline в Spark MLlib объединяет последовательность этапов обработки данных, подготовки признаков и обучения модели	Для чего используется Pipeline в Spark MLlib?	ПК-6
51	Batch processing используется для обработки накопленных данных за	В чем различие между пакетной и потоковой обработкой данных?	ОПК-2

	определенный период, streaming processing используется для обработки данных по мере их поступления		
52	Event time отражает время фактического возникновения события, processing time отражает время обработки события системой	В чем различие между event time и processing time в потоковой обработке данных?	ОПК-2
53	Late events представляют собой события, поступившие в систему обработки позже ожидаемого времени	Что такое late events в потоковой аналитике больших данных?	ОПК-2
54	Watermark позволяет ограничить время ожидания запоздавших событий и управлять состоянием потоковой обработки	Для чего используется watermark в Spark Structured Streaming?	ПК-6
55	Exactly-once означает, что каждое событие учитывается в результате обработки ровно один раз, несмотря на возможные сбои системы	Что означает семантика обработки exactly-once в потоковых системах?	ОПК-2
56	Kafka используется как распределенная платформа для приема, хранения и передачи потоковых событий	Какую роль Apache Kafka может выполнять в архитектуре обработки больших данных?	ПК-6
57	Data lake хранит данные разных форматов в исходном или слабо обработанном виде, а data warehouse хранит структурированные данные, подготовленные для аналитических запросов	В чем различие между data lake и data warehouse?	ОПК-2
58	Lakehouse сочетает гибкость data lake с возможностями управления, транзакционности и аналитической обработки, характерными для data warehouse	Что такое архитектура lakehouse в системах больших данных?	ОПК-2

59	Data mart представляет собой предметно-ориентированную витрину данных, подготовленную для аналитических задач конкретного подразделения или направления	Что такое витрина данных в аналитической архитектуре?	ОПК-2
60	Data catalog используется для описания, поиска, классификации и управления метаданными наборов данных	Для чего используется каталог данных в Big Data-системах?	ОПК-2
61	Метаданные описывают структуру, происхождение, формат, владельца, качество и правила использования данных	Что такое метаданные в системах управления большими данными?	ОПК-2
62	Data lineage показывает происхождение данных, историю их преобразований и связи между источниками, промежуточными наборами и итоговыми результатами	Что означает понятие data lineage?	ОПК-2
63	Data quality включает полноту, точность, согласованность, актуальность, уникальность и достоверность данных	Какие характеристики обычно учитываются при оценке качества данных?	ОПК-2
64	Constraint validation позволяет проверять данные на соответствие заданным ограничениям, например типам, диапазонам значений, уникальности и отсутствию пропусков	Для чего применяется проверка ограничений данных в процессе подготовки больших данных?	ПК-6
65	Мониторинг данных и моделей позволяет выявлять ухудшение качества данных, изменение распределений, сбои обработки и снижение качества аналитических результатов	Для чего используется мониторинг в задачах Big Data Analysis and Mining?	ОПК-2

66	Reproducibility означает возможность повторить обработку данных или обучение модели и получить сопоставимый результат при тех же условиях	Что означает воспроизводимость вычислительного эксперимента в анализе больших данных?	ОПК-2
67	Версионирование данных и кода позволяет отслеживать изменения, воспроизводить результаты экспериментов и сравнивать разные варианты обработки	Зачем используется версионирование в проектах анализа больших данных?	ОПК-2
68	Sampling применяется для получения представительной части большого набора данных с целью предварительного анализа, отладки алгоритмов или снижения вычислительных затрат	Для чего используется выборка данных при анализе больших массивов?	ОПК-2
69	Stratified sampling сохраняет пропорции классов или групп в выборке, что важно при несбалансированных данных	В чем особенность стратифицированной выборки?	ОПК-2
70	Дисбаланс классов возникает, когда объекты одного класса существенно преобладают над объектами другого класса, что может привести к смещению модели	Что такое дисбаланс классов в задачах классификации больших данных?	ОПК-2
Номер задания	Правильный ответ	Содержание вопроса	Компетенции
71	Нормализация приводит числовые признаки к сопоставимому масштабу, что повышает корректность работы многих алгоритмов анализа данных	Для чего применяется нормализация признаков при подготовке данных к интеллектуальному анализу?	ОПК-2
72	One-hot encoding преобразует категориальные признаки в набор	Какой метод кодирования категориальных признаков создает отдельный бинарный столбец для каждого	ПК-6

	бинарных признаков	значения категории? А) Scaling Б) One-hot encoding В) Sampling Г) Checkpointing	
73	VectorAssembler объединяет несколько признаков в один вектор признаков, используемый моделями Spark MLlib	Для чего используется компонент VectorAssembler в Spark MLlib?	ПК-6
74	Estimator обучается на данных и создает модель, Transformer преобразует входные данные в новый набор данных	В чем различие между Estimator и Transformer в Spark MLlib?	ПК-6
75	Cross-validation используется для более надежной оценки качества модели путем обучения и проверки на нескольких разбиениях данных	Для чего применяется кросс-валидация при построении моделей интеллектуального анализа данных?	ОПК-2
76	Grid search перебирает заданные комбинации гиперпараметров модели для выбора наилучшего варианта по выбранной метрике качества	Что такое подбор гиперпараметров методом grid search?	ОПК-2
77	Accuracy показывает долю правильных предсказаний среди всех объектов	Что показывает метрика accuracy в задаче классификации?	ОПК-2
78	Precision показывает долю правильно предсказанных положительных объектов среди всех объектов, отнесенных моделью к положительному классу	Что показывает метрика precision в задаче классификации?	ОПК-2
79	Recall показывает долю правильно найденных положительных объектов среди всех реально положительных объектов	Что показывает метрика recall в задаче классификации?	ОПК-2
80	F1-мера является гармоническим средним precision и recall	Что представляет собой F1-мера в задаче классификации?	ОПК-2
81	RMSE показывает среднеквадратичную ошибку прогноза и сильнее штрафует крупные ошибки	Что показывает метрика RMSE в задаче регрессии?	ОПК-2

82	MAE показывает среднюю абсолютную величину ошибки прогноза	Что показывает метрика MAE в задаче регрессии?	ОПК-2
83	R^2 показывает долю вариации целевой переменной, объясненную моделью	Что показывает коэффициент детерминации R^2 в задаче регрессии?	ОПК-2
84	Silhouette score оценивает качество кластеризации на основе сходства объектов внутри кластера и различия между кластерами	Для чего используется метрика silhouette score в задачах кластеризации?	ОПК-2
85	K-means разбивает объекты на заданное число кластеров путем минимизации расстояния объектов до центров кластеров	Опишите основную идею алгоритма k-means	ОПК-2
86	FP-growth используется для поиска частых наборов элементов и построения ассоциативных правил	Какой алгоритм применяется в Spark MLlib для поиска частых наборов элементов? А) K-means Б) FP-growth В) Linear Regression Г) Random Forest	ПК-6
87	Random Forest строит ансамбль деревьев решений и объединяет их результаты для повышения устойчивости и качества предсказания	В чем состоит основная идея алгоритма Random Forest?	ОПК-2
88	Логистическая регрессия используется для решения задач классификации, чаще всего бинарной классификации	Для решения какого типа задач обычно применяется логистическая регрессия?	ОПК-2
89	Линейная регрессия используется для моделирования зависимости числовой целевой переменной от одного или нескольких признаков	Для чего применяется линейная регрессия в задачах анализа данных?	ОПК-2
90	Confusion matrix показывает количество верных и ошибочных классификаций по каждому классу	Для чего используется матрица ошибок в задаче классификации?	ОПК-2
91	ROC-кривая показывает зависимость доли истинно положительных классификаций от доли	Что показывает ROC-кривая в задаче классификации?	ОПК-2

	ложноположительных классификаций при разных порогах принятия решения		
92	AUC характеризует обобщенную способность модели различать классы: чем выше AUC, тем лучше модель разделяет объекты разных классов	Что означает показатель AUC при оценке качества классификационной модели?	ОПК-2
93	PR-кривая показывает зависимость precision от recall и особенно полезна при оценке моделей на несбалансированных данных	Для чего используется precision-recall curve в задачах классификации?	ОПК-2
94	Oversampling увеличивает число объектов миноритарного класса, undersampling уменьшает число объектов мажоритарного класса	В чем различие между oversampling и undersampling при работе с несбалансированными данными?	ОПК-2
95	Feature selection используется для отбора наиболее значимых признаков и снижения размерности данных	Для чего применяется отбор признаков при построении моделей анализа больших данных?	ОПК-2
96	PCA преобразует исходные признаки в новые некоррелированные компоненты, сохраняющие наибольшую часть дисперсии данных	В чем состоит основная идея метода главных компонент?	ОПК-2
97	Dimensionality reduction снижает число признаков, уменьшая вычислительную сложность и риск переобучения модели	Для чего применяется снижение размерности в задачах Big Data Analysis and Mining?	ОПК-2
98	Outlier detection направлен на выявление объектов, существенно отличающихся от основной массы данных	Что понимается под обнаружением выбросов в интеллектуальном анализе больших данных?	ОПК-2
99	Обучение без учителя применяется, когда в данных отсутствуют заранее заданные целевые метки	В каком случае используется обучение без учителя?	ОПК-2
100	Обучение с учителем используется,	В каком случае применяется обучение с учителем в	ОПК-2

	когда для объектов известны правильные ответы или целевые значения	задачах Data Mining?	
101	Apache Spark может читать данные из HDFS, локальной файловой системы, объектных хранилищ, баз данных и потоковых источников	Из каких источников Apache Spark может загружать данные для последующей обработки?	ПК-6
102	select выбирает столбцы, filter отбирает строки по условию, groupBy группирует данные для последующей агрегации	Укажите назначение операций select, filter и groupBy в Spark DataFrame API	ПК-6
103	explain выводит логический и физический план выполнения запроса или операции в Spark	Для чего используется метод explain в Apache Spark?	ПК-6
104	Narrow dependency не требует перераспределения данных между узлами, wide dependency обычно приводит к shuffle	В чем различие между узкой и широкой зависимостью в Apache Spark?	ПК-6
105	Broadcast-переменные позволяют эффективно передавать неизменяемые данные на все executors без многократной пересылки	Для чего используются broadcast-переменные в Apache Spark?	ПК-6
106	Accumulator используется для накопления агрегированных значений в процессе распределенных вычислений	Для чего применяются accumulators в Apache Spark?	ПК-6
107	persist позволяет выбрать уровень хранения данных, а cache обычно использует хранение в памяти по умолчанию	В чем различие между cache и persist в Apache Spark?	ПК-6
108	Spark-приложение состоит из driver-программы, SparkContext или SparkSession, набора transformations и	Назовите основные элементы Spark-приложения	ПК-6

	actions, выполняемых на executors		
109	ML Pipeline обеспечивает последовательное выполнение этапов подготовки данных, преобразования признаков, обучения и применения модели	Какую роль выполняет ML Pipeline при реализации моделей в Spark MLlib?	ПК-6
110	TrainValidationSplit выполняет разбиение данных на обучающую и проверочную части для выбора параметров модели по заданной метрике	Для чего используется TrainValidationSplit в Spark MLlib?	ПК-6
111	LogisticRegression в Spark MLlib используется для построения классификационной модели на основе признакового вектора и целевой метки	Для чего применяется класс LogisticRegression в Spark MLlib?	ПК-6
112	DecisionTreeClassifier строит дерево решений, в котором выбор класса выполняется на основе последовательности проверок значений признаков	В чем состоит основная идея классификатора DecisionTreeClassifier в Spark MLlib?	ПК-6
113	RandomForestClassifier объединяет несколько деревьев решений и использует голосование для повышения качества и устойчивости классификации	Как работает RandomForestClassifier в задачах классификации больших данных?	ПК-6
114	GBTClsifier строит ансамбль деревьев последовательно, исправляя ошибки предыдущих моделей	В чем состоит принцип работы градиентного бустинга в задачах классификации?	ОПК-2
115	LinearRegression используется для построения модели зависимости числовой целевой переменной от набора признаков	Для чего применяется LinearRegression в Spark MLlib?	ПК-6
116	RegressionEvaluator применяется для	Для чего используется RegressionEvaluator в Spark	ПК-6

	оценки качества регрессионной модели по выбранной метрике, например RMSE, MAE или R^2	MLlib?	
117	MulticlassClassificationEvaluator позволяет оценивать качество классификационной модели по метрикам accuracy, F1, precision или recall	Для чего применяется MulticlassClassificationEvaluator в Spark MLlib?	ПК-6
118	BinaryClassificationEvaluator используется для оценки моделей бинарной классификации, чаще всего по метрикам areaUnderROC или areaUnderPR	Для чего используется BinaryClassificationEvaluator в Spark MLlib?	ПК-6
119	ParamGridBuilder задает набор комбинаций гиперпараметров для подбора оптимальной модели	Какую роль выполняет ParamGridBuilder в Spark MLlib?	ПК-6
120	CrossValidator выполняет кросс-валидацию модели на нескольких разбиениях данных и выбирает наилучшую комбинацию параметров	Для чего используется CrossValidator в Spark MLlib?	ПК-6
121	StringIndexer преобразует категориальный текстовый признак в числовые индексы	Для чего используется StringIndexer при подготовке данных в Spark MLlib?	ПК-6
122	OneHotEncoder преобразует категориальные индексы в бинарные векторы признаков	Для чего используется OneHotEncoder в Spark MLlib?	ПК-6
123	StandardScaler масштабирует числовые признаки, приводя их к сопоставимому виду для последующего обучения модели	Для чего применяется StandardScaler при построении моделей анализа данных?	ПК-6
124	Imputer используется для заполнения пропущенных значений в числовых признаках	Для чего применяется Imputer в Spark MLlib?	ПК-6

125	Tokenizer разбивает текст на отдельные слова или токены для последующей обработки	Для чего используется Tokenizer при анализе текстовых данных?	ПК-6
126	HashingTF преобразует набор токенов в числовой вектор признаков на основе хеширования	Какую задачу решает HashingTF при обработке текстовых данных в Spark MLlib?	ПК-6
127	IDF уменьшает вес часто встречающихся слов и повышает значимость более информативных терминов	Для чего применяется IDF при построении текстовых признаков?	ОПК-2
128	TF-IDF используется для числового представления текста с учетом частоты термина и его информативности в коллекции документов	Что такое TF-IDF в задачах анализа текстовых данных?	ОПК-2
129	ALS используется для построения рекомендательных систем на основе матричной факторизации	Какой алгоритм Spark MLlib применяется для построения рекомендательных систем? А) ALS Б) K-means В) PCA Г) StringIndexer	ПК-6
130	Рекомендательная система прогнозирует предпочтения пользователя и предлагает объекты, которые могут быть ему интересны	Что такое рекомендательная система в задачах интеллектуального анализа больших данных?	ОПК-2
131	DataFrameWriter используется для записи обработанных данных в файлы, таблицы или внешние хранилища	Для чего применяется DataFrameWriter в Apache Spark?	ПК-6
132	Partition pruning позволяет Spark читать только те партиции данных, которые необходимы для выполнения запроса	Что такое partition pruning при обработке больших данных?	ПК-6
133	Predicate pushdown переносит фильтрацию данных ближе к источнику хранения, уменьшая объем считываемых данных	Для чего используется predicate pushdown в Spark SQL?	ПК-6
134	Bucketing распределяет данные по	Что такое bucketing в системах обработки больших	ПК-6

	корзинам на основе значений одного или нескольких столбцов и может ускорять соединения и агрегации	данных?	
135	Сериализация преобразует объекты в формат, пригодный для передачи по сети или сохранения, что важно для распределенных вычислений	Почему сериализация важна при выполнении Spark-приложений?	ПК-6
136	Executor memory определяет объем памяти, доступный исполнителю Spark для выполнения задач и хранения промежуточных данных	Для чего настраивается параметр executor memory в Apache Spark?	ПК-6
137	Driver memory определяет объем памяти, доступный управляющей программе Spark-приложения	Что задает параметр driver memory в Apache Spark?	ПК-6
138	Количество executors и cores влияет на степень параллелизма, скорость обработки и использование ресурсов кластера	Как количество executors и cores влияет на выполнение Spark-приложения?	ПК-6
139	Spark UI позволяет анализировать выполнение jobs, stages и tasks, выявлять shuffle, перекос данных, ошибки и узкие места производительности	Для чего используется Spark UI при разработке и отладке приложений обработки больших данных?	ПК-6
140	DAG visualization помогает увидеть последовательность операций и зависимости между этапами вычислений	Для чего применяется визуализация DAG в Apache Spark?	ПК-6
141	Data preprocessing снижает влияние шумов, пропусков, выбросов и несогласованности данных на качество модели	Почему предобработка данных является обязательным этапом интеллектуального анализа больших данных?	ОПК-2
142	Missing values могут искажать статистические характеристики	Почему необходимо обрабатывать пропущенные значения перед построением модели анализа данных?	ОПК-2

	данных и приводить к ошибкам обучения модели		
143	Выбросы могут существенно смещать параметры модели и ухудшать качество прогнозирования или кластеризации	Как выбросы влияют на результаты интеллектуального анализа данных?	ОПК-2
144	Балансировка классов позволяет уменьшить смещение модели в сторону преобладающего класса	Для чего выполняется балансировка классов при построении классификационной модели?	ОПК-2
145	Train-test split позволяет отделить данные для обучения модели от данных для независимой проверки ее качества	Для чего используется разбиение данных на обучающую и тестовую выборки?	ОПК-2
146	Validation set используется для выбора параметров модели и настройки алгоритма без использования тестовой выборки	Для чего применяется валидационная выборка при разработке модели анализа данных?	ОПК-2
147	Гиперпараметры задаются до обучения модели и управляют процессом обучения, а параметры модели настраиваются в ходе обучения на данных	В чем различие между параметрами модели и гиперпараметрами?	ОПК-2
148	Интерпретируемость модели означает возможность объяснить, какие признаки и зависимости повлияли на полученный результат	Что означает интерпретируемость модели в задачах Data Mining?	ОПК-2
149	Explainable AI направлен на объяснение решений моделей машинного обучения и повышение доверия к результатам анализа	Для чего используются методы explainable AI при интеллектуальном анализе данных?	ОПК-2
150	Robustness означает устойчивость модели к шуму, выбросам, изменению части данных и незначительным	Что такое устойчивость модели в задачах анализа больших данных?	ОПК-2

	изменениям входных признаков		
151	Scalability означает способность системы или алгоритма сохранять приемлемую производительность при росте объема данных, числа пользователей или вычислительных узлов	Что означает масштабируемость в задачах обработки и анализа больших данных?	ОПК-2
152	Horizontal scaling предполагает увеличение числа узлов в кластере, vertical scaling предполагает увеличение ресурсов одного узла	В чем различие между горизонтальным и вертикальным масштабированием?	ОПК-2
153	Fault tolerance позволяет системе продолжать работу или восстанавливаться после отказа отдельных узлов, задач или компонентов	Для чего необходима отказоустойчивость в Big Data-системах?	ОПК-2
154	Data locality означает выполнение вычислений как можно ближе к месту хранения данных, что снижает затраты на передачу данных по сети	Что такое data locality в распределенной обработке данных?	ПК-6
155	Cluster manager распределяет ресурсы между приложениями и управляет запуском вычислительных процессов в кластере	Какую роль выполняет cluster manager в архитектуре Apache Spark?	ПК-6
156	Standalone, YARN, Mesos, Kubernetes	Какие режимы или менеджеры кластера могут использоваться для запуска Apache Spark?	ПК-6
157	SparkSession является основной точкой входа для работы с DataFrame, Spark SQL и источниками данных в современных Spark-приложениях	Для чего используется SparkSession в Apache Spark?	ПК-6
158	SparkContext представляет соединение с кластером и используется для создания RDD и управления	Что такое SparkContext в Apache Spark?	ПК-6

	выполнением Spark-приложения		
159	collect возвращает все данные на driver и может привести к нехватке памяти при работе с большими наборами данных	Почему операцию collect следует осторожно использовать при обработке больших данных в Spark?	ПК-6
160	take возвращает ограниченное число элементов, что безопаснее для предварительного просмотра данных, чем загрузка всего набора на driver	Для чего используется операция take в Apache Spark?	ПК-6
161	count запускает вычисления и возвращает количество строк или элементов в распределенном наборе данных	Что делает операция count в Apache Spark?	ПК-6
162	distinct удаляет дублирующиеся записи, но может приводить к shuffle при распределенной обработке	Для чего используется операция distinct и почему она может быть затратной на больших данных?	ПК-6
163	dropDuplicates удаляет повторяющиеся строки полностью или по выбранным столбцам DataFrame	Для чего применяется метод dropDuplicates в Spark DataFrame API?	ПК-6
164	withColumn используется для создания нового столбца или изменения существующего столбца в DataFrame	Для чего применяется метод withColumn в Spark DataFrame API?	ПК-6
165	orderBy и sort выполняют сортировку данных и могут быть затратными из-за необходимости перераспределения данных между узлами	Почему сортировка больших данных в Spark может требовать значительных вычислительных ресурсов?	ПК-6
166	Агрегация объединяет данные по заданным группам и вычисляет итоговые показатели, например сумму, среднее, минимум или максимум	Что такое агрегация данных в задачах аналитической обработки больших данных?	ОПК-2
167	Window functions позволяют выполнять вычисления по группам строк с учетом порядка, не сворачивая	Для чего применяются оконные функции при анализе больших данных?	ОПК-2

	результат до одной строки на группу		
168	Ranking functions используются для присвоения порядкового номера или ранга объектам внутри группы	Для чего применяются функции ранжирования в аналитических запросах?	ОПК-2
169	Time series analysis используется для выявления закономерностей, трендов, сезонности и аномалий в данных, упорядоченных во времени	Для чего применяется анализ временных рядов в Big Data Analysis and Mining?	ОПК-2
170	Агрегация по временным интервалам позволяет анализировать события за минуты, часы, дни или другие периоды и выявлять динамику изменения показателей	Для чего используется группировка событий по временным интервалам при анализе потоковых и исторических данных?	ОПК-2
171	Логирование позволяет фиксировать ход выполнения обработки, ошибки, предупреждения и служебную информацию для последующего анализа	Для чего используется логирование в задачах обработки больших данных?	ПК-6
172	Оркестрация пайплайнов обеспечивает управление порядком выполнения задач, зависимостями, расписанием запусков и обработкой ошибок	Что означает оркестрация пайплайнов обработки данных?	ОПК-2
173	Apache Airflow используется для планирования, запуска и контроля выполнения связанных задач обработки данных	Для чего может применяться Apache Airflow в проектах Big Data?	ПК-6
174	DAG в Airflow описывает направленный ациклический граф задач и зависимостей между ними	Что представляет собой DAG при оркестрации пайплайнов данных?	ПК-6
175	Инкрементальная обработка обрабатывает только новые или измененные данные, снижая вычислительные затраты	В чем преимущество инкрементальной обработки данных по сравнению с полной переобработкой всего набора?	ОПК-2

176	Полная переобработка пересчитывает весь набор данных заново, а инкрементальная обработка учитывает только изменения с момента последнего запуска	В чем различие между полной и инкрементальной обработкой данных?	ОПК-2
177	Change Data Capture фиксирует изменения в источнике данных и позволяет передавать только новые, измененные или удаленные записи	Что такое Change Data Capture в архитектуре обработки данных?	ОПК-2
178	Data ingestion представляет собой процесс получения данных из источников и загрузки их в систему хранения или обработки	Что означает термин data ingestion в системах Big Data?	ОПК-2
179	ELT предполагает сначала загрузку данных в хранилище, а затем их преобразование внутри аналитической платформы	В чем состоит отличие ELT от классического ETL-подхода?	ОПК-2
180	Data validation позволяет проверить корректность, полноту, типы, диапазоны и согласованность данных до их дальнейшей обработки	Для чего выполняется валидация данных в пайплайнах Big Data?	ОПК-2
181	Schema-on-read означает, что структура данных интерпретируется в момент чтения, а не жестко задается при записи	Что означает принцип schema-on-read в хранилищах больших данных?	ОПК-2
182	Schema-on-write означает, что данные должны соответствовать заданной схеме уже на этапе записи в хранилище	Что означает принцип schema-on-write в аналитических хранилищах данных?	ОПК-2
183	Объектное хранилище используется для хранения больших объемов файлов и неструктурированных данных с доступом через API	Для чего применяются объектные хранилища в современных Big Data-архитектурах?	ОПК-2

184	Apache Iceberg, Delta Lake и Apache Hudi обеспечивают управление таблицами в lakehouse-архитектуре, поддержку транзакций, версионирования и эволюции схемы	Какие технологии могут использоваться для построения lakehouse-таблиц поверх больших хранилищ данных?	ПК-6
185	ACID-транзакции обеспечивают атомарность, согласованность, изолированность и долговечность операций с данными	Что означает поддержка ACID-транзакций в lakehouse-таблицах?	ОПК-2
186	Time travel позволяет обращаться к предыдущим версиям таблицы и воспроизводить состояние данных на определенный момент времени	Для чего используется механизм time travel в lakehouse-системах?	ОПК-2
187	Компактизация объединяет большое количество мелких файлов в более крупные, что повышает эффективность чтения и обработки данных	Для чего выполняется compaction в хранилищах больших данных?	ПК-6
188	Small files problem возникает при хранении большого количества мелких файлов, что увеличивает нагрузку на метаданные и снижает эффективность обработки	Что такое проблема мелких файлов в HDFS и lakehouse-хранилищах?	ПК-6
189	Z-ordering размещает связанные данные рядом по нескольким столбцам, что может ускорять фильтрацию и аналитические запросы	Для чего используется Z-ordering в системах хранения аналитических данных?	ПК-6
190	Data skipping позволяет пропускать файлы или блоки данных, которые не могут содержать нужные значения по условиям запроса	Что такое data skipping при выполнении аналитических запросов к большим данным?	ПК-6
191	Data governance определяет правила управления данными, ответственность,	Что означает управление данными в проектах Big Data?	ОПК-2

	доступ, качество, безопасность и жизненный цикл данных		
192	Data owner отвечает за смысл, качество и правила использования данных в предметной области	Какую роль выполняет владелец данных в системе управления большими данными?	ОПК-2
193	Data steward отвечает за сопровождение данных, контроль качества, описание метаданных и соблюдение правил работы с наборами данных	Какую роль выполняет data steward в процессах управления данными?	ОПК-2
194	Контроль доступа ограничивает операции с данными в соответствии с ролями пользователей и требованиями безопасности	Для чего используется разграничение прав доступа в Big Data-системах?	ПК-6
195	Маскирование данных скрывает чувствительные значения, сохраняя возможность безопасной обработки и анализа данных	Для чего применяется маскирование данных при подготовке аналитических наборов?	ОПК-2
196	Анонимизация удаляет или изменяет персональные признаки так, чтобы снизить риск идентификации конкретного человека	Что такое анонимизация данных в задачах анализа больших данных?	ОПК-2
197	Псевдонимизация заменяет идентификаторы искусственными значениями, при этом связь с исходными данными может быть восстановлена при наличии дополнительной информации	Чем псевдонимизация отличается от полной анонимизации данных?	ОПК-2
198	Data contract фиксирует структуру, формат, правила качества и ожидания потребителя относительно набора данных	Для чего используются контракты данных в современных аналитических системах?	ОПК-2
199	SLA определяет ожидаемый уровень	Что означает SLA применительно к пайплайнам и	ОПК-2

	сервиса, например допустимое время обработки, доступность данных или срок обновления витрины	аналитическим данным?	
200	SLO задает конкретные измеримые целевые показатели качества сервиса, например процент успешных запусков или максимальную задержку обновления данных	Что такое SLO в контексте эксплуатации Big Data-пайплайнов?	ОПК-2
201	Data observability позволяет отслеживать состояние данных, пайплайнов, задержки, сбои, пропуски и изменение распределений	Для чего используется наблюдаемость данных в Big Data-системах?	ПК-6
202	Freshness показывает, насколько данные актуальны и когда они были последний раз обновлены	Что означает показатель freshness при контроле качества данных?	ОПК-2
203	Completeness показывает долю заполненных и доступных значений в наборе данных	Что означает полнота данных при оценке их качества?	ОПК-2
204	Consistency показывает отсутствие противоречий между связанными данными, источниками и правилами предметной области	Что означает согласованность данных в системах Big Data?	ОПК-2
205	Uniqueness показывает отсутствие необоснованных дубликатов объектов или записей в наборе данных	Что означает уникальность данных при контроле качества набора?	ОПК-2
206	Validity показывает соответствие данных заданным форматам, типам, диапазонам и ограничениям	Что означает валидность данных при подготовке больших данных к анализу?	ОПК-2
207	Accuracy данных показывает степень соответствия значений реальному состоянию объекта или процесса	Что означает точность данных в задачах Big Data Analysis and Mining?	ОПК-2
208	Drift detection позволяет выявлять изменение распределения признаков,	Для чего используется обнаружение дрейфа данных и модели?	ОПК-2

	целевой переменной или качества модели во времени		
209	Alerting автоматически уведомляет ответственных специалистов о сбоях, нарушениях качества данных или отклонениях метрик пайплайна	Для чего используется механизм оповещений в эксплуатации Big Data-процессов?	ПК-6
210	Dashboard позволяет визуально контролировать состояние данных, выполнение пайплайнов, качество моделей и ключевые аналитические показатели	Для чего применяются информационные панели в проектах анализа больших данных?	ПК-6

2. Описание шкалы оценивания

Успеваемость студентов по каждой дисциплине оценивается в ходе текущего контроля и промежуточной аттестации.

3. Критерии оценивания компетенций*

3.1. Оценка компетенции ОПК-2

Компетенция ОПК-2 направлена на формирование способности создавать, исследовать и совершенствовать математические модели, концепции, теории и методы, применяемые в задачах анализа больших данных и интеллектуального извлечения закономерностей из данных. В рамках дисциплины Big Data Analysis and Mining компетенция формируется через изучение математических и алгоритмических основ обработки больших данных, методов классификации, регрессии, кластеризации, поиска ассоциативных правил, выявления аномалий, анализа потоковых данных и оценки качества моделей. Особое внимание уделяется исследованию применимости моделей к большим, разнородным и динамически изменяющимся наборам данных.

Критерии достижения ОПК-2 включают:

- знание основных математических моделей и методов, применяемых в интеллектуальном анализе больших данных;
- понимание принципов построения моделей классификации, регрессии, кластеризации, поиска закономерностей и обнаружения аномалий;
- умение выбирать математическую модель с учетом структуры, объема, качества и особенностей исходных данных;
- способность исследовать свойства модели, оценивать ее качество, устойчивость, интерпретируемость и ограничения;
- умение сравнивать альтернативные методы анализа данных и обосновывать выбор наиболее подходящего подхода;

- способность совершенствовать существующие методы обработки и анализа больших данных с учетом требований масштабируемости, воспроизводимости и вычислительной эффективности;
- навык интерпретации результатов моделирования и формулирования обоснованных выводов о применимости модели к исследовательским и прикладным задачам.

Компетенция оценивается через тестовые задания, ситуационные задачи, письменные ответы, анализ прикладных сценариев Big Data Analysis and Mining, интерпретацию результатов интеллектуального анализа данных, обоснование выбора моделей и методов, а также выполнение заданий, связанных с исследованием и сравнением математических моделей на больших наборах данных.

3.2. Оценка компетенции ПК-6

Компетенция ПК-6 направлена на формирование способности использовать современные методы разработки и реализации алгоритмов математических моделей с применением языков программирования и пакетов прикладных программ моделирования. В рамках дисциплины Big Data Analysis and Mining компетенция формируется через изучение и практическое применение инструментов распределенной обработки и анализа больших данных, включая Hadoop, Apache Spark, PySpark, Spark SQL, Spark MLlib и другие современные программные средства, используемые для реализации алгоритмов обработки, моделирования и интеллектуального анализа данных.

Критерии достижения ПК-6 включают:

- знание современных языков программирования, программных библиотек и пакетов прикладных программ, применяемых для анализа больших данных и реализации математических моделей;
- понимание принципов разработки алгоритмов обработки больших данных с учетом распределенной архитектуры, параллельных вычислений, объема данных и вычислительных ограничений;
- умение реализовывать алгоритмы предварительной обработки данных, фильтрации, агрегации, соединения, трансформации признаков и построения аналитических наборов данных;
- способность применять Hadoop и Apache Spark для хранения, обработки и анализа больших данных в пакетном и потоковом режимах;
- умение использовать PySpark, Spark SQL и Spark MLlib для реализации конкретных алгоритмов классификации, регрессии, кластеризации, поиска ассоциативных правил, обнаружения аномалий и оценки качества моделей;
- способность проводить вычислительные эксперименты, настраивать параметры алгоритмов, анализировать результаты выполнения программ и выявлять причины ошибок или снижения качества модели;
- умение оценивать производительность, масштабируемость, воспроизводимость и вычислительную эффективность реализованных алгоритмов;
- способность обосновывать выбор языка программирования, программного пакета, библиотеки или вычислительной среды для решения конкретной задачи Big Data Analysis and Mining.

Компетенция оценивается через практические задания, тестовые вопросы, ситуационные задачи, разработку программных решений, анализ результатов выполнения алгоритмов, работу с Hadoop и Apache Spark, реализацию моделей с использованием PySpark, Spark SQL и Spark MLlib, а также письменное обоснование выбора методов, инструментов и параметров реализации алгоритмов математических моделей.

3.3. Оценка уровня сформированности

Уровень достижения компетенций оценивается по четырёхбалльной шкале:

- **2 (низкий)** — компетенция не сформирована, действия фрагментарны или некорректны;
- **3 (достаточный)** — компетенция сформирована на базовом уровне, имеются неточности;
- **4 (хороший)** — уверенное применение знаний и умений, допущены незначительные ошибки;
- **5 (высокий)** — компетенция сформирована в полном объёме, решения обоснованы, результаты устойчивы.

Компетенции ОПК-1 и ОПК-3 развиваются в тесной связке, обеспечивая подготовку обучающихся к решению реальных задач в области анализа данных и интеллектуальных технологий.