МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ
(МИНОБРНАУКИ РОССИИ)

ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ
ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ
«ДОНБАССКИЙ ГОСУДАРСТВЕННЫЙ ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ»
(ФГБОУ ВО «ДонГТУ»)

Факультет информационных технологий и автоматизации производственных процессов

Кафедра интеллектуальных систем и информационной безопасности

УТВЕРЖДАЮ
И.о. проректора по учебной работе
Д.В. Мулов

РАБОЧАЯ ПРОГРАММА ДИСЦИПЛИНЫ

Анализ больших данных
(наименование дисциплины)

09.04.01 Информатика и вычислительная техника
(код, наименование специальности)

Искусственный интеллект и цифровые двойники предприятий
(магистерская программа)

Квалификация магистр
(бакалавр/специалист/магистр)

Форма обучения очная
(очная, очно-заочная, заочная)

Алчевск, 2024

1 Цели и задачи изучения дисциплины

Цели дисциплины. Целью изучения дисциплины «Анализ больших данных» является изучение основных методов и моделей для анализа больших данных.

Задачи изучения дисциплины:

- приобретение студентами знаний о технологиях подготовки, хранения, обработки и анализа больших данных;
- приобретение практических навыков работы большими данными.

Дисциплина направлена на формирование профессиональной (ПК-4) компетенции выпускника.

2 Место дисциплины в структуре образовательной программы

Логико-структурный анализ дисциплины – курс входит обязательную часть БЛОКА 1 «Дисциплины (модули)» подготовки студентов по специальности 09.04.01 Информатика и вычислительная техника (Искусственный интеллект и цифровые двойники предприятий).

Дисциплина реализуется кафедрой интеллектуальных систем и информационной безопасности. Основывается на базе дисциплин: «Высшая математика», «Методы анализа данных», изученных обучающимися при прохождении подготовки по программе бакалавриата (специалитета).

Является основой для изучения следующих дисциплин: «Инжиниринг данных», «Выполнение и защита выпускной квалификационной работы».

Для изучения дисциплины необходимы компетенции, сформированные у студента для решения профессиональных задач деятельности, связанных с научно-исследовательской работы.

Курс является фундаментом для ориентации студентов в сфере научных исследований.

Общая трудоемкость освоения дисциплины составляет 5 зачетных единиц, 180 ак.ч. Программой дисциплины предусмотрены лекционные (36 ч.), практические (36 ч.) занятия и самостоятельная работа студента (108 ч.).

Дисциплина изучается на 1 курсе в 1 семестре. Форма промежуточной аттестации – экзамен.

3 Перечень результатов обучения по дисциплине, соотнесённых с планируемыми результатами освоения образовательной программы

Процесс изучения дисциплины «Анализ больших данных» направлен на формирование компетенции, представленной в таблице 1.

Таблица 1 –Компетенции, обязательные к освоению

Содержание компетенции	Код компетенции	Код и наименование индикатора достижения компетенции
Способен руководить проектами по созданию комплексных систем на основе аналитики больших данных в различных отраслях	ПК-4	ПК-4.2. Применяет варианты использования больших данных, определений, словарей и эталонной архитектуры больших данных при руководстве проектами по построению комплексных систем на основе аналитики больших данных в различных отраслях.

4 Объём и виды занятий по дисциплине

Общая трудоёмкость учебной дисциплины составляет 5 зачётных единиц, 180 ак.ч.

Самостоятельная работа студента (СРС) включает проработку материалов лекций, подготовку к практическим занятиям, текущему контролю, выполнение индивидуального задания, самостоятельное изучение материала, выполнение курсовой работы и подготовку к экзамену.

При организации внеаудиторной самостоятельной работы по данной дисциплине используются формы и распределение бюджета времени на СРС для очной формы обучения в соответствии с таблицей 2.

Таблица 2 – Распределение бюджета времени на СРС

Вид учебной работы	Всего ак.ч.	Ак.ч. по семестрам
		1
Аудиторная работа, в том числе:	72	72
Лекции (Л)	36	36
Практические занятия (ПЗ)	-	-
Лабораторные работы (ЛР)	36	36
Курсовая работа/курсовой проект	-	-
Самостоятельная работа студентов (СРС), в том числе:	108	108
Подготовка к лекциям	9	9
Подготовка к лабораторным работам	18	18
Подготовка к практическим занятиям / семинарам	-	-
Выполнение курсовой работы / проекта	20	20
Расчетно-графическая работа (РГР)	-	-
Реферат (индивидуальное задание)	-	-
Домашнее задание	-	-
Подготовка к контрольным работам	-	-
Подготовка к коллоквиуму	-	-
Аналитический информационный поиск	10	10
Работа в библиотеке	15	15
Подготовка к экзамену (диф.зачету)	36	36
Промежуточная аттестация – экзамен (Э), дифзачет (Д/з)	Э, Д/з	Э, Д/з
Общая трудоемкость дисциплины		
ак.ч.	180	180
з.е.	5	5

5 Содержание дисциплины

С целью освоения компетенции, приведенной в п.3 дисциплина разбита на 6 тем:

– тема 1 (Введение в большие данные. Классификация задач анализа данных);

– тема 2 (Жизненный цикл аналитики данных);

– тема 3 (Высокопроизводительные вычисления);

– тема 4 (Масштабирование и многоуровневое хранение данных);

– тема 5 (Визуализация данных и результатов анализа);

– тема 6 (Сложные методы аналитики).

Виды занятий по дисциплине и распределение аудиторных часов для очной формы приведены в таблице 3.

В таблице 4 приведено распределение видов занятий и распределение аудиторных часов для выполнения курсовой работы.

Таблица 3 – Виды занятий по дисциплине и распределение аудиторных часов (очная форма обучения)

№ п/п	Наименование темы (раздела) дисциплины	Содержание лекционных занятий	Трудоемкость в ак.ч.	Темы практических занятий	Трудоемкость в ак.ч.	Тема лабораторных занятий	Трудоемкость в ак.ч.
1	2	3	4	5	6	7	8
1	Введение в большие данные. Задачи анализа данных. Жизненный цикл аналитики данных	Перегрузка информацией и Data Mining. Типы закономерностей. Модели вместо законов Системы и модели. Модели информационно-развивающихся систем. Виды знаний и способы их представления. Классы систем Data Mining	2 2	–	–	Подбор наборов больших данных для анализа	4
2	Основы языка Python	Особенности языка. Синтаксис и базовые конструкции. Структуры данных. Модули. Ветвления и циклы. Элементы функционального программирования. Объектно-ориентированное программирование. Средства визуализации, библиотека matplotlib.	6	–	–	Основы Python	6

Продолжение таблицы 3

1	2	3	4	5	6	7	8
3	Высокопроизводительные вычисления	Классификация архитектур вычислительных систем (по числу потоков команд и данных). Архитектурные свойства высокопроизводительных ВС. Показатели эффективности структуры ВС. Типовые структуры ВС. Вычислительные кластеры (computer cluster).	4	–	–	Библиотеки Python для анализа больших данных	6
4	Масштабирование и многоуровневое хранение данных	NoSQL. Масштабируемость. Репликация. CAP – теорема. Основы NoSQL. СУБД, поддерживающие NoSQL.	8	–	–	Изучение технологии NoSQL	8
5	Визуализация данных и результатов анализа	Области использования визуализации. Типы и задачи визуализации. Требования к визуализации. Традиционные виды визуализации. Графики и диаграммы. Инфографика. Презентация и анализ данных. Интерактивный сторителлинг. Дашборды и бизнес аналитика. Визуализация в медицине и науке. Карты и картограммы. Облако тегов. Кластерграмма. Исторический поток. Пространственный поток. Известные решения в области визуализации.	8	–	-	Визуализация результатов обработки больших данных	6

Окончание таблицы 3

1	2	3	4	5	6	7	8
6	Сложные методы аналитики	Методы матричного анализа. Оптимизация. Вероятность. Основные вероятностные формулы. Закон арксинуса. Математическая статистика как некорректная обратная задача теории вероятностей. Многомерный нормальный закон. Генерация случайных чисел. Метод наименьших квадратов в линейной модели измерений. Множественный регрессионный анализ. Главные компоненты и факторный анализ. Дискриминантный анализ. Анализ канонических корреляций. Дискриминантные информанты и классификация. Оценка вероятностей ошибочной классификации. Классификация на основе линейных дискриминантных форм. Кластеризация. Выбор метрики. Метод k средних. Метод опорных векторов.	6	–	-	Глубокий анализ больших данных	6
Всего аудиторных часов			36	-		36	

6 Фонд оценочных средств для проведения текущего контроля успеваемости и промежуточной аттестации студентов по дисциплине

6.1 Критерии оценивания

В соответствии с Положением о кредитно-модульной системе организации образовательного процесса ФГБОУ ВО «ДонГТУ» (https://www.dstu.education/images/structure/license_certificate/polog_kred_modul.pdf) при оценивании сформированности компетенций по дисциплине используется 100-балльная шкала.

Перечень компетенций по дисциплине и способы оценивания знаний приведены в таблице 5.

Таблица 5 – Перечень компетенций по дисциплине и способы оценивания знаний

Код и наименование компетенции	Способ оценивания	Оценочное средство
ПК-4	Экзамен Дифференцированный зачет	Комплект контролирующих материалов для экзамена и дифзачета

Всего по текущей работе в семестре студент может набрать 100 баллов, в том числе:

– лабораторные работы – всего 100 баллов.

Экзаменационная оценка проставляется автоматически, если студент набрал в течении семестра не менее 60 баллов и отчитался за каждую контрольную точку. Минимальное количество баллов по каждому из видов текущей работы составляет 60% от максимального.

Экзамен по дисциплине «Анализ больших данных» проводится по результатам работы в семестре. В случае, если полученная в семестре сумма баллов не устраивает студента, во время сессии студент имеет право повысить итоговую оценку либо в форме устного собеседования по приведенным ниже вопросам (п.п. 6.5), либо в результате тестирования.

Шкала оценивания знаний при проведении промежуточной аттестации приведена в таблице 6.

Таблица 6 – Шкала оценивания знаний

Сумма баллов за все виды учебной деятельности	Оценка по национальной шкале зачёт/экзамен
0-59	Не зачтено/неудовлетворительно
60-73	Зачтено/удовлетворительно
74-89	Зачтено/хорошо
90-100	Зачтено/отлично

6.2 Домашнее задание

Домашнее задание не предусмотрено.

6.3 Темы для рефератов (презентаций) – индивидуальное задание

Реферат (индивидуальное задание) не предусмотрен.

6.4 Оценочные средства для самостоятельной работы и текущего контроля успеваемости

Тема 1 Введение в большие данные. Классификация задач анализа данных

1) Изложите краткую историю больших данных.

2) Опишите подходы в работе с данными.

3) Какова этика работы с данными?

4) Опишите прогресс нейронных сетей.

5) В чем связь анализа данных и искусственного интеллекта?

Тема 2 Жизненный цикл аналитики данных

1) Опишите принципы обработки больших данных.

2) Хранение и обработка структурированных данных. Базовые принципы.

3) Каковы базовые принципы использования инструментов для обработки структурированных данных: Excel и SQL?

4) Как осуществляется хранение и обработка неструктурированных данных?

5) Какие Вы знаете технологии для обработки неструктурированных данных?

6) Как производится кодировка категориальных данных?.

Тема 3 Высокопроизводительные вычисления

1) Что такое Business Intelligence и Data Science?

2) Каковы базовые принципы, общее, отличия технологий для обработки больших данных: Python, Apache Hadoop, Apache Spark, Apache Storm.

4) Какие Вы знаете нейросетевые архитектуры?

5) Что такое машинное обучение?

Тема 4 Масштабирование и многоуровневое хранение данных

1) Что такое майнинг данных?

2) Что такое Deep Learning?

3) Что такое нейросетевые архитектуры?

4) Что такое обучение с подкреплением?

5) Что такое сверточные нейронные сети?

Тема 5 Визуализация данных и результатов анализа

1) Как используется анализ больших данных в бизнесе?

2) Как используется анализ больших данных в медицинской информатике?

3) Как используется анализ больших данных в физике элементарных частиц?

4) Как используется анализ больших данных в цифровых изданиях и семантической разметке?

5) Как используется анализ больших данных в компьютерном зрении?

Тема 6 Сложные методы аналитики

1) Что такое линейная регрессия: определение, формулы?

2) Что такое логистическая регрессия: определение, формулы?

3) Что такое решающие деревья: определение, схема?

4) Как используется глубокое обучение в обработке текстов?

5) Как используется машинное обучение в лингвистике?

6.5 Вопросы для подготовки к экзамену

1) Что означает термин «Big Data» в информационных технологиях?
2) Что является основной целью обработки Big Data?
3) Кто и в каком году впервые ввел термин «Big Data»?
4) Какие главные характеристики Big Data?
5) Какие данные занимают больше мировой памяти относительно остальных?
6) Какие понятия содержит в себе принцип трех "V"?
7) Что является примером квази-структурированных данных?
8) Чем характеризуются "Большие данные"?
9) Что является главным результатом процесса Business Intelligence?
10) Что означает термин «Business Intelligence» в информационных технологиях?
11) Расшифруйте аббревиатуру OLAP.

12) Что относится к средствам предоставления информации в Business Intelligence?

13) Что относится к средствам интеграции в «Business Intelligence»?

14) Какие цели ставит перед собой Data Science?

15) Что такое жизненный цикл аналитики данных?

16) Дайте определение термину «предиктивное моделирование»?

17) Что такое ETL?

18) Какова роль BI-аналитика в проекте?

19) Что такое Apache Hadoop? В чем преимущества решений на базе Hadoop?

20) Что такое MapReduce? Какими достоинствами и недостатками обладает MapReduce?

21) Какому основному принципу следует HDFS? Какой размер блока по умолчанию в HDFS? Какие функции выполняет NameNode в HDFS?

22) Какой узел отвечает за репликацию данных в Hadoop? Какие компоненты содержит Slave узел в Hadoop? Какие компоненты содержит Master узел в Hadoop?

23) Какие компоненты являются частями HDFS?

24) Для чего используется автономный режим Hadoop?

25) Какой режим необходим для того, чтобы на локальной машине использовать Hadoop как кластер, состоящий из одного узла?

26) Что является отличительной особенностью NoSQL? В каком случае стоит применять NoSQL хранилища?

27) Что, согласно теореме CAP, возможно обеспечить в любой реализации распределённых вычислений?

28) Какое свойство означает, что транзакции не нарушают согласованность данных, то есть они переводят базу данных из одного корректного состояния в другое?

29) Какой способ хранения данных используется в MongoDB?

30) Что относится к плюсам репликации?

31) Что относится к преимуществам нереляционных БД?

32) Что такое шардинг?

33) Какие три свойства фигурируют в определении теоремы CAP?

34) Для чего нужна визуализация? В чем состоят основные задачи визуализации?

35) Какие традиционные виды визуализации Вы знаете?

36) Что такое дедупликация данных?

37) Какие требования предъявляются к визуализации? Какие типы визуализации Вы знаете?

38) Чем анализ больших данных отличается от традиционного анализа?

39) Какие основные типы Data Mining Вы знаете?

40) Какие категории Web Mining можно выделить? В чем основная задача Web Content Mining?

41) В чем состоят основные задачи интеллектуального анализа текстов?

42) К каким алгоритмам классификации относится метод ближайших соседей?

43) Что является целью кластеризации?

44) С помощью какого алгоритма можно найти ассоциативное правило?

45) Что подразумевается под определением "статистический вывод"?

46) Чем отличаются ошибки первого и второго рода?

47) Что является результатом решения задачи регрессии?

48) Что такое α-error?

6.6 Примерные темы курсовых работ

1) Методы сбора и обработки данных предприятия/производственного процесса.
2) Анализ технико-экономических данных предприятия.
3) Прогнозные модели многомерных показателей предприятия на основе методов искусственного интеллекта.
4) Применение машинного обучения в настройке цифровых двойников технологических процессов производства.
5) Интеллектуальная система прогнозирования износа оборудования.
6) Интеллектуальная система планирования текущего и профилактического ремонта оборудования.
7) Интеллектуальная система бизнес-аналитики для потребностей технолога предприятия.
8) Интеллектуальные модели для промышленного интернета вещей.
9) Советующая система поддержки принятия решений технолога.
10) Интеллектуальная система оптимизации режима работы производственного оборудования.
11) Система компьютерного зрения для контроля технологического процесса промышленного предприятия.

12) Интеллектуальные методы визуализации анализа качества продукции предприятия.

13) Разработка цифровых двойников участников образовательного процесса в ВУЗе.

14) Модуль интеллектуального семантического анализа текста

15) Интеллектуальные методы и модели кластеризации больших данных.

Для каждой темы студент получает уточняющее задание: вид отрасли промышленности или конкретное предприятие. Числовые данные студент находит самостоятельно, используя открытые зарубежные и отечественные источники данных.

ЗАДАНИЕ

Создать программу для анализа больших данных с выводом результатов в виде приложения с многооконным интерфейсом. Особенности реализации:

1) Данные загружаются из файла (*.txt или *.xls или *.xml).

2) Полноценный интерфейс: главное меню, панель инструментов, строка состояния, несколько окон:

а) ввод и редактирование данных (табличная форма), с возможностью сохранения измененных значений в новом файле;

б) отчет-таблица + текст;

в) отчет-график + текст;

г) о программе.

3) Исходные данные и результаты сохранять в базе данных

4) Отчеты сохранять в виде файла *.html (текст, графики, таблицы) или *.pdf.

5) Комментарии в программе обязательны!

По завершении работы предоставляется объяснительная записка в объеме 30-40 страниц в *печатном и электронном виде* в формате Word 2003, оформленная согласно ГОСТ, а также рабочая программа (*все* файлы проекта и скомпилированный exe-*файл*).

Формулы в записке набирать в редакторе формул.

Сканированные рисунки не допускаются.

Ссылки на литературу обязательны. Количество источников – 3-5!

При использовании Python обязательным является создание оболочки, например, средствами Qt.

Структура пояснительной записки

Раздел	Приблизительное количество страниц
Титульный лист	1
Задание и календарный план	1
РЕФЕРАТ	1
ВВЕДЕНИЕ	1-2
1 АНАЛИЗ ЗАДАЧИ	
1.1 Теоретические сведения, необходимые для выполнения работы	5-6
1.2 Описание исходного набора данных	2-3
1.3 Описание переменных и констант, функций, используемых библиотек и программ	6-8
2 РАЗРАБОТКА СТРУКТУРЫ ПРОГРАММЫ	
2.1 Разработка структуры программы и алгоритмов расчетов	3-5
2.2 Разработка интерфейса пользователя	3-5
2.3 Разработка алгоритмов расчетов	5-7
3 РАЗРАБОТКА ПРОГРАММЫ	
3.1 Создание функций, реализующих алгоритмы расчетов	3-4
3.2 Реализация программного кода управления программой	2-3
3.3 Тестирование программы	1-2
ВЫВОДЫ	1
ПРИЛОЖЕНИЕ А ИСХОДНЫЕ ДАННЫЕ	1-2
ПРИЛОЖЕНИЕ Б ТЕХНИЧЕСКОЕ ЗАДАНИЕ	3-4
ПРИЛОЖЕНИЕ В ФОРМЫ ПРОГРАММЫ	2-4
ПРИЛОЖЕНИЕ Г ПРОГРАММНЫЙ КОД	15-20
ПРИЛОЖЕНИЕ Д ТЕСТОВЫЙ ПРИМЕР	2-4
ВСЕГО	60-80

Примечания:

1. При использовании сторонних библиотек (не входящих стандартные пакеты) привести их полное описание в отдельном приложении (Приложение Е).

2. При описании программы руководствоваться ГОСТ 19.402-78 «Описание программы».

7. Учебно-методическое и информационное обеспечение дисциплины

7.1 Рекомендуемая литература

Основная литература

1. Келлехер, Д. Наука о данных: базовый курс : учебное пособие / Д. Келлехер. – Москва : Альпина Паблишер, 2020. – 224 с. URL: https://ugolok.vercel.app/books/ai_ds_bd/dzhon_kelleher_brendan_tirni_nauka_o_dannih_bazovii_kurs_alipina.pdf (Дата обращения 26.08.2024).

2. Уэс Маккинни. Python и анализ данных: Первичная обработка данных с применением pandas, NumPy и Jupiter / пер. с англ. А. А. Слинкина. 3-е изд. – М.: МК Пресс, 2023. – 536 с.: ил. URL: gstu.by/sites/default/files/files/resources/2021/09/makkini.pdf?ysclid=m0dp0a8sbk386511381 (Дата обращения 26.08.2024).

Дополнительная литература

1. Белов, В.С. Информационно-аналитические системы: основы проектирования и применения: учебно-практическое пособие / В.С. Белов. − Москва: Евразийский открытый институт, 2010. – 111с. URL: https://shpora1.do.am/_ld/2/255_-_.pdf (Дата обращения 26.08.2024).

Учебно-методические материалы и пособия

1. Бизянов, Е.Е. Методы анализа данных: лабораторный практикум / Е.Е. Бизянов, А.С. Закутный ; Каф. Специализированных компьютерных систем. – Алчевск: ГОУ ВО ЛНР ДонГТИ, 2023. – 91 с. URL: https://library.dstu.education/download.php?rec=132246.

7.2 Базы данных, электронно-библиотечные системы, информационно-справочные и поисковые системы

1. Научная библиотека ДонГТУ : официальный сайт.— Алчевск. — URL: library.dstu.education.— Текст : электронный.
2. Научно-техническая библиотека БГТУ им. Шухова : официальный сайт. — Белгород. — URL: http://ntb.bstu.ru/jirbis2/.— Текст : электронный.
3. Консультант студента : электронно-библиотечная система.— Москва. — URL: http://www.studentlibrary.ru/cgi-bin/mb4x.— Текст : электронный.
4. Сайт кафедры ИСИБ http://scs.dstu.education

8 Материально-техническое обеспечение дисциплины

Материально-техническая база обеспечивает проведение всех видов деятельности в процессе обучения, соответствует требованиям ФГОС ВО.

Материально-техническое обеспечение представлено в таблице 9.

Таблица 9 – Материально-техническое обеспечение

Наименование оборудованных учебных кабинетов	Адрес (местоположение) учебных кабинетов
Специальные помещения: *Мультимедийная аудитория. (60 посадочных мест)*, оборудованная специализированной (учебной) мебелью (скамья учебная –20 шт., стол– 1 шт., доска аудиторная– 1 шт.), учебное ПК (монитор + системный блок), мультимедийная стойка с оборудованием – 1 шт., широкоформатный экран. Аудитории для проведения лекций: *Компьютерные классы (22 посадочных места),* оборудованный учебной мебелью, компьютерами с неограниченным доступом к сети Интернет, включая доступ к ЭБС: ПК– 12 шт.; Доска – 1 шт.	ауд. *207* корп. *4* ауд. *211* корп. *4*

Лист согласования рабочей программы дисциплины

Разработал
И.о. заведующего кафедрой
интеллектуальных систем и
информационной безопасности — Е.Е. Бизянов
(должность) (подпись) (Ф.И.О.)

И.о. заведующего кафедрой
интеллектуальных систем и
информационной безопасности — Е.Е. Бизянов
(наименование кафедры) (подпись) (Ф.И.О.)

Протокол № 1 заседания кафедры
интеллектуальных систем и
информационной безопасности от 27.08.20 24 г.

И.о. декана факультета
информационных технологий и
автоматизации производственных
процессов — В.В Дьячкова
(подпись) Ф.И.О.)

Согласовано

Председатель методической комиссии по
направлению подготовки 09.04.01
Информатика и вычислительная техника — Е.Е. Бизянов
(подпись) (Ф.И.О.)

Начальник учебно-методического центра — О.А.Коваленко
(подпись Ф.И.О.)

Лист изменений и дополнений

<table>
<tr><td colspan="2">Номер изменения, дата внесения изменения, номер страницы для внесения изменений</td></tr>
<tr><td>ДО ВНЕСЕНИЯ ИЗМЕНЕНИЙ:</td><td>ПОСЛЕ ВНЕСЕНИЯ ИЗМЕНЕНИЙ:</td></tr>
<tr><td colspan="2">Основание:</td></tr>
<tr><td colspan="2">Подпись лица, ответственного за внесение изменений</td></tr>
</table>