Федеральное государственное автономное образовательное учреждение высшего образования
«Российский университет дружбы народов»

Факультет физико-математических и естественных наук

Рекомендовано МССН
09.00.00 «Информатика и вычислительная техника»

РАБОЧАЯ ПРОГРАММА ДИСЦИПЛИНЫ

Наименование дисциплины

Современные проблемы анализа больших данных

Рекомендуется для направления подготовки

09.06.01 — Информатика и вычислительная техника

Направленность программ (профилей)

«Теоретические основы информатики»

Квалификация (степень) выпускника

Исследователь. Преподаватель-исследователь.

1. Цели и задачи дисциплины:

Цель курса: формирование у аспирантов профиля «Теоретические основы информатики» универсальных и профессиональных компетенций, на основе понимания методов работы с большими данными

К основным задачам изучения дисциплины относятся:

- овладение теоретическими знаниями и практическими навыками в области современных методов работы с большими данными;
- овладение способами исследования в области больших данных;
- применение соответствующих алгоритмов в процессе разработки информационно-вычислительных систем, предназначенных для работы с большими данными.

2. Место дисциплины в структуре ОП ВО.

Дисциплина относится к вариативной части блока 1 «Дисциплины (модули)», дисциплины по выбору.

В таблице № 1 приведены предшествующие и последующие дисциплины, направленные на формирование компетенций дисциплины в соответствии с матрицей компетенций ОП ВО.

№ п/п	Шифр и наименование компетенции	Предшествующие дисциплины	Последующие дисциплины (группы дисциплин)
Универсальные компетенции			
	УК-1	История и философия науки Методология научных исследований	Научные исследования, Научно-исследовательская практика, Подготовка к сдаче и сдача государственного экзамена, Представление научного доклада об основных результатах подготовленной научно-квалификационной работы (диссертации)
Общепрофессиональные компетенции			
	ОПК-1, ОПК-2, ОПК-3, ОПК-5	История и философия науки Методология научных исследований	Научно-исследовательская практика, Научные исследования, Подготовка к сдаче и сдача государственного экзамена, Представление научного доклада об основных результатах подготовленной научно-квалификационной работы (диссертации)
Профессиональные компетенции (вид профессиональной деятельности_- научно-исследовательский)			
	ПК-1	Методология научных исследований	Научно-исследовательская практика Научные исследования Представление научного доклада об основных результатах подготовленной научно-квалификационной работы (диссертации)
Профессионально-специализированные компетенции специализации_____			
	-	-	-

УК-1 — способность к критическому анализу и оценке современных научных достижений, генерированию новых идей при решении исследовательских и практических

задач, в том числе в междисциплинарных областях;

ОПК-1 — владеть методологией теоретических и экспериментальных исследований в области профессиональной деятельности;

ОПК-2 – владеть культурой научного исследования, в том числе с использованием современных информационно-коммуникационных технологий;

ОПК-3 – способность к разработке новых методов исследования и их применению в самостоятельной научно-исследовательской деятельности в области профессиональной деятельности;

ОПК-5 – способность объективно оценивать результаты исследований и разработок, выполненных другими специалистами и в других научных учреждениях ;

ПК-1 — способность самостоятельно проводить научные исследования в области теоретических основ информатики, применять полученные результаты в научных исследованиях и других областях.

3. Требования к результатам освоения дисциплины:

Процесс изучения дисциплины направлен на формирование следующих компетенций: УК-1; ОПК-1; ОПК-2; ОПК-3; ОПК-5; ПК-1

В результате освоения дисциплины аспирант должен:

Знать основные принципы обработки больших массивов данных, способы их представления и хранения.

Уметь формулировать задачи анализа больших данных; выбирать адекватные алгоритмы их решения.

Владеть технологиями разработки алгоритмов и программными системами анализа больших данных.

4. Объем дисциплины и виды учебной работы

Общая трудоемкость дисциплины составляет _____4_____ зачетных единиц.

Вид учебной работы	Всего часов	Очн.ф.о.	Заочн.ф.о.
		Семестры	курс
		3	2
Аудиторные занятия (всего)	60	60	60
В том числе:	-	-	
Лекции	20	20	20
Практические занятия (ПЗ)	40	40	40
Семинары (С)			
Лабораторные работы (ЛР)			
Самостоятельная работа (всего)	84	84	84
Общая трудоемкость час	144	144	144
зач. ед.	4	4	4

5. Содержание дисциплины

5.1. Содержание разделов дисциплины

№ п/п	Наименование раздела дисциплины	Содержание раздела
1	Введение в интеллектуальный анализ данных и большие данные	Тема 1. Интеллектуальный анализ данных. Большие данные. Способы масштабирования анализа больших данных. Наборы больших данных. Числовые и категориальные признаки. Основные этапы

		интеллектуального анализа больших данных. Потоковая передача данных из источников. Предварительная обработка данных. Очистка данных. Пропущенные значения. Зашумленные данные. Нормализация данных. Стохастическое обучение. Пакетный градиентный спуск. Стохастический градиентный спуск (SGD). Определение параметров алгоритма SGD.
2	Обучение без учителя	Тема 1. Методы машинного обучения без учителя. Снижение размерности данных при помощи алгоритма PCA. Кластеризация больших данных при помощи алгоритма K-средних. Допущения алгоритма. Подбор оптимальной величины K. Масштабирование алгоритма K-средних. Алгоритм LDA и его масштабирование.
3	Метод опорных векторов	Тема 1. Метод опорных векторов (SVM). Гиперплоскости. Разделяющая гиперплоскость. Маржа и опорные вектора. Кусочно-линейная функция потерь и ее варианты. Реализация SVM для больших данных на основе SGD. Отбор признаков посредством регуляризации. Добавление нелинейности в алгоритм SGD. Доводка гиперпараметров SGD.
4	Деревья классификации и регрессии	Тема 1. Обучение дерева решений. Агрегация выборок. Случайный лес и экстремально рандомизированный лес. Экстремально рандомизированные деревья и большие наборы данных. Алгоритм CART и бустинг. Алгоритм XGBoost. Регрессия на основе XGBoost. Потоковая передача больших наборов данных посредством XGBoost. Стохастический градиентный бустинг и сеточный поиск.
5.	Глубокое обучение с большими данными	Тема 1. Искусственные нейронные сети. Архитектура нейронной сети. Параллелизация в нейронных сетях. Регуляризация в нейронных сетях. Гиперпараметрическая оптимизация в нейронных сетях. Глубокое обучение с большими данными. Сеточный поиск. Автокодировщики. Тема 2. Глубокое обучение с библиотекой TensorFlow. Операции TensorFlow. Инкрементное глубокое обучение с большими данными. Сверточные нейронные сети (CNN) в TensorFlow. Сверточный слой. Объединяющий слой. Полносвязный слой. Обучение сети CNN при помощи инкрементной тренировки. Вычисления на GPU.
6.	Интеллектуальный анализ данных в рас-пределенных вычислительных средах	Тема 1. Распределенная вычислительная среда Hadoop. Архитектура Hadoop. Распределенная файловая система HDFS. Вычислительная парадигма MapReduce. Тема 2. Интеллектуальный анализ данных на платформе Spark. Распространение переменных по узлам кластера. Предобработка данных в среде Spark. Машинное обучение с платформой Spark. Библиотека pySpark.

5.2 Разделы дисциплин и виды занятий
Для очной формы обучения

№ п/п	Наименование раздела дисциплины	Лекц.	Практ. и лаб. зан.	СРС	Все-го час.
1	Введение в интеллектуальный анализ данных и большие данные	2	4	8	**14**
2	Обучение без учителя	4	8	16	**28**
3	Метод опорных векторов	4	8	16	**28**
4	Деревья классификации и регрессии	4	8	12	**24**
5	Глубокое обучение с большими данными	2	4	16	**22**
6	Интеллектуальный анализ данных в распределенных вычислительных средах	4	8	16	**28**
	ИТОГО	**20**	**40**	**84**	**144**

Для заочной формы обучения

№ п/п	Наименование раздела дисциплины	Лекц.	Практ. и лаб. зан.	СРС	Все-го час.
1	Введение в интеллектуальный анализ данных и большие данные	2	4	8	**14**
2	Обучение без учителя	4	8	16	**28**
3	Метод опорных векторов	4	8	16	**28**
4	Деревья классификации и регрессии	4	8	12	**24**
5	Глубокое обучение с большими данными	2	4	16	**22**
6	Интеллектуальный анализ данных в распределенных вычислительных средах	4	8	16	**28**
	ИТОГО	**20**	**40**	**84**	**144**

6-7. Лабораторные и практические занятия

№ п/п	№ раздела дисциплины	Тематика практических занятий	Трудоемкость (час.)	
			Очн. ф.о.	Заочн. ф.о.
1.	1	1. Расчет статистических показателей и визуализация заданного набора данных	4	4
2.	2	1. Решение задачи кластеризации данных	8	8

		при помощи алгоритма K-средних.		
3.	3	1. Решение задачи классификации данных при помощи метода опорных векторов.	8	8
4.	4	1. Решение задачи классификации данных при помощи обучения дерева решений.	8	8
5.	5	1. Решение задачи классификации данных при помощи байесовской классификации и классификации по ближайшим соседям. 2. Решение задач классификации и прогнозирования данных при помощи регресионного анализа.	4	4
6.	6.	1. Поиск ассоциативных правил при помощи алгоритмов Apriori, Eclat, Declat, FPGrowth 2. Использование метода главных компонент для снижения размерности данных.	8	8

8. Материально-техническое обеспечение дисциплины:

Мультимедийная учебная аудитория для проведения лекционных занятий. Компьютерные (дисплейные) классы с доступом к сети Интернет и электронно-образовательной среде Университета для выполнения обучающимися лабораторных работ по дисциплине, для проведения обучающимися самостоятельной работы и компьютерного тестирования обучающихся (при необходимости).

9. Информационное обеспечение дисциплины.

a.) программное обеспечение:

ОС Linux, офисный пакет LibreOffice (лицензия MPL-2.0), ПО для просмотра pdf (например, evince (лицензия GPL-2+ CC-BY-SA-3.0)), GNU Midnight Commander (Лицензия GNU GPL 3), редакторы emacs (лицензия GPL) или vi (лицензия BSD), FreeFem++ (Лицензия LGPL-2.1), TeXLive (Лицензия GPL-2 LPPL-1.3c TeX), Sagemath (Лицензия GPLv3), система компьютерной алгебры MAXIMA (лицензия GPL-2 GPL-2+), SciLab (Лицензия CeCILL (свободная, совместимая с GNU GPL v2)).

b.) Базы данных, информационно-справочные и поисковые системы

1. ТУИС http://esystem.pfur.ru
2. Сайт библиотеки РУДН http://lib.rudn.ru/
3. Электронная библиотека РГБ http://www.rsl.ru/
4. Общероссийский математический портал mathnet.ru
5. NIST Цифровая энциклопедия математических функций (https://dlmf.nist.gov)
6. Старейший ресурс по численным методом в сети Numerical recipes (http://numerical.recipes/)
7. Библиографическая и реферативная база данных и инструмент для отслеживания цитируемости статей, опубликованных в научных изданиях:

- Scopus (https://www.scopus.com).
- Web of Science (http://www.isiknowledge. Com)
- Zentralblatt MATH (zbMATH) (https://zbmath.org)

c.) Облачные сервисы:

- CoCalc (https://cocalc.com) - веб-платформа для облачных вычислений и управления курсами для вычислительной математики, является частью проекта Sage, поддерживает редактирование рабочих листов Sage, документов LaTeX и блокнотов Jupyter, открывает доступ к экспериментам в консоли Linux (Ubuntu 18.04.2 LTS).

- ShareLaTeX (https://ru.sharelatex.com) - онлайн редактор LaTeX, не требует установки, поддерживает совместную работу в реальном времени.
- WolframAlpha (https://www.wolframalpha.com) — онлайн система компьютерной алгебры и база знаний
- Math Partner (http://mathpar.cloud.unihub.ru/ru) — язык и веб-платформа для облачных вычислений, разработанный группой Г.И. Малошонка.

10. Учебно-методическое обеспечение дисциплины:

а) основная литература

1. Data mining // [Электронный ресурс] URL: https://www.intuit.ru/studies/courses/6/6/info, режим доступа: свободный.

б) дополнительная литература

1. Введение в аналитику больших массивов данных // [Электронный ресурс] URL: https://www.intuit.ru/studies/courses/12385/1181/info, режим доступа: свободный.

11. Методические указания для обучающихся по освоению дисциплины.

Учебным планом на изучение дисциплины отводится 1 семестр. В течение семестра выполняются практические работы, домашние задания и проводятся контрольные мероприятия. В конце семестра производится итоговый контроль знаний – зачет с оценкой.

12. Фонд оценочных средств для проведения промежуточной аттестации обучающихся по дисциплине (модулю).

ФОС по дисциплине представлен в приложении к данной программе.

Программа составлена в соответствии с требованиями ОС ВО РУДН.

Разработчики:

к.ф.-м.н., доцент кафедры информационных технологий — М.Б. Фомин

Заведующий кафедрой
информационных технологий, д.ф.-м.н. — Ю.Н. Орлов

Директор направления
Заведующий кафедрой
прикладной информатики и теории вероятностей, д.т.н., проф. — К.Е. Самуйлов

Федеральное государственное автономное образовательное учреждение высшего образования
«Российский университет дружбы народов»

Факультет физико-математических и естественных наук

Кафедра прикладной информатики и теории вероятностей

ФОНД ОЦЕНОЧНЫХ СРЕДСТВ

ПО УЧЕБНОЙ ДИСЦИПЛИНЕ

Программное обеспечение для проведения научных исследований

(наименование дисциплины)

09.06.01 Информатика и вычислительная техника
(код и наименование направления подготовки)

Современные проблемы анализа больших данных
(наименование профиля подготовки)

Исследователь. Преподаватель-исследователь
Квалификация (степень) выпускника

Паспорт фонда оценочных средств по дисциплине Современные проблемы анализа больших данных
название

Направление: 09.06.01 Информатика и вычислительная техника. Профиль «Теоретические основы информатики»
шифр название

Код контролируемой компетенции или ее части	Контролируемый раздел дисциплины	Контролируемая тема дисциплины	ФОСы (формы контроля уровня освоения ООП)			Баллы темы	Баллы раздела
			Аудиторная работа	СРС			
			Выполнение ЛР	Выполнение ДЗ	Экзамен/Зачет		
УК-1; ОПК-1; ОПК-2; ОПК-3; ОПК-5; ПК-1	Раздел 1: Введение в интеллектуальный анализ данных и большие данные	Тема 1. Интеллектуальный анализ данных. Большие данные. Способы масштабирования анализа больших данных. Наборы больших данных. Числовые и категориальные признаки. Основные этапы интеллектуального анализа больших данных. Потоковая передача данных из источников. Предварительная обработка данных. Очистка данных. Пропущенные значения. Зашумленные данные. Нормализация данных. Стохастическое обучение. Пакетный градиентный спуск. Стохастический градиентный спуск (SGD). Определение параметров алгоритма SGD.	10		2	12	12
	Раздел 2: Обучение без учителя	Тема 1. Методы машинного обучения без учителя. Снижение размерности данных при помощи алгоритма PCA. Кластеризация больших данных при помощи алгоритма K-средних. Допущения алгоритма. Подбор оптимальной величины K. Масштабирование алгоритма K-средних. Алгоритм LDA и его масштабирование.	10		2	12	12
	Раздел 3: Метод опорных векторов	Тема 1. Метод опорных векторов (SVM). Гиперплоскости. Разделяющая гиперплоскость. Маржа и опорные вектора. Кусочно-линейная функция потерь и ее варианты. Реализация SVM для больших данных на основе SGD. Отбор признаков посредством регуляризации. Добавление нелинейности в алгоритм SGD. Доводка гиперпараметров SGD.	10		2	12	12
	Раздел 4: Деревья классификации и	Тема 1. Обучение дерева решений. Агрегация выборок. Случайный лес и экстремально рандомизированный лес. Экстремально	10		2	12	12

	регрессии	рандомизированные деревья и большие наборы данных. Алгоритм CART и бустинг. Алгоритм XGBoost. Регрессия на основе XGBoost. Потоковая передача больших наборов данных посредством XGBoost. Стохастический градиентный бустинг и сеточный поиск.					
	Раздел 5: Глубокое обучение с большими данными	Тема 1. Искусственные нейронные сети. Архитектура нейронной сети. Параллелизация в нейронных сетях. Регуляризация в нейронных сетях. Гиперпараметрическая оптимизация в нейронных сетях. Глубокое обучение с большими данными. Сеточный поиск. Автокодировщики.	10		3	13	52
		Тема 2. Глубокое обучение с библиотекой TensorFlow. Операции TensorFlow. Инкрементное глубокое обучение с большими данными. Сверточные нейронные сети (CNN) в TensorFlow. Сверточный слой. Объединяющий слой. Полносвязный слой. Обучение сети CNN при помощи инкрементной тренировки. Вычисления на GPU.	10		3	13	
	Раздел 6: Интеллектуальный анализ данных в распределенных вычислительных средах	Тема 1. Распределенная вычислительная среда Hadoop. Архитектура Hadoop. Распределенная файловая система HDFS. Вычислительная парадигма MapReduce.	10		3	13	52
		Тема 2. Интеллектуальный анализ данных на платформе Spark. Распространение переменных по узлам кластера. Предобработка данных в среде Spark. Машинное обучение с платформой Spark. Библиотека pySpark.	10		3	13	
		ИТОГО:	80		20	100	100

Процесс изучения дисциплины направлен на формирование следующих компетенций УК-1; ОПК-1; ОПК-2; ОПК-3; ОПК-5; ПК-1

(в соответствии с ОС ВО РУДН)

УК-1 — способность к критическому анализу и оценке современных научных достижений, генерированию новых идей при решении исследовательских и практических задач, в том числе в междисциплинарных областях;

ОПК-1 — владеть методологией теоретических и экспериментальных исследований в области профессиональной деятельности;

ОПК-2 – владеть культурой научного исследования, в том числе с использованием современных информационно-коммуникационных технологий;

ОПК-3 – способность к разработке новых методов исследования и их применению в самостоятельной научно-исследовательской деятельности в области профессиональной деятельности;

ОПК-5 – способность объективно оценивать результаты исследований и разработок, выполненных другими специалистами и в других научных учреждениях ;

ПК-1 — способность самостоятельно проводить научные исследования в области теоретических основ информатики, применять полученные результаты в научных исследованиях и других областях.

Примерный перечень оценочных средств

п/п	Наименование оценочного средства	Краткая характеристика оценочного средства	Представление оценочного средства в фонде
		Аудиторная работа	
1	Лабораторная работа	Система практических заданий, направленных на формирование практических навыков у обучающихся	Фонд практических заданий
2	Тест *	Система стандартизированных заданий (вопросов), позволяющая автоматизировать процедуру измерения уровня знаний и умений обучающегося.	База тестовых заданий
3	Опрос *	Средство контроля, организованное как специальная беседа преподавателя с обучающимся на темы, связанные с изучаемой дисциплиной, и рассчитанное на выяснение объема знаний обучающегося по определенному разделу или теме.	Вопросы по темам/разделам дисциплины
4	Экзамен *	Оценка работы обучающегося в течение семестра (года, всего срока обучения и др.) и призван выявить уровень, прочность и систематичность полученных им теоретических и практических знаний, приобретения навыков самостоятельной работы, развития творческого мышления, умение синтезировать полученные знания и применять их в решении практических задач.	Примеры заданий/вопросов, пример экзаменационного билета
		Самостоятельная работа	
1	Подготовка отчетов по результатам выполнения лабораторных работ	Форма проверки качества выполнения обучающимися лабораторных работ в соответствии с утвержденной программой.	Фонд практических заданий

Оценивание результатов освоения дисциплины производится в соответствии с балльно-рейтинговой системой.

Балльно-рейтинговая система оценки уровня знаний

Сводная оценочная таблица дисциплины

Контролируемый раздел дисциплины	**Контролируемая тема дисциплины**	**ФОСы (формы контроля уровня освоения ООП)**			**Баллы темы**	**Баллы раздела**
		Аудиторная работа		**Самостоятельная работа**		
		Выполнение ЛР	**Опрос**	**Проект**		
Математическое моделирование	Основные принципы математического моделирования.	0	5	0	5	10
	Использование специализированного ПО в научной работе. Свободное программное обеспечение.	0	5		5	
Специализированное программное обеспечение для научных исследований	ПО для решения задач линейной алгебры.	5	0	10	35	35
	ПО для решения систем нелинейных уравнений	5	0			
	ПО для исследования динамических систем.	10	0			
	ПО для решения задач механики сплошных тел и математической физики	5	0			
Оформление результатов научных исследований	Набор и верстка научных работ в издательской системе LaTeX.	5		10	30	30
	Подготовка презентаций в издательской системе LaTeX	5				
	Подготовка графического контента, представляющего результаты научной работы	5				
	Набор и верстка диссертации и автореферата	5				
Научные базы данных	Общая методика библиографического поиска	5	0	10	25	25
	Математические ресурсы в сети Интернет	0	10			
	Итого:	**50**	**20**	**30**	**100**	**100**

Таблица соответствия баллов и оценок

Баллы БРС	Традиционные оценки РФ	Оценки ECTS
95 - 100	5	A
86 - 94		B
69 - 85	4	C
61 - 68	3	D
51 - 60		E
31 - 50	2	FX
0 - 30		F
51-100	Зачет	Passed

Правила применения БРС

1. Раздел (тема) учебной дисциплины считаются освоенными, если обучающийся набрал более 50 % от возможного числа баллов по этому разделу (теме).
2. Обучающийся не может быть аттестован по дисциплине, если он не освоил все темы и разделы дисциплины, указанные в сводной оценочной таблице дисциплины.
3. По решению преподавателя и с согласия обучающегося, не освоивших отдельные разделы (темы) изучаемой дисциплины, в течение учебного семестра могут быть повторно проведены мероприятия текущего контроля успеваемости или выданы дополнительные учебные задания по этим темам или разделам. При этом обучающимся за данную работу засчитывается минимально возможный положительный балл (51 % от максимального балла).
4. При выполнении обучающимся дополнительных учебных заданий или повторного прохождения мероприятий текущего контроля полученные им баллы засчитываются за конкретные темы. Итоговая сумма баллов не может превышать максимального количества баллов, установленного по данным темам (в соответствии с приказом Ректора № 564 от 20.06.2013). По решению преподавателя предыдущие баллы, полученные обучающимся по учебным заданиям, могут быть аннулированы.
5. График проведения мероприятий текущего контроля успеваемости формируется в соответствии с календарным планом курса. Обучающиеся обязаны сдавать все задания в сроки, установленные преподавателем.
6. Время, которое отводится обучающемуся на выполнение мероприятий текущего контроля успеваемости, устанавливается преподавателем. По завершение отведенного времени обучающийся должен сдать работу преподавателю, вне зависимости от того, завершена она или нет.
7. Использование источников (в том числе конспектов лекций и лабораторных работ) во время выполнения контрольных мероприятий возможно только с разрешения преподавателя.
8. Отсрочка в прохождении мероприятий текущего контроля успеваемости считается уважительной только в случае болезни обучающегося, что подтверждается наличием у него медицинской справки, заверенной круглой печатью в поликлинике № 25, предоставляемой преподавателю не позднее двух недель после выздоровления. В этом случае выполнение контрольных

мероприятий осуществляется после выздоровления обучающегося в срок, назначенный преподавателем. В противном случае, отсутствие обучающегося на контрольном мероприятии признается не уважительным.

9. Обучающийся допускается к итоговому контролю знаний с любым количеством баллов, набранных в семестре.
10. Если в итоге за семестр аспирант получил менее 51 балла, то аспиранту разрешается добор необходимого (до 51) количества баллов путем повторного одноразового выполнения предусмотренных контрольных мероприятий, при этом по усмотрению преподавателя аннулируются соответствующие предыдущие результаты. Ликвидация задолженностей проводится в период теоретического обучения в сроки по согласованию с деканатом.

Критерии оценки по дисциплине

95-100 баллов:

- своевременное выполнение практических (лабораторных) работ, предоставление проекта, выполненных на высоком профессиональном уровне;
- успешная защита проекта, продемонстрировавшая глубокое понимание затронутых в нем вопросов;

86- 94 балла:

- своевременное выполнение практических (лабораторных) работ, предоставление проекта, выполненных на хорошем профессиональном уровне;
- успешная защита проекта, продемонстрировавшая глубокое понимание затронутых в нем вопросов;

69-85 баллов:

- своевременное выполнение практических (лабораторных) работ, предоставление проекта, выполненных на приемлемом профессиональном уровне;
- успешная защита проекта, продемонстрировавшая хорошее понимание наиболее важных из затронутых в нем вопросов.

51-68 баллов:

- своевременное выполнение практических (лабораторных) работ, предоставление проекта, выполненных на удовлетворительном уровне;
- успешная защита проекта, продемонстрировавшая удовлетворительное понимание наиболее важных из затронутых в нем вопросов.

31 - 50 баллов – НЕ ЗАЧТЕНО:

- несвоевременное выполнение практических (лабораторных) работ, проекта;
- защита проекта, продемонстрировавшая неудовлетворительное понимание наиболее важных из затронутых в нем вопросов.

0-30 баллов, НЕ ЗАЧТЕНО:

- несвоевременное предоставление проекта;
- игнорирование занятий по дисциплине по неуважительной причине.

Аудиторные занятия

Аудиторные занятия проводятся в форме лекций, опросов по пройденному материалу и обсуждений лабораторных работ. Оценивается освоение пройденного материала и умение его применять при выполнении практических заданий.

Темы и содержание лабораторных работ

Лабораторная работа № 1. Расчет статистических показателей и визуализация заданного набора данных.
Задание:

- Дано математическое ожидание **a** и корреляционная матрица **R** двумерного гауссовского распределения.
- Постройте **n** значений случайных признаков **X** и **Y**, имеющих двумерное гауссовское распределение с математическим ожиданием **a** и корреляционной матрицей **R**.
- Визуализируйте построенный набор данных на плоскости в виде набора точек с координатами $\{(x_i, y_i)\}$, $i=1,\ldots,n$.
- Вычислите и выведите на экран для построенных данных математические ожидания, дисперсии, а также корреляцию между данными.
- Считайте статистические данные для двух признаков из заданного набора данных репозитария UCI. Если в записи значение какого-либо из признаков не определено (символ “?”), то следует пропустить данную запись.
- Изобразите считанные из набора данные в виде точек на плоскости.
- Вычислите и выведите на экран для двух признаков математические ожидания, дисперсии, а также корреляцию между признаками.
- Подготовьте отчет о выполнении заданий лабораторной работы.

Лабораторная работа № 2. Решение задачи кластеризации данных при помощи алгоритма K-средних.
Задание:

- Для двух классов объектов даны значения количеств объектов (точек) для каждого класса ($\mathbf{n_1}$ и $\mathbf{n_2}$), значения векторов математических ожиданий для каждого класса ($\mathbf{a_1}$ и $\mathbf{a_2}$) и корреляционные матрицы для каждого класса ($\mathbf{R_1}$ и $\mathbf{R_2}$) для моделируемой выборки из гауссовских случайных векторов.
- Для каждого из классов постройте значения случайных признаков **X** и **Y**, имеющие двумерное гауссовское распределение с математическим ожиданием $\mathbf{a_i}$ и корреляционной матрицей $\mathbf{R_i}$.
- Изобразите построенные данные на плоскости в виде точек с координатами $\{(x_i, y_i)\}$, $i=1,\ldots,n$ и раскрасьте их разными цветами (красным и синим) для разных классов.
- Проведите кластеризацию построенных объектов с помощью алгоритма **k** средних для случая, когда количество кластеров равно двум.
- Изобразите на плоскости кластеризованные объекты разными цветами для точек разных кластеров и центры кластеров.
- Найдите для построенной кластеризации таблицу сопряженности (contingency table) и вычислите показатель чистоты кластеризации.
- Считайте статистические данные для двух признаков и класса из заданного набора данных репозитария UCI. Если в записи значение какого-либо из признаков или класса не определено (символ “?”), то следует пропустить данную запись.

- Изобразите считанные из набора данные в виде точек на плоскости и раскрасьте их разными цветами для разных классов.
- Проведите кластеризацию построенных объектов с помощью алгоритма **k** средних для случая, когда количество кластеров равно количеству классов.
- Изобразите на плоскости кластеризованные объекты разными цветами для точек разных кластеров и центры кластеров.
- Найдите для построенной кластеризации таблицу сопряженности (contingency table) и вычислите показатель чистоты кластеризации.
- Подготовьте отчет о выполнении заданий лабораторной работы.

Лабораторная работа № 3. Решение задачи классификации данных при помощи метода опорных векторов.
Задание:

- Считайте данные (два признака и метки) из заданного набора данных репозитария UCI.
- Разбейте метки на два класса (положительный и отрицательный) и подготовьте набор данных для обучения бинарного классификатора.
- При необходимости масштабируйте значения признаков при помощи StandardScaler.
- Случайным образом разделите считанные из набора данные на обучающую выборку и контрольную выборку в соотношении 80% на 20%.
- Изобразите обучающую выборку на плоскости в виде точек с координатами (x_i, y_i) с использованием разных цветов для положительного и отрицательного класса.
- Обучите бинарный классификатор метода опорных векторов LinearSVC на обучающей выборке.
- Произведите классификацию объектов контрольной выборки с помощью обученного классификатора LinearSVC.
- Изобразите объекты контрольной выборки на плоскости разными цветами для положительного и отрицательного классов (на одном рисунке) и для произведенной классификации (на другом рисунке).
- Для проведенной бинарной классификации постройте ROC-кривую и выведите на рисунке показатель AUC ROC (площадь под кривой).
- Подготовьте отчет о выполнении заданий лабораторной работы.

Лабораторная работа № 4. Решение задачи классификации данных при помощи обучения дерева решений.
Задание:

- Считайте статистические данные для двух признаков и класса из заданного набора данных репозитария UCI. Если в какой-либо записи значение какого-либо из признаков или класса не определено (символ “?”), то следует пропустить данную запись.
- Случайным образом разделите считанные из набора данные на обучающую выборку и контрольную выборку в соотношении 70% на 30%.
- Изобразите обучающую выборку на плоскости в виде точек с координатами координатами $\{(x_i, y_i)\}$, $i=1,\ldots,n$ и раскрасьте их разными цветами для разных классов.
- Обучите классификатор DecisionTreeClassifier на обучающей выборке, ограничивая глубину дерева значением 5.
- Выполните визуализацию полученного дерева решений.
- Произведите классификацию объектов контрольной выборки с помощью обученного классификатора DecisionTreeClassifier.

- Изобразите объекты контрольной выборки на плоскости разными цветами для исходных классов (на одном рисунке) и для произведенной классификации (на другом рисунке).
- Вычислите и выведите на экран показатель точности классификации.
- Подготовьте отчет о выполнении заданий лабораторной работы.

Лабораторная работа № 5. Решение задачи классификации данных при помощи байесовской классификации и классификации по ближайшим соседям.
Задание:

- Для двух классов объектов даны значения количеств объектов (точек) для каждого класса ($\mathbf{n_1}$ и $\mathbf{n_2}$), значения векторов математических ожиданий для каждого класса ($\mathbf{a_1}$ и $\mathbf{a_2}$) и корреляционные матрицы для каждого класса ($\mathbf{R_1}$ и $\mathbf{R_2}$) для моделируемой выборки из гауссовских случайных векторов.
- Для каждого из классов постройте значения случайных признаков **X** и **Y**, имеющие двумерное гауссовское распределение с математическим ожиданием $\mathbf{a_i}$ и корреляционной матрицей $\mathbf{R_i}$. Объедините построенные данные для двух классов в единый набор.
- Случайным образом разделите полученные данные на обучающую выборку и контрольную выборку в соотношении 80% на 20%.
- Изобразите обучающую выборку в трёхмерном пространстве в виде точек с координатами $\mathbf{\{(x_i, y_i, z_i)\}}$, $\mathbf{i=1,\ldots,n}$ и раскрасьте их разными цветами (например, красным и синим) для разных классов.
- Произведите классификацию объектов контрольной выборки, используя данные о классах объектов из обучающей выборки, с помощью алгоритма наивной байесовской классификации.
- Изобразите объекты контрольной выборки в трёхмерном пространстве разными цветами и маркерами для исходных классов (на одном рисунке) и для произведенной классификации (на другом рисунке).
- Вычислите и выведите на экран показатель точности классификации.
- Считайте статистические данные для трех признаков и класса из заданного набора данных репозитария UCI. Если в какой-либо записи значение какого-либо из признаков или класса не определено (символ “?”), то следует пропустить данную запись.
- Случайным образом разделите считанные из набора данные на обучающую выборку и контрольную выборку в соотношении 75% на 25%.
- Изобразите обучающую выборку в трёхмерном пространстве в виде точек с координатами координатами $\mathbf{\{(x_i, y_i, z_i)\}}$, $\mathbf{i=1,\ldots,n}$ и раскрасьте их разными цветами для разных классов.
- Произведите классификацию объектов контрольной выборки, используя данные о классах объектов из обучающей выборки, с помощью алгоритма **K** ближайших соседей для **K**, равного удвоенному количеству классов в наборе, но не менее 10.
- Изобразите объекты контрольной выборки в трёхмерном пространстве разными цветами для исходных классов (на одном рисунке) и для произведенной классификации (на другом рисунке).
- Вычислите и выведите на экран показатель точности классификации.
- Подготовьте отчет о выполнении заданий лабораторной работы.

Лабораторная работа № 6. Решение задач классификации и прогнозирования данных при помощи регресионного анализа.
Задание:

- Даны количество объектов (точек) **n**, параметры **a** и **b** и дисперсия гауссовского белого шума σ^2.
- Смоделируйте точки $\{(x_i, y_i)\}$, $i=1,\ldots,n$ согласно модели $y=ax+b+\varepsilon$, где в качестве ε используется гауссовский белый шум (нормально распределенная случайная величина) с нулевым математическим ожиданием и заданной дисперсией σ^2. Значения x_i выбираются через равные промежутки на отрезке [0;1].
- Визуализируйте на одном графике точки (x_i, y_i) и прямую $y=ax+b$ при $x \in [0, 1]$.
- Случайным образом разделите полученные данные на обучающую выборку и контрольную выборку в соотношении 75% на 25%.
- Постройте модель линейной регрессии $y=a'x+b'+\varepsilon$ на обучающей выборке.
- Выведите на экран полученные значения a', b' и сравните их с первоначальными значениями **a**, **b**.
- Визуализируйте на одном графике точки (x_i, y_i) из контрольной выборки и прямую $y = a'x+b'$ при $x \in [0, 1]$.
- Вычислите и выведите на экран показатели MSE, MAE и коэффициент детерминации.
- Считайте данные для независимой переменной (предиктора) и зависимой переменной из заданного набора данных репозитария UCI.
- Масштабируйте зависимую переменную на диапазон от 0.001 до 0.999
- Используйте для построения модели три подхода:
    - линейную регрессию
    - полиномиальную регрессию (степень полинома degree=2)
    - преобразование зависимой переменной при помощи логистической функции и последующего применения линейной регрессии
- Случайным образом разделите считанные из набора данные на обучающую выборку и контрольную выборку в соотношении 70% на 30%.
- Изобразите обучающую выборку на плоскости в виде точек с координатами (x_i, y_i).
- Постройте на обучающей выборке различные модели прогнозирования значений зависимой переменной.
- Визуализируйте на одном графике разными цветами точки (x_i, y_i) из контрольной выборки, а также точки с прогнозируемыми значениями зависимой переменной для трех моделей, соединенные линиями (для улучшения картинки может потребоваться сортировка точек контрольной выборки по возрастанию независимой переменной).
- Вычислите и сравните значения показателей MSE, MAE и коэффициента детерминации для различных моделей. Определите лучшую модель по показателю коэффициента детерминации.
- Подготовьте отчет о выполнении заданий лабораторной работы.

Лабораторная работа № 7. Поиск ассоциативных правил при помощи алгоритмов Apriori, Eclat, Declat, FPGrowth.
Задание:

- Скачайте заданный набор данных из репозитария UCI. Считайте из набора данных для последующего анализа только категориальные признаки, исключая числовые признаки.
- Проанализируйте набор данных и оставьте в наборе 10 категориальных, исключая признаки с неопределенными значениями и принимающие одно и то же значение для всех записей наборе данных.
- Преобразуйте записи набора данных в записи транзакционной базы данных (список) следующим образом:

    - в качестве первого элемента добавьте идентификатор транзакции (порядковый номер записи в наборе)
    - в качестве второго элемента добавьте список, составленный из значений признаков записи набора с учетом указанных ниже преобразований и отсортированный по возрастанию значений
    - если признак принимает логические (булевые) значения (True/False), то подставьте вместо значения True название признака, вместо значения False ничего подставлять не нужно, например, если признак имеет название age_gt_60, подставьте вместо этих значения t (True) значение age_gt_60
    - если признак принимает несколько вариантов значения, то подставьте вместо текущего значения конкатенацию названия признака и его текущего значения, например, если признак имеет название ar_c и возможные значения признака normal, elevated, absent, то подставьте вместо этих значений значения ar_c_normal, ar_c_elevated, ar_c_absent
- В качестве минимального значения поддержки примите значение, равное 1/4 количества записей в наборе данных (относительная поддержка 0.25).
- Используя алгоритм, указанный в Вашем варианте, найдите популярные наборы данных для указанного выше значения минимальной поддержки.
- Подготовьте отчет о выполнении заданий лабораторной работы.

Лабораторная работа № 8. Использование метода главных компонент для снижения размерности данных.
Задание:

- Проанализируйте заданный набор данных из репозитария UCI. Считайте из набора данных для последующего анализа только числовые признаки.
- Если в данных значение какого-либо из признаков не определено (символ “?”), то пропустите данную запись.
- Найдите 5 признаков, имеющих наибольшую дисперсию. Если числовых признаков меньше, чем 5, то используйте все числовые признаки.
- Для набора данных, состоящего из пяти признаков с наибольшей дисперсией, найдите размерность метода главных компонент, для которой доля объясняемой дисперсии будет не менее 95%.
- Пользуясь методом главных компонент, снизьте размерность набора данных до двух признаков и изобразите полученный набор данных в виде точек на плоскости. Отображайте точки различных классов разными цветами.
- Подготовьте отчет о выполнении заданий лабораторной работы.